BAYESIAN FUSION OF PROBABILISTIC FORECASTS

Dissertation von
Susanne Gabriele Trick

zur Erlangung des Grades
Doktor rerum naturalium
(Dr. rer. nat.)

Centre for Cognitive Science
Fachbereich Humanwissenschaften
Technische Universität Darmstadt

Erstgutachter: Prof. Constantin Rothkopf, PhD
Zweitgutachter: Prof. Dr. Dominik Endres

Darmstadt 2024

Trick, Susanne Gabriele: Bayesian Fusion of Probabilistic Forecasts
Darmstadt, Technische Universität Darmstadt
Jahr der Veröffentlichung der Dissertation auf TUprints: 2024
URN: urn:nbn:de:tuda-tuprints-273857
Tag der mündlichen Prüfung: 26.04.2024

In loving memory of my mom.
You told me I will make it.

ACKNOWLEDGMENTS

I am deeply grateful for all the support I received while working on this thesis.

First and foremost, I would like to thank my supervisor Prof. Constantin Rothkopf for giving me the opportunity to do a PhD in his lab. Thank you for giving me the freedom to work on the projects I like. Thank you for inspiring and motivating me since the start of my bachelor's studies, encouraging me in the lows, and celebrating the highs with me. Your intuition and advice were invaluable for the works presented in this thesis, and I have learned a lot from you.

I would also like to thank Prof. Dominik Endres for being my second reviewer. I am honored to have your expertise in evaluating my thesis. Many thanks also go to Prof. Frank Jäkel and Prof. Loes van Dam for being members of my thesis committee.

Prof. Frank Jäkel also deserves special thanks for super long, interesting, fruitful, and amusing talks, for inspiration and advice, and for enjoying the magic and pain of math together with me.

I want to thank Dr. Dorothea Koert, who supported me as a friend, colleague, and team leader. Thank you for believing in me, encouraging me, giving valuable advice, and having my back while I was writing my thesis. I am grateful for all the fun we had with robots, the milestone meetings we survived, for traveling the world with you, and for endless conversations through the nights.

Thanks to my favorite office neighbor Dr. Dirk Balfanz for organizing grants and projects, for providing wise advice for research and real life, for enjoyable talks, and for constantly encouraging me. Thanks to our secretary Inge Galinski for her all-round support in all organizational questions and for providing me with lovingly handmade knitwear.

I thank my thesis students and student assistants, in particular Lisa Scherf, Tabea Wilke, Franziska Herbert, Vilja Lott, Simon Kohaut, and Philip Wolff. Thank you for your trust in me and for supporting my research, it was a pleasure to work with you. Likewise, I would like to thank my colleagues in the PIP lab, in the IKIDA team, as well as in the other groups of the Centre for Cognitive Science for supporting me with research questions and for making the work more fun. Special thanks go to Dr. Nils Neupärtl, Vildan Salikutluk, Matthias Schultheis, Fabian Kessler, Tobias Thomas, Inga Ibs, and Lisa Scherf.

Beyond that, I am deeply grateful to Dr. Florian Kadner, the best colleague and best friend I could imagine. Thank you for always being just one phone call away, for cheering me up, for crying with me and for being happy with me, for making me laugh out loud, for giving me pinky promises when I was worried, and for supporting me unconditionally. Words cannot express how much our friendship means to me.

I am also very grateful to my friends and family, in particular my mom and dad, my sister Johanna, and my brother Tom, for emotional support, for celebrating the successes and overcoming the setbacks with me, for cheering me up when I was worried or frustrated, and for always encouraging and empowering me. Special thanks go to my dad for being a role-model as a researcher and a wise advisor. Finally, I am deeply grateful to my partner Daniel for loving me, for being my biggest fan and my resting place, and for being patient when I could not stop working in the evenings.

Research in this thesis was supported by the German Federal Ministry of Education and Research (BMBF) in the projects Kobo34 (Grant No. 16SV7984) and IKIDA (Grant No. 01IS20045).

ABSTRACT

Due to pervasive noise and ambiguity, our world is dominated by uncertainty. In order to face uncertain perception, action, and decision making, humans have an internal representation of their uncertainty and communicate it in interactions with other humans. Furthermore, humans combine information from different information sources to reduce their uncertainty about the world's state. Specifically in perceptual tasks people have been shown to integrate redundant sensory cues by weighting different cues according to their uncertainty to maximally reduce the uncertainty of the integrated sensory estimate, as described by Bayes' theorem.

Like humans, Artificial Intelligence (AI) systems and robots as their embodied form should represent, consider, reduce, and communicate uncertainty in order to cope with our uncertain world. In particular, this can increase the safety of critical AI applications and improve the quality of the interaction between human and AI, e.g., in AI-supported human decision making or human-robot collaboration in industry or caregiving settings. Quantifying an AI system's uncertainty can be realized with probabilistic methods, e.g., probabilistic classifiers that output categorical distributions over all classes. Also, AI systems often combine different information sources: Classifier ensembles, which combine multiple individual classifiers in order to improve classification performance, are known to be the most successful classification methods. However, classifier ensembles are usually optimized to merely maximize the classification performance instead of reducing the uncertainty. Thus, an open question is how to optimally combine probabilistic forecasts provided by classifiers while explicitly considering and correctly reducing their uncertainty, similar to how humans combine multiple cues in perception. Since the individual classifiers in an ensemble are usually correlated, particular focus should be put on the combination of correlated classifiers. This thesis investigates how to optimally combine the outputs of probabilistic classifiers. It provides a normative Bayesian model that formalizes how to optimally fuse individual classifiers according to their properties, such as uncertainty, bias, and variance, given different assumptions. Moreover, our model explicitly considers the correlation of the individual classifiers. It models the classifiers' correlations with a newly introduced probability distribution, the correlated Dirichlet distribution. The resulting Correlated Fusion Model quantifies how classification uncertainty should be reduced through Bayes optimal classifier combination depending on the individual classifiers' uncertainty, bias, variance, and correlation and outperforms related Bayesian classifier fusion models on simulated and real data sets.

A special case of the correlated Dirichlet distribution introduced for modeling correlated probabilistic classifiers is the bivariate beta distribution. The bivariate beta distribution models two beta-distributed random variables with a positive correlation. Thus, it is particularly interesting for modeling binary probabilistic classifiers but is also of general interest in statistics. While the bivariate beta distribution has been proposed before, previous work used an approximate and sometimes inaccurate method to compute the distribution's covariance and correlation and estimate its parameters. Therefore, in this thesis, we derive all product moments and the exact covariance and introduce an algorithm for estimating the bivariate beta distribution's parameters using moment matching.

A promising application of Bayes optimal fusion of multiple probabilistic classifiers is multimodal human-robot interaction. Since humans interact multimodally using modalities such as speech, gestures, and gaze directions, an intuitive and natural interaction between humans and robots requires robots to also interact multimodally. In particular, a robot should be able to process people's uncertain multimodal signals, e.g., about their intentions, and correctly combine its uncertainties about them. However, present approaches for multimodal intention recognition in human-robot interaction do not focus on how to correctly consider individual modalities' uncertainties and reduce uncertainty. Therefore, in this thesis, we recognize human intentions from multimodal data using probabilistic classifiers for each modality whose output distributions are combined Bayes optimally. We present three applications of Bayes optimal classifier fusion to different human-robot interaction scenarios. We first detect human intentions from multimodal data including speech, gestures, gaze directions, and scene objects. In an interaction task between a human and a 7-Degrees-of-Freedom robot arm, we show that adding more modalities contributes to increased detection performance and reduced uncertainty. Second, we apply Bayesian fusion to enable humans to teach a 7-Degrees-of-Freedom robot arm using multimodal action advice given by speech and gestures for interactive reinforcement learning. Evaluations with human participants show that the learning speed can be improved significantly compared to other methods. Third, we learn to detect humans' intention to start an interaction with a robot, the intention for interaction, from natural human behavior. We recorded a multimodal data set including speech and body poses with human participants in a collaborative task with a two-armed assistive robot. We compare different unimodal and multimodal classifiers and show that the intention for interaction can be detected better from multimodal data using Bayesian classifier fusion.

Bayes optimal fusion methods can not only be applied to combine classifiers but also to combine subjective probability estimates provided by humans. Such probabilistic estimates or forecasts, e.g., provided by experts, are of particular importance in many domains, such as finance, politics, engineering, meteorology, and public health, and can further be used to build rule-based AI systems. While combining forecasts is known to increase forecasting performance, as for classifier fusion, there is a need for a normative model that defines how to correctly combine human forecasts while explicitly considering their uncertainty. In this thesis, we present a family of normative Bayesian models for the aggregation of subjective probability estimates, which are closely related to our normative Bayesian model for classifier fusion. We model the forecasting behavior of individual forecasters with beta distributions, implicitly calibrate their probability estimates, and combine them accordingly in order to obtain the Bayes optimal uncertainty of the fused forecast. However, the proposed fusion models disregard the correlation between the forecasters, reduce too much uncertainty, and are thus overconfident. Therefore, in a second step, we extend these models to a Bayesian model for the combination of correlated subjective probability estimates. By explicitly representing the skills of the individual forecasters and the difficulties of individual queries for which forecasts are provided, this model can represent the correlation between individual forecasters and can consider it when fusing forecasts. As a consequence, its fusion performance is improved compared to our previous models and related fusion models.

In summary, this thesis investigates the fundamental computational problem of combining uncertain probabilistic forecasts that humans, robots, as well as AI systems in general are facing. While human perception unconsciously integrates multiple sensory cues to provide an optimal percept, here we develop normative Bayesian fusion models for combining probabilistic forecasts provided by classifiers or humans. The proposed models define how probabilistic forecasts should be fused Bayes optimally, in particular if the forecasts are correlated. We demonstrate that the developed algorithms outperform related fusion methods and successfully apply them in multimodal human-robot interaction.

ZUSAMMENFASSUNG

Aufgrund von allgegenwärtigem Rauschen und Ambiguitäten wird unsere Welt von Unsicherheit beherrscht. Um der Unsicherheit in Wahrnehmung, Handlungen und Entscheidungen begegnen zu können, haben Menschen eine interne Repräsentation ihrer Unsicherheit und kommunizieren diese in Interaktionen mit anderen Menschen. Darüber hinaus kombinieren Menschen Informationen aus verschiedenen Informationsquellen, um ihre Unsicherheit über den Zustand der Welt zu verringern. Insbesondere bei Wahrnehmungsaufgaben wurde gezeigt, dass Menschen redundante sensorische Hinweisreize integrieren, indem sie verschiedene Hinweisreize entsprechend ihrer Unsicherheit gewichten. So reduzieren sie die Unsicherheit der integrierten sensorischen Schätzung maximal, wie es der Satz von Bayes beschreibt.

Ebenso wie Menschen sollten auch Systeme der künstlichen Intelligenz (KI) und Roboter in ihrer verkörperten Form Unsicherheit repräsentieren, berücksichtigen, reduzieren und kommunizieren, um mit unserer unsicheren Welt zurechtzukommen. Dies kann insbesondere die Sicherheit kritischer KI-Anwendungen erhöhen und die Qualität der Interaktion zwischen Mensch und KI verbessern, z.B. bei KI-gestützter menschlicher Entscheidungsfindung oder Mensch-Roboter-Kollaboration in der Industrie oder im Pflegebereich. Die Quantifizierung der Unsicherheit eines KI-Systems kann mit probabilistischen Methoden erfolgen, z.B. mit probabilistischen Klassifizierern, die kategoriale Wahrscheinlichkeitsverteilungen über alle Klassen ausgeben. Außerdem kombinieren KI-Systeme oft verschiedene Informationsquellen: Ensembles aus Klassifizierern, die mehrere einzelne Klassifizierer kombinieren, um die Performance der Klassifizierung zu verbessern, gelten als die erfolgreichsten Klassifizierungsmethoden. Allerdings werden solche Ensembles in der Regel darauf optimiert, lediglich die Performance der Klassifizierung zu maximieren, anstatt die Unsicherheit zu reduzieren. Eine offene Frage ist daher, wie man probabilistische Vorhersagen von Klassifizierern optimal kombinieren kann, wobei deren Unsicherheit explizit berücksichtigt und korrekt reduziert werden soll, ähnlich wie der Mensch bei der Wahrnehmung mehrere Hinweisreize kombiniert. Da die einzelnen Klassifizierer in einem Ensemble in der Regel korreliert sind, sollte dabei besonderes Augenmerk auf die Kombination von korrelierten Klassifizierern gelegt werden. In dieser Arbeit wird untersucht, wie die Ausgaben probabilistischer Klassifizierer optimal kombiniert werden können. Es wird ein normatives Bayesianisches Modell vorgestellt, das formalisiert, wie einzelne Klassifizierer entsprechend ihrer Eigenschaften wie Unsicherheit, Bias und Varianz unter verschiedenen Annahmen optimal kombiniert werden können. Außerdem berücksichtigt das vorgeschlagene Modell ausdrücklich die Korrelation der einzelnen Klassifizierer, indem es deren Korrelationen mit einer neu eingeführten Wahrscheinlichkeitsverteilung modelliert, der korrelierten Dirichlet-Verteilung. Das daraus resultierende Fusionsmodell quantifiziert, wie die Klassifizierungsunsicherheit durch eine optimale Bayesianische Kombination von Klassifizierern in Abhängigkeit von deren Unsicherheit, Bias und Varianz und der Korrelation der einzelnen Klassifizierer reduziert werden sollte. Darüber hinaus übertrifft das vorgeschlagene Modell verwandte Bayesianische Fusionsmodelle auf simulierten und realen Datensätzen.

Ein Spezialfall der korrelierten Dirichlet-Verteilung, die zur Modellierung korrelierter probabilistischer Klassifizierer eingeführt wurde, ist die bivariate Beta-Verteilung. Die bivariate Beta-Verteilung modelliert zwei Beta-verteilte Zufallsvariablen mit einer positiven Korrelation. Sie ist daher besonders interessant für die Modellierung binärer probabilistischer Klassifizierer, ist aber ebenso von allgemeinem Interesse in der Statistik. Die bivariate Beta-Verteilung wurde bereits früher eingeführt. In früheren Arbeiten wurde jedoch eine ungefähre und mitunter ungenaue Methode verwendet, um die Kovarianz und Korrelation der Verteilung zu berechnen und ihre Parameter zu schätzen. In dieser Arbeit werden daher alle Produktmomente und die exakte Kovarianz hergeleitet. Basierend darauf wird ein Algorithmus zur Schätzung der Parameter der bivariaten Beta-Verteilung mit Hilfe von Moment-Matching vorgestellt.

Eine vielversprechende Anwendung der Bayes-optimalen Fusion mehrerer probabilistischer Klassifizierer ist die multimodale Mensch-Roboter-Interaktion. Da Menschen multimodal interagieren, indem sie Modalitäten wie Sprache, Gesten und Blickrichtungen verwenden, erfordert eine intuitive und natürliche Interaktion zwischen Menschen und Robotern, dass auch Roboter multimodal interagieren. Insbesondere sollte ein Roboter in der Lage sein, die unsicheren multimodalen Signale des Menschen, z.B. über seine Intentionen, zu verarbeiten und seine Unsicherheiten darüber korrekt zu kombinieren. Bisherige Ansätze zur multimodalen Intentionserkennung in der Mensch-Roboter-Interaktion konzentrieren sich jedoch nicht darauf, wie die Unsicherheiten der einzelnen Modalitäten korrekt berücksichtigt und die Unsicherheit reduziert werden kann. Daher werden in dieser Arbeit menschliche Intentionen aus multimodalen Daten erkannt, indem probabilistische Klassifizierer für die einzelnen Modalitäten gelernt und deren Ausgabeverteilungen optimal nach Bayes kombiniert werden. Es werden drei Anwendungen der Bayes-optimalen Fusion von Klassifizierern für verschiedene Mensch-Roboter-Interaktionsszenarien vorgestellt. Erstens werden menschliche Intentionen aus multimodalen Daten wie Sprache, Gesten, Blickrichtungen und Szenenobjekten erkannt. In einer Interaktionsaufgabe zwischen einem Menschen und einem Roboterarm mit sieben Freiheitsgraden wird gezeigt, dass das Hinzufügen weiterer Modalitäten zu einer verbesserten Erkennung von Intentionen und einer geringeren Unsicherheit beiträgt. Zweitens wird die Bayesianische Fusion angewandt, um es Menschen zu ermöglichen, einem Roboterarm mit sieben Freiheitsgraden multimodale Handlungsempfehlungen in Form von Sprache und Gesten für interaktives Verstärkungslernen zu geben. Auswertungen mit Versuchspersonen zeigen, dass die Lerngeschwindigkeit im Vergleich zu anderen Methoden signifikant verbessert werden kann. Drittens wird die Intention des Menschen, eine Interaktion mit einem Roboter zu beginnen, aus natürlichem menschlichem Verhalten erkannt. Ein multimodaler Datensatz wurde aufgezeichnet, der Sprache und Körperhaltungen von Versuchspersonen in einer kollaborativen Aufgabe mit einem zweiarmigen Assistenzroboter umfasst. Auf diesem Datensatz werden verschiedene unimodale und multimodale Klassifizierer verglichen und es wird gezeigt, dass die Intention zur Interaktion besser aus multimodalen Daten unter Verwendung der Bayesianischen Fusion von Klassifizierern erkannt werden kann.

Bayes-optimale Fusionsmethoden können nicht nur für die Kombination von Klassifizierern angewandt werden, sondern auch für die Kombination subjektiver Wahrscheinlichkeitsschätzungen, die von Menschen stammen. Solche probabilistischen Schätzungen oder Vorhersagen, z.B. von Experten, sind in vielen Bereichen von besonderer Bedeutung, unter anderem im Finanzwesen, in der Politik, im Ingenieurwesen, in der Meteorologie und im Gesundheitswesen, und können darüber hinaus zum Aufbau regelbasierter KI-Systeme verwendet werden. Obwohl bekannt ist, dass die Kombination einzelner Vorhersagen die Vorhersage-Performance erhöht, besteht wie bei der Fusion von Klassifizierern ein Bedarf an einem normativen Modell, das definiert, wie menschliche Vorhersagen unter expliziter Berücksichtigung ihrer Unsicherheit korrekt kombiniert werden sollen. In dieser Arbeit wird eine Familie von normativen Bayesianischen Modellen für die Kombination von subjektiven Wahrscheinlichkeitsschätzungen vorgestellt, die eng mit dem vorgestellten normativen Bayesianischen Modell für die Fusion von Klassifizierern verwandt sind. Das Vorhersageverhalten der einzelnen menschlichen Experten wird mit Beta-Verteilungen modelliert und ihre Wahrscheinlichkeitsschätzungen implizit kalibriert und entsprechend kombiniert, um die Bayes-optimale Unsicherheit der fusionierten Vorhersage zu erhalten. Die vorgeschlagenen Fusionsmodelle vernachlässigen jedoch die Korrelation zwischen den Experten und reduzieren daher zu viel Unsicherheit. Aus diesem Grund werden diese Modelle in einem zweiten Schritt zu einem Bayesianischen Modell für die Kombination von korrelierten subjektiven Wahrscheinlichkeitsschätzungen erweitert. Durch die explizite Repräsentation der Fähigkeiten der einzelnen Experten und der Schwierigkeiten der einzelnen Fragestellungen, für die Vorhersagen gemacht werden, kann dieses Modell die Korrelation zwischen den einzelnen Experten repräsentieren und bei der Fusion von Vorhersagen berücksichtigen. Dies verbessert die Performance der fusionierten Vorhersage im Vergleich zu unserem vorherigen Modell und anderen verwandten Fusionsmodellen.

Zusammenfassend untersucht diese Arbeit das grundlegende computationale Problem der Kombination unsicherer probabilistischer Vorhersagen, mit dem Menschen, Roboter und KI-Systeme im Allgemeinen konfrontiert sind. Während die menschliche Wahrnehmung unbewusst verschiedene sensorische Hinweisreize integriert, um ein optimales Ergebnis zu erhalten, werden hier normative Bayesianische Fusionsmodelle für die Kombination probabilistischer Vorhersagen entwickelt, wobei die Vorhersagen sowohl von Klassifizierern als auch von Menschen stammen können. Die vorgeschlagenen Modelle definieren, wie probabilistische Vorhersagen Bayes-optimal fusioniert werden sollten, insbesondere wenn die Vorhersagen korreliert sind. Außerdem wird gezeigt, dass die entwickelten Algorithmen verwandte Fusionsmethoden übertreffen und erfolgreich in der multimodalen Mensch-Roboter-Interaktion angewandt werden können.

CONTENTS

1 Introduction 1
1.1 Uncertainty in Classifier Ensembles 2
1.2 Bayesian Classifier Fusion for Multimodal Human-Robot Interaction . . . 3
1.3 Bayesian Combination of Subjective Probability Estimates 4
1.4 Outline 4
1.5 Contributions 5

2 Background 7
2.1 Uncertainty 7
2.2 Probability 9
2.3 Probability Theory 9
2.4 Bayesian Inference 11
2.5 Graphical Models as Generative Models 12
2.6 Parameter Estimation 15
2.7 Cue Integration as Parameter Estimation 17
2.8 Classifier Fusion as Parameter Estimation 21

3 Normative Bayesian Classifier Fusion 27
3.1 Related Work 28
3.2 Bayesian Models of Classifier Fusion 29
3.2.1 Independent Opinion Pool 29
3.2.2 Independent Fusion Model 31
3.2.3 Correlated Fusion Model 37
3.3 Evaluation 44
3.3.1 Simulated Data Sets 45
3.3.2 Real Data Sets 46
3.3.3 Comparison to the Model by Pirs and Strumbelj (2019) 49
3.4 Discussion and Conclusion 51
3.5 Appendix 53
3.5.1 Model Parameters Used for Evaluation 53

4 Parameter Estimation for a Bivariate Beta Distribution 59
4.1 Construction of a Bivariate Beta Distribution with Arbitrary Beta Marginals 62
4.2 Joint Distribution 64
4.3 Moments 64
4.4 Correlation and Covariance 66
4.5 Parameter Inference 68
4.5.1 Moment Matching 69
4.5.2 Simulation Study 70
4.5.3 Application on a Real Data Set of Correlated Forecasts 71
4.5.4 Relationship Between δ_1 and the Correlation 73
4.6 Generalization to the Correlated Dirichlet Distribution 74

4.7 Conclusion 75

5 Bayesian Fusion for Intention Recognition in Human-Robot Interaction 77
5.1 Related Work 78
5.2 Multimodal Intention Recognition 80
5.2.1 Classifier Fusion with Independent Opinion Pool 81
5.2.2 Classifiers for Single Modalities 82
5.3 Experimental Evaluation 85
5.4 Conclusions 90

6 Interactive Reinforcement Learning with Bayesian Fusion of Multimodal Advice 91
6.1 Related Work 93
6.2 Multimodal IOP-Based Advice For Interactive Reinforcement Learning . . 94
6.2.1 RL Module 96
6.2.2 Classifiers for Individual Modalities 96
6.2.3 Fusion Module 97
6.2.4 Action Selection Module 98
6.3 Experimental Evaluation 98
6.3.1 Advantages of Bayesian Fusion 98
6.3.2 Grid World 99
6.3.3 Pancake Scenario 101
6.4 Conclusion 104

7 Multimodal Detection of the Intention for Interaction in Human-Robot Interaction 107
7.1 Related Work 109
7.2 Task-Independent IFI Detection 110
7.2.1 Data Collection 110
7.2.2 Data Preprocessing 113
7.2.3 IFI Classification 115
7.3 Experimental Evaluation 117
7.3.1 IFI Data Set 117
7.3.2 Performance Measures 117
7.3.3 Classification Performances 118
7.3.4 Detailed Analysis of Human Behavior and IFI Models 120
7.4 Conclusion 121

8 A Normative Model for Bayesian Combination of Subjective Probability Estimates 123
8.1 Modeling Probability Estimates with Beta Distributions 127
8.1.1 Hierarchical Beta Fusion Model 128
8.1.2 Non-Hierarchical Beta Fusion Model 129
8.1.3 Beta Calibration 130
8.1.4 Hierarchical Symmetric Beta Fusion Model 131
8.1.5 Non-Hierarchical Symmetric Beta Fusion Model 132
8.2 Evaluation 133

8.2.1 Data Sets . . . 133
8.2.2 Cross-Validation . . . 134
8.2.3 Performance Measures . . . 135
8.2.4 Model Performances . . . 136
8.2.5 Beta Calibration vs LLO . . . 140
8.3 Discussion and Conclusion . . . 142

9 Bayesian Combination of Correlated Subjective Probability Estimates 145
9.1 The Skill-Difficulty Correlated Fusion Model . . . 148
9.1.1 Generative Model . . . 148
9.1.2 Interpretation of the Model's Parameters as Skill and Difficulty . . 151
9.1.3 Correlations in the Model . . . 152
9.1.4 Parameter Inference and Fusion . . . 154
9.2 Evaluation . . . 154
9.2.1 Knowledge Test Confidence Data Set . . . 155
9.2.2 Cross-Validation . . . 155
9.2.3 Parameter Interpretation and Model Fit . . . 156
9.2.4 Fusion Performance . . . 161
9.3 Discussion and Conclusion . . . 164

10 General Discussion 167
10.1 General Findings . . . 168
10.2 Implications . . . 169
10.2.1 Mathematics . . . 169
10.2.2 Artificial Intelligence . . . 170
10.2.3 Robotics . . . 171
10.2.4 Cognitive Science . . . 173
10.3 Limitations and Future Work . . . 175

Bibliography 179

LIST OF FIGURES

Figure 2.1 An exemplary graphical model consisting of six stochastic variables $x_1, \dots, x_6$. 13

Figure 2.2 Different versions of an exemplary generative model of a simple coin-flipping experiment. 14

Figure 2.3 Generative models of cue integration with dependent and independent cues $c_1, \dots, c_n$ generated from state s. 17

Figure 2.4 Exemplary Bayesian integration of two cues c_1 and c_2 in order to estimate the state of the world s. 19

Figure 2.5 Generative models of classification with dependent and independent features $x_1, \dots, x_n$ generated from truth value y. 21

Figure 3.1 Two examples of fusion with Independent Opinion Pool. 30

Figure 3.2 Graphical model of the Independent Fusion Model (IFM). 32

Figure 3.3 Marginal and joint densities of correlated Dirichlet distributions with selected parameter values leading to low correlations. 40

Figure 3.4 Marginal and joint densities of correlated Dirichlet distributions with selected parameter values leading to varying and high correlations. 41

Figure 3.5 Graphical model of the Correlated Fusion Model (CFM). 42

Figure 3.6 Detailed graphical model of the Correlated Fusion Model (CFM) for $K = 2$ classifiers. 42

Figure 3.7 Detailed graphical model of the proposed Correlated Fusion Model for $K > 2$ classifiers. 43

Figure 3.8 Fusion of the base distributions $\boldsymbol{x}_i^1 = \boldsymbol{x}_i^2 = (0.6, 0.2, 0.2)$ using the CFM assuming different marginal parameters and correlations. 45

Figure 3.9 Fusion performances on simulated data in terms of mean entropy and log-loss. 47

Figure 3.10 Fusion performances on real data in terms of mean entropy and log-loss. 50

Figure 4.1 Joint densities of bivariate beta distributions with selected parameter values. 65

Figure 4.2 Different parameters can generate similar data. 69

Figure 4.3 The correlations r implied by different values of correlation parameter δ_1 compared to the desired correlation $\hat{r}$ given estimates of the marginal parameters a_1, a_2, b_1, b_2. 70

Figure 4.4 Results of a simulation study for evaluating the proposed moment matching approach for parameter inference. 71

Figure 4.5 The predicted probabilities from the chess classifier data set and the inferred bivariate beta distribution. 72

Figure 4.6 The exact correlation compared to the approximate correlation for all parameter configurations used in the simulation study. 74

Figure 5.1 An overview over the proposed multimodal intention recognition system, which fuses the modalities speech, gestures, gaze directions, and scene objects. 78
Figure 5.2 Examples of classifier fusion with Independent Opinion Pool. 82
Figure 5.3 The experiment setup in which the fusion system was evaluated. . . 85
Figure 5.4 Comparison of all possible combinations of base classifiers regarding accuracy, mean entropy, and score difference. 87
Figure 5.5 Mean entropies resulting from fusing all different combinations of base classifiers with a weighted sum. 87
Figure 5.6 Entropies of all distributions resulting from classifier combinations including the speech or gesture classifier. 88
Figure 5.7 An exemplary interaction between human and robot. 89
Figure 6.1 An overview of the proposed approach Multimodal IOP-Based Advice for Interactive Reinforcement Learning (MIA-IRL). 92
Figure 6.2 Comparison of IOP in MIA-IRL and the fusion method in C-IRL on exemplary base distributions. 99
Figure 6.3 Learning curves of non-interactive Q-learning, C-IRL, and MIA-IRL for a grid world with simulated advice in the first 10 episodes. . . . 100
Figure 6.4 The experimental setup used for evaluating MIA-IRL. 102
Figure 6.5 Learning curves of non-interactive Q-learning, C-IRL, and MIA-IRL individually for all 10 participants of the pancake experiment. . . . 103
Figure 6.6 Learning curves of non-interactive Q-learning, C-IRL, and MIA-IRL averaged over all 10 participants of the pancake experiment. 104
Figure 7.1 An overview of the proposed approach for automatically detecting a human Intention for Interaction (IFI). 108
Figure 7.2 The experimental setup, in which we collected our IFI data set. . . 111
Figure 7.3 The features and classifiers used for IFI detection. 114
Figure 7.4 The neck positions of all subjects in the experiment space while showing an IFI. 118
Figure 7.5 Boxplots of the individual subjects' F1 scores for the 4 best models of each body pose-only, speech-only, feature fusion, and IOP fusion classifiers. 120
Figure 8.1 Graphical models of one of the models proposed by Turner et al. (2014), Calibrate Then Average, and an exemplary normative fusion model. 125
Figure 8.2 Graphical models of the hierarchical and non-hierarchical beta fusion models. 129
Figure 8.3 Graphical models of the hierarchical and non-hierarchical symmetric beta fusion models. 132
Figure 8.4 Model performances on the Turner data set according to Brier score, 0-1 loss, and mean absolute error. 137
Figure 8.5 Model performances on the Knowledge Test Confidence data set according to Brier score, 0-1 loss, and mean absolute error. 138
Figure 8.6 Model performances on the reduced Turner data set consisting of a subset of the 20 forecasters of the Turner data set that provided the most forecasts. 139

Figure 8.7 Model performances on the reduced Knowledge Test Confidence data set consisting of a subset of the first 10 forecasters of KTeC data set. 139
Figure 8.8 The empirical, beta calibration, and LLO curves of two exemplary forecasters from the Knowledge Test Confidence data set together with the respective asymmetric and symmetric beta distributions. . 141
Figure 9.1 The graphical model of the Skill-Difficulty Correlated Fusion Model. 150
Figure 9.2 Simulated forecasts of two forecasters, x^1 and x^2, for 20000 queries. 153
Figure 9.3 A comparison of the skills of different forecasters and the difficulties of different queries from the KTeC data set. 157
Figure 9.4 The forecasts provided by three exemplary forecasters along with their estimated distribution according to our Skill-Difficulty Correlated Fusion Model. 158
Figure 9.5 The forecasts provided for six exemplary queries along with their estimated distribution according to our Skill-Difficulty Correlated Fusion Model. 160
Figure 9.6 Fusion performance of the Skill-Difficulty Correlated Fusion Model in comparison to related Bayesian fusion models on the KTeC data set. 163

LIST OF TABLES

Table 3.1 Comparison of the performances of the CFM and the model proposed by Pirs and Strumbelj (2019). 51
Table 3.2 Model parameters of the Correlated Fusion Model that we used to generate the first simulated data set (SIM 1) for five correlation levels from $r \approx 0$ to $r \approx 1$. 54
Table 3.3 Model parameters of the Correlated Fusion Model that we used to generate the second simulated data set (SIM 2) for five correlation levels from $r \approx 0$ to $r \approx 1$. 55
Table 3.4 Model parameters of the Correlated Fusion Model that we used to generate the third simulated data set (SIM 3) for five correlation levels from $r \approx 0$ to $r \approx 1$. 56
Table 3.5 Model parameters of the Correlated Fusion Model that we inferred for the real data sets with $K = 2$ base classifiers. 57
Table 3.6 Model parameters of the Correlated Fusion Model that we inferred for the real data set with $K = 3$ classifiers, Bookies C. 58
Table 7.1 Mean F1 scores, recalls, and precisions of the best performing IFI classifiers for different modalities. 119

ACRONYMS

ACE	Aggregative Contingency Estimation
AI	Artificial Intelligence
API	Application Programming Interface
ATC	Average Then Calibrate
B	Non-hierarchical Beta Fusion Model
BC	Beta Calibration
CDF	Cumulative Density Function
CFM	Correlated Fusion Model
C-IRL	IRL approach proposed by Cruz et al. (2018)
CNN	Convolutional Neural Network
CTA	Calibrate Then Average
CTALO	Calibrate Then Average using Log-Odds
DAG	Directed Acyclic Graph
DBCC	Dependent Bayesian Classifier Combination
DK	Development Kit
GMM	Gaussian Mixture Model
HB	Hierarchical Beta Fusion Model
HCTA	Hierarchical Calibrate Then Average
HCTALO	Hierarchical Calibrate Then Average on Log-Odds
HRI	Human-Robot Interaction
HSB	Hierarchical Symmetric Beta Fusion Model
IARPA	Intelligence Advanced Research Projects Activity
IBCC	Independent Bayesian Classifier Combination
IFI	Intention for Interaction
IFM	Independent Fusion Model
IOP	Independent Opinion Pool
IRL	Interactive Reinforcement Learning
KI	Künstliche Intelligenz

kNN	k Nearest Neighbors
KTeC	Knowledge Test Confidence
LLO	Linear-in-Log-Odds
LOO	Leave-One-Out
MAE	Mean Absolute Error
MCMC	Markov Chain Monte Carlo
MDP	Markov Decision Process
MIA-IRL	Multimodal IOP-Based Advice for Interactive Reinforcement Learning
MIA-IRL-3	MIA-IRL using 3 advice classifiers
MLE	Maximum-Likelihood Estimation
MLP	Multilayer Perceptron
PAVG	Probit Average
PDF	Probability Density Function
PMF	Probability Mass Function
ProMP	Probabilistic Movement Primitive
RL	Reinforcement Learning
ROC	Receiver Operating Characteristic
ROS	Robot Operating System
SB	Non-hierarchical Symmetric Beta Fusion Model
SDCFM	Skill-Difficulty Correlated Fusion Model
SDK	Software Development Kit
SVM	Support Vector Machine
ULINOP	Unweighted Linear Opinion Pool
7-DoF	7-Degrees-of-Freedom

1 INTRODUCTION

Optimal behavior is always Bayesian.

Wei Ji Ma (2019)

Due to pervasive noise and ambiguity, everything in our world is uncertain (D. R. Bach & Dolan, 2012; Lindley, 2013; Vilares & Kording, 2011). As a consequence, uncertainty is inherent in human perception, thoughts, decisions, and actions (Dong & Hayes, 2012; Harris & Wolpert, 1998; Knill & Pouget, 2004; Kochenderfer, 2015; Lindley, 2013). While uncertainty can obstruct goal achievement, representing and quantifying it can improve human decisions and actions accordingly (D. R. Bach & Dolan, 2012). Therefore, humans have a representation of their uncertainty to handle sensory information, to choose their actions accordingly, and to make judgments (Knill & Pouget, 2004; Koblinger et al., 2021; W. J. Ma & Jazayeri, 2014). Also, they are used to expressing and perceiving uncertainties in interactions with other humans (Krahmer & Swerts, 2005).

Furthermore, humans combine information from different sources in order to reduce their uncertainty about the world's state (Ernst & Bülthoff, 2004). By integrating redundant sensory cues, e.g., from different modalities such as audition and vision, they reduce their perceptual uncertainty. In particular, there is much evidence that humans optimally combine sensory information from different cues according to Bayes' rule (Alais & Burr, 2004; Ernst & Banks, 2002; Gepshtein & Banks, 2003; Gepshtein et al., 2005; Girshick & Banks, 2009; Helbig & Ernst, 2007; Hillis et al., 2002; Hillis et al., 2004; Knill & Saunders, 2003; Landy & Kojima, 2001; Saunders & Chen, 2015; Scarfe & Hibbard, 2011; Svarverud et al., 2010; Watt et al., 2005). By this, they weight different sensory cues according to their uncertainty and maximally reduce the uncertainty of the integrated sensory estimate (Ernst & Bülthoff, 2004; Landy et al., 2011; Scarfe, 2022).

In order to cope with an uncertain world, like humans, Artificial Intelligence (AI) systems as well as robots as their embodied form should represent, consider, reduce, and communicate uncertainty (Abdar et al., 2021; Baek et al., 2023; Z. Huang et al., 2021; Kompa et al., 2021). This can particularly be crucial for critical applications, such as autonomous driving or medical image analysis (Guo et al., 2017), but also for a successful interaction between humans and AI, e.g., for AI-supported human decision making or human-robot collaborations in industrial or caregiving settings. For such human-AI interactions it has been shown that quantifying uncertainty and providing uncertainty information together with predictions avoids over- and under-relying in AI models (Bhatt et al., 2021), increases trustworthiness (Bhatt et al., 2021; Kompa et al., 2021), and improves the performance of human decision-makers (Bansal et al., 2021; Joslyn & LeClerc, 2013; Nadav-Greenberg & Joslyn, 2009; Roulston et al., 2006). Also, embodied human-AI interaction, i.e., human-robot interaction, can be made safer if uncertainty is considered, represented, and communicated (Baek & Kröger, 2023; Baek et al., 2023). However, the quantification, representation, and communication of uncertainty is not a matter of course in AI systems. Particularly with recent non-probabilistic deep learning systems, uncertainty quantification in Machine Learning and AI has declined, a fact that has been noted because of its danger in critical applications (Guo et al., 2017; Kompa et al., 2021).

1.1 Uncertainty in Classifier Ensembles

AI systems can represent and communicate their uncertainty by outputting probability distributions over the target variable. For example, probabilistic classifiers output categorical probability distributions over all classes instead of only the discrete predicted labels (K. P. Murphy, 2022). Also, it is common practice to combine different information sources to increase an AI system's performance. For example, classifier ensembles combine different individual classifiers to ensembles in order to improve classification performance (Bishop, 2006; Dietterich, 2000; Hamed & Akbari, 2018; Kittler et al., 1998; Mohandes et al., 2018; Pirs & Strumbelj, 2019). In fact, these classifier ensembles are the most successful classification methods, e.g., in machine learning competitions (Kuncheva, 2014; Pirs & Strumbelj, 2019). However, they only focus on improving the classification performance in terms of, e.g., accuracy and do not explicitly consider normative uncertainty reduction. In particular, it is not clear how to optimally combine the outputs of individual probabilistic classifiers in order to correctly reduce their uncertainty, similar to how humans combine multiple cues in perception. Therefore, a fundamental question in AI research is how to optimally combine probabilistic classifiers while explicitly considering and reducing uncertainty. Since classifiers trained on the same target are usually correlated (Jacobs, 1995; Kim & Ghahramani, 2012), which can lead to overestimated uncertainty reduction if disregarded (Wilson, 2017), particular focus should thereby be placed on the combination of correlated classifiers. A solution to this problem has versatile applications in all domains where classification is used.

In this thesis, we present a normative Bayesian model for combining the outputs of probabilistic classifiers. Using this model, we formalize how to obtain the correct uncertainty of ensembles of multiple classifiers given different assumptions. For Bayes optimal classifier fusion, our model can consider the individual classifiers' properties, such as uncertainty, bias, and variance. In addition, we explicitly model the correlation between individual classifiers. For modeling this correlation, we introduce a new probability distribution, the correlated Dirichlet distribution. The resulting Correlated Fusion Model quantifies how classification uncertainty should be reduced through Bayes optimal classifier fusion depending on the individual classifiers' uncertainty, bias, variance, and correlation. In addition, it outperforms related Bayesian classifier fusion methods on simulated and real data sets.

A special case of the correlated Dirichlet distribution that we introduce for modeling correlated probabilistic classifiers is the bivariate beta distribution. It models two beta-distributed random variables with a positive correlation and is therefore particularly interesting for modeling binary correlated probabilistic classifiers. However, it is also of general interest in statistics. While the bivariate beta distribution has already been proposed previously, prior work used an approximate and inaccurate method to compute the distribution's covariance and correlation and estimated its parameters using these inaccurate measures. In this thesis, we derive all product moments and the exact covariance and correlation of the bivariate beta distribution, which can be computed numerically. Using these derivations, we propose an algorithm for estimating the distribution's parameters using moment matching.

1.2 Bayesian Classifier Fusion for Multimodal Human-Robot Interaction

The correct fusion of multiple probabilistic classifiers that explicitly and optimally considers uncertainty can contribute to human-robot interaction, e.g., with collaborative robots in industrial settings (L. Wang et al., 2020) or assistive robots in elderly assistance and caregiving (Abbasi et al., 2019; Rodomagoulakis et al., 2016). Humans interact multimodally in interactions with other humans (Bunt et al., 1998; Quek et al., 2002; Rodomagoulakis et al., 2016; Turk, 2014). Specifically, they do not only perceive their environment by integrating redundant cues from different modalities (Ernst & Bülthoff, 2004), but also express themselves using multiple redundant modalities, such as speech, gestures, and gaze directions (Barthelmess et al., 2006; Chandrasekaran et al., 2009; De Ruiter et al., 2012; So et al., 2009; Todisco et al., 2021). In order to enable an intuitive and natural interaction between humans and robots, we need to come up with algorithms that enable robots to also interact multimodally (Goodrich & Schultz, 2008; Stiefelhagen et al., 2004). While multimodal interaction between humans and robots can increase interaction performance and robustness (Mollaret et al., 2016; Rodomagoulakis et al., 2016), here we strive to also reduce uncertainty, similar to how human perception uses multimodal information to reduce uncertainty. In particular, robots should be able to perceive humans' uncertain multimodal signals, expressing e.g., their intentions, and combine them in order to reduce their uncertainty in the correct way. However, there is limited previous research on multimodal intention recognition in human-robot interaction that considers the individual modalities' uncertainties in order to reduce the robot's uncertainty in the Bayes optimal way.

To close this gap, in this thesis we detect human intentions from multiple modalities by training probabilistic classifiers for each modality and combining their output distributions Bayes optimally to a combined distribution representing the correct uncertainty. As a consequence, the uncertainty of the robot about the humans' expressed signals, e.g., their intentions, can be correctly reduced according to the individual modalities' uncertainties. We introduce three approaches that demonstrate how human-robot interaction can benefit from Bayes optimal classifier fusion. First, we recognize human intentions from multimodal data including speech, gestures, gaze directions, and scene objects. We evaluate the proposed multimodal intention recognition system in an interaction task between a human and a 7-Degrees-of-Freedom (7-DoF) robot arm and show that adding more modalities contributes to increased detection performance and reduced uncertainty. Second, we enable humans to teach a 7-DoF robot arm using interactive reinforcement learning. The human can provide multimodal advice using the modalities speech and gestures, which are fused using Bayesian classifier fusion. We evaluate the approach with 10 human participants and show that the learning speed can be improved significantly compared to alternative approaches that use different classifier fusion methods. Third, we learn to recognize a special human intention from natural human behavior: people's intention to start an interaction with a robot, which we call the intention for interaction. With

21 human participants, we recorded multimodal data including speech and body poses in a collaborative task with a two-armed assistive robot. On the resulting data set, we train different unimodal and multimodal classifiers and show that the intention for interaction can be detected better from multimodal data, using Bayesian classifier fusion.

1.3 Bayesian Combination of Subjective Probability Estimates

Bayesian fusion methods that explicitly and correctly consider uncertainty can not only be applied to fuse the outputs of probabilistic classifiers but also to combine subjective probability estimates provided by humans. Human's subjective probability estimates or forecasts, especially those given by experts, are of particular importance in many different domains, such as finance, business, marketing, politics, engineering, meteorological, ecological, and environmental science, as well as public health (McAndrew et al., 2021), and can be used to build rule-based AI systems (Masri et al., 2019). In contrast to classifiers, they do not require large data sets but can rely on human experience and intuition for providing information (McAndrew et al., 2021). Combining human forecasts usually increases the prediction performance compared to single forecasts (Budescu & Chen, 2015; McAndrew et al., 2021; Satopää, 2022; Turner et al., 2014). However, increased performance in terms of correctly predicted events is not sufficient in many cases. For example, imagine multiple doctors providing their estimate on the probability of a patient's diagnosis. Knowing the correct uncertainty of their combined probability estimate can be crucial for optimal treatment of a patient. Thus, also in the field of human forecast aggregation the uncertainty of individual forecasters should be explicitly modeled and correctly considered for their combination. In fact, defining a normative model for combining human forecasts is designated to be an open challenge in forecasting research (McAndrew et al., 2021).

As part of this thesis, we introduce a family of normative Bayesian models for the combination of subjective probability estimates provided by human forecasters. The models, which are closely related to our normative Bayesian model for classifier fusion, model the forecasting behavior of individual forecasters with beta distributions, implicitly calibrate them, and combine their forecasts accordingly in order to obtain the Bayes optimal uncertainty of the fused forecast. However, since the proposed fusion models disregard the correlation between the forecasters, they reduce too much uncertainty and are thus overconfident. For this reason, in a second step, we extend these models to a Bayesian model for the combination of correlated subjective probability estimates. This model explicitly represents the skills of the individual forecasters as well as the difficulties of individual queries for which forecasts are provided. By this, it can explicitly model the correlation between the individual forecasters and consider it when fusing forecasts, which improves the model's fusion performance compared to our previous model and related fusion models.

1.4 Outline

This thesis is structured as follows. In Chapter 2 we start by providing the background of the works presented in the thesis, consisting of the most important concepts, definitions, and mathematical foundations, as well as prior research findings. Chapter 3 introduces our normative Bayesian framework for classifier fusion. Given sequentially more general assumptions, here we derive how probabilistic classifiers should be fused Bayes optimally given their individual properties, i.e., their uncertainty, bias, and variance, as well as their correlation. In Chapter 4 we propose an algorithm for parameter estimation for the bivariate beta distribution, which is a special case of the correlated Dirichlet distribution introduced in Chapter 3 for classifier fusion. In Chapter 5 we apply the Bayesian fusion method Independent Opinion Pool, which was derived in Chapter 3, to multimodal intention recognition in human-robot interaction. Accordingly, Chapter 6 investigates how Bayesian fusion with Independent Opinion Pool can improve interactive reinforcement learning with multimodal human action advice in human-robot interaction. In Chapter 7 we recognize a specific human intention, i.e., the human intention to start an interaction with a robot, from multimodal data. We show that fusing individual classifiers for the respective modalities with the Bayesian fusion method Independent Opinion Pool results in the best recognition performance. Instead of fusing classifiers, in Chapter 8 we fuse forecasts provided by humans. In particular, here we introduce a family of normative Bayesian fusion models, which model human forecasts with beta distributions and implicitly calibrate them when fusing. While the Bayesian models in Chapter 8 assume independent human forecasters, in Chapter 9 we extend these models to a Bayesian model for fusing correlated human forecasts, which explicitly models the skills of the forecasters as well as the difficulties of the queries the forecasts are provided for and by this the correlation between forecasts. Finally, in Chapter 10 we discuss the methods and findings provided in this thesis as well as their implications and limitations and suggest interesting directions for future work.

1.5 Contributions

The chapters in this thesis are based on previous publications, i.e., Trick and Rothkopf (2022), Trick, Rothkopf, and Jäkel (2023b), Trick et al. (2019), Trick et al. (2022), Trick, Lott, et al. (2023), and Trick, Rothkopf, and Jäkel (2023a), and may contain previously published text and figures.

– Chapter 3: Normative Bayesian Classifier Fusion

  This work was published in
  Trick, S., & Rothkopf, C. A. (2022). Bayesian classifier fusion with an explicit model of correlation, In *Proceedings of the 25th international conference on artificial intelligence and statistics (AISTATS)*. PMLR. https://proceedings.mlr.press/v151/trick22a.html.

– Chapter 4: Parameter Estimation for a Bivariate Beta Distribution

  This work was published in
  Trick, S., Rothkopf, C. A., & Jäkel, F. (2023b). Parameter estimation for a bivariate beta distribution with arbitrary beta marginals and positive correlation. *METRON*, 1–18. https://doi.org/10.1007/s40300-023-00247-2.

- Chapter 5: Bayesian Fusion for Intention Recognition in Human-Robot Interaction

  This work was published in
  Trick, S., Koert, D., Peters, J., & Rothkopf, C. A. (2019). Multimodal uncertainty reduction for intention recognition in human-robot interaction, In *Proceedings of the 2019 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE. `https://doi.org/10.1109/IROS40897.2019.8968171`.
  ©2019 IEEE. All rights reserved. Reprinted with permission.[1]

- Chapter 6: Interactive Reinforcement Learning with Bayesian Fusion of Multimodal Advice

  This work was published in
  Trick, S., Herbert, F., Rothkopf, C. A., & Koert, D. (2022). Interactive reinforcement learning with Bayesian fusion of multimodal advice. *IEEE Robotics and Automation Letters*, *7*(3), 7558–7565. `https://doi.org/10.1109/LRA.2022.3182100`.
  ©2022 IEEE. All rights reserved. Reprinted with permission.[1]

- Chapter 7: Multimodal Detection of the Intention for Interaction in Human-Robot Interaction

  This work was published in
  Trick, S., Lott, V., Scherf, L., Rothkopf, C. A., & Koert, D. (2023). What can I help you with: Towards task-independent detection of intentions for interaction in a human-robot environment, In *Proceedings of the 2023 32nd IEEE international conference on robot and human interactive communication (RO-MAN)*. IEEE. `https://doi.org/10.1109/RO-MAN57019.2023.10309347`.
  ©2023 IEEE. All rights reserved. Reprinted with permission.[1]

- Chapter 8: A Normative Model for Bayesian Combination of Subjective Probability Estimates

  This work was published in
  Trick, S., Rothkopf, C. A., & Jäkel, F. (2023a). A normative model for Bayesian combination of subjective probability estimates. *Judgment and Decision Making*, *18*, e40. `https://doi.org/10.1017/jdm.2023.39`.

- Chapter 9: Bayesian Combination of Correlated Subjective Probability Estimates

  This work is in preparation for submission.

[1] Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

2 BACKGROUND

This thesis examines how probabilistic forecasts provided by classifiers or humans should be combined in order to deal with the ubiquitous uncertainty present in our world. We propose a Bayesian model for the combination of probabilistic classifiers that normatively specifies how to obtain the correct uncertainty of the ensemble. Furthermore, we show how Bayesian fusion considering uncertainty can increase performance in relevant applications in human-robot interaction scenarios and investigate Bayesian models for the aggregation of subjective probability estimates provided by humans. In the following chapter, we provide the most important concepts, definitions, mathematical foundations, and prior research findings that form the basis of our work.

In Section 2.1, we start with discussing the ubiquitous uncertainty in our world and define the two sources of uncertainty, aleatoric and epistemic uncertainty. Section 2.2 gives a definition of probability as a quantification of uncertainty, followed by some mathematical foundations of probability theory in Section 2.3. Building upon these mathematical foundations, Section 2.4 introduces Bayes' theorem and Bayesian inference. The basics of graphical models for visualizing probability distributions, and in particular generative models, are presented in Section 2.5, while the subsequent Section 2.6 discusses the parameter estimation methods that are used for the Bayesian models in this thesis. In Section 2.7 we review prior work on how humans combine sensory cues in order to reduce their uncertainty about their environment's current state, which can be described with a generative model and Bayesian parameter inference. Finally, in Section 2.8 we provide some definitions of classification and classifier fusion, which can also be formalized as parameter estimation in a generative model.

2.1 Uncertainty

The world is full of uncertainty (Lindley, 2013). First and foremost, human perception is uncertain (Fiser et al., 2010; Knill & Pouget, 2004; W. J. Ma et al., 2006; Vilares & Kording, 2011). We can never be certain about the current state of our environment because the sensory information we receive is noisy (Faisal et al., 2008; Fiser et al., 2010; Kochenderfer, 2015; Vilares & Kording, 2011). In addition, the sensory input can be ambiguous since different states of the world can generate the same input to our sensory system (Fiser et al., 2010; Kersten & Yuille, 2003; Vilares & Kording, 2011) and the same state can generate infinitely many inputs (Kersten & Yuille, 2003). For example, different three-dimensional objects in the real world can cause the same two-dimensional image on the human retina (Ernst & Bülthoff, 2004), and a bicycle seen from different perspectives can cause different two-dimensional images on the human retina (Kersten & Yuille, 2003). As a result, perception can be seen as unconscious inference, where the real state of the world is constantly inferred from uncertain observations (Helmholtz, 1924). In the human nervous system, uncertainty is even explicitly represented (Fiser et al., 2010). However, perceptual uncertainty is not only relevant for human perception but also for machine perception. Just like humans, machines also need to deal with uncertain input due to noisy and ambiguous observations.

In addition to perceptual uncertainty caused by imperfect observations, there is also uncertainty that is caused by lack of knowledge. Factual statements can be known to be true or false, but for a majority of statements we do not know the correct answer, so we are uncertain (Lindley, 2013). For example, predictions of future states are always uncertain, including weather forecasts and forecasts of outcomes of elections or the course of a patient's disease. Again, uncertainty through imperfect knowledge affects both humans and machines.

Perceptual uncertainty and lack of knowledge make us humans and also machines that support us uncertain about the current and future state of the world, which in turn renders decision making uncertain (Dong & Hayes, 2012; Nadav-Greenberg & Joslyn, 2009). This uncertainty in decision making is even amplified by the fact that the consequences of decisions are not necessarily deterministic (Kochenderfer, 2015). Moreover, even human motor action execution is uncertain since noise in the firing motor neurons causes deviations of the executed trajectories from the desired trajectories (Faisal et al., 2008; Harris & Wolpert, 1998).

The sources of the different uncertainties listed above can be divided into aleatoric uncertainty and epistemic uncertainty (Bhatt et al., 2021; Der Kiureghian & Ditlevsen, 2009; Hora, 1996; Hüllermeier & Waegeman, 2021; Kompa et al., 2021; K. P. Murphy, 2022). Aleatoric or aleatory uncertainty describes the uncertainty that arises from the inherent randomness and intrinsic variability of the data (Hüllermeier & Waegeman, 2021; K. P. Murphy, 2022), or the natural, unpredictable variation of a system's performance (Hora, 1996). The term aleatoric is derived from the Latin word alea, which translates to dice (Der Kiureghian & Ditlevsen, 2009; K. P. Murphy, 2022). Aleatoric uncertainty is a property of the data-generating process (Hüllermeier & Waegeman, 2021) or simply the data and is therefore also termed data uncertainty (K. P. Murphy, 2022). Aleatoric uncertainty is irreducible (Hora, 1996; Hüllermeier & Waegeman, 2021). In particular, it cannot be reduced by collecting more data or additional information. A popular example of aleatoric uncertainty is flipping a coin: We know that the probability for heads is 0.5, so we know the data-generating model of the task, but we can never know the outcome of the next flip, no matter how often we repeat flipping (Hüllermeier & Waegeman, 2021; K. P. Murphy, 2022). Other examples of aleatoric uncertainty are noisy data or a class overlap of data when classifying (Bhatt et al., 2021). In particular, perceptual uncertainty caused by noisy or ambiguous data or variability in motor action execution due to noise in the firing motor cells can be considered aleatoric uncertainty.

Epistemic uncertainty is caused by a lack of data or knowledge (Hora, 1996). In particular, it arises from our lack of knowledge about the underlying mechanisms generating our data (K. P. Murphy, 2022), the correct model to select (e.g., a linear or polynomial curve to fit) or the model's parameters (Hüllermeier & Waegeman, 2021; Kompa et al., 2021). While the term epistemic uncertainty is derived from the Greek word for knowledge (Der Kiureghian & Ditlevsen, 2009), this kind of uncertainty is also called model uncertainty (K. P. Murphy, 2022) since it is a property of the model (or the agent/human): The model has not enough knowledge or has not seen enough data to be certain. In contrast to aleatoric uncertainty, epistemic uncertainty can be reduced by observing additional data or getting new information (Hora, 1996; Hüllermeier & Waegeman, 2021). In the coin-flipping example discussed above, there is epistemic uncertainty if we do not know if the coin is fair. In this case, we are uncertain about the data-generating model and need to collect more data, i.e., repeatedly flip the coin, to reduce our epistemic uncertainty. The above-mentioned uncertainty about the correctness of factual statements because of limited knowledge is another example of epistemic uncertainty.

2.2 Probability

Uncertainty can be quantified with probability (Bishop, 2006; Gelman et al., 2013; Lindley, 1987; K. P. Murphy, 2012, 2022). However, note that there are two different interpretations of the probability $Pr(A)$ of an event A, the frequentist and the Bayesian interpretation. Frequentists interpret a probability as the relative frequency of event A's occurrence (Bishop, 2006; K. P. Murphy, 2022). If, e.g., we flip a fair coin and event A is "the coin lands heads", then the probability for this event is $Pr(A) = 0.5$. Given the frequentist interpretation of probability, this means that if we repeatedly flip the coin, it will land heads for about 50% of all flips, the more accurate, the more often we flip it (K. P. Murphy, 2022).

In contrast, according to the Bayesian interpretation of probability, $Pr(A)$ is used to quantify our uncertainty if event A will occur or not (Bishop, 2006; K. P. Murphy, 2012, 2022), it is a subjective belief rather than a frequency (Bertsekas & Tsitsiklis, 2008). For the coin-flipping example from above, the Bayesian interpretation of the probability $Pr(A) = 0.5$ is that for the next coin flip, it is equally likely that the coin will land heads or tails (K. P. Murphy, 2022). Therefore, with Bayesian probabilities we can quantify our uncertainty about the outcome of the next coin flip. In particular, one big strength of the Bayesian as opposed to the frequentist view on probabilities is that it serves to express uncertainties about events that cannot be repeated, e.g., if the polar ice caps will be melted by the next 10 years (Bishop, 2006; K. P. Murphy, 2012, 2022). Thus, we can use probability to quantify all uncertainties listed above in Section 2.1, i.e., perceptual uncertainty, uncertainty due to imperfect knowledge, uncertainty in decision making and action execution. In fact, Lindley (1987) even states that "the only satisfactory description of uncertainty is probability" and other attempts to quantify uncertainty, e.g., fuzzy logic (Zadeh, 1983) or Dempster-Shafer theory (Shafer, 1976) are unnecessary.

2.3 Probability Theory

In the following, we present the basic concepts and rules of probability that will be used throughout this thesis. We mainly follow the presentation by K. P. Murphy (2022). A random variable X represents a quantity of interest that is unknown and/or can change. If X can only take a finite set of values, it is called a discrete random variable. Examples are representing the outcome of flipping a coin or the class label of an image to be classified. The probability of X taking value x is defined as $Pr(X = x)$ and $p(x) := Pr(X = x)$ is the probability mass function (PMF) of X. The PMF assigns a probability to each possible value x with $0 \leq p(x) \leq 1$ and $\sum p(x) = 1$. A discrete random variable can be modeled with a discrete probability distribution, such as the Bernoulli distribution, the binomial distribution, the categorical distribution, or the Poisson distribution, which then determines the shape of its PMF.

If X represents a real-valued quantity, it is called a continuous random variable. X could for example represent a size, a temperature, or a duration. Since continuous random variables can take all possible values on the continuous scale, the one specific predefined value x will never occur. Therefore, for continuous random variables we do not assign probabilities to single values but to intervals. In particular, we define the cumulative distribution function (CDF) of X as $P(x) \coloneqq Pr(X \leq x)$ and by this can compute the probability that X is in the interval $(a, b]$ as $Pr(a < X \leq b) = P(b) - P(a)$. If we derive the CDF, we obtain the probability density function (PDF) as $p(x) \coloneqq \frac{\mathrm{d}}{\mathrm{d}x} P(x)$. Using the PDF instead of the CDF, the probability of X being in the interval $(a, b]$ can be computed using a finite integral: $Pr(a < X \leq b) = \int_a^b p(x)\,\mathrm{d}x$. For a very small interval $(a, a+\delta]$ with $\delta > 0$ the probability of X being in the interval can be approximated by $Pr(x < X \leq x + \delta) = \delta p(x)$. Thus, although not directly mapping probabilities to values of X, the PDF provides an intuition of these probabilities. Continuous probability distributions, such as the Gaussian distribution, the gamma distribution, or the beta distribution, can be used to model a continuous random variable.

According to K. P. Murphy (2022), as above, random variables should be represented with capital letters, e.g., X, while the values they can take on should be represented with small letters, e.g., x. However, other notations also allow random variables to be directly described with small letters (Bishop, 2006), which improves readability and is therefore used for the following definitions.

Two (or more) random variables x and y can also be represented jointly as the joint distribution $p(x, y)$, which assigns a probability (density or mass) value to all possible combinations of x and y. Given the joint distribution $p(x, y)$, we can obtain the marginal distribution of x by summing over all possible values of y for discrete random variables,

$$p(x) = \sum_y p(x, y), \tag{2.1}$$

or integrating over all possible values of y for continuous random variables respectively,

$$p(y) = \int p(x, y)\,\mathrm{d}y. \tag{2.2}$$

Summing or integrating over all possible values of y to obtain the marginal distribution of x is also called marginalizing over y.

The conditional distribution of y given x is defined as

$$p(y|x) = \frac{p(x, y)}{p(x)} \tag{2.3}$$

and represents the probability distribution of y given that the value of x is given. By rewriting equation (2.3) we obtain the product rule:

$$p(x, y) = p(x)p(y|x), \tag{2.4}$$

which can be extended to the chain rule of probability for n random variables $x_1, \ldots, x_n$:

$$p(x_1, \ldots, x_n) = p(x_1)p(x_2|x_1)p(x_3|x_1, x_2) \ldots p(x_n|x_1, \ldots, x_{n-1}). \tag{2.5}$$

The chain rule allows representing a complex joint distribution as a product of conditional distributions.

If the joint distribution $p(x, y)$ can be written as the product $p(x)p(y)$, x and y are independent. If they are only independent given a third variable z, they are conditionally independent given z, and it holds that $p(x, y|z) = p(x|z)p(y|z)$. If x and y are not independent, they might be correlated. The correlation between two random variables x and y can be quantified with the Pearson correlation coefficient r, which is defined as

$$r := \frac{\mathrm{Cov}(x, y)}{\sqrt{\mathrm{Var}(x)\mathrm{Var}(y)}} \tag{2.6}$$

with $\mathrm{Var}(x)$ as the variance of x. $\mathrm{Cov}(x, y)$ defines the covariance between x and y. It is defined as

$$\mathrm{Cov}(x, y) := \mathrm{E}((x - \mathrm{E}(x))(y - \mathrm{E}(y))) = \mathrm{E}(xy) - \mathrm{E}(x)\mathrm{E}(y), \tag{2.7}$$

with $\mathrm{E}(x)$ as the expected value of x, and quantifies how strongly x and y are linearly related. Covariance can take all values in $\mathbb{R}$, while correlation is in $[-1, 1]$. If $r = 1$, x and y show a perfect positive correlation, i.e., they form a line when plotted against each other. If $r = -1$, x and y show a perfect negative correlation and if $r = 0$, x and y are uncorrelated. A correlation of $r = 0$, however, does not necessarily mean that x and y are independent. It just indicates that there is no linear relation between x and y.

2.4 Bayesian Inference

Combining equations (2.3) and (2.4) we obtain Bayes' rule:

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}. \tag{2.8}$$

It computes the probability distribution of x given y. y might be, e.g., some observed data, while x is the latent or hidden variable of interest we want to infer. $p(x)$ is called the prior distribution, which represents the distribution of x we assume before knowing anything about the value of y, i.e., before seeing any data. It might be learned from past experience. $p(y|x)$ is called the likelihood. It computes how likely the observed data y are given different values of x and is therefore a function of x, while y is fixed. Note that the likelihood does not necessarily sum or integrate to 1, so it is not a distribution. $p(y)$ is known as the marginal likelihood (K. P. Murphy, 2022) or the normalization constant (Bishop, 2006). It can be obtained by marginalizing the joint distribution $p(x, y)$ over x and normalizes the term to a probability distribution. $p(y)$ does not depend on x, so it is a constant that in many cases does not need to be computed explicitly. The resulting probability distribution $p(x|y)$ is the posterior distribution of x given observed data y and our prior information about x. If we see new data the next time, our posterior can serve as prior. Thus, Bayes' rule can be used to continuously update our belief about a random variable x when we observe new data. Bayesian inference using Bayes' rule is normative, which means that it defines how a rational agent should change its posterior over x when observing new data y (M. D. Lee & Wagenmakers, 2014). However, note that the inferred posterior $p(x|y)$ is only correct given that the likelihood $p(y|x)$ and prior $p(x)$ are good descriptions of the process that generated the data y, i.e., the data-generating process (W. J. Ma, 2019).

In many cases, after inferring a posterior belief over x using Bayes' rule, the agent is required to take an action a according to this belief. This action can be a movement or an answer to a question but also simply an estimate of the value of x. In order to choose an action a, a cost function or objective function $\mathrm{C}(x, a)$ can be defined that quantifies the cost of action a given the true value of x and should thus be minimized. Then, action a can be chosen by minimizing the expected cost, for discrete x as

$$\mathrm{E}(\mathrm{C}(a)) = \sum_x p(x|y)\mathrm{C}(x, a) \tag{2.9}$$

and for continuous x as

$$\mathrm{E}(\mathrm{C}(a)) = \int p(x|y)\mathrm{C}(x, a)\,\mathrm{d}x, \tag{2.10}$$

with $p(x|y)$ being the posterior distribution obtained using Bayes' rule (W. J. Ma, 2019; K. P. Murphy, 2022; Rothkopf et al., 2010). Since it minimizes the expected cost, the action chosen in this way is the optimal, normative action (W. J. Ma, 2019; K. P. Murphy, 2022; Rothkopf et al., 2010). If a is just an estimate of the value of x, we can directly specify the optimal estimate. If x is discrete and we aim to estimate its value correctly, i.e., the cost function is $\mathrm{C}(x, a) = 0$ if $x = a$ and $\mathrm{C}(x, a) = 1$ otherwise, the optimal estimate is the mode of the posterior $p(x|y)$. If x is continuous and we want to minimize the squared error between our estimate and the true value of x, i.e., $\mathrm{C}(x, a) = (x - a)^2$, the optimal estimate is the mean of the posterior $p(x|y)$ (W. J. Ma, 2019; K. P. Murphy, 2022). Designing appropriate cost functions for complex desired behavior can be challenging, e.g., in robotics (Englert et al., 2017). Inferring cost functions from observed behavior can be realized using inverse reinforcement learning (Rothkopf & Dimitrakakis, 2011) or inverse optimal control (Schultheis et al., 2021; Straub et al., 2023).

2.5 Graphical Models as Generative Models

Joint probability distributions can be illustrated as graphical models. For providing important foundations on graphical models needed in this thesis, we will mainly follow the presentation of Bishop (2006) and M. D. Lee and Wagenmakers (2014). Graphical models visualize the structure of a probabilistic model, i.e., how the different random variables in a joint distribution are related to each other. They are composed of nodes, which represent random variables, and edges between the nodes, which represent relations between these variables.

If the edges are undirected links between nodes, the graphical model is called an undirected graphical model or Markov random field. If, in contrast, the edges are directed and visualized as arrows, we call the graphical model a directed graphical model. If it additionally does not include any cycles, it is a directed acyclic graph (DAG). In this thesis, we refer to DAGs when we talk about graphical models.

DAGs represent dependence relationships between random variables with arrows between their respective nodes. Nodes can be visualized differently, depending on the variable's properties. However, the notation is not completely standardized. Here, we just differentiate between stochastic and deterministic nodes with single- and double-bordered nodes.

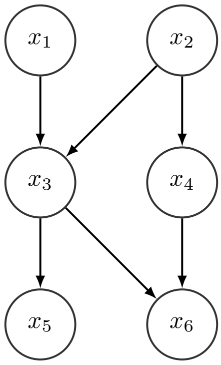


Figure 2.1: An exemplary graphical model consisting of six stochastic variables $x_1, \dots, x_6$.

Variables that are just set to a fixed specified value are visualized without a surrounding node. Plates around nodes represent repetitions, e.g., they define a probabilistic relationship for every element x_i of a vector $\boldsymbol{x}$.

An edge pointing from variable x_1 to variable x_2 indicates that x_1 is a parent node of x_2 and x_2 is a child node of x_1 respectively. The parent-child relationships in a graphical model are of particular importance for simplifying the joint distribution visualized with the graphical model.

As we showed in (2.5), the joint distribution of n random variables $x_1, \dots, x_n$ can be represented as a product of conditional distributions, one for each variable conditioned on all other variables with lower indices. This factorization is independent of how the specific graphical model looks like. However, if we know the graphical model, the joint distribution can be factorized as a product of conditional distributions, where the distribution for each variable x_i is just conditioned on its parent nodes $\mathrm{pa}(x_i)$:

$$p(x_1, \dots, x_n) = \prod_{i=1}^{n} p(x_i | \mathrm{pa}(x_i)). \tag{2.11}$$

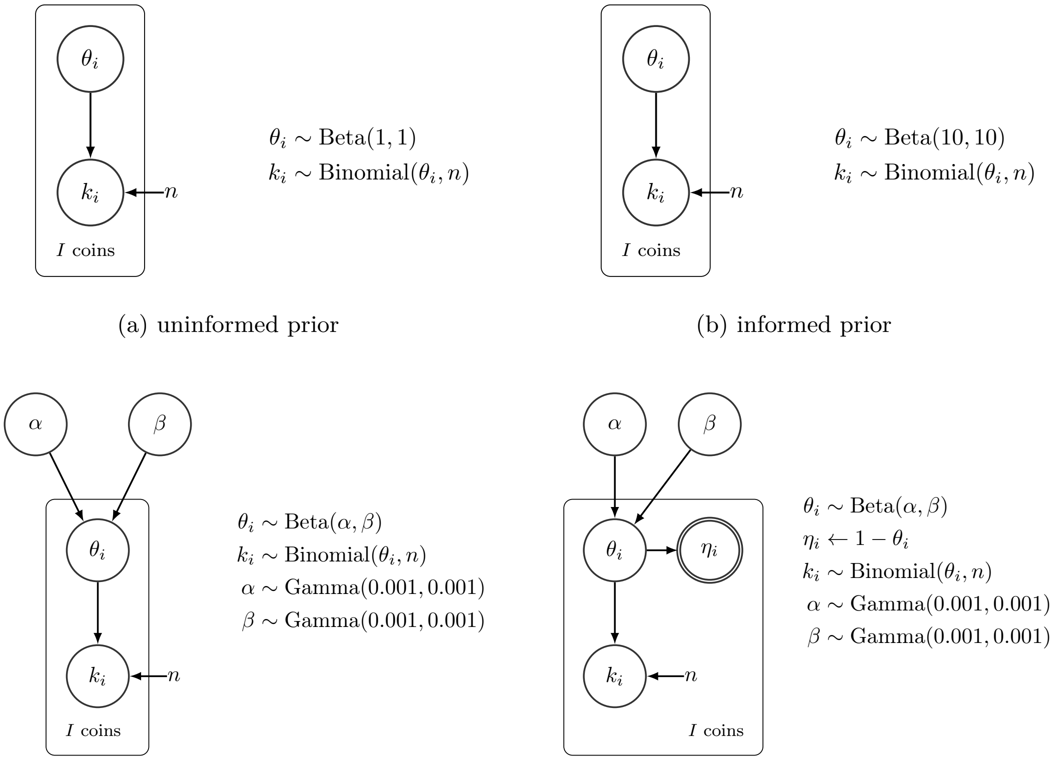
The factorization for the exemplary graphical model in Figure 2.1 is

$$p(x_1, \dots, x_6) = p(x_1)p(x_2)p(x_3|x_1, x_2)p(x_4|x_2)p(x_5|x_3)p(x_6|x_3, x_4). \tag{2.12}$$

Random variables in probabilistic models can be observed or latent. The values of observed variables are known, e.g., given as data. Latent variables' values are unknown. Sometimes we want to infer their values using the observed variables, sometimes their exact values are irrelevant and they are just introduced to construct a complex joint distribution from simple conditional distributions.

In most applications of graphical models, the random variables with lower indices represent latent variables, and variables with higher indices are observed variables represented as terminal nodes of the graphical model. These models are called generative models and model the data-generating process.

Figure 2.2 shows different versions of an exemplary generative model of a simple coin-flipping experiment. We flip I coins n times and observe k_i heads for coin i. Having seen


θ_i
k_i
n
I coins
$\theta_i \sim \mathrm{Beta}(1,1)$
$k_i \sim \mathrm{Binomial}(\theta_i, n)$
θ_i
k_i
n
I coins
$\theta_i \sim \mathrm{Beta}(10,10)$
$k_i \sim \mathrm{Binomial}(\theta_i, n)$
α
β
θ_i
k_i
n
I coins
$\theta_i \sim \mathrm{Beta}(\alpha,\beta)$
$k_i \sim \mathrm{Binomial}(\theta_i, n)$
$\alpha \sim \mathrm{Gamma}(0.001, 0.001)$
$\beta \sim \mathrm{Gamma}(0.001, 0.001)$
α
β
θ_i
η_i
k_i
n
I coins
$\theta_i \sim \mathrm{Beta}(\alpha,\beta)$
$\eta_i \leftarrow 1 - \theta_i$
$k_i \sim \mathrm{Binomial}(\theta_i, n)$
$\alpha \sim \mathrm{Gamma}(0.001, 0.001)$
$\beta \sim \mathrm{Gamma}(0.001, 0.001)$

(a) uninformed prior

(b) informed prior

(c) hierarchical prior

(d) hierarchical prior & deterministic node

Figure 2.2: Different versions of an exemplary generative model of a simple coin-flipping experiment. I coins are flipped n times, while k_i heads are observed for coin i. k_i is assumed to be binomial-distributed with parameters θ_i and n. For the latent variable θ_i, the probability for heads, different prior distributions can be assumed: an uninformed prior Beta(1,1) (a), an informed prior Beta(10,10) (b), which assumes a fair coin with θ_i most likely close to 0.5, or a hierarchical prior Beta(α, β) (c) with uninformed hyperpriors on the hyperparameters α and β. In (d), the hierarchical model (c) additionally includes a deterministic variable η_i, the probability for tails.

these data we now want to know if the coins are fair. We assume that the observed number of heads k_i is generated from a binomial distribution with parameters θ_i and the number of total flips n. However, we do not know the value of θ_i, it is a latent parameter. Our prior knowledge of θ_i before seeing any data k_i is captured in its prior, for which we have different options:

- If we expect all possible values for θ_i in the range $[0, 1]$ to be equally likely, we set an uninformative $\text{Beta}(1, 1)$ prior on θ_i, which is the uniform distribution between 0 and 1 (Figure 2.2(a)).
- Since we know that usually coins are fair and their θ_i is thus close to 0.5, we can also choose an informative prior $\text{Beta}(10, 10)$, with its mode at 0.5 (Figure 2.2(b)).
- We can also consider the parameters of the beta prior as latent variables, put a prior on them, and infer them. The parameters of the prior are hyperparameters. In the example in Figure 2.2(c) the so-called hyperpriors on the hyperparameters α and β are uninformed gamma distributions with shape and rate set to 0.001, $\text{Gamma}(0.001, 0.001)$. If a model has a prior on hyperparameters, it is called a hierarchical generative model or a hierarchical Bayesian model.

In Figure 2.2(d) we added a deterministic variable η_i to our generative model. It simply represents the probability for tails, which is $1 - \theta_i$, which does not add much value to our exemplary model but nevertheless shows how we will represent deterministic variables in the graphical models shown in this thesis.

2.6 Parameter Estimation

As discussed in Section 2.5, probabilistic models consist of observed and latent variables, e.g., the number of observed heads k_i and the unknown probability for heads θ_i of a coin i. For inferring the value of the latent variable θ_i from the observed variable k_i or, more generally, for estimating some model parameters $\boldsymbol{\theta}$ from observed data D, there are different estimation methods.

A very popular parameter estimation method is Maximum-Likelihood Estimation (MLE), which maximizes the likelihood $p(D|\boldsymbol{\theta})$ of the data D given $\boldsymbol{\theta}$ (K. P. Murphy, 2022),

$$\boldsymbol{\theta}_{\text{MLE}} = \arg\max_{\boldsymbol{\theta}} p(D|\boldsymbol{\theta}). \tag{2.13}$$

For many models, it is computationally difficult to maximize the likelihood $p(D|\boldsymbol{\theta})$. However, if we are able to define the respective distribution's moments in closed form or can at least compute them efficiently, the simple and fast method moment matching can be used as an alternative (K. P. Murphy, 2022). For moment matching, we set up K equations with K as the number of parameters to be inferred, i.e., the size of parameter vector $\boldsymbol{\theta}$. In each of them, we equate some theoretical moment μ_k with some empirical moment $\hat{\mu}_k$ computed from observed data D. Solving these equations for $\boldsymbol{\theta}$, analytically or via optimization, provides estimates for the K parameters in $\boldsymbol{\theta}$.

Both MLE and moment matching provide estimates for parameter values $\boldsymbol{\theta}$ given observed data D. However, they do not consider prior information that we might have about $\boldsymbol{\theta}$, and they do only provide point estimates of $\boldsymbol{\theta}$. MLE and moment matching cannot provide the uncertainty over $\boldsymbol{\theta}$ given data D. In contrast, by estimating the posterior distribution over $\boldsymbol{\theta}$ given data D, $p(\boldsymbol{\theta}|D)$, Bayesian parameter inference considers prior distributions $p(\boldsymbol{\theta})$. In addition, it provides a probability distribution over $\boldsymbol{\theta}$ instead of a point estimate, which shows how uncertain we are about the estimated value of $\boldsymbol{\theta}$ given data D. Since Bayesian inference is normative (M. D. Lee & Wagenmakers, 2014), Bayesian parameter inference is also normative or optimal, meaning that the inferred posterior $p(\boldsymbol{\theta}|D)$ defines a rational agent's belief about parameters $\boldsymbol{\theta}$ given data D and the defined generative model.

If it is computationally difficult to compute the posterior distribution $p(\boldsymbol{\theta}|D)$, we can resort to Markov Chain Monte Carlo (MCMC) sampling. MCMC techniques draw samples from a posterior distribution and by this allow characterizing it without being able to compute it analytically (Van Ravenzwaaij et al., 2018). The drawn samples form a Markov chain, which means that each sample is only dependent on the previous sample. If we draw a sufficient number of samples, MCMC sampling is guaranteed to converge to the desired posterior distribution, i.e., the distribution of the drawn samples is the posterior distribution. Since the Markov chain has to first approach the region of high probability, where it then converges, the first samples of the Markov chain are usually discarded as burn-in (M. D. Lee & Wagenmakers, 2014; K. P. Murphy, 2012).

A special form of MCMC sampling is Gibbs sampling. Gibbs sampling samples from a joint distribution, i.e., from an unnormalized posterior distribution, by iteratively sampling from conditional distributions. For Gibbs sampling, we first initialize all latent variables in our model, e.g., our parameter $\boldsymbol{\theta}$. Observed variables are set to the respective observed values. Then, for N iterations we repeatedly sample the latent variables given all other (latent and observed) variables' current values from their respective conditional distribution (Algorithm 2.1). Therefore, the applicability and efficiency of Gibbs sampling is determined by how readily samples can be drawn from the conditional distributions (Bishop, 2006). Note that in Algorithm 2.1 sampling can be skipped for observed variables since their values are already known.

Algorithm 2.1 Gibbs Sampling

```
Initialize $\boldsymbol{x}^{(0)}$ containing all $M$ variables
for $n = 1, \dots, N$ do
    $x_1^{(n)} \sim p(x_1 | x_2 = x_2^{(n-1)}, x_3 = x_3^{(n-1)}, \dots, x_M = x_M^{(n-1)})$
    $x_2^{(n)} \sim p(x_2 | x_1 = x_1^{(n-1)}, x_3 = x_3^{(n-1)}, \dots, x_M = x_M^{(n-1)})$
    $\vdots$
    $x_M^{(n)} \sim p(x_M | x_1 = x_1^{(n-1)}, x_2 = x_2^{(n-1)}, \dots, x_{M-1} = x_{M-1}^{(n-1)})$
end for
```

While the required conditional distributions can be derived by hand to implement Gibbs sampling according to Algorithm 2.1 (with a good manual provided by Yildirim (2012)), there are also probabilistic programming libraries that perform Gibbs sampling automatically. An example library is JAGS (Plummer, 2003), where we only have to specify our model (as e.g., in the example in Figure 2.2) and define which variables are observed and which are latent to sample from the posterior distribution of the latent ones.

2.7 Cue Integration as Parameter Estimation

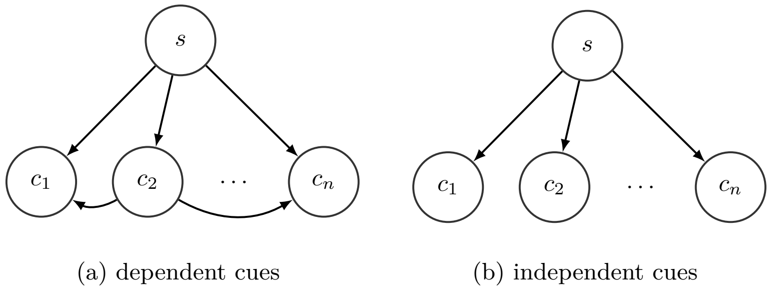
Bayes' theory, probabilistic modeling, and the respective methods for parameter estimation can be used to formally describe human behavior. A well-studied example of this is cue integration. Our environment usually provides multiple redundant sources of information that can be used to assess its properties (Landy et al., 2011). These information sources, termed cues, are sensory information that cause specific sensory measurements (Ernst & Bülthoff, 2004). Multiple cues can be different features from the same sense or modality or information from different modalities such as vision, audition, or touch (Ernst & Bülthoff, 2004; Landy et al., 2011). For example, depth can be estimated from multiple visual cues, such as texture, shading, or disparity (E. B. Johnston et al., 1993), whereas the position of another person can be inferred using multimodal cues by seeing her and hearing her talk. Likewise, we integrate auditory and visual cues when listening to other persons by hearing what they say and reading their lips (W. J. Ma et al., 2009). Also, the size of an object can be estimated using vision and touch cues (Ernst & Banks, 2002; Gepshtein et al., 2005; Hillis et al., 2002).

Single cues provide uncertain information because sensory information is noisy (Faisal et al., 2008; Fiser et al., 2010; Kochenderfer, 2015) and ambiguous (Fiser et al., 2010) (Section 2.1). Thus, since one cannot perfectly estimate the environment's properties using one single cue, it is beneficial to integrate multiple cues. In particular, cue integration can increase the reliability of an estimate and thus decrease its uncertainty if the cues are integrated in a rational way (Landy et al., 2011).

The generative model of cue integration is illustrated in the graphical model in Figure 2.3(a). The state of the world s generates some cues $c_1, \dots, c_n$, which might depend on each other in some way. Integrating the cues $c_1, \dots, c_n$ is equivalent to estimating parameter s, the unknown world state, from observed values of the cues $c_1, \dots, c_n$. Integrating the cues in a rational or normative way is achieved with Bayesian parameter inference as in Section 2.4. Thus, the model in Figure 2.3(a) is a normative model of cue integration.

Given multiple cues $c_1, \dots, c_n$, we obtain an estimate of the current state of the world s by computing the posterior distribution

$$p(s|c_1, \dots, c_n) \propto p(c_1, \dots, c_n|s)p(s) \tag{2.14}$$



(a) dependent cues

(b) independent cues

Figure 2.3: Generative models of cue integration with dependent (a) and independent (b) cues $c_1, \dots, c_n$ generated from state s.

from the joint likelihood $p(c_1, \ldots, c_n|s)$ and our prior on s $p(s)$. If the cues are conditionally independent given s, the generative model of cue integration simplifies to the graphical model shown in Figure 2.3(b), and we can simplify the posterior to a product of the prior $p(s)$ and the likelihoods of individual cues c_i $p(c_i|s)$

$$p(s|c_1, \ldots, c_n) \propto p(s) \prod_{i=1}^{n} p(c_i|s). \tag{2.15}$$

If we assume conditionally independent unbiased cues given state s and a uniform prior $p(s)$ and further assume that the individual cues c_i are Gaussian distributed with means μ_i and variances σ_i^2 given state s, i.e., their likelihoods $p(c_i|s)$ are

$$p(c_i|s) \propto \mathcal{N}(\mu_i, \sigma_i^2), \tag{2.16}$$

according to (2.15), the posterior over s is a product of the Gaussian likelihoods $\mathcal{N}(\mu_i, \sigma_i^2)$, which is proportional to a Gaussian distribution

$$p(s|c_1, \ldots, c_n) \propto \mathcal{N}\left(\frac{\sum_{i=1}^{n} \frac{\mu_i}{\sigma_i^2}}{\sum_{i=1}^{n} \frac{1}{\sigma_i^2}}, \frac{1}{\sum_{i=1}^{n} \frac{1}{\sigma_i^2}}\right) \quad \text{(Bromiley, 2003).} \tag{2.17}$$

The above posterior can be rewritten as

$$p(s|c_1, \ldots, c_n) \propto \mathcal{N}(\mu_{\text{integrated}}, \sigma^2_{\text{integrated}}) \tag{2.18}$$

with mean

$$\mu_{\text{integrated}} = \sum_{i=1}^{n} w_i \mu_i \tag{2.19}$$

with weights

$$w_i = \frac{\frac{1}{\sigma_i^2}}{\sum_{j=1}^{n} \frac{1}{\sigma_j^2}} \quad \text{(Cochran, 1937)} \tag{2.20}$$

and variance

$$\sigma^2_{\text{integrated}} = \frac{1}{\sum_{i=1}^{n} \frac{1}{\sigma_i^2}}. \tag{2.21}$$

Note that in this formulation with weights w_i the posterior can be computed as a linear combination of the individual estimates, which are the means μ_i of the Gaussian likelihoods in (2.16).

If we soften our assumption of a uniform prior but instead assume a Gaussian prior on s, the posterior $p(s|c_1, \ldots, c_n)$ is a product of Gaussian likelihoods and the Gaussian prior and thus again a Gaussian with mean and variance according to (2.19) and (2.21). The prior can be straightforwardly treated as an additional cue in this case. An example of the integration of two cues c_1 and c_2 with a Gaussian prior on s is given in Figure 2.4.

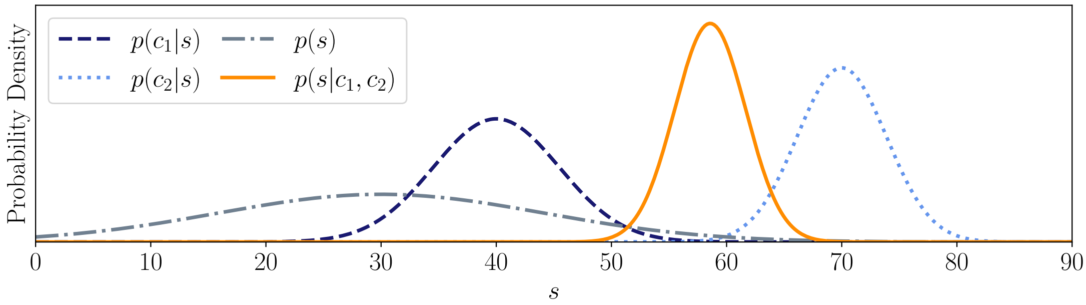


Figure 2.4: Exemplary Bayesian integration of two cues c_1 and c_2 in order to estimate the state of the world s. c_1 and c_2 are Gaussian distributed given s with $\mu_1 = 40$, $\sigma_1^2 = 30$ and $\mu_2 = 70$, $\sigma_2^2 = 15$. The prior on s is also Gaussian with $\mu_{\text{prior}} = 30$ and $\sigma^2_{\text{prior}} = 200$. The posterior on s integrating cues c_1 and c_2, $p(s|c_1, c_2)$, is also a Gaussian distribution with $\mu_{\text{integrated}} = 58.571$ and $\sigma^2_{\text{integrated}} = 9.524$, computed according to equations (2.19) and (2.21). The respective weights are $w_1 = 0.317$, $w_2 = 0.635$, $w_{\text{prior}} = 0.048$, showing that the more uncertain cue c_1 has a lower impact on the integrated estimate. Also, we see uncertainty reduction through cue integration since the integrated variance $\sigma^2_{\text{integrated}}$ is lower than the individual cues' variances.

The uncertainty of the individual cues can also be expressed as their reliability, which is the inverse of their variance $r_i = \frac{1}{\sigma_i^2}$. Writing the weights w_i using reliabilities instead of variances as

$$w_i = \frac{r_i}{\sum_{j=1}^{n} r_j} \tag{2.22}$$

shows that they are proportional to the cues' reliabilities r_i. Thus, a more reliable or certain cue with lower variance has a higher impact on the integrated estimate. This is also illustrated in the example in Figure 2.4.

The reliability of the integrated estimate is the sum of the individual cues' reliabilities,

$$r_{\text{integrated}} = \sum_{i=1}^{n} r_i. \tag{2.23}$$

Consequently, the reliability of the integrated estimate is higher than the individual cues' reliabilities, its variance and thus its uncertainty is reduced, which can also be seen in Figure 2.4. In particular, integrating cues in this way minimizes the variance of the integrated estimate (Ernst & Bülthoff, 2004; Landy et al., 2011; Scarfe, 2022) and is known as minimum variance unbiased estimator (Landy et al., 2011; Scarfe, 2022).

Prior work has found much evidence that humans integrate cues in this statistically optimal way according to a weighted linear combination of cues with weights proportional to the cues' reliabilities. Thus, human behavior in cue integration can be well described with this computational normative model. Optimal integration of different modalities' cues has been shown for vision and touch for estimating size (Ernst & Banks, 2002; Gepshtein et al., 2005; Hillis et al., 2002), the shape of real objects (Helbig & Ernst, 2007), or distances (Gepshtein & Banks, 2003). Likewise, visual and audio cues are integrated optimally for estimating the spatial localization of audio-visual stimuli (Alais & Burr, 2004). Optimal integration of different features of the same modality has been shown for various visual features. Multiple texture cues, i.e., frequency and orientation or contrast and orientation, are integrated optimally for estimating the relative location of edges (Landy & Kojima, 2001), texture and motion cues for estimating the three-dimensional shape of objects (Scarfe & Hibbard, 2011), texture, stereo, and motion cues for estimating the three-dimensional location of objects (Svarverud et al., 2010), and two focus cues, i.e., eye accommodation and the gradient of retinal blur, for estimating the slant of a surface (Watt et al., 2005). In addition, much evidence for statistically optimal cue integration according to weighted linear combinations was found for the integration of texture and disparity cues for estimating surface slant (Girshick & Banks, 2009; Hillis et al., 2002; Hillis et al., 2004; Knill & Saunders, 2003; Saunders & Chen, 2015). In contrast to all other studies, Saunders and Chen (2015) also integrates an informative Gaussian prior, treating it like an additional cue, as explained above.

Nearly all cue integration approaches assume conditionally independent cues or a very small correlation that allows approximating cue integration with the independence assumption (Scarfe, 2022). While assuming conditional independence of the cues given the true state s as for the linear weighting above is mathematically convenient, it is not always realistic. If the cues are information from different modalities or senses (such as vision and touch in the work of Ernst and Banks (2002)), the independence assumption is plausible (Ernst & Bülthoff, 2004; Oruç et al., 2003), but not if the cues are distinct features from the same modality (Oruç et al., 2003). Oruç et al. (2003) showed how to correct the weights in (2.19) for correlated cues. In particular, the weights should be computed as in (2.22) but with corrected reliabilities. For two correlated cues, the corrected reliabilities are reduced by a term proportional to their correlation ρ,

$$r_i' = r_i - \rho\sqrt{r_1 r_2}, \quad i = 1, 2. \tag{2.24}$$

This linear weighting with corrected reliabilities still maximizes the integrated reliability. However, this integrated reliability is lower than it would be for uncorrelated cues (Oruç et al., 2003). For some human subjects, Oruç et al. (2003) showed that they integrate the two cues texture gradient and linear perspective for estimating the slant of a plane according to a linear combination with corrected weights as in (2.24). However, the respective correlations were derived from the inferred weights of the subjects instead of estimated separately. Other subjects' behavior could not be described well with a linear combination with weights as in (2.24). Future research may investigate more broadly how humans integrate correlated cues.

While in multiple studies humans were shown to integrate cues according to a linear weighting approach, note that for all those studies, it is not clear if humans internally perform linear weighting while computing appropriate weights from some perceived stimulus properties that co-vary with cue uncertainty or if they represent and multiply the Gaussian likelihoods' densities in their brains (Knill & Pouget, 2004). However, a strong argument for the latter are perceptual problems for which the assumptions given above are not met, e.g., no Gaussian distributions can be assumed and thus the linear weighting approach is not appropriate. Nonetheless, also for such problems, human behavior is well described by the resulting normative models of cue integration.

A popular example of this is the work of Saunders and Knill (2001), who examine how humans integrate stereo and skew symmetry cues for estimating the orientation of a surface. While they assume stereo and skew symmetry cues to be conditionally independent, the likelihood for skew symmetry is strictly non-Gaussian. Therefore, they propose a non-linear Bayesian model for cue integration according to (2.15) with the non-Gaussian likelihood for skew symmetry and a Gaussian likelihood for stereo. Their model describes human cue integration well, unlike the weighted linear combination approach in (2.19).

In any case, regardless of any assumptions on independence or special distributions for the cues' likelihoods, Bayes' rule (2.14) with appropriate likelihood functions and prior distributions is the rational choice for cue integration. The cues as well as the target variable s might be continuous or discrete, and appropriate likelihood functions can model the cues' distributions conditioned on the state s. In particular, with a joint likelihood for all cues, as in (2.14), independence as well as dependence between the cues can be modeled.

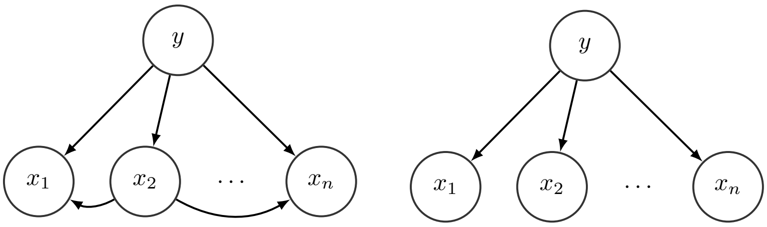
Depending on the assumptions about the data-generating process, i.e., whether the state s is discrete or continuous, how the cues are generated from s, and whether they are independent or dependent, it can be difficult to define a normative model, i.e., to choose appropriate likelihood functions and priors. Also, depending on this choice, inference in the model and thus cue integration can be more difficult than with Gaussian likelihoods. In particular, the posterior over s might not be available in closed form, in which case we have to resort to approximate parameter estimation algorithms, such as Gibbs sampling (Section 2.6).

2.8 Classifier Fusion as Parameter Estimation

Similarly to cue integration, classification can also be seen as parameter estimation in a generative model. Here, the generative model consists of a discrete truth value y that generates some features $x_1, \ldots, x_n$, which might depend on each other in some way (Figure 2.5(a)). The truth value y can take on C different values, where C is the number of possible classes to be distinguished. The task in classification is to estimate the latent truth value y from observed features $x_1, \ldots, x_n$.

As for cue integration, estimating the truth value y from features $x_1, \ldots, x_n$ in a normative or optimal way is achieved by inferring the posterior distribution of y given $x_1, \ldots, x_n$

$$p(y|x_1, \ldots, x_n) \propto p(x_1, \ldots, x_n|y)p(y) \tag{2.25}$$



(a) dependent features

(b) independent features

Figure 2.5: Generative models of classification with dependent (a) and independent (b) features $x_1, \ldots, x_n$ generated from truth value y.

from the joint likelihood $p(x_1, \ldots, x_n|y)$ and some prior on y $p(y)$. If conditional independence of the features given y can be assumed, the generative model of classification reduces to the graphical model shown in Figure 2.5(b), and the posterior distribution is proportional to a product of the prior $p(y)$ and the likelihoods of individual features x_i $p(x_i|y)$

$$p(y|x_1, \ldots, x_n) \propto p(y) \prod_{i=1}^{n} p(x_i|y). \tag{2.26}$$

Classifiers are algorithms to estimate the truth value y from observed features $x_1, \ldots, x_n$. In particular, they are functions f that map the input features $x_1, \ldots, x_n$ to a truth value or class label y (K. P. Murphy, 2022),

$$f : x_1, \ldots, x_n \to y. \tag{2.27}$$

To quantify the uncertainty of a classifier's prediction y for specific features $x_1, \ldots, x_n$, probabilistic classifiers do not directly output a discrete truth value y but a categorical probability distribution $p(y|x_1, \ldots, x_n)$ over all possible classes,

$$f : x_1, \ldots, x_n \to p(y|x_1, \ldots, x_n), \tag{2.28}$$

which is a vector of dimensionality C consisting of probabilities $p_1, \ldots, p_C$ (K. P. Murphy, 2022). p_i represents the conditional probability that y is of value i given $x_1, \ldots, x_n$, $p(y = i|x_1, \ldots, x_n)$. Accordingly, all probabilities p_i sum to 1. A discrete predicted label $\hat{y}$ can be obtained from the returned categorical distribution by selecting the most likely class, $\hat{y} = \arg\max_i p(y = i|x_1, \ldots, x_n)$.

A training set of L labeled instances $\{\boldsymbol{x_l}, y_l\}_{l=1}^{L}$ with $\boldsymbol{x_l}$ as the feature vector $\boldsymbol{x} = [x_1, \ldots, x_n]$ of instance l can serve to train the classifier. The trained classifier can then be used to classify a test set of M unlabeled instances $\{\boldsymbol{x_m}\}_{m=1}^{M}$, i.e., to return their predicted labels $\hat{y}_m$ or the respective categorical probability distributions $p(y|\boldsymbol{x_m})$.

Training and classification can be realized by defining a generative model as in Figure 2.5 with appropriate likelihood functions and priors, learning its parameters from observed training data, and inferring the posterior $p(y|x_1, \ldots, x_n)$ of new unseen features $x_1, \ldots, x_n$ according to (2.25) or (2.26) (Bishop, 2006). Examples of such generative classifiers are the Naïve Bayes classifier and logistic regression as a special case of a Naïve Bayes classifier (Bishop, 2006). However, sometimes it is hard to find and infer a generative model of the data. In these cases, discriminative models can be used, which assume some functional form of the probability $p(y|x_1, \ldots, x_n)$ and directly estimate its parameters from observed training data (Bishop, 2006). Logistic Regression is usually used as a discriminative model, where its functional form of $p(y|x_1, \ldots, x_n)$, which can also be derived with a generative model, is directly fitted to the training data (K. P. Murphy, 2022). Other examples of discriminative probabilistic classifiers are decision trees or Multilayer Perceptrons with a logistic or softmax final layer (K. P. Murphy, 2022).

Individual classifiers can be fused to improve the classification performance (Bishop, 2006; Dietterich, 2000; Hamed & Akbari, 2018; Kittler et al., 1998; Mohandes et al., 2018; Pirs & Strumbelj, 2019). Classifier fusion methods combine individual classifiers either on the data level, on the feature level, or on the decision level (Mohandes et al., 2018). For data fusion, the data used to train the individual classifiers, which might come from different sensors or modalities, are combined. The combined collection of data can then be preprocessed to features and used for training a classifier. In the generative model in Figure 2.5 the features $x_1, \ldots, x_n$ thus comprise the data used to train different individual classifiers. For feature fusion, the feature vectors of the individual classifiers are concatenated and then used to train a single classifier, so the features $x_1, \ldots, x_n$ in the generative model in Figure 2.5 are features used to train different individual classifiers. The advantage of data fusion and feature fusion is that by using raw data or features, interactions between data or features of different individual classifiers can be considered for classification with the fused classifiers (Tulyakov et al., 2008). However, the increased amount of data and long feature vectors increase the complexity of training and classification (Tulyakov et al., 2008). Also, if a new individual classifier is added to an existing ensemble of classifiers, the complete ensemble has to be retrained (Schuldhaus et al., 2013). And finally, if the individual classifiers all use the same data or features, data or feature fusion is useless. In contrast, for decision fusion, the individual classifiers with their individual data and features are trained, and their outputs on new, unseen data are fused using a specific fusion method, which then returns the fused result. Thus, decision fusion can also be described with the generative model of classification in Figure 2.5 if the features $x_1, \ldots, x_n$ are the outputs of n individual classifiers, from which the truth value y needs to be inferred. Decision fusion allows efficient training of individual classifiers of low complexity which can then be combined to obtain a high-performing fused classifier. Decision fusion is even possible without knowing the features and exact classification method used for the individual classifiers (Tulyakov et al., 2008). The individual classifications can also be performed by human forecasters. Moreover, additional classifiers can be sequentially added, and others can be replaced without retraining the complete ensemble, allowing flexibility and modularity.

Decision fusion can be performed on three levels. On the abstract level, discrete class labels are combined, while on the rank level, a ranking of the different possible classes is used as input for the fusion method. On the score level, the categorical probability distributions returned by probabilistic classifiers are fused (Mohandes et al., 2018), which allows considering the individual classifiers' uncertainties quantified in their output distributions. A fused classifier F that performs decision fusion on the score level can be formalized as

$$F : x_1 = f_1(\boldsymbol{z}) = p_1(y|\boldsymbol{z}), \ldots, x_n = f_n(\boldsymbol{z}) = p_n(y|\boldsymbol{z}) \rightarrow y, \tag{2.29}$$

It estimates the truth value of y given features $x_1, \ldots, x_n$, where the features are the categorical probability distributions over y given some feature vector $\boldsymbol{z}$ returned by each individual classifier f_j, $p_j(y|\boldsymbol{z})$, $j = 1, \ldots, n$. In most cases, decision fusion on the score level also returns a probability distribution over the truth value y given the individual classifiers' probabilistic outputs $x_1, \ldots, x_n$, which results in a fused classifier

$$F : x_1 = f_1(\boldsymbol{z}) = p_1(y|\boldsymbol{z}), \ldots, x_n = f_n(\boldsymbol{z}) = p_n(y|\boldsymbol{z}) \rightarrow p(y|x_1, ..., x_n), \tag{2.30}$$

Note that the individual classifiers f_j can also observe different feature vectors $\boldsymbol{z}_j$, e.g., from different modalities. While the performance of such a fused classifier F is of course dependent on the performances of the individual classifiers f_j, the fusion method used for combining their outputs is of particular importance.

Several fusion methods have been proposed. Among the most popular ensemble methods are Bagging (Breiman, 1996) and Boosting (Freund & Schapire, 1996). However, they do not only combine outputs of multiple individual classifiers as it is done for decision fusion but also train them either with randomly generated training subsets for Bagging or with iteratively changing weights on training examples depending on previous misclassifications for Boosting.

In contrast, decision fusion methods assume the individual classifiers to be given and fixed. The fusion methods can be seen as algorithms to approximate the truth value y from observed outputs $x_1 = f_1(\boldsymbol{z}), \ldots, x_n = f_n(\boldsymbol{z})$ of n individual classifiers f_j given some feature vector $\boldsymbol{z}$. There is a large number of ad-hoc fusion rules that apply a specific function to combine individual classifier outputs. On the abstract fusion level, majority voting or weighted majority voting with weights reflecting the individual classifiers' performance are commonly used ad-hoc fusion rules (Mohandes et al., 2018). However, since we are particularly interested in the classifications' uncertainties, here we focus on fusion methods operating on the score level, i.e., fusing the categorical probability distributions returned by the individual classifiers. The most common ad-hoc fusion methods on the score level are the sum rule, the median rule, and the product rule, which combine the probability distributions returned by the individual classifiers using a sum, a product, or the median respectively, and the min and max rules, which use the minimum or maximum probability for each class out of all returned individual probability distributions to build the fused classifier output (Kittler et al., 1998; Mi et al., 2016). For all these rules, the final fused output vector has to be normalized to sum to 1 to result in a probability distribution if knowledge of the uncertainty of the fused output is required. The sum rule can also be extended to weighted averages that assign different weights to different classifiers, e.g., according to their performance or reliability. L. Xu and Amari (2009) provide some examples of weighted averages for decision fusion on the score level.

In addition to these ad-hoc fusion rules, other approaches explicitly learn a mapping between the n individual classifiers' outputs $x_1 = f_1(\boldsymbol{z}), \ldots, x_n = f_n(\boldsymbol{z})$ and the truth label y. These approaches can be subsumed as stacking approaches. For stacking or stacked generalization (Wolpert, 1992), the outputs of the individual base classifiers on some (test) data set D are used to build a new data set D', consisting of the base classifiers' outputs and the respective truth values. Using this new data set D', a meta classifier can be trained, which classifies an example based on the outputs of the base classifiers for this respective example. Thus, this meta classifier learns how the base classifiers behave for different truth values. In particular, it can learn the base classifiers' bias, which defines a systematic deviation of their classifications from the truth value, and their variance, which describes the variability of their classifications, and can also correct for a potential bias. Stacking can be performed on all levels of decision fusion, the abstract, rank, and score level. However, stacking works best on the score level using probabilistic base classifiers (Ting & Witten, 1999). In this case, the meta classifier can also learn the base classifiers' uncertainty in addition to bias and variance. While usually stacking is only done on two levels, the base and the meta level, it can be extended to a stacking pipeline of multiple levels. Likewise, stacking can also be performed with a single base classifier. In this case, the respective meta classifier can improve the classification performance because it learns the bias of the base classifier and can correct for it and can also use information on variance and uncertainty to provide better classifications. In fact, Wolpert (1992) suggests using stacking in almost any classification problem in order to increase the classification performance.

The actual classification method of the meta classifier can be chosen arbitrarily. There are approaches that fuse probabilistic base classifiers but use non-probabilistic meta classifiers, e.g., multi-response linear regression (Ting & Witten, 1999) or decision templates, which learn the average outputs of all base classifiers for different classes and compute a distance measure between the templates and a new unseen example for classification (Kuncheva et al., 2001; Mi et al., 2016). However, although fusing probabilistic classifiers, these meta classifiers do not provide the uncertainty of their final fused classification. Therefore, other stacking approaches use probabilistic meta classifiers, such as Naïve Bayes (Ting & Witten, 1999), neural networks (D.-S. Lee & Srihari, 1995; Mohammed et al., 2021), or decision trees (Elmannai et al., 2022; Ting & Witten, 1999), or meta classifiers that at least provide estimates of probability, e.g., Support Vector Machines (Elmannai et al., 2022), nearest neighbor classifiers (Elmannai et al., 2022; Ting & Witten, 1999), or Random Forests (Elmannai et al., 2022). In addition, the meta classifiers can be Bayesian models, which explicitly define and learn a generative model of the probabilistic classifier outputs, i.e., how the outputs $x_1 = f_1(\boldsymbol{z}), \ldots, x_n = f_n(\boldsymbol{z})$ of n base classifiers are generated from the truth value y according to Figure 2.5. Using this generative model, the posterior over y given new unseen classifier outputs $x_1 = f_1(\boldsymbol{z}), \ldots, x_n = f_n(\boldsymbol{z})$ can be estimated according to (2.25) or (2.26). Given that the assumed model correctly describes the data-generating process, fusing classifier outputs in this way is normative. Stacking approaches using Bayesian models as meta classifiers are also called supra-Bayesian approaches for classifier fusion. Examples of such approaches on the score level that are fusing probabilistic classifiers are the works of Nazabal et al. (2016) and Pirs and Strumbelj (2019).

As can be seen, there are many different possible methods for fusing probabilistic classifier outputs, including supra-Bayesian approaches that compute a posterior distribution over the truth value y given individual classifiers' outputs in a generative model. Still, an open question is how to optimally combine probabilistic classifier outputs in order to correctly consider their uncertainty and reduce the final fused classification's uncertainty, as shown by human participants in many experiments on cue integration in human perception (Section 2.7). Whereas for many of these cue integration tasks performed by humans it has been shown that the considered cues can be assumed to be conditionally independent and Gaussian-distributed, this is not the case if we fuse categorical output distributions of probabilistic classifiers. Here, different assumptions about the generative model need to be made. In particular, different prior distributions and likelihood functions need to be chosen to be able to fuse probabilistic classifiers in a normative way. Also, a potential correlation between individual base classifiers should be considered since classifiers trained on the same target are usually correlated (Jacobs, 1995; Kim & Ghahramani, 2012). This dependence between individual classifiers' outputs significantly complicates the definition of a generative model for classifier fusion according to Figure 2.5(a).

Modeling probabilistic classifier outputs in such a generative model in order to optimally combine them with Bayes' rule will be discussed in the next chapter, with a special focus on a potential correlation between the individual classifiers to be fused (Chapter 3). Chapter 4 discusses the bivariate beta distribution, which can, e.g., be used to model correlated outputs of binary classifiers, and provides a method for estimating its parameters. Chapters 5 – 7 will subsequently present three applications of Bayes optimal classifier fusion in different human-robot interaction tasks. In addition, the idea of probabilistically modeling classifiers is transferred to the combination of subjective probability estimates provided by human forecasters in Chapters 8 and 9.

3 NORMATIVE BAYESIAN CLASSIFIER FUSION

Classification is one of the fundamental tasks in machine learning with broad applicability in many domains. The most successful classification methods, e.g., in machine learning competitions, have proven to be classifier ensembles, which combine different classifiers to improve classification performance (Dietterich, 2000; Kittler et al., 1998; Kuncheva, 2014; Mohandes et al., 2018; Pirs & Strumbelj, 2019). Apart from the selection and training of individual classifiers, the fusion method used for classifier combination is of particular importance for the success of an ensemble, as individual classifiers can be biased or highly variable. Such fusion methods can equivalently be applied for fusing human experts' opinions. However, for convenience, most common fusion methods assume independent classifiers (Mohandes et al., 2018; Schubert et al., 2004), although in practice, classifiers trained on the same target as well as human experts are highly correlated (Jacobs, 1995; Kim & Ghahramani, 2012; Winkler et al., 2019).

Different strategies for coping with correlated classifiers have been proposed, such as selecting only those classifiers with the lowest correlation (Faria et al., 2013; Goebel & Yan, 2004; Petrakos et al., 2000; Prabhakar & Jain, 2002; Singh et al., 2018), explicitly decorrelating the classifiers before fusion (Ulaş et al., 2012), or weighting them according to their correlation (Lacoste et al., 2014; Safont et al., 2019; Srinivas et al., 2009; Terrades et al., 2009). While there are several non-Bayesian models of improved fusion of correlated classifiers (Baertlein et al., 2001; Drakopoulos & Lee, 1988; Kam et al., 1991; A. J. Ma et al., 2013; Sundaresan et al., 2011; Veeramachaneni et al., 2008), Kim and Ghahramani (2012) introduced a Bayesian model for fusing dependent discrete classifier outputs, albeit not probabilistic outputs, thereby disregarding valuable information about the uncertainty of decisions. Pirs and Strumbelj (2019) extend the work of Kim and Ghahramani (2012) by allowing probabilistic classifier outputs. But, their focus is on outperforming related fusion algorithms using an approximate model of dependent classifiers rather than developing a theoretically justified normative model of how correlated classifier fusion should work. In particular, Pirs and Strumbelj (2019) conclude that a fusion method should not outperform the base classifiers if these are highly correlated. However, while it is known that there should be no fusion gain for a correlation of $r = 1$ between classifiers (Baertlein et al., 2001; Drakopoulos & Lee, 1988; Kuncheva & Jain, 2000; Petrakos et al., 2000; Tumer & Ghosh, 1995; Zhou, 2012), this has not been shown for probabilistic classifiers. Here, we clarify how the correlation between classifiers affects uncertainty reduction through fusion in general, which is well known in the case of fusing independent probabilistic classifier outputs (Andriamahefa, 2017).

Therefore, in order to show how correlated probabilistic classifier outputs should be fused Bayes optimally, in this work we introduce a hierarchical fully Bayesian normative model of the fusion of correlated probabilistic classifiers. We model the classifiers to be fused with a new correlated Dirichlet distribution, which is able to model Dirichlet-distributed random vectors with positive correlation. We derive this model by progressively generalizing its assumptions and, in this course, also show how to fuse independent classifiers Bayes optimally. Also, we show that existing fusion methods such as Independent Opinion Pool are special cases of this model. Evaluations on simulated and real data reveal that fusion should reduce uncertainty the less, the higher the classifiers are correlated. In particular, if the classifiers' correlation is 1, there should be no uncertainty reduction through fusion. Still, since we learn a model of each base classifier, this does not necessarily mean that the fused distribution equals the base distributions. Empirical evaluations show the approach's superiority on real-world fusion problems.

The rest of the chapter is structured as follows. Section 3.1 discusses related work. In Section 3.2 we derive Bayes optimal fusion behavior given successively more general assumptions, including Independent Opinion Pool, a Bayesian model for fusing independent classifiers, and the proposed Bayesian model for fusing correlated classifiers. These models are evaluated on simulated and real data sets in Section 3.3. In Section 3.4 we conclude and discuss limitations and future work.

3.1 Related Work

Bayesian models of classifier fusion are known as Supra-Bayesian fusion approaches (Jacobs, 1995). For combining expert opinions, they have already been proposed before machine learning methods emerged. Considering the opinions as data, a probability distribution is learned over them, conditional on the true outcome. From this expert model, a decision maker can compute the likelihood of observed opinions and combine it with its prior using Bayes' rule. The resulting posterior distribution over the possible outcomes is the fusion result (Genest, Zidek, et al., 1986). For instance, Lindley (1985), French (1980), and Winkler (1981) modeled experts' opinions using a multivariate normal distribution, which enabled explicit modeling of their correlations, while Jouini and Clemen (1996) used copulas to model experts' correlations.

Such Supra-Bayesian approaches have also been proposed for classifier fusion. Kim and Ghahramani (2012) model independent discrete classifier outputs by learning a multinomial distribution over each row of the classifiers' confusion matrices, conditioned on the true class label. This Independent Bayesian Classifier Combination Model (IBCC) is additionally extended to a Dependent Bayesian Classifier Combination Model (DBCC), which uses Markov networks to model correlations. Inference is realized with Gibbs sampling, and training is unsupervised. Several authors have extended the work of Kim and Ghahramani (2012). However, most of them extend the IBCC method, which assumes independent classifiers. For example, Simpson et al. (2013) infer the IBCC parameters with variational inference instead of Gibbs sampling. Hamed and Akbari (2018) instead presented a supervised extension of IBCC. Ueda et al. (2014) additionally introduce another latent variable into the original IBCC model that determines a classifier's effectiveness, i.e., whether it always outputs the same label for a class or varies considerably. Still, as in the work of Kim and Ghahramani (2012), this line of work considers discrete classifier outputs without utilizing classifiers' uncertainties for fusion. Thus, Nazabal et al. (2016) introduced a Bayesian model for fusing probabilistic classifiers that output categorical distributions instead of only discrete class labels. The output distributions of each classifier are modeled with a Dirichlet distribution conditioned on the true class label. Parameter inference is realized with Gibbs sampling on labeled training data. However, similar to the approaches above, the model assumes independent base classifiers and disregards potential correlations.

In contrast, Pirs and Strumbelj (2019) explicitly model correlations between probabilistic classifiers. They transform the classifiers' categorical output distributions with the inverse additive logistic transform and model the resulting real-valued vectors with mixtures of multivariate normal distributions with means and covariances conditioned on the true class labels. While Pirs and Strumbelj (2019) show that this model outperforms other Bayesian fusion methods on most data sets, the model does not provide a normative account of how fusion of correlated probabilistic classifiers should work Bayes optimally. In particular, they conclude that a fused classifier cannot outperform the base classifiers if these are highly correlated and provide empirical evidence for this conclusion based on one data set. However, this has not been proven for probabilistic classifiers, where a special focus should be on uncertainty reduction through fusion. To investigate how this uncertainty reduction should be affected by correlation, we propose a normative hierarchical Bayesian generative model of the fusion of correlated probabilistic classifiers. The model's structure resembles the structure presented by Pirs and Strumbelj (2019) up to a newly introduced conjugate prior of the categorical distribution, a correlated Dirichlet distribution for jointly modeling the classifier outputs. In contrast to Pirs and Strumbelj (2019), we do not require any transformation of the classifier outputs or mixture distributions and show that the fused classifier can outperform the base classifiers, even for highly correlated base classifiers.

3.2 Bayesian Models of Classifier Fusion

Throughout this work, we assume K base classifiers $C_k, k = 1, \dots, K$ to be given and fixed. For a given example i, each base classifier C_k receives observation o_i^k with corresponding true class label $t_i = 1, \dots, J$. Based on observation o_i^k, each classifier C_k outputs the respective probability distribution $p(t_i|o_i^k)$, which is a J-dimensional categorical distribution. The goal of the present work is to fuse these given classifier outputs $p(t_i|o_i^k)$ in order to obtain $p(t_i|o_i^1, \dots, o_i^K)$. Accordingly, in the following we investigate Bayes optimal fusion methods with successively more general assumptions. In Section 3.2.1 we start with assuming independent classifiers whose behavior is not known. In Section 3.2.2 we proceed by modeling each individual classifier's behavior while still assuming independence. The resulting Independent Fusion Model is finally extended to the Correlated Fusion Model in Section 3.2.3, which explicitly models classifiers' correlations. Our implementation of the proposed fusion methods is publicly available at `https://github.com/RothkopfLab/Bayesian_Correlated_Classifier_Fusion`.

3.2.1 Independent Opinion Pool

By applying Bayes' rule we can transform the sought $p(t_i|o_i^1, \dots, o_i^K)$ to

$$p(t_i|o_i^1, \dots, o_i^K) = \frac{p(o_i^1, \dots, o_i^K|t_i)p(t_i)}{p(o_i^1, \dots, o_i^K)}. \tag{3.1}$$

If we assume conditional independence of the observation o_i^k given the true class t_i and expand the fraction by $p(t_i)^{K-1}$, we can reformulate this to

$$p(t_i|o_i^1,\dots,o_i^K) = \frac{\prod_{k=1}^K p(o_i^k|t_i)p(t_i)^K}{p(o_i^1,\dots,o_i^K)p(t_i)^{K-1}}. \tag{3.2}$$

By again applying Bayes' rule and commutativity we get

$$\begin{aligned} p(t_i|o_i^1,\dots,o_i^K) &= \frac{\prod_{k=1}^K \left(\frac{p(t_i|o_i^k)p(o_i^k)}{\cancel{p(t_i)}}\cancel{p(t_i)}\right)}{p(o_i^1,\dots,o_i^K)p(t_i)^{K-1}} \\ &= \frac{\prod_{k=1}^K p(t_i|o_i^k)p(o_i^k)}{p(o_i^1,\dots,o_i^K)p(t_i)^{K-1}} \\ &= \frac{\prod_{k=1}^K p(t_i|o_i^k)\prod_{k=1}^K p(o_i^k)}{p(o_i^1,\dots,o_i^K)p(t_i)^{K-1}} \\ &= \underbrace{\frac{\prod_{k=1}^K p(o_i^k)}{p(o_i^1,\dots,o_i^K)}}_{\text{constant}} \frac{\prod_{k=1}^K p(t_i|o_i^k)}{p(t_i)^{K-1}}, \end{aligned} \tag{3.3}$$

where the first fraction is constant, which allows us to rewrite the expression as

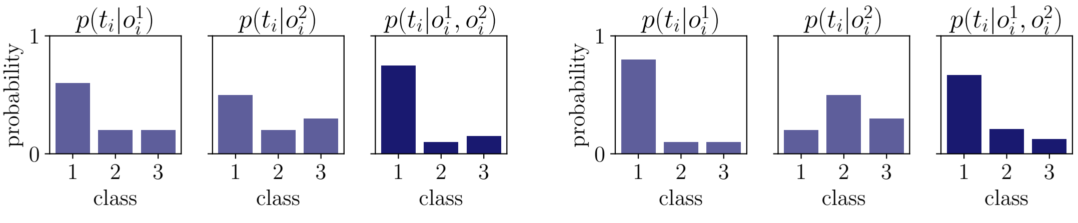
$$p(t_i|o_i^1,\dots,o_i^K) \propto \frac{\prod_{k=1}^K p(t_i|o_i^k)}{p(t_i)^{K-1}}. \tag{3.4}$$

When assuming an uninformed prior on $p(t_i)$ this simplifies to a product of the categorical probability distributions outputted by the individual base classifiers

$$p(t_i|o_i^1,\dots,o_i^K) \propto \prod_{k=1}^K p(t_i|o_i^k), \tag{3.5}$$

which needs to be renormalized to sum to 1.

Thus, if we assume that the observations and with them the outputs of all base classifiers are conditionally independent given t_i with an uninformed prior on t_i, the Bayes optimal



(a) (b)

Figure 3.1: Two examples of fusion with Independent Opinion Pool. In (a), fusing two non-conflicting distributions leads to uncertainty reduction, while in (b) the fusion of two conflicting distributions increases the uncertainty of the fused distribution. In addition, (b) shows that a base distribution's impact on the fused distribution is determined by its uncertainty. The less uncertain first distribution $p(t_i|o_i^1)$ affects the fused distribution $p(t_i|o_i^1, o_i^2)$ more than the more uncertain second distribution $p(t_i|o_i^2)$.

fusion rule is a renormalized product of the individual base classifiers' categorical output distributions as in (3.5).

This fusion rule, known as Independent Opinion Pool (IOP) (Berger, 1985), leads to intuitive results regarding uncertainty. Non-conflicting base distributions reinforce each other in a way that the fused categorical distribution's uncertainty is reduced (Andriamahefa, 2017), which is shown in Figure 3.1(a). Equivalently, fusing conflicting distributions increases the uncertainty of the fused distribution (Andriamahefa, 2017), which can be seen in Figure 3.1(b). Moreover, a base distribution's impact on the fused result distribution depends on its uncertainty. In particular, the more uncertain a base distribution, the less it affects the resulting fused distribution (Hayman & Eklundh, 2002), as also seen in Figure 3.1(b).

3.2.2 Independent Fusion Model

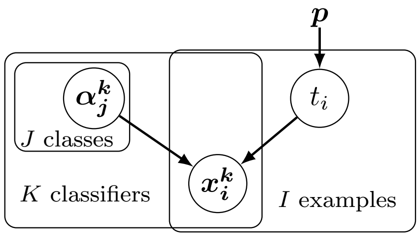
Although IOP is Bayes optimal given that the prior over t_i is uninformed and the base classifiers' observations and with them their output distributions are conditionally independent given the true label t_i, it is an ad-hoc method. Thus, only information given by the current output distributions can be exploited for fusion. The individual classifiers' properties, their bias, variance, and uncertainty, cannot be considered. Therefore, the Independent Fusion Model (IFM) additionally models the behavior of the classifiers to be fused. Since modeling each classifier's behavior requires considering their categorical output distributions as data, here we assume them as given and fixed and define them as $\boldsymbol{x_i^k} = p(t_i|o_i^k)$ for base classifier C_k and example i. We still assume an uninformed prior over t_i and conditional independence of the classifiers' output distributions $\boldsymbol{x_i^k}$.

By observing multiple training examples of classifier outputs $\boldsymbol{x_i^k}$, a probability distribution over them conditional on the true class label t_i can be learned, $p(\boldsymbol{x_i^k}|t_i)$. We set this distribution to be a Dirichlet distribution, which is the conjugate prior of the categorical distribution. Thus, if t_i can take J different values, each base classifier's outputs are modeled by J Dirichlet distributions, $p(\boldsymbol{x_i^k}|t_i = 1), \dots, p(\boldsymbol{x_i^k}|t_i = J)$. The graphical model of the proposed IFM is shown in Figure 3.2. The true label t_i of example i is modeled with a categorical distribution with parameter $\boldsymbol{p}$. If sufficient knowledge about the data is available, the prior $\boldsymbol{p}$ over true labels t_i can be chosen accordingly. For the subsequent experiments we chose an uninformed prior with $\boldsymbol{p} = (\frac{1}{J}, \dots, \frac{1}{J})$. $\boldsymbol{\alpha}$ holds the parameters of the Dirichlet distributions that model the classifiers' outputs. $\boldsymbol{\alpha_j^k}$ with $\alpha_{j\,l}^k > 0$ for $l = 1, \dots, J$ thereby contains the parameters of the Dirichlet distribution over the outputs of classifier C_k if $t_i = j$. Hence, the output $\boldsymbol{x_i^k}$ of classifier C_k for example i with true label $t_i = j$ is Dirichlet-distributed with parameter vector $\boldsymbol{\alpha_j^k}$.

A similar model was proposed by Nazabal et al. (2016). However, their model uses more parameters since they chose the parameters of Dirichlet distributions to be a product of two parameters.

3.2.2.1 *Parameter Inference*

For learning the classifier model parameters $\boldsymbol{\alpha}$, the posterior distribution over $\boldsymbol{\alpha}$ conditioned on observed classifier outputs $\boldsymbol{x}$ and the corresponding true labels $\boldsymbol{t}$, $p(\boldsymbol{\alpha}|\boldsymbol{x}, \boldsymbol{t})$, needs to be inferred. The training data $\boldsymbol{x}$ consist of I examples composed of K categorical



$$t_i \sim \text{Categorical}(\boldsymbol{p})$$
$$\boldsymbol{x_i^k} | t_i = j \sim \text{Dirichlet}(\boldsymbol{\alpha_j^k})$$

Figure 3.2: Graphical model of the Independent Fusion Model (IFM).

output distributions $\boldsymbol{x_i^k}$, and $\boldsymbol{t}$ holds I true labels t_i respectively. Inference is performed with Gibbs sampling. As an uninformed prior for all elements of $\boldsymbol{\alpha_j^k}$ we chose a vague gamma prior with shape and rate set to 10^{-3}. Of course, one could choose any other prior given additional domain knowledge about the data.

We implement Gibbs sampling using the standard inference tool JAGS (Plummer, 2003), which allows sampling given a definition of the generative model of the IFM (as shown in Figure 3.2). In the following, we take the means of inferred posterior distributions as point estimates for $\boldsymbol{\alpha_j^k}$.

3.2.2.2 Normative Fusion Behavior

For fusion, the posterior distribution over t_i given all K classifier outputs $\boldsymbol{x_i^k}$ and the learned model parameters $\boldsymbol{\alpha}$, $p(t_i|\boldsymbol{x_i^1}, \dots, \boldsymbol{x_i^K}, \boldsymbol{\alpha})$, needs to be inferred. Since the IFM is a generative model for independent categorical classifier outputs, performing fusion in this way is Bayes optimal given the model assumptions. The posterior fused distribution can be derived analytically:

The joint distribution of the Independent Fusion Model shown in Figure 3.2 is

$$p(\boldsymbol{x_i}, \boldsymbol{\alpha}, t_i) = p(t_i|\boldsymbol{p})p(\boldsymbol{\alpha}) \prod_{k=1}^{K} p(\boldsymbol{x_i^k}|\boldsymbol{\alpha_j^k}, t_i). \tag{3.6}$$

Since we observe $\boldsymbol{\alpha}$ and assume the prior over t_i, $p(t_i|\boldsymbol{p})$, to be uninformed, this can be simplified to

$$p(\boldsymbol{x_i}, \boldsymbol{\alpha}, t_i) \propto \prod_{k=1}^{K} p(\boldsymbol{x_i^k}|\boldsymbol{\alpha_j^k}, t_i). \tag{3.7}$$

The fusion rule can be obtained by computing the posterior probability of $t_i = j$ for $j = 1, \dots, J$ given the categorical distributions $\boldsymbol{x_i}$ and the respective $\boldsymbol{\alpha}$ learned before,

$$p(t_i = j|\boldsymbol{x_i}, \boldsymbol{\alpha_j}) \propto p(t_i = j, \boldsymbol{x_i}, \boldsymbol{\alpha_j}) \tag{3.8}$$

$$\propto \prod_{k=1}^{K} p(\boldsymbol{x_i^k}|\boldsymbol{\alpha_j^k}, t_i = j) \tag{3.9}$$

$$= \prod_{k=1}^{K} \mathrm{Dirichlet}(\boldsymbol{x_i^k}; \boldsymbol{\alpha_j^k}) \tag{3.10}$$

$$= \prod_{k=1}^{K} \frac{1}{\mathrm{B}(\boldsymbol{\alpha_j^k})} \prod_{l=1}^{J} (x^{k}_{i\,l})^{\alpha^{k}_{j\,l}-1}. \tag{3.11}$$

This unnormalized posterior probability can now be computed for all $t_i = j$ for $j = 1, \ldots, J$, and normalizing these values to make them sum to 1 gives the posterior fused categorical distribution.

As (3.10) and (3.11) show, using the IFM, we do not multiply the categorical output distributions of the base classifiers, such as for IOP, but their probabilities conditioned on the modeling Dirichlet distributions. Thus, fusion can take into account the variances and uncertainties of the base classifiers as well as potential learned biases.

How variance and uncertainty are considered for fusion can be demonstrated with the following example. If a classifier C_1 is modeled by three Dirichlet distributions with parameters $\boldsymbol{\alpha_1^1} = (a+n, a, a)$ for $t_i = 1$, $\boldsymbol{\alpha_2^1} = (a, a+n, a)$ for $t_i = 2$, $\boldsymbol{\alpha_3^1} = (a, a, a+n)$ for $t_i = 3$, and a classifier C_2 is modeled equivalently with $\boldsymbol{\alpha_1^2} = (b+m, b, b)$, $\boldsymbol{\alpha_2^2} = (b, b+m, b)$, $\boldsymbol{\alpha_3^2} = (b, b, b+m)$, with $a, b, n, m > 0$, we can reformulate the general fusion rule (3.11) with $K = 2$ and $J = 3$ to

$$p(t_i = j|\boldsymbol{x_i}, \boldsymbol{\alpha_j}) \propto \prod_{k=1}^{2} \frac{1}{\mathrm{B}(\boldsymbol{\alpha_j^k})} \prod_{l=1}^{3} (x^{k}_{i\,l})^{\alpha^{k}_{j\,l}-1} \tag{3.12}$$

$$\propto \prod_{k=1}^{2} \prod_{l=1}^{3} (x^{k}_{i\,l})^{\alpha^{k}_{j\,l}-1} \tag{3.13}$$

$$= (x^{1}_{i\,1})^{\alpha^{1}_{j\,1}-1} (x^{1}_{i\,2})^{\alpha^{1}_{j\,2}-1} (x^{1}_{i\,3})^{\alpha^{1}_{j\,3}-1} (x^{2}_{i\,1})^{\alpha^{2}_{j\,1}-1} (x^{2}_{i\,2})^{\alpha^{2}_{j\,2}-1} (x^{2}_{i\,3})^{\alpha^{2}_{j\,3}-1}. \tag{3.14}$$

If we now exemplarily compute this for $j = 1$, we get

$$p(t_i = 1|\boldsymbol{x_i}, \boldsymbol{\alpha_1}) \propto (x^{1}_{i\,1})^{a+n-1} (x^{1}_{i\,2})^{a-1} (x^{1}_{i\,3})^{a-1} (x^{2}_{i\,1})^{b+m-1} (x^{2}_{i\,2})^{b-1} (x^{2}_{i\,3})^{b-1} \tag{3.15}$$

$$= (x^{1}_{i\,1})^{n} (x^{2}_{i\,1})^{m} \underbrace{(x^{1}_{i\,1} x^{1}_{i\,2} x^{1}_{i\,3})^{a-1} (x^{2}_{i\,1} x^{2}_{i\,2} x^{2}_{i\,3})^{b-1}}_{\text{constant}} \tag{3.16}$$

$$\propto (x^{1}_{i\,1})^{n} (x^{2}_{i\,1})^{m}. \tag{3.17}$$

Equivalently, for $j = 2$ and $j = 3$ we get

$$p(t_i = 2|\boldsymbol{x_i}, \boldsymbol{\alpha_2}) \propto (x^{1}_{i\,2})^{n} (x^{2}_{i\,2})^{m} \tag{3.18}$$

$$p(t_i = 3|\boldsymbol{x_i}, \boldsymbol{\alpha_3}) \propto (x^{1}_{i\,3})^{n} (x^{2}_{i\,3})^{m} \tag{3.19}$$

and can thus simplify the general fusion rule in (3.11) to

$$p(t_i = j|\boldsymbol{x_i}, \boldsymbol{\alpha_j}) \propto (x^{1}_{i\,j})^{n} (x^{2}_{i\,j})^{m} \tag{3.20}$$

for $j = 1, 2, 3$, while again, the resulting values must be normalized to sum to 1. This case, which was not considered by Nazabal et al. (2016), is of particular interest, because if we set parameters $n = m = 1$, the IFM reduces to IOP (Section 3.2.1). However, increasing n and m results in lower uncertainty of the fused distribution if non-conflicting base distributions are fused. In addition, if $n > m$, C_1 has a higher impact on the fused result than C_2.

How n and m are related to variance and uncertainty of a classifier can be quantified with two properties of the Dirichlet distribution, its precision and the entropy of its mean, which is a categorical distribution. Classifier variance, which defines how concentrated the classifier's output distributions are around its average output distribution, can be quantified with the precision of its corresponding Dirichlet distributions. The precision of a Dirichlet distribution with parameter $\boldsymbol{\alpha}$ is defined as $s = \sum_j \alpha_j$. It is higher, the more concentrated the distribution is around the Dirichlet's mean $\boldsymbol{\mu}$ (J. Huang, 2005). Accordingly, the precision s_j^k of classifier k is the sum of all elements in $\boldsymbol{\alpha_j^k}$ for each $j = 1, \ldots, 3$. In general, of course, for different values of j the precision can differ. However, in the example we consider here, for simplicity it is the same for all $j = 1, \ldots, 3$ and therefore can be regarded as a measure for the classifier's variance. The higher the precision, the closer the categorical samples, i.e., the classifier outputs, are to the mean and thus the lower is the classifier's variance.

Classifier uncertainty can be described by the entropies of the modeling Dirichlet distributions' means. The mean of a J-dimensional Dirichlet distribution with parameter $\boldsymbol{\alpha}$ is a categorical distribution defined as $\boldsymbol{\mu} = \left[\frac{\alpha_1}{\sum_l \alpha_l}, \ldots, \frac{\alpha_J}{\sum_l \alpha_l}\right]$. Its entropy $\mathrm{H}_{\boldsymbol{\mu}} = -\sum_j \mu_j \log(\mu_j)$ is the higher, the more uncertain is the Dirichlet distribution's mean. Accordingly, a classifier's uncertainty can be quantified by the entropies of its respective means $\boldsymbol{\mu_j^k}$ for $j = 1, \ldots, 3$. Again, in general the entropies of the means can be different for different values of j, but due to the chosen example parameters the means' entropies are equal for $j = 1, \ldots, 3$. Therefore, we can regard this mean entropy $\mathrm{H}_{\boldsymbol{\mu_j^k}}$ as the mean entropy of the modeled classifier. The lower it is, the lower is the average uncertainty of the respective classifier.

If we increase n while a remains fixed, the precision of C_1's modeling Dirichlet distributions increases, implying a lower variance of classifier C_1. In addition, its mean entropy decreases, which we show in the following:

The mean of classifier C_1 for $j=1$ is $\boldsymbol{\mu_1^1} = \left[\frac{a+n}{3a+n}, \frac{a}{3a+n}, \frac{a}{3a+n}\right]$. Thus, its entropy is

$$\begin{aligned}
\mathrm{H}_{\boldsymbol{\mu_1^1}} &= -\left(\frac{a+n}{3a+n}\cdot\log\left(\frac{a+n}{3a+n}\right) + \frac{a}{3a+n}\cdot\log\left(\frac{a}{3a+n}\right) + \frac{a}{3a+n}\cdot\log\left(\frac{a}{3a+n}\right)\right)\\
&= -\frac{1}{3a+n}((a+n)\cdot\log(a+n) - (a+n)\cdot\log(3a+n) + a\cdot\log(a)\\
&\quad - a\cdot\log(3a+n) + a\cdot\log(a) - a\cdot\log(3a+n))\\
&= -\frac{1}{3a+n}(a\cdot\log(a+n) + n\cdot\log(a+n) - a\cdot\log(3a+n) - n\cdot\log(3a+n)\\
&\quad + a\cdot\log(a) - a\cdot\log(3a+n) + a\cdot\log(a) - a\cdot\log(3a+n))\\
&= -\frac{1}{3a+n}(a\cdot(\log(a+n) - \log(3a+n) + \log(a) - \log(3a+n) + \log(a)\\
&\quad - \log(3a+n)) + n\cdot(\log(a+n) - \log(3a+n)))\\
&= -\frac{1}{3a+n}\left(a\cdot\log\left(\frac{a^2(a+n)}{(3a+n)^3}\right) + n\cdot\log\left(\frac{a+n}{3a+n}\right)\right).
\end{aligned} \tag{3.21}$$

Differentiating $\mathrm{H}_{\boldsymbol{\mu_1^1}}$ w.r.t. n yields

$$\mathrm{H}'_{\boldsymbol{\mu_1^1}}(n) = \frac{a\left(\log\left(\frac{a^2(a+n)}{(3a+n)^3}\right) - 3\log\left(\frac{a+n}{3a+n}\right)\right)}{(3a+n)^2}. \tag{3.22}$$

The derivative $\mathrm{H}'_{\boldsymbol{\mu_1^1}}(n)$ is negative for all $a, n > 0$, since

$$\begin{aligned}
\mathrm{H}'_{\boldsymbol{\mu_1^1}}(n) = \frac{a\left(\log\left(\frac{a^2(a+n)}{(3a+n)^3}\right) - 3\log\left(\frac{a+n}{3a+n}\right)\right)}{(3a+n)^2} &< 0\\
\Leftrightarrow \log\left(\frac{a^2(a+n)}{(3a+n)^3}\right) - 3\log\left(\frac{a+n}{3a+n}\right) &< 0\\
\Leftrightarrow \log\left(\frac{a^2(a+n)}{(3a+n)^3}\right) - \log\left(\frac{(a+n)^3}{(3a+n)^3}\right) &< 0\\
\Leftrightarrow \log\left(\frac{a^2(a+n)}{(3a+n)^3}\cdot\frac{(3a+n)^3}{(a+n)^3}\right) &< 0\\
\Leftrightarrow \log\left(\frac{a^2}{(a+n)^2}\right) &< 0\\
\Leftrightarrow \frac{a^2}{(a+n)^2} &< 1\\
\Leftrightarrow a^2 &< (a+n)^2\\
\Leftrightarrow a &< a+n\\
\Leftrightarrow 0 &< n.
\end{aligned} \tag{3.23}$$

Hence, $\mathrm{H}_{\boldsymbol{\mu_1^1}}$ is decreasing if n increases, which means that higher values for n decrease the mean uncertainty of classifier C_1. Since higher values for n lead to a higher fusion impact of classifier C_1 and higher uncertainty reduction of the fused distribution, this means that a low variance and a low uncertainty of a classifier increase its fusion impact and uncertainty reduction.

If we instead increase a while n remains fixed, again the precision of C_1's modeling Dirichlet distributions increases. The variance of classifier C_1 thus decreases. In contrast, its mean entropy increases, which can be shown if we differentiate the mean entropy w.r.t. a,

$$\mathrm{H}'_{\boldsymbol{\mu}_{\mathbf{1}}^{\mathbf{1}}}(a) = \frac{-n\left(\log\left(\frac{a^2(a+n)}{(3a+n)^3}\right) - 3\log\left(\frac{a+n}{3a+n}\right)\right)}{(3a+n)^2}. \tag{3.24}$$

The derivative $\mathrm{H}'_{\boldsymbol{\mu}_{\mathbf{1}}^{\mathbf{1}}}(a)$ is positive for all $a, n > 0$, since

$$\begin{aligned}
\mathrm{H}'_{\boldsymbol{\mu}_{\mathbf{1}}^{\mathbf{1}}}(a) = \frac{-n\left(\log\left(\frac{a^2(a+n)}{(3a+n)^3}\right) - 3\log\left(\frac{a+n}{3a+n}\right)\right)}{(3a+n)^2} & > 0 \\
\Leftrightarrow \log\left(\frac{a^2(a+n)}{(3a+n)^3}\right) - 3\log\left(\frac{a+n}{3a+n}\right) & < 0 \\
\Leftrightarrow \log\left(\frac{a^2(a+n)}{(3a+n)^3}\right) - \log\left(\frac{(a+n)^3}{(3a+n)^3}\right) & < 0 \\
\Leftrightarrow \log\left(\frac{a^2(a+n)}{(3a+n)^3} \cdot \frac{(3a+n)^3}{(a+n)^3}\right) & < 0 \\
\Leftrightarrow \log\left(\frac{a^2}{(a+n)^2}\right) & < 0 \\
\Leftrightarrow \frac{a^2}{(a+n)^2} & < 1 \\
\Leftrightarrow a^2 & < (a+n)^2 \\
\Leftrightarrow a & < a+n \\
\Leftrightarrow 0 & < n.
\end{aligned} \tag{3.25}$$

Consequently, in addition to a decreased variance, the mean entropy $\mathrm{H}_{\boldsymbol{\mu}_{\mathbf{1}}^{\mathbf{1}}}$ and with it the classifier's uncertainty increases if we increase a and keep n fixed. Accordingly, decreasing a while n remains fixed leads to a decreased precision and hence an increased variance, while the mean entropy and with it the classifier's uncertainty decreases. Since according to (3.20) a (and b for classifier C_2) does not affect the fusion behavior, a classifier with a low variance and a high uncertainty thus has the same fusion impact as a classifier with a high variance and a low uncertainty. Regarding fusion, variance and uncertainty cancel out each other.

The IFM does not only consider the individual classifiers' variance and uncertainty for fusion but also their potential biases. The bias of a classifier terms the extent to which the average prediction of the classifier deviates from the true class label. A classifier C_k is biased if for its Dirichlet parameters it applies that $\arg\max \boldsymbol{\alpha}_{\boldsymbol{j}}^{\boldsymbol{k}} \neq j$ for some class j. As a consequence, also for the Dirichlet's categorical mean $\boldsymbol{\mu}_{\boldsymbol{j}}^{\boldsymbol{k}}$ it applies that $\arg\max \boldsymbol{\mu}_{\boldsymbol{j}}^{\boldsymbol{k}} \neq j$. Hence, on average, the classifier would misclassify class j as another class.

The example classifiers introduced above can be modified in order to show how biased classifiers are fused. Accordingly, in the following we derive the fusion rule for classifiers C_1 and C_2 with parameters $\boldsymbol{\alpha}_{\mathbf{1}}^{\mathbf{1}} = (a+n, a, a)$ for $t_i = 1$, $\boldsymbol{\alpha}_{\mathbf{2}}^{\mathbf{1}} = (a, a+n, a)$ for $t_i = 2$, and $\boldsymbol{\alpha}_{\mathbf{3}}^{\mathbf{1}} = (a, a, a+n)$ for $t_i = 3$ for C_1 and $\boldsymbol{\alpha}_{\mathbf{1}}^{\mathbf{2}} = (b, b+m, b)$, $\boldsymbol{\alpha}_{\mathbf{2}}^{\mathbf{2}} = (b+m, b, b)$, and $\boldsymbol{\alpha}_{\mathbf{3}}^{\mathbf{2}} = (b, b, b+m)$ for C_2 with $a, b, n, m > 0$. C_2 is a biased classifier; on average it predicts class 2 if the true label is $t_i = 1$ and class 1 if $t_i = 2$.

Given these model parameters, for $t_i = 1$ (3.20) can be transformed to

$$\begin{aligned} p(t_i = 1|\boldsymbol{x_i}, \boldsymbol{\alpha_1}) &\propto (x^1_{i\,1})^{a+n-1}(x^1_{i\,2})^{a-1}(x^1_{i\,3})^{a-1}(x^2_{i\,1})^{b-1}(x^2_{i\,2})^{b+m-1}(x^2_{i\,3})^{b-1} && (3.26)\\ &= (x^1_{i\,1})^{n}(x^2_{i\,2})^{m}\underbrace{(x^1_{i\,1}x^1_{i\,2}x^1_{i\,3})^{a-1}(x^2_{i\,1}x^2_{i\,2}x^2_{i\,3})^{b-1}}_{\text{constant}} && (3.27)\\ &\propto (x^1_{i\,1})^{n}(x^2_{i\,2})^{m} && (3.28)\end{aligned}$$

Equivalently, for $t_i = 2$ and $t_i = 3$ we get

$$\begin{aligned} p(t_i = 2|\boldsymbol{x_i}, \boldsymbol{\alpha_2}) &\propto (x^1_{i\,2})^{n}(x^2_{i\,1})^{m} && (3.29)\\ p(t_i = 3|\boldsymbol{x_i}, \boldsymbol{\alpha_3}) &\propto (x^1_{i\,3})^{n}(x^2_{i\,3})^{m} && (3.30)\end{aligned}$$

As can be seen, if classifier C_2 assigns a high probability to class 1, i.e., $x^2_{i\,1}$ is high, our model interprets this as evidence for $t_i = 2$. Without having learned the classifier's bias inherent in the learned Dirichlet parameters, high values for $x^2_{i\,1}$ would, however, be evidence for $t_i = 1$. In particular, this would be the case if Independent Opinion Pool (Section 3.2.1) was used.

Note that if we set $K = 1$ in (3.11), the IFM can also be used as a meta classifier for a single classifier C_1. This meta classifier classifies a given example i based on C_1's output distribution $\boldsymbol{x_i^1}$. Thus, we only learn a Dirichlet model of classifier C_1 instead of multiple classifiers. Conditioned on the learned model parameters $\boldsymbol{\alpha^1}$ and the single base classifier's output distribution $\boldsymbol{x_i^1}$, then the posterior distribution over all possible class labels, $p(t_i = j|\boldsymbol{x_i^1}, \boldsymbol{\alpha_j^1})$, is computed, which is the meta classifier's result.

3.2.3 Correlated Fusion Model

The IFM introduced in Section 3.2.2 enables optimal fusion of categorical output distributions of conditionally independent base classifiers, considering the base classifiers' uncertainty, bias, and variance. However, in practice most classifiers trained on the same target are highly correlated (Jacobs, 1995; Kim & Ghahramani, 2012). Therefore, we extend the IFM to a Correlated Fusion Model (CFM) to explicitly model the correlations between different classifiers' outputs. As in the IFM, we also model the categorical classifier outputs $\boldsymbol{x_i^k}$ given the true label t_i as a probability distribution. However, instead of modeling all classifiers independently with individual Dirichlet distributions, we model the joint distribution $p(\boldsymbol{x_i^1}, \ldots, \boldsymbol{x_i^K}|t_i)$ with a new correlated Dirichlet distribution that can express correlations between the classifiers' outputs.

3.2.3.1 Correlated Dirichlet Distribution

For modeling correlated classifiers' categorical output distributions with their conjugate prior, a distribution is required that can model correlations between marginally Dirichlet-distributed random variables. While previous generalizations of the Dirichlet distribution focused on more flexible *correlations between individual random vector entries* $x_1, \ldots, x_J$ of a Dirichlet variate $\boldsymbol{x}$ (Connor & Mosimann, 1969; Linderman et al., 2015; Wong, 1998), here we introduce a correlated Dirichlet distribution that models *correlations between two random vectors* $\boldsymbol{x^1} = (x^1_1, \ldots, x^1_J)$ and $\boldsymbol{x^2} = (x^2_1, \ldots, x^2_J)$ with arbitrary marginal Dirichlet distributions.

A J-dimensional correlated Dirichlet distribution is thereby constructed from $3J$ independent gamma variates $A_1^1, \dots, A_J^1, A_1^2, \dots, A_J^2, D_1, \dots, D_J$ with shape parameters $\alpha_1^1 - \delta_1, \dots, \alpha_J^1 - \delta_J, \alpha_1^2 - \delta_1, \dots, \alpha_J^2 - \delta_J, \delta_1, \dots, \delta_J$ with $\alpha_l^1, \alpha_l^2, \delta_l > 0, \alpha_l^1, \alpha_l^2 > \delta_l$, and equal rate parameter 1. $\boldsymbol{x^1} = (x_1^1, \dots, x_J^1)$ and $\boldsymbol{x^2} = (x_1^2, \dots, x_J^2)$ with:

$$x_l^k = \frac{A_l^k + D_l}{\sum_{n=1}^J A_n^k + D_n}, \qquad l = 1, \dots, J, k = 1, 2, \tag{3.31}$$

are marginally Dirichlet-distributed with Dirichlet($\boldsymbol{x^1}; \alpha_1^1, \dots, \alpha_J^1$) and Dirichlet($\boldsymbol{x^2}; \alpha_1^2, \dots, \alpha_J^2$). Their positive correlation, i.e., positive correlations between x_l^1 and x_l^2 for $l = 1, \dots, J$, is generated by the shared variables $D_1, \dots, D_J$ with the correlation parameters $\delta_1, \dots, \delta_J$. If δ_l tends to zero for $l = 1, \dots, J$, $\boldsymbol{x^1}$ and $\boldsymbol{x^2}$ are independent and each follow a standard Dirichlet distribution. If $\boldsymbol{x^1}$ and $\boldsymbol{x^2}$ have the same marginal distributions with $\boldsymbol{\alpha^1} = \boldsymbol{\alpha^2}$, their correlation tends to 1 if $\boldsymbol{\delta}$ tends to $\boldsymbol{\alpha^1} = \boldsymbol{\alpha^2}$. Thus, if $\boldsymbol{x^1}$ and $\boldsymbol{x^2}$ have different marginal distributions, the correlation is limited below 1.

The correlated Dirichlet distribution can also be constructed as a pairwise combination of three independent Dirichlet distributions, which might serve as a more intuitive interpretation of the correlated Dirichlet distribution and its correlations.

To show this we transform the $3J$ independent gamma-distributed random variables $A_1^1, \dots, A_J^1$, $A_1^2, \dots, A_J^2$, $D_1, \dots, D_J$ into three independent gamma- and three independent Dirichlet-distributed random variables $U_1, U_2, U_3, \boldsymbol{W_1}, \boldsymbol{W_2}, \boldsymbol{W_3}$ with

$$\begin{aligned}
U_1 &= \sum_{l=1}^J A_l^1, & & U_1 \sim \text{Gamma}(\upsilon_1, 1) \\
U_2 &= \sum_{l=1}^J A_l^2, & & U_2 \sim \text{Gamma}(\upsilon_2, 1) \\
U_3 &= \sum_{l=1}^J D_l, & & U_3 \sim \text{Gamma}(\upsilon_3, 1) \\
W_{1l} &= \frac{A_l^1}{\sum_{j=1}^J A_j^1}, \quad l = 1, \dots, J, & & \boldsymbol{W_1} \sim \text{Dirichlet}(\alpha_1^1 - \delta_1, \dots, \alpha_J^1 - \delta_J) \\
W_{2l} &= \frac{A_l^2}{\sum_{j=1}^J A_j^2}, \quad l = 1, \dots, J, & & \boldsymbol{W_2} \sim \text{Dirichlet}(\alpha_1^2 - \delta_1, \dots, \alpha_J^2 - \delta_J) \\
W_{3l} &= \frac{D_l}{\sum_{j=1}^J D_j}, \quad l = 1, \dots, J, & & \boldsymbol{W_3} \sim \text{Dirichlet}(\delta_1, \dots, \delta_J)
\end{aligned} \tag{3.32}$$

with

$$\upsilon_1 = \sum_{i=1}^J \alpha_i^1 - \delta_i, \qquad \upsilon_2 = \sum_{i=1}^J \alpha_i^2 - \delta_i \qquad \upsilon_3 = \sum_{i=1}^J \delta_i. \tag{3.33}$$

With these definitions we can then rewrite construction (3.31) as

$$\begin{aligned}
\boldsymbol{x^1} &= \frac{U_1}{U_1 + U_3} \cdot \boldsymbol{W_1} + \frac{U_3}{U_1 + U_3} \cdot \boldsymbol{W_3} = X'\boldsymbol{W_1} + (1 - X')\boldsymbol{W_3} \\
\boldsymbol{x^2} &= \frac{U_2}{U_2 + U_3} \cdot \boldsymbol{W_2} + \frac{U_3}{U_2 + U_3} \cdot \boldsymbol{W_3} = Y'\boldsymbol{W_2} + (1 - Y')\boldsymbol{W_3}.
\end{aligned} \tag{3.34}$$

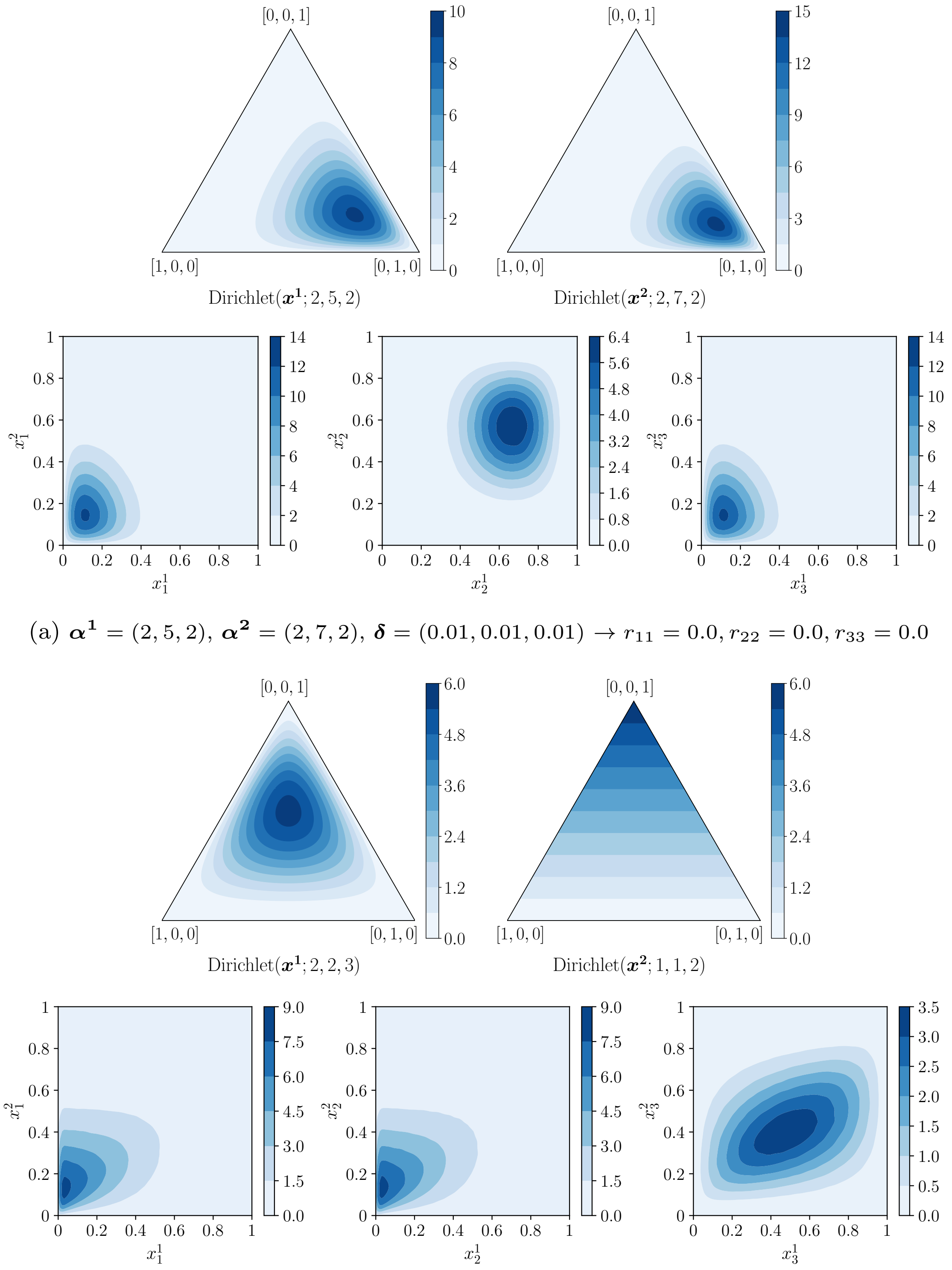
Thus, the correlated Dirichlet distribution can be constructed as a pairwise combination of the three Dirichlet distributions Dirichlet$(\alpha_1^1 - \delta_1, \ldots, \alpha_J^1 - \delta_J)$, Dirichlet$(\alpha_1^2 - \delta_1, \ldots, \alpha_J^2 - \delta_J)$, and Dirichlet$(\delta_1, \ldots, \delta_J)$. If the correlation parameters $\delta_1, \ldots, \delta_J$ tend to 0, weights X' and Y' tend to 1 and we obtain two independent Dirichlet distributions for $\boldsymbol{x^1}$ and $\boldsymbol{x^2}$, Dirichlet$(\alpha_1^1 - \delta_1, \ldots, \alpha_J^1 - \delta_J)$ and Dirichlet$(\alpha_1^2 - \delta_1, \ldots, \alpha_J^2 - \delta_J)$, which is then Dirichlet$(\alpha_1^1, \ldots, \alpha_J^1)$ and Dirichlet$(\alpha_1^2, \ldots, \alpha_J^2)$. If instead the correlation parameters tend to the marginal parameters, weights X' and Y' tend to 0 and $\boldsymbol{x^1}$ and $\boldsymbol{x^2}$ follow the same marginal Dirichlet distribution and have a correlation close to 1.

Figures 3.3 and 3.4 show four examples of correlated Dirichlet distributions with different marginal distributions and correlations. The shown examples demonstrate that the correlated Dirichlet can model different or equal marginal Dirichlet distributions for $\boldsymbol{x^1}$ and $\boldsymbol{x^2}$ and correlations between 0 (Figure 3.3(a)) and 1 (Figure 3.4(b)). As Figure 3.4(a) shows, the correlation can also differ for different dimensions of the correlated Dirichlet distribution.

While no closed-form solution for the correlated Dirichlet distribution is available, sampling from it is straightforward so that it can be applied to the CFM. Figure 3.5 shows the CFM's general graphical model. The only difference to the IFM in Figure 3.2 is that classifier outputs $\boldsymbol{x_i^1}, \ldots, \boldsymbol{x_i^K}$ are jointly correlated-Dirichlet-distributed with parameters $\boldsymbol{\alpha_j^k}$ and $\boldsymbol{\delta_j}$ if $t_i = j$. As in the IFM, $\boldsymbol{\alpha_j^k}$ with $\alpha_{j\,l}^k > 0$ holds the parameter vector of the marginal Dirichlet distribution of classifier C_k if $t_i = j$. The new parameter $\boldsymbol{\delta_j}$ is responsible for the pairwise correlation between the classifier outputs if $t_i = j$. Its dimensionality is $1 \times J$ for $K = 2$ and $(\binom{K}{2} + 1) \times J$ for $K > 2$ classifiers. For the reduced case of $K = 2$ classifiers, Figure 3.6 additionally shows a more detailed graphical model of the CFM including the latent variables of the correlated Dirichlet distribution. For $K = 2$, it must hold that $\delta_{j\,l} > 0$ and $\delta_{j\,l} < \alpha_{j\,l}^k$ for $l = 1, \ldots, J, k = 1, \ldots, K$. Figure 3.7 shows the detailed graphical model of the CFM given in Figure 3.6 for $K > 2$ classifiers. $\boldsymbol{\alpha_j^k}$ holds the marginal parameters of classifier C_k's Dirichlet model if $t_i = j$. $\boldsymbol{\delta_j^{km}}$ holds the correlation parameters that determine the pairwise correlations between classifier C_k and all other classifiers C_m, $m = 1, \ldots, K, m \neq k$ if $t_i = j$. Therefore, it applies that ${\delta_j^{km}}_l = {\delta_j^{mk}}_l$ and equivalently ${D_j^{km}}_{il} = {D_j^{mk}}_{il}$. $\boldsymbol{\delta_j^a}$ holds the common correlation parameters between all classifiers $C_1, \ldots, C_K$ if $t_i = j$. Thus, note that for the special case of $K = 2$ classifiers $\boldsymbol{\delta_j}$ only consists of $\boldsymbol{\delta_j^a}$.

3.2.3.2 Parameter Inference

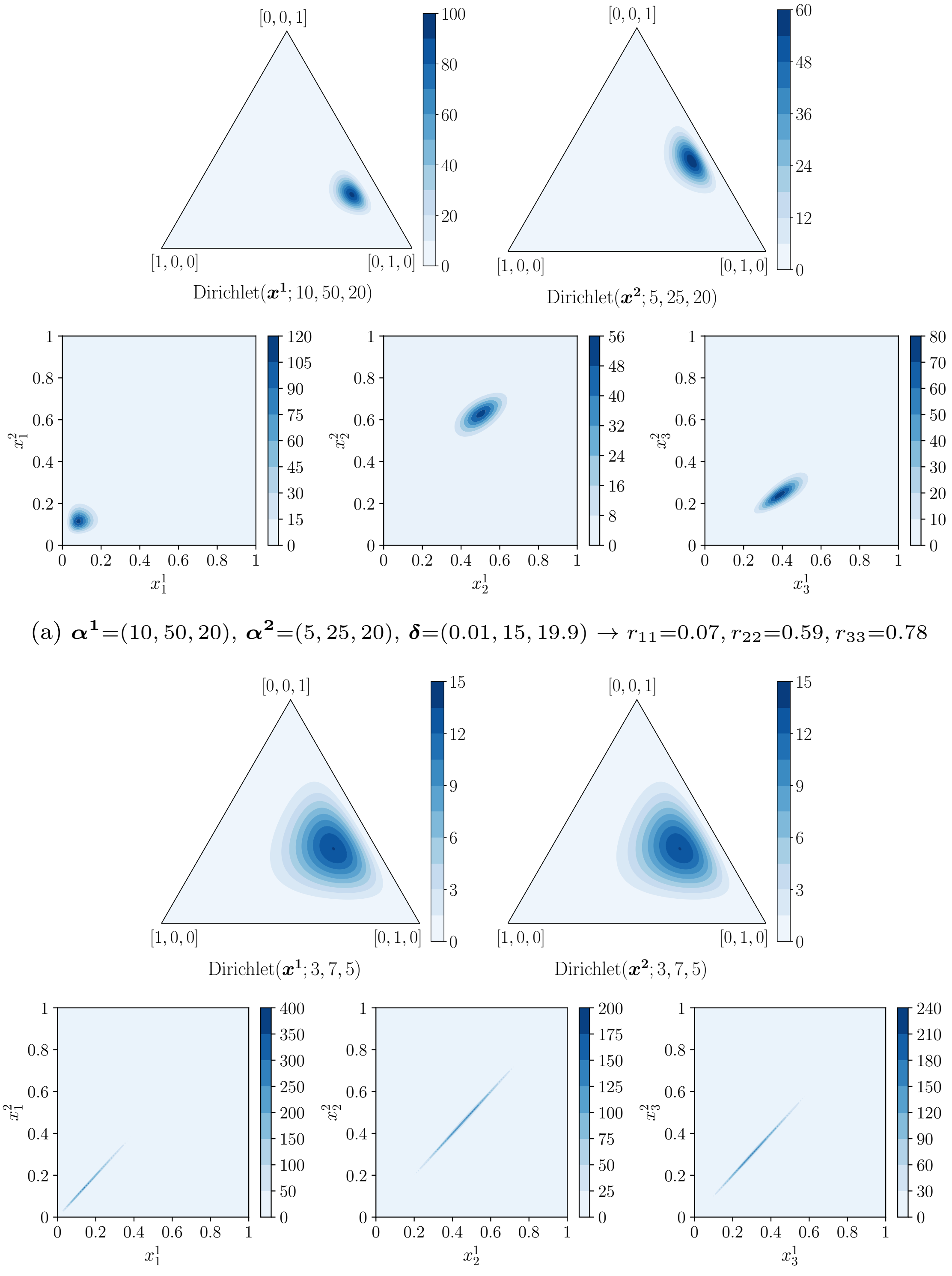
We learn the joint classifier model by inferring the posterior distribution over parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\delta}$ given observed classifier outputs $\boldsymbol{x}$ and their true labels $\boldsymbol{t}$, $p(\boldsymbol{\alpha}, \boldsymbol{\delta} | \boldsymbol{x}, \boldsymbol{t})$, using Gibbs sampling. For all elements of $\boldsymbol{\alpha_j^k}$ and $\boldsymbol{\delta}_j$, we chose a vague gamma prior with shape and rate set to 10^{-3}, which, however, can be set differently according to prior knowledge about the data. To increase robustness, inference can also be split up in two steps by first inferring the marginal Dirichlet parameters $\boldsymbol{\alpha}$ as described in Section 3.2.2.1 and subsequently inferring the posterior distribution over the correlation parameters given the inferred marginal parameters, $p(\boldsymbol{\delta} | \boldsymbol{x}, \boldsymbol{t}, \boldsymbol{\alpha})$. This step-wise inference gives the same results as full inference on data generated from the CFM, but was observed to be more robust empirically on real data since it guarantees correctly inferred marginal distributions.



(a) $\boldsymbol{\alpha}^{\mathbf{1}} = (2, 5, 2)$, $\boldsymbol{\alpha}^{\mathbf{2}} = (2, 7, 2)$, $\boldsymbol{\delta} = (0.01, 0.01, 0.01) \to r_{11} = 0.0, r_{22} = 0.0, r_{33} = 0.0$

(b) $\boldsymbol{\alpha}^{\mathbf{1}} = (2, 2, 3)$, $\boldsymbol{\alpha}^{\mathbf{2}} = (1, 1, 2)$, $\boldsymbol{\delta} = (0.4, 0.4, 0.9) \to r_{11} = 0.26, r_{22} = 0.24, r_{33} = 0.29$

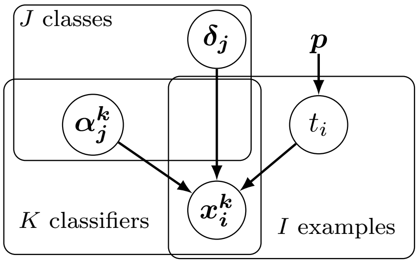
Figure 3.3: Marginal and joint densities of correlated Dirichlet distributions with selected parameter values leading to low correlations. The simplexes display the marginal Dirichlet distributions of $\boldsymbol{x}^{\mathbf{1}}$ and $\boldsymbol{x}^{\mathbf{2}}$, while the joint densities of x_l^1 and x_l^2, $l = 1, \ldots, 3$, are shown for each dimension of the correlated Dirichlet distribution. In (a) $\boldsymbol{x}^{\mathbf{1}}$ and $\boldsymbol{x}^{\mathbf{2}}$ are independent with dimension-wise correlations $r_{11} = r_{22} = r_{33} = 0.0$ between x_l^1 and x_l^2, $l = 1, \ldots, 3$, and different marginals. (b) shows different marginals with low correlations $r_{11} = 0.26, r_{22} = 0.24, r_{33} = 0.29$. The joint density plots were created with kernel density estimation based on 10^7 samples.



(a) $\boldsymbol{\alpha}^1$=(10, 50, 20), $\boldsymbol{\alpha}^2$=(5, 25, 20), $\boldsymbol{\delta}$=(0.01, 15, 19.9) → r_{11}=0.07, r_{22}=0.59, r_{33}=0.78

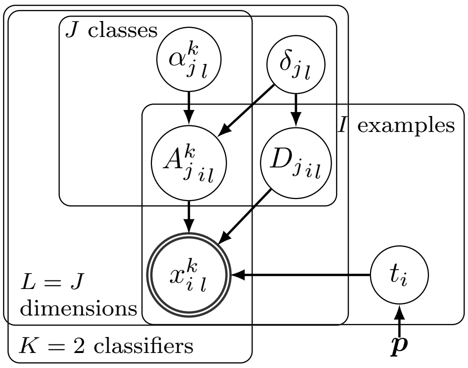
(b) $\boldsymbol{\alpha}^1 = (3, 7, 5)$, $\boldsymbol{\alpha}^2 = (3, 7, 5)$, $\boldsymbol{\delta} = (2.9, 6.9, 4.9) \rightarrow r_{11} = 0.97, r_{22} = 0.98, r_{33} = 0.98$

Figure 3.4: Marginal and joint densities of correlated Dirichlet distributions with selected parameter values leading to varying and high correlations. The simplexes display the marginal Dirichlet distributions of $\boldsymbol{x}^1$ and $\boldsymbol{x}^2$, while the joint densities of x_l^1 and x_l^2, $l = 1, \ldots, 3$, are shown for each dimension of the correlated Dirichlet distribution. (a) shows different marginals with different dimension-wise correlations $r_{11} = 0.07, r_{22} = 0.59, r_{33} = 0.78$ between x_l^1 and x_l^2, $l = 1, \ldots, 3$. (b) shows equal marginals with correlations close to 1, $r_{11} = 0.97, r_{22} = 0.98, r_{33} = 0.98$. The joint density plots were created with kernel density estimation based on 10^7 samples.



$$t_i \sim \text{Categorical}(\boldsymbol{p})$$
$$\boldsymbol{x_i^k} | t_i = j \sim \text{CorrDirichlet}(\boldsymbol{\alpha_j^k}, \boldsymbol{\delta_j})$$

Figure 3.5: Graphical model of the Correlated Fusion Model (CFM).



$$t_i \sim \text{Categorical}(\boldsymbol{p})$$
$$A^k_{j\,il} \sim \text{Gamma}(\alpha^k_{j\,l} - \delta_{j\,l}, 1)$$
$$D_{j\,il} \sim \text{Gamma}(\delta_{j\,l}, 1)$$
$$x^k_{i\,l} | t_i{=}j \leftarrow \frac{A^k_{j\,il} + D_{j\,il}}{\sum_{n=1}^{J} A^k_{j\,in} + D_{j\,in}}$$

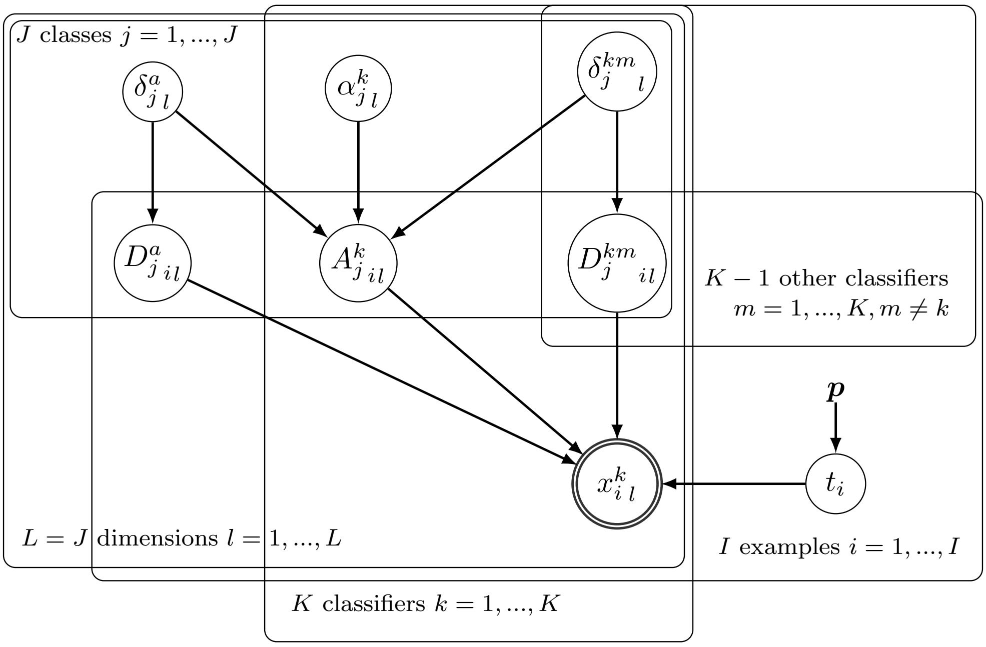
Figure 3.6: Detailed graphical model of the Correlated Fusion Model (CFM) for $K = 2$ classifiers including all latent variables of the correlated Dirichlet distribution.

As for the IFM, we implement Gibbs sampling using the standard inference tool JAGS (Plummer, 2003). Since $x^k_{i\,l}$ in the CFM is a deterministic variable, and inference tools such as JAGS do not allow deterministic variables to be observed, as commonly done, we inserted another random variable into the CFM. This additional variable $x^{k*}_{i\,l}$ is normally distributed with $x^{k*}_{i\,l} \sim \mathcal{N}(x^k_{i\,l}, \epsilon)$ and $\epsilon = 10^{-4}$. In the following, as for the IFM, we use the means of the posterior distributions inferred with Gibbs sampling as point estimates for $\boldsymbol{\alpha_j^k}$ and $\boldsymbol{\delta_j}$.

3.2.3.3 Normative Fusion Behavior

The fusion of K categorical base distributions $\boldsymbol{x_i^1}, \ldots, \boldsymbol{x_i^K}$ is performed by inferring the posterior distribution over the true label t_i conditioned on the base distributions $\boldsymbol{x_i^k}$ and the learned model parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\delta}$, $p(t_i | \boldsymbol{x_i^1}, \ldots, \boldsymbol{x_i^K}, \boldsymbol{\alpha}, \boldsymbol{\delta})$. Different from the IFM, here we cannot derive the fused distribution analytically because we do not have a closed-form solution for the probability density function of the correlated Dirichlet distribution. However, by assuming $\boldsymbol{\alpha}, \boldsymbol{\delta}$, and $\boldsymbol{x_i^1}, \ldots, \boldsymbol{x_i^K}$ to be observed, inference of latent t_i can be performed with Gibbs sampling using JAGS (Plummer, 2003). From a sufficient number of samples of t_i we can infer the categorical distribution over t_i, which is the fused result. Alternatively inferring t_i with variational methods in order to speed up fusion is left for future work.

Note that if we let all correlation parameters $\boldsymbol{\delta_j}$ tend to zero, the CFM reduces to the IFM, and its fusion behavior coincides with the one we derived analytically for the IFM



$$
\begin{aligned}
t_i &\sim \text{Categorical}(\boldsymbol{p}) \\
A^{k}_{j\,il} &\sim \text{Gamma}(\alpha^{k}_{j\,l} - \delta^{a}_{j\,l} - \sum_{\substack{m=1\\m\neq k}}^{K} \delta^{km}_{j\quad l}, 1) \\
D^{km}_{j\quad il} &\sim \text{Gamma}(\delta^{km}_{j\quad l}, 1) \\
D^{a}_{j\,il} &\sim \text{Gamma}(\delta^{a}_{j\,l}, 1) \\
x^{k}_{i\,l}|t_i = j &\leftarrow \frac{A^{k}_{j\,il} + D^{a}_{j\,il} + \sum_{\substack{m=1\\m\neq k}}^{K} D^{km}_{j\quad il}}{\sum_{n=1}^{J} A^{k}_{j\,in} + \sum_{n=1}^{J} D^{a}_{j\,in} + \sum_{\substack{m=1\\m\neq k}}^{K} D^{km}_{j\quad in}}
\end{aligned}
$$

Figure 3.7: Detailed graphical model of the proposed Correlated Fusion Model for $K > 2$ classifiers. $\boldsymbol{\alpha^{k}_{j}}$ holds the marginal parameters of classifier C_k's Dirichlet model if $t_i = j$. $\boldsymbol{\delta^{km}_{j}}$ holds the correlation parameters that determine the pairwise correlations between classifier C_k and all other classifiers C_m, $m = 1, \ldots, K, m \neq k$ if $t_i = j$. Thus, $\delta^{km}_{j\quad l} = \delta^{mk}_{j\quad l}$ and equivalently $D^{km}_{j\quad il} = D^{mk}_{j\quad il}$. $\boldsymbol{\delta^{a}_{j}}$ holds the common correlation parameters between all classifiers $C_1, \ldots, C_K$ if $t_i = j$.

in Section 3.2.2.2. Thus, variance, uncertainty, and bias of individual classifiers similarly influence the fusion when fusing with the CFM. Additionally, in contrast to previous fusion algorithms, our model can be used to investigate how uncertainty reduction through fusion should be affected by the correlation of the fused classifiers in a normative way. We examine this in detail with the two examples in the following.

Specifically, we compare the fusion behavior of the CFM for systematically varied correlations between two base classifiers. The blue bars in Figure 3.8 show an example where the marginal parameters of the correlated Dirichlet distributions are chosen to replicate IOP fusion behavior for zero correlation ($n = m = 1$ in (3.20)). The higher the correlation between the two classifiers, the smaller is the uncertainty reduction through fusion. In particular, there is no uncertainty reduction if the correlation is $r = 1$. In this case, the fused distribution equals the two base distributions.

The orange bars in Figure 3.8 show the fusion results given different correlation levels for marginal parameters that imply increased uncertainty reduction compared to IOP ($n = m = 2$ in (3.20)) for zero correlation because of lower classifier variance and uncertainty. As can be seen, there is also less uncertainty reduction, the higher the correlation between both classifiers. However, for $r = 1$, the fused distribution is not identical to the two base distributions; its uncertainty is reduced despite the high correlation. Yet, the reason for this is not fusion but the Dirichlet models we learned for each individual classifier. The resulting fused distribution for $r = 1$ is similar to the resulting distributions we get if we use the IFM as a meta classifier individually for each base distribution (see end of Section 3.2.2.2). Hence, the fusion of two highly correlated classifiers does not additionally reduce the uncertainty. This also applies to the first example. However, in this case, due to the chosen marginal distributions, the meta classifier results are equal to the base distributions.

Both examples reveal that the uncertainty reduction through fusion should decrease progressively if the base classifiers' correlation increases. For a correlation of $r = 1$, fusion should not reduce the uncertainty at all. Still, the fused distribution might be less uncertain than the base distributions since uncertainty cannot only be reduced by fusion but also as a result of modeling each individual classifier's behavior, i.e., bias, variance, and uncertainty.

3.3 Evaluation

We evaluate our model on simulated and real data sets. The fused distributions returned by the CFM are compared to those of the IFM and IOP and the base distributions. In addition, we compare the fusion performances to the performances of each classifier's meta classifier and the related method proposed by Pirs and Strumbelj (2019). As performance measures, we consider entropy for quantifying uncertainty reduction through fusion and log-loss for quantifying correctness of classifications. The log-loss, which is a standard measure for the performance of probabilistic classifiers (Vovk, 2015), penalizes wrong classifications according to their uncertainty, thus considering both correctness and uncertainty of a classifier.

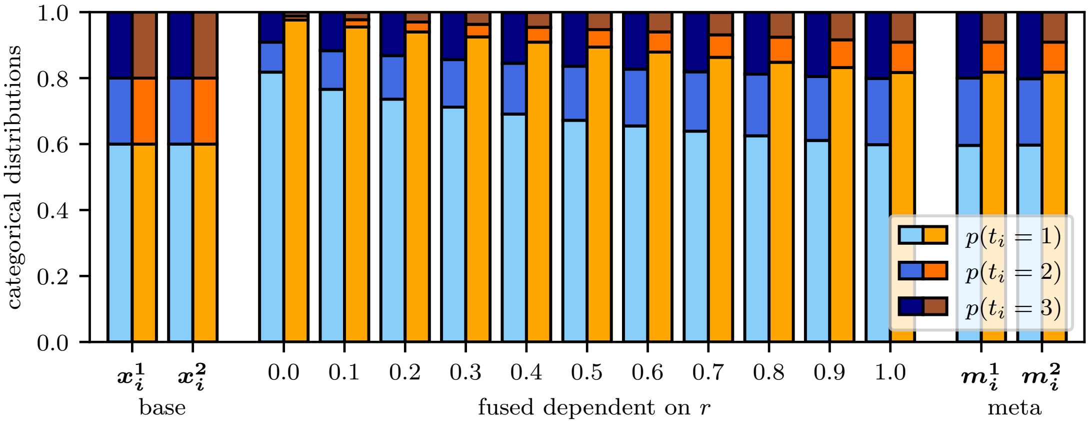


Figure 3.8: Fusion of the base distributions $\boldsymbol{x_i^1}$=$\boldsymbol{x_i^2}$=$(0.6, 0.2, 0.2)$ using the CFM assuming different marginal parameters and correlations. Each bar represents a categorical distribution consisting of the probabilities for $p(t_i = 1)$, $p(t_i = 2)$, $p(t_i = 3)$. For the blue bars we assume IOP marginal parameters, for the orange bars we assume marginals that imply stronger uncertainty reduction. We progressively increase the assumed correlation between classifiers from 0.0 to 1.0 and show the corresponding fused distributions as well as the results of the meta classifiers $\boldsymbol{m_i^1}$ and $\boldsymbol{m_i^2}$.

3.3.1 Simulated Data Sets

We created different simulated data sets by generating random samples of output distributions $\boldsymbol{x_i^1}$ and $\boldsymbol{x_i^2}$ of $K = 2$ classifiers for different given marginal parameters $\boldsymbol{\alpha}$, correlation parameters $\boldsymbol{\delta}$, and true class labels t_i with $J = 3$ possible outcomes according to the generative model of the CFM (Figure 3.6). To show the normative fusion behavior depending on the base classifiers' correlation, for three sets of marginal parameters $\boldsymbol{\alpha}$, we chose different correlation parameters $\boldsymbol{\delta}$ respectively that correspond to the correlations 0.0, 0.25, 0.5, 0.75, 1.0 between the two classifiers' outputs. For all five correlation levels, we generated 25 simulated random test sets on which we evaluate, each consisting of 60 test examples (20 per class) composed of two categorical distributions and their corresponding class label. Since the true parameters of the data were known, no training data were required. We chose the marginal parameters to represent three prototype cases of classifier models in order to demonstrate that the effect of correlation on the fusion behavior also depends on the individual classifiers' marginal Dirichlet models. One of the chosen classifier models leads to IOP fusion for zero correlation (SIM 1), one represents two classifiers with decreased variance (SIM 2), and one represents two biased classifiers (SIM 3).

For the first simulated data set SIM 1, we determine the marginal parameters $\boldsymbol{\alpha}$ of the CFM such that it reduces to IOP if $r = 0$. As shown in Figure 3.9(a), therefore, the results of IOP and the IFM are equal regarding entropy and log-loss. The shown entropies reveal that the higher the correlation between the classifiers is, the more uncertainty is reduced by fusing with IOP or the IFM. In contrast, when fusing with the CFM, we see less uncertainty reduction through fusion for higher correlations. Particularly, for $r = 1$, there is no uncertainty reduction. The mean entropy is the same as for the two meta classifiers. Also, the CFM's mean log-loss is equal to the meta classifiers' log-loss if $r = 1$. Thus, as expected, we see no change in performance through fusion for highly correlated classifiers when using the CFM. Since we chose the marginals according to IOP fusion, the CFM's performance also equals the performances of the base classifiers. In general, the CFM performs best at all correlation levels. Particularly for high correlations, it outperforms the other fusion methods, which assume independence, overestimate uncertainty reduction, and therefore perform even worse than the base classifiers.

The second simulated data set SIM 2 was generated setting the CFM's marginal parameters $\boldsymbol{\alpha}$ according to the example in (3.20) with $n = m = 4$, which leads to increased uncertainty reduction through fusion in comparison to IOP for independent classifiers, since the modeled base classifiers' variance is decreased. Accordingly, Figure 3.9(b) shows significantly lower mean entropies for the IFM than for IOP for all correlation levels. In contrast, for the CFM, the fused distributions' mean entropy increases with the correlation such as for SIM 1. If $r = 1$, the CFM again shows the same entropy as the two meta classifiers. Hence, the fusion of two highly correlated base classifiers does not reduce the uncertainty. This is confirmed by the log-loss (Figure 3.9(b)). However, in contrast to SIM 1, here, the meta classifiers' performances are increased compared to the base classifiers, and uncertainty is reduced. Therefore, the CFM outperforms the base classifiers also for a correlation of $r = 1$. Note that, again, the CFM achieves the lowest log-loss and thus the best performance for all correlation levels.

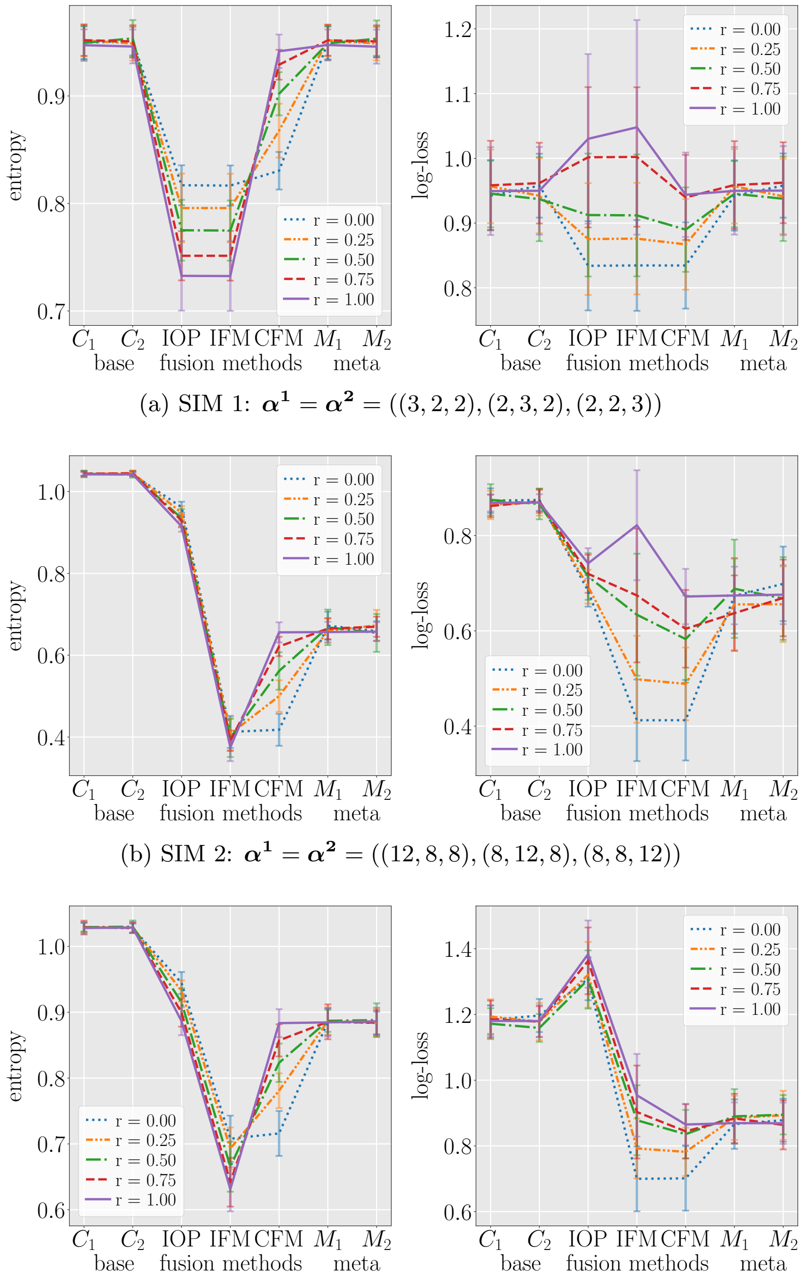
For the third simulated data set SIM 3 we generated classifier outputs of two biased classifiers, which on average predict class 3 if $t_i = 2$ and vice versa. In Figure 3.9(c) we see similar fusion results for SIM 3 as for the other simulated data sets SIM 1 and SIM 2 in Figure 3.9(a) and (b): less uncertainty reduction for higher correlations, no fusion gain for $r = 1$, and best performance of the CFM compared to other fusion methods. In addition, for the biased data set SIM 3, we observe a performance decline of IOP compared to the base classifiers according to log-loss, since IOP reinforces the mainly wrong classifications. In contrast, the IFM and CFM have learned the bias and thus compensate for it. This demonstrates the superiority of learning classifier models over ad-hoc methods.

3.3.2 Real Data Sets

In addition to simulated data sets, we also evaluated the CFM on 6 real data sets, Bookies A, Bookies B, DNA A, DNA B, DNA C, and Bookies C. Bookies A and Bookies B are each constructed from the odds of two bookmakers for football matches. The target variable has three possible outcomes (home, draw, away), and for each match, the odds were transformed to a 3-dimensional categorical probability distribution by normalizing their reciprocals. Thus, each bookie is considered as a base classifier and each example in the data sets is composed of two categorical distributions and a true class label. Bookmakers' predictions were also used for evaluations in the related work by Pirs and Strumbelj (2019).

Bookies A contains predictions of two bookmakers (B365 and BW) for football matches of the English Premier League[1] from 14 seasons from 2005 to 2019. Excluding matches with missing odds, the data set comprises 5317 examples in total. The correlation between the bookmakers' predictions is approximately 1; it ranges from 0.955 to 0.993 in different dimensions and for different values of t_i.

[1] https://www.football-data.co.uk/englandm.php



(a) SIM 1: $\boldsymbol{\alpha^1} = \boldsymbol{\alpha^2} = ((3,2,2),(2,3,2),(2,2,3))$

(b) SIM 2: $\boldsymbol{\alpha^1} = \boldsymbol{\alpha^2} = ((12,8,8),(8,12,8),(8,8,12))$

(c) SIM 3: $\boldsymbol{\alpha^1} = \boldsymbol{\alpha^2} = ((7,5,5),(5,5,7),(5,7,5))$

Figure 3.9: Fusion performances on simulated data in terms of mean entropy and log-loss. We compare the performances of base classifiers C_1, C_2, the three fusion methods IOP, IFM, and CFM, and the meta classifiers M_1, M_2. We show the fusion behavior for five levels of correlation between the base classifiers and different marginal model parameters, implying IOP fusion (a), higher reinforcement due to decreased classifier variance and uncertainty (b), and the fusion of two biased classifiers (c). Standard deviations are shown as error bars. Note that we connect the means for better readability, although there are no evaluated points in between them.

Bookies B consists of the predictions of two bookmakers (B365 and BW) for matches of the German Bundesliga[2] from 14 seasons from 2005 to 2019. Similar to the Bookies A data set, we excluded matches with missing odds, totaling to 4278 matches. The correlation between the bookmakers' predictions is approximately 1; it ranges from 0.955 to 0.996 in different dimensions and for different values of t_i.

The DNA data set from the StatLog project[3], which was also chosen for evaluations in the related work by Pirs and Strumbelj (2019) and Kim and Ghahramani (2012), was used to construct three more data sets for evaluating the CFM. The original DNA data set contains DNA sequences in which splice junctions are detected. It consists of 3188 examples with 60 attributes and a target variable with $J = 3$ possible outcomes. For each data set DNA A, DNA B, DNA C, we trained $K = 2$ different classifiers on this data set. Their categorical output distributions on the corresponding test data set form the respective data set DNA A, DNA B, DNA C.

For DNA A, we trained two highly correlated classifiers by using the same classification method (kNN) and the same training data but different hyperparameters ($k = 120$ and $k = 150$). For training we used 10-fold cross-validation. The output distributions in the 10 test splits form the DNA A data set, totaling to 3188 examples. The correlation between both base classifiers is approximately 1; it ranges from 0.962 to 0.986 for different dimensions and values for t_i.

For DNA B, we trained two classifiers by using the same classification method (kNN, $k = 50$) but different training data. Each classifier was trained on 5% of the DNA data set, their classifications on the remaining 90% of the data (2869 examples) formed the DNA B data set. The correlation between both base classifiers ranges from 0.463 to 0.709 for different dimensions and values for t_i.

DNA C was created by training two different classifiers, one kNN classifier ($k = 50$) and one Random Forest classifier, on the same training set composed of 5% of the DNA data set. The classifiers' output distributions on the remaining 95% of the data (3030 examples) construct the DNA C data set. The correlation between the base classifiers' output distributions ranges from 0.5 to 0.693 in different dimensions and for different values of t_i.

While all of the above data sets consist of the output distributions of only $K = 2$ classifiers, we additionally evaluate the CFM on a data set consisting of $K = 3$ classifiers, Bookies C. Bookies C is equivalent to Bookies A but additionally includes a third bookmaker's (IW) predictions. Thus, it contains the predictions of three bookmakers (B365, BW, IW) for football matches of the English Premier League[4] from 14 seasons from 2005 to 2019. As Bookies A, excluding matches with missing odds, the data set comprises 5317 examples in total. Also, the correlation between all three bookmakers' predictions is approximately 1.

We randomly split each real data set into test and training set, while the test set contains 60 examples (20 per class) and the training set all remaining ones. On each random training split the model parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\delta}$ were inferred, which were then used to fuse the distributions in the test set. The random splitting was repeated five times with different

[2] https://www.football-data.co.uk/germanym.php

[3] https://archive.ics.uci.edu/ml/datasets/Molecular+Biology+(Splice-junction+Gene+Sequences)

[4] https://www.football-data.co.uk/englandm.php

random seeds, and means and standard deviations of the resulting performance measures were computed. For all real data sets, the performances of base classifiers, different fusion methods, and meta classifiers in terms of entropy and log-loss are shown in Figure 3.10. Figure 3.10(a) shows the results for all data sets with $K = 2$ base classifiers, while Figure 3.10(b) shows the results for the Bookies C data set with $K = 3$ base classifiers.

For the three highly correlated data sets with $K = 2$ base classifiers, Bookies A, Bookies B, and DNA A, the CFM's performance is equal to the performances of the meta classifiers, both regarding entropy and log-loss. Thus, also on real data we confirm that fusion causes no uncertainty reduction and no change in performance if the base classifiers are highly correlated. However, this does not necessarily result in equal performances of the CFM and the base classifiers. Depending on the Dirichlet models learned for the individual classifiers, the CFM can still outperform highly correlated base classifiers, which we see for DNA A. Also, the CFM can perform worse than the base classifiers, e.g., for Bookies B, which is an effect of too similar Dirichlet models for different class labels t_i.

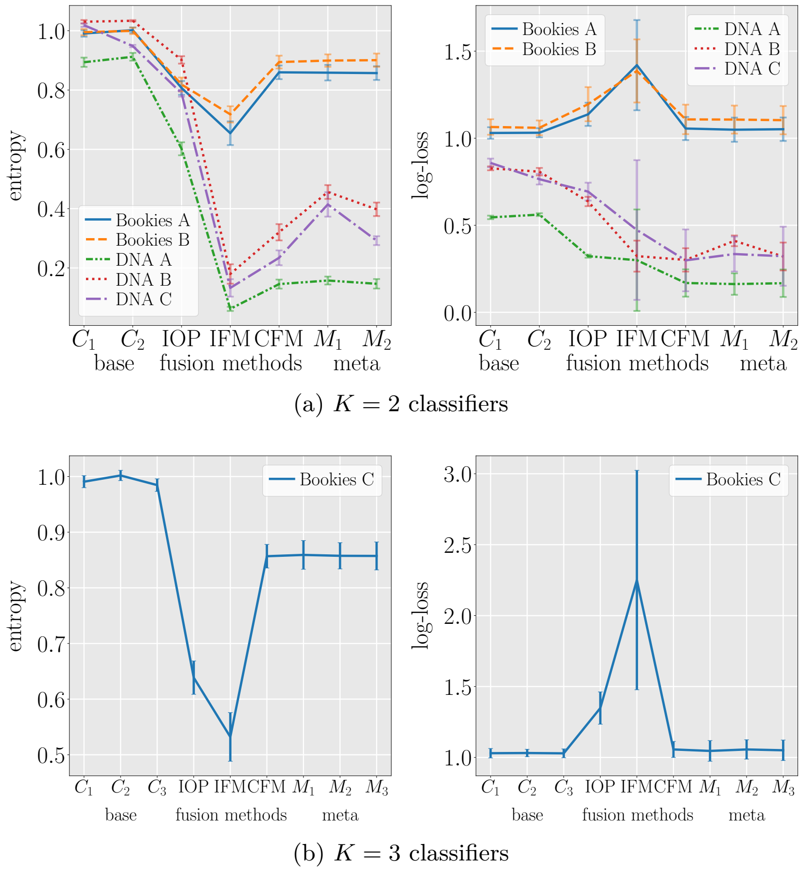
For the less correlated data sets with $K = 2$ base classifiers, DNA B and C, we see that the CFM reduces less uncertainty than the IFM but is more certain than the meta classifiers. Also, the CFM performs best of all fusion methods and better than base and meta classifiers.

For the Bookies C data set, which consists of $K = 3$ highly correlated base classifiers, Figure 3.10(b) shows the same behavior as for the data sets Bookies A and Bookies B in Figure 3.10(a). The CFM's performance is equal to the performance of all three meta classifiers. Thus, also when fusing three highly correlated classifiers fusion with the CFM causes no uncertainty reduction and no change in performance. Moreover, also for three classifiers, the IFM and IOP perform worse than the CFM since they assume independence and overestimate uncertainty reduction.

3.3.3 Comparison to the Model by Pirs and Strumbelj (2019)

The model introduced by Pirs and Strumbelj (2019), which relies on modeling transformed classifier outputs with a multivariate normal mixture, is the only comparable Bayesian method for fusing correlated probabilistic classifiers. Contrary to Pirs and Strumbelj (2019), on simulated and real data we show that although fusion should not reduce the uncertainty if $r = 1$, in a normative framework fused classifiers can outperform highly correlated base classifiers due to the models learned for the individual classifiers. Moreover, we additionally compared the performances of the CFM and Pirs' model in terms of log-loss. As can be seen in Table 3.1, the CFM outperforms on all tested simulated and real data sets.

A limitation of the proposed algorithm for inference in the CFM is slow fusion as a result of Gibbs sampling. Therefore, in addition to their performance we also compared the CFM and the model by Pirs and Strumbelj (2019) in terms of required time for fusion. Fusing all 60 base distributions in the first random test split of data set DNA B requires 940.92 seconds when using the CFM with 120 parallel chains with 175.000 burn-in samples and 175.000 samples each. Fusing the same test split with the model by Pirs and Strumbelj (2019) takes only 3.53 seconds. However, note that we intentionally decided to use a large number of samples to guarantee correctness of the fusion results, whereas time efficiency



(a) $K = 2$ classifiers

(b) $K = 3$ classifiers

Figure 3.10: Fusion performances on real data in terms of mean entropy and log-loss. For real data sets with $K = 2$ (a) or $K = 3$ (b) base classifiers we compare the performances of base classifiers $C_1, \ldots, C_K$, the fusion methods IOP, IFM, and CFM, and the meta classifiers $M_1, \ldots, M_K$. Standard deviations are shown as error bars. Note that we connect the means for better readability, although there are no evaluated points in between them.

Table 3.1: Comparison of the performances of the CFM and the model proposed by Pirs and Strumbelj (2019). We compared performances in terms of log-loss on the simulated data sets SIM 1, SIM 2, SIM 3 with different correlation levels r and the six real data sets Bookies A, Bookies B, DNA A, DNA B, DNA C, and Bookies C. The table shows means μ and standard deviations σ of the models' achieved log-losses.

data set	CFM ($\mu \pm \sigma$)	Pirs & Strumbelj's model ($\mu \pm \sigma$)
SIM 1 $_{r=0.0}$	0.834 ± 0.067	0.915 ± 0.03
SIM 1 $_{r=0.25}$	0.867 ± 0.07	0.925 ± 0.041
SIM 1 $_{r=0.5}$	0.89 ± 0.065	0.938 ± 0.039
SIM 1 $_{r=0.75}$	0.94 ± 0.066	0.955 ± 0.043
SIM 1 $_{r=1.0}$	0.944 ± 0.065	0.96 ± 0.056
SIM 2 $_{r=0.0}$	0.412 ± 0.085	0.582 ± 0.048
SIM 2 $_{r=0.25}$	0.489 ± 0.076	0.607 ± 0.051
SIM 2 $_{r=0.5}$	0.583 ± 0.092	0.66 ± 0.065
SIM 2 $_{r=0.75}$	0.604 ± 0.082	0.687 ± 0.048
SIM 2 $_{r=1.0}$	0.672 ± 0.058	0.717 ± 0.041
SIM 3 $_{r=0.0}$	0.701 ± 0.098	0.836 ± 0.047
SIM 3 $_{r=0.25}$	0.782 ± 0.073	0.869 ± 0.039
SIM 3 $_{r=0.5}$	0.836 ± 0.074	0.887 ± 0.04
SIM 3 $_{r=0.75}$	0.844 ± 0.082	0.901 ± 0.057
SIM 3 $_{r=1.0}$	0.865 ± 0.063	0.893 ± 0.043
Bookies A	1.056 ± 0.067	1.165 ± 0.035
Bookies B	1.108 ± 0.085	1.176 ± 0.052
DNA A	0.169 ± 0.078	0.177 ± 0.021
DNA B	0.301 ± 0.067	0.421 ± 0.043
DNA C	0.298 ± 0.178	0.351 ± 0.092
Bookies C	1.056 ± 0.056	1.297 ± 0.046

is not in the scope of this work but left for future investigations. We conclude that the CFM should be chosen if correct fusion is the goal. If instead fast fusion is the goal the method by Pirs and Strumbelj (2019) can be selected with the risk of incorrect fusion and performance losses.

3.4 Discussion and Conclusion

In this work, we derived Bayes optimal fusion behavior for probabilistic classifiers, which explicitly considers the classifiers' uncertainty, bias, variance, and correlation. The resulting Correlated Fusion Model (CFM) is derived assuming successively more general assumptions. In particular, it subsumes known independent fusion models as special cases, which are each optimal given their specific assumptions.

Independent Opinion Pool (IOP) is the Bayes optimal fusion model if the base classifiers' observations and with them their output distributions are conditionally independent and we assume an uninformed prior. It considers the uncertainty of individual base classifiers for fusion and reduces uncertainty in the correct way. Moreover, since it is an ad-hoc method, IOP does not require any further information about the base classifiers expect from their current output distribution at hand. In particular, no training data are required to apply IOP, which makes it suitable for applications with sparse data.

The Independent Fusion Model (IFM) is the Bayes optimal fusion model if we assume independent base classifiers and an uninformed prior over classes, but, in contrast to IOP, have some knowledge about the base classifiers' properties, i.e., their bias, variance, and uncertainty, from observed training data. The IFM models these properties with Dirichlet distributions conditioned on the true class label. Thus, for fusion, it can not only consider the uncertainty of the base classifiers for the current example at hand, but also their general uncertainty, bias, and variance.

Compared to IOP and the IFM, the Correlated Fusion Model (CFM) is the most general fusion model. While also assuming an uninformed prior, the CFM does not assume independent base classifiers and can explicitly model their correlation using a new correlated Dirichlet distribution. Thus, it can not only learn the classifiers' individual properties from training data, but also the correlations between different classifiers. By this, the CFM normatively specifies how to fuse classifiers considering their bias, variance, uncertainty, and correlation. In particular, with the CFM we showed that uncertainty reduction through fusion should be the lower, the higher the correlation between the classifiers is, resulting in no uncertainty reduction through fusion if $r = 1$. However, this does not necessarily lead to equal performances of the fused classifier and the base classifiers if a model for each classifier is learned.

A limitation of the proposed Correlated Fusion Model is that the improvements in handling uncertainties come at the price of a high number of required parameters. Additionally, the inference algorithm proposed in this work, which uses Gibbs sampling, is computationally expensive and therefore slower compared to alternative previous models and their inference algorithms. For future work, we thus plan to investigate alternatives to inference via Gibbs sampling to speed up the inference for fusion.

Still, the proposed normative fusion model offers a new perspective on Bayesian combination of probabilistic classifiers, thereby clarifying how the correlation between classifiers affects uncertainty reduction through fusion and subsuming well known pioneering expert opinion aggregation techniques. Since it additionally outperforms the only comparable model on all tested data sets, it should be the method of choice if correct Bayes optimal fusion is the goal. However, as classification could potentially be used in conjunction with data and tasks with negative societal impact, we encourage responsible deployment of the proposed approach.

3.5 Appendix

3.5.1 Model Parameters Used for Evaluation

We evaluated the Correlated Fusion Model on simulated as well as on real data sets. In the following, we present the model parameters that we chose for generating the simulated data sets (Section 3.5.1.1) and that were inferred for the real data sets (Section 3.5.1.2).

3.5.1.1 Parameters for the Simulated Data Sets

The parameters we used for generating the simulated data sets used for evaluation in Section 4.1 are presented in Table 3.2 for the first simulated data set (SIM 1), Table 3.3 for the second simulated data set (SIM 2), and Table 3.4 for the third simulated data set (SIM 3). Note that the shown correlations can only be generated approximately with the presented parameters.

3.5.1.2 Parameters for the Real Data Sets

The parameters of the Correlated Fusion Model that we inferred for the five real data sets with $K = 2$ base classifiers Bookies A, Bookies B, DNA A, DNA B, and DNA C are presented in Table 3.5.

For the three data sets Bookies A, Bookies B, and DNA A, the correlation parameters $\boldsymbol{\delta}$ are very close to the marginal parameters $\boldsymbol{\alpha^1}$ and $\boldsymbol{\alpha^2}$, modeling a correlation close to $r = 1$ between the two classifiers.

In contrast, for the data sets DNA B and DNA C, we see that the correlation parameters $\boldsymbol{\delta}$ differ more from the marginal parameters $\boldsymbol{\alpha^1}$ and $\boldsymbol{\alpha^2}$. This reflects the lower correlation between the corresponding base classifiers in these data sets.

Table 3.6 shows the parameters of the real data set Bookies C, which consists of the predictions of $K = 3$ bookmakers. Since all three are highly correlated, the common correlation parameters $\boldsymbol{\delta^a}$ are close to the marginal parameters in $\boldsymbol{\alpha^1}$, $\boldsymbol{\alpha^2}$, $\boldsymbol{\alpha^3}$, while the pairwise correlation parameters are close to 0.

Table 3.2: Model parameters of the Correlated Fusion Model that we used to generate the first simulated data set (SIM 1) for five correlation levels from $r \approx 0$ to $r \approx 1$. $\boldsymbol{\alpha^1}$ holds the marginal Dirichlet parameters of classifier C_1, $\boldsymbol{\alpha^2}$ the ones of C_2, and $\boldsymbol{\delta}$ the correlation parameters of the correlated Dirichlet distribution. The j-th row of each parameter matrix holds the parameters modeling the classifier outputs of examples with true label $t_i = j$.

correlation	$\boldsymbol{\alpha^1}$	$\boldsymbol{\alpha^2}$	$\boldsymbol{\delta}$
$r \approx 0.0$	$\begin{bmatrix} 3 & 2 & 2 \\ 2 & 3 & 2 \\ 2 & 2 & 3 \end{bmatrix}$	$\begin{bmatrix} 3 & 2 & 2 \\ 2 & 3 & 2 \\ 2 & 2 & 3 \end{bmatrix}$	$\begin{bmatrix} 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 \end{bmatrix}$
$r \approx 0.25$	$\begin{bmatrix} 3 & 2 & 2 \\ 2 & 3 & 2 \\ 2 & 2 & 3 \end{bmatrix}$	$\begin{bmatrix} 3 & 2 & 2 \\ 2 & 3 & 2 \\ 2 & 2 & 3 \end{bmatrix}$	$\begin{bmatrix} 0.75 & 0.5 & 0.5 \\ 0.5 & 0.75 & 0.5 \\ 0.5 & 0.5 & 0.75 \end{bmatrix}$
$r \approx 0.5$	$\begin{bmatrix} 3 & 2 & 2 \\ 2 & 3 & 2 \\ 2 & 2 & 3 \end{bmatrix}$	$\begin{bmatrix} 3 & 2 & 2 \\ 2 & 3 & 2 \\ 2 & 2 & 3 \end{bmatrix}$	$\begin{bmatrix} 1.5 & 1 & 1 \\ 1 & 1.5 & 1 \\ 1 & 1 & 1.5 \end{bmatrix}$
$r \approx 0.75$	$\begin{bmatrix} 3 & 2 & 2 \\ 2 & 3 & 2 \\ 2 & 2 & 3 \end{bmatrix}$	$\begin{bmatrix} 3 & 2 & 2 \\ 2 & 3 & 2 \\ 2 & 2 & 3 \end{bmatrix}$	$\begin{bmatrix} 2.25 & 1.5 & 1.5 \\ 1.5 & 2.25 & 1.5 \\ 1.5 & 1.5 & 2.25 \end{bmatrix}$
$r \approx 1.0$	$\begin{bmatrix} 3 & 2 & 2 \\ 2 & 3 & 2 \\ 2 & 2 & 3 \end{bmatrix}$	$\begin{bmatrix} 3 & 2 & 2 \\ 2 & 3 & 2 \\ 2 & 2 & 3 \end{bmatrix}$	$\begin{bmatrix} 2.9 & 1.9 & 1.9 \\ 1.9 & 2.9 & 1.9 \\ 1.9 & 1.9 & 2.9 \end{bmatrix}$

Table 3.3: Model parameters of the Correlated Fusion Model that we used to generate the second simulated data set (SIM 2) for five correlation levels from $r \approx 0$ to $r \approx 1$. $\boldsymbol{\alpha}^{\mathbf{1}}$ holds the marginal Dirichlet parameters of classifier C_1, $\boldsymbol{\alpha}^{\mathbf{2}}$ the ones of C_2, and $\boldsymbol{\delta}$ the correlation parameters of the correlated Dirichlet distribution. The j-th row of each parameter matrix holds the parameters modeling the classifier outputs of examples with true label $t_i = j$.

correlation	$\boldsymbol{\alpha}^{\mathbf{1}}$	$\boldsymbol{\alpha}^{\mathbf{2}}$	$\boldsymbol{\delta}$
$r \approx 0.0$	$\begin{bmatrix} 12 & 8 & 8 \\ 8 & 12 & 8 \\ 8 & 8 & 12 \end{bmatrix}$	$\begin{bmatrix} 12 & 8 & 8 \\ 8 & 12 & 8 \\ 8 & 8 & 12 \end{bmatrix}$	$\begin{bmatrix} 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 \end{bmatrix}$
$r \approx 0.25$	$\begin{bmatrix} 12 & 8 & 8 \\ 8 & 12 & 8 \\ 8 & 8 & 12 \end{bmatrix}$	$\begin{bmatrix} 12 & 8 & 8 \\ 8 & 12 & 8 \\ 8 & 8 & 12 \end{bmatrix}$	$\begin{bmatrix} 3 & 2 & 2 \\ 2 & 3 & 2 \\ 2 & 2 & 3 \end{bmatrix}$
$r \approx 0.5$	$\begin{bmatrix} 12 & 8 & 8 \\ 8 & 12 & 8 \\ 8 & 8 & 12 \end{bmatrix}$	$\begin{bmatrix} 12 & 8 & 8 \\ 8 & 12 & 8 \\ 8 & 8 & 12 \end{bmatrix}$	$\begin{bmatrix} 6 & 4 & 4 \\ 4 & 6 & 4 \\ 4 & 4 & 6 \end{bmatrix}$
$r \approx 0.75$	$\begin{bmatrix} 12 & 8 & 8 \\ 8 & 12 & 8 \\ 8 & 8 & 12 \end{bmatrix}$	$\begin{bmatrix} 12 & 8 & 8 \\ 8 & 12 & 8 \\ 8 & 8 & 12 \end{bmatrix}$	$\begin{bmatrix} 9 & 6 & 6 \\ 6 & 9 & 6 \\ 6 & 6 & 9 \end{bmatrix}$
$r \approx 1.0$	$\begin{bmatrix} 12 & 8 & 8 \\ 8 & 12 & 8 \\ 8 & 8 & 12 \end{bmatrix}$	$\begin{bmatrix} 12 & 8 & 8 \\ 8 & 12 & 8 \\ 8 & 8 & 12 \end{bmatrix}$	$\begin{bmatrix} 11.9 & 7.9 & 7.9 \\ 7.9 & 11.9 & 7.9 \\ 7.9 & 7.9 & 11.9 \end{bmatrix}$

Table 3.4: Model parameters of the Correlated Fusion Model that we used to generate the third simulated data set (SIM 3) for five correlation levels from $r \approx 0$ to $r \approx 1$. $\boldsymbol{\alpha^1}$ holds the marginal Dirichlet parameters of classifier C_1, $\boldsymbol{\alpha^2}$ the ones of C_2, and $\boldsymbol{\delta}$ the correlation parameters of the correlated Dirichlet distribution. The j-th row of each parameter matrix holds the parameters modeling the classifier outputs of examples with true label $t_i = j$.

correlation	$\boldsymbol{\alpha^1}$	$\boldsymbol{\alpha^2}$	$\boldsymbol{\delta}$
$r \approx 0.0$	$\begin{bmatrix} 7 & 5 & 5 \\ 5 & 5 & 7 \\ 5 & 7 & 5 \end{bmatrix}$	$\begin{bmatrix} 7 & 5 & 5 \\ 5 & 5 & 7 \\ 5 & 7 & 5 \end{bmatrix}$	$\begin{bmatrix} 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 \end{bmatrix}$
$r \approx 0.25$	$\begin{bmatrix} 7 & 5 & 5 \\ 5 & 5 & 7 \\ 5 & 7 & 5 \end{bmatrix}$	$\begin{bmatrix} 7 & 5 & 5 \\ 5 & 5 & 7 \\ 5 & 7 & 5 \end{bmatrix}$	$\begin{bmatrix} 1.75 & 1.25 & 1.25 \\ 1.25 & 1.25 & 1.75 \\ 1.25 & 1.75 & 1.25 \end{bmatrix}$
$r \approx 0.5$	$\begin{bmatrix} 7 & 5 & 5 \\ 5 & 5 & 7 \\ 5 & 7 & 5 \end{bmatrix}$	$\begin{bmatrix} 7 & 5 & 5 \\ 5 & 5 & 7 \\ 5 & 7 & 5 \end{bmatrix}$	$\begin{bmatrix} 3.5 & 2.5 & 2.5 \\ 2.5 & 2.5 & 3.5 \\ 2.5 & 3.5 & 2.5 \end{bmatrix}$
$r \approx 0.75$	$\begin{bmatrix} 7 & 5 & 5 \\ 5 & 5 & 7 \\ 5 & 7 & 5 \end{bmatrix}$	$\begin{bmatrix} 7 & 5 & 5 \\ 5 & 5 & 7 \\ 5 & 7 & 5 \end{bmatrix}$	$\begin{bmatrix} 5.25 & 3.75 & 3.75 \\ 3.75 & 3.75 & 5.25 \\ 3.75 & 5.25 & 3.75 \end{bmatrix}$
$r \approx 1.0$	$\begin{bmatrix} 7 & 5 & 5 \\ 5 & 5 & 7 \\ 5 & 7 & 5 \end{bmatrix}$	$\begin{bmatrix} 7 & 5 & 5 \\ 5 & 5 & 7 \\ 5 & 7 & 5 \end{bmatrix}$	$\begin{bmatrix} 6.9 & 4.9 & 4.9 \\ 4.9 & 4.9 & 6.9 \\ 4.9 & 6.9 & 4.9 \end{bmatrix}$

Table 3.5: Model parameters of the Correlated Fusion Model that we inferred for the real data sets with $K = 2$ base classifiers. $\boldsymbol{\alpha}^{\mathbf{1}}$ holds the marginal Dirichlet parameters of classifier C_1, $\boldsymbol{\alpha}^{\mathbf{2}}$ the ones of C_2, and $\boldsymbol{\delta}$ the correlation parameters of the correlated Dirichlet distribution. The j-th row of each parameter matrix holds the parameters modeling the classifier outputs of examples with true label $t_i = j$. Since for different train/test set splits, the inferred parameters are slightly different, here we show the mean parameters over all five splits for all real data sets.

data set	$\boldsymbol{\alpha}^{\mathbf{1}}$	$\boldsymbol{\alpha}^{\mathbf{2}}$	$\boldsymbol{\delta}$
Bookies A	$\begin{bmatrix} 6.460 & 3.365 & 2.847 \\ 5.860 & 4.018 & 4.149 \\ 3.877 & 3.459 & 4.638 \end{bmatrix}$	$\begin{bmatrix} 7.150 & 3.781 & 3.262 \\ 6.706 & 4.572 & 4.801 \\ 4.464 & 3.922 & 5.281 \end{bmatrix}$	$\begin{bmatrix} 6.426 & 3.343 & 2.823 \\ 5.833 & 3.993 & 4.123 \\ 3.853 & 3.435 & 4.612 \end{bmatrix}$
Bookies B	$\begin{bmatrix} 7.239 & 3.853 & 3.459 \\ 7.087 & 4.670 & 4.923 \\ 4.788 & 3.784 & 4.773 \end{bmatrix}$	$\begin{bmatrix} 7.673 & 4.086 & 3.786 \\ 7.409 & 4.880 & 5.251 \\ 5.112 & 4.009 & 5.127 \end{bmatrix}$	$\begin{bmatrix} 7.210 & 3.832 & 3.437 \\ 7.058 & 4.647 & 4.898 \\ 4.765 & 3.763 & 4.748 \end{bmatrix}$
DNA A	$\begin{bmatrix} 9.655 & 2.564 & 3.59 \\ 3.585 & 12.398 & 4.153 \\ 3.432 & 3.123 & 7.743 \end{bmatrix}$	$\begin{bmatrix} 10.301 & 2.955 & 4.108 \\ 4.177 & 13.527 & 4.899 \\ 3.848 & 3.544 & 8.645 \end{bmatrix}$	$\begin{bmatrix} 9.616 & 2.543 & 3.564 \\ 3.558 & 12.345 & 4.125 \\ 3.409 & 3.1 & 7.712 \end{bmatrix}$
DNA B	$\begin{bmatrix} 13.176 & 9.762 & 12.408 \\ 9.235 & 20.014 & 14.673 \\ 7.141 & 8.442 & 16.428 \end{bmatrix}$	$\begin{bmatrix} 16.403 & 9.584 & 14.081 \\ 11.073 & 21.295 & 16.027 \\ 7.840 & 8.335 & 16.226 \end{bmatrix}$	$\begin{bmatrix} 9.004 & 6.664 & 7.163 \\ 5.838 & 13.53 & 8.122 \\ 4.968 & 5.453 & 10.163 \end{bmatrix}$
DNA C	$\begin{bmatrix} 17.313 & 10.022 & 19.258 \\ 12.478 & 20.759 & 22.879 \\ 8.517 & 7.881 & 21.761 \end{bmatrix}$	$\begin{bmatrix} 10.359 & 4.097 & 9.337 \\ 4.335 & 9.307 & 9.938 \\ 5.528 & 4.838 & 19.167 \end{bmatrix}$	$\begin{bmatrix} 8.014 & 4.000 & 8.936 \\ 4.237 & 7.569 & 8.836 \\ 4.989 & 3.790 & 14.264 \end{bmatrix}$

Table 3.6: Model parameters of the Correlated Fusion Model that we inferred for the real data set with $K = 3$ classifiers, Bookies C. $\boldsymbol{\alpha^1}$ holds the marginal Dirichlet parameters of classifier C_1, $\boldsymbol{\alpha^2}$ the ones of C_2, and $\boldsymbol{\alpha^3}$ the ones of C_3. The $\boldsymbol{\delta}$ parameters hold the correlation parameters of the correlated Dirichlet distribution. $\boldsymbol{\delta^{12}}$ defines the pairwise correlation between C_1 and C_2, $\boldsymbol{\delta^{13}}$ between C_1 and C_3, and $\boldsymbol{\delta^{23}}$ between C_2 and C_3. $\boldsymbol{\delta^a}$ holds the common correlation parameters for all three classifiers. The j-th row of each parameter matrix holds the parameters modeling the classifier outputs of examples with true label $t_i = j$. Since for different train/test set splits, the inferred parameters are slightly different, here we show the mean parameters over all five splits.

data set	$\boldsymbol{\alpha^1}$	$\boldsymbol{\alpha^2}$	$\boldsymbol{\alpha^3}$
Bookies C	$\begin{bmatrix} 6.456 & 3.363 & 2.845 \\ 5.856 & 4.016 & 4.148 \\ 3.872 & 3.455 & 4.631 \end{bmatrix}$	$\begin{bmatrix} 7.14 & 3.775 & 3.258 \\ 6.694 & 4.564 & 4.792 \\ 4.456 & 3.914 & 5.27 \end{bmatrix}$	$\begin{bmatrix} 6.033 & 3.343 & 2.573 \\ 5.484 & 3.757 & 3.814 \\ 3.594 & 3.235 & 4.329 \end{bmatrix}$
	$\boldsymbol{\delta^{12}}$	$\boldsymbol{\delta^{13}}$	$\boldsymbol{\delta^{23}}$
	$\begin{bmatrix} 0.448 & 0.260 & 0.288 \\ 0.394 & 0.277 & 0.349 \\ 0.298 & 0.234 & 0.319 \end{bmatrix}$	$\begin{bmatrix} 0.037 & 0.024 & 0.025 \\ 0.031 & 0.025 & 0.03 \\ 0.027 & 0.025 & 0.031 \end{bmatrix}$	$\begin{bmatrix} 0.034 & 0.025 & 0.026 \\ 0.031 & 0.026 & 0.027 \\ 0.028 & 0.023 & 0.027 \end{bmatrix}$
	$\boldsymbol{\delta^a}$		
	$\begin{bmatrix} 5.931 & 3.049 & 2.5 \\ 5.392 & 3.681 & 3.731 \\ 3.514 & 3.166 & 4.245 \end{bmatrix}$		

4

PARAMETER ESTIMATION FOR A BIVARIATE BETA DISTRIBUTION

A special case of the correlated Dirichlet distribution introduced in the previous Chapter 3 is the bivariate beta distribution. Instead of two marginally Dirichlet-distributed random vectors, it models two marginally beta-distributed random variables with a positive correlation. Those beta-distributed random variables can be proportions or probabilities, e.g., probabilistic forecasts from human experts or classifiers, such as in Chapter 3, but only for binary problems. Such forecasts, either provided by humans or machine learning algorithms, are important in many domains, among them finance and economics, business and marketing, politics, public health, engineering, and meteorological, ecological, and environmental science (McAndrew et al., 2021).

In theory, these probability estimates can be modeled with an arbitrary distribution on the interval $[0, 1]$, such as the beta distribution, beta-generated distributions, the Kumaraswamy distribution, or any distribution on the real numbers transformed through the logistic function. However, in many cases the probabilities provided by different experts or classifiers will be correlated (Jacobs, 1995), e.g., because classifiers have been trained on the same data or experts have similar knowledge. In order to be able to model such correlated probabilities, therefore, a bivariate distribution is needed. For example, one can use bivariate generalizations of the Kumaraswamy distribution (Arnold & Ghosh, 2017), bivariate beta-generated distributions (Samanthi & Sepanski, 2019), or a multivariate Gaussian with logistic transformations (Pirs & Strumbelj, 2019). However, the most common choice for modeling such probabilities is the beta distribution, since it is the standard distribution for probabilities in Bayesian statistics and is simply more familiar to practitioners than the other distributions (Johnson et al., 1995; Magnussen, 2004). Therefore, in this work we focus on bivariate beta distributions.

While a multitude of constructions for bivariate beta distributions have been proposed, they have different constraints and properties, which limit their applicability. We will first review previous constructions of bivariate beta distributions together with their respective properties and then examine the most promising construction that can model arbitrary beta marginals with a positive correlation (Magnussen, 2004), which is a special case of the correlated Dirichlet distribution introduced in Chapter 3. For this construction of modeling arbitrary beta marginals with positive correlation, so far, there has not been an exact method for parameter inference and it has thus rarely been used. Here, we therefore introduce a new estimation method for this bivariate beta distribution with arbitrary beta marginals and positive correlation.

Many bivariate beta distributions have been proposed in the literature (Arnold & Ghosh, 2017; Arnold & Ng, 2011; Bran-Cardona et al., 2011; David Sam Jayakumar et al., 2019; El-Bassiouny & Jones, 2009; Gupta et al., 2011; Gupta & Wong, 1985; Jones, 2002; Koutoumanou et al., 2017; Libby & Novick, 1982; Magnussen, 2004; Nadarajah &

Kotz, 2005; Nadarajah et al., 2017; Olkin & Liu, 2003; Olkin & Trikalinos, 2015; Orozco-Castañeda et al., 2012; Samanthi & Sepanski, 2019; Sarabia & Castillo, 2006; Ting Lee, 1996). Among them, some approaches have been derived from general families of bivariate distributions, such as the Farlie-Gumbel-Morgenstern family of distributions (e.g. Gupta & Wong, 1985) and the Sarmanov family of distributions (e.g. Ting Lee, 1996). Others derived a bivariate beta distribution from a bivariate extension of the F distribution (El-Bassiouny & Jones, 2009; Jones, 2002), whereas Nadarajah and Kotz (2005) proposed different bivariate beta distributions constructed from products of univariate beta-distributed random variables. In general, using different copulas one can construct different bivariate distributions with the same beta marginals (Koutoumanou et al., 2017; Samanthi & Sepanski, 2019). However, as beta-distributed random variables can easily be constructed from normalized gamma-distributed random variables, it is natural to try and generalize this construction to the bivariate case. In this vein, several authors have introduced correlations through shared gamma-distributed random variables (Arnold & Ghosh, 2017; Arnold & Ng, 2011; Magnussen, 2004; Olkin & Liu, 2003; Olkin & Trikalinos, 2015). The most straightforward case of this construction has been studied by Olkin and Liu (2003), building on the work of Libby and Novick (1982). They use three independent gamma-distributed random variables U_1, U_2, U_3 with respective shape parameters $\upsilon_1, \upsilon_2, \upsilon_3$ and same scale parameter to construct

$$X' = \frac{U_1}{U_1 + U_3} \quad \text{and} \quad Y' = \frac{U_2}{U_2 + U_3}. \tag{4.1}$$

Using the standard construction of beta variates from gamma variates, the joint distribution of the random variables X' and Y' is a bivariate beta distribution with marginal distributions Beta(υ_1, υ_3) for X' and Beta(υ_2, υ_3) for Y'. The correlation between X' and Y', which is obtained through the shared latent variable U_3 and its parameter υ_3, is in the range [0,1]. For high values of υ_3, the correlation tends to 0 whereas for low values of υ_3, it tends to 1. However, if υ_3 is high, the values of X' and Y' also tend to 0 and if υ_3 is low they tend to 1 accordingly. This behavior severely limits the usefulness of the distribution for most applications. A further limitation is the constraint that the marginal distributions share the same second parameter υ_3. Thus, the bivariate beta distribution proposed by Olkin and Liu does not allow for arbitrary beta marginals, which limits its flexibility in modeling probability forecasts.

Arnold and Ng (2011) proposed a more flexible construction for a bivariate beta distribution. They use five independent gamma-distributed random variables $U_1, \ldots, U_5$ with shape parameters $\upsilon_1, \ldots, \upsilon_5$ and scale parameter 1 to define two correlated random variables

$$X = \frac{U_1 + U_3}{U_1 + U_3 + U_4 + U_5} \quad \text{and} \quad Y = \frac{U_2 + U_4}{U_2 + U_3 + U_4 + U_5} \tag{4.2}$$

with marginal distributions Beta$(\upsilon_1 + \upsilon_3, \upsilon_4 + \upsilon_5)$ for X and Beta$(\upsilon_2 + \upsilon_4, \upsilon_3 + \upsilon_5)$ for Y. Compared to Olkin and Liu (2003), this construction of a bivariate beta distribution can generate all correlations in the range [-1,1] and marginal distributions with differing second parameters. Nevertheless, because of how the two marginals share parameters, not all combinations of parameters of the marginal beta distributions are possible. For example, the marginals Beta$(10, 4)$ for X and Beta$(1, 1)$ for Y cannot be obtained.

Olkin and Trikalinos (2015) base their construction of a bivariate beta distribution on the Dirichlet distribution. $U = (U_{00}, U_{01}, U_{10}, U_{11})$ is drawn from a 4-dimensional Dirichlet distribution with parameters $\upsilon_1, \dots, \upsilon_4$. By just using three of its components, U_{00}, U_{01}, U_{10}, new random variables

$$X = U_{00} + U_{10} \quad \text{and} \quad Y = U_{01} + U_{10} \tag{4.3}$$

are constructed, with marginal distributions $\text{Beta}(\upsilon_1 + \upsilon_3, \upsilon_2 + \upsilon_4)$ for X and $\text{Beta}(\upsilon_2 + \upsilon_3, \upsilon_1 + \upsilon_4)$ for Y. As Dirichlet-distributed random variables can also be constructed from gamma random variables, we can equivalently construct X and Y in equation (4.3) from four independent gamma-distributed random variables $U_1, \dots, U_4$ with shape parameters $\upsilon_1, \dots, \upsilon_4$ and equal scale parameter 1, with

$$X = \frac{U_1 + U_3}{U_1 + U_2 + U_3 + U_4} \quad \text{and} \quad Y = \frac{U_2 + U_3}{U_1 + U_2 + U_3 + U_4}. \tag{4.4}$$

As can easily be seen from this construction, all correlations in the range [-1,1] can be generated. In particular, the correlation tends to -1 if υ_3 and υ_4 tend to 0 and U_3 and U_4 will be negligible compared to U_1 and U_2. In this case $X \approx \frac{U_1}{U_1+U_2} \approx 1 - Y$. Similarly, the higher the values of υ_3 and υ_4 relative to υ_1 and υ_2, the more negligible U_1 and U_2 will be and the correlation increases to 1 until $X \approx \frac{U_3}{U_3+U_4} \approx Y$. Less obviously, a correlation of 0 is obtained in case $\upsilon_1 \cdot \upsilon_2 = \upsilon_3 \cdot \upsilon_4$ (Olkin & Trikalinos, 2015). Still, this construction of a bivariate beta distribution does not allow arbitrary beta marginal distributions. Since all υ_i are constrained to be positive, for some combinations of marginal distributions the resulting system of linear equations for the parameters υ_i has no solution. For example, the two marginals $\text{Beta}(2, 2)$ for X and $\text{Beta}(1, 1)$ for Y cannot be generated, regardless of their correlation.

Magnussen (2004) introduced yet another construction based on six gamma variates. While all the constructions thus far constrain the parameters of the marginal beta distributions, this construction does allow for arbitrary beta marginals with positive correlation, thus providing the necessary flexibility to model probability forecasts. Magnussen's distribution is a special case of a more general 8-parameter bivariate beta distribution introduced by Arnold and Ng (2011) and reviewed in Arnold and Ghosh (2017), which even allows for positive and negative correlations. However, in many applications, it is enough to model positive correlations, for which the less complex 6-parameter distribution is sufficient. For example, if X and Y are probability estimates elicited from two skilled forecasters, we do not expect negative correlations. But we do want to allow for the possibility that their marginal forecasts have different distributions that should not be tied together by parameter constraints on the marginals. Hence, the bivariate beta distribution proposed by Magnussen (2004), which can model arbitrary beta-distributed marginals with a positive correlation, is an appropriate distribution for modeling correlated probability forecasts.

However, just like any other distribution, the bivariate beta distribution can only be used if its parameters can be estimated correctly. While Magnussen (2004) proposes a moment matching approach for fitting the distribution's parameters, this approach relies on a rough and sometimes inaccurate approximation for the covariance. Also, Magnussen (2004) did not discuss the fact that very similar data can be generated with different parameter values, which makes it hard to statistically infer the parameters of the bivariate beta distribution from data without constraining the distribution.

Therefore, in this work we introduce an alternative approach for estimating this bivariate beta distribution's parameters. First, we will derive the full joint distribution, which is missing in the work of Magnussen (2004), probably because it is intractable. We will then clarify the relationship between Magnussen's distribution and the Olkin-Liu distribution (Olkin & Liu, 2003). Using this relationship with the Olkin-Liu distribution we derive all product moments and in particular the exact covariance function (and in passing we correct a small mistake in the product moments from Olkin and Liu (2003)). For parameter inference, we propose to match moments numerically using the exact covariance we derived. While other estimation methods such as Bayesian inference could be used (Crackel & Flegal, 2017), here we focus on moment matching due to its simplicity and efficiency. In order to make parameter inference unambiguous, we additionally show how to reasonably constrain the distribution's parameters. We evaluate the proposed parameter estimation method in a simulation study and demonstrate its practical use on a real data set consisting of predictions from two correlated forecasters. In addition, we discuss the relationship between the distribution's parameters and the correlation. Finally, we show how to extend the bivariate beta distribution to the Dirichlet distribution introduced in Chapter 3, for which the parameter estimation method proposed in this work can also be applied.

The remainder of the chapter is structured as follows. In Section 4.1 we show how to construct the bivariate beta distribution with arbitrary beta marginals. We continue with deriving its joint distribution in Section 4.2, its moments in Section 4.3, and correlation and covariance in Section 4.4. In Section 4.5, we propose and evaluate our approach for parameter inference. Finally, Section 4.6 shows how to generalize the bivariate beta distribution to a correlated Dirichlet distribution.

4.1 Construction of a Bivariate Beta Distribution with Arbitrary Beta Marginals

Magnussen (2004) uses six independent gamma-distributed random variables A_1, A_2, B_1, B_2, D_1, D_2 that are distributed according to

$$\begin{aligned} A_i &\sim \mathrm{Gamma}(\alpha_i, 1) \quad i = 1, 2 \\ B_i &\sim \mathrm{Gamma}(\beta_i, 1) \quad i = 1, 2 \\ D_i &\sim \mathrm{Gamma}(\delta_i, 1) \quad i = 1, 2, \end{aligned} \tag{4.5}$$

to construct two bivariate-beta-distributed random variables

$$X = \frac{A_1 + D_1}{A_1 + A_2 + D_1 + D_2} \quad \text{and} \quad Y = \frac{B_1 + D_1}{B_1 + B_2 + D_1 + D_2}. \tag{4.6}$$

The resulting marginal distributions of X and Y are Beta(a_1, a_2) and Beta(b_1, b_2) with

$$a_1 = \alpha_1 + \delta_1 \qquad a_2 = \alpha_2 + \delta_2 \qquad b_1 = \beta_1 + \delta_1 \qquad b_2 = \beta_2 + \delta_2. \tag{4.7}$$

The marginals follow immediately from the definition because the sum of gamma random variables of the same scale is gamma-distributed with the same scale but with the original shape parameters summed. In contrast to other constructions that were discussed above (Arnold & Ng, 2011; Olkin & Liu, 2003; Olkin & Trikalinos, 2015), this construction allows for arbitrary marginal distributions. In particular, when δ_1 and δ_2 tend to zero, we can model arbitrary independent marginal distributions Beta(α_1, α_2) and Beta(β_1, β_2).

Since all parameters $\alpha_1, \alpha_2, \beta_1, \beta_2, \delta_1, \delta_2$ need to be positive by definition, for fixed marginal distributions $\mathrm{Beta}(a_1, a_2)$ for X and $\mathrm{Beta}(b_1, b_2)$ for Y it must hold that $\delta_1 < \delta_1^{\max} = \min(a_1, b_1)$ and $\delta_2 < \delta_2^{\max} = \min(a_2, b_2)$. Therefore, for most marginal distributions the maximum correlation that can be generated is below 1. The higher the difference between two marginal distributions, the lower the possible maximum correlation. A perfect correlation approaching 1 can, of course, only be generated for equal marginal distributions, i.e., if $a_1 = b_1$ and $a_2 = b_2$ and α_1, α_2, β_1, and β_2 tend to 0, as also noted by Magnussen (2004). Note that this limitation applies to other bivariate distributions that do not allow for arbitrary marginal beta distributions as well (e.g., Olkin & Trikalinos, 2015).

The construction of this bivariate beta distribution can also be seen as a pairwise combination of three beta distributions. First transform the six independent gamma-distributed random variables (4.5) into three independent gamma- and three independent beta-distributed random variables,

$$\begin{aligned}
U_1 &= A_1 + A_2, & U_1 &\sim \mathrm{Gamma}(\upsilon_1, 1)\\
U_2 &= B_1 + B_2, & U_2 &\sim \mathrm{Gamma}(\upsilon_2, 1)\\
U_3 &= D_1 + D_2, & U_3 &\sim \mathrm{Gamma}(\upsilon_3, 1)\\
W_1 &= \frac{A_1}{A_1 + A_2}, & W_1 &\sim \mathrm{Beta}(\alpha_1, \alpha_2)\\
W_2 &= \frac{B_1}{B_1 + B_2}, & W_2 &\sim \mathrm{Beta}(\beta_1, \beta_2)\\
W_3 &= \frac{D_1}{D_1 + D_2}, & W_3 &\sim \mathrm{Beta}(\delta_1, \delta_2)
\end{aligned} \tag{4.8}$$

with

$$\upsilon_1 = \alpha_1 + \alpha_2 \qquad \upsilon_2 = \beta_1 + \beta_2 \qquad \upsilon_3 = \delta_1 + \delta_2. \tag{4.9}$$

With these definitions we can then rewrite construction (4.6) as

$$\begin{aligned}
X &= \frac{U_1}{U_1 + U_3} \cdot W_1 + \frac{U_3}{U_1 + U_3} \cdot W_3 = X'W_1 + (1 - X')W_3\\
Y &= \frac{U_2}{U_2 + U_3} \cdot W_2 + \frac{U_3}{U_2 + U_3} \cdot W_3 = Y'W_2 + (1 - Y')W_3,
\end{aligned} \tag{4.10}$$

where X' and Y' are defined as in (4.1) but with υ_1, υ_2, and υ_3 as in (4.9). Furthermore, X' and Y' are independent of W_1, W_2, W_3. If parameters δ_1 and δ_2 and with them U_3 tend to 0, $X \approx W_1$ and $Y \approx W_2$ are independent with marginal distributions $\mathrm{Beta}(\alpha_1, \alpha_2)$ for X and $\mathrm{Beta}(\beta_1, \beta_2)$ for Y. Mixing in the shared component W_3 by increasing the values of parameters δ_1 and δ_2 increases the correlation between X and Y. If U_1 and U_2 are negligible compared to U_3 because δ_1 and δ_2 dominate the parameters, the correlation will be close to 1 with $X \approx W_3 \approx Y$ and hence X and Y have the same marginal distribution $\mathrm{Beta}(\delta_1, \delta_2)$, as mentioned before.

4.2 Joint Distribution

X and Y in (4.10) are linear transformations of X' and Y'. Given W_1, W_2, W_3 it is easy to recover X' and Y' from observed X and Y,

$$\begin{aligned} X' &= \frac{X - W_3}{W_1 - W_3} = \frac{|X - W_3|}{|W_1 - W_3|} = f_1(X) \\ Y' &= \frac{Y - W_3}{W_2 - W_3} = \frac{|Y - W_3|}{|W_2 - W_3|} = f_2(Y). \end{aligned} \tag{4.11}$$

Note that we can ignore the sign because according to (4.10) X is always between W_1 and W_3 and Y between W_2 and W_3, so that the numerator and denominator always have the same sign.

As X' and Y' jointly follow the Olkin-Liu distribution (Olkin & Liu, 2003),

$$p'(x', y') = \frac{(x')^{\upsilon_1 - 1} (1 - x')^{\upsilon_2 + \upsilon_3 - 1} (y')^{\upsilon_2 - 1} (1 - y')^{\upsilon_1 + \upsilon_3 - 1}}{B(\upsilon_1, \upsilon_2, \upsilon_3) (1 - x'y')^{\upsilon_1 + \upsilon_2 + \upsilon_3}}, \tag{4.12}$$

where $B(\upsilon_1, \upsilon_2, \upsilon_3) = \frac{\Gamma(\upsilon_1)\Gamma(\upsilon_2)\Gamma(\upsilon_3)}{\Gamma(\upsilon_1 + \upsilon_2 + \upsilon_3)}$, the joint distribution of X and Y given W_1, W_2, W_3 is

$$\begin{aligned} p(x, y \mid w_1, w_2, w_3) &= \left| \frac{df_1(x)}{dx} \frac{df_2(y)}{dy} \right| p' \left(f_1(x), f_2(y) \right) \\ &= \frac{1}{|w_1 - w_3||w_2 - w_3|} \\ &\quad \frac{\left(\frac{|x - w_3|}{|w_1 - w_3|} \right)^{\upsilon_1 - 1} \left(1 - \frac{|x - w_3|}{|w_1 - w_3|} \right)^{\upsilon_2 + \upsilon_3 - 1} \left(\frac{|y - w_3|}{|w_2 - w_3|} \right)^{\upsilon_2 - 1} \left(1 - \frac{|y - w_3|}{|w_2 - w_3|} \right)^{\upsilon_1 + \upsilon_3 - 1}}{B(\upsilon_1, \upsilon_2, \upsilon_3) \left(1 - \frac{|x - w_3|}{|w_1 - w_3|} \frac{|y - w_3|}{|w_2 - w_3|} \right)^{\upsilon_1 + \upsilon_2 + \upsilon_3}} \\ &= \frac{|w_1 - w_3||w_2 - w_3|}{B(\upsilon_1, \upsilon_2, \upsilon_3)} \\ &\quad \cdot \frac{|x - w_3|^{\upsilon_1 - 1} |x - w_1|^{\upsilon_2 + \upsilon_3 - 1} |y - w_3|^{\upsilon_2 - 1} |y - w_2|^{\upsilon_1 + \upsilon_3 - 1}}{(|w_1 - w_3||w_2 - w_3| - |x - w_3||y - w_3|)^{\upsilon_1 + \upsilon_2 + \upsilon_3}} \end{aligned} \tag{4.13}$$

with x between w_1 and w_3 and y between w_2 and w_3 according to (4.10). We have not been able to integrate out w_1, w_2, w_3 from their joint distribution with x and y. However, we suspect that even if the joint density for X and Y could be expressed in terms of special functions, computing those might not be efficient enough for parameter inference for which we will resort to moment matching. Example plots with smoothed samples for the joint density are shown in Figure 4.1 for several parameter settings showing different marginal distributions for X and Y and different correlations between X and Y. Sampling from the bivariate beta distribution is realized with JAGS (Plummer, 2003).

4.3 Moments

As the marginal distributions for X and Y are beta-distributed, their moments are readily available, even in closed form. Computation of the product moments $\mathrm{E}(X^k Y^l)$ is more challenging but can be realized with help of the work of Olkin and Liu (2003). Looking at construction (4.10), X and Y are a linear combination of independent beta-distributed

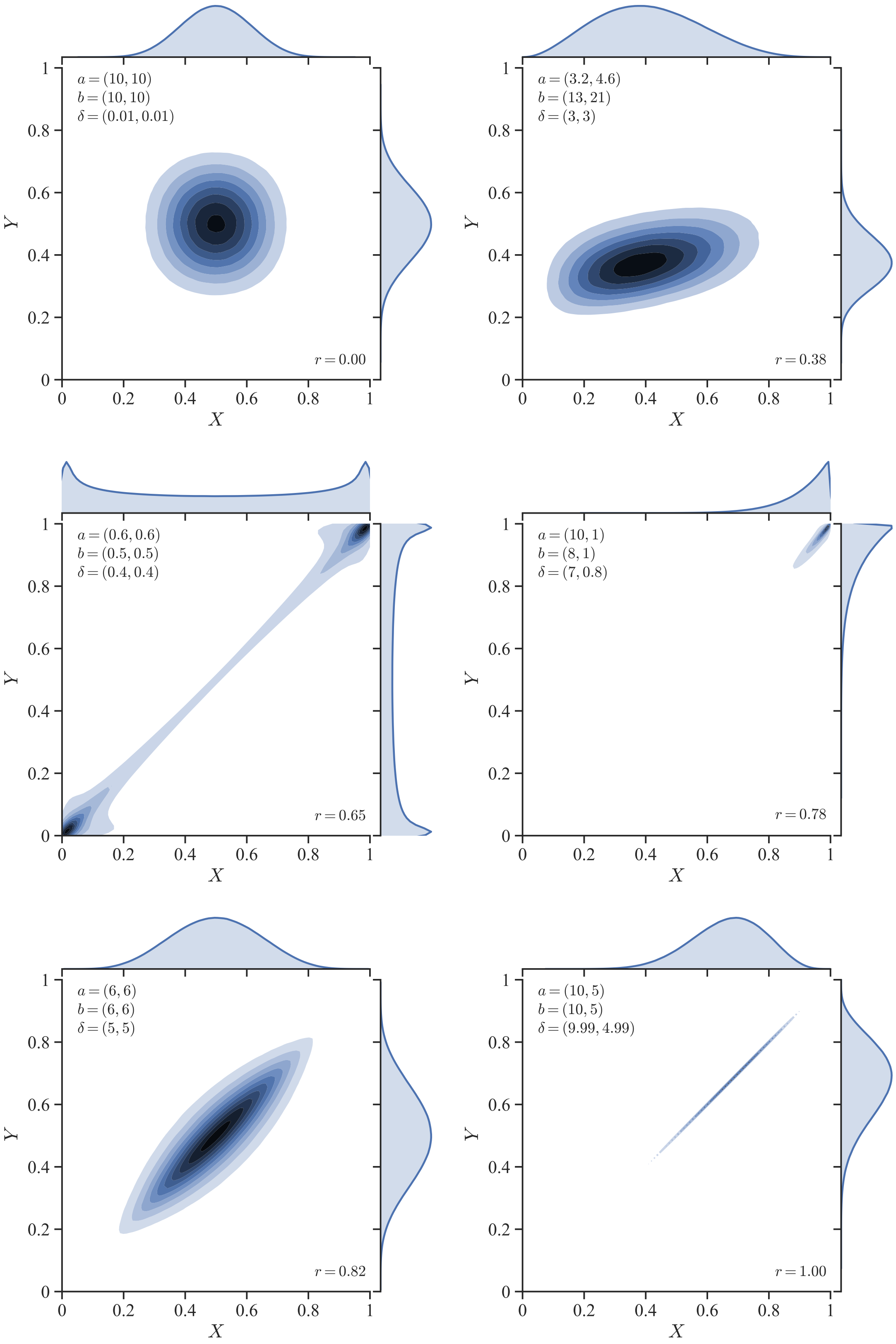


Figure 4.1: Joint densities of bivariate beta distributions with selected parameter values. The plots were created with kernel density estimation based on 10 million samples of the respective distributions. Note that the smoothing is inaccurate at the borders and produces artifacts close to zero and one as a consequence of smoothing with a symmetric kernel.

random variables W_1, W_2, W_3 with weights X' and Y'. Thus, we can express the product moments as

$$\mathrm{E}(X^k Y^l) = \mathrm{E}\left((X'W_1 + (1-X')W_3)^k \, (Y'W_2 + (1-Y')W_3)^l\right). \tag{4.14}$$

Since X' and Y' are independent of W_1, W_2, W_3, it is possible to compute the expectation if the moments of W_1, W_2, W_3, X', Y' and the product moments of X' and Y' are known. W_1, W_2, W_3 as well as the marginals of X' and Y' are beta-distributed, so their moments can be computed straightforwardly in closed form. Furthermore, X' and Y' are jointly Olkin-Liu distributed according to (4.12), and Olkin and Liu (2003) have shown how to compute their product moments. However, note that the derivation of $\mathrm{E}((X')^k(Y')^l)$ in equation (2.2) in the work of Olkin and Liu (2003) is incorrect and should read

$$\begin{aligned}
\mathrm{E}\left((X')^k(Y')^l\right) &= \sum_{j=0}^{\infty} dA(j) \frac{B(\upsilon_1+k+j, \upsilon_2+\upsilon_3)}{B(\upsilon_1+j, \upsilon_2+\upsilon_3)} \frac{B(\upsilon_2+l+j, \upsilon_1+\upsilon_3)}{B(\upsilon_2+j, \upsilon_1+\upsilon_3)} && (4.15)\\
&= \frac{\Gamma(\upsilon_1+\upsilon_3)\Gamma(\upsilon_2+\upsilon_3)\Gamma(\upsilon_1+k)\Gamma(\upsilon_2+l)\Gamma(\Upsilon)}{\Gamma(\upsilon_1)\Gamma(\upsilon_2)\Gamma(\upsilon_3)\Gamma(\Upsilon+k)\Gamma(\Upsilon+l)} \\
&\quad \sum_{j=0}^{\infty} \frac{(\upsilon_1+k)_j(\upsilon_2+l)_j(\Upsilon)_j}{(\Upsilon+k)_j(\Upsilon+l)_j} \frac{1}{j!} && (4.16)\\
&= h \cdot {}_3F_2(\upsilon_1+k, \upsilon_2+l, \Upsilon; \Upsilon+k, \Upsilon+l; 1), && (4.17)
\end{aligned}$$

with

$$\begin{aligned}
d &= \frac{\Gamma(\upsilon_1+\upsilon_3)\Gamma(\upsilon_2+\upsilon_3)}{\Gamma(\upsilon_3)\Gamma(\Upsilon)} && (4.18)\\
A(j) &= \frac{\Gamma(\upsilon_1+j)}{\Gamma(\upsilon_1)} \frac{\Gamma(\upsilon_2+j)}{\Gamma(\upsilon_2)} \frac{\Gamma(\Upsilon)}{\Gamma(\Upsilon+j)} \frac{1}{j!} && (4.19)\\
h &= \frac{\Gamma(\upsilon_1+\upsilon_3)\Gamma(\upsilon_2+\upsilon_3)\Gamma(\upsilon_1+k)\Gamma(\upsilon_2+l)\Gamma(\Upsilon)}{\Gamma(\upsilon_1)\Gamma(\upsilon_2)\Gamma(\upsilon_3)\Gamma(\Upsilon+k)\Gamma(\Upsilon+l)}, && (4.20)
\end{aligned}$$

where $\Upsilon = \upsilon_1 + \upsilon_2 + \upsilon_3$ and ${}_pF_q$ is the generalized hypergeometric function. Equations (4.15), (4.18), and (4.19) are taken directly from Olkin and Liu (2003) with $a = \upsilon_1, b = \upsilon_2, c = \upsilon_3$. Equations (4.16), (4.17), and (4.20) are our corrections of their equations.

4.4 Correlation and Covariance

The correlation r between X and Y is

$$r = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}} \tag{4.21}$$

with the known variances of the beta marginals

$$\begin{aligned}
\mathrm{Var}(X) &= \frac{a_1 a_2}{(a_1+a_2)^2(a_1+a_2+1)} \\
&= \frac{(\alpha_1+\delta_1)(\alpha_2+\delta_2)}{(\alpha_1+\delta_1+\alpha_2+\delta_2)^2(\alpha_1+\delta_1+\alpha_2+\delta_2+1)} \\
\mathrm{Var}(Y) &= \frac{b_1 b_2}{(b_1+b_2)^2(b_1+b_2+1)} \\
&= \frac{(\beta_1+\delta_1)(\beta_2+\delta_2)}{(\beta_1+\delta_1+\beta_2+\delta_2)^2(\beta_1+\delta_1+\beta_2+\delta_2+1)}.
\end{aligned} \tag{4.22}$$

For the covariance of X and Y Magnussen (2004) gives an approximate solution, namely

$$\operatorname{Cov}(X,Y) \approx \frac{a_1 a_2 \delta_2 + (1+b_1)(1+b_2)\delta_1}{(a_1+b_1)(a_2+b_2)(1+a_1+b_1)(1+a_2+b_2)}, \tag{4.23}$$

where a_1, a_2, b_1, and b_2 are defined based on $\alpha_1, \alpha_2, \beta_1, \beta_2, \delta_1, \delta_2$ as in (4.7). This approximation is inaccurate for small values of these parameters, e.g., for $a_1 = a_2 = b_1 = b_2 = 4, \delta_1 = \delta_2 = 3$, the approximated covariance is $\operatorname{Cov}(X,Y) = 0.024$, while the true covariance computed from 10^6 samples of the bivariate beta distribution is $\operatorname{Cov}(X,Y) = 0.020$. This might seem like a small difference but it results in an overestimated correlation of $r = 0.854$ as opposed to the true correlation of $r = 0.730$. Even more worryingly, for $a_1 = a_2 = b_1 = b_2 = 1, \delta_1 = \delta_2 = \frac{4}{5}$, the approximated covariance is $\operatorname{Cov}(X,Y) = \frac{1}{9}$, which results in an estimated correlation of $r = \frac{4}{3}$, which is greater than 1 and therefore wrong by definition.

Given the connection to the Olkin-Liu distribution (Olkin & Liu, 2003), which we derived in Section 4.3, we therefore proceed to compute the exact covariance between X and Y:

$$\operatorname{Cov}(X,Y) = \operatorname{E}(XY) - \operatorname{E}(X)\operatorname{E}(Y) \tag{4.24}$$

where

$$\begin{aligned} \operatorname{E}(X) &= \frac{a_1}{a_1+a_2} = \frac{\alpha_1+\delta_1}{\alpha_1+\delta_1+\alpha_2+\delta_2} \\ \operatorname{E}(Y) &= \frac{b_1}{b_1+b_2} = \frac{\beta_1+\delta_1}{\beta_1+\delta_1+\beta_2+\delta_2} \end{aligned} \tag{4.25}$$

are readily available as the means of the beta marginals. We can compute $\operatorname{E}(XY)$ from (4.14) with $k = l = 1$, which results in

$$\begin{aligned} \operatorname{E}(XY) &= \operatorname{E}\left((X'W_1 + (1-X')W_3)(Y'W_2 + (1-Y')W_3)\right) \\ &= \operatorname{E}(X'Y')\operatorname{E}(W_1)\operatorname{E}(W_2) + \left(\operatorname{E}(X') - \operatorname{E}(X'Y')\right)\operatorname{E}(W_1)\operatorname{E}(W_3) \\ &\quad + \left(\operatorname{E}(Y') - \operatorname{E}(X'Y')\right)\operatorname{E}(W_2)\operatorname{E}(W_3) \\ &\quad + \left(1 - \operatorname{E}(Y') - \operatorname{E}(X') + \operatorname{E}(X'Y')\right)\operatorname{E}(W_3^2) \end{aligned} \tag{4.26}$$

with the moments of the beta marginals from (4.8)

$$\begin{aligned} \operatorname{E}(W_1) &= \frac{\alpha_1}{\alpha_1+\alpha_2} \\ \operatorname{E}(W_2) &= \frac{\beta_1}{\beta_1+\beta_2} \\ \operatorname{E}(W_3) &= \frac{\delta_1}{\delta_1+\delta_2} \\ \operatorname{E}(W_3^2) &= \frac{\delta_1(\delta_1+1)}{(\delta_1+\delta_2+1)(\delta_1+\delta_2)} \end{aligned} \tag{4.27}$$

and (4.10)

$$\begin{aligned} \operatorname{E}(X') &= \frac{\upsilon_1}{\upsilon_1+\upsilon_3} \\ \operatorname{E}(Y') &= \frac{\upsilon_2}{\upsilon_2+\upsilon_3} \end{aligned} \tag{4.28}$$

with υ_1, υ_2, and υ_3 as defined in (4.9). $\mathrm{E}(X'Y')$ can be specialized from the moments (4.17) above with $k = l = 1$:

$$\begin{aligned} \mathrm{E}(X'Y') &= h\, {}_3F_2(\upsilon_1 + 1, \upsilon_2 + 1, \Upsilon; \Upsilon + 1, \Upsilon + 1; 1) \qquad \text{with} \\ h &= \frac{\Gamma(\upsilon_1 + \upsilon_3)\Gamma(\upsilon_2 + \upsilon_3)}{\Gamma(\upsilon_3)\Gamma(\Upsilon + 1)} \cdot \frac{\upsilon_1 \upsilon_2}{\Upsilon} \end{aligned} \tag{4.29}$$

and $\Upsilon = \upsilon_1 + \upsilon_2 + \upsilon_3$ as before. According to Olkin and Liu (2003) there is no closed-form solution for (4.17) and (4.29). However, using the generalized hypergeometric function the product moments and thus the covariance between X and Y can be computed numerically (e.g., by using the hyper function in sympy or the HypergeometricPFQ function in Mathematica). Note that with Raabe's test one can show that the generalized hypergeometric function ${}_3F_2$ in (4.29) converges. According to Raabe's test the series converges if

$$\lim_{j\to\infty} \rho_j > 1 \text{ with } \rho_j = j\left(\frac{c_j}{c_{j+1}} - 1\right), \tag{4.30}$$

where c_j is the j-th element of the series described by the generalized hypergeometric function ${}_3F_2$ in (4.29), which is

$$c_j = \frac{(\upsilon_1 + 1)_j(\upsilon_2 + 1)_j(\Upsilon)_j}{(\Upsilon + 1)_j(\Upsilon + 1)_j} \frac{1}{j!}. \tag{4.31}$$

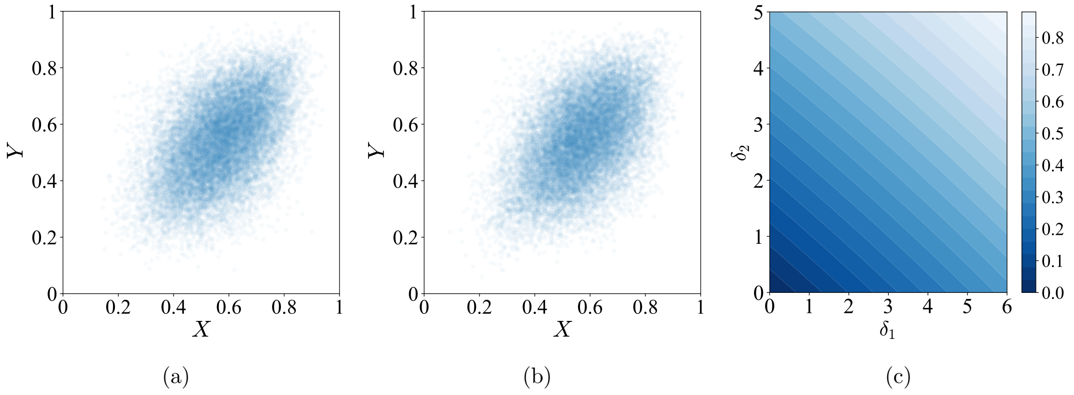
Since $\lim_{j\to\infty} \rho_j = 1 + \upsilon_3 > 1$, the generalized hypergeometric function ${}_3F_2$ in (4.29) converges. However, this analysis also suggests that convergence of the series might be very slow for small $\upsilon_3 = \delta_1 + \delta_2$.

4.5 Parameter Inference

Magnussen (2004) used the method of moments to infer the parameters a_1, a_2, b_1, b_2 for the marginal distributions. Given these parameters he then matched the empirical correlation to the correlation for the parameters δ_1, δ_2 (given marginal parameters a_1, a_2, b_1, b_2) using the approximate solution for the covariance given in (4.23). However, as shown in Section 4.4 this approximation can lead to very inaccurate correlation estimates. An additional problem with the distribution is that very similar data can be generated with different parameter values, as one can see in Figure 4.2(a) and (b). Increasing δ_1 and simultaneously decreasing δ_2 or vice versa while keeping the marginal parameters a_1, a_2 and b_1, b_2 fixed, can result in very similar correlations, which is shown in Figure 4.2(c). Since two distributions with very different parameters can lead to extremely similar data, it is hard to statistically infer the parameters δ_1 and δ_2 from data: The empirical correlation alone does not provide enough constraints and differences in higher moments can be subtle.

Therefore, in order to make parameter inference unambiguous, we decided to constrain the 6-parameter bivariate beta distribution to five parameters: two for each marginal and one parameter to control the correlation. A reasonable way to constrain δ_1 and δ_2 is to set

$$\delta_2 = \frac{\delta_2^{\max}}{\delta_1^{\max}} \delta_1 \tag{4.32}$$



(a)

(b)

(c)

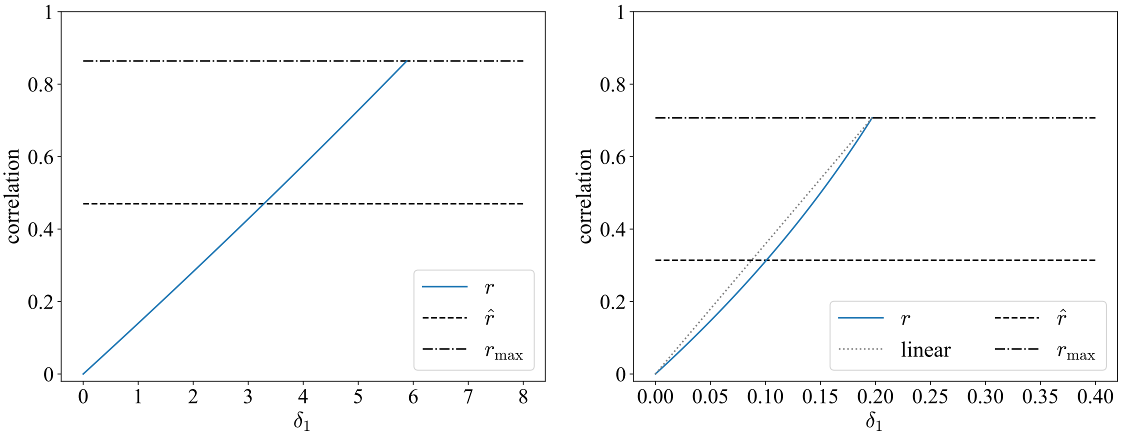
Figure 4.2: Different parameters can generate similar data. (a) and (b) show samples generated from two bivariate beta distributions with the same marginal parameters $a = (8, 6)$, $b = (6, 5)$ and correlation parameters $\delta = (3.28, 2.73)$ for (a) and $\delta = (5, 1.6)$ for (b). Both, the data in (a) and (b) show a correlation of $r = 0.47$. Correspondingly, (c) shows the correlations generated by different combinations of δ_1 and δ_2 given marginal parameters $a = (8, 6)$ and $b = (6, 5)$. Different combinations of δ_1 and δ_2 can lead to very similar correlations.

with $\delta_1^{\max} = \min(a_1, b_1), \delta_2^{\max} = \min(a_2, b_2)$, because this enables the maximum possible correlation between X and Y when the maximum values for δ_1 and δ_2 are attained and the shared component between X and Y is as big as it can be without violating the marginal constraints.

4.5.1 Moment Matching

Using this constraint (4.32), the model-inherent constraints $\delta_1 < \delta_1^{\max}, \delta_2 < \delta_2^{\max}$, and the formula for the correlation derived in Section 4.4, we can now optimize the parameters numerically to match the empirical moments. First, the marginal parameters a_1, a_2, b_1, b_2 are obtained using the standard procedure of moment matching for the beta distribution. Given the estimated marginal parameters, an estimate of δ_1 (and with it δ_2) can be obtained numerically by minimizing the quadratic deviation between the theoretical correlation and the empirical correlation. To avoid the undefined cases $\delta_1 \leq 0$ and $\delta_1 \geq \delta_1^{\max}$ we bound the optimization between ε and $\delta_1^{\max} - \varepsilon$ with $\varepsilon = 0.001$. Unless the empirical correlation is bigger than the maximum correlation that can be attained with the matched marginals or smaller than 0, it can be matched exactly for some δ_1. Otherwise δ_1 will take on its maximum value $\delta_1^{\max} - \varepsilon$ or its minimal value ε. We implement inference in Python using the package mpmath (Johansson et al., 2013) for the numerical computation of the generalized hypergeometric function and the package scipy (Virtanen et al., 2020) for optimization.

As an example we used this numerical moment matching approach on 5000 data points generated with parameters $a_1 = 8$, $a_2 = 6$, $b_1 = 6$, $b_2 = 5$, $\delta_1 = 3.28$, $\delta_2 = 2.73$, equivalent to the data shown in Figure 4.2(a). We inferred the parameter values $\hat{a}_1 = 8.143, \hat{a}_2 = 6.193, \hat{b}_1 = 5.882, \hat{b}_2 = 4.931, \hat{\delta}_1 = 3.286, \hat{\delta}_2 = 2.754$. Figure 4.3(a) shows the correlations implied by different values for parameter δ_1 compared to the desired correlation of $\hat{r} = 0.47$. As one can see, the difference between r and $\hat{r}$ is zero for the inferred $\hat{\delta}_1 = 3.286$. Due to our constraint (4.32), $\hat{\delta}_2 = 2.754$ can be computed from δ_1. In this case there is an almost



(a)

(b)

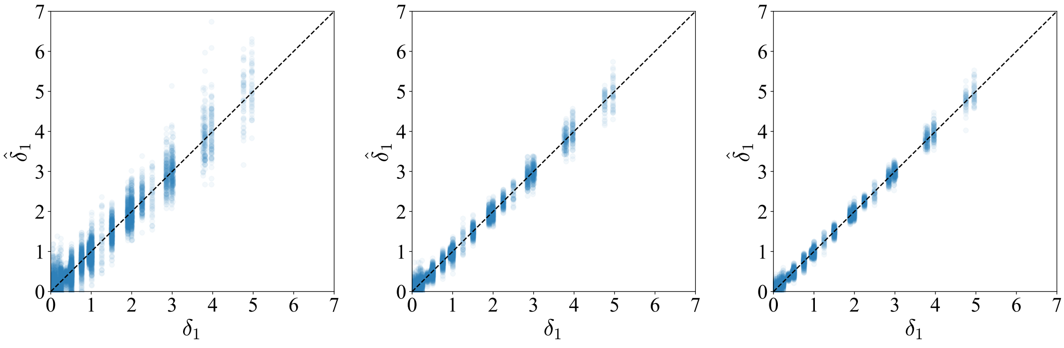
Figure 4.3: The correlations r implied by different values of correlation parameter δ_1 compared to the desired correlation $\hat{r}$ given estimates of the marginal parameters a_1, a_2, b_1, b_2. δ_2 is not displayed since it can be computed from δ_1 using constraint (4.32). $r_{\max}$ is the maximum correlation that can be reached for the given marginal parameters. (a) shows the first inference example with $a_1 = 8.143, a_2 = 6.193, b_1 = 5.882, b_2 = 4.931$, $\hat{r} = 0.47$, and $r_{\max} = 0.864$. The difference between r and $\hat{r}$ is zero for $\delta_1 = 3.286$. In (b) we show the second inference example with $a_1 = 0.197, a_2 = 0.903, b_1 = 0.403, b_2 = 1.017$, $\hat{r} = 0.314$, and $r_{\max} = 0.707$. The difference between r and $\hat{r}$ is zero for $\delta_1 = 0.101$. The dotted linear reference line additionally shows that the relationship between the correlation and δ_1 is non-linear.

linear relationship between δ_1 and r but this is not true in general, especially for smaller parameter values.

This can be seen in a second example where we applied our moment matching approach on 5000 data points generated with parameters $a_1 = 0.2$, $a_2 = 0.9$, $b_1 = 0.4$, $b_2 = 1$, $\delta_1 = 0.1$, $\delta_2 = 0.45$ and received the inferred parameter values $\hat{a}_1 = 0.197, \hat{a}_2 = 0.903, \hat{b}_1 = 0.403, \hat{b}_2 = 1.017, \hat{\delta}_1 = 0.101, \hat{\delta}_2 = 0.462$. As seen in Figure 4.3(b), for this second example the relationship between δ_1 and the correlation is not well approximated by a linear function, in contrast to the first example in Figure 4.3(a). Still, inference works as for the first example shown above.

4.5.2 Simulation Study

The performance of the proposed approach for parameter inference was evaluated in a simulation study. We generated data from a bivariate beta distribution using different marginal parameters, correlation parameters, and different numbers of generated samples N and inferred δ_1 from these data using the proposed moment matching approach. For half of all considered simulations, X and Y were chosen to have the same marginal parameters to be able to generate the full range of correlations from 0 to 1, hence $a_1 = b_1, a_2 = b_2$. All possible combinations of the values in $[0.5, 1, 2, 3, 4, 5]$ for a_1 and a_2 were tested while omitting symmetric cases with $a_1 \leq a_2$, resulting in 21 different marginal distributions. For the remaining simulations we considered differing marginal distributions. We chose a subset of 7 marginal distributions from the set of marginals above as $[[0.5, 0.5], [1, 1], [1, 4], [2, 2], [2, 4], [3, 3], [4, 5]]$ and tested inference for all $\binom{7}{2} = 21$



(a) $N = 100$

(b) $N = 500$

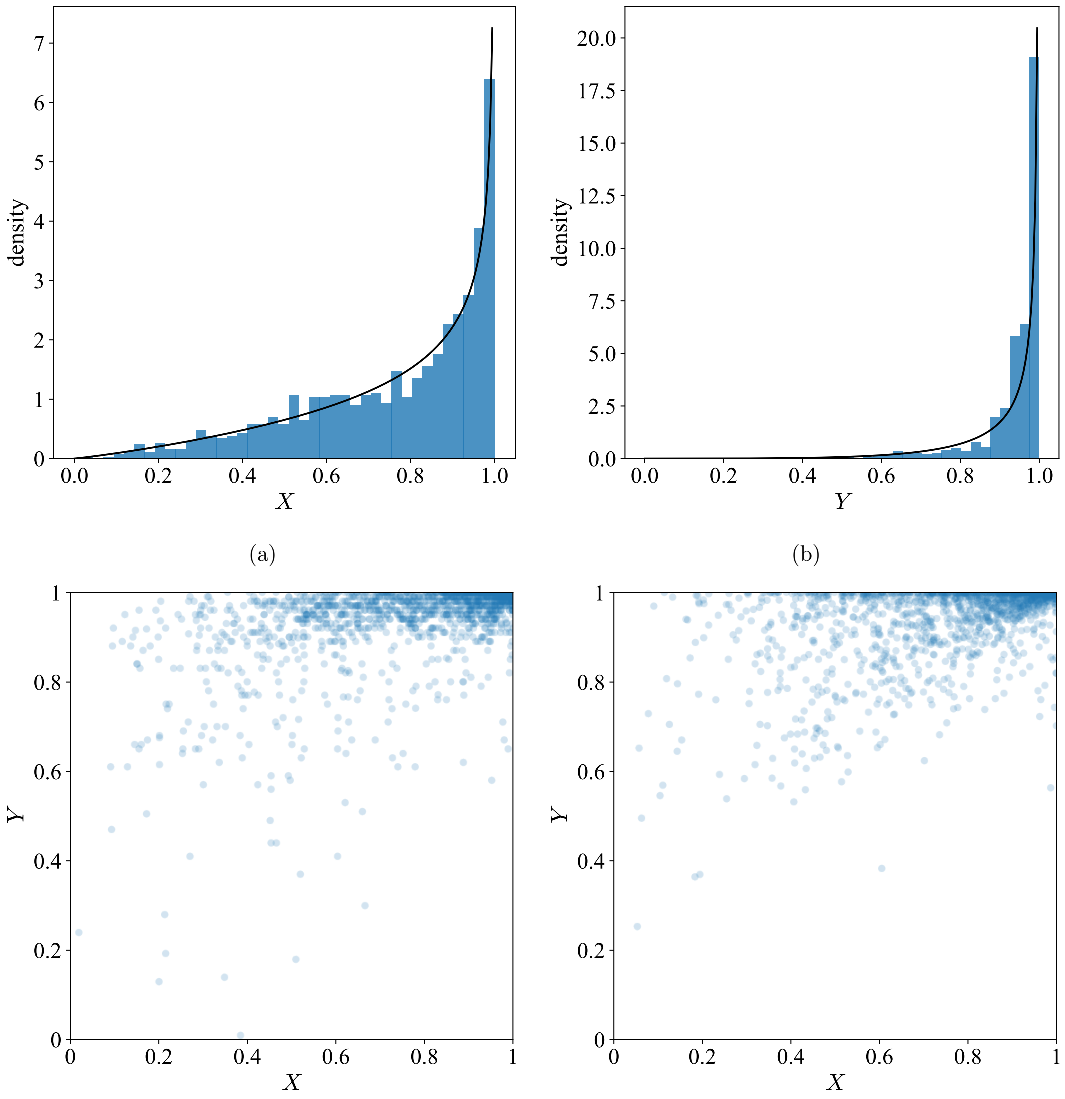
(c) $N = 1000$

Figure 4.4: Results of a simulation study for evaluating the proposed moment matching approach for parameter inference. We generated data from a bivariate beta distribution using different marginal distributions with different correlations and different numbers of generated samples N and inferred δ_1 from these data. We show inferred $\hat{\delta}_1$ against true δ_1 for $N = 100$ samples (a), $N = 500$ samples (b), and $N = 1000$ samples (c). $\hat{\delta}_1$ matches δ_1, the better the higher the number of available data samples N.

combinations of different marginals in this set. Thus, in total we inferred parameters for 42 combinations of marginals. The correlation parameters δ_1 were chosen as $p \cdot \delta_1^{\max}$ with $p = 0.01, 0.05, 0.25, 0.5, 0.75, 0.95, 0.99$ respectively, in order to show how inference works for data with different correlations between 0 and the maximum possible correlation for the respective marginals. δ_2 was obtained using the constraint in (4.32). To test the effect of the number of samples N on the accuracy of parameter inference we evaluated with $N = 100, 500, 1000$. For each of the $42 \cdot 7 \cdot 3 = 882$ parameter settings, we repeated the data simulation and inference process 50 times, resulting in 44100 inference results. These results are displayed in Figure 4.4, for $N = 100, 500, 1000$ in (a), (b), (c) respectively. We can see that the inferred $\hat{\delta}_1$ match the true value of δ_1, the better the higher the number of available data samples N. The average standard deviations of all inferred $\hat{\delta}_1$ are 0.171 for $N = 100$, 0.078 for $N = 500$, and 0.055 for $N = 1000$. Note that although the generalized hypergeometric function ${}_3F_2$ in (4.29) is guaranteed to converge, as we showed in Section 4.4, it can happen that its numerical computation fails due to very slow convergence. In our simulation study, this error occurred for 2.5% of all inference computations, only for low values of δ_1.

4.5.3 Application on a Real Data Set of Correlated Forecasts

The bivariate beta distribution has broad applicability in many fields as diverse as Bayesian analysis, where it can model the correlation among priors for Binomial distributions (Arnold & Ng, 2011), the modeling of proportions of hardwood forests over time, where it serves to estimate decadal changes in the relative land use of a region (Magnussen, 2004), the modeling of proportions of electorate voting in a two candidate election, proportions of substances in mixtures, or brand shares (Gupta et al., 2011), and utility assessment (Libby & Novick, 1982). Furthermore, the bivariate beta distribution can be used for modeling probabilities produced by two correlated forecasters. Correlations between forecasters are quite common, e.g., two bookmakers who base their odds on common information will produce correlated odds. For the same reason experts in risk assessment will often produce



(a) (b) (c) (d)

Figure 4.5: The predicted probabilities from the chess classifier data set and the inferred bivariate beta distribution. (a) and (b) show the marginal distributions of X representing classifier 1 (Bayes Net), and Y representing classifier 2 (Random Forest). For both marginal distributions we show the relative frequencies of the predictions as well as the beta densities inferred with moment matching. (c) jointly shows X and Y, which are correlated with approximately $\hat{r} = 0.483$. (d) shows a simulated data set consisting of 1527 samples of a bivariate beta distribution with the parameters inferred for the chess classifier data set.

correlated forecasts. Similarly, different machine classifiers produce correlated predictions when trained on the same data (Jacobs, 1995; Kim & Ghahramani, 2012). These correlations should be taken into account when their predictions are combined, e.g., in different techniques for classifier fusion (Trick & Rothkopf, 2022, see Chapter 3), since combining correlated classifiers can otherwise lead to overconfidence and high generalization error (Ueda & Nakano, 1996). Here, we use such a data set consisting of the predictions of two classifiers as an illustrative example of the application of the proposed inference method. Two classifiers, a Bayes Net and a Random Forest, were trained on Alen Shapiro's chess (King-Rook vs. King-Pawn) data set (Dua & Graff, 2017; Shapiro, 1987).[1] The task is to predict if King+Pawn will achieve a draw or win in a chess match against King+Rook based on 36 categorical features of a chess position. For training both classifiers, we used 10-fold cross-validation. The two classifiers' predictions on the respective 10 test sets form the data set we evaluate on. We only considered the predicted probabilities of winning King+Pawn or draw for all 1527 match instances that ended either with a win of King+Pawn or a draw. The predicted probabilities of winning King+Rook for the matches actually won by King+Rook might follow a different distribution and are therefore excluded in our example. With the bivariate beta distribution we now model the probabilities the two classifiers predicted for a win of King+Pawn or a draw.

Figure 4.5 shows the data and the inferred distribution. X is the predicted probability for King+Pawn winning or a draw of classifier 1, a Bayes Net, and Y the predicted probability of classifier 2, a Random Forest. In Figure 4.5(a) and (b) we show histograms of the predictions of X and Y together with marginal beta densities that were inferred with moment matching: The parameters are $\hat{a}_1 = 2.094, \hat{a}_2 = 0.64$ for (a) and $\hat{b}_1 = 4.44, \hat{b}_2 = 0.288$ for (b). Figure 4.5(c) jointly shows X and Y with a correlation of approximately $\hat{r} = 0.483$. Matching this correlation, too, as described in Section 4.5.1, we obtain $\hat{\delta}_1 = 1.723$ and $\hat{\delta}_2 = 0.237$. The corresponding correlation is $r = 0.483$, which matches the data's empirical correlation up to numerical precision. Figure 4.5(d) shows a simulated data set consisting of 1527 samples drawn from a bivariate beta distribution with the inferred parameters, $\hat{a}_1 = 2.094, \hat{a}_2 = 0.64, \hat{b}_1 = 4.44, \hat{b}_2 = 0.288, \hat{\delta}_1 = 1.723$ and $\hat{\delta}_2 = 0.237$. As can be seen, the generated data set is similar to the real data set in Figure 4.5(c) and the classifiers' predictions can thus be modeled reasonably well with this bivariate beta distribution.

4.5.4 Relationship Between δ_1 and the Correlation

We found empirically that over a large range of parameters the following approximate relationship holds:

$$r \approx \frac{\delta_1}{\delta_1^{\max}} r_{\max}, \tag{4.33}$$

while $\delta_1^{\max} = \min(a_1, b_1)$ and $r_{\max}$ is the maximum possible correlation for the given marginals. While this relationship could be used for approximate inference of δ_1, we still recommend the moment matching approach proposed in Section 4.5.1 that gives exact results. However, the shown relationship allows interpreting the correlation parameter δ_1. Particularly for equal marginal distribution with $a_1 = b_1$ and $a_2 = b_2$, for which

[1] https://archive.ics.uci.edu/ml/datasets/Chess+%28King-Rook+vs.+King-Pawn%29

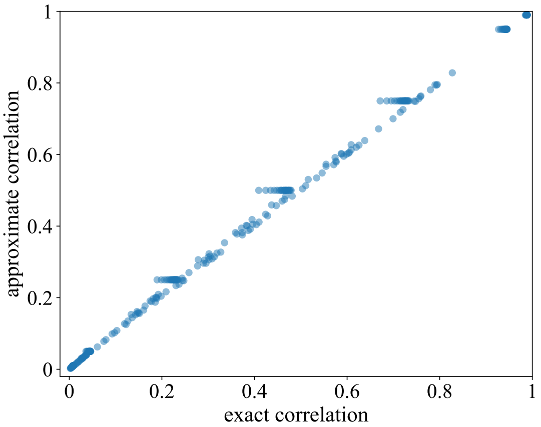


Figure 4.6: The exact correlation computed as in Section 4.4 compared to the approximate correlation computed with equation (4.33) for all parameter configurations used in the simulation study in Section 4.5.2.

$r_{\max} = 1$, this interpretation of δ_1 is very simple: The fraction of $\frac{\delta_1}{\delta_1^{\max}}$ approximately matches the generated correlation. For example, if $a_1 = b_1 = 2$ and $a_2 = b_2 = 4$, for $\delta_1 = 1$ we generate a correlation of $r = 0.468 \approx \frac{1}{2} r_{\max} = 0.5$ with $r_{\max} = 1$. For differing marginal distributions, interpreting δ_1 is more difficult because $r_{\max}$ must be computed numerically using the formulas given above. If, e.g., $a_1 = a_2 = 2$ and $b_1 = 1, b_2 = 4$, with $\delta_1 = 0.5$, we generate a correlation of $0.3 \approx \frac{0.5}{1} r_{\max} = 0.313$ with $r_{\max} = 0.627$. In Figure 4.6, we plot the exact correlation and the approximated correlation computed with (4.33) for all parameter configurations used in the simulation study in Section 4.5.2. As can be seen, the relationship in (4.33) holds for all parameter values. The plateaus seen for approximate correlations of 0.25, 0.5, and 0.75 are a consequence of choosing $\delta_1 = p \cdot \delta_1^{\max}$ with $p = 0.25, 0.5, 0.75$ in the simulation study (Section 4.5.2) leading to approximate correlations of 0.25, 0.5, 0.75 for all simulations with equal marginals. We leave it as an open problem to show when approximation (4.33) holds to what accuracy.

4.6 Generalization to the Correlated Dirichlet Distribution

The bivariate beta distribution can be generalized to a correlated Dirichlet distribution (Trick & Rothkopf, 2022, see Chapter 3) in order to model two positively correlated random vectors $\boldsymbol{X} = (X_1, \ldots, X_k)$ and $\boldsymbol{Y} = (Y_1, \ldots, Y_k)$ with the two marginal vectors being Dirichlet-distributed. A k-dimensional correlated Dirichlet distribution can be constructed from $3k$ gamma-distributed random variables $A_1, \ldots, A_k$, $B_1, \ldots, B_k$, $D_1, \ldots, D_k$ with $3k$ parameters $\alpha_1, \ldots, \alpha_k$, $\beta_1, \ldots \beta_k$, $\delta_1, \ldots \delta_k$ distributed according to

$$\begin{aligned} A_i &\sim \text{Gamma}(\alpha_i, 1) \quad i = 1, \ldots, k \\ B_i &\sim \text{Gamma}(\beta_i, 1) \quad i = 1, \ldots, k \\ D_i &\sim \text{Gamma}(\delta_i, 1) \quad i = 1, \ldots, k. \end{aligned} \tag{4.34}$$

These random variables are used to construct the correlated Dirichlet-distributed random variables $\boldsymbol{X} = (X_1, \ldots, X_k)$ and $\boldsymbol{Y} = (Y_1, \ldots, Y_k)$ with

$$X_i = \frac{A_i + D_i}{\sum_{i=1}^k A_i + \sum_{i=1}^k D_i} \qquad \text{and} \qquad Y_i = \frac{B_i + D_i}{\sum_{i=1}^k B_i + \sum_{i=1}^k D_i}. \tag{4.35}$$

The two resulting marginal distributions are Dirichlet$(\boldsymbol{X}; \alpha_1 + \delta_1, \ldots, \alpha_k + \delta_k)$ and Dirichlet$(\boldsymbol{Y}; \beta_1 + \delta_1, \ldots, \beta_k + \delta_k)$.

Analogous to the example of the bivariate beta distribution in Section 4.5.3, this correlated Dirichlet distribution can be used for modeling non-binary probabilistic predictions of experts, sensors, or classifiers. This is particularly useful for Bayesian approaches to classifier or expert fusion, which are the reason why we started working on this distribution in the first place. For example, in Chapter 3 we apply it to classifier fusion, but instead of using moment matching – as developed here – we use rather inefficient Markov chain methods to sample from the posterior distribution over the parameters (Trick & Rothkopf, 2022, see Chapter 3). Being able to explicitly model the correlation between probabilistic classifiers or probability estimates given by human experts with the correlated Dirichlet distribution allows Bayes optimal fusion of classifiers or experts, avoids overconfidence of the ensemble and thereby improves its performance. Another application of the correlated Dirichlet distribution is the generation of stochastic matrices with individual rows or columns being Dirichlet-distributed and correlated, which can be beneficial for Markov processes, in optimal control, or reinforcement learning.

The derivations of the product moments and the exact covariance of the correlated Dirichlet distribution are analogous to the derivations for the bivariate beta distribution shown in this work. Thus, the parameters of the correlated Dirichlet distribution can also be estimated using the proposed moment matching approach, extended to the higher dimensionality of the Dirichlet distribution.

4.7 Conclusion

In this work, we discussed a bivariate beta distribution that can model arbitrary beta-distributed marginals with a positive correlation, which is constructed from six independent gamma-distributed random variates. While previous work used an approximate and sometimes inaccurate method to compute the distribution's covariance and estimate its parameters, here, we derived all product moments and the exact covariance, which can be computed numerically. Based on this analysis we presented an algorithm for estimating the parameters of the distribution using moment matching and additionally constrained the distribution's parameters in order to make parameter inference unambiguous. We evaluated the proposed inference method in a simulation study, demonstrated its practical use on a data set consisting of predictions from two correlated forecasters, and discussed the relationship between the distribution's parameters and the correlation. Furthermore, we generalized the bivariate beta distribution to a correlated Dirichlet distribution, for which the proposed parameter estimation method can be used analogously.

5 BAYESIAN FUSION FOR INTENTION RECOGNITION IN HUMAN-ROBOT INTERACTION

A prevalent challenge for our society is an increasing number of elderly people in need of care while facing a shortage of nursing staff (Kochskaemper, 2018). A promising solution may come from investigating technical solutions such as assistive robots that can improve the quality of life of elderly people, not just by supporting caregivers but also by directly providing assistance to affected elderly people. In particular, such assistive robots can potentially enable even physically handicapped people to stay in their own habitual environments for a longer time by facilitating harmful or arduous everyday life tasks.

In order to guarantee trouble-free cooperation between human and robot, it is necessary that the robot recognizes human intentions (Hofmann & Williams, 2007). As humans exploit multiple modalities such as speech, body language, and situational cues for understanding intentions (Schrempf et al., 2007), it is reasonable to also take advantage of multimodal data in automatic intention recognition.

Multimodal intention recognition is beneficial in two kinds of ways. On the one hand, it offers the possibility to compensate for limited or missing modalities, e.g., a speech disorder as a consequence of a stroke. On the other hand, through the integration of information of several modalities the uncertainty about the true intention to be recognized can be decreased (Ernst & Banks, 2002). This is particularly important in the interaction between elderly and potentially vulnerable people and robotic systems to ensure safety. However, the reduction of uncertainty through the use of multiple modalities for intention recognition was not the focus of previous works.

In order to explicitly consider and appropriately reduce uncertainty, Bayesian classifier fusion as derived in Chapter 3 can be applied to multimodal intention recognition in human-robot interaction. Since in human-robot interaction tasks it is difficult to obtain large training data sets, the ad-hoc fusion rule Independent Opinion Pool, which we discussed in Section 3.2.1, is of particular interest for human-robot interaction scenarios. In particular, its conditional independence assumption is plausible for data from different modalities, at least shown in the context of human perception (Oruç et al., 2003).

Accordingly, here, we propose a multimodal intention recognition system that reduces uncertainty using classifier fusion with the Bayesian fusion method Independent Opinion Pool (Berger, 1985) (Figure 5.1). Thereby, we consider the four modalities speech, gestures, gaze directions and scene objects. For all modalities an individual classifier is trained that returns a categorical probability distribution over all possible intentions. The resulting distributions are then fused with Independent Opinion Pool (Berger, 1985). Through the application of this method for classifier fusion the uncertainty about the intention to be predicted can be lowered even if the classifiers for the single modalities are individually inaccurate or uncertain. This is shown in a collaborative task where the human is sup-

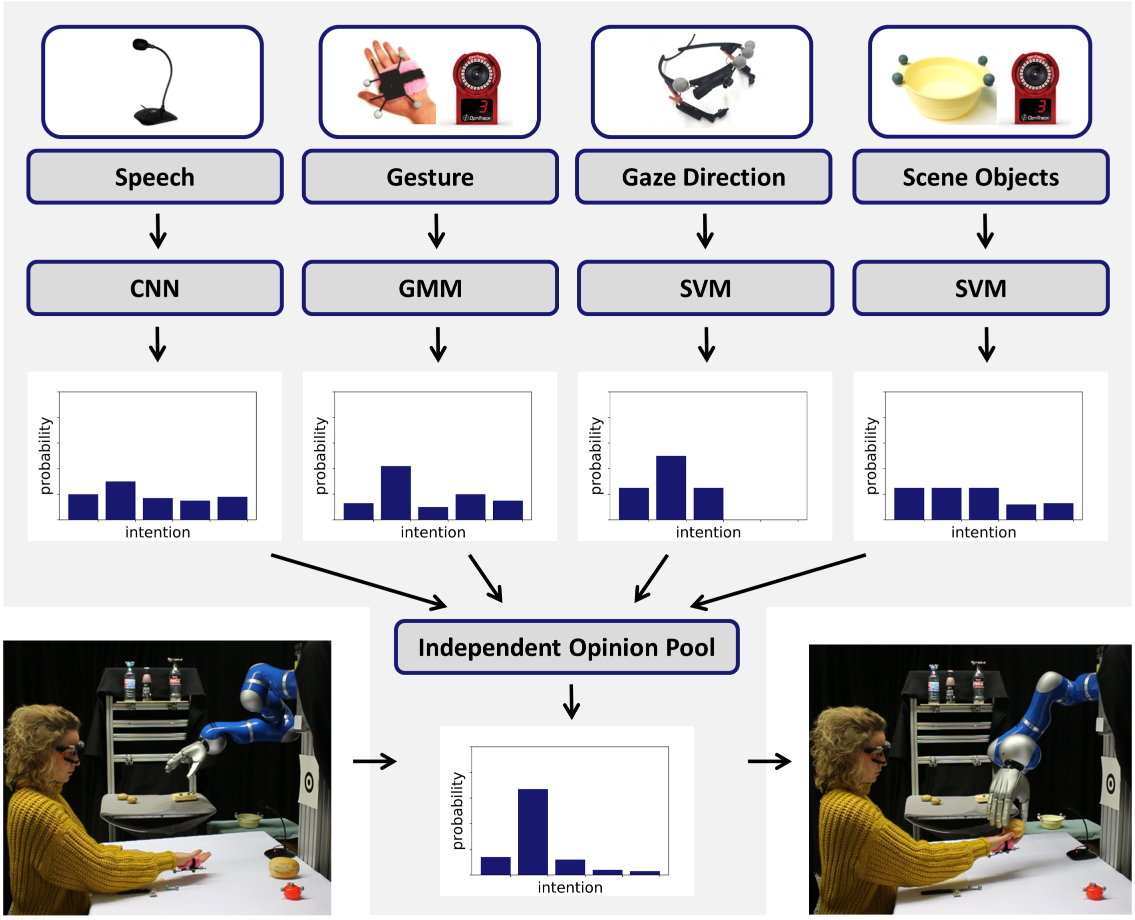


Figure 5.1: An overview over the proposed multimodal intention recognition system, which fuses the modalities speech, gestures, gaze directions, and scene objects. For all modalities individual classifiers are learned. Their output distributions are fused with the Bayesian fusion method Independent Opinion Pool (Berger, 1985) in order to reduce the overall system's uncertainty.

ported by a 7-DoF robot arm in preparing some food in a kitchen scenario while human intentions are recognized online using the four considered modalities.

The rest of the chapter is structured as follows. In Section 5.1 related work is discussed. Section 5.2 explains the chosen approach including methods for classifier fusion and intention recognition from the four considered modalities. In Section 5.3 we evaluate this approach in a collaborative task with a 7-DoF robot and present corresponding results. Finally, we conclude with Section 5.4 and discuss possible future work.

5.1 Related Work

Since the combined utilization of multiple modalities can improve the performance and robustness of a system (Nweke et al., 2019; Rodomagoulakis et al., 2016), there are several approaches that have already dealt with multimodal intention recognition in human-robot interaction. The most popular combination of modalities in these works is the use of speech commands together with gestures (Mollaret et al., 2016; Rodomagoulakis et al., 2016; Stiefelhagen et al., 2004; Vaufreydaz et al., 2016; Zlatintsi et al., 2018), sometimes additionally combined with head poses (Mollaret et al., 2016; Stiefelhagen et al., 2004) or otherwise face information and movement speed of the respective person (Vaufreydaz et al., 2016). Gaze was considered as a modality for intention recognition in two previous approaches, used either in combination with body pose (Yu et al., 2015) or with speech and button presses (Bannat et al., 2009). Scene objects were only considered in relation to the human, not as passive parts of the scene (Dutta & Zielinska, 2018; Kelley, Browne, et al., 2012; Kelley, Tavakkoli, et al., 2012). They were additionally combined with body poses (Dutta & Zielinska, 2018) or with the considered objects' states (Kelley, Browne, et al., 2012; Kelley, Tavakkoli, et al., 2012). Two further proposed systems that considered completely different modalities are the work of W. Xu et al. (2015), which combined force sensor and laser rangefinder data for recognizing motion intentions, and the work of Kulic and Croft (2003), which employed multiple physiological signals such as blood volume pressure for the automatic recognition of human approval of a robot's actions. As can be seen, there are several systems that consider multiple modalities for intention recognition. However, none of the works has combined the four modalities speech, gestures, gaze directions, and scene objects as it is done in this work.

Among all the approaches discussed so far, there are some that not only deal with intention recognition in human-robot interaction but especially address elderly assistance, which is of particular interest in this work. Vaufreydaz et al. (2016) worked on the automatic detection of the intention to interact with a robot. Considered data were face size and position, speech, shoulder pose rotation, and movement speed. These data were concatenated to form a single feature vector which was classified either by a Support Vector Machine or a Neural Network. Thus, feature fusion was conducted instead of classifier fusion, which would fuse the outputs of individual classifiers. Mollaret et al. (2016) also dealt with the recognition of an intention for interaction with an assistive robot. Using head and shoulder orientation and voice activity the corresponding intention could be inferred with a Hidden Markov Model. Here, raw data instead of features were used for fusion but again not the outputs of individual classifiers.

W. Xu et al. (2015) proposed a walking-aid robot and in this context focused on recognizing human intentions in terms of intended walking velocities. Data were captured from force sensors and a laser rangefinder and the estimated velocities were fused with a Kalman filter. Even though here, outputs of individual estimators were fused, the study only dealt with a continuous size instead of category labels.

Rodomagoulakis et al. (2016), on the other hand, fused outputs of classifiers working on discrete categories. In order to enable people with limited mobility to interact with a robotic rollator, they recognized intentions considering speech and gestures. The respective classifiers' output scores for all possible intentions were fused by a weighted linear combination with tunable weights while the intention with the highest fused score was predicted. Another work operating on classifier outputs for fusion by Zlatintsi et al. (2018) proposed an intention recognition system for an assistive bathing robot based on speech and gestures. They applied a late fusion scheme meaning that an intention was chosen as the detected one if it was ranked highest by the speech classifier and among the two highest ranked intentions according to observed gestures. Although these two works fuse discrete classifiers' outputs, these outputs are not treated probabilistically, so uncertainty reduction is not possible.

In fact, all approaches regarded so far perform modality fusion for intention recognition. Some even do so by fusing outputs of individual classifiers. However, none of them considers uncertainty for fusion or attempts to reduce the uncertainty of the final decision. Instead, they are exclusively concerned with improving the system's accuracy and robustness. This is also the case for works about multimodal intention recognition in other contexts than elderly assistance (Yu et al., 2015).

Some specialized approaches for audio-visual speech recognition (Gurban & Thiran, 2008; H. Liu et al., 2014) considered uncertainty by performing uncertainty-based weighting for the fusion of multiple classifiers' outputs. In these works, the respective two categorical probability distributions returned by two individual classifiers for audio and visual input were combined by a weighted sum. The respective weights were computed from the individual distributions' uncertainties, quantified e.g., with entropy (Gurban & Thiran, 2008). Consequently, the more uncertain distribution got the lower weight and by this had a lower influence on the fused distribution. Whereas it is desirable to consider uncertainty for determining the individual distributions' impact on the fusion result, a weighted sum of categorical distributions does not necessarily reduce uncertainty, because it results in an average of the distributions which is less or equally certain by definition. However, uncertainty reduction is one of the biggest advantages of fusing different classifiers' distributions (Nweke et al., 2019).

Consequently, another method is needed that not only determines each distribution's impact on the fused resulting distribution based on its uncertainty but also reduces the uncertainty of this fused distribution. A suitable method that meets this requirement is the Bayesian fusion method Independent Opinion Pool (Andriamahefa, 2017; Berger, 1985), which we use in our approach (Section 5.2.1). It basically multiplies the individual probability distributions for fusion. The method has already been applied for different fusion tasks, among them the fusion of geological data from different measurement locations (Elsaesser, 2007), of laser rangefinder data for semantic labeling of places (L. Shi et al., 2010), and of camera data for robust robot navigation (Stepan et al., 2005). However, Independent Opinion Pool has not been used for multimodal intention recognition so far in order to explicitly reduce uncertainty. All in all, to the best of our knowledge there is no work that uses the four modalities we consider for intention recognition together with a method that targets uncertainty reduction.

5.2 Multimodal Intention Recognition

In this work, an approach for multimodal intention recognition is introduced which focuses on reducing the uncertainty about the intention to be recognized. As intentions we define atomic action intents for predefined tasks such as the intention for a handover of an object. Since we represent these intentions as discrete categories, recognizing them can be seen as a classification problem. Our approach applies classifier fusion which fuses the outputs of individual and independent base classifiers instead of e.g., fusing directly the raw data or respective feature vectors. For this reason, for each of the four considered modalities, a classifier was trained on its own data from the respective modality. Each individual classifier returns a categorical probability distribution over all possible intentions as output, which contains a probability for each possible intention. All of these base classifiers could perform intention recognition on their own. However, their output distributions are fused in order to decrease uncertainty and improve performance. An overview of the proposed approach is shown in Figure 5.1.

5.2.1 Classifier Fusion with Independent Opinion Pool

Our principal motivation for combining multiple modalities is uncertainty reduction. First, the fusion of two non-conflicting distributions, in the most extreme case two equal distributions, should result in a fused distribution with a lower entropy than those of the respective base distributions. Second, the uncertainty of each base distribution, e.g., in terms of its entropy, should determine the influence of the distribution on the fused result in a way that an uncertain distribution's influence is lower.

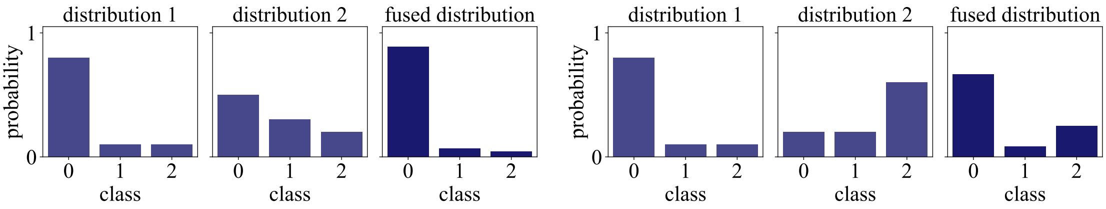
In order to achieve uncertainty reduction as it is described above, we apply the Bayesian fusion method Independent Opinion Pool (Berger, 1985) for fusion of the n categorical probability distributions $p(y|x_i)$ over intentions y given modality data x_i. This method assumes that the base classifiers' modality data x_i and with them their output distributions are conditionally independent given the true class label, which is the true underlying intention y in our case. Furthermore, a uniform distribution over all possible classes $p(y)$ is assumed a priori. By applying Bayes' rule with these assumptions, the fusion can be conducted by simply multiplying the underlying base distributions returned by each classifier and renormalizing the resulting categorical distribution so that it sums to 1,

$$p(y|x_1, \ldots, x_n) \propto \frac{\prod_{i=1}^{n} p(y|x_i)}{p(y)^{n-1}} \overset{\substack{\text{uniform} \\ \text{prior}}}{\propto} \prod_{i=1}^{n} p(y|x_i). \tag{5.1}$$

Given the two assumptions conditional independence and an uninformed prior, fusing classifiers in this way is Bayes optimal, meaning that the resulting posterior fused distribution is the correct fusion result.

Two advantages of Independent Opinion Pool for fusion are reinforcement and mitigation (Andriamahefa, 2017). Reinforcement describes the first criterion for uncertainty reduction we set in the previous paragraph. In case the distributions returned by the individual base classifiers are non-conflicting and thus predict the same class, the uncertainty and with it the entropy of the resulting fused distribution is reduced compared to those of the base distributions (Figure 5.2(a)). Mitigation means that conflicting base distributions cause a fused distribution with a higher uncertainty (Figure 5.2(b)). This might seem to be a contradiction to this work's goal of uncertainty reduction but is indeed desirable as in case that cues from different modalities are conflicting the resulting fused distribution should reflect this.

The second criterion for uncertainty reduction is also accomplished. Using Independent Opinion Pool for fusion, each base distribution's uncertainty determines its impact on the fusion result. The fusion impact of an uncertain base distribution is lower than that of a more certain one (Figure 5.2(b)). Hereby, the fusion impact of each base distribution is only dependent on its current uncertainty and is thus recomputed online for every new multimodal data example.



(a) (b)

Figure 5.2: Examples of classifier fusion with Independent Opinion Pool (Berger, 1985). When used for classifier fusion, Independent Opinion Pool leads to reinforcement, which means that two non-conflicting distributions result in a more certain fused distribution (a). It also leads to mitigation, meaning that the fusion of two conflicting distributions causes an increased uncertainty. Meanwhile, the more certain distribution has a higher impact on the fused distribution and is thus decisive (b).

5.2.2 Classifiers for Single Modalities

The individual classification methods we use for intention recognition from the four modalities speech, gestures, gaze directions, and scene objects are presented in the following. However, it is not the focus of this work to develop new methods for classifying intentions from the different single modalities' data. Instead, the power of multimodality for uncertainty reduction is demonstrated. Thus, the base classifiers for the considered modalities are designed as simple as possible and build upon existing methods. In fact, they can be easily replaced by any other classifiers which output categorical distributions, and additional classifiers might be added sequentially in order to further improve the system.

5.2.2.1 Speech

Speech is a meaningful modality for intention recognition as it is effortless and intuitive for humans (X. Liu et al., 2016). For intention recognition from speech, keyword spotting is applied. This enables the recognition of simple keywords that are related to intentions while simultaneously allowing humans to flexibly formulate the command sentences. Since a probability distribution over all keywords is required as output, which in many popular frameworks is not available, an open-source framework called Honk (Tang & Lin, 2017) is used. Honk builds upon a Convolutional Neural Network proposed in the work of Sainath and Parada (2015) that consists of two convolutional layers and one final softmax layer. As input for the network, Mel-Frequency Cepstrum Coefficient features are used. Its implementation is realized with PyTorch.

For training, we recorded keyword utterances from 16 people (8 female, 8 male), while each keyword, e.g., "bowl" for the intention to get a bowl, was repeated ten times. In addition to recordings of the keywords of interest also eleven other words were recorded that are likely to be part of possible command sentences in order to reduce false alarms. 15% of the training data were taken from these unknown words, another 15% were taken from example recordings of noise sounds. To increase robustness, with a probability of 0.8 these noise sounds were also added to the training examples. 80% of all data were taken for training and 10% each for testing and validation.

Since the network is trained on single keywords but keywords need to be detected within complete command sentences, multiple probability distributions are obtained for one query sentence. To combine them, for each intention the maximal probability value in all distributions is taken to constitute the final probability distribution. This is motivated by the assumption that each sentence contains only one keyword to which the highest probability should be assigned.

The device for recording speech is a USB microphone (Klim) that captures mono sound with a sample rate of 16000 Hz.

5.2.2.2 Gestures

Since a majority of human communication is nonverbal, gestures provide valuable information about intentions, especially when referring to objects (Canal et al., 2015). Here, we realize intention recognition from gestures based on the method Mixture of Interaction Primitives (Ewerton et al., 2015). In this method, gesture trajectories are represented with Probabilistic Movement Primitives (ProMPs) (Paraschos et al., 2013) which approximate each trajectory position by a linear combination of Gaussian basis functions and weights $\boldsymbol{w}$. Using this representation one can learn a probability distribution over multiple demonstrated trajectories. Mixture of Interaction Primitives (Ewerton et al., 2015) extends ProMPs to be usable with multiple gestures and two interacting agents, e.g., a human and a robot. For this, a Gaussian Mixture Model (GMM) over human and robot trajectories represented as ProMPs is learned, in which each mixture component represents one interaction pattern between human and robot. In addition to inferring a learned gesture from an observed human trajectory, the method is also able to estimate the most likely response trajectory of the robot conditioned on an observed human trajectory. By this, the robot's movement can be adapted to the actually shown human gesture, e.g., with respect to a common end position in a handover task. For more details on the used approach the interested reader is referred to (Ewerton et al., 2015).

We apply Mixture of Interaction Primitives for intention recognition by representing gestures together with corresponding robot reaction trajectories as ProMPs and learning a respective GMM with one mixture component for every intention. Thus, in addition to recognizing intentions from human gestures, we can also generate a corresponding robot trajectory as a reaction to the intention recognized by the fused classifier considering all four modalities.

The parameters of the GMM are originally learned from unlabeled data with the Expectation Maximization algorithm (Ewerton et al., 2015). However, as we work with labeled data, we estimate the GMM's parameters with Maximum Likelihood Estimation. In addition, just the last point of the observed trajectory is taken for gesture classification which is sufficient for differentiating the gestures in our experiment. For training of the gesture classifier, 30 examples of reaching motions were demonstrated by one human subject. The resulting trajectories were captured with the motion tracking system Optitrack which uses cameras and passive reflective markers attached to the human wrist. For training of the robot reaction movements, kinesthetic teaching was applied.

5.2.2.3 Gaze Directions

Previous studies revealed fixations to be strongly task-dependent and predictive for future actions and intentions (Admoni & Srinivasa, 2016; C.-M. Huang et al., 2015; Mennie et al., 2007; Rothkopf et al., 2007), which motivates us to infer human intentions from gaze directions. For this, we apply a Support Vector Machine (SVM) that is inspired by two existing approaches about intention recognition from gaze (Admoni & Srinivasa, 2016; C.-M. Huang et al., 2015). Considered features are the distances between the human's 3D gaze vector and all locations of interest in the scene, which are mainly object locations. The mean gaze vector is computed from the last 900 samples of the recorded gaze directions during a trial. The motivation for working with this mean is that it indirectly includes information about the most recently fixated location and the number and duration of fixations towards this location, which were all stated to be important features for intention recognition in the work of C.-M. Huang et al. (2015). Additionally considering the distances between locations of interest and the gaze vector is an idea presented in the work of Admoni and Srinivasa (2016). Using the described features, a multiclass SVM with linear kernel was trained using the package sklearn. In contrast to just the predicted class label, which is the usual output of an SVM, this package also provides a probability distribution over all possible labels as output.

The used training data included 30 labeled examples per intention, each with a duration of five seconds with sample rate 250Hz. For recording, one person was seated in the evaluation scenario (Section 5.3) and asked for several robot assists, e.g., an object handover, while shifting its gaze towards the location of interest, i.e., the object itself. Gaze direction is recorded by a monocular head-mounted eye tracker (Pupil Labs) which uses infrared lights and eye cameras for inferring the gaze vector. The device is additionally equipped with reflective markers in order to be trackable by the Optitrack system also used for gesture tracking, because we need the gaze vector in scene coordinates rather than just related to the eye tracker itself. For integration of the eye tracker in the overall system an open-source ROS plugin is used (Qian, 2016).

5.2.2.4 Scene Objects

Scene objects are objects that are passive parts of the scene but can still give hints about possible human intentions (P. Bach et al., 2014). For estimating intentions from such objects we build upon an approach that deals with scene classification from observed objects (Luo & Xu, 2016). The reason for this choice is that other approaches which directly address intention recognition only consider objects manipulated by a human (Dutta & Zielinska, 2018; Kelley, Browne, et al., 2012) rather than passive scene objects. Thus, an SVM was chosen as classifier with input feature vectors containing the horizontal distances of all available scene objects to a pre-defined center point on a working area in front of the human. Objects that are positioned outside this working area are set to the same pre-defined value. Other possible features, such as the raw positions of objects or just Boolean values indicating whether an object is in the working area or not, performed worse than the chosen approach. The multiclass SVM is again implemented using the package sklearn and again a probability distribution over all intentions is returned instead of just one predicted intention.

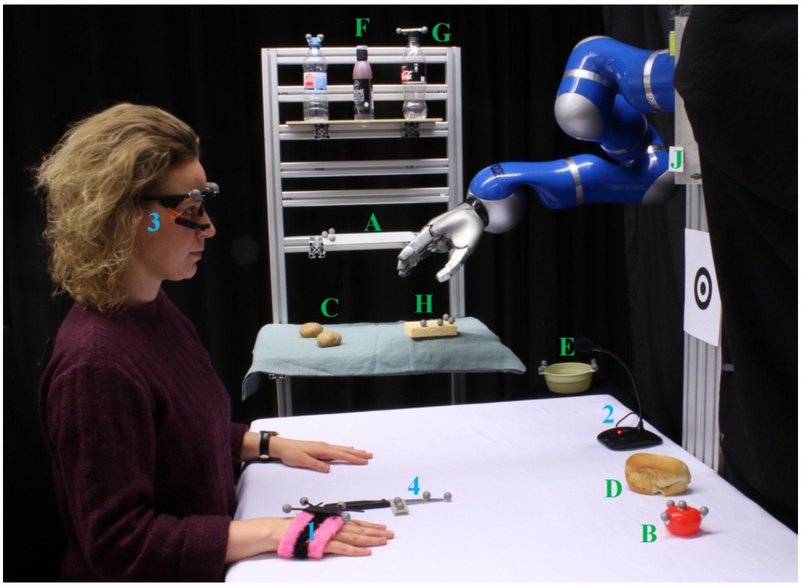


Figure 5.3: The experiment setup in which the fusion system was evaluated. In a kitchen scenario, the human can request different actions from the robot by displaying intentions through the four considered modalities. For capturing gestures and scene objects the human's hand (1) and the scene objects (e.g., 4) are equipped with markers. Speech is captured with a microphone (2) and gaze with a head-mounted eye tracker (3). The ten recognizable intentions are the handover of the board (A), tomato (B), potato (C), roll (D), bowl (E), dressing (F), coke (G), or towel (H). Additionally, there is the intention to stand up for which location (J) is fixated.

Training was conducted on 50 recordings of different scene object placements for each intention, e.g., a glass for the intention to get some coke. Object positions are gathered with the motion tracking system Optitrack. For this, each object is equipped with four markers in a unique geometric pattern that enables the system to distinguish between objects. Used scene objects in our interaction scenario are a cutting board, a tomato, a bowl, a bottle of each coke and water, a sponge, a glass, and a knife.

5.3 Experimental Evaluation

For evaluation of the proposed multimodal intention recognition system we chose a kitchen scenario in which a 7-DoF robot arm assists a human in preparing some food, e.g., a salad (Figure 5.3). The human sits at a table with several task-relevant objects placed around him that are not easily reachable from a seated position. The robot can assist by handing over requested objects or helping to stand up by reaching out its arm as a prop. The nine recognizable intentions are to receive a cutting board, a tomato, a potato, a roll, a bowl, a bottle of dressing or coke, or a towel, and to get support for standing up. These intentions are deliberately chosen to be ambiguous, e.g., with respect to their positioning or sound, in order to obtain uncertain results when using the individual base classifiers only. For all intentions ten example data sets were recorded, each consisting of a speech recording, the last point of the shown hand gesture, a list of the shown gaze directions and the positions of the scene objects of interest. Consequently, 90 multimodal examples from one human subject are available for evaluating the proposed intention recognition system.

Two measures are chosen to quantify the uncertainty of a categorical distribution. First, Shannon entropy is used, which is a well-known measure for the uncertainty entailed in a distribution. It is maximal for a maximally uncertain distribution, that is uniform over all intentions, and minimal for a maximally certain distribution that assigns all probability mass to one intention. Second, a measure called score difference (Potamianos & Neti, 2000) is applied for measuring uncertainty. It computes the difference between the two highest probabilities in the distribution. Thus, unlike entropy, score difference does not consider the complete distribution's uncertainty but quantifies the actual uncertainty in making a hard decision for one intention. Thereby, a low score difference indicates a high uncertainty.

Although uncertainty reduction is the focus of our approach, it is also essential to guarantee correct classification. Thus, in addition to entropy and score difference also the accuracy of the multimodal and unimodal classifiers is evaluated. Figure 5.4 shows accuracy, entropy, and score difference of the underlying fused distributions for all possible combinations of base classifiers as well as for the four single base classifiers. As can be seen, the accuracy of the fused result combining all modalities (0.94) is higher than that of the single classifiers (speech: 0.63, gesture: 0.86, gaze: 0.59, objects: 0.79). Only one other combination of base classifiers, namely gesture-gaze-objects, has a slightly higher accuracy. In general, eight out of eleven combinations of base classifiers result in a higher accuracy than all of the base classifiers do. This already indicates the superiority of multimodal over unimodal intention recognition.

If additionally considering uncertainty, both entropy and score difference show the lowest mean uncertainty for the fusion of all four modalities compared to all other possible combinations. Two other combinations of base classifiers, namely speech-gesture-objects and gesture-gaze-objects, show a similarly low uncertainty which is nevertheless higher than that resulting from fusing all modalities. In general, it can be seen that except from one classifier combination the uncertainties of the fused distributions are considerably lower than that of all single base classifiers. The only exception is the fusion of speech and gaze classifiers, which is slightly more uncertain than the most certain base classifier, the gesture classifier. Yet, its uncertainty is reduced in comparison to the two actually fused individual classifiers for speech and gaze. Note that these two classifiers are the least accurate and most uncertain of all four classifiers. This demonstrates the power of multimodal classifier fusion for intention recognition as proposed here. Even inaccurate and uncertain classifiers such as speech and gaze classifiers contribute to uncertainty reduction and better performance if added to a multimodal intention recognition system. Moreover, also combinations of less than all four base classifiers already improve performance and reduce uncertainty.

To further quantify the reduction in uncertainty through fusion of multiple modalities we compared our approach to the weighted sum approach. In this alternative method motivated by Rodomagoulakis et al. (2016) and Gurban and Thiran (2008), each distribution's weight is its inverse entropy. Thus, the higher a distribution's uncertainty, the lower is its impact on the fused result. Figure 5.5 shows the entropies of all possible combinations of base classifiers when fused with a weighted sum approach instead of Independent Opinion Pool. Although uncertainty is considered for fusion, there is no reduction of uncertainty compared to the base classifiers. In addition, the fused result combining all four classifiers is much more uncertain than with the proposed Independent Opinion Pool approach.

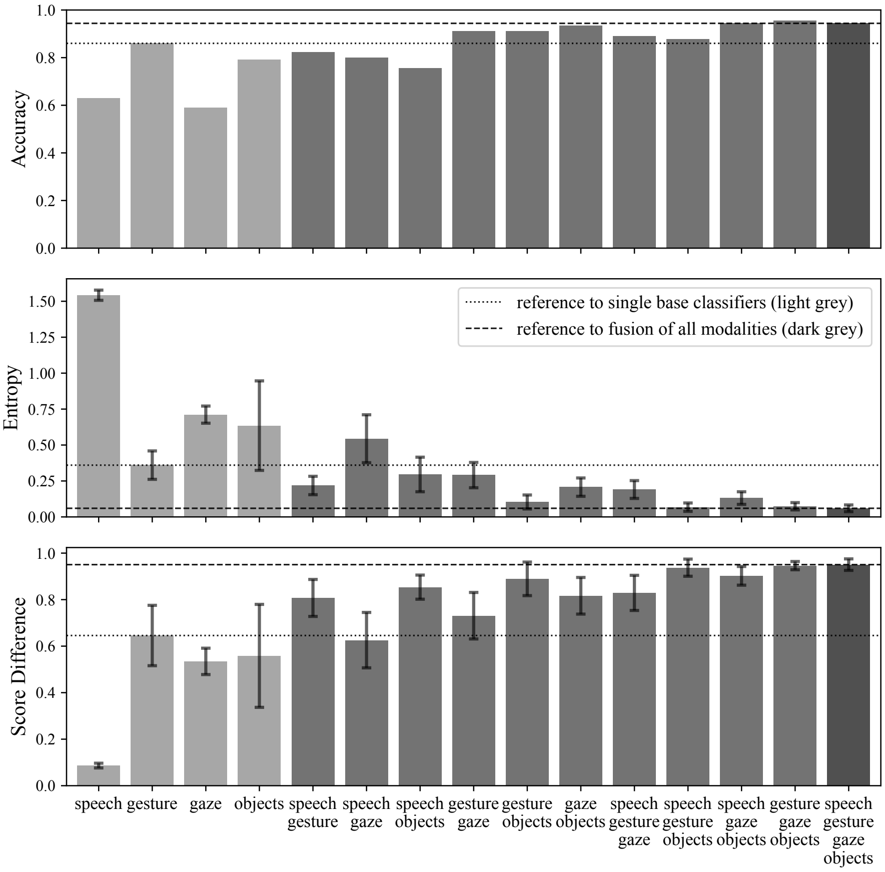


Figure 5.4: Comparison of all possible combinations of base classifiers regarding accuracy, mean entropy, and score difference. Corresponding variances are plotted as error bars. It is seen that uncertainty is reduced through classifier fusion, in particular it is lowest for the fusion of all four modalities, according to both measures entropy and score difference. Accuracy is also increased through fusion compared to the single classifiers.

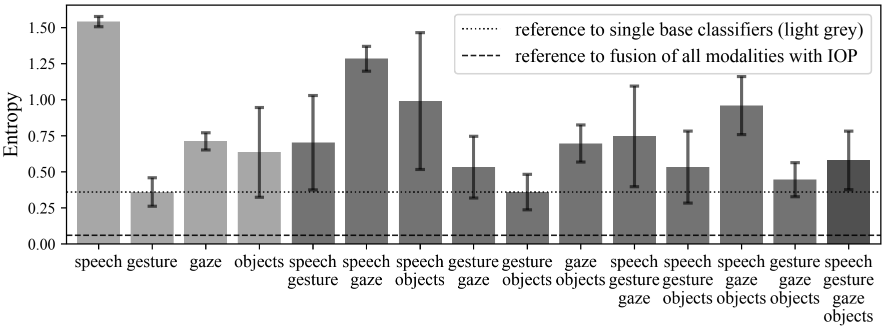


Figure 5.5: Mean entropies resulting from fusing all different combinations of base classifiers with a weighted sum. In comparison to fusing with Independent Opinion Pool (Figure 5.4), the weighted sum approach reduces less uncertainty.

So far, the results imply that the proposed multimodal approach for intention recognition reduces the uncertainty of the overall system, even better than the weighted sum approach. However, just means and variances of entropy and score difference over all test examples were taken into consideration. We additionally need to analyze whether uncertainty reduction is also accomplished for individual fusion examples. For speech and gesture classifiers, Figure 5.6 shows the uncertainties in terms of entropy of the generated categorical distributions of all 90 test examples, differentiated according to whether they are generated by just the single classifier or by fusion of multiple classifiers.

One can see that for the most uncertain classifier, the speech classifier, the fused distributions are always less uncertain than the single base distribution for all recorded test examples, no matter how many of the three other modalities are added for fusion. In particular, the entropy of the distribution fused from all four modalities is lowest for nearly all examples. The only exceptions are examples from the intentions tomato and roll. Yet, this is easily explainable as these two intentions are often confounded by the base classifiers, which leads to conflicting base distributions. This in turn results in a higher uncertainty of the fused distribution which is desirable as different opinions of base classifiers should be reflected in the resulting fused distribution.

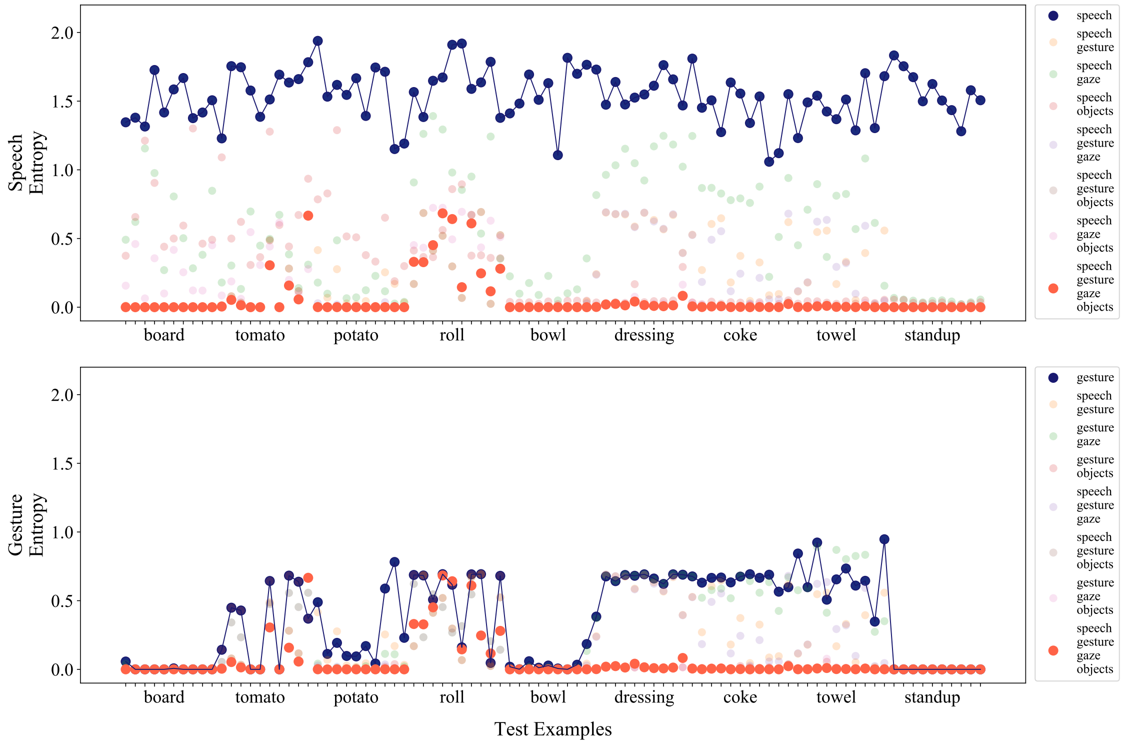


Figure 5.6: Entropies of all distributions resulting from classifier combinations including the speech (top) or gesture (bottom) classifier over all 90 test examples of the nine intentions. We see that for a large majority of examples already combinations of two or three classifiers reduce the uncertainty compared to the uncertainty of the single speech or gesture classifier. The fusion of all four modalities causes the strongest uncertainty reduction. The intentions tomato and roll show higher uncertainties since the base classifiers often confound them.

For the most certain one of the base classifiers, the gesture classifier, similar results can be shown, but not as strong as for the speech classifier. The gesture classifier is already quite certain on its own, which is demonstrated by the much smaller overall entropy values. Apart from some exceptions, again, the fused distributions from all possible combinations with the other three modalities are less uncertain than the single gesture classifier, and in the majority of cases the fusion of all four modalities results in the lowest entropies near to 0. In contrast to the speech classifier seen before, for the gesture classifier there are some examples with higher entropies for the distribution resulting from the fusion of four modalities compared to the single classifier's distribution, but all these examples repeatedly come from the two ambiguous intentions tomato and roll.

These cases are especially interesting as the examples that are classified incorrectly by the fused distribution combining all possible modalities mostly are examples of the intention roll. Consequently, the intention with the most uncertain fused distributions is also the intention with the most incorrect classifications. As an uncertain misclassification is more desirable than a certain one, this is desirable behavior.

Our proposed intention recognition approach was not just evaluated quantitatively on recorded multimodal data but also online in a real interaction with the 7-DoF robot arm. For this, the kitchen task was performed cooperatively by a human and a robot. This means that in order to prepare a salad the human expresses the different intentions using the four modalities and after having recognized the correct intention the robot reacts accordingly. As a reaction, the robot moves to the respective location for this intention, e.g., the position of a requested object, and grasps it. Subsequently, it executes a trajectory in order to hand over the respective object or help the human to stand up. This trajectory was learned from demonstrations and is conditioned on the last point of the shown human movement as was explained in Section 5.2.2.2. This means that the most likely robot movement given the last point of the observed human trajectory is executed, which leads to an adaptation of the robot movement to the human. This is especially beneficial for our handover tasks. The complete interaction process is shown exemplarily in Figure 5.7 for the intention to get a roll. The human expresses this intention by uttering a command containing the word "roll", reaching out its arm in the roll's direction, fixating it and having placed the scene objects board and knife in the working area. The robot recognizes the correct intention and subsequently moves towards the roll, grasps it and hands it over to the human by executing the inferred trajectory. As well as on prerecorded data, the proposed intention recognition system using speech, gestures, gaze directions, and scene

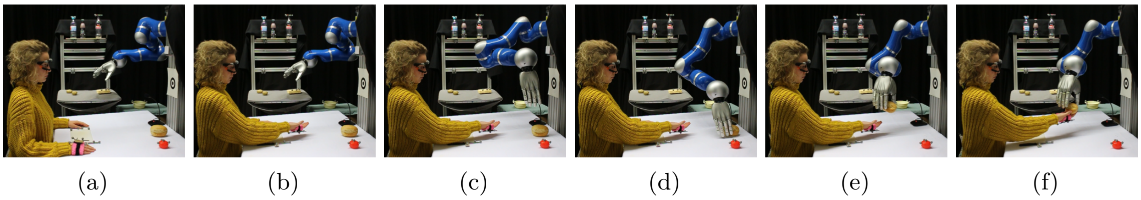

(a) (b) (c) (d) (e) (f)

Figure 5.7: An exemplary interaction between human and robot. The human shows the intention roll by uttering a command containing the word "roll", reaching out its arm for the roll, fixating the roll and having placed the scene objects board and knife in the working area (b). As soon as this intention is recognized, the robot moves towards the roll (c), grasps it (d), and subsequently executes the inferred trajectory (e) to hand it over to the human (f).

objects was shown to work well also in online interaction with a real robot for all considered intentions.

5.4 Conclusions

In this work, we introduce a multimodal approach for intention recognition to be applicable in elderly assistance. In contrast to existing works, we focus on uncertainty reduction in a way that the combination of modalities makes the system more certain about the intention to be recognized. For this, the categorical output distributions of individual classifiers for the four different modalities speech, gestures, gaze directions, and scene objects are fused using the Bayesian fusion method Independent Opinion Pool. We evaluate our approach in a cooperative kitchen task between a human and a 7-DoF robot arm. The results show that uncertainty can be reduced through the use of multiple modalities. Even very inaccurate and uncertain classifiers can contribute to uncertainty reduction, better performance, and robustness when added to a multimodal system.

This shows that the proposed approach allows well-performing and more certain intention recognition using simple and easily trained base classifiers that only require little training data. Additionally, it is particularly important for elderly assistance since even if complex classifiers are available they might be challenged by data from elderly people, which can increase their uncertainty and error rate.

Although the proposed intention recognition system considers the uncertainty of the individual modalities' classifiers, it is limited to the information provided by the classifiers' current output distributions and has no further knowledge of the classifiers' general performance, e.g., their bias, variance, and uncertainty. Therefore, an interesting line for future work is to extend the proposed approach to also consider the individual classifiers' general behavior for fusion, which can be learned from training data if available. Also, the assumption that the individual modalities' classifiers are conditionally independent given the true intention should be investigated carefully in future work. Another limitation is that although the uncertainty over the intention to be recognized is explicitly represented and reduced in a Bayes optimal way, it is not exploited for the robot's reaction. The intention with the highest probability is recognized and the corresponding action is executed accordingly. However, a reaction according to the actual uncertainty of the recognition could further improve the quality of interaction between human and robot. Therefore, for future work we plan to exploit the knowledge of the decision uncertainty in a way that the robot reacts according to its uncertainty about the situation.

6 INTERACTIVE REINFORCEMENT LEARNING WITH BAYESIAN FUSION OF MULTIMODAL ADVICE

Classical industrial robots are typically designed to perform very specific and mostly repetitive tasks. In contrast, future assistive robots, which are intended to support humans in their daily lives, will be challenged by a multitude of different tasks. Since usually not all of these tasks may be known explicitly beforehand, a key component for such robots is the ability for self-improvement at runtime and adaptation to human preferences and new situations at hand.

Even though Reinforcement Learning (RL) (Kormushev et al., 2013; Sutton & Barto, 2018) offers a powerful methodology for robots to learn from direct interaction with their environment, in many practical robotic applications large state and action spaces as well as costly sample collection prevent the use of RL algorithms. This is where the novel research field of interactive RL (IRL) (Knox & Stone, 2008; Thomaz et al., 2005) aims to improve learning speed and convergence of RL algorithms by incorporating human feedback (Knox & Stone, 2008) or advice (Cruz et al., 2015) into the learning process.

To facilitate a beneficial interaction of everyday users with such IRL systems it is particularly important to enable ways for more natural and intuitive communication of human advice during the learning process (Lin et al., 2020). Since humans are used to teaching other humans using natural cues such as speech, gestures, body language, gaze, or facial expressions (Song et al., 2012), it is a central question how to best integrate such natural interaction channels into IRL algorithms. In particular, as we showed in Chapter 5 for human intention recognition, exploiting all available multimodal data can in general increase a decision's accuracy and decrease its uncertainty (Trick et al., 2019, see Chapter 5).

Accordingly, Cruz et al. (2018) introduced an IRL approach (termed C-IRL hereafter) which allows humans to give advice using the modalities speech and gestures. For C-IRL the authors trained an individual probabilistic classifier for each of the two advice modalities and then fused the resulting output distributions. The used fusion method reduces the decision's uncertainty if both modalities' classifiers detect non-conflicting advice and increases the uncertainty otherwise.

However, C-IRL only considers the confidence values of the predicted most likely class label and only utilizes probabilities above a certain threshold, thereby discarding valuable information of the single base classifiers' distributions. Additionally, in C-IRL the computation of the fused confidences is not theoretically founded but based on a heuristic tailored to exactly two modalities, and it is not discussed how to extend this computation to more modalities.

The Bayesian fusion method Independent Opinion Pool (Andriamahefa, 2017; Berger, 1985), which we derived in Chapter 3, can overcome these limitations. While we successfully applied it for human intention recognition in Chapter 5, Independent Opinion Pool

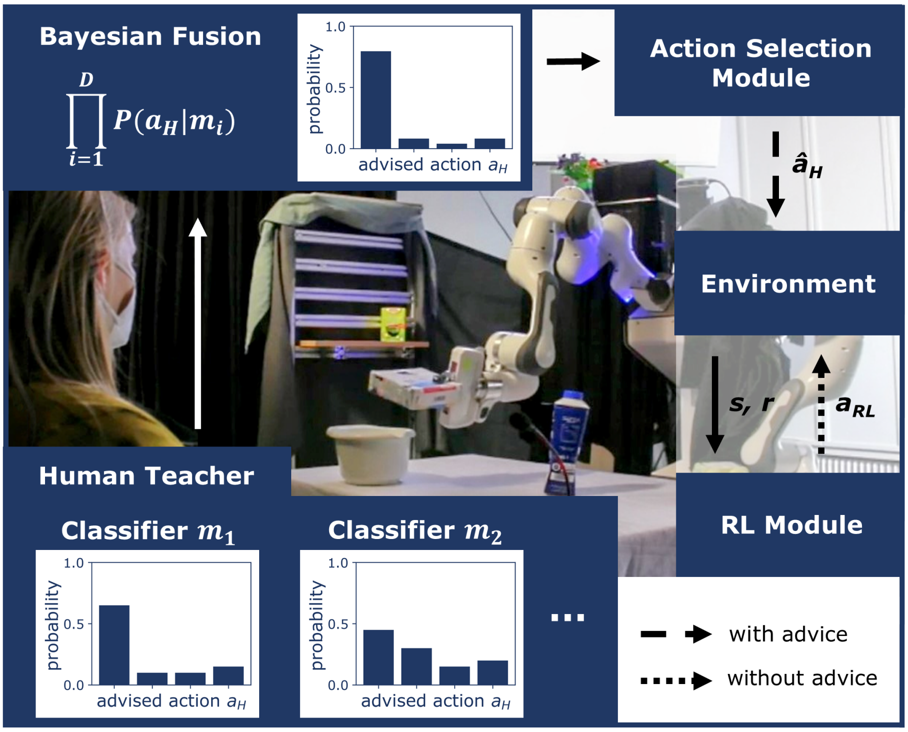


Figure 6.1: An overview of the proposed approach Multimodal IOP-Based Advice for Interactive Reinforcement Learning (MIA-IRL). MIA-IRL uses the Bayesian method Independent Opinion Pool (IOP) to combine the output distributions of the single modalities' base classifiers m_i. From the fused distribution we sample an estimated human action advice $\hat{a}_H$ to execute on the robot. When no human advice is given we use the action a_{RL} suggested by the base policy of our RL-Module.

can also be used to fuse action advice from different modalities in order to improve multimodal IRL. Accordingly, the main contribution of this work is a new multimodal IRL algorithm that uses Independent Opinion Pool (IOP) for combining the advice modalities' classifiers' output distributions (Figure 6.1). As IOP combines individual classifiers' output distributions Bayes optimally, it reduces uncertainty correctly. Additionally, the proposed method allows straightforward generalization to more than two modalities, which is not clear in C-IRL. Because our method takes advantage of all available information of the base classifiers' distributions and computes the Bayes optimal uncertainty in the fused distribution, the action selection can be done probabilistically instead of just executing the most likely action. We evaluate our method in direct comparison to C-IRL in a simulated grid world scenario and on a real-world human-robot interaction (HRI) task, in which human participants teach a 7-DoF robot arm. The experimental evaluations show that our method clearly outperforms C-IRL, particularly in the case of partially wrong outputs of the modalities' base classifiers. Thus, we show that Bayesian fusion of modalities increases the robustness of multimodal IRL.

The rest of the chapter is structured as follows. In Section 6.1 we discuss related work. Section 6.2 introduces our novel IRL approach using Bayesian fusion of multiple input modalities. In Section 6.3 we present the experimental evaluation on theoretical corner cases, in a simulated grid world, and in a real HRI scenario. Finally, we summarize our findings and discuss future research directions in Section 6.4.

6.1 Related Work

Traditionally, Interactive Reinforcement Learning allows a human trainer to give feedback on the action a robot just performed (Blumberg et al., 2002; Kaplan et al., 2002; Knox & Stone, 2008). In contrast to this feedback-driven approach, humans also try to guide the robot on future actions by giving advice (Thomaz et al., 2005). Accordingly, several IRL approaches were proposed that include human advice instead of or in addition to feedback (Cruz et al., 2015; Knox et al., 2013; Koert et al., 2020; Kuhlmann et al., 2004; Maclin & Shavlik, 1996; Thomaz et al., 2006). However, in many approaches the human advisors are not able to communicate their advice over natural interaction channels. In the work of Maclin and Shavlik (1996) the human teacher needs to use a specific programming language to interact with the learning agent. Thomaz et al. (2006) proposed a computer mouse as input device for human advice, while Knox et al. (2013) instead uses a remote control. Koert et al. (2020) chose a graphical user interface provided on a tablet computer as input modality for advice.

More intuitive modalities for interacting with the learning agent were proposed by Kuhlmann et al. (2004) and Cruz et al. (2015), who used speech as input source, or Veeriah et al. (2016) and Gordon et al. (2016), who used facial feedback. However, humans use multiple modalities to express their intentions (Schrempf & Hanebeck, 2005) and also their advice (Cruz et al., 2018). Accordingly, several approaches exploit multimodal input data for IRL (Cruz et al., 2016; Cruz et al., 2018; Leite et al., 2011; Qureshi et al., 2016; Weber, Ritschel, Aslan, et al., 2018; Weber, Ritschel, Lingenfelser, et al., 2018). In order to teach an empathic chess partner for children, Leite et al. (2011) combine human facial features with task-related features, e.g., if the human is winning or losing. The modalities are fused at the feature level, which, however, impedes generalization by exchanging or adding modalities.

In contrast, Qureshi et al. (2016) propose combining the data from depth and grayscale images for a robot to learn social behavior. For both modalities, two individual Q-functions are learned, which are averaged for fusion. Weber, Ritschel, Lingenfelser, et al. (2018) and Weber, Ritschel, Aslan, et al. (2018) combine facial and audio features in order to learn how to entertain people. The probability for laughing is computed individually from visual and audio cues, and the resulting probabilities are averaged for fusion. While these approaches can be straightforwardly generalized by exchanging the modalities or their respective classifiers, or by adding additional modalities, by averaging individual modalities' results, they cannot account for the uncertainty of the individual modalities' classifiers. For instance, a less certain modality has the same impact on the fused result as a more certain one and a decision's uncertainty cannot be reduced through fusion.

Cruz et al. (2018) also use multimodal input channels for IRL, however, they explicitly consider the individual modalities' uncertainties. In their framework C-IRL, a human teacher can give advice using the two modalities speech and gestures. For each modality a separate probabilistic classifier was trained, which outputs the predicted label of the detected advice and a corresponding confidence value. The individual classifiers' outputs are combined by a heuristic fusion rule that chooses the label with the higher confidence value if the classifiers are conflicting. Furthermore, they compute a fused confidence to decrease a decision's uncertainty in case both classifiers are non-conflicting and increase it otherwise (Cruz et al., 2016; Cruz et al., 2018).

Although this seems to be a reasonable fusion behavior, Cruz et al. (2018) do not provide any mathematical foundation for their fusion rule, it is not sufficiently motivated why one should use exactly this function for updating the fused confidence. Moreover, their method discards valuable information by only considering the confidence values of the most likely classes instead of entire probability distributions and by not utilizing probabilities below a manually set threshold. Additionally, their fusion method, in particular their function for updating the fused confidence, is explicitly designed for fusing two modalities and does not straightforwardly transfer to more modalities.

In contrast to Cruz et al. (2018), we propose to use a Bayesian fusion approach. Bayesian inference was already used for inferring reward functions in inverse reinforcement learning from successive feedback (Jeon et al., 2020), but not for fusing multimodal action advice for IRL. Here, we propose to use the Bayesian fusion method Independent Opinion Pool (IOP) (Andriamahefa, 2017; Berger, 1985). IOP provides uncertainty reduction for non-conflicting output distributions, is theoretically founded on Bayes' rule, exploits all classifier information by considering entire probability distributions, allows sampling from the fused distribution for action selection, and is applicable to an arbitrary number of additional modalities. IOP has already been successfully applied for multimodal human intention recognition (Trick et al., 2019, see Chapter 5), and in this work we leverage its advantages for multimodal IRL.

6.2 Multimodal IOP-Based Advice For Interactive Reinforcement Learning

In this work, we propose a new approach for Interactive Reinforcement Learning with multiple input modalities. Specifically, our novel method Multimodal IOP-Based Advice for Interactive Reinforcement Learning (MIA-IRL) uses the Bayesian fusion method Independent Opinion Pool (IOP) (Andriamahefa, 2017; Berger, 1985) to incorporate multiple probabilistic base classifiers' distributions over human advice into an RL algorithm. In this section, we explain the main components of our approach, which are also illustrated in Figure 6.1. We describe the agent's interaction with its environment as a Markov Decision Process (MDP) and as a core deploy a standard RL algorithm, such as Q-Learning, in our RL Module (Section 6.2.1). We then assume a human teacher that wants to communicate intended action advice a_H to suggest to the robot which action should be performed next. The human's action advice is recognized using multiple modalities. For each modality m_i an individual base classifier is trained, which is assumed to output a categorical distribution over all possible actions $p(a_H|m_i)$ (Section 6.2.2). Subsequently, the categorical distributions returned by all D base classifiers are fused within the Fusion Module using IOP (Section 6.2.3). By sampling from the fused categorical distribution $p(a_H|m_1, \ldots, m_D)$ we obtain an estimate for the action proposed by the human $\hat{a}_H$, which the RL agent then executes (Section 6.2.4). If no advice is given, the action proposed by the RL Module a_{RL} is chosen (Section 6.2.1). For the experiments in this work, human advice was provided in the first N episodes of learning. However, MIA-IRL could straightforwardly also incorporate distributed advice if an advisor is available over the complete learning process. Our MIA-IRL approach is summarized in Algorithm 6.1. An implementation of the proposed approach is publicly available at `https://github.com/RothkopfLab/MIA-IRL`.

Algorithm 6.1 MIA-IRL

```
Require: max number of steps per episode M
init $Q$-table $Q[s,a]$=0   $\forall s, a$ if $a$ possible in $s$, else $-\infty$
init visits per state $v[s] = 0 \quad \forall s$
init discount factor $\gamma$ and exploration rate $\varepsilon$
init episode counter $e = 0$
while $Q$ not converged do
    init steps per episode counter $j = 0$
    $s$ = random init state
    while episode not finished and $j <$ M do
        $v[s] = v[s] + 1$
        $\alpha = 1/v[s]$
        if human advice provided then
            for modalities $m_i = m_1, m_2, \ldots, m_D$ do
                $p(a_H|m_i) = \mathit{ModalityClassifier}(m_i)$
            end for
            $p(a_H|m_1, \ldots, m_D) = \mathit{FusionModule}(p(a_H|m_1), \ldots, p(a_H|m_D), Q[s])$
            $a$ = sample from distribution $p(a_H|m_1, \ldots, m_D)$
        else
            $a$ = choose $\varepsilon$-greedy action $a$ from $Q[s,a]$
        end if
        execute action $a$, get reward $r$ and next state $s'$
        $Q[s,a] = Q[s,a] + \alpha(r + \gamma \max_{a'} Q[s',a'] - Q[s,a])$
        $s = s'$
        $j = j + 1$
    end while
    $e = e + 1$
end while
```

6.2.1 RL Module

The learning agent's interaction with its environment is represented as a Markov Decision Process (MDP). Thus, in a state s it takes an action a, gets a reward r, and transits to the next state s'. The agent's goal is to learn an optimal policy $\pi(s)$ in order to receive the expected maximum discounted total future reward. For the experiments in this work, we used tabular Q-learning, which, however, could be replaced by other RL algorithms for different applications. The Q-function is updated according to

$$Q(s,a) \leftarrow Q(s,a)+\alpha(s)(r+\gamma \max_{a'} Q(s',a')-Q(s,a)) \tag{6.1}$$

and we chose a hand-tuned discount factor $\gamma = 0.98$ and an adaptive learning rate $\alpha(s) = 1/v(s)$, which is common in literature (Koert et al., 2020; Sutton & Barto, 2018), where $v(s)$ is the number of times the learning agent has visited state s so far. If no human advice is given, during learning the agent chooses actions according to an ε-greedy policy with ε set to 0.1 for our experiments.

6.2.2 Classifiers for Individual Modalities

For MIA-IRL we assume base classifiers for each modality m_i that output a categorical distribution $p(a_H|m_i)$ over all possible actions a_H. For the HRI experiments in this work, we exemplarily used two classifiers for the modalities speech and gestures. Since this work's focus is on demonstrating the benefits of applying the Bayesian fusion method IOP to IRL, these classifiers are based on off-the-shelf existing approaches. They can be straightforwardly replaced by other classifiers that return categorical distributions. In particular, adding more modalities is also possible from the mathematical formulations of the fusion method in MIA-IRL.

6.2.2.1 Speech

In our experiments, we chose speech as one of our modalities since it is mostly effortless and intuitive for humans to use for communicating their intentions (X. Liu et al., 2016). In particular, we use keyword spotting where each keyword is assigned to an action; e.g., "milk" is the keyword for getting some milk. We use the framework Honk (Tang & Lin, 2017), which returns a categorical distribution over all keywords, also including the categories "silence" and "unknown". Honk is based on a Convolutional Neural Network (CNN) with two convolutional layers, one softmax layer, and Mel-Cepstrum Coefficient features as input and is implemented in Pytorch. For training, we recorded 10 keyword utterances per word from 13 people. In addition to the 7 intended keywords (milk, flour, flower, bowl, roll, shelf, pour) we also recorded some unknown words, such as "please" or "give", that are likely to be used if people formulate their advice as a sentence. Also, noise and silence sounds were used for training. An amount of 20% of the training data for all keywords was added to the training set from the unknown words, correspondingly also 30% from the silence recordings. With a probability of 0.8, noise was added to training samples. 80% of all recorded data were taken for training, 10% each for testing and validation. Since in our experiments subjects were briefed to only use the defined keywords, before fusing the speech classifier into MIA-IRL we exclude the categories "silence" and "unknown" from the output distribution and renormalize to obtain a categorical distribution over all possible actions.

6.2.2.2 Gestures

Besides speech commands, humans also use nonverbal cues to communicate intentions, in particular when they refer to objects (Canal et al., 2015). Therefore we also chose arm gestures as an advice modality. The gestures are predefined, namely pointing gestures for objects and a 2-arm symbolic gesture for the action pour. Using an RGB-D camera (Intel Realsense D435), the human skeleton is tracked based on Openpose (Cao et al., 2021). Missing skeleton frames are interpolated using univariate splines. The tracked joint positions of arms and shoulders are aligned with the neck joint and scaled to uniform length in order to become invariant to the human-camera distance. Since we assume a gesture duration of 1 second with a skeleton tracking frame rate of 30Hz, the resulting 30 samples of respective upper body joint positions for a gesture are transformed into a single vector as features for classification. As a classification method we chose a multiclass Support Vector Machine (SVM) with a polynomial kernel of degree 2 (C=1, γ=0.1), implemented using the machine learning framework Sklearn in Python. As class labels, we provide the possible actions. The trained SVM does not only return the predicted advised action but also provides a categorical probability distribution as output.

6.2.3 Fusion Module

The categorical output distributions returned by the base classifiers are fused using Independent Opinion Pool (IOP) (Andriamahefa, 2017; Berger, 1985). IOP fuses D categorical probability distributions $p(a_H|m_i)$ over advised actions a_H given modality data $m_i, i = 1, \ldots, D$ by multiplying them and renormalizing the resulting vector to sum to 1 in order to obtain a categorical distribution. Thus, the resulting fused distribution is

$$p(a_H|m_1, \ldots, m_D) \propto \prod_{i=1}^{D} p(a_H|m_i). \tag{6.2}$$

Assuming conditional independence of observations from different modalities m_i and with them the categorical output distributions $p(a_H|m_i)$ returned by each modality's classifier and an uninformed prior $p(a_H)$ over actions a_H, this fusion method can be derived as probabilistically optimal by applying Bayes' rule. Its advantages are uncertainty reduction through fusion and uncertainty-dependent fusion impact (Andriamahefa, 2017; Hayman & Eklundh, 2002). If the categorical base distributions to be fused are non-conflicting, the fused distribution is less uncertain than the base distributions, i.e., its entropy is lower. If instead the base distributions are conflicting, the resulting fused distribution's uncertainty is increased. Moreover, the less uncertain base distribution has a higher impact on the fused distribution than the more uncertain base distribution.

Since in the defined MDP some actions are impossible in specific states, in addition to multiplying the base distributions according to IOP, the fusion module additionally excludes the probabilities of these impossible actions from the fused distribution. Then the remaining probabilities are renormalized to sum to 1. Algorithm 6.2 shows the complete functionality of the proposed fusion module.

Algorithm 6.2 Fusion Module

```
Require: classifiers' output distributions $p(a_H|m_i)$, $Q[s,:]$
// multiply distributions
$p(a_H|m_1,\dots,m_D) = \prod_{i=1}^{D} p(a_H|m_i)$
// remove unavailable actions
for actions $a = 0, 1, \dots$ do
    if $Q[s,a] == -\infty$ then
        remove entry $p(a_H|m_1,\dots,m_D)[a]$
    end if
end for
renormalize $p(a_H|m_1,\dots,m_D)$ to sum to 1
return $p(a_H|m_1,\dots,m_D)$
```

6.2.4 Action Selection Module

While the proposed fusion module outputs a categorical probability distribution over all possible actions, the RL algorithm requires a discrete action to be executed. If we just chose the action with the highest probability, we would discard valuable information about the decision's uncertainty, which we intentionally wanted to consider by using probabilistic classifiers. Therefore, we propose sampling from the fused categorical distribution $p(a_H|m_1,\dots,m_D)$ to obtain a probabilistically selected action $\hat{a}_H$ to be executed by the RL agent. If two actions' probabilities are quite similar after fusion, by sampling, each of them could be chosen to be executed instead of only the one with the slightly higher probability. Thus, we account for the system's uncertainty about the human's advice. Also, this action selection allows additional exploration, which is particularly helpful in case of imperfect base classifiers.

6.3 Experimental Evaluation

In this section, we present the results of the experimental evaluation of our approach involving Bayesian fusion of multimodal advice. In Section 6.3.1 we show the main advantages of our fusion method IOP in MIA-IRL in comparison to the fusion method in the related approach C-IRL (Cruz et al., 2018) on artificial base distributions. Next, we compare the performances of MIA-IRL, non-interactive RL, and C-IRL in a simulated grid world environment (Section 6.3.2) and in an HRI task with a 7-DoF robot arm and 10 human subjects (Section 6.3.3). For all comparisons between MIA-IRL and C-IRL we replaced the fusion module and the action selection module accordingly, while using the same RL module and base classifiers.

6.3.1 Advantages of Bayesian Fusion

As mentioned in Section 6.1, the fusion method proposed in C-IRL (Cruz et al., 2018) shares some desirable properties with our method MIA-IRL such as uncertainty reduction and uncertainty-dependent fusion impact. However, because C-IRL does not consider the complete base classifier distributions and does not fuse according to Bayes' rule, there exist particular situations in which the IOP fusion, which we propose for MIA-IRL, shows

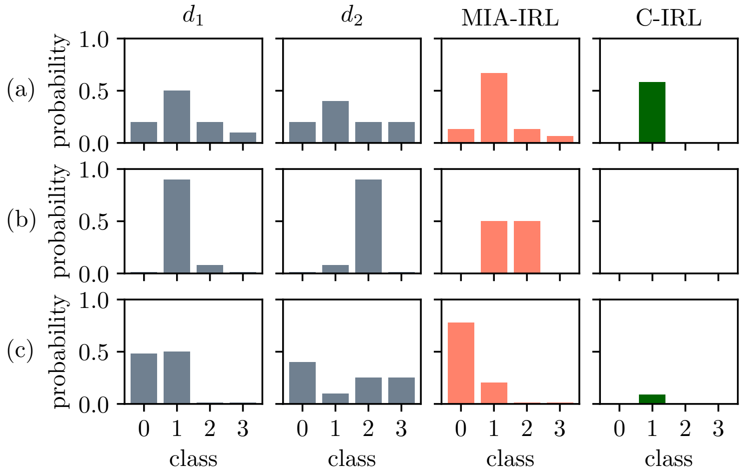


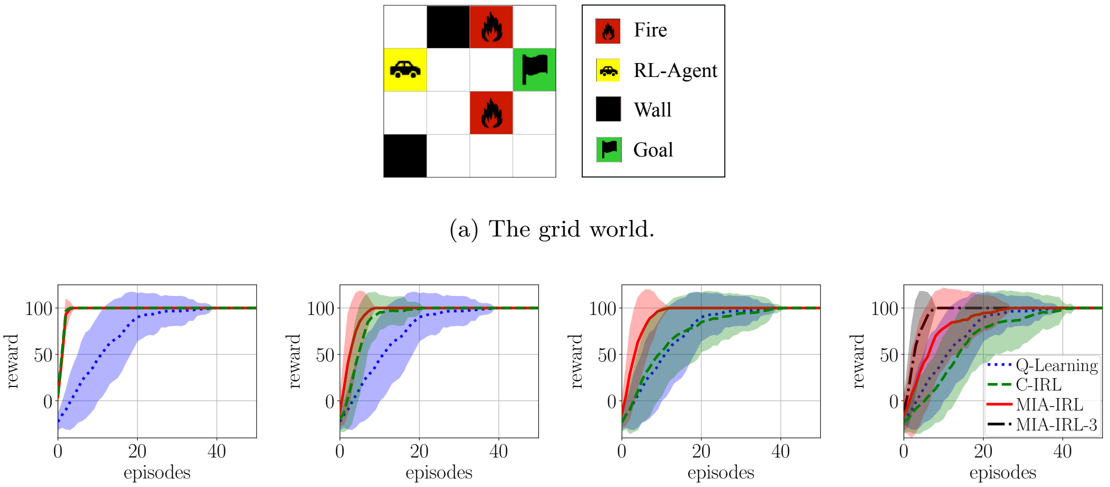
Figure 6.2: Comparison of IOP in MIA-IRL and the fusion method in C-IRL (Cruz et al., 2018) on exemplary base distributions d_1 and d_2. C-IRL disregards information by discarding all probabilities but the highest one (a), returns no information about which class to choose for conflicting distributions (b), or even chooses a different class than Bayes optimal IOP (c).

clear advantages. Three exemplary cases for such situations are shown in Figure 6.2. In Figure 6.2(a) C-IRL reduces uncertainty such as MIA-IRL but discards all probabilities apart from the highest one. However, MIA-IRL, which is Bayes optimal, assigns non-zero probabilities to all possible actions. Therefore, C-IRL's fusion method neglects the uncertainty that should be reflected in the fused distribution. It would never choose classes 0, 2, or 3, although there is a small probability that one of these classes is the correct one. Figure 6.2(b) shows two conflicting base distributions. Fusion with IOP in MIA-IRL results in a distribution that assigns the same probability to classes 1 and 2. However, when fused with C-IRL all classes have a probability of 0. Thus, C-IRL disregards the fact that only classes 1 and 2 should be considered and classes 0 and 3 can be neglected. In Figure 6.2(c), fusing two conflicting base distributions, the fusion rule in C-IRL would even choose a different action than the one selected by Bayes optimal IOP in MIA-IRL. Most likely, this would lead to a misclassification by C-IRL.

These three examples highlight the theoretical advantages of the IOP fusion used in MIA-IRL compared to the fusion method in C-IRL (Cruz et al., 2018). We argue here, that these advantages lead to an increase in learning speed for IRL in particular in cases, where base classifiers may partially output wrong distributions, which we demonstrate in the following sections for a simulated grid world and a real HRI task.

6.3.2 Grid World

We first evaluate MIA-IRL in a simulated 4 × 4 grid world environment (Figure 6.3(a)) where an agent is supposed to reach a goal while avoiding falling into one of two fires. If the agent falls into a fire, the episode ends and the agent receives a negative reward of -100. Otherwise, if the agent reaches the goal marked by the green flag, it receives a positive reward of 100. An episode may also end with a zero reward if the number of required steps in one episode exceeds 15 steps.



(a) The grid world.

(b) C_1 and C_2 are correct. (c) C_2 confuses right/left. (d) C_2 always incorrect. (e) In 20% both incorrect.

Figure 6.3: Learning curves of non-interactive Q-learning, C-IRL (Cruz et al., 2018), and MIA-IRL for the grid world in (a) with simulated advice in the first 10 episodes. Mean rewards (lines) and standard deviations (shaded areas) over 50 runs are shown. In (b) both classifiers C_1 and C_2 advise correct actions. (c) shows the results for a correct C_1 and a C_2 confusing actions "right" and "left". In (d) C_2 additionally confuses actions "up" and "down", and in (e) in 20% of cases both classifiers fail. As (c) – (e) show, MIA-IRL clearly outperforms C-IRL if the individual modalities' classifiers partially fail. The additional curve in (e) (MIA-IRL-3) shows how performance improves by including a third classifier to MIA-IRL.

We provide simulated advice in the form of two randomly generated categorical distributions, which simulate two individual modalities' classifier outputs. This simulated advice is given during the first 10 episodes of learning.

First, we simulate correct non-conflicting categorical distributions as advice, i.e., we randomly generate two categorical distributions in which the probability for the correct action is always above 0.5. The resulting learning curves for non-interactive Q-learning, C-IRL, and MIA-IRL are shown in Figure 6.3(b). Here, we plot the mean and standard deviation for the reward per episode averaged over 50 repeated runs, while for each episode we evaluate the policy 100 times and average the obtained rewards. MIA-IRL (red) as well as C-IRL (green) converge faster than standard non-interactive Q-learning (blue). A Kruskal-Wallis-Test on the convergence times of the three compared approaches showed a significant difference (p<0.001). The Conover-Posthoc-Test additionally provided evidence that MIA-IRL converges significantly faster than standard Q-learning (p<0.001). However, the convergence times of MIA-IRL and C-IRL do not differ significantly. Thus, in the case of non-conflicting correct outputs of both modalities' individual classifiers human advice speeds up learning compared to non-interactive RL, but the particular IRL approach, either C-IRL or MIA-IRL using IOP, does not significantly influence the learning speed.

However, as real-world classifiers for human advice cannot be assumed to be always correct, next we simulate a case where one base classifier C_1 always outputs a correct distribution, while the second classifier C_2 confuses the actions "right" and "left". If the correct action is "right", for C_2 a distribution with a probability above 0.5 for action "left" is randomly generated and vice versa. Figure 6.3(c) shows that in this case MIA-IRL (red) converges faster than both non-interactive Q-learning (blue) and C-IRL (green). A Kruskal-Wallis significance test showed a significant difference between the convergence times of the three compared approaches (p<0.001). The Conover-Posthoc-Test revealed a significant difference between MIA-IRL and non-interactive Q-learning, C-IRL and non-interactive Q-learning, and MIA-IRL and C-IRL (p<0.001). Accordingly, MIA-IRL is more robust against partially incorrect classifier output in this case.

This effect is even stronger in a third simulated experiment, where C_2 is assumed to also confuse actions "up" and "down" in addition to "left" and "right", while C_1 is still assumed correct. Figure 6.3(d) shows the corresponding learning curves. The convergence times of all approaches are significantly different (Kruskal-Wallis-Test, p<0.001). According to a Conover-Posthoc-Test, there is no significant difference between the convergence times of non-interactive Q-learning and C-IRL (p=0.34), but a significant difference between MIA-IRL and non-interactive Q-learning (p<0.001) and MIA-IRL and C-IRL (p<0.001).

If we further modify the third simulated experiment in a way that in 20% of cases both classifiers fail, MIA-IRL still outperforms C-IRL and Q-learning significantly (p<0.001), as shown in Figure 6.3(e). Also, since MIA-IRL is straightforwardly extendable to more than two classifiers, we can easily add a third classifier, which is correct in 60% of cases. MIA-IRL-3 using 3 advice classifiers significantly outperforms MIA-IRL and C-IRL with 2 classifiers (p<0.001). We expect that adding more classifiers to MIA-IRL can further increase the robustness of advice detection and by this the learning speed, depending on the quality of individual classifiers.

We conclude from the simulated experiments that if individual classifiers partially fail in detecting the correct human advice, MIA-IRL clearly outperforms C-IRL.

6.3.3 Pancake Scenario

In addition to the simulated grid world scenario, we also evaluated our approach in a real HRI scenario where human subjects can advise a 7-DoF robot arm using speech and gestures. Here, the goal of the task is that the robot should learn to assist a human in preparing a pancake batter. The task is solved successfully once the robot gets flour and milk from a nearby shelf and pours them into a bowl. The state of the robot is defined by the position of the arm, which can be AT-BOWL or AT-SHELF, the current object in the robot's hand (or the hand being empty), the positions of the objects and the current state of the bowl, which indicates if ingredients have already been poured inside. In our experiments, the objects flour, flower, and roll are always placed on the shelf, whereas the position of the milk changes between the shelf and the table between different episodes. In total, this results in 320 possible states. There are 7 actions, i.e., GO-SHELF, GET-MILK, GET-FLOUR, GO-BOWL, POUR, GET-FLOWER, GET-ROLL. The robot receives a reward of 100 if the task is solved successfully and a negative reward of -100 in case of a failure, which happens when the robot pours wrong ingredients such as flowers or the roll into the bowl or if the robot tries to get objects when already having another object in its hand. The action POUR does not only include pouring the respective ingredient into the bowl but also placing it close to the bowl on the table afterward. If the maximum number of 20 steps per episode is exceeded the episode ends with zero reward. Figure

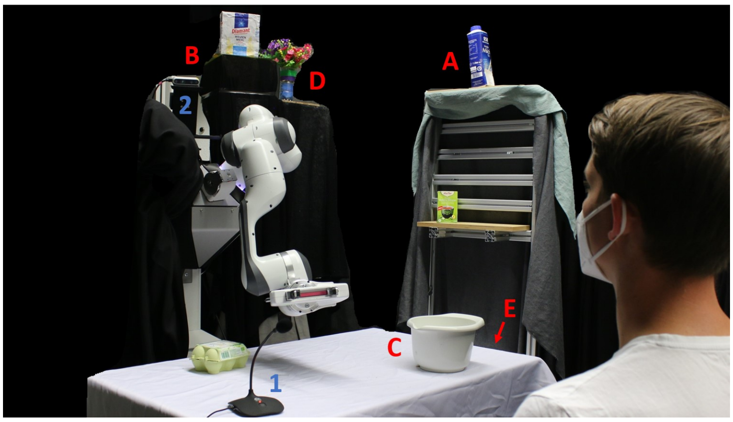


Figure 6.4: The experimental setup used for evaluating MIA-IRL. A human is seated at a table to teach a 7-DoF robot arm to prepare pancake batter. The robot's task is to pour the required ingredients milk (A) and flour (B) into a bowl (C). Flowers (D) and a roll (E, not visible from shown perspective) should not be picked by the robot. The human can give advice by speech commands recorded with a microphone (1) and gestures recorded by a depth camera (2).

6.4 shows the full task setup. For each of the 7 actions, a speech classifier is trained to recognize a corresponding keyword and a gesture classifier to recognize a corresponding gesture. Details on the classifiers used for the experiments of this work can be found in Section 6.2.2.1 and Section 6.2.2.2 respectively. The experimental setup was designed in a particular way to evaluate IRL in cases where classifiers may confuse intended actions. For instance, some objects are placed close to each other to cause similar pointing gestures, e.g., flour and flower, as can be seen in Figure 6.4. Moreover, we chose actions with similar-sounding keywords, i.e., the keyword "roll" to get a roll (similar to "bowl") and the keyword "flower" to get a flower (similar to "flour").

In the described experimental setup, we conducted experiments with 10 human participants (4 female, 6 male, 3 aged 18-25, 7 aged 26-35), who advised the robot in preparing pancake batter.[1] After a short briefing, during which the participants got familiar with the required gestures and keywords as well as the robot's movements, they carried out two experiment blocks, interrupted by a short break. In each block, the participant gave advice over the first 20 episodes, using gestures and speech commands. This choice of 20 episodes of human advice was made after preliminary experiments, and is a trade-off between performance increase and time required for each participant. In one of the blocks, MIA-IRL was used for learning and in the other block, the related method C-IRL (Cruz et al., 2018) was applied. To eliminate sequence effects, 5 participants started with MIA-IRL, 5 with C-IRL.

For each method, after the 20 initial episodes with human advice, we let the RL agent finish the learning until convergence of the average rewards per episode. Here, we average over 50 individual runs of learning to cancel out randomness. In each episode the learned policy is evaluated 100 times, and the resulting learning curves are compared between MIA-IRL and C-IRL. In addition, we also evaluated standard non-interactive Q-learning as a baseline.

[1] The experiments were approved by the ethics committee of TU Darmstadt on September 21, 2021 (approval code EK44/2021).

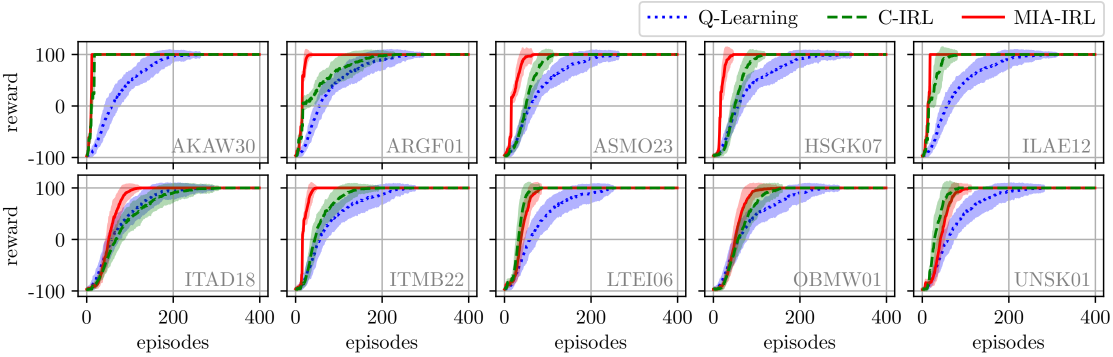


Figure 6.5: Learning curves of non-interactive Q-learning, C-IRL (Cruz et al., 2018), and MIA-IRL individually for all 10 participants of the pancake experiment. Each plot is labeled with the respective participant's code. Mean rewards (lines) and the corresponding standard deviations (shaded areas) over 50 runs of learning with human advice given in the first 20 episodes are shown.

The individual resulting learning curves of all participants for MIA-IRL, C-IRL, and non-interactive Q-learning are shown in Figure 6.5. For all participants MIA-IRL converges faster than non-interactive Q-learning. For 4 participants MIA-IRL and C-IRL perform similarly, while for the remaining 6 participants MIA-IRL outperforms C-IRL. The differences between participants are caused by subject-dependent variation of base classifier distributions. Particularly classifications of flour are crucial since flour is necessary for success but ambiguous for speech (similar sound flower) and pointing gestures (located next to flower). For AKAW30, LTEI06, OBMW01, and UNSK01, MIA-IRL and C-IRL perform similarly, since for all of them one classifier detects flour accurately and with high certainty, while the other one is uncertain. Thus, for both methods the certain base classifier is decisive, while the fusion method, either MIA-IRL or C-IRL, has only little impact. In contrast, e.g., for ITMB22, MIA-IRL performs best, since the gesture classifier is uncertain between flour and flower with flower more likely and the speech classifier is even more uncertain with flour more likely. C-IRL favors the more certain gesture classifier and thus fails often, while MIA-IRL's fusion more often correctly chooses flour. For ARGF01 C-IRL suddenly diverges, since it only learned to solve the task from one of two starting states (milk on table). Thus, at an average reward of 50 only MIA-IRL, which learned more also for the other starting state, continues its steep increase. The base classifiers' output distributions here often match the example in Figure 6.2(b), where MIA-IRL outperforms C-IRL.

In addition to the learning curves of individual participants, Figure 6.6 shows the mean learning curves over all 10 participants for MIA-IRL (red), C-IRL (green), and non-interactive Q-learning (blue). MIA-IRL converges faster than both C-IRL and standard Q-learning. The Mann-Whitney-U-Test for independent samples showed a significant difference between the convergence times of MIA-IRL and standard Q-learning (p<0.001) and between C-IRL and standard Q-learning (p<0.001). The Wilcoxon-Signed-Rank-Test for dependent samples confirmed a significant difference between the convergence times of MIA-IRL and C-IRL (p<0.001). Thus, MIA-IRL clearly outperforms C-IRL also in real experiments with human advisors and real classifiers. In particular, the experiments show

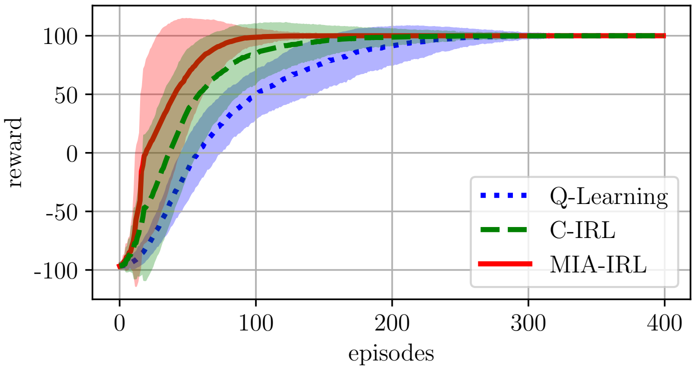


Figure 6.6: Learning curves of non-interactive Q-learning, C-IRL (Cruz et al., 2018), and MIA-IRL averaged over all 10 participants of the pancake experiment. Both mean rewards (lines) and the corresponding standard deviations (shaded areas) over 50 runs of all 10 participants are shown. MIA-IRL converges significantly faster than standard Q-learning and C-IRL.

again that MIA-IRL is more robust against misclassifications of given human advice and conflicting outputs of the individual modalities' classifiers.

6.4 Conclusion

In this work, we proposed MIA-IRL, a novel Interactive Reinforcement Learning approach that enables humans to advise a robot via multiple modalities, such as speech and gestures. In contrast to previous work, we fuse the modalities' classifiers' output distributions with the method Independent Opinion Pool, which can be derived as Bayes optimal and explicitly considers the individual modalities' uncertainties correctly. Importantly, this also allows probabilistic action selection through sampling from the resulting fused distribution, instead of just choosing the most probable action, and straightforward integration of more than two modalities. In a simulated grid world scenario as well as in an HRI experiment with human participants and a real robot we showed that our approach clearly outperforms the closest related state-of-the-art approach (Cruz et al., 2018). In particular, MIA-IRL is more robust against misclassifications of the modalities' individual classifiers. Thus, MIA-IRL lays an improved solid foundation for future development of multimodal IRL.

A limitation of MIA-IRL is that although it exploits the uncertainty represented by the fused distribution by selecting the executed action by sampling from it, it always selects an action, no matter how uncertain the fused distribution is. Therefore, for future work, we want to further exploit the uncertainty represented by the fused distribution. For instance, one could include an active request for additional information if the given advice is too uncertain in order to reduce the risk for catastrophic failures. Additionally, since MIA-IRL is limited to the uncertainty of the current output distributions of each modality's classifier, we plan to extend our fusion module to explicitly consider the properties of the individual base classifiers, such as their bias, variance, and uncertainty, as well as potential correlations between them. Another promising line for future work is to evaluate MIA-IRL with additional modalities such as gaze or facial expressions, and add an additional module that learns and preserves human advice over time to enable reusing the given advice during the entire learning process.

7

MULTIMODAL DETECTION OF THE INTENTION FOR INTERACTION IN HUMAN-ROBOT INTERACTION

In the previous chapters 5 and 6, we recognized human intentions for intuitive human-robot interaction or suggestions of actions for interactive reinforcement learning. In these two approaches, the respective intentions or suggestions are specific for the particular task at hand. For example, the intention to get a bowl from a robot is specifically defined for the task to prepare a salad together with a robot. It is not necessarily transferable to other human-robot interaction tasks. Moreover, the human signals such as their gestures or speech commands for the respective intentions or action suggestions are predefined for the specific task at hand instead of being learned from natural human behavior.

However, future assistive robots will face interactions with humans in various different scenarios, such as support in household tasks (Graf et al., 2004), elderly caregiving (Santhanaraj & MM, 2021), or shared workspaces in industry (Vojić, 2020). In these different application scenarios a wide range of possible interactions between human and robot might occur, including a large number of task-specific intentions not all of which can be predefined.

Nonetheless, there are intentions that re-occur in different potential interactions. In particular, many interactions between human and robot have in common that they should be started on request of the human, i.e., once the human shows an *Intention for Interaction* (Mollaret et al., 2015). Accordingly, here we propose an approach for detecting this human Intention for Interaction (IFI) from natural human behavior.

As we already noted in the previous chapters 5 and 6 for task-specific intentions, humans naturally communicate their intentions using multiple modalities (Jaimes & Sebe, 2007). From a young age, humans attribute intentions to specific motions (Blakemore & Decety, 2001) and communicate their intentions using body language (Gaschler et al., 2012). Additionally, speech is a common channel for communicating intentions since it is common between humans (Frydrychowicz & Matejczuk, 2006). Thus, using multimodal data such as body poses and speech can enable automatic detection of a human's Intention for Interaction (Foster et al., 2017; Mollaret et al., 2015; Mollaret et al., 2016; Vaufreydaz et al., 2016). Moreover, combining multimodal data from speech and body movements can increase the accuracy and robustness of intention recognition (Jaimes & Sebe, 2007) and reduce its uncertainty (Trick et al., 2019, see Chapter 5).

Previous approaches for detecting the Intention for Interaction from multimodal data such as body poses and speech (Bohus & Horvitz, 2009; Foster et al., 2017; Mollaret et al., 2015; Mollaret et al., 2016; Vaufreydaz et al., 2016; Q. Xu et al., 2013) focused on specific tasks done by the humans (Bohus & Horvitz, 2009; Foster et al., 2017; Q. Xu et al., 2013), considered only one interaction type between human and robot (Bohus & Horvitz, 2009; Foster et al., 2017; Vaufreydaz et al., 2016; Q. Xu et al., 2013), or assumed invariability

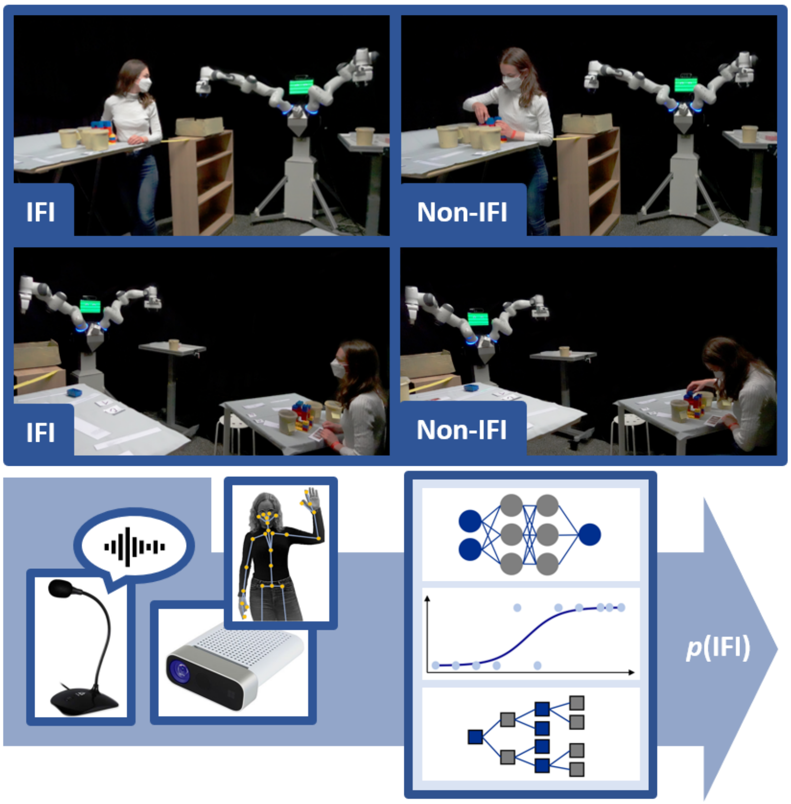


Figure 7.1: An overview of the proposed approach for automatically detecting a human Intention for Interaction (IFI). We designed a human-robot interaction experiment that elicits human behavior communicating IFI over varying tasks, interaction types, and positions and orientations towards the robot. We collected multimodal data including speech and body poses of 21 human subjects, on which we trained probabilistic multimodal classifiers that detect task-independent IFIs.

of the human's position (Foster et al., 2017; Mollaret et al., 2015; Mollaret et al., 2016), which might limit their generalizability.

Here, instead, we investigate how humans show their Intention for Interaction towards a robot in a more versatile experimental setup. Specifically, we recorded the natural behavior of 21 human subjects that were asked to perform different kinds of tasks and interactions with the robot, including sitting and standing positions as well as different orientations towards the robot. On the recorded data we trained different multimodal classifier models (Figure 7.1).

Overall, the main contributions of this work are the following. First, we designed a human-robot interaction experiment that elicits human behavior communicating IFI over varying tasks, interaction types, and positions. Second, in this setup we collected a new data set of natural human behavior while interacting with the robot and analyzed the occurrences of human IFIs. Finally, we trained task-, interaction-, and position-independent classifiers for IFI detection on the collected data. In particular, we systematically compare and quantitatively evaluate unimodal and multimodal classifiers that are able to detect IFIs from natural human behavior. For multimodal classification, we compare feature fusion and decision fusion using the Bayesian fusion method Independent Opinion Pool (Berger, 1985), which we applied for multimodal intention recognition and interactive reinforcement learning in Chapters 5 and 6.

The rest of this chapter is structured as follows. Section 7.1 discusses related work. In Section 7.2, we describe our proposed experiment, the collected data, and the trained IFI classification models. In Section 7.3, we present the evaluation of our data set and the trained classifiers' performances. Finally, we conclude and outline future work in Section 7.4.

7.1 Related Work

Various approaches deal with the detection of engagement in human-robot interaction, which terms how interactors begin, continue, and end their connection with each other while interacting (Sidner et al., 2005). Some of these approaches try to detect human engagement during an ongoing interaction with a robot to adapt their interaction strategy accordingly (Ishii et al., 2013; Nakano & Ishii, 2010; Ooko et al., 2011; Rossi et al., 2021; Z. Zhang et al., 2022). Others explicitly deal with the start of an interaction, but with the robot instead of the human as the initiator. In the works of Satake et al. (2009), C. Shi et al. (2015), Z. Zhang et al. (2021), and Ito et al. (2020) the appropriate time and position for the robot to start a conversation with a human is detected, e.g., from human walking patterns (Satake et al., 2009) or nonverbal signs and face information (Ito et al., 2020; Z. Zhang et al., 2021).

For detecting an Intention for Interaction actively expressed by the human, Cesta et al. (2007) rely on explicit vocal commands, i.e., questions, which the humans need to utter to signalize their IFI to a robot. Similarly, Burger et al. (2012) use explicit gestures and language commands to detect an IFI. However, predefined commands for starting an interaction require precise instructions and render the interaction less natural and intuitive (Li et al., 2012).

Therefore, several approaches investigate how to detect a human's Intention for Interaction from their natural behavior, without giving any instructions on how to signal the IFI (Bohus & Horvitz, 2009; Foster et al., 2017; Mollaret et al., 2015; Mollaret et al., 2016; Vaufreydaz et al., 2016; Q. Xu et al., 2013). Some of them only use body poses (Q. Xu et al., 2013) or face information (Bohus & Horvitz, 2009), others use multimodal data consisting of body poses, such as the human's torso angle or head position, and speech (Foster et al., 2017; Mollaret et al., 2015; Mollaret et al., 2016; Vaufreydaz et al., 2016).

Q. Xu et al. (2013) performed a Wizard-of-Oz study in a reception desk scenario. They recorded visual cues, e.g., distance, upper body pose, and face direction, and trained two SVM classifiers to detect the intention to start a conversation (=IFI) and to end a conversation. IFI could be detected with precision 0.83 and recall 0.72. However, they only considered one situation, i.e., people at a reception desk, with no other specific tasks than showing an IFI and only one interaction type, i.e., information consultation. Moreover, people were always standing when showing an IFI.

Foster et al. (2017) also collected data to automatically detect an IFI. In a bar scenario, a human could order a drink after expressing an IFI towards the robot bartender. Recorded data were head position and rotation, the torso angle towards the robot, and speech. As

Q. Xu et al. (2013), they considered only one situation with the only task to show an IFI and only one possible interaction type, i.e., ordering a drink at a bar. Also, there was not much variability in the location of the human during shown IFIs, since he or she was always starting the interaction at the bar in front of the robot bartender.

This invariability of the human's positions during data collection can also be found in the works of Mollaret et al. (2015) and Mollaret et al. (2016). They recorded humans' head and shoulder orientation and vocal activity while showing an IFI with the goal of receiving help from the robot, e.g., to find some missing items. Human subjects were seated at a predefined distance to the robot.

In contrast, Bohus and Horvitz (2009) collected data with more variability in the human's location. While recording face position, rotation, width, and height, the human was allowed to move freely in the experiment space and could approach a Kiosk-like robot avatar. Still, the only interaction type possible for the human was to play a game on the robot. This game required the humans to closely approach the robot, making it difficult to generalize to IFIs expressed for different interaction types (e.g., conversation) from different positions in the experiment space and distances to the robot.

Similarly, Vaufreydaz et al. (2016) collected data in a home-like environment, where people could do different tasks, either freely walk around or sit and play cards, while their speech, body pose, and face were recorded. However, as in the work of Bohus and Horvitz (2009) humans could also only interact with the robot by playing a game on the robot's tablet. Therefore, they only recorded data for the intention for an interaction, for which the human needs to closely approach the robot. This might not generalize well to IFIs expressed for other interactions such as conversations and from a further distance.

All related approaches presented above base their IFI detection on data sets that are either limited in the variability of the human's positions towards the robot or were devised for single tasks and interactions between human and robot.

7.2 Task-Independent IFI Detection

We introduce a new versatile data set for detecting human IFIs when interacting with a robot. We carefully designed our experimental setup for data collection to include different tasks, and elicited various types of interactions. The human subjects expressed the provoked IFIs from different positions towards the robot, sitting and standing, and acted in a natural way.[1] Based on this data set we classify IFIs from multimodal speech and body pose data. In the following, we describe our data collection (Section 7.2.1), the preprocessing of the recorded data set (Section 7.2.2), and the classifiers trained for IFI detection (Section 7.2.3).

7.2.1 Data Collection

To collect data about how humans initiate an interaction with a robot, we conducted a Wizard-of-Oz experiment. Human subjects were instructed to build a tower made of building blocks following step-by-step instructions. If they needed help, they could ask

[1] The recorded data set is available at `https://osf.io/mvzsa/`.

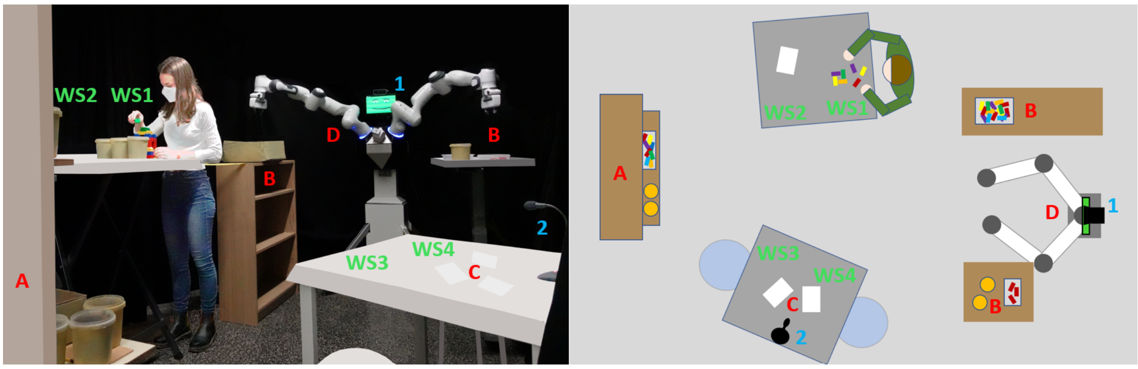


Figure 7.2: The experimental setup, in which we collected our IFI data set. The subjects worked at four different workspaces (WS1-WS4) on a standard and a standing table, i.e., sitting and standing, facing or not facing the robot. Using instruction cards (C), the subjects built a tower of building blocks. If they needed help, the robot Kobo (D) supported them, e.g., by handing over items from its shelves (B). In addition, subjects could get required items from the self-service shelf (A). We recorded body poses with an RGB-D camera mounted above a tablet showing Kobo's face (1) and speech via a Microphone (2).

the two-armed robot Kobo for help, which was covertly controlled by the experimenter. The subjects did not know the goal of the experiment, i.e., collecting IFI data. They were free to act completely natural – standing, sitting, or walking – and interacted with the robot in different ways, either talking or handing over objects.

7.2.1.1 Experimental Setup

The experiment was performed in a lab environment with two tables as the subjects' workspaces, three shelves, and the robot Kobo (Figure 7.2). Kobo is equipped with two 7-DoF Panda arms (Franka Emika) and a tablet (Samsung Galaxy S6) mounted between the arms that displayed a smiling face. Additionally, an RGB-D camera (Azure Kinect) was installed above the tablet, covering the experimental space from Kobo's perspective. Next to Kobo, we placed two shelves with containers holding items that Kobo could hand to the subjects, e.g., additional building blocks. Opposite of the robot, a self-service shelf was placed where subjects could take required items such as building blocks or boxes by themselves.

A standing table and a standard table were placed between the self-service shelf and Kobo (Figure 7.2). Each table provided two workspaces for the subjects, two standing and two sitting, two facing the robot and two facing away from it. The order of workspaces for different subjects was changed to increase variability of the data. On all workspaces, numbered instruction cards were placed that helped the subjects with subtasks for building the tower. On the standard table, a microphone was placed to record audio data.

Throughout the experiment, the experimenter was not visible to the subjects, so they could only interact with Kobo. Through Kobo's RGB-D camera, the experimenter watched the experiment and triggered Kobo's reactions accordingly.

7.2.1.2 Experimental Procedure

After a short briefing and informed consent, the experiment started with a familiarization phase with the robot Kobo. The subjects were shown how Kobo speaks and moves. They were guided through a demonstration interaction with Kobo and instructed that the robot senses their movements and speech.

Within the experiment, the subjects had to build a tower out of building blocks. This task consisted of 10 subtasks that the subjects had to complete in a specific order. The subtasks were explained on instruction cards lying on the workspaces (see Figure 7.2, red C). After completing a subtask, subjects had to get in touch with Kobo, which then told them at which workspace they should continue with which instruction card. The subtasks were designed to be versatile and included labeling containers, sorting blocks into containers, moving required items, and assembling building blocks. After successfully building the tower, which was checked by Kobo, they tidied up the workspaces. In order to increase the variability of our data and thus the generalizability of IFI detection, we carefully designed the experiment's subtasks to include tasks that were done sitting or standing, either facing the robot or not facing it, or required walking around.

We are particularly interested in the situations in which the subjects contacted Kobo and thus showed IFIs. Such IFIs could be caused by 4 different situations: 1) the subject wanted to know the next instruction from Kobo, 2) the subject was explicitly asked to interact with Kobo in their instruction card, e.g., ask Kobo for a special building block, 3) some instructions were intentionally wrong to make the subject ask Kobo, e.g., the instructions included building blocks that were not provided, 4) required materials were missing, e.g., the provided pen was out of ink, or 5) the subject had spontaneous questions we did not anticipate.

Also, the interactions initiated by the shown IFIs varied. Some were pure conversations, which could be performed from anywhere in the experiment space, others required handovers of items such as building blocks and thus required close interaction.

7.2.1.3 Recorded Data

We collected data from 22 participants (12 male, 10 female), aged between 18 and 35.[2] Each participant received a 6-digit subject code to anonymously refer to in this work. Most participants (12) had no experience with robots, 9 had only some prior contact with robots (1-10 times), and one participant was experienced with robots. The time required for the experiment varied between 18 and 27 minutes. All participants talked German to Kobo and were naive about the experiment's objective of collecting IFI data.

The complete system, including the control of all sensors and the teleoperation of the robot, was realized using ROS. We recorded an RGB video from Kobo's perspective using the Kinect Azure RGB-D camera with the corresponding ROS timestamp for each video frame. The video was later used to label the data (Section 7.2.1.4). Using the Azure Kinect camera and the Body Tracking SDK of the Azure Kinect DK we also recorded positions (x, y, z) of 5 upper body parts of the participants in relation to the robot: their neck, both eyes, and both shoulders. We recorded the body position data at a frame rate

[2] The experiments were approved by the ethics committee of TU Darmstadt on October 28, 2020 (approval code EK39/2020).

of 30Hz and saved the corresponding ROS timestamp for each frame. For recording sound, we used a USB Microphone (Klim) that captured Mono sound with 48000Hz and stored a ROS timestamp for every 100 frames.

7.2.1.4 Labeling

We labeled the recorded data watching the video along with the synchronized audio data. After labeling, we subtracted 1 second of all label timestamps in order to take into account the labeling person's reaction time, which resulted in a precise mapping between label timestamps and shown IFIs. We distinguished between three labels: *IFI-Start*, *Interaction-Start*, and *Interaction-End*. *IFI-Start* labels the moment in which the IFI can be first noticed, either based on the body pose or because the subject starts talking to Kobo. *Interaction-Start* labels the end of the IFI and the start of the interaction between the subject and Kobo, which is marked by Kobo's first response to the subject's request. Accordingly, *Interaction-End* labels the end of the interaction, after which the subject starts working on its own again. The order of labels is always *IFI-Start*, *Interaction-Start*, *Interaction-End*. We decided to use these three labels in order to separate IFIs from the actual interaction. We only wanted to learn the subjects' behavior when they show an IFI, independent of which concrete interaction follows. Therefore, using these labels we exclude all interactions (from labels *Interaction-Start* to *Interaction-End*) from the data.

7.2.2 Data Preprocessing

The recorded multimodal data were synchronized and preprocessed into features. For body poses, we chose three features, i.e., the distance between the human and Kobo and the human's head and shoulder orientation with respect to Kobo (Section 7.2.2.1). For speech, two features were considered: speech activity recognition and hotword detection for the robot's name "Kobo" (Section 7.2.2.2). An overview of the used features is given in Figure 7.3.

7.2.2.1 Body Poses

The recorded body pose data were resampled to a framerate of 30Hz, missing frames, which could be caused by sporadic failures of the skeleton tracking, were interpolated. The label timestamps were matched to the body pose data frames, and all frames between *Interaction-End* (or experiment's start) to *IFI-Start* were labeled as non-IFI, all frames between *IFI-Start* and *Interaction-Start* as IFI and between *Interaction-Start* and *Interaction-End* as Interaction. In order to take into account that there might be some movements that express an IFI even before the labeler reacted, we used an intention buffer k as a global hyperparameter and also labeled the k frames before *IFI-Start* as IFI. In our evaluations we tested $k \in \{0, 5\}$.

From the synchronized body pose data we extract the three features distance, shoulder orientation, and head orientation with respect to Kobo. The distance is computed as the euclidean distance between the subject's neck and Kobo in the x-y plane. For computing the shoulder orientation with respect to the robot in degrees, an auxiliary line connecting both shoulders is constructed. The normal vector in the middle of this line is calculated in the x-y plane. Additionally, the vector between Kobo and the middle of the shoulder-

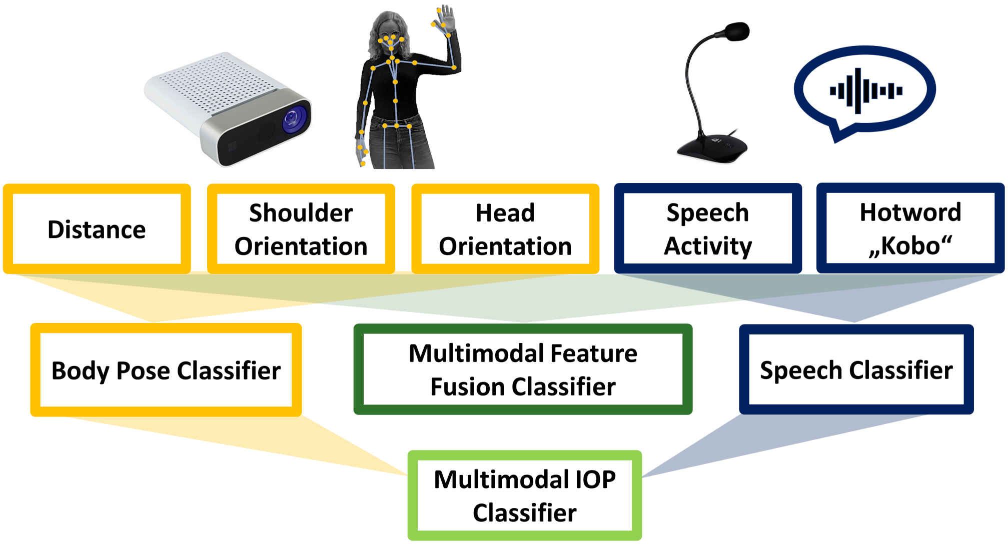


Figure 7.3: The features and classifiers used for IFI detection. From skeleton tracking data provided by an RGB-D camera we extracted the three body pose features distance, shoulder orientation, and head orientation. For speech, we extracted speech activity detection and hotword detection. We trained unimodal classifiers for each body poses and speech and multimodal classifiers, either using feature fusion or decision fusion of the unimodal classifiers' outputs with the Bayesian fusion method Independent Opinion Pool (IOP).

connecting line is determined. The angle between those two vectors in degrees represents the shoulder orientation. The values range from 0° to 180°, where for 0° the human directly faces Kobo and for 180° the human is facing away from Kobo. The head orientation is computed similarly to the shoulder orientation, but with eye instead of shoulder positions.

We used a sliding window approach for classification. For each subject we first obtained all sequences from *Interaction-End* (or the start of the experiment) to *Interaction-Start* in order to exclude all sequences labeled as Interaction. For each frame in those sequences, we created a window of the last n frames with window sizes $n \in \{30, 60\}$, corresponding to 1 or 2 seconds respectively. The features distance, shoulder orientation, and head orientation were additionally smoothed with a median filter with size 5 and standardized by subtracting the mean and scaling to unit variance. These windows formed the data to be classified.

7.2.2.2 Speech

To synchronize the speech data with the body pose data, we mapped speech windows to the body pose windows described in Section 7.2.2.1 using the saved speech timestamps and the timestamps of the body pose windows. As for the body pose windows, this was done for intention buffers $k \in \{0, 5\}$ and window sizes $n \in \{30, 60\}$. As features we used speech activity detection, as in the works of Mollaret et al. (2015), Mollaret et al. (2016), Vaufreydaz et al. (2016), and Foster et al. (2017), and hotword detection with hotword “Kobo”. Thus, for each speech window, we checked if speech could be recognized and if the speech contained the word “Kobo”. Speech activity detection was realized using the Python package SpeechRecognition with the Google Speech Recognition API for German language. If speech is recognized in a window, its value is 1, else 0. Additionally, we detect the hotword “Kobo” as a second feature. If the recognized text or one of the provided alternatives for a window contained the hotword, the value was set to 1, else to 0.

7.2.3 IFI Classification

Using the data and features specified above, for detecting IFIs we compare different classifiers using body pose and speech data. We trained body pose-only and speech-only classifiers as well as multimodal classifiers considering both body pose and speech data. For multimodal classification, we compare feature fusion and decision fusion using the Bayesian fusion method Independent Opinion Pool (IOP) (Berger, 1985). Figure 7.3 gives an overview of the considered classifiers.

We only use probabilistic classifiers that map the input data to probabilities for IFI since this allows quantification of uncertainty, which can be useful particularly in intention recognition for human-robot interaction (Trick et al., 2019, see Chapter 5). Also, using probabilistic classifiers maximizes the flexibility in choosing methods for decision fusion since probabilistic outputs can be mapped to discrete outputs but not vice versa.

In this work, we evaluate three probabilistic classification methods: Logistic Regression, Multilayer Perceptron (MLP), and Decision Tree. All classifiers are implemented using the Python package sklearn.

For all classifiers hyperparameters had to be set. As explained in Section 7.2.2, we vary intention buffer $k \in \{0, 5\}$, which determines how many frames before each *IFI-Start* label are regarded as IFI, and window size of the sliding windows $n \in \{30, 60\}$. The two-layer MLP classifier has two additional hyperparameters, the number of neurons in the first layer $s \in \{4, 6, 8\}$ and in the second layer $t \in \{4, 6, 8\}$. The Decision Tree classifier has one additional hyperparameter $d \in \{3, 5, 7\}$, determining the tree's maximum depth. The best hyperparameters for each classifier were found using an exhaustive grid search.

We train the classifiers using leave-one-out cross-validation, i.e., a classifier is repeatedly trained on the data of all but one subject and evaluated on the remaining subject. By this, we investigate the IFI detection performance for new, unseen subjects and thus the generalizability of our approach.

7.2.3.1 Body Pose Classifier

The body pose classifier C_{body} maps body pose features to the probability for IFI,

$$C_{\text{body}} : x_{\text{body}} \to p(\text{IFI}|x_{\text{body}}), \tag{7.1}$$

where x_{body} is the body pose feature vector with dimensionality $I \times n \cdot 3$ for I windows of window size n and three body pose features distance, shoulder orientation, and head orientation, as described in Section 7.2.2.1.

7.2.3.2 Speech Classifier

The speech classifier C_{speech}, which maps speech features to IFI probabilities, is defined as

$$C_{\text{speech}} : x_{\text{speech}} \to p(\text{IFI}|x_{\text{speech}}), \tag{7.2}$$

where x_{speech} is the speech feature vector with dimensionality $I \times 2$ for I windows and 2 speech features, speech activity and hotword detection, as described in Section 7.2.2.2.

7.2.3.3 Multimodal Classifiers

Besides training individual classifiers for the modalities body poses and speech, we also consider multimodal classifiers that return the probability for IFI given speech and body pose data. We compare feature fusion and decision fusion using the Bayesian fusion method Independent Opinion Pool (IOP).

Feature Fusion of Body Poses and Speech

For multimodal IFI classification using feature fusion we concatenate the body pose features x_{body} and the speech features x_{speech} to one feature vector $x_{\text{body+speech}}$ with dimensionality $I \times n \cdot 3 + 2$ for I windows of window size n with three body pose features and two speech features. The multimodal feature fusion classifier C_{FF} maps this combined feature vector to the probability for IFI,

$$C_{\text{FF}} : x_{\text{body+speech}} \rightarrow p(\text{IFI}|x_{\text{body+speech}}). \tag{7.3}$$

IOP Fusion of Body Poses and Speech

In contrast to feature fusion, where different modalities' data are fused at the feature level, multimodal data can also be fused by decision fusion, i.e., by combining the outputs of individual classifiers using a specific fusion rule. The advantage of decision fusion is high modularity since the individual classifiers can be replaced and additional classifiers can be added straightforwardly. Here, we fuse an individual body pose classifier with an individual speech classifier using the fusion method Independent Opinion Pool (IOP) (Berger, 1985). Thus, the multimodal IOP classifier C_{IOP} maps two IFI probabilities given by body pose and speech classifier to the probability for IFI given body pose and speech features,

$$C_{\text{IOP}} : p(\text{IFI}|x_{\text{body}}), p(\text{IFI}|x_{\text{speech}}) \rightarrow p(\text{IFI}|x_{\text{body}}, x_{\text{speech}}). \tag{7.4}$$

IOP fuses M categorical distributions $p(\text{IFI}|x_{\text{mod}_1}), \ldots, p(\text{IFI}|x_{\text{mod}_M})$ returned by different modalities' IFI classifiers by multiplying them and renormalizing the resulting vector. Thus, the resulting fused distribution is

$$p(\text{IFI}|x_{\text{mod}_1}, \ldots, x_{\text{mod}_M}) \propto \prod_{m=1}^{M} p(\text{IFI}|x_{\text{mod}_m}). \tag{7.5}$$

For our classifiers, this simplifies to

$$p(\text{IFI}|x_{\text{body}}, x_{\text{speech}}) \propto p(\text{IFI}|x_{\text{body}}) \cdot p(\text{IFI}|x_{\text{speech}}). \tag{7.6}$$

Assuming conditional independence of the modality data x_{speech} and x_{body} given IFI and thus of the returned probability distributions $p(\text{IFI}|x_{\text{mod}_m})$ and an uninformed prior over classes IFI and non-IFI, IOP is the probabilistically optimal fusion method according to

Bayes' rule. Thus, given the assumptions, it correctly reduces the decision's uncertainty if the body pose and speech classifier predict the same class and correctly increases the decision's uncertainty if they are conflicting. In previous works on multimodal intention recognition, we already showed that fusing individual modalities with IOP can increase recognition performance (Trick et al., 2022, see Chapter 6; Trick et al., 2019, see Chapter 5).

7.3 Experimental Evaluation

We analyze the recorded data set (Section 7.3.1) and evaluate the performances of the IFI classifiers we trained on this data set (Section 7.3.3) regarding F1 score, recall, and precision (Section 7.3.2). Furthermore, we discuss the behavior of different characteristic human subjects and its impact on the IFI models' performances (Section 7.3.4).

7.3.1 IFI Data Set

The recorded IFI data set consists of 21 subjects since we excluded one subject due to mislabeling. In total, the data set contains 405 IFIs. On average, a subject showed 19 IFIs, minimally 13 and maximally 25. Different subjects showed different numbers of IFIs because some had more questions about the tasks, while others needed less support. For instance, subject ARBN30 somehow managed to use the pen out of ink that we intentionally provided to make the subjects ask for another pen.

Figure 7.4 shows all subjects' positions during IFIs, blue if they were standing or walking, red if they were sitting. As intended with our experiment setup and design, subjects showed IFIs from very different positions in the experiment space, including close IFIs from about 1m to the robot and IFIs further away from about 3m. From the subjects' neck heights we inferred that 18.7% of all IFIs were shown sitting. As we designed half of the tasks at the two sitting workspaces, this is less than we expected. The reason is that some subjects did not take a seat while working at the standard table and some stood up for IFIs. At workspaces WS2 and WS3 subjects are facing Kobo while doing their tasks, while at workspaces WS1 and WS4 they are not. Thus, the data set includes IFIs with different orientations towards the robot.

7.3.2 Performance Measures

For evaluating the proposed IFI classifiers' performances we report recall, precision, and F1 score. Accuracy is not reported, since the data set is imbalanced in favor of non-IFI examples, leading to high accuracy if always predicting non-IFI. For computing recall, precision, and F1 score, the classifiers' probabilistic predictions are discretized with a threshold of 0.5. Recall and precision are not computed based on the number of frames labeled as IFI but on the number of IFIs. Since it is not required that every single frame of an IFI is detected for the robot to react appropriately, an IFI is considered as detected if at least one frame of it is detected as IFI. Also, a predicted IFI is only considered a false positive if none of its frames is labeled as IFI. Thus, we do not penalize IFIs detected a little earlier than they were labeled. Also, when computing the precision, two detected

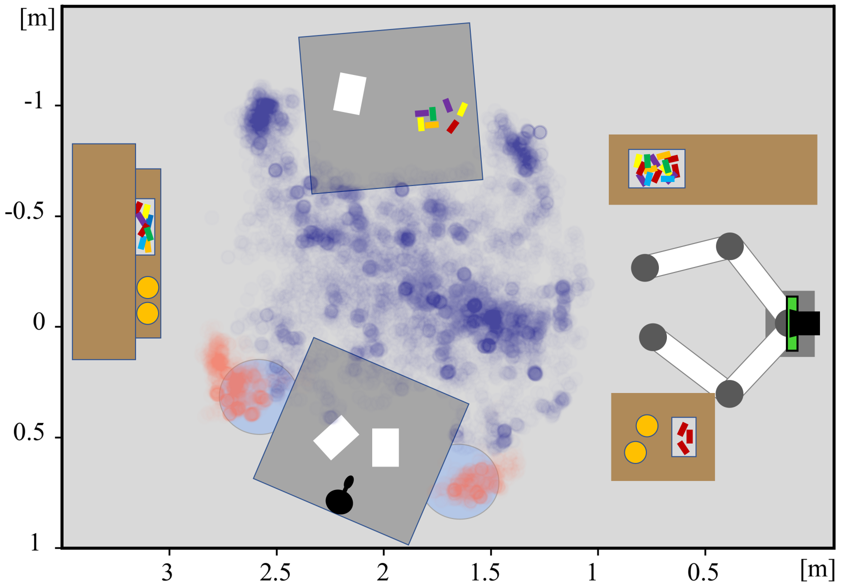


Figure 7.4: The neck positions of all subjects in the experiment space while showing an IFI. As can be seen, subjects showed IFIs from very different positions, close to the robot and further away, standing or walking (blue) and sitting (red). We only plotted every fifth data point to increase readability.

IFIs that are less than 1 second apart are merged and treated as one detected IFI. By this, we avoid counting a just shortly interrupted false IFI detection as multiple false positives.

7.3.3 Classification Performances

Table 7.1 shows mean F1 scores, recalls, and precisions of the best performing IFI classifiers according to F1 score for the body pose, speech, and multimodal classifiers using feature fusion or decision fusion with IOP, and compares the different tested classification methods Logistic Regression, MLP, and Decision Tree. The best-performing body pose classifier is an MLP with window size n=60, intention buffer k=0, and s=t=6 neurons in both layers (F1=0.638, recall=0.671, precision=0.67). Logistic Regression and Decision Tree perform slightly worse, Logistic Regression with a lower recall of 0.593 and Decision Tree with a lower precision of 0.584. For the modality speech the results are similar for all classification methods. Logistic Regression, MLP, and Decision Tree classifiers all achieve the same maximum scores, F1=0.772, recall=0.698, precision=0.919. These scores are also constant over different hyperparameters of the classifiers as long as the hyperparameter window size is set to n=60. This is because the speech classifiers detect an IFI as soon as speech is recognized, and speech recognition works better for n=60 than for n=30. Of course, this can cause problems if more than one human is present in the experiment setup, as we also discuss in Section 7.4. Among the multimodal classifiers that fuse body pose and speech data using feature fusion, the highest mean F1 score (F1=0.794, recall=0.876, precision=0.74) is achieved by Logistic Regression with hyperparameters window size n=60 and intention buffer k=0. For fusing the two modalities with decision fusion using IOP, we compare two different combinations of classifiers. First, we fuse the two best-performing individual classifiers for body poses and speech, thus the MLP classifier with window size n=60,

Table 7.1: Mean F1 scores, recalls, and precisions of the best performing IFI classifiers for different modalities.

	Body Pose Only		
	Logistic Regression	MLP	Decision Tree
F1	0.609	**0.638**	0.593
Recall	0.593	**0.671**	0.647
Precision	0.661	**0.67**	0.584

	Speech Only		
	Logistic Regression	MLP	Decision Tree
F1	**0.772**	**0.772**	**0.772**
Recall	**0.698**	**0.698**	**0.698**
Precision	**0.919**	**0.919**	**0.919**

	Multimodal Feature Fusion		
	Logistic Regression	MLP	Decision Tree
F1	**0.794**	0.789	0.76
Recall	**0.876**	0.887	0.898
Precision	**0.74**	0.729	0.686

	Multimodal IOP Fusion	
	IOP (2 Best Classifiers)	IOP Best
F1	0.788	**0.811**
Recall	0.723	**0.778**
Precision	0.895	**0.88**

intention buffer k=0, and s=t=6 neurons in both layers for body poses and one of the speech classifiers with window size n=60. The best resulting fused IFI classifier achieves F1=0.788, recall=0.723, and precision=0.895. This performance is achieved when fusing the body pose MLP with one of the following four speech classifiers: a Logistic Regression classifier with window size n=60 and intention buffer k=0, or a Decision Tree with window size n=60, intention buffer k=0 and a maximum depth of $d \in \{3, 5, 7\}$. However, testing all possible combinations of individual classifiers for body poses and speech, we found combinations that outperform the combination of the two best individual classifiers resulting in F1=0.811, recall=0.778, and precision=0.88. To achieve these scores, we combined a body pose MLP with window size n=60, intention buffer k=5, and s=t=8 neurons in both layers with one of the following four speech classifiers: a Logistic Regression classifier with window size n=60 and intention buffer k=5, or a Decision Tree with window size n=60, intention buffer k=5, and a maximum depth of $d \in \{3, 5, 7\}$.

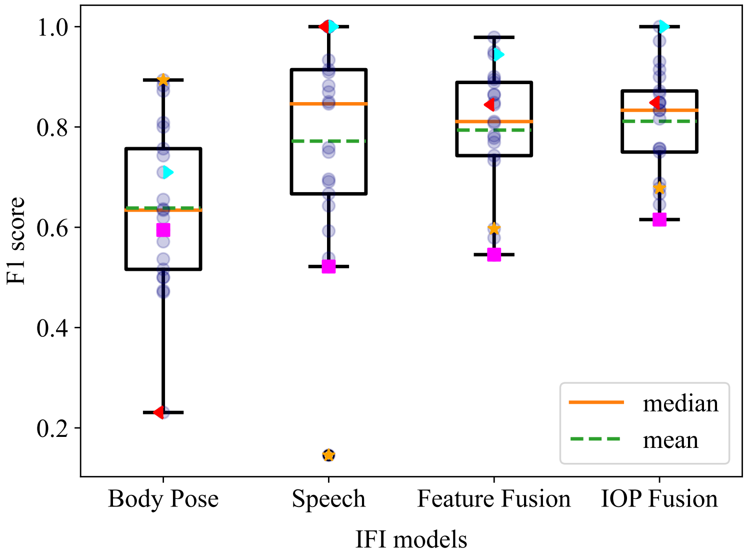


Figure 7.5: Boxplots of the individual subjects' F1 scores for the four best models of each body pose-only, speech-only, feature fusion, and IOP fusion classifiers. All subjects' F1 scores are visualized (blue dots), and some characteristic subjects are highlighted: LZBB10 (red triangle left), EKOK04 (yellow star), RNZA17 (pink square), IEBD07 (cyan triangle right).

7.3.4 Detailed Analysis of Human Behavior and IFI Models

The F1 scores of the best body pose-only, speech-only, and multimodal feature fusion and IOP fusion classifiers are visualized for all 21 subjects in Figure 7.5. Some characteristic subjects' scores are highlighted in order to explain their behavior and the behavior of the different IFI models.

For subject LZBB10 (red triangle left), the body pose classifier shows very poor performance, while the speech classifier achieves F1=1. The reason for this is that LZBB10 did not express his IFIs using body poses but kept looking at what he was doing. Instead, LZBB10 expressed all IFIs using speech, while uttering the hotword "Kobo" for 11 of in total 20 IFIs. Interestingly, he was thereby not only using "Kobo" to start his speech commands – as people are doing when controlling smart devices such as Alexa – but in a more natural way. Despite the body pose classifier's poor performance, since LZBB10 always talked to the robot to start an interaction, the multimodal classifiers perform above average.

In contrast, EKOK04 (yellow star) expressed IFIs very clearly using body poses. To start an interaction, he walked straight to the robot and addressed it head on. Therefore, the body pose classifier performs best for EKOK04 (F1=0.894). In addition to showing body poses, EKOK04 also talked when expressing IFIs. However, unlike LZBB10, he never uttered the hotword "Kobo". Although EKOK04 talked as other subjects did, the speech classifier performs worst for EKOK04. The reason for this is suboptimal speech recognition. While for all other subjects, the speech activity recognition mostly correctly detected speech, for EKOK04 it often missed or incorrectly detected speech. Due to this poor speech recognition, also the multimodal classifiers perform poorly for EKOK04.

Subject RNZA17 (pink square) is not clear in expressing IFIs using body poses or speech. For example, for some IFIs she just looked at the robot without talking or turning to it. This behavior is reflected in the poor performances of all classifiers.

Finally, the IFIs of IEBD07 (cyan triangle right) can be classified well by all classifiers. The body pose classifier does not perform as well as for other participants (F1=0.71), since IEBD07 expresses IFIs by turning to the robot but not walking towards it as e.g., EKOK04. The speech classifier performs perfectly with F1=1 since IEBD07 talks to Kobo for signaling IFIs, while using the hotword "Kobo" for 11 of 18 IFIs. Combining body pose and speech data, the multimodal classifiers also perform very well for IEBD07.

The above analysis of the behavior of different IFI models and human subjects shows that humans express IFIs in various ways. Some are very clear in characteristic body poses, others rely more on speech, some use the hotword "Kobo" frequently, others never use it. Despite this variability in behavior, we could learn models for classifying IFIs. In particular, multimodal models, which perform best on our data set, achieve promising results. Although the speech-only classifier already performs quite well, additionally considering body poses increases the performance and in particular reduces the number of poor-performing outliers. While decision fusion using the fusion method IOP achieves the highest mean F1 score (F1=0.811), the difference to feature fusion (F1=0.794) is not significant (p=0.22, Wilcoxon-Signed-Rank-Test). However, besides performance, decision fusion using IOP is also more flexible in adding or exchanging individual classifiers than feature fusion.

7.4 Conclusion

In this work, we proposed an experimental setup where we investigate natural human behavior when expressing an Intention for Interaction towards a robot. We recorded body poses from RGB-D data and audio data and obtained a data set with in total 405 IFIs from 21 human subjects. In contrast to related approaches, our data set is not limited to a specific task or interaction type and includes standing and sitting IFIs as well as IFIs in different distances and orientations to the robot. Using the recorded data, we trained unimodal and multimodal probabilistic classifiers and compared different approaches for multimodal fusion, i.e., feature fusion and decision fusion. We showed that IFIs can be automatically detected from natural human behavior across different tasks and interaction types, while the best performance could be achieved with a multimodal classifier based on decision fusion with the Bayesian fusion method Independent Opinion Pool.

While working with the data, we noticed some limitations in our data collection: The microphone should have been placed close to the robot to avoid volume differences at different workspaces, and the data should have been labeled by more than one person to measure the labels' quality. Another limitation of the current data set is that only one human is present in the scene. For future work, we consider it interesting to also analyze data of scenarios with multiple humans present, who might also start interactions with each other and thus make it more challenging to detect IFIs towards the robot. Besides, a more advanced speech classifier could consider semantic meaning and prosodic speech features in addition to speech activity and hotword detection, and additionally tracking the human's gaze direction could make it possible to accurately detect if the human is looking at the robot. Another promising line for future work is to validate our IFI classifiers on interactions completely outside of our data set. Further, in future work it should be investigated how the IFI classifiers' uncertainty can be exploited for an appropriate reaction of the robot.

8

A NORMATIVE MODEL FOR BAYESIAN COMBINATION OF SUBJECTIVE PROBABILITY ESTIMATES

So far, in this thesis we proposed a normative Bayesian framework for fusing the probabilistic outputs of classifiers while explicitly considering their uncertainty. This is a fundamental task in machine learning and can be applied in various different domains, e.g., for intuitive human-robot interaction, as we showed in Chapters 5 to 7. However, in addition to classifiers, also humans can provide probabilistic forecasts in order to provide their uncertainty, e.g., as a subjective probability estimate of an event's occurrence or the correctness of a statement.

Such subjective probability estimates provided by human experts are of particular importance in many different domains such as finance, business, marketing, politics, engineering, meteorological, ecological, and environmental science, and public health (McAndrew et al., 2021). While statistical models are usually limited in applicability by requiring sufficiently large and complete data sets, human forecasts can overcome this limitation taking advantage of human experience and intuition (McAndrew et al., 2021). The probability estimates can be given either as forecasts of events, e.g., rain probabilities in meteorological science or probabilities for the outcomes of geopolitical events such as elections (Graefe, 2018; Turner et al., 2014), other binary classifications, or the quantification of the experts' confidence on a prediction or the answer to a specific question (Karvetski et al., 2013; Prelec et al., 2017).

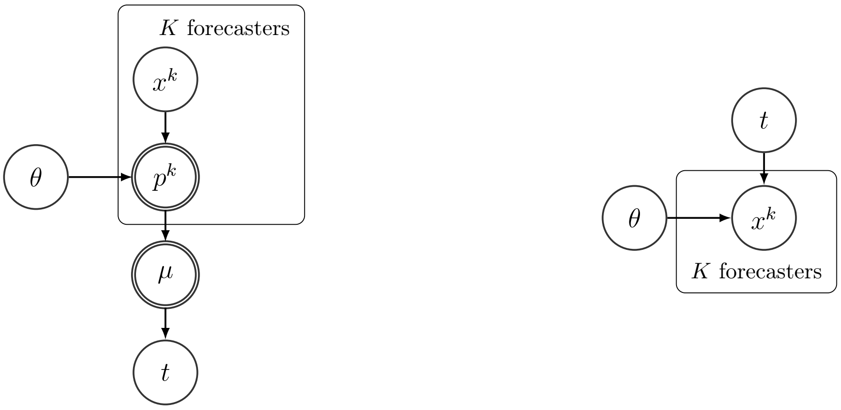
We assume humans to have internal beliefs, which are expressed as the subjective probability estimates they provide. However, we do not necessarily assume that they compute their beliefs by doing Bayesian inference in their heads (Griffiths & Tenenbaum, 2006; M. D. Lee, 2018a). Subjective probability estimates provided by humans are often miscalibrated meaning that they are overconfident or underconfident (Morgan, 2014). A well-calibrated forecaster's probability estimates match the respective relative frequency of occurrence, i.e., $100x\%$ of the answers for which the forecaster predicts probability x are correct (Brenner et al., 1996). In contrast, a miscalibrated forecaster's probability estimate is more (overconfidence) or less (underconfidence) extreme (Morgan, 2014). Measuring the calibration of forecasters, quantifying it with a calibration function, and recalibrating their given probabilities using this calibration function can improve forecasts (Graham, 1996).

Compared to individual forecasts, combining forecasts usually increases performance (Budescu & Chen, 2015; McAndrew et al., 2021; Satopää, 2022; Turner et al., 2014), since a group of forecasters provides more information than a single forecaster (Clemen & Winkler, 1999). A distinction is made between behavioral and mathematical aggregation of forecasts. While behavioral aggregation can be subsumed as a process in which the forecasters negotiate a consensus (Minson et al., 2018; Silver et al., 2021), mathematical aggregation involves mathematical rules or models to combine individual forecasts to an aggregated forecast (Clemen & Winkler, 1999; Hanea et al., 2021; Wilson, 2017). In this work, we focus on mathematical aggregation.

A popular choice for mathematical aggregation of probability estimates are linear opinion pools. While unweighted linear opinion pools, i.e., simple averages, often perform surprisingly well (Turner et al., 2014), numerous weighted linear opinion pools have been designed, because not all opinions are necessarily of the same value. The weights can be selected based on the forecasters' performance (Budescu & Chen, 2015; Cooke, 1991; Hanea et al., 2021), the coherence of their answers (Karvetski et al., 2013), or by the number of cues available to them (Budescu & Rantilla, 2000), or can be optimized for maximum performance (Ranjan & Gneiting, 2010). In the latter approach by Ranjan and Gneiting (2010) they additionally calibrate the weighted linear opinion pool by transforming it with the cumulative distribution function of the beta distribution. To deal with under- and overconfidence of linear opinion pools, also trimmed opinion pools (Grushka-Cockayne et al., 2017) or methods for (anti-) extremizing linear opinion pools (Baron et al., 2014; Lichtendahl Jr et al., 2022) have been introduced. In addition to linear pooling, there are also multiplicative pooling methods, e.g., Independent Opinion Pool (Berger, 1985), which is a renormalized product of the forecasts, or logarithmic or geometric pooling (Berger, 1985; Dietrich & List, 2016), which is a weighted product of forecasts. Both, linear and multiplicative pooling methods can also be used with transformations of the forecasts, e.g., as a probit average (Satopää et al., 2023) or a geometric mean of odds (Satopää et al., 2014).

Although, as seen above, there are already many different methods for combining probability forecasts, according to the review by McAndrew et al. (2021) an open challenge is a "normative theory for how to combine expert opinions into a single consensus distribution" (McAndrew et al., 2021). Therefore, in this work we propose a normative model for combining probability forecasts that models the behavior of individual forecasters and fuses them accordingly.

Several Bayesian statistical models for combining subjective probability estimates have recently been introduced (Hanea et al., 2021; M. D. Lee & Danileiko, 2014; Satopää, 2022; Turner et al., 2014). M. D. Lee and Danileiko (2014) ask the forecasters for percentages or probabilities (e.g., 'What percentage of the world's water is not freshwater?') and thus consider probabilities as ground truth. Their model is unsupervised, so they do not use any historical seed queries from which the forecasters' behavior can be learned. Hanea et al. (2021) and Satopää (2022) consider binary truth values as ground truth, but also work with unsupervised models. In contrast, Turner et al. (2014), who also assume binary ground truth values, propose supervised models for combining probability forecasts. In particular, they investigate whether it is better first to calibrate the given probabilities and then average the recalibrated probability estimates or first to average them and recalibrate the resulting average. For calibration they use the Linear-in-Log-Odds (LLO) calibration function. Turner et al. (2014) compare non-hierarchical models and hierarchical models and models based on probability or log-odds. They evaluated all of these models on one data set consisting of a total of 11551 answers to 176 geopolitical questions, and conclude that first recalibrating and then combining the recalibrated probability estimates results in the best performance. Although their approach uses Bayesian inference to infer the best parameters for different combination rules, the combination rules themselves are not motivated normatively. In Figure 8.1(a) we show a simplified graphical model of one of the



(a) Calibrate Then Average Model

(b) Normative Fusion Model

Figure 8.1: Graphical models of one of the models proposed by Turner et al. (2014), Calibrate Then Average, (a) and an exemplary normative fusion model (b). The models are simplified to the forecasts x^k of K forecasters to only one query with truth value t.

fusion models presented by Turner et al. (2014), Calibrate Then Average. In this model the forecasts x^k of K forecasters are calibrated using a deterministic calibration function with parameters θ (which is the LLO function in their model) to obtain the calibrated forecasts p^k. These calibrated forecasts p^k are then fused to the fused forecast μ using a deterministic fusion method, i.e., averaging. The truth value t is drawn from a Bernoulli distribution with parameter μ. Note that Turner's approach thus models how the truth value t is generated from the forecast data x^k, as it is usually done in a discriminative model.

In contrast, a normative fusion model as we propose it in this work (Figure 8.1(b)) expresses how the true value t generates the forecasts x^k, i.e., the data-generating process, in a generative model. The forecasts x^k are generated from some probability distribution conditioned on the true label t with the respective distribution parameters θ. Thus, after learning the parameters θ from labeled training data, the model represents the forecasting behavior of the forecasters conditioned on t. In particular, as was shown for a model combining classifier outputs (Trick & Rothkopf, 2022, see Chapter 3), it models the forecasters' bias, variance, and uncertainty. In addition, the normative fusion model implicitly calibrates the forecasts without the need for an explicit calibration function. New forecasts x^k are fused using Bayes' rule by inferring the posterior probability of t given the forecasts x^k and the learned model parameters θ. This is normative fusion behavior, because Bayesian inference is normative. Of course, the parameters θ can also be modeled for each forecaster individually as θ^k, which allows modeling each forecaster's individual forecasting behavior conditioned on t.

Lindley (1985) proposed such a normative model for combining probability estimates (nicely explained by Jacobs, 1995). He transforms the probabilities to log-odds and models these log-odds with Gaussian distributions conditioned on the truth value t that indicates whether the respective event occurred or not. For fusing the predictions, the posterior probability of t given the learned Gaussian models and the predictions to be fused needs to be inferred using Bayes' rule.

Another natural way to model the combination of probability estimates normatively is to model the probabilities directly with a beta distribution without the need for any transformation. As far as we know, a Bayesian model for combining human probability estimates using beta distributions has not been worked out in detail yet, but it has been mentioned in an example in the book of Berger (1985). Steyvers et al. (2014) modeled probability forecasts with beta distributions, however, not for fusion or calibration of probability estimates but for evaluating forecast performances using ROC curves. Here, we propose a normative model for combining probability estimates that models the probabilities with a beta distribution conditioned on the true label t of each forecast.

In this vein, we will show that modeling probabilistic forecasts with beta distributions conditioned on the true label t implicitly calibrates them. This calibration function named beta calibration has recently been introduced in a machine learning context (Ji et al., 2020; Kull et al., 2017). The LLO calibration function, used by Turner et al. (2014) and M. D. Lee and Danileiko (2014), can be shown to be a special case of the beta calibration function.

Turner et al. (2014) evaluate their Bayesian fusion models on the The Good Judgment data set by the IARPA Aggregative Contingency Estimation (ACE) System, which is the most popular data set for evaluating forecast aggregation methods and is used for evaluation in many approaches on forecast aggregation (Budescu & Chen, 2015; Hanea et al., 2021; Satopää, 2022; Steyvers et al., 2014; Turner et al., 2014; J. Wang et al., 2021). It includes questions on the probability of future geopolitical events, such as the outcome of elections. Turner et al. (2014) only evaluated on a subset of this data set including only binary events and only similarly framed questions.

This data set used by Turner et al. (2014) consists of 176 events or questions, which are answered by 1290 forecasters. While the high number of events seems to be beneficial for modeling the forecasters' behavior, all forecasters only provided forecasts for a subset of these 176 questions. On average, a forecaster provides only 8.25 answers and 221 (169/122) forecasters only provide 1 (2/3) answer. These low numbers of forecasts per person is unfavorable for modeling the behavior of individual forecasters. Also, modeling their behavior conditioned on whether the event has occurred or not is difficult with this data set since only 37 out of 176 events are positive events that actually occurred.

Several other data sets consisting of probability estimates of multiple forecasters also show similar drawbacks. The ACE-IDEA data set (Hanea et al., 2021) includes forecasts on 155 events, but on average each forecaster only replied to about 19 queries. Other data sets consist of less queries to be predicted or answered (Graefe, 2018; Hanea et al., 2021; Karvetski et al., 2013; Prelec et al., 2017), of which the highest number of queries is about 80 (Prelec et al., 2017). However, 80 answers per forecaster is still a small number for modeling the forecasters' behavior, particularly if we divide the data into training and test sets and model the answers to true/false queries separately. Also, this data set is not publicly available.

Since the available data sets do not provide much information about single forecasters, in this work we publish a new data set consisting of 180 true and false statements, for each of which 85 forecasters provide their confidence on the statement's correctness.

Thus, the contribution of this work is four-fold: First, we present a family of natural normative models for aggregating probability forecasts based on the beta distribution. Second, we introduce beta calibration in expert fusion contexts and show the connection between the widely-used LLO calibration function and the beta distribution. Third, for evaluating our normative models, we provide a new data set consisting of subjective probability forecasts that includes a sufficient number of data points to model the behavior of individual forecasters. Fourth, we systematically evaluate the proposed normative beta fusion models on a data set by Turner et al. (2014) and our new data set, compare them to the models proposed in the work of Turner et al. (2014), and provide some general findings about the respective performance of the considered models regarding several different scores.

The remainder of this chapter is structured as follows. In Section 8.1 we systematically derive different variants of the normative beta fusion model, including hierarchical and non-hierarchical as well as asymmetric and symmetric beta fusion models, and show how the beta calibration function naturally arises in this framework. Section 8.2 presents the new Knowledge Test Confidence data set and an evaluation of the proposed models on this and another data set. In addition, the beta calibration function is evaluated. Finally, in Section 8.3 we discuss the results and outline our model's limitations and possible directions for future work.

8.1 Modeling Probability Estimates with Beta Distributions

In the following, we assume that K human subjects (forecasters) provide probability estimates (forecasts) on N binary queries. These queries can be questions about future events, e.g., 'Will XY happen?', factual statements, or other binary classifications, and must have in common that they can be answered with 'true' or 'false'. The forecasters provide a probability for each item, which can be interpreted as either their belief that the respective event will occur or their confidence in their answer's correctness. In the latter case, a probability of 0 means that the forecaster is 100% certain that the correct answer is 'false', while a probability of 1 indicates that the forecaster is 100% certain the answer is 'true'.

The forecast given by forecaster k for query n is formalized as x_n^k with $n = 1, \ldots, N$, $k = 1, \ldots, K$. The true label t_n for query n, which is the ground truth, can take values 0 or 1, where 0 indicates truth value 'false' and 1 indicates truth value 'true'. We assume the forecasts x_n^k to be conditionally independent given the true label t_n. For a discussion of this assumption, which does not hold in general, we refer to Section 8.3.

The beta distribution is the natural choice for modeling proportions and probabilities, because it is the standard distribution for probabilities in Bayesian statistics and the conjugate prior of the Bernoulli distribution. Therefore, here we model the forecasts x_n^k with a beta distribution conditioned on the true label $t_n = j$, $j \in \{0, 1\}$. We thus assume that humans have some skill to differentiate between true and false queries.

Whereas usually the beta distribution is parameterized with two shape parameters α and β, in the following we will parameterize it alternatively with mean μ and the proportion ρ of the maximum possible variance given μ, which is $\mu(1 - \mu)$. This reparametrization increases the parameters' interpretability and computationally improves the performance of Gibbs sampling by reducing correlations between the variables. Thus, the forecasts x_n^k are modeled with a beta distribution conditioned on the true label $t_n = j$, $j \in \{0, 1\}$, with parameter μ_j^k as the beta distribution's mean and parameter ρ_j^k as the proportion of the beta distribution's maximum variance,

$$x_n^k | t_n = j \sim \text{Beta}'(\mu_j^k, \rho_j^k), \tag{8.1}$$

where $\text{Beta}'(\mu_j^k, \rho_j^k)$ is identical to $\text{Beta}(\alpha_j^k, \beta_j^k)$ with

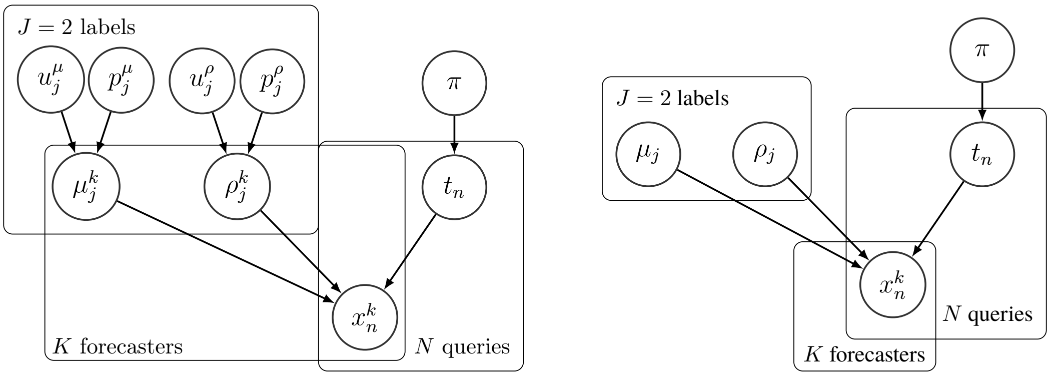
$$\begin{aligned} \alpha_j^k &= \mu_j^k \eta_j^k \\ \beta_j^k &= (1 - \mu_j^k)\eta_j^k \\ \eta_j^k &= \frac{\mu_j^k(1 - \mu_j^k)}{\nu_j^k} - 1 \\ \nu_j^k &= \rho_j^k \mu_j^k (1 - \mu_j^k), \end{aligned} \tag{8.2}$$

and ν_j^k is the beta distribution's variance. The true label t_n is modeled with a Bernoulli distribution with parameter π.

8.1.1 Hierarchical Beta Fusion Model

In many cases, forecasters provide their forecasts to only a subset of the available queries. In order to also be able to accurately model the forecasting behavior of forecasters who only provided a small number of forecasts, we propose the Hierarchical Beta Fusion Model. This allows taking advantage of the statistical properties of hierarchical models and their psychological interpretations. Their statistical properties are increased statistical power since parameter inference is based on more data that share information across groups, and reduced variance of parameter estimates, known as shrinkage (Britten et al., 2021). In addition, they allow modeling of individual differences between forecasters, make it possible to model forecasters who only provided few forecasts, and provide information on the distribution of the forecasters' individual parameters (M. D. Lee, 2018b). In particular, the model's hyperparameters indicate how the forecasters behave on average, how variable their behavior is, and the associated hyperpriors make assumptions about the forecasters' average behavior and variability explicit.

In the Hierarchical Beta Fusion Model, we use a beta prior on the model parameter μ_j^k, parameterized with mean u_j^μ and maximum variance proportion p_j^μ. The proportion of the maximum variance ρ_j^k is also modeled with a beta distribution with mean u_j^ρ and maximum variance proportion p_j^ρ. To avoid values of ρ_j^k too close to 0 or 1 that may get the Gibbs sampler in trouble, we constrained this beta distribution and all other beta priors on maximum variance proportions ρ, p between 0.001 and 0.999. As prior distribution for the proportion π of true queries we chose a uniform beta distribution Beta(1,1). The graphical model of the Hierarchical Beta Fusion Model is shown in Figure 8.2(a). A complete overview over the corresponding modeling distributions and priors, also including the priors of the hyperparameters $u_j^\mu, p_j^\mu, u_j^\rho, p_j^\rho$, is given as



(a) Hierarchical Beta Fusion Model

(b) Non-Hierarchical Beta Fusion Model

Figure 8.2: Graphical models of the hierarchical (a) and non-hierarchical (b) beta fusion models.

$$\begin{aligned} t_n &\sim \text{Bernoulli}(\pi) & \pi &\sim \text{Beta}(1,1) \\ x_n^k | t_n = j &\sim \text{Beta}'(\mu_j^k, \rho_j^k) & u_j^\mu &\sim \text{Beta}(1,1) \\ \mu_j^k &\sim \text{Beta}'(u_j^\mu, p_j^\mu) & p_j^\mu &\sim \text{Beta}(1,1) \\ \rho_j^k &\sim \text{Beta}'(u_j^\rho, p_j^\rho) & u_j^\rho &\sim \text{Beta}(1,1) \\ & & p_j^\rho &\sim \text{Beta}(1,1). \end{aligned} \tag{8.3}$$

Based on labeled training data, the model parameters μ_j^k, ρ_j^k, and π can be inferred using Gibbs sampling (e.g., by specifying this model in JAGS (Plummer, 2003)). Since we assume that human forecasters are on average consistent between different queries, we can use the learned parameters to infer the posterior probability of the true label t_n of new unseen forecasts x_n^k as the fusion result. Inference of t_n can either also be realized using Gibbs sampling, or the posterior probability of t_n can be computed analytically using the closed-form probability density function of the beta distribution:

$$p(t_n = j | \boldsymbol{x_n}, \boldsymbol{\mu_j}, \boldsymbol{\rho_j}, \pi) \propto \pi^j (1-\pi)^{1-j} \prod_{k=1}^{K} \text{Beta}'(x_n^k; \mu_j^k, \rho_j^k). \tag{8.4}$$

8.1.2 Non-Hierarchical Beta Fusion Model

The normative Hierarchical Beta Fusion model represents the forecasting behavior of each forecaster individually, i.e., for each forecaster an individual set of beta parameters μ_j^k, ρ_j^k is learned. In this way, we can model interindividual differences between forecasters, e.g., different levels of expertise, and can exploit these learned properties for fusion by giving more weight to a better-performing forecaster. However, since the related approach by Turner et al. (2014) also compared hierarchical and non-hierarchical versions of their fusion models, we also compare our normative Hierarchical Beta Fusion Model to a non-hierarchical version of the model, which assumes exchangeable forecasters that behave similarly. In this non-hierarchical Beta Fusion Model, we model the forecasts x_n^k with the same beta distributions for all forecasters with shared parameters μ_j and ρ_j for $k = 1, \ldots, K$ and $j \in \{0, 1\}$. Thus, we learn only two beta distributions for all forecasters: The beta distribution with mean μ_0 and maximum variance proportion ρ_0 models the forecasting behavior of all forecasters for false queries, the beta distribution with mean μ_1 and maximum variance proportion ρ_1 models their forecasting behavior for true queries. The priors on μ_j and ρ_j are uniform distributions Beta(1,1). As for the Hierarchical Beta Fusion Model in Section 8.1.1 the prior for proportion π is an uninformative beta prior Beta(1,1). We illustrate the graphical model of the non-hierarchical Beta Fusion Model in Figure 8.2(b). All corresponding modeling distributions and priors can be summarized as

$$\begin{aligned} t_n &\sim \text{Bernoulli}(\pi) \\ x_n^k | t_n = j &\sim \text{Beta}'(\mu_j, \rho_j) \end{aligned} \qquad \begin{aligned} \mu_j &\sim \text{Beta}(1,1) \\ \rho_j &\sim \text{Beta}(1,1) \\ \pi &\sim \text{Beta}(1,1). \end{aligned} \tag{8.5}$$

Again, given labeled training data we can infer the model parameters μ_j, ρ_j, and π using Gibbs sampling, and the fused result for some unseen forecasts x_n^k of multiple forecasters is the posterior probability over t_n given the learned model parameters and forecasts x_n^k. Using equation (8.4) as for the Hierarchical Beta Fusion Model and dropping index k for μ and ρ allows the analytical computation of the posterior over t_n.

8.1.3 Beta Calibration

A forecaster is well-calibrated if their probability estimate matches the respective relative frequency of occurrence, i.e., if $100x\%$ of the statements to which the forecaster assigns a probability of x are true or $100x\%$ of the events to which the forecaster assigns a probability of x occur (Brenner et al., 1996). The calibration of a human forecaster can be measured empirically by binning the provided probability estimates and computing the proportions of true events for each bin. It is customary to illustrate this relationship with the so-called calibration curve, which plots the proportions of true events as a function of the human forecast probabilities. If the resulting calibration curve is the identity function, the forecaster is perfectly calibrated. If not, a function can be fitted to the empirical calibration curve. This calibration function can then also be used to recalibrate probability estimates, i.e., to correct for overconfident or underconfident judgments (Turner et al., 2014).

While various different functions can serve as calibration functions, the Linear-in-Log-Odds (LLO) function is frequently used. For example, M. D. Lee and Danileiko (2014) and Turner et al. (2014) explicitly include the LLO function in their Bayesian fusion models for calibrating the provided forecasts. However, by modeling the provided forecasts with a probability distribution, one can also calibrate them implicitly. This means that given a probabilistic generative fusion model, as we provide in this work, the calibration function is not chosen empirically, but instead the normative calibration function for the respective model can be derived.

With the Hierarchical Beta Fusion Model we model forecasts x_n^k of forecaster k conditioned on the true label t_n with a beta distribution

$$p(x_n^k = x | t_n = j) = \frac{x^{\alpha_j^k - 1}(1-x)^{\beta_j^k - 1}}{\mathrm{B}(\alpha_j^k, \beta_j^k)}. \tag{8.6}$$

The corresponding calibration function, which is called the beta calibration function, can be derived using Bayes' rule

$$\begin{aligned} \mathrm{BC}(x) = p(t_n = 1 | x_n^k = x) &= \frac{\frac{x^{\alpha_1^k-1}(1-x)^{\beta_1^k-1}}{\mathrm{B}(\alpha_1^k,\beta_1^k)}\pi}{\frac{x^{\alpha_1^k-1}(1-x)^{\beta_1^k-1}}{\mathrm{B}(\alpha_1^k,\beta_1^k)}\pi + \frac{x^{\alpha_0^k-1}(1-x)^{\beta_0^k-1}}{\mathrm{B}(\alpha_0^k,\beta_0^k)}(1-\pi)} \\ &= \frac{1}{1 + \frac{\mathrm{B}(\alpha_1^k,\beta_1^k)}{\mathrm{B}(\alpha_0^k,\beta_0^k)} \frac{(1-x)^{\beta_0^k-\beta_1^k}}{x^{\alpha_1^k-\alpha_0^k}} \frac{1-\pi}{\pi}} \end{aligned} \tag{8.7}$$

with $\pi = p(t_n = 1)$ as introduced above. The function in (8.7) has been first introduced by Kull et al. (2017) in the context of machine learning for calibrating the probabilistic outputs of classification algorithms.

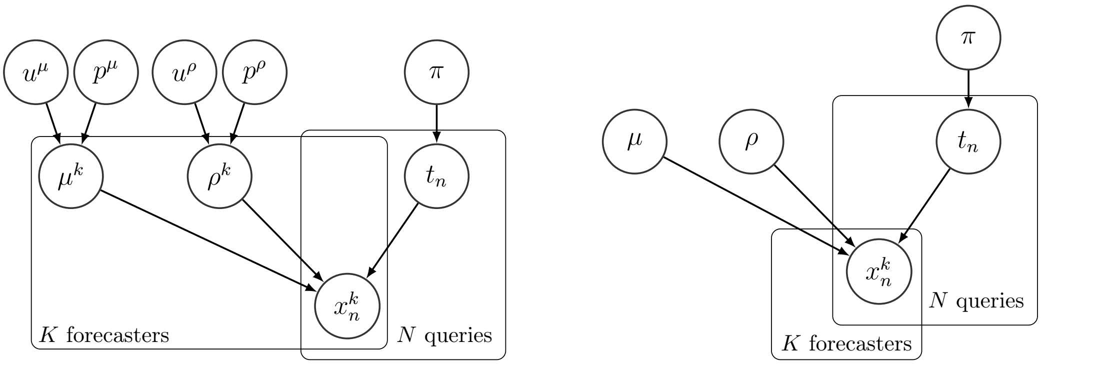
Interestingly, the Linear-in-Log-Odds (LLO) calibration function used by Turner et al. (2014) and M. D. Lee and Danileiko (2014) can be derived as a special case of the beta calibration function (Kull et al., 2017). If we constrain the beta distributions to be symmetric around $\frac{1}{2}$, i.e., $\alpha = \alpha_0^k = \beta_1^k$ and $\beta = \beta_0^k = \alpha_1^k$, the resulting calibration function is

$$\begin{aligned} \mathrm{LLO}(x) = p(t_n = 1 | x_n^k = x) &= \frac{\frac{\pi}{1-\pi}x^{\beta-\alpha}}{\frac{\pi}{1-\pi}x^{\beta-\alpha} + (1-x)^{\beta-\alpha}} \\ &= \frac{\delta x^\gamma}{\delta x^\gamma + (1-x)^\gamma} \end{aligned} \tag{8.8}$$

with $\delta = \frac{\pi}{1-\pi}$ and $\gamma = \beta - \alpha$. Beta calibration in (8.7) is more flexible than LLO calibration in (8.8) because it does not assume symmetric beta distributions for $t_n = 0$ and $t_n = 1$. Thus, beta calibration can consider that the forecasters' behavior might be different for different truth values t_n. In contrast, LLO calibration assumes symmetric beta distributions for $t_n = 0$ and $t_n = 1$ and therefore symmetric forecasting behavior for true and false queries, which might not hold for real forecasts. A more detailed comparison of beta calibration and LLO calibration including example calibration functions can be found in Section 8.2.5.

8.1.4 Hierarchical Symmetric Beta Fusion Model

Since modeling the forecasts with symmetric beta distributions results in calibrating them with the LLO calibration function (see Section 8.1.3), we are interested in comparing the original beta fusion model using asymmetric beta distributions to a symmetric beta fusion model using symmetric beta distributions, which assumes humans to show symmetric forecasting behavior given true or false queries. In the Hierarchical Symmetric Beta Fusion Model we thus model forecasts x_n^k with two symmetric beta distributions with parameters $\mu_0^k = \mu^k$, $\rho_0^k = \rho^k$ and $\mu_1^k = 1 - \mu^k$, $\rho_1^k = \rho^k$. We set a beta prior on μ^k with mean u^μ



(a) Hierarchical Symmetric Beta Fusion Model (b) Non-Hierarchical Symmetric Beta Fusion Model

Figure 8.3: Graphical models of the hierarchical (a) and non-hierarchical (b) symmetric beta fusion models.

and maximum variance proportion p^μ and model ρ^k with a beta distribution with mean u^ρ and maximum variance proportion p^ρ. Similar to the asymmetric beta fusion models, the proportion π of true queries is modeled with an uninformed prior Beta(1,1). Figure 8.3(a) shows the graphical model of the Hierarchical Symmetric Beta Fusion Model. All modeling distributions and priors are given as

$$\begin{aligned} t_n &\sim \text{Bernoulli}(\pi) & \pi &\sim \text{Beta}(1,1) \\ x_n^k|t_n = 0 &\sim \text{Beta}'(\mu^k, \rho^k) & u^\mu &\sim \text{Beta}(1,1) \\ x_n^k|t_n = 1 &\sim \text{Beta}'(1-\mu^k, \rho^k) & p^\mu &\sim \text{Beta}(1,1) \\ \mu^k &\sim \text{Beta}'(u^\mu, p^\mu) & u^\rho &\sim \text{Beta}(1,1) \\ \rho^k &\sim \text{Beta}'(u^\rho, p^\rho) & p^\rho &\sim \text{Beta}(1,1). \end{aligned} \tag{8.9}$$

The model parameters μ^k, ρ^k, and π are estimated from labeled training data using Gibbs sampling. For fusing unseen forecasts x_n^k the posterior probability of their true label t_n can be computed analytically given the learned model parameters:

$$p(t_n = 0|\boldsymbol{x_n}, \boldsymbol{\mu}, \boldsymbol{\rho}, \pi) \propto (1-\pi) \prod_{k=1}^{K} \text{Beta}'(x_n^k; \mu^k, \rho^k) \tag{8.10}$$

$$p(t_n = 1|\boldsymbol{x_n}, \boldsymbol{\mu}, \boldsymbol{\rho}, \pi) \propto \pi \prod_{k=1}^{K} \text{Beta}'(x_n^k; 1-\mu^k, \rho^k). \tag{8.11}$$

8.1.5 Non-Hierarchical Symmetric Beta Fusion Model

As for the asymmetric beta fusion model, we also examine a non-hierarchical version of the symmetric beta fusion model. In the non-hierarchical Symmetric Beta Fusion Model all forecasters' forecasts x_n^k are modeled with the same two symmetric beta distributions for $t_n = 0$ and $t_n = 1$ with parameters $\mu_0 = \mu$, $\rho_0 = \rho$ and $\mu_1 = 1-\mu$, $\rho_1 = \rho$. The priors for μ and ρ are uniform distributions Beta(1,1). As in all models, we set an uninformative prior

Beta(1,1) on the proportion of true queries π. The graphical model of the non-hierarchical Symmetric Beta Fusion Model is presented in Figure 8.3(b). An overview of all modeling distributions and priors is given as

$$\begin{aligned} t_n &\sim \text{Bernoulli}(\pi) & \mu &\sim \text{Beta}(1,1) \\ x_n^k | t_n = 0 &\sim \text{Beta}'(\mu, \rho) & \rho &\sim \text{Beta}(1,1) \\ x_n^k | t_n = 1 &\sim \text{Beta}'(1-\mu, \rho) & \pi &\sim \text{Beta}(1,1). \end{aligned} \tag{8.12}$$

The model parameters μ, ρ, and π are estimated from labeled training data using Gibbs sampling. For fusing unseen forecasts x_n^k the posterior probability of their true label t_n can be computed analytically using equations (8.10) and (8.11) with dropping index k for parameters μ, ρ.

8.2 Evaluation

We evaluated the four proposed Bayesian models on two data sets consisting of forecasts provided by human subjects, the Turner data set and a new data set that we collected (Section 8.2.1). Using leave-one-out cross-validation (Section 8.2.2) we compare their Brier scores, 0-1 losses, and mean absolute errors as performance measures (Section 8.2.3). In addition, we also compared the performances of the proposed fusion models to the fusion models presented by Turner et al. (2014). For implementing them, we adopted the JAGS code they provided for their models and all specifications given with respect to sampling, e.g., the number of samples and initial values. Motivated by the results of the models' comparison presented in Section 8.2.4, we compare beta calibration and LLO calibration in Section 8.2.5.

8.2.1 Data Sets

We evaluated the performances of the proposed Bayesian fusion models and the reference models by Turner et al. (2014) on two data sets, namely the Turner data set (Section 8.2.1.1), which is the data set Turner et al. (2014) used for evaluating their fusion models, and our new data set (Section 8.2.1.2).

8.2.1.1 Turner Data Set

The Turner data set[1] is a subset of the The Good Judgment data set,[2] containing 176 geopolitical statements in the form of 'Will event X happen by date Y?', e.g., 'There will be a military coup in Venezuela in 2011'. All statements are binary, so they are either true or false, but at the time of data collection, all events were unresolved, so their outcome could not be known yet. After completion of the study, 37 of 176 statements turned out to be true, while the remaining 139 statements resolved as false.

Human subjects could reply to the given items through a web page. They provided their estimate of the probability that the respective statement will resolve to true for as many

[1] `https://webfiles.uci.edu/msteyver/codeanddata/forecastingdata.csv`

[2] `https://dataverse.harvard.edu/dataverse/gjp`

statements as they wanted. The provided probabilities are between 0 and 1, accurate to 2 decimal places. To avoid problems with estimates of exactly 0 or 1, we preprocessed these estimates and changed them to 0.001 and 0.999 respectively, as also done in the work of Turner et al. (2014).[3] 1290 subjects provided a total number of 11551 probability estimates. The maximum number of replies per subject is 127, on average a subject provided 8.25 probability estimates.

8.2.1.2 Knowledge Test Confidence Data Set

In this work, we publish a new data set called the Knowledge Test Confidence (KTeC) data set.[4] It consists of the confidence judgments of 85 forecasters to 180 knowledge statements of which 90 statements are true and 90 statements are false. There are easy statements, e.g., 'Elephants are mammals', and hard statements, at least for our participant pool, e.g., 'Port Moresby is the capital of Papua New Guinea'.

The data were collected in a probabilistic modeling class at the University of Osnabrück and were part of the lessons on proper scoring rules. The students attending this class were asked to generate statements that are easy to understand, do not contain negations, and cover the whole range from easy to hard statements. They were told that for an easy query 80-90% of their peers should know the statement's truth value, for hard queries only 60-70%. Most of the resulting statements test general knowledge, some are specific to student life in Osnabrück, and some were deliberately designed as trick questions (e.g., 'The official language of the United States is English'). The students who provided the statements and other students in a couple of following years voluntarily and completely anonymously provided their confidence on the truth of each statement through an online questionnaire. A confidence of 0 indicates that a subject is convinced that the statement is wrong, whereas a confidence of 1 indicates a strong belief that the statement is correct. Subjects could provide their confidences not on a continuous scale, but in 11 steps of 0.1. As for the Turner data set, for the following evaluations we again preprocessed 0 to 0.001 and 1 to 0.999.[5]

85 students provided a total number of 15300 probability estimates. Thus, each subject replied to all 180 statements.

8.2.2 Cross-Validation

In order to evaluate the different models on the data sets described above (Section 8.2.1) we split the data into training and test sets using leave-one-out (LOO) cross-validation. While Turner et al. (2014) evaluated with 10-fold cross-validation, we preferred leave-one-out cross-validation over k-fold cross-validation since it allows training the model on almost the entire data set, which reduces the evaluation bias. Also, it comes with zero randomness in the partitioning of the data and is therefore straightforwardly reproducible.

[3] We also evaluated different score corrections, namely (0.01,0.99) and (0.025,0.975), which did not change the results significantly.

[4] The data set is provided at `https://osf.io/ae25w/`.

[5] We also evaluated different score corrections, namely (0.01,0.99) and (0.025,0.975), which did not change the results significantly.

If a data set consists of M queries, we obtain M LOO training sets, each including $M-1$ data points. For each of the resulting M training sets, we inferred the posterior distribution of each model's parameters using Gibbs sampling. We implement Gibbs Sampling for inference using JAGS (Plummer, 2003). For fitting the model parameters given labeled training data we ran two parallel chains, each consisting of 1000 samples with a burn-in of 1000 samples.

For fusing the one example in the test set, one could now use the means of the obtained posterior distributions as point estimates for the parameters and compute the posterior over the true label t_n analytically given these point estimates. However, in order to consider the uncertainty of our parameter estimates, we computed the posterior over t_n analytically for each sample of the model parameters' posterior distribution, as for example in (8.4), and averaged all obtained posteriors to the final fused forecast.

8.2.3 Performance Measures

As performance measures, we consider Brier score, 0-1 loss, and mean absolute error. The Brier score (Brier et al., 1950) is a popular metric for quantifying human forecast performance, used by e.g., Karvetski et al. (2013), Turner et al. (2014), Hanea et al. (2021), and Satopää (2022). It is a strictly proper scoring rule (A. H. Murphy, 1973), meaning that it is optimized when people report their true beliefs of the probability instead of intentionally providing more or less extreme probabilities. The Brier score is defined as the mean squared error between the predicted probabilities x_n and the true labels t_n,

$$\mathrm{BS} = \frac{1}{N} \sum_{n=1}^{N} (x_n - t_n)^2. \tag{8.13}$$

Thus, the best attainable Brier score is 0 and the worst is 1. Interestingly, if a forecaster always provides 0.5 as her estimate, the resulting Brier score will be 0.25. Thus, a model should at least achieve a Brier score below 0.25.

It is controversial whether the Brier score is a suitable metric for comparing the performances of different forecasting systems, since as a strictly proper scoring rule it was originally developed in order to measure if forecasters report their true beliefs, not to compare different forecasters (Steyvers et al., 2014). Also, it can be dominated by outliers (Canbek et al., 2022), though not as much as e.g., the log-loss. Still, it is commonly used for comparing forecasters' performances (Baron et al., 2014; Hanea et al., 2021; Karvetski et al., 2013; Ranjan & Gneiting, 2010; Satopää, 2022; Turner et al., 2014), so we report it here, too.

0-1 loss describes the proportion of incorrect forecasts to the total number of all forecasts,

$$\mathrm{L}_{01} = \frac{1}{N} \sum_{n=1}^{N} \begin{cases} 1 & \text{if} \quad |t_n - x_n| \geq 0.5 \\ 0 & \text{else} \end{cases}, \tag{8.14}$$

and thus ranges between 0 and 1, with lower values indicating better performances. In comparison to the Brier score, the 0-1 loss is more easily interpretable and directly compares the forecasters' performances. However, it disregards their uncertainty by considering a forecast as correct, if its corresponding probability is closer to the true label t_n of the respective query.

To overcome the limitations of Brier score and 0-1 loss, we additionally evaluate the different fusion methods in terms of mean absolute error (Canbek et al., 2022; Ferri et al., 2009). Mean absolute error (MAE) measures the absolute difference between the forecasted probability and the true label:

$$\mathrm{MAE} = \frac{1}{N} \sum_{n=1}^{N} |x_n - t_n|. \tag{8.15}$$

Similar to the Brier score (mean squared error) and 0-1 loss it ranges between 0 and 1, with lower values indicating higher performance. However, in contrast to the Brier score, which can be dominated by outliers, MAE is more robust to outliers (Canbek et al., 2022). Also, it is straightforwardly interpretable and a more natural and intuitive metric for comparing different fusion models without disregarding their uncertainty. However, note that MAE is an improper scoring rule (Buja et al., 2005), so it incentivizes overconfident forecasts.

8.2.4 Model Performances

On both data sets described in Section 8.2.1 we evaluate the four Bayesian fusion models introduced in Section 8.1 in terms of Brier score, 0-1 loss, and mean absolute error (MAE). In addition, we compare our fusion methods' performances to the models by Turner et al. (2014). Turner et al. (2014) present several Bayesian fusion models, which explicitly consider the calibration of forecasts with the LLO calibration function. Their key question is whether it is better first to average the forecasts or first to calibrate them. Hence, the proposed models are three non-hierarchical models, Average Then Calibrate (ATC), Calibrate Then Average (CTA), Calibrate Then Average using Log-Odds (CTALO), and two hierarchical models, Hierarchical Calibrate Then Average (HCTA), and Hierarchical Calibrate Then Average on Log-Odds (HCTALO).[6] For reference, they also evaluate the performance of Unweighted Linear Opinion Pool (ULINOP) as a baseline. Since ULINOP is known to be biased towards 0.5 (Baron et al., 2014), we additionally evaluate the performance of Probit Average (PAVG) (Satopää et al., 2023) as another benchmark. For PAVG we first transform all forecasts with probit, then average the transformed forecasts, and finally transform this average back to probability score. Here, we investigate how a normative model that implicitly calibrates the forecasts by modeling them with beta distributions will perform relative to all these models. In particular, we investigate whether the normative approach increases the performance of the fused forecast.

Figure 8.4 shows the means and standard errors of the mean of Brier score, 0-1 loss, and mean absolute error (MAE) on the Turner data set (Section 8.2.1.1). The beta fusion models introduced in this work are abbreviated as HB (Hierarchical Beta Fusion Model), B (non-hierarchical Beta Fusion Model), HSB (Hierarchical Symmetric Beta Fusion Model), and SB (non-hierarchical Symmetric Beta Fusion Model). According to Brier score, the three non-hierarchical Turner models ATC, CTA, and CTALO perform best with BS $\approx$ 0.125. These results are different to the results reported by Turner et al. (2014), which were obtained using 10-fold cross-validation and favored the HCTALO model. The best beta model is HSB with BS = 0.151, which performs comparably to HB and the hierarchical

[6] In our evaluations, we used the JAGS code provided in the work of Turner et al. (2014). However, note that in their work the implementation of HCTALO is significantly different from the implementation of HCTA.

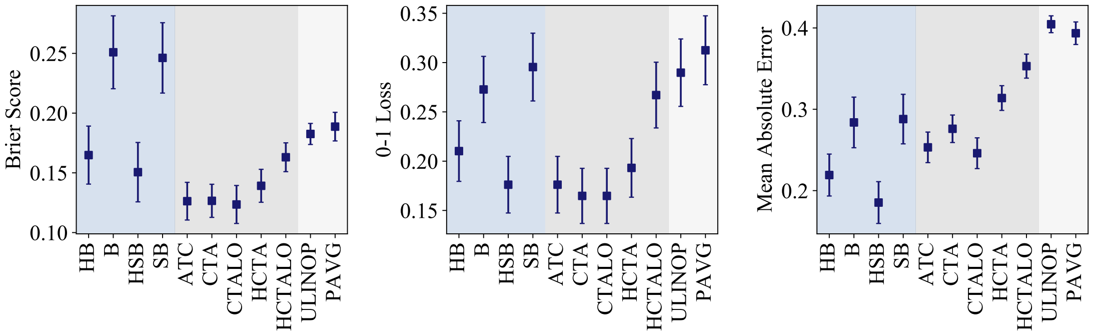


Figure 8.4: Model performances on the Turner data set according to Brier score, 0-1 loss, and mean absolute error. We compare the scores' means and standard errors of the mean of our beta fusion models, the Hierarchical Beta Fusion Model (HB), the non-hierarchical Beta Fusion Model (B), the Hierarchical Symmetric Beta Fusion Model (HSB), and the non-hierarchical Symmetric Beta Fusion Model (SB), the models by Turner et al. (2014), Average Then Calibrate (ATC), Calibrate Then Average (CTA), Calibrate Then Average using Log-Odds (CTALO), Hierarchical Calibrate Then Average (HCTA), and Hierarchical Calibrate Then Average on Log-Odds (HCTALO), and the two baseline methods Unweighted Linear Opinion Pool (ULINOP) and Probit Average (PAVG).

Turner models HCTA and HCTALO. The non-hierarchical beta models B and SB perform worst with Brier scores of about 0.25, which is close to the performance of a forecaster that always forecasts 0.5. As per 0-1 loss, the ranking of the models is different. HSB performs similarly to ATC, CTA, CTALO, and HCTA with $\mathrm{L}_{01} = 0.176$, HB is approaching ($\mathrm{L}_{01} = 0.21$). The hierarchical Turner model HCTALO ($\mathrm{L}_{01} = 0.267$) performs clearly worse than both hierarchical beta models. According to MAE, both hierarchical beta models HB (MAE = 0.219) and HSB (MAE = 0.185) perform best. All non-hierarchical models, Turner and beta models, perform similarly (MAE ≈ 0.25), while the hierarchical Turner models perform worst, similarly to ULINOP and PAVG with MAE ≈ 0.4. Consistently over all performance measures, HSB outperforms HB. Also, the hierarchical beta models outperform the non-hierarchical beta models, while the non-hierarchical Turner models outperform the hierarchical Turner models.

In Figure 8.5 we compare the performances of beta and Turner fusion models on the newly introduced Knowledge Test Confidence (KTeC) data set. Compared to the results on the Turner data set, the differences between the models' performances are generally smaller over all three performance measures Brier score, 0-1 loss, and MAE. Based on Brier score HCTALO performs best with BS = 0.125, but CTALO, ATC, HSB, and PAVG perform quite similarly. HB and HCTA perform worst with Brier scores of BS = 0.179 and BS = 0.185. However, as per 0-1 loss CTA performs worst, and HSB and HCTALO are performing best with $\mathrm{L}_{01} = 0.156$ and $\mathrm{L}_{01} = 0.15$. According to MAE, all beta fusion models clearly outperform the Turner models and perform quite comparably. Still, HSB is again the best performing beta fusion model with an MAE of 0.153.

In the evaluations shown so far we combine the forecasts of 1290 forecasters for the Turner data set and 85 forecasters for the KTeC data set. For both data sets the Hierarchical Symmetric Beta Fusion model (HSB) is best or among the best models according to 0-1 loss or MAE. However, according to Brier score some Turner models outperform HSB on both data sets, which might indicate that HSB or the beta models in general are overconfident.

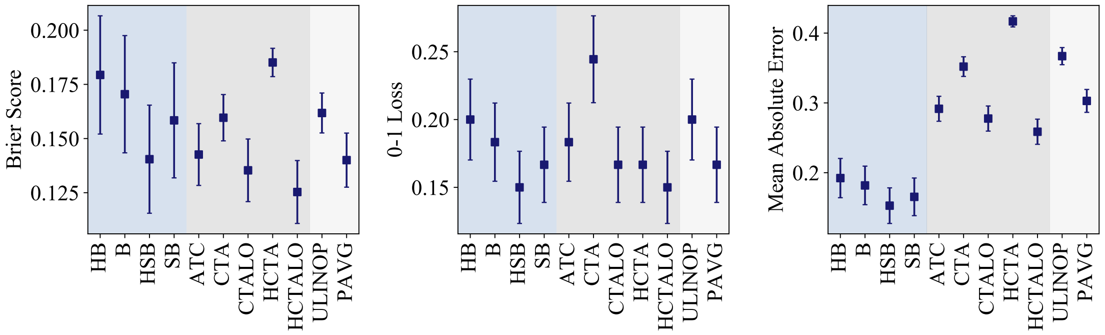


Figure 8.5: Model performances on the Knowledge Test Confidence (KTeC) data set according to Brier score, 0-1 loss, and mean absolute error. We compare the scores' means and standard errors of the mean of our beta fusion models, the Hierarchical Beta Fusion Model (HB), the non-hierarchical Beta Fusion Model (B), the Hierarchical Symmetric Beta Fusion Model (HSB), and the non-hierarchical Symmetric Beta Fusion Model (SB), the models by Turner et al. (2014), Average Then Calibrate (ATC), Calibrate Then Average (CTA), Calibrate Then Average using Log-Odds (CTALO), Hierarchical Calibrate Then Average (HCTA), and Hierarchical Calibrate Then Average on Log-Odds (HCTALO), and the two baseline methods Unweighted Linear Opinion Pool (ULINOP) and Probit Average (PAVG).

To investigate this, we also consider two subsets of the two data sets that contain fewer forecasters since fusing a lower number of forecasters should attenuate overconfidence. For the Turner data set the subset of forecasters must be chosen with care, since all forecasters only provided forecasts for only a subset of queries. To be able to evaluate fusion methods, we must guarantee that all queries are answered by at least two forecasters, which we can then fuse. In order to do so, we selected the 20 forecasters providing the most forecasts. The results on the reduced Turner data set are shown in Figure 8.6. As for the full Turner data set, the hierarchical beta fusion models outperform the non-hierarchical ones, achieve a 0-1 loss ($\mathrm{L}_{01} = 0.171$ for HB and $\mathrm{L}_{01} = 0.165$ for HSB) comparable to the best Turner model ATC ($\mathrm{L}_{01} = 0.176$), and outperform all other models clearly according to MAE with $\mathrm{MAE} \approx 0.2$. In addition, and in contrast to the full Turner data set, here HB ($\mathrm{BS} = 0.129$) and HSB ($\mathrm{BS} = 0.128$) also perform comparably to the best Turner model ATC ($\mathrm{BS} = 0.132$) regarding Brier score. Thus, for the reduced Turner data set the hierarchical beta fusion models HB and HSB are among the best fusion models for all performance measures.

For the KTeC data set it is more straightforward to create a reduced data set since all forecasters replied to all queries. Therefore, we simply selected the first 10 forecasters as a subset. In Figure 8.7 we see that similar to the reduced Turner data set, on the reduced KTeC data set HSB is the best model according to all three performance measures ($\mathrm{BS} = 0.124, \mathrm{L}_{01} = 0.15, \mathrm{MAE} = 0.192$). As per Brier score and 0-1 loss, HCTALO ($\mathrm{BS} = 0.132, \mathrm{L}_{01} = 0.156$) performs comparably, but regarding MAE, again all beta fusion models clearly outperform all Turner models. Similar to the full KTeC data set, the hierarchical symmetric beta model (HSB) generally outperforms the asymmetric one (HB) and the hierarchical beta models achieve better scores than the non-hierarchical ones.

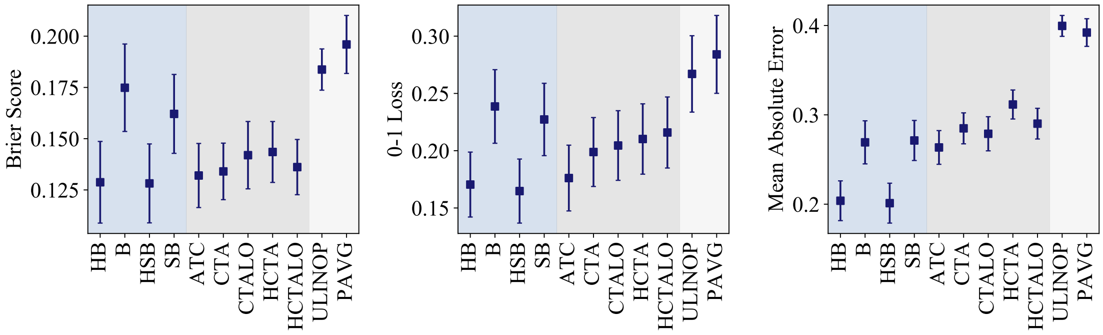


Figure 8.6: Model performances on the reduced Turner data set consisting of a subset of the 20 forecasters of the Turner data set that provided the most forecasts. We compare the means and standard errors of the mean of Brier scores, 0-1 losses, and mean absolute errors of our beta fusion models, the Hierarchical Beta Fusion Model (HB), the non-hierarchical Beta Fusion Model (B), the Hierarchical Symmetric Beta Fusion Model (HSB), and the non-hierarchical Symmetric Beta Fusion Model (SB), the models by Turner et al. (2014), Average Then Calibrate (ATC), Calibrate Then Average (CTA), Calibrate Then Average using Log-Odds (CTALO), Hierarchical Calibrate Then Average (HCTA), and Hierarchical Calibrate Then Average on Log-Odds (HCTALO), and the two baseline methods Unweighted Linear Opinion Pool (ULINOP) and Probit Average (PAVG).

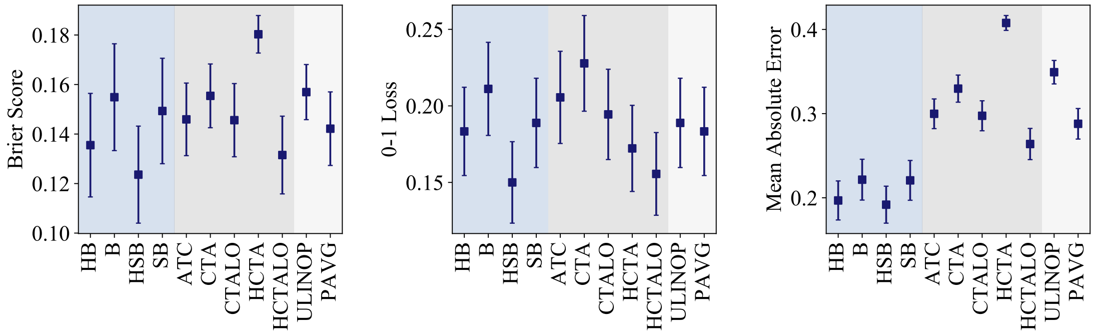


Figure 8.7: Model performances on the reduced Knowledge Test Confidence (KTeC) data set consisting of a subset of the first 10 forecasters of KTeC data set. We compare the means and standard errors of the mean of Brier scores, 0-1 losses, and mean absolute errors of our beta fusion models, the Hierarchical Beta Fusion Model (HB), the non-hierarchical Beta Fusion Model (B), the Hierarchical Symmetric Beta Fusion Model (HSB), and the non-hierarchical Symmetric Beta Fusion Model (SB), the models by Turner et al. (2014), Average Then Calibrate (ATC), Calibrate Then Average (CTA), Calibrate Then Average using Log-Odds (CTALO), Hierarchical Calibrate Then Average (HCTA), and Hierarchical Calibrate Then Average on Log-Odds (HCTALO), and the two baseline methods Unweighted Linear Opinion Pool (ULINOP) and Probit Average (PAVG).

8.2.5 Beta Calibration vs LLO

The results presented in Section 8.2.4 show that the Hierarchical Symmetric Beta Fusion Model (HSB) outperforms the Hierarchical Beta Fusion Model (HB). While this outcome is rather unexpected, since HSB constrains the modeling beta distributions to be symmetric and is therefore less expressive than HB, we can explain it with the calibration functions that are implied by modeling the forecasts with symmetric or asymmetric beta distributions. As shown in Section 8.1.3, by modeling the forecasts with beta distributions conditioned on the true label, we implicitly calibrate them using the beta calibration function (8.7). If the beta distributions are symmetric, the beta calibration function reduces to the LLO calibration function (8.8).

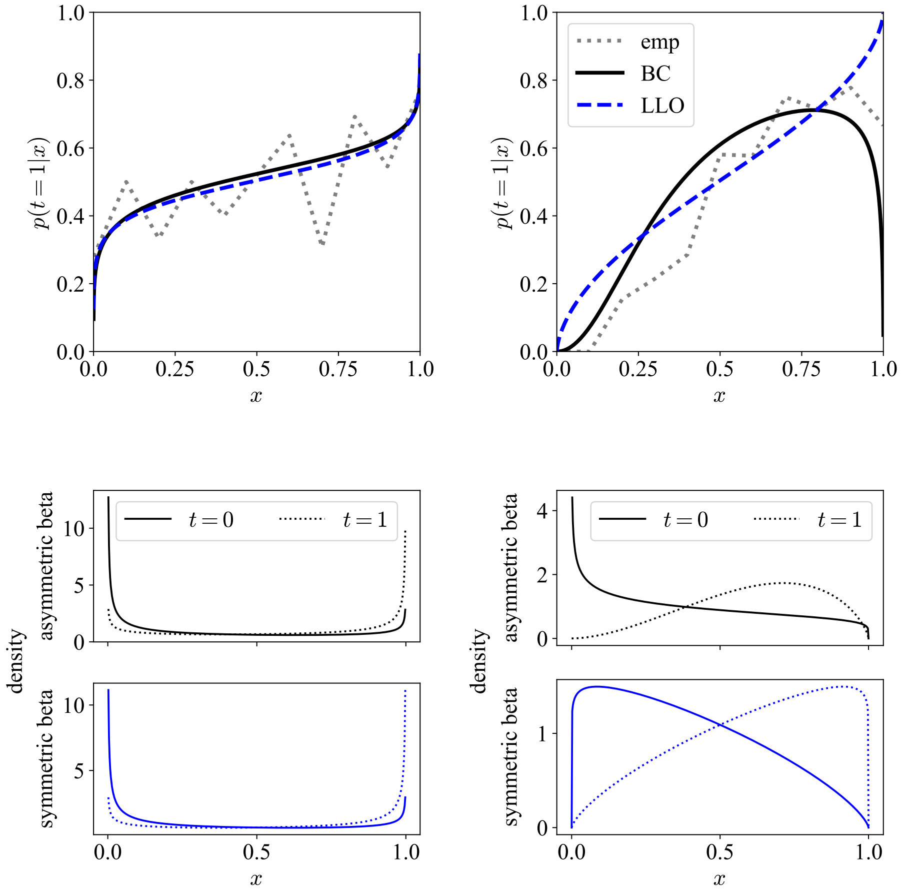
In most cases the LLO and beta calibration function do not differ significantly. However, there are special cases, in which the beta calibration deviates drastically from the LLO calibration function. From (8.8) we can directly see that $\mathrm{LLO}(0) = 0$ and $\mathrm{LLO}(1) = 1$ if $\gamma = \beta - \alpha > 0$ with $\alpha = \alpha_0^k = \beta_1^k$ and $\beta = \beta_0^k = \alpha_1^k$. The latter condition should hold for the most forecasters, since otherwise they would be biased towards always predicting the wrong answer.

In contrast, also for such unbiased forecasters, for which $\alpha_0^k < \beta_0^k$ and $\alpha_1^k > \beta_1^k$, the beta calibration function $\mathrm{BC}(x)$ is not always defined at $x = 0$ and $x = 1$, depending on the beta parameters. In particular, looking at (8.7) we see that

$$\begin{aligned} &\lim_{x\to 0} \mathrm{BC}(x) = 1 \quad \text{if} \quad \alpha_1^k - \alpha_0^k < 0 \quad \text{and} \\ &\lim_{x\to 1} \mathrm{BC}(x) = 0 \quad \text{if} \quad \beta_0^k - \beta_1^k < 0. \end{aligned} \tag{8.16}$$

In Figure 8.8 we show the calibration curves of two exemplary forecasters from the Knowledge Test Confidence (KTeC) data set (top row) together with the densities of the respective beta distributions for $t = 0$ and $t = 1$ when assuming asymmetric or symmetric beta distributions (bottom row). The respective parameters of the beta distributions that model their forecasts and define their calibration curves are taken from training split 1 of the cross-validation done for HB and HSB described in Section 8.2.2. Figure 8.8(a) shows the calibration curves and corresponding beta distributions of forecaster 57 with parameters $\alpha_0^{57} = 0.41, \beta_0^{57} = 0.65, \alpha_1^{57} = 0.67, \beta_1^{57} = 0.47$ for the asymmetric Hierarchical Beta Fusion Model (HB) or the beta calibration function respectively and parameters $\alpha_0^{57} = \beta_1^{57} = 0.44, \beta_0^{57} = \alpha_1^{57} = 0.65$ for the Hierarchical Symmetric Beta Fusion Model (HSB) and the LLO calibration function. We can see that the learned beta distributions for HB (asymmetric) and HSB (symmetric) look very similar. Also, LLO and beta calibration curves look very similar since $\alpha_1^{57} - \alpha_0^{57} > 0$ and $\beta_0^{57} - \beta_1^{57} > 0$.

In contrast, in Figure 8.8(b) we see that for forecaster 46, the beta calibration curve looks very different from the LLO curve. Since the forecaster is modeled with parameters $\alpha_0^{46} = 0.73, \beta_0^{46} = 1.16, \alpha_1^{46} = 2.81, \beta_1^{46} = 1.73$ for HB or beta calibration and $\alpha_0^{46} = \beta_1^{46} = 1.07, \beta_0^{46} = \alpha_1^{46} = 1.72$ for HSB or LLO, $\beta_0^{46} - \beta_1^{46} < 0$ and $\mathrm{BC}(x)$ tends to 0 for $x \to 1$. This can also be seen in the corresponding beta distributions in the bottom of Figure 8.8(b). If we assume asymmetric beta distributions, at $x = 1$ the probability density for $t = 0$ is higher than the probability density for $t = 1$, leading to a beta calibration function that tends to 0 for $x \to 1$. If we assume symmetric beta distributions, this does not happen. In this case, fusing with HB and thereby calibrating with beta calibration



(a) BC and LLO similar

(b) BC and LLO different

Figure 8.8: The empirical, beta calibration (BC), and LLO curves of two exemplary forecasters from the Knowledge Test Confidence data set (top row) together with the respective asymmetric and symmetric beta distributions (bottom row). (a) shows the calibration curves and respective beta distributions of forecaster 57, where LLO and beta calibration curves are similar. (b) shows the calibration curves and beta distributions of forecaster 46, for which the beta calibration function tends to 0 for $x \to 1$ causing miscalibrations.

can induce miscalibration of forecasts close to 1 that lead to forecasting the opposite of the forecast that was originally provided. This can result in worse performance of the HB fusion model in comparison to the HSB fusion model, which we see in the results presented in Section 8.2.4.

8.3 Discussion and Conclusion

In this work, we presented a family of normative generative models for fusing probability forecasts. Since uncertainty over probabilities is commonly modeled with the beta distribution, in our normative fusion models we model each forecaster's probability forecasts with beta distributions conditioned on their true label. We compare different variants of this model including hierarchical and non-hierarchical as well as asymmetric and symmetric beta fusion models. Given the respective model, new unseen probability estimates can be fused by inferring their true label. The obtained fused forecast is Bayes optimal given the model's assumptions. While previous approaches explicitly calibrate the considered forecasts using the Linear-in-Log-Odds (LLO) calibration function (M. D. Lee & Danileiko, 2014; Turner et al., 2014), the proposed beta fusion models implicitly calibrate the probability estimates provided by the forecasters with the beta calibration function, which accommodates the LLO calibration function as a special case.

We evaluated the proposed models on a data set by Turner et al. (2014) and the newly introduced Knowledge Test Confidence (KTeC) data set, also including two smaller subsets of these two data sets. In this vein, we also compared the proposed beta fusion models to the models by Turner et al. (2014), which fuse forecasts by averaging and calibrating them using the LLO calibration function. Looking at the results of all four data sets, i.e., the two full data sets and their respective reduced subsets, we can observe some general findings.

The hierarchical beta models generally outperform the non-hierarchical beta models. This is expected behavior because the hierarchical models are able to model each individual forecaster's behavior, which can be different, while the non-hierarchical models assume exchangeable forecasters, model all forecasters' collective behavior, and therefore discard valuable information. However, the Turner and KTeC data set differ in the magnitude of the difference between the performances of non-hierarchical and hierarchical beta fusion models. For the Turner data this difference is greater than for the KTeC data set since in the Turner data set the beta parameters of different forecasters are more variable than in the KTeC data set. Therefore, modeling all forecasters in the Turner data set with the same beta parameters causes comparably inferior performance.

Among the beta fusion models, the Hierarchical Symmetric Beta Fusion Model (HSB) shows the best performance. In particular, it outperforms the (asymmetric) Hierarchical Beta Fusion Model (HB), although HB does not constrain the modeling beta distributions to be symmetric and is therefore more expressive than HSB. As we discussed in Section 8.2.5, the reason for this is the calibration functions implied by the beta distributions, beta calibration for HB and LLO for HSB. Depending on the parameters learned for HB, beta calibration can lead to miscalibration of forecasts which causes worse performance of HB compared to HSB. Therefore, and in line with the results presented in Section 8.2.4, we recommend using the Hierarchical Symmetric Beta Fusion Model (HSB) instead of the Hierarchical Beta Fusion Model, and LLO calibration instead of beta calibration accordingly.

The conclusions discussed above are all consistent for the three performance measures Brier score, 0-1 loss, and MAE. However, as we report in Section 3.4, different measures suggest different models as the best-performing model. As we mentioned in Section 3.3 it is not clear which measure should be preferred. Brier score is commonly used (Baron et al., 2014; Hanea et al., 2021; Karvetski et al., 2013; Ranjan & Gneiting, 2010; Satopää, 2022; Turner et al., 2014), but criticized for not being appropriate for comparing forecasters, because it was developed for measuring whether forecasters report their true beliefs (Steyvers et al., 2014). Also, Brier score can be dominated by outliers (Canbek et al., 2022). On the other hand, the 0-1 loss directly compares the forecasters' performances and is easily interpretable but disregards their uncertainty. MAE considers the forecasters' uncertainty, is straightforwardly interpretable, and is robust to outliers (Canbek et al., 2022). However, it is an improper scoring rule (Buja et al., 2005) and incentivizes overconfident forecasts.

Since Brier score, 0-1 loss, and MAE have different strengths and weaknesses, we reported all three measures in our work. Interestingly, the differences between the results according to different measures reveal something about the models' properties. The hierarchical beta fusion models always outperform all other models regarding MAE. Thus, at least according to this measure, the normative models outperform the models proposed by Turner et al. (2014). On the two full data sets, some Turner models achieve lower, i.e., better Brier scores than the HSB, but these are different models depending on the respective data set, ATC, CTA, CTALO, and HCTA on the Turner data set, HCTALO on the KTeC data set. Still, according to 0-1 loss HSB is always competing with these best Turner models and outperforms them in terms of MAE. On the reduced data sets, also HSB's Brier score is better or competing to all other models.

The main reason why the hierarchical beta models (HB and HSB) achieve better MAE scores but worse Brier scores than some of Turner's models on the full data sets is their overconfidence. This overconfidence is a direct result of the conditional independence assumption of our beta fusion models (given the label, each subject provides independent probabilistic forecasts). This assumption is not met in the data. Since the forecasters respond to the same questions and share information and knowledge, their forecasts are not independent but tend to increase when other forecasters' forecasts increase. For example, the forecasters in our participant pool for the KTeC data set did not know the capital of Papua New Guinea but knew the answers to questions related to their university. Since they were students from Osnabrück in Germany, they shared knowledge about their university but consistently had little geographic knowledge about Papua New Guinea. If forecasts are combined assuming conditional independence, the fused forecast's uncertainty is usually reduced. This effect becomes stronger when more forecasts are fused. Unfortunately, if forecasts that are not actually independent are fused in this way, the fused forecast can be more confident than it should be (Trick & Rothkopf, 2022, see Chapter 3). Those overconfident forecasts can lead to high Brier scores, because the Brier score drastically punishes wrong forecasts with high confidence. If fewer forecasts are combined, the fused forecast is less overconfident, which is why the hierarchical beta models achieve better Brier scores than Turner's models on the two reduced data sets. The beta fusion models' overconfidence is also the reason why the standard errors of the mean, shown as error bars in the figures, are larger for the beta fusion models than for Turner's models regarding Brier score and MAE. More confident fusion models lead to higher variability in the Brier and MAE scores of different splits of LOO cross-validation and thus to higher standard errors in these measures.

Since the forecasts of multiple human experts are rarely independent (Winkler et al., 2019; Wiper & French, 1995), future work on combining forecasts should include the possibility to model correlations between forecasters to take into account the shared questions and knowledge of the forecasters, which will further increase the performance of the fused forecast. This could be realized, for example, using correlated beta distributions (Arnold & Ghosh, 2017; Moschen & Carvalho, 2023; Trick, Rothkopf, & Jäkel, 2023b, see Chapter 4).

9

BAYESIAN COMBINATION OF CORRELATED SUBJECTIVE PROBABILITY ESTIMATES

As we already stated in Chapter 8, subjective probability estimates provided by human experts play an important role in many different domains, among them finance, politics, engineering, meteorology, environmental science, and public health (McAndrew et al., 2021). In contrast to machine learning algorithms, they provide knowledge based on human intuition and experience without access to large data sets (McAndrew et al., 2021). In this way, they can steer decisions and facilitate planning and dealing with risks. The subjective probability estimates can be either predictions for future events (Graefe, 2018; Turner et al., 2014), e.g., election outcomes or weather phenomena, a quantification of an expert's belief in the truth of a statement (Karvetski et al., 2013; Prelec et al., 2017), or any other binary classifications. Consistent with Chapter 8, in the following, we will refer to these probability estimates as forecasts. Accordingly, the human subjects providing them will be termed forecasters. The subject of the forecast, e.g., an event to be predicted or a statement to be judged, will be referred to as a query.

It is well known that if multiple forecasts from different forecasters are available for the query at hand, combining these individual forecasts usually increases performance (Bennett et al., 2018; Budescu & Chen, 2015; McAndrew et al., 2021; Satopää, 2022; Satopää et al., 2023; Turner et al., 2014), known as the wisdom-of-the-crowds effect (Surowiecki, 2005). Multiple different mathematical fusion rules exist, which can result in different performances of the fused forecast (Satopää et al., 2016). Many of them assume the individual forecasters to be independent (Wilson, 2017) or at least do not explicitly consider a possible correlation between the provided probability estimates. Popular examples of such methods are linear opinion pools, i.e., unweighted or weighted averages. While unweighted linear opinion pools, i.e., standard averages, can already achieve solid performance (Clemen, 1989; Turner et al., 2014), weighted linear opinion pools try to increase fusion performance by giving each forecaster a different weight. The weights can be determined by the forecasters' individual performance (Cooke, 1991; Hanea et al., 2021), their forecasts' coherence (Karvetski et al., 2013), or the number of cues available to them (Budescu & Rantilla, 2000), or can be optimized for maximum performance (Ranjan & Gneiting, 2010). Because linear opinion pools can be over- or underconfident, also trimmed (Grushka-Cockayne et al., 2017) and extremized (Baron et al., 2014) linear opinion pools have been proposed. Besides linear opinion pools there are also multiplicative opinion pools, which combine forecasts using a product instead of a sum. Examples are Independent Opinion Pool (Berger, 1985), which explicitly assumes independent forecasts and multiplies the individual forecasts (see Chapter 3), or geometric (logarithmic) pooling (Berger, 1985; Dietrich & List, 2016), which is a weighted product of the forecasts. Linear as well as multiplicative opinion pools can also be used with transformations of the probability forecasts. Examples are the probit average (Satopää et al., 2023), which avoids a bias of the simple average towards 0.5, or a geometric mean of odds (Satopää et al., 2014).

In addition to the linear and multiplicative pooling methods listed above, also Bayesian models for forecast aggregation have been proposed that do not consider correlations between forecasters or explicitly assume independent forecasters (Hanea et al., 2021; Trick, Rothkopf, & Jäkel, 2023a, see Chapter 8; Turner et al., 2014). Hanea et al. (2021) propose an unsupervised Bayesian model that does not rely on historical seed questions to learn the forecasters' behavior. In contrast, Turner et al. (2014) introduces a family of different supervised Bayesian models, which either first calibrate the forecasts and then fuse them using an unweighted linear opinion pool or vice versa. Finally, our work proposed in the previous Chapter 8 explicitly assumes independent forecasts and models them with beta distributions conditioned on the truth value (Trick, Rothkopf, & Jäkel, 2023a, see Chapter 8).

However, human forecasters are usually correlated (Berger, 1985; Hogarth, 1978; Lichtendahl Jr et al., 2022; Wilson & Farrow, 2018; Winkler et al., 2019; Wiper & French, 1995). The correlation between forecasters can be attributed to similar data seen (Morris, 1986; Winkler, 1981; Winkler et al., 2019), similar training (Lichtendahl Jr et al., 2022; Winkler, 1981; Winkler et al., 2019), and/or similar methodology, such as statistical procedures as an aid for forecasting (Lichtendahl Jr et al., 2022; Morris, 1986; Winkler, 1981; Winkler et al., 2019).

If the provided forecasts are correlated but the used fusion method assumes independence, the fused forecast might be overconfident, because the amount of unique information is overestimated and thus uncertainty is reduced too much (Trick & Rothkopf, 2022, see Chapter 3; Trick, Rothkopf, & Jäkel, 2023a, see Chapter 8; Wilson, 2017). Therefore, forecast aggregation methods should explicitly consider the forecasters' correlation to avoid overconfidence and improve the fused forecast's performance (Wilson, 2017). In fact, considering the correlation of forecasters has been declared as one of the major challenges in forecast aggregation according to the review by McAndrew et al. (2021).

There are already several approaches that explicitly propose fusion methods for correlated forecasts. Budescu and Chen (2015) introduced a weighted linear opinion pool with higher weights for forecasters that are less correlated to the other forecasters. Satopää (2022) and J. Wang et al. (2021) propose Bayesian models for combining correlated forecasts. However, their models are unsupervised, so they cannot learn the individual behavior of the forecasters, including their correlation, from historical seed questions. In contrast, the model by Babic et al. (2022) relies on historical data but assumes the complete evidence that caused the forecasts to be known, which includes the amount of shared information that caused the correlation between the forecasters.

Since the concrete evidence underlying the provided forecasts is difficult to come by in practice, some Bayesian approaches model the unknown information that caused the forecasts as latent variables (Di Bacco et al., 2003; Lichtendahl Jr et al., 2022; Satopää et al., 2016). Di Bacco et al. (2003) consider two forecasters that provide their probabilistic forecasts for an event H. Their model assumes some knowledge that both forecasters share and some knowledge that only specific forecasters have, represented as events F, G_1, and G_2. The provided forecasts are transformed to odds and modeled with a log-normal distribution, jointly with the ratios of the posterior distributions of H given F, H given G_1, and H given G_2. However, their approach is rather theoretical. It remains unclear how to get the model parameters from prior experience. Satopää et al. (2016) propose a partial information framework for the aggregation of K forecasters. They model the information underlying the forecasts as particles of information, either positive or negative. If the sum of all particles X_S is positive, the event happens. Each forecaster observes some particle subset B_i, and the subsets of different forecasters can overlap, which generates a correlation between their forecasts. The sums of particles $X_S, X_{B_1}, \ldots, X_{B_K}$ are modeled with a multivariate normal distribution with mean 0 and covariances equal to the number of shared particles in different subsets. The sum of particles X_{B_i} is transformed to a probability forecast with probit transformation. With their model, Satopää et al. (2016) show when averages of forecasts should be extremized, i.e., shifted towards 0 or 1, and how this is dependent on how much information the forecasters share. With a similar goal, Lichtendahl Jr et al. (2022) also model the information causing the forecasts as information particles x_i. Some information particles are private for individual forecasters, some are shared between all forecasters. All information particles x_i as well as the target variable x_t, whose value determines the truth value t, are distributed under the same distribution with a conjugate prior on its parameters. By applying Bayes' rule, the posterior probability of the target variable x_t is computed as the fusion result given some sufficient statistics of observed information particles x_i, which are in turn computed from the observed probabilistic forecasts provided by the forecasters. The unknown sufficient statistics from the shared information particles are integrated out. Two specific conjugate models are discussed, Beta-Bernoulli and Normal-Normal. However, to be applicable to real data the model is simplified: It is assumed that the probability of the truth value t given all provided forecasts is a generalized linear model.

In contrast to the models presented above, which explicitly model the latent information causing the provided forecasts, other Bayesian models only model the provided probability forecasts as data (Bordley, 1982; Clemen & Winkler, 1987; French, 1980). French (1980) as well as Clemen and Winkler (1987) transfer the probability estimates to log-odds and model them with a multivariate Gaussian distribution conditioned on the truth value t. With their model and some historical training data, they can learn how individual forecasters behave for true and false queries, i.e., their bias, variance, and uncertainty. Also, using the Gaussian distribution they can model pairwise correlation between individual forecasters. Bordley (1982) uses the same multivariate Gaussian model approach in order to derive the weights for a multiplicative fusion rule considering the correlation of forecasters.

While transforming the probability forecasts to log-odds and modeling them with a multivariate Gaussian distribution is mathematically convenient and enables straightforward representation of correlations between forecasts, another possibility is to model them directly with a beta distribution without any transformation. The beta distribution is commonly used to model probabilities since it is the standard distribution over probabilities in Bayesian statistics and the conjugate prior of the Bernoulli distribution. Accordingly, in the previous Chapter 8 we modeled probabilistic forecasts with a beta distribution conditioned on the truth value t (Trick, Rothkopf, & Jäkel, 2023a, see Chapter 8) and used this model for normative aggregation of probability forecasts. However, in that model, we assumed the forecasts provided by different forecasters to be conditionally independent given the truth value t. This assumption, which usually does not hold in reality, caused our beta fusion model to be overconfident on two real-world data sets. Thus, for correlated forecasts this model is not normative. It does not formalize how to obtain the correct fused uncertainty.

Therefore, in the present work, we introduce a Bayesian model for combining probabilistic forecasts that considers the correlation between forecasts. The new model also represents the forecasts with a beta distribution conditioned on the truth value t. However, in our model, we additionally assume that the forecast provided by a forecaster for a query depends on an interplay of the forecaster's skill and the query's difficulty. Moreover, we assume that the correlation between forecasts provided by different forecasters can be attributed to the fact that they answer to the same queries since for all forecasters some queries are easy and others are hard. The resulting Skill-Difficulty Correlated Fusion Model explicitly models the forecasters' skills and the queries' difficulties and thereby models the correlation between forecasts, which can then be considered for fusion. Fusing forecasts according to the Skill-Difficulty Correlated Fusion Model is normative, given that the model assumptions are correct. In particular, correlated forecasts can be fused normatively.

The remainder of the chapter is structured as follows. Section 9.1 introduces our new Skill-Difficulty Correlated Fusion Model, including its generative model and a discussion of its parameters, the correlations that can be represented, as well as parameter inference and fusion. In Section 9.2 we present our evaluations of the model, including a detailed description of the used data set and cross-validation, an evaluation of how the estimated model parameters can be interpreted and how our model fits the data, and a comparison of the Skill-Difficulty Correlated Fusion Model's fusion performance to related Bayesian models. Finally, we discuss our results and outline conclusions, limitations, and ideas for future work in Section 9.3.

9.1 The Skill-Difficulty Correlated Fusion Model

We propose the Skill-Difficulty Correlated Fusion Model, a Bayesian generative model for combining probabilistic forecasts provided by human forecasters, which explicitly models the forecasters' skills and the queries' difficulties in order to model the correlation between forecasts. After introducing the generative model in Section 9.1.1, we discuss the model parameters' interpretation in Section 9.1.2 and the correlations that can be represented using the model in Section 9.1.3. In Section 9.1.4 we outline parameter inference and fusion with the model.

9.1.1 Generative Model

We assume K human forecasters to provide subjective probability estimates, i.e., forecasts, on N binary queries. The queries can be factual statements that are either true or false, questions that can be answered with yes or no, or any other binary classification task with a truth value of either 1 or 0. For each query, the probability estimates provided by the forecasters quantify their belief in their answer's correctness. A probability of 0 indicates that the forecaster is completely certain that the query's truth value is 0 (e.g., false/no), whereas a probability of 1 means that the forecaster assumes a truth value of 1 (e.g., true/yes) with full certainty.

We formalize the forecast provided by forecaster k for query n as $x_n^k \in [0,1]$ with $n = 1,\dots,N$, $k = 1,\dots,K$. The truth value of query n is formalized as $t_n \in \{0,1\}$ for $n = 1,\dots,N$. Since the beta distribution is the natural choice for modeling probabilities, our previous model introduced in Chapter 8 (Trick, Rothkopf, & Jäkel, 2023a, see Chapter 8) assumes the forecasts x_n^k to be beta-distributed conditioned on the truth value $t_n \in \{0,1\}$, as

$$x_n^k | t_n = j \sim \mathrm{Beta}(\alpha_j^k, \beta_j^k), \quad j = 0,1. \tag{9.1}$$

After learning the parameters α_j^k and β_j^k from labeled training data, this model represents the forecasting behavior of each forecaster conditioned on the truth value t_n, including their bias, uncertainty, and variance. In particular, forecaster k's bias can be expressed with the beta distribution's mean $\mu_j^k = \alpha_j^k/(\alpha_j^k + \beta_j^k)$. If $\mu_0^k < 0.5$ and $\mu_1^k > 0.5$, we define forecaster k to be unbiased since he on average provides correct forecasts. The mean forecast μ_j^k can also quantify the uncertainty of forecaster k: The closer it is to 0 or 1, the less uncertain the forecaster is on average. However, the uncertainty of the actual forecasts provided by forecaster k is also dependent on his variance, which determines his forecasts' concentration around the mean μ_j^k. This variance can be expressed with the modeling beta distribution's precision $p_j^k = \alpha_j^k + \beta_j^k$, also known as its concentration parameter (J. Huang, 2005). This precision is the higher, the lower the beta distribution's variance is. Still, note that it is not the inverse of the distribution's variance. While explicitly considering the learned behavior of the forecasts, i.e., their bias, uncertainty, and variance, new forecasts x_u^k for a previously unseen query u can be fused by inferring the truth value t_u given the forecasts x_u^k and the learned parameters α_j^k and β_j^k using Bayes' rule. Note that by modeling bias, variance, and uncertainty with a beta distribution conditioned on the truth value t_n, the forecasters are also implicitly calibrated when they are fused. This means that they are corrected for over- or underconfident forecasts, which do not match the relative frequency of occurrence, e.g., of a predicted event or a judged true statement (Trick, Rothkopf, & Jäkel, 2023a, see Chapter 8).

In previous work, four variants of this beta fusion model were compared (Trick, Rothkopf, & Jäkel, 2023a, see Chapter 8). While in (9.1) the forecasts are modeled with two completely different beta distributions for $t_n = 0$ and $t_n = 1$, the symmetric variant of the model represents the forecasts with symmetric beta distributions for $t_n = 0$ and $t_n = 1$, i.e., as

$$\begin{aligned} x_n^k | t_n = 0 &\sim \mathrm{Beta}(\alpha^k, \beta^k) \\ x_n^k | t_n = 1 &\sim \mathrm{Beta}(\beta^k, \alpha^k). \end{aligned} \tag{9.2}$$

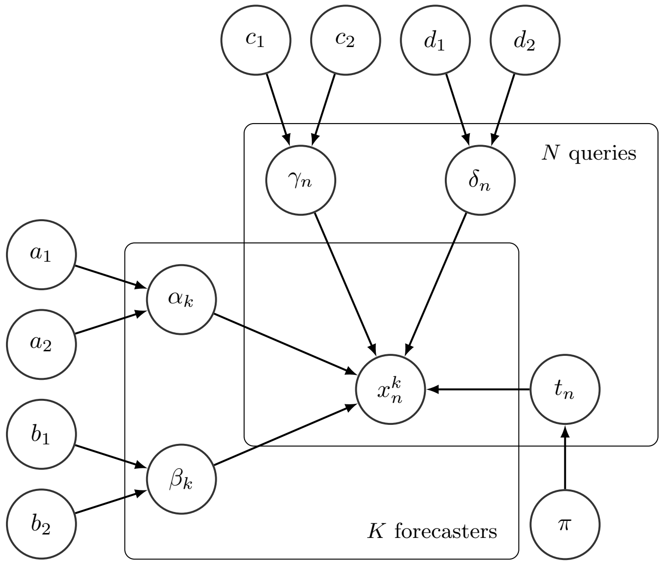
Thus, it assumes the forecasts provided for queries with truth values $t_n = 0$ and the ones for queries with truth values $t_n = 1$ to be symmetric around 0.5. This symmetric beta fusion model is of special interest because it implicitly calibrates the forecasts according to a well-known calibration function, the Linear-in-Log-Odds calibration function (Trick, Rothkopf, & Jäkel, 2023a, see Chapter 8). For both the asymmetric (9.1) and the symmetric (9.2) variant of the model, which model each forecaster k separately with hierarchical models, we also investigated non-hierarchical model variants, which consider the forecasters as exchangeable and only learn one parameter set for all of them.

All variants of the beta fusion model shown above assume the forecasts x_n^k provided by different forecasters k to be conditionally independent given the truth value t_n. However, for real-world forecasting data, this assumption is usually not met (Berger, 1985; Hogarth, 1978; Lichtendahl Jr et al., 2022; Wilson & Farrow, 2018; Winkler et al., 2019; Wiper & French, 1995). Accordingly, our evaluations on two data sets also showed that the beta fusion models are overconfident, which deteriorates their performance. Therefore, in this work we extend the beta fusion model to consider correlations between forecasters.

In line with the previous beta fusion models, in our new Skill-Difficulty Correlated Fusion Model we also model the forecasts x_n^k with a beta distribution conditioned on the truth value t_n. Likewise, we also assume symmetric beta distributions for modeling forecasts with truth value $t_n = 0$ and $t_n = 1$ as in (9.2), because previous evaluations showed that symmetric beta modeling leads to more robust calibration and higher fusion performance (Trick, Rothkopf, & Jäkel, 2023a, see Chapter 8). However, we do not assume that forecast x_n^k by forecaster k for query n is generated merely by the forecaster's behavior and can thus be modeled with a beta distribution with parameters α^k, β^k specific for this forecaster k. Instead, here we assume that forecast x_n^k by forecaster k for query n is generated by both forecaster k's properties, which we call his skill, and query n's properties, which we call its difficulty. Accordingly, we model x_n^k with a beta distribution with parameters α_k and β_k, specific for forecaster k, and parameters γ_n and δ_n, specific for query n. In particular, if $t_n = 0$, x_n^k is modeled with a beta distribution with parameters $\alpha_k + \gamma_n$ and $\beta_k + \delta_n$. If $t_n = 1$ the parameters are interchanged to $\beta_k + \delta_n$ and $\alpha_k + \gamma_n$,

$$\begin{aligned} x_n^k | t_n = 0 &\sim \mathrm{Beta}(\alpha_k + \gamma_n, \beta_k + \delta_n) \\ x_n^k | t_n = 1 &\sim \mathrm{Beta}(\beta_k + \delta_n, \alpha_k + \gamma_n). \end{aligned} \tag{9.3}$$

As priors for the parameters $\alpha_k, \beta_k, \gamma_n, \delta_n$ we chose gamma distributions with hyperparameters $a_1, a_2, b_1, b_2, c_1, c_2, d_1, d_2$ respectively. Their uninformed hyperpriors are vague gamma distributions with shape and rate set to 0.001. The prior distribution on the truth



$$\begin{aligned} t_n &\sim \mathrm{Bernoulli}(\pi) \\ x_n^k | t_n = 0 &\sim \mathrm{Beta}(\alpha_k + \gamma_n, \beta_k + \delta_n) \\ x_n^k | t_n = 1 &\sim \mathrm{Beta}(\beta_k + \delta_n, \alpha_k + \gamma_n) \\ \alpha_k &\sim \mathrm{Gamma}(a_1, a_2) \\ \beta_k &\sim \mathrm{Gamma}(b_1, b_2) \\ \gamma_n &\sim \mathrm{Gamma}(c_1, c_2) \\ \delta_n &\sim \mathrm{Gamma}(d_1, d_2) \\ a_1 &\sim \mathrm{Gamma}(0.001, 0.001) \\ a_2 &\sim \mathrm{Gamma}(0.001, 0.001) \\ b_1 &\sim \mathrm{Gamma}(0.001, 0.001) \\ b_2 &\sim \mathrm{Gamma}(0.001, 0.001) \\ c_1 &\sim \mathrm{Gamma}(0.001, 0.001) \\ c_2 &\sim \mathrm{Gamma}(0.001, 0.001) \\ d_1 &\sim \mathrm{Gamma}(0.001, 0.001) \\ d_2 &\sim \mathrm{Gamma}(0.001, 0.001) \\ \pi &\sim \mathrm{Beta}(1, 1) \end{aligned}$$

Figure 9.1: The graphical model of the Skill-Difficulty Correlated Fusion Model.

value t_n is a Bernoulli distribution with parameter π, which is the proportion of true queries with truth value 1. The hyperprior on π is a uniform beta distribution Beta(1,1). The graphical model of the Skill-Difficulty Correlated Fusion Model is shown in Figure 9.1.

9.1.2 Interpretation of the Model's Parameters as Skill and Difficulty

The parameters α_k and β_k represent something akin to the skill of forecaster k in the sense that they determine the bias, variance, and uncertainty of forecasts provided by forecaster k independently of the query at hand n. Therefore, in the following, α_k and β_k will be termed skill parameters. Equivalently, the parameters γ_n and δ_n represent the difficulty of query n since they determine the bias, variance, and uncertainty of forecasts provided for query n independently of the specific forecasters providing them. Accordingly, we call γ_n and δ_n difficulty parameters.

For interpreting the skill parameters α_k and β_k isolated from the difficulty parameters γ_n and δ_n, we can set the difficulty parameters γ_n and δ_n close to 0, meaning that only the forecasters' properties are determining the forecasts. In this case, forecasts x_n^k are distributed according to Beta(α_k, β_k) for $t_n = 0$ and Beta(β_k, α_k) for $t_n = 1$. Note that with this parametrization our Skill-Difficulty Correlated Fusion Model equals the symmetric variant of the previously proposed independent beta fusion model in (9.2). If for real data the difficulty parameters γ_n and δ_n are not equal to 0, the distributions Beta(α_k, β_k) for $t_n = 0$ and Beta(β_k, α_k) for $t_n = 1$ do not describe the forecasts x_n^k. However, they describe the forecasts we would assume if the queries' difficulties had no influence on forecaster k's forecasts. Forecaster k's bias, which in part defines his skill, is determined by the mean of this beta distribution for $t_n = 0$, which is $\mu_k^{\alpha\beta} = \alpha_k/(\alpha_k + \beta_k)$. This skill mean describes the average forecast we would expect if the queries' difficulties had no influence on forecaster k's forecasts. If $\mu_k^{\alpha\beta} < 0.5$, we define forecaster k to be unbiased because his mean forecast is closer to the truth value t_n. Thus, if $\beta_k > \alpha_k$, then forecaster k is unbiased. Forecaster k's skill is also determined by his uncertainty, which is also quantified by the skill mean $\mu_k^{\alpha\beta}$. The more uncertain, i.e., the closer to 0.5 it is, the more uncertain is forecaster k on average. However, the skill mean $\mu_k^{\alpha\beta}$ can only provide information on the average uncertainty, not on the actual uncertainties of different forecasts provided, since there might be variance in the provided forecasts. The variance or variability of forecaster k around his skill mean, which is also part of his skill, can be straightforwardly quantified with the beta distribution's skill precision $p_k^{\alpha\beta} = \alpha_k + \beta_k$. Thereby, lower precisions indicate higher variance around the mean.

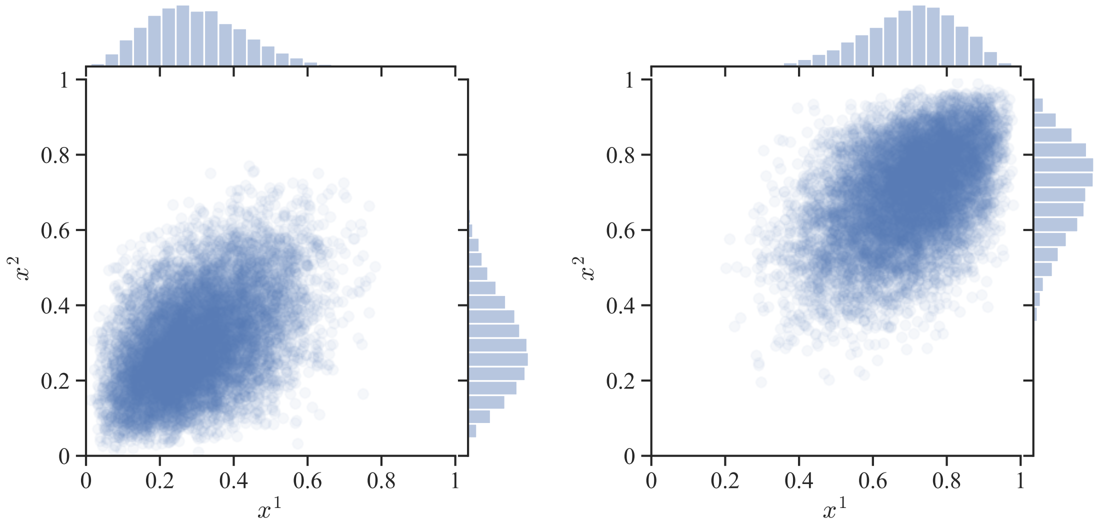
The skill parameters α_k and β_k and the derived skill mean $\mu_k^{\alpha\beta}$ and skill precision $p_k^{\alpha\beta}$ can identify prototypical forecasters. Given that we assume the queries' difficulties to have no impact on his forecasts, a highly skilled forecaster provides unbiased and certain forecasts with low variance and will thus show a skill mean $\mu_k^{\alpha\beta}$ close to 0 with high skill precision. An uncertain forecaster provides forecasts close to 0.5 with low variance with $\mu_k^{\alpha\beta}$ close to 0.5 and a high skill precision. If $\mu_k^{\alpha\beta}$ is close to 1 with high skill precision, forecaster k is a wrong forecaster with a strong bias and thus always provides incorrect forecasts with low uncertainty and low variance and might not have understood the task correctly or is malingering.

The difficulty parameters γ_n and δ_n can be interpreted when setting the skill parameters α_k and β_k close to 0. As a consequence, x_n^k are distributed according to $\text{Beta}(\gamma_n, \delta_n)$ for $t_n = 0$ and $\text{Beta}(\delta_n, \gamma_n)$ for $t_n = 1$. Although these distributions might not model real forecasts, which are also determined by the skills of the forecasters providing them, they describe the forecasts we would assume to be provided for query n if the forecasters' skills had no impact on the forecasts. The mean of this distribution for $t_n = 0$, the difficulty mean $\mu_n^{\gamma\delta} = \gamma_n/(\gamma_n + \delta_n)$, quantifies the average forecast for query n if the forecasters' skills had no impact on the forecasts provided for query n. It determines the bias of the forecasts provided for query n, which partly defines its difficulty. If the difficulty mean $\mu_n^{\gamma\delta} < 0.5$, hence if $\delta_n > \gamma_n$, the forecasts for query n are unbiased. Thus, the forecasters on average provide correct answers for the query, so it is rather easy. Its difficulty is, however, also determined by the uncertainty of the forecasts provided. The average uncertainty of the forecasts provided for query n is quantified by the uncertainty of the difficulty mean $\mu_n^{\gamma\delta}$. More extreme difficulty means closer to 0 or 1 show lower average uncertainty of the forecasts provided for a query. The difficulty precision $p_n^{\gamma\delta} = \gamma_n + \delta_n$ quantifies how concentrated the forecasts provided for query n are around the difficulty mean $\mu_n^{\gamma\delta}$, i.e., their variance.

Given difficulty mean $\mu_n^{\gamma\delta}$ and difficulty precision $p_n^{\gamma\delta}$ we can identify special queries. If $\mu_n^{\gamma\delta}$ is close to 0 with high difficulty precision, query n is a very easy query, for which forecasters provide unbiased and certain forecasts with low variance, given that we assume the forecasters' skills to have no impact on the forecasts provided for it. For difficult queries we distinguish between two types of queries, unknown and trick queries. For unknown queries, $\mu_n^{\gamma\delta}$ is close to 0.5 with high difficulty precision, so the forecasters do not know the answer to the query and therefore provide uncertain forecasts close to 0.5 with low variance. In contrast, for trick queries they provide certain but biased forecasts with low variance because they think they know the correct answer but do not. In this case, $\mu_n^{\gamma\delta}$ is close to 1 with high difficulty precision.

9.1.3 Correlations in the Model

The Skill-Difficulty Correlated Fusion Model can model positive correlations between forecasts. On the one hand, it can model the correlation between forecasts provided by two forecasters l and m for different queries, x^l and x^m, with fixed skill parameters α_l and β_l for forecaster l and α_m and β_m for forecaster m but variable difficulty parameters γ_n and δ_n. This correlation is caused by the forecasters seeing the same queries, e.g., easy queries, unknown queries, or trick queries. On the other hand, the Skill-Difficulty Correlated Fusion Model can represent the correlation between the forecasts provided by multiple forecasters for two queries p and q, x_p and x_q, with fixed difficulty parameters γ_p and δ_p for query p and γ_q and δ_q for query q but variable skill parameters α_k and β_k. These forecasts for queries p and q might be correlated because they come from the same forecasters. In this work, we will use the Skill-Difficulty Correlated Fusion Model to combine forecasts provided by different forecasters for the same query while considering the potential correlation between these forecasters over different queries. Therefore, here we focus on the first kind of correlation mentioned above: the correlation between the forecasts provided by two forecasters l and m for different queries, x^l and x^m.



(a) $t = 0$

(b) $t = 1$

Figure 9.2: Simulated forecasts of two forecasters, x^1 and x^2, for 20000 queries, 10000 of which with truth value $t = 0$ (a) and 10000 with truth value $t = 1$ (b). The two forecasters are equally skilled with skill parameters $\alpha_1 = \alpha_2 = 3$ and $\beta_1 = \beta_2 = 10$. The difficulty parameters γ_n and δ_n for different n are drawn from a gamma distribution with parameters $c_1 = 2, c_2 = 0.5, d_1 = 3, d_2 = 0.5$. The resulting correlation between x^1 and x^2 is 0.47 for $t = 0$ and $t = 1$.

As shown in (9.3), the forecasts of two forecasters l and m for a query n, x_n^l and x_n^m are conditionally independent given γ_n and δ_n by definition, because they are independent draws from a beta distribution. Accordingly, if all modeled queries have the same difficulty, so γ_n and δ_n are the same for all n, there is no correlation between the forecasts provided by different forecasters. If additionally $\gamma_n = \delta_n = 0$ for all n, our model equals the independent beta fusion model with symmetric beta distributions shown in (9.2). While x_n^l and x_n^m are conditionally independent given γ_n and δ_n, multiple forecasts provided by forecasters l and m for different queries, x^l and x^m, are, however, not unconditionally independent. If the queries have different difficulties, as we expect in reality, x^l and x^m are positively correlated. An example of such a model parametrization is shown in Figure 9.2.

The correlation between two forecasters is modulated by how much the forecasts provided by these forecasters are determined by their individual skills in comparison to the shared queries' difficulties. This trade-off between the influence of individual forecasters' skills and the queries' difficulties is quantified by the relation between skill parameters α_k, β_k and difficulty parameters γ_n, δ_n. Higher difficulty parameters γ_n and δ_n compared to α_k and β_k lead to higher influence of the shared queries' difficulties on the provided forecasts compared to the forecasters' individual skills. Thus, the correlation between different forecasters' provided forecasts increases. If in contrast γ_n and δ_n stay the same, so the difficulties of the queries remain unchanged, but the skill parameters α_k and β_k are increased, meaning that the forecasts are more determined by the forecasters' individual skills and less dependent on the difficulty of shared queries, the correlation between the forecasters is decreased.

The correlation r^{lm} between the forecasts of two forecasters l and m for different queries, x^l and x^m, can be derived mathematically. It is defined as

$$r^{lm} = \frac{\mathrm{Cov}(x^l, x^m)}{\sqrt{\mathrm{Var}(x^l)\mathrm{Var}(x^m)}} \tag{9.4}$$

with

$$\mathrm{Cov}(x^l, x^m) = \mathrm{E}(x^l x^m) - \mathrm{E}(x^l)\mathrm{E}(x^m). \tag{9.5}$$

The expected values, variances, as well as the product moments need to be computed separately for $t_n = 0$ and $t_n = 1$. However, as we assume our model to be symmetric for $t_n = 0$ and $t_n = 1$ and the correlation is thus the same for $t_n = 0$ and $t_n = 1$, we will only show the derivations for $t_n = 0$ in the following. Since in the following derivations we marginalize over all queries, we omit the index n for t, γ, and δ. The expected value and variance of x^l and the product moment of x^l and x^m given $t = 0$ are

$$\mathrm{E}(x^l|t=0) = \int\int \frac{\alpha_l + \gamma}{\alpha_l + \beta_l + \gamma + \delta} \frac{c_2^{c_1}}{\Gamma(c_1)} \gamma_n^{c_1 - 1} \frac{d_2^{d_1}}{\Gamma(d_1)} \delta^{d_1 - 1} e^{-c_2\gamma - d_2\delta} \,\mathrm{d}\gamma\,\mathrm{d}\delta, \tag{9.6}$$

$$\begin{aligned}\mathrm{Var}(x^l|t=0) = \int\int &\frac{\mathrm{B}(\alpha_l + \gamma + 2, \beta_l + \delta)}{\mathrm{B}(\alpha_l + \gamma, \beta_l + \delta)} \frac{c_2^{c_1}}{\Gamma(c_1)} \gamma^{c_1 - 1} \frac{d_2^{d_1}}{\Gamma(d_1)} \delta^{d_1 - 1} e^{-c_2\gamma - d_2\delta} \,\mathrm{d}\gamma\,\mathrm{d}\delta \\ &- \mathrm{E}(x^l|t=0)^2,\end{aligned} \tag{9.7}$$

$$\begin{aligned}\mathrm{E}(x^l x^m|t=0) = \int\int &\frac{\alpha_l + \gamma}{\alpha_l + \beta_l + \gamma + \delta} \frac{\alpha_m + \gamma}{\alpha_m + \beta_m + \gamma + \delta} \\ &\cdot \frac{c_2^{c_1}}{\Gamma(c_1)} \gamma^{c_1 - 1} \frac{d_2^{d_1}}{\Gamma(d_1)} \delta^{d_1 - 1} e^{-c_2\gamma - d_2\delta} \,\mathrm{d}\gamma\,\mathrm{d}\delta.\end{aligned} \tag{9.8}$$

We were not able to solve the integrals above but can compute them (and with them the correlation) numerically using the Python package mpmath (Johansson et al., 2013).

9.1.4 Parameter Inference and Fusion

From labeled training data the posterior distribution over the parameters $\alpha_k, \beta_k, \gamma_n, \delta_n$ as well as the hyperparameters $a_1, a_2, b_1, b_2, c_1, c_2, d_1, d_2$ and the proportion of true queries π can be inferred using Gibbs sampling.

For fusing the forecasts of K forecasters $x_u^1, \ldots, x_u^K$ for a new unseen query u, we infer the posterior distribution over its truth value t_u given the forecasts to be fused and the previously learned posterior distributions over the model parameters. Since we do not know the difficulty of the new query u and thus its difficulty parameters γ_u and δ_u, we can only condition on the learned hyperparameters c_1, c_2, d_1, d_2 together with the learned skill parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ of all forecasters and the proportion of true queries π. The resulting posterior fused forecast $p(t_u|x_u^1, \ldots, x_u^K, \boldsymbol{\alpha}, \boldsymbol{\beta}, c_1, c_2, d_1, d_2, \pi)$ can be inferred using Gibbs sampling. Note that by fusing the forecasts in this way, the individual forecasters' skills as well as their correlation are considered for fusion.

9.2 Evaluation

We evaluate the proposed Skill-Difficulty Correlated Fusion Model on the Knowledge Test Confidence data set (Section 9.2.1) using leave-one-out cross-validation (Section 9.2.2). For exemplary forecasters and queries we show how our model's parameters can be interpreted and how the model fits the data (Section 9.2.3). In addition, we evaluate the fusion performance of the Skill-Difficulty Correlated Fusion Model in comparison to previously proposed Bayesian fusion models in terms of Brier score, mean absolute error, and entropy (Section 9.2.4).

9.2.1 Knowledge Test Confidence Data Set

The Knowledge Test Confidence (KTeC) data set[1] (Trick, Rothkopf, & Jäkel, 2023a, see Chapter 8) includes subjective probability estimates of 85 human forecasters on 180 queries. Each forecaster provided a forecast to all 180 queries, resulting in a total number of 15300 subjective probability estimates.

The queries are knowledge statements that are either false ($t_n = 0$) or true ($t_n = 1$), while the number of false and true statements is balanced. There are easy queries known to the majority of forecasters, such as 'Elephants are mammals' ($t_n = 1$), and hard queries. Hard queries are either unknown to most forecasters, e.g., '*Being and Nothingness* was written in 1943' ($t_n = 1$), or are trick questions that most forecasters judge incorrectly, e.g., 'The official language of the United States is English' ($t_n = 0$).

The forecasters were students from University of Osnabrück. Their probabilistic forecasts are their confidence judgments on the given statements. A confidence of 0 indicates that the forecaster is 100% certain that the statement is wrong, whereas a confidence of 1 indicates that the forecaster is 100% certain that the statement is correct. The forecasts could be provided in 11 discrete steps of 0.1 between 0 and 1. For the following evaluations we preprocessed forecasts of 0 and 1 to 0.001 and 0.999 to avoid computational problems. Note that in previous work, we showed that this score correction has no significant influence on the fusion results (Trick, Rothkopf, & Jäkel, 2023a, see Chapter 8). In the KTeC data set, the forecasts provided by different forecasters are correlated. For queries with truth value 0 the pairwise correlation between two forecasters is on average 0.293, for queries with truth value 1 it is on average 0.333. More details on the data set, e.g., how the query statements were designed, can be found in the work of Trick, Rothkopf, and Jäkel (2023a, see Chapter 8).

9.2.2 Cross-Validation

We evaluate the Skill-Difficulty Correlated Fusion Model on the KTeC data set described above using leave-one-out (LOO) cross-validation. Accordingly, we split the data set, consisting of 180 queries, into 180 training and test sets. Each training set is composed of the forecasts of all 85 forecasters on 179 queries, while the forecasts on the one remaining query build the respective test sets.

On each training set we inferred the posterior distributions of the model parameters α_k, β_k, γ_n, δ_n, a_1, a_2, b_1, b_2, c_1, c_2, d_1, d_2, π using Gibbs sampling. For our evaluations, we implemented Gibbs sampling using JAGS (Plummer, 2003) and ran one chain with 1000 samples and a burn-in of 1000 samples.

[1] The KTeC data set is available at `https://osf.io/ae25w/`.

Given the model parameters estimated on the training set, the forecasts $x_u^1, \dots, x_u^K$ on the one remaining query u in the respective test set are fused by inferring the posterior distribution of its truth value t_u, $p(t_u | x_u^1, \dots, x_u^K, \boldsymbol{\alpha}, \boldsymbol{\beta}, c_1, c_2, d_1, d_2, \pi)$. As for parameter inference, inference is realized using Gibbs sampling in JAGS. In order to consider the uncertainty of the previously inferred parameters, here, we do not consider their point estimates as observed variables but uniformly sample from their posterior distributions' samples. We ran 100 parallel chains with 800 samples and a burn-in of 800 samples each.

9.2.3 Parameter Interpretation and Model Fit

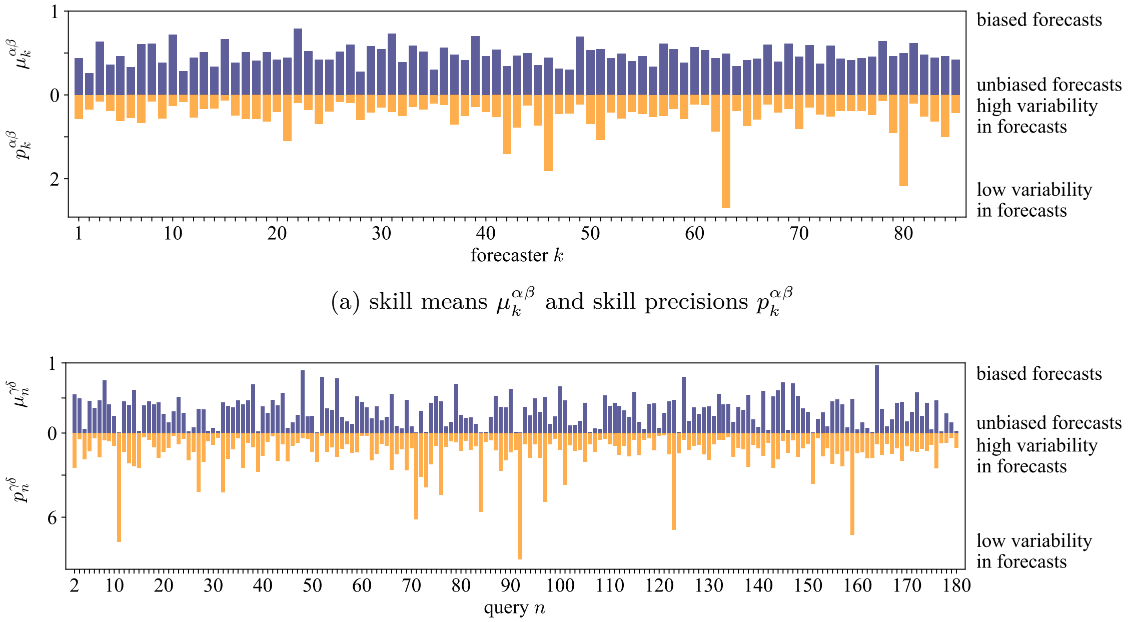
In Section 9.1.2 we outlined how the skill parameters α_k, β_k and the difficulty parameters γ_n, δ_n can be interpreted in theory. In the following, we will analyze and interpret the parameters inferred for training split 1 of LOO cross-validation on the KTeC data set. We discuss what the inferred parameters reveal about our data set and show how the parameters and with them the Skill-Difficulty Correlated Fusion Model fit the data. The parameter values specified in the following are the means of the posterior distributions inferred with Gibbs Sampling.

The skill hyperparameters inferred for our model, a_1, a_2, b_1, b_2, show that the forecasters in the KTeC data set are on average unbiased and provide more certain than uncertain forecasts. From the estimated values $a_1 = 2.74$, $a_2 = 10.98$, $b_1 = 1.86$, $b_2 = 6.46$ we can infer mean $\alpha_k = 0.25$ and mean $\beta_k = 0.29$, leading to an unbiased and uncertain skill mean $\mu_k^{\alpha\beta} = 0.46 < 0.5$. Because of the low skill precision $p_k^{\alpha\beta} = 0.54$, the average provided forecasts are, however, not concentrated at the uncertain skill mean but are more certain, tending towards 0 and 1. The variances of α_k and β_k are low with $\text{Var}(\alpha_k) = 0.02$ and $\text{Var}(\beta_k) = 0.04$, so the different forecasters perform similarly.

A more precise picture of the forecasters' skills in the KTeC data set can be obtained by looking at Figure 9.3(a), which shows the skill means $\mu_k^{\alpha\beta}$ and the skill precisions $p_k^{\alpha\beta}$ of all 85 forecasters in training split 1, computed from their individual skill parameters α_k and β_k. As can be seen, there are unbiased ($\mu_k^{\alpha\beta} < 0.5$) and biased ($\mu_k^{\alpha\beta} > 0.5$) forecasters and forecasters with low and high skill precision $p_k^{\alpha\beta}$. However, the forecasters' skills are fairly similar, as indicated by the low variance of α_k and β_k reported above.

The inferred difficulty hyperparameters c_1, c_2, d_1, d_2 with values $c_1 = 0.83$, $c_2 = 1.85$, $d_1 = 1.37$, $d_2 = 1.29$ reveal that the mean difficulty parameters are $\gamma_n = 0.45$ and $\delta_n = 1.06$. Thus, on average the difficulty mean is $\mu_n^{\gamma\delta} = 0.3 < 0.5$ with a rather high average difficulty precision of $p_n^{\gamma\delta} = 1.51$, showing that the forecasts provided for the queries in the KTeC data set are on average unbiased with medium uncertainty and rather low variance. Thus, the queries are on average rather easy. The variances of γ_n and δ_n are higher than those of α_k and β_k with $\text{Var}(\gamma_n) = 0.24$ and $\text{Var}(\delta_n) = 0.82$, so the queries' difficulties in the KTeC data set are more variable than the forecasters' skills.

This can also be seen in Figure 9.3(b), which shows the difficulty means $\mu_n^{\gamma\delta}$ and difficulty precisions $p_n^{\gamma\delta}$ of all 179 queries in training split 1, computed from their difficulty parameters γ_n and δ_n. Compared to the skills in Figure 9.3(a), the difficulties are more diverse. Still, there are much more rather easy queries than hard queries, indicated by more low difficulty means. These findings are consistent with the interpretations of the inferred difficulty hyperparameters above.



(a) skill means $\mu_k^{\alpha\beta}$ and skill precisions $p_k^{\alpha\beta}$

(b) difficulty means $\mu_n^{\gamma\delta}$ and difficulty precisions $p_n^{\gamma\delta}$

Figure 9.3: A comparison of the skills of different forecasters and the difficulties of different queries from the KTeC data set. We show the skill means $\mu_k^{\alpha\beta}$ (blue) and skill precisions $p_k^{\alpha\beta}$(orange) of all 85 forecasters (a) and the difficulty means $\mu_n^{\gamma\delta}$ (blue) and difficulty precisions $p_n^{\gamma\delta}$(orange) of queries 2-180 (b) in training split 1 of LOO cross-validation. Low means indicate unbiased forecasts provided by forecaster k or for query n, high means indicate biased forecasts. Low precisions indicate high variability in forecasts provided by forecaster k or for query n, high precisions indicate low variability in forecasts.

In order to analyze individual forecasters' skills and queries' difficulties, in the following, we will show the provided forecasts by exemplary forecasters (Section 9.2.3.1) and for prototypical queries (Section 9.2.3.2) from training split 1 of LOO cross-validation and explain how the respective inferred skill or difficulty parameters and with them our Skill-Difficulty Correlated Fusion Model fit the shown data.

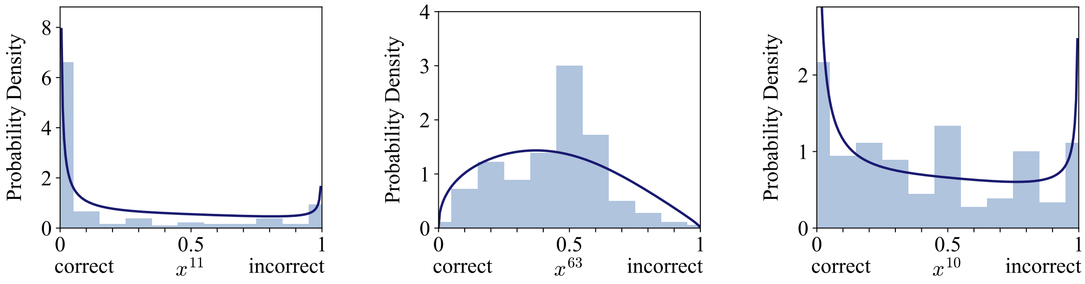
9.2.3.1 Forecasters' Skills

Figure 9.4 shows the provided forecasts of three exemplary forecasters, forecaster 11, forecaster 63, and forecaster 10, for the 179 queries in training split 1. To be able to show their skill without distinguishing queries with truth label $t_n = 0$ and $t_n = 1$, we inverted all forecasts given for queries with truth value $t_n = 1$. Thus, a forecast $x_n^k < 0.5$ can be considered a correct forecast, the lower, the more certain. We plot the relative frequency of the forecasts provided by forecaster k for all queries in a histogram. The curves plotted over the data show the learned distributions according to our model in (9.3) over all forecasts provided by forecaster k. While the skill parameters α_k and β_k are thus fixed, for each single forecast in the shown data the difficulty parameters γ_n and δ_n in (9.3) are different because they are provided for different queries. Therefore, the shown distribution is a mixture of beta distribution, whose mixture components are 179 beta distributions according to (9.3) with skill parameters α_k and β_k for the respective forecaster k and the difficulty parameters γ_n and δ_n of all 179 queries in the training split. All mixture weights are equal because all queries have the same impact on the distribution. Since we inverted all forecasts for queries with truth value $t_n = 1$ in order to show a forecaster's skill regardless of the respective queries' truth values, in the mixture of beta distribution we use the parametrization for $t_n = 0$ in (9.3) for all mixture components.

Note that the shown data and distributions for the exemplary forecasters in Figure 9.4 are a result of both forecasters' skills and queries' difficulties. They would have looked different if the shown forecasters had provided forecasts for different queries. In particular, as the inferred hyperparameters c_1, c_2, d_1, d_2 show, the queries are on average rather easy. Therefore, the forecasters provide better forecasts than they would for more difficult queries, given their skill.

Forecaster 11 shown in Figure 9.4(a) is a skilled forecaster, who judges most of the given queries correctly and with high confidence. Matching the data, the corresponding estimated skill parameters generating the shown mixture of beta distribution are $\alpha_{11} = 0.05$ and $\beta_{11} = 0.12$. Thus, the corresponding skill mean is $\mu_{11}^{\alpha\beta} = 0.294 < 0.5$, showing that on average forecaster 11 provides unbiased forecasts. While the mean uncertainty of forecasts x^{11} is medium, due to the low values of α_{11} and β_{11}, the corresponding skill precision is low with $p_{11}^{\alpha\beta} = 0.17$. Accordingly, the forecasts are not concentrated around the skill mean, so forecaster 11 provides certain forecasts, most of which close to 1.

Forecaster 63 in Figure 9.4(b) is a very uncertain forecaster with parameters $\alpha_{63} = 1.33$ and $\beta_{63} = 1.37$. On average, forecaster 63 also provides unbiased forecasts with skill mean $\mu_{63}^{\alpha\beta} = 0.49$. However, his skill mean is very close to 0.5. Together with the high skill precision $p_{63}^{\alpha\beta} = 2.7$, the skill parameters reflect that forecaster 63 provides rather uncertain forecasts.



(a) Forecaster 11

(b) Forecaster 63

(c) Forecaster 10

Figure 9.4: The forecasts provided by three exemplary forecasters along with their estimated distribution according to our Skill-Difficulty Correlated Fusion Model. We show the relative frequency of the forecasts provided by forecaster 11 (a), 63 (b), and 10 (c) on the 179 queries in training split 1 as histograms. Forecasts for queries with truth value $t_n = 1$ are inverted in order to display a forecaster's skill regardless of the queries' truth values. Thus, a forecast is correct if $x_n^k < 0.5$. The curves plotted over the data illustrate the estimated distributions over all forecasts provided by forecaster k according to our model in (9.3). Since the difficulty parameters γ_n, δ_n are different for every single forecast in the shown data, the shown distributions are equally-weighted mixture of beta distributions consisting of 179 components according to (9.3) for $t_n = 0$ with skill parameters α_k, β_k for the respective forecaster k and the difficulty parameters γ_n, δ_n of the 179 queries in the training split.

For forecaster 10 shown in Figure 9.4(c) we inferred the skill parameters $\alpha_{10} = 0.19$ and $\beta_{10} = 0.08$. Thus, forecaster 10's skill mean is $\mu_{10}^{\alpha\beta} = 0.7$ with a skill precision of $p_{10}^{\alpha\beta} = 0.27$. The high skill mean above 0.5 indicates that forecaster 10 is on average biased and provides incorrect forecasts, given that we assume the queries' difficulties to have no impact on his forecasts. Interestingly, this cannot be seen when only looking at the data. The provided forecasts and the inferred distribution over them shown in Figure 9.4(c) suggest that forecaster 10 provides more correct forecasts than incorrect forecasts. The reason for this discrepancy is that there are more easy than hard queries in the KTeC data set, as noted above. Thus, an advantage of the proposed Skill-Difficulty Correlated Fusion Model is that it quantifies the forecasters' skills independently of the queries, so we can identify a biased forecaster even though the forecasts he provided do not look biased.

For all three forecasters shown in Figure 9.4 we see that the learned distributions over the forecasts provided by the specific forecasters, illustrated by the blue curves, fit the data visualized by the histograms well.

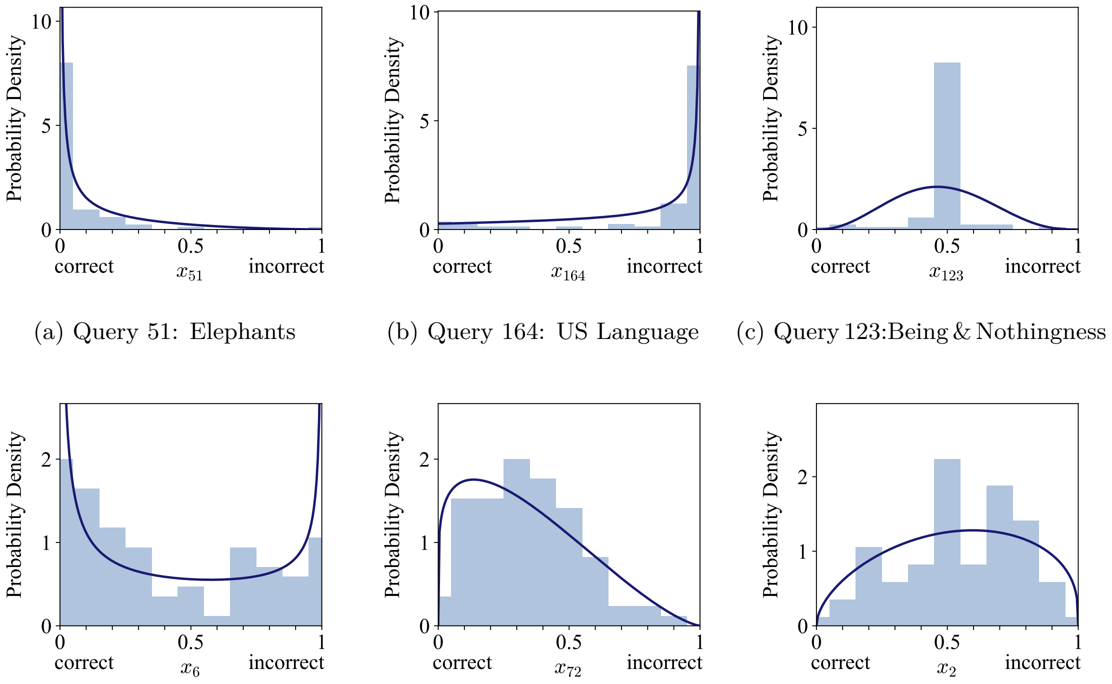
9.2.3.2 Queries' Difficulties

In Figure 9.5 we show the forecasts provided by all 85 forecasters for six specific queries in training split 1 of LOO cross-validation. As for the forecasters in Figure 9.4 we also inverted the forecasts for queries with truth value $t_n = 1$ in order to be able to illustrate the difficulty of the query regardless of its truth value t_n. Accordingly, forecasts $x_n^k < 0.5$ are correct forecasts. We plot the relative frequency of the respective forecasts provided by all forecasters for the specific query as histograms. The curves plotted over the histograms are the distributions according to our model in (9.3) over all forecasts provided for the specific query. While here, the difficulty parameters γ_n and δ_n are fixed for the specific query n, the skill parameters α_k and β_k are different for different forecasts since they are provided by different forecasters. Thus, the shown distributions are equally-weighted mixture of beta distributions. Their 85 mixture components are beta distributions according to (9.3) for $t_n = 0$ with difficulty parameters γ_n and δ_n for the respective query n and the skill parameters α_k and β_k of all 85 forecasters in the training split.

As for the forecasters in Figure 9.4, the data and distributions shown in Figure 9.5 would have looked different if the shown queries had been forecasted by different forecasters. Since the forecasters are more skilled than unskilled according to the estimated hyperparameters a_1, a_2, b_1, b_2, the forecasts for the shown queries are slightly better than they would be with less skilled forecasters, given their difficulty.

Query 51 ('Elephants are mammals', $t_{51} = 1$) shown in Figure 9.5(a) is a prototypical easy query. Almost all forecasters provided unbiased and certain forecasts. This is also reflected in the query's parameters $\gamma_{51} = 0.02$ and $\delta_{51} = 2.03$. Its corresponding difficulty mean $\mu_{51}^{\gamma\delta} = 0.01$ is very close to 0 with a high difficulty precision $p_{51}^{\gamma\delta} = 2.05$, implying unbiased, i.e., correct, and certain forecasts with low variance.

In contrast, query 164 ('The official language of the United States is English', $t_{164} = 0$) in Figure 9.5(b) is a trick query, which is answered incorrectly, i.e., with a strong bias, and with high certainty by the majority of forecasters. Its difficulty parameters are $\gamma_{164} = 0.77$ and $\delta_{164} = 0.03$. Thus, its difficulty mean is $\mu_{164}^{\gamma\delta} = 0.96$, close to 1, showing that the forecasts for query 164 are strongly biased and on average very certain. Since the difficulty



(a) Query 51: Elephants

(b) Query 164: US Language

(c) Query 123:Being & Nothingness

(d) Query 6: US Flag

(e) Query 72: Diamonds S. Africa

(f) Query 2: Rum Jamaica

Figure 9.5: The forecasts provided for six exemplary queries along with their estimated distribution according to our Skill-Difficulty Correlated Fusion Model. We show the relative frequency of the forecasts provided by all 85 forecasters for queries 51 (a), 164 (b), 123 (c), 6 (d), 72 (e), and 2 (f) in training split 1 as histograms. Forecasts for queries with truth value $t_n = 1$ are inverted in order to display a query's difficulty regardless of its truth value. Thus, a forecast $x_n^k < 0.5$ is a correct forecast. The curves plotted over the data are the distributions over the forecasts provided for the respective query according to our model in (9.3). Since the skill parameters α_k, β_k are different for every single forecast in the shown data, the shown distributions are equally-weighted mixture of beta distributions consisting of 85 components according to (9.3) for $t_n = 0$ with difficulty parameters γ_n, δ_n for the respective query n and the skill parameters α_k, β_k of the 85 forecasters.

precision $p_{164}^{\gamma\delta} = 0.8$ is additionally below 1, the forecasts are concentrated at 1, while some forecasters also provided correct forecasts close to 0.

Query 123 (‘*Being and Nothingness* was written in 1943’, $t_{123} = 1$) in Figure 9.5(c) is an unknown query for which most forecasters provide a forecast of 0.5. The corresponding difficulty parameters $\gamma_{123} = 3.27, \delta_{123} = 3.65$ reflect this. They result in an uncertain difficulty mean of $\mu_{123}^{\gamma\delta} = 0.47$ with a high difficulty precision of $p_{123}^{\gamma\delta} = 6.92$, leading to mostly uncertain forecasts concentrated around 0.5.

For query 6 (‘The national flag of the United States of America consists of thirteen horizontal stripes and fifty small five-pointed stars’, $t_6 = 1$) in Figure 9.5(d) more variable forecasts are provided. The query can be regarded as a rather easy query with forecasts being on average correct, i.e., unbiased, reflected in the difficulty parameters $\gamma_6 = 0.14, \delta_6 = 0.24$ with a difficulty mean of $\mu_6^{\gamma\delta} = 0.37$, lower than 0.5. Since the difficulty precision $p_6^{\gamma\delta} = 0.38$ is below one, the forecasts are not concentrated around the mean but tend toward 0 and 1, resulting in variable forecasts, many of which uncertain.

Similar to query 6, query 72 (‘South Africa is the world’s largest diamond producing country’, $t_{72} = 1$) shown in Figure 9.5(e) is also an uncertain unbiased query, which is on average forecasted correctly. Its difficulty parameters $\gamma_{72} = 0.98, \delta_{72} = 2.14$ result in a difficulty mean $\mu_{72}^{\gamma\delta} = 0.31$, similar to that of query 6. However, the difficulty precision $p_{72}^{\gamma\delta} = 3.12$ is much higher, so the forecasts are more concentrated around the difficulty mean $\mu_{72}^{\gamma\delta} = 0.31$.

Query 2 (‘Rum is Jamaica’s principal export’, $t_2 = 0$) in Figure 9.5(f) is an example of a trick query, for which uncertain forecasts are provided. Accordingly, its difficulty parameters are $\gamma_2 = 1.35$ and $\delta_2 = 1.12$, with a difficulty mean $\mu_2^{\gamma\delta} = 0.55 > 0.5$, showing a bias and high average uncertainty, and a high difficulty precision $p_2^{\gamma\delta} = 2.47$, leading to uncertain forecasts concentrated around the uncertain difficulty mean.

For all six shown queries in Figure 9.5 we see that the estimated distributions over the forecasts provided for the specific queries, shown with the blue curves drawn over the data, fit the data well.

9.2.4 Fusion Performance

To evaluate the fusion performance, we compare the Skill-Difficulty Correlated Fusion Model to the independent beta fusion models we proposed in previous work (Trick, Rothkopf, & Jäkel, 2023a, see Chapter 8) and the related fusion models by Turner et al. (2014) on the KTeC data set.

The independent beta fusion model represents the forecasts x_n^k with a beta distribution with parameters α_j^k, β_j^k conditioned on the truth value $t_n = j$ according to (9.1). Different variants of this model are either hierarchical or non-hierarchical and model the forecasts with either asymmetric or symmetric beta distributions for $t_n = 0$ and $t_n = 1$. The resulting four beta fusion models are the Hierarchical Beta Fusion Model (HB), the non-hierarchical Beta Fusion Model (B), the Hierarchical Symmetric Beta Fusion Model (HSB), and the non-hierarchical Symmetric Beta Fusion Model (SB).

In our previous work (Trick, Rothkopf, & Jäkel, 2023a, see Chapter 8), we also compared the performances of our independent beta fusion models against the fusion models proposed by Turner et al. (2014). Their approach compares different Bayesian models that either first calibrate the provided forecasts and then fuse them or first fuse the forecasts and then calibrate the fused forecast. Fusion is realized with averaging. Hierarchical and non-hierarchical models as well as fusion of log-odds or probabilities are compared. The resulting models are Average Then Calibrate (ATC), Calibrate Then Average (CTA), Calibrate Then Average on Log-Odds (CTALO), Hierarchical Calibrate Then Average (HCTA), and Hierarchical Calibrate Then Average on Log-Odds (HCTALO). Turner et al. (2014) additionally compared their models to the baseline method Unweighted Linear Opinion Pool (ULINOP). ULINOP can be augmented with the baseline method Probit Average (PAVG) (Satopää et al., 2023), which transforms the forecasts with probit transformation before averaging them. Details on the implementation and evaluation of the independent beta fusion models, the Turner fusion models, and the two baseline methods can be found in the work of Trick, Rothkopf, and Jäkel (2023a, see Chapter 8).

Since the forecasts provided by different forecasters in the KTeC data set are correlated, in our comparison we do not only consider mere forecasting performance but put a special focus on the forecasts' uncertainty and potential overconfidence. In order to do this, we quantify fusion performance with the performance measures Brier score, mean absolute error, and entropy.

The Brier score (Brier et al., 1950) is commonly used for measuring the performance of human forecasters (Baron et al., 2014; Hanea et al., 2021; Karvetski et al., 2013; Ranjan & Gneiting, 2010; Satopää, 2022; Turner et al., 2014). It describes the mean squared error between the provided probability forecasts x_n and the corresponding truth values t_n,

$$\mathrm{BS} = \frac{1}{N}\sum_{n=1}^{N}(x_n - t_n)^2. \tag{9.9}$$

A lower Brier score indicates a better performance, with $\mathrm{BS} = 0$ being the best and $\mathrm{BS} = 1$ being the worst attainable score. The Brier score is a strictly proper scoring rule (A. H. Murphy, 1973) and is thus optimized if the forecasters provide their true beliefs of the probability instead of deliberately making them more or less extreme. Thus, it punishes over- and underconfident forecasts.

Mean absolute error (MAE) (Canbek et al., 2022; Ferri et al., 2009) quantifies the absolute difference between the provided forecast x_n and the truth value t_n,

$$\mathrm{MAE} = \frac{1}{N}\sum_{n=1}^{N}|x_n - t_n|. \tag{9.10}$$

Accordingly, it also ranges between 0 and 1, with $\mathrm{MAE} = 0$ being the best and $\mathrm{MAE} = 1$ being the worst possible score. Compared to the Brier score, MAE is more straightforwardly interpretable and more robust to outliers (Canbek et al., 2022). However, unlike Brier score, it is not a proper scoring rule (Buja et al., 2005) and thus rewards overconfident forecasts.

In contrast to Brier score and MAE, (Shannon) entropy does not directly measure the forecasting performance but only the uncertainty of the forecasts, independent of their correctness. It is defined for a single binary forecast as $\mathrm{h}(x_n) = -x_n \log(x_n) - (1 - x_n)\log(1 - x_n)$, so for N considered queries we compute the mean entropy H as

$$\mathrm{H} = \frac{1}{N} \sum_{n=1}^{N} -x_n \log(x_n) - (1 - x_n)\log(1 - x_n). \tag{9.11}$$

The minimum mean entropy is 0 and indicates fully certain forecasts being either 0 or 1. The maximum mean entropy for binary forecasts is approximately 0.7 and is obtained if all forecasts are 0.5 and thus fully uncertain. Since entropy explicitly quantifies uncertainty, it is well-suited to check for overconfident forecasts.

Figure 9.6 shows the means and standard errors of the mean of Brier score, mean absolute error, and entropy on the KTeC data set for all considered fusion models, i.e., the Skill-Difficulty Correlated Fusion Model, the independent beta fusion models, the models proposed by Turner et al. (2014), and the baseline methods ULINOP and PAVG. The main result from evaluating the independent beta fusion models is that the Hierarchical Symmetric Beta Fusion Model performs best of all beta fusion models. This is also shown in Figure 9.6 for Brier score and MAE. However, while in terms of MAE it also clearly outperforms the models by Turner et al. (2014), its Brier score (BS = 0.141) is worse than that of the best Turner model HCTALO (BS = 0.125). Since MAE rewards overconfidence and Brier score punishes it, these results are a clear indication of overconfidence. The entropy plot in Figure 9.6 further strengthens this conclusion because all beta fusion models show entropies close to 0, indicating very certain fused forecasts. In contrast, the fusion models by Turner et al. (2014) show higher entropies between 0.4 and 0.7 and thus result in more uncertain fused forecasts.

Our Skill-Difficulty Correlated Fusion Model, which models the correlations between forecasters, performs comparably to all independent beta fusion models and better than the

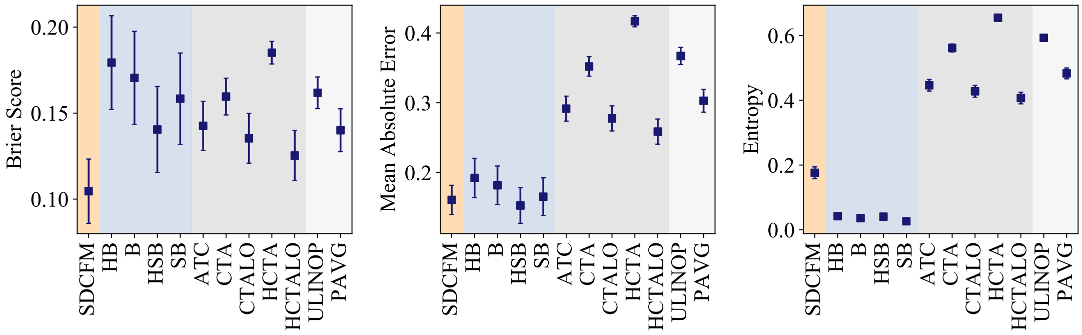


Figure 9.6: Fusion performance of the Skill-Difficulty Correlated Fusion Model in comparison to related Bayesian fusion models on the KTeC data set. We compare the means and standard errors of the mean of Brier score, mean absolute error, and entropy of the Skill-Difficulty Correlated Fusion Model (SDCFM), the four independent beta fusion models including the Hierarchical Beta Fusion Model (HB), the non-hierarchical Beta Fusion Model (B), the Hierarchical Symmetric Beta Fusion Model (HSB), and the non-hierarchical Symmetric Beta Fusion Model (SB), the models by Turner et al. (2014), Average Then Calibrate (ATC), Calibrate Then Average (CTA), Calibrate Then Average using Log-Odds (CTALO), Hierarchical Calibrate Then Average (HCTA), and Hierarchical Calibrate Then Average on Log-Odds (HCTALO), and the two baseline methods Unweighted Linear Opinion Pool (ULINOP) and Probit Average (PAVG).

Turner models according to MAE, with MAE = 0.161. However, while the independent beta fusion models cannot outperform the Turner models in terms of Brier score, the Skill-Difficulty Correlated Fusion Model performs best of all considered fusion models with a Brier score of BS = 0.105, compared to BS = 0.125 for the second best model HCTALO by Turner et al. (2014). Looking at entropy, we also see that the Skill-Difficulty Correlated Fusion Model is less uncertain than the Turner models, but more uncertain than the independent beta fusion models with an entropy of H = 0.176.

9.3 Discussion and Conclusion

In this work, we introduced a Bayesian model for combining correlated human forecasts. The model assumes that a forecast provided by a forecaster for some query is dependent on both the forecaster's properties, which we call his skill, and the query's properties, which we call its difficulty. Further, we assume that the forecasts provided by different forecasters are correlated because they provide forecasts for the same queries. Accordingly, the proposed Skill-Difficulty Correlated Fusion Model represents the forecasts with a beta distribution conditioned on their truth value with skill parameters specific for each forecaster and difficulty parameters specific for each query. By explicitly modeling the forecasters' skills as well as the queries' difficulties, the Skill-Difficulty Correlated Fusion Model can model the correlation between the forecasts provided by different forecasters. Thus, compared to previous models that assume independence, the new model can better represent human forecasts, which are often correlated (Berger, 1985; Hogarth, 1978; Lichtendahl Jr et al., 2022; Wilson & Farrow, 2018; Winkler et al., 2019; Wiper & French, 1995). Also, it explicitly quantifies the skills of individual forecasters and the difficulties of individual queries and thus allows analyzing and comparing different forecasters independently of the queries their forecasts are provided for or different queries independently of the forecasters who provide their respective forecasts. After learning the model parameters from observed training data, the Skill-Difficulty Correlated Fusion Model can be used to combine new unseen forecasts provided by different forecasters by inferring the posterior distribution over their truth value. Given that the model assumptions are correct, fusion with the Skill-Difficulty Correlated Fusion Model is normative. Thus, the forecasters' skills as well as their correlation can be normatively considered for fusion.

We evaluated the Skill-Difficulty Correlated Fusion Model on a data set consisting of 85 forecasters who each provided forecasts for 180 queries. We compared our model's fusion performance in terms of Brier score, mean absolute error, and entropy to related Bayesian fusion models, which do not consider correlation (Trick, Rothkopf, & Jäkel, 2023a, see Chapter 8; Turner et al., 2014). While the previously proposed independent beta fusion models (Trick, Rothkopf, & Jäkel, 2023a, see Chapter 8) turned out to be overconfident and therefore outperformed the models proposed by Turner et al. (2014) in terms of mean absolute error but could not outperform them in terms of Brier score, the Skill-Difficulty Correlated Fusion Model performs best of all compared fusion models according to both mean absolute error and Brier score. In addition, it leads to more uncertain fused forecasts than the independent beta fusion models, shown by a higher entropy. Thus, modeling correlations between forecasters in our Skill-Difficulty Correlated Fusion Model fixes the overconfidence problem of the independent beta fusion models.

For exemplary forecasters and queries we showed how the Skill-Difficulty Correlated Fusion Model quantifies the forecasters' and queries' properties, i.e., their skills and difficulties, and how it fits the forecasting data. However, as seen for example in Figure 9.4(c), our model cannot represent a peak at 0.5 in the histogram of provided forecasts, unless all forecasts are concentrated around 0.5 anyway. This peak or blip in the histogram is called the 50-50 blip and is caused by the forecasters' high epistemic uncertainty (Bruine de Bruin et al., 2002; Fischhoff & Bruine De Bruin, 1999; Karvetski et al., 2013). If the forecasters have absolutely no knowledge about the truth value of a query, i.e., they have high epistemic uncertainty, they choose 0.5. Even for skilled forecasters this will happen for quite a few queries, at least if the queries' topics differ from each other. If the forecasts are predictions of the occurrence of future events, whose outcomes cannot be known yet, the 50-50 blip can even be amplified, because it comprises both the forecasts of people that have absolutely no knowledge, i.e., with high epistemic uncertainty, and the predictions that are explicitly intended to express a probability of 0.5, i.e., with high aleatoric uncertainty (Karvetski et al., 2013). Thus, future work should extend the Skill-Difficulty Correlated Fusion Model to explicitly represent a potential peak of forecasts at 0.5, i.e., the 50-50 blip. This might not only improve the model's fit to real forecasting data but also further improve the fusion performance due to better modeling of the forecasts.

10 GENERAL DISCUSSION

This thesis provides a normative framework for Bayesian fusion of probabilistic forecasts, either provided by classifiers or humans. First, we introduced a Bayesian model of classifier fusion that formalizes how to optimally combine the categorical output distributions returned by probabilistic classifiers and, in particular, how to optimally reduce their uncertainty. Given progressively more general assumptions, we derived the optimal fusion behavior of classifiers according to their individual properties, i.e., their uncertainty, bias, and variance. In addition, we explicitly modeled the correlation between individual classifiers with a newly introduced probability distribution, the correlated Dirichlet distribution. With the proposed framework we could specify how uncertainty should be reduced depending on the individual classifiers' uncertainty, bias, variance, and correlation. Moreover, our model outperformed the closest related Bayesian fusion model on simulated as well as real data sets.

Second, we proposed a parameter estimation method for the bivariate beta distribution, which is a special case of the correlated Dirichlet distribution that we introduced for representing correlated classifiers. Instead of correlated categorical distributions, the bivariate beta distribution models simple probabilities with a potential correlation, thus e.g., the outputs of binary probabilistic classifiers. While this distribution has already been introduced in previous work, the proposed parameter estimation method is only approximate and inaccurate. Therefore, in this thesis, we derived the bivariate beta distribution's product moments and exact covariance, which can be computed numerically. Using the derived covariance, we proposed parameter inference using moment matching.

Third, we applied the Bayesian classifier fusion method Independent Opinion Pool, which is a special case of the normative Bayesian classifier fusion model derived in this thesis, to multimodal intention recognition in human-robot interaction. We trained individual probabilistic intention classifiers for four different modalities, i.e., speech, gestures, gaze directions, and scene objects, and fused them Bayes optimally using Independent Opinion Pool. By this, we enabled a 7-DoF robot arm to recognize intentions from multimodal data while explicitly considering each modality's uncertainty and reducing the uncertainty over the intention to be recognized.

Fourth, we investigated if Bayes optimal classifier fusion using Independent Opinion Pool can also improve interactive reinforcement learning using multimodal human advice. We combined individual probabilistic action advice classifiers for the two modalities speech and gestures and fused them in a Bayes optimal way using Independent Opinion Pool. In order to exploit the uncertainty of the resulting fused distribution, instead of choosing the action with the highest probability, we chose the action to be executed by sampling from it. In a simulated grid-world scenario as well as a real-world interactive task between a human and a 7-DoF robot arm with 10 human participants, we could show that the proposed approach outperforms the closest related approach in terms of learning speed.

Fifth, we learned to recognize a specific human intention from natural human behavior, which frequently occurs in interactions between human and robot: the intention to start an interaction with a robot, or in short, the intention for interaction. In an experiment with 21 human subjects we collected natural human behavior data while the human subjects repeatedly showed an intention for interaction towards a two-armed robot. In contrast to previous approaches, we included different tasks and interactions and different human positions and orientations towards the robot in our data set in order to be able to generalize the detection to different situations. We recorded multimodal data from the two modalities speech and body poses and compared different unimodal and multimodal probabilistic classifiers. Bayesian fusion of the unimodal speech and body pose classifiers using Independent Opinion Pool resulted in the best detection performance.

Sixth, we applied Bayesian fusion to subjective probability estimates provided by humans. In a normative generative model we represented the human forecasts with beta distributions conditioned on their truth value. We compared different variants of this model, i.e., hierarchical and non-hierarchical models as well as models that assume symmetric or asymmetric beta distributions. While previous approaches explicitly calibrate given human forecasts using specific calibration functions, we show that our model implicitly calibrates forecasts with the beta calibration function, which accommodates commonly used calibration functions as a special case. For evaluating the proposed family of normative models, we introduced a new forecasting data set. On this and another data set, we showed that the best of our normative models, a hierarchical model that assumes symmetric beta distributions, performs comparably to related Bayesian models. However, because our models mistakenly assume conditional independence of the forecasts provided by different forecasters given the truth value, they reduce too much uncertainty and are thus overconfident.

Therefore, seventh, we proposed a Bayesian model for combining correlated subjective probability estimates. This model also represents the forecasts with beta distributions conditioned on their truth value, but with special parameters representing the skill of each forecaster as well as the difficulty of each query for which forecasts are provided. By explicitly modeling the forecasters' skills and the queries' difficulties, the correlation between forecasts provided by different forecasters can be modeled and considered for fusion respectively. Evaluations on a data set consisting of human forecasts showed that our model can quantify the forecasters' skills and the queries' difficulties. Moreover, its fusion performance is improved compared to our previous models for fusing independent forecasts and other related models.

10.1 General Findings

In general, this thesis's starting point is the fundamental computational problem of combining uncertainties in the form of probabilistic forecasts. Humans, robots, as well as other AI systems and agents are constantly facing uncertainty in our world (D. R. Bach & Dolan, 2012; Lindley, 2013; Russell & Norvig, 2010; Vilares & Kording, 2011), e.g., in their perception, decision making, or when receiving estimates provided by other systems or humans. This uncertainty can be reduced by combining multiple sources of information. In this thesis, we used human perception, which can be conceptualized as reducing uncertainty in a rational way, as an inspiration for improving artificial information processing systems including AI systems. In particular, we developed rational, i.e., Bayesian, computational models and inference algorithms for AI systems in order to enable them to combine uncertain information in a rational way.

We proposed normative Bayes optimal fusion of probabilistic forecasts either provided by classifiers or humans using Bayesian methods and models. For this, we built on existing methods, e.g., the Bayesian fusion method Independent Opinion Pool, and introduced new Bayesian models, for which we also proposed a new probability distribution. While these models quantify how the forecasts should be fused, i.e., how uncertainty should be reduced correctly, the works included in this thesis demonstrate that Bayes optimal fusion also improves performance compared to other fusion methods. It is correct, and it performs well.

In particular, our proposed Bayesian fusion model that explicitly models the correlation between different individual classifiers with a correlated Dirichlet distribution outperforms the related model by Pirs and Strumbelj (2019) (Chapter 3). Further, Bayesian fusion using Independent Opinion Pool increases the learning speed of multimodal interactive reinforcement learning in human-robot interaction compared to the related heuristic fusion approach used by Cruz et al. (2018) (Chapter 6). When detecting a human intention for interaction in human-robot interaction from multimodal data, decision fusion of the unimodal intention classifiers using the Bayesian Independent Opinion Pool is better suited than feature fusion (Chapter 7). Finally, our Bayesian model for combining correlated human forecasts also outperforms all related models proposed by Turner et al. (2014) (Chapter 9).

While we showed that fusion assuming independence can already produce good results (Chapter 5-7), this thesis particularly revealed that considering the correlation between probabilistic forecasts is crucial because it avoids overestimating uncertainty reduction, which leads to overconfidence (Chapter 3, 8, 9). While this was already known previously (Wilson, 2017), the models presented in this thesis also explicitly provide normative solutions to this problem (Chapter 3, 9). In particular, we quantify how correlated probabilistic forecasts should be fused and how this fusion affects uncertainty reduction (Chapter 3).

10.2 Implications

The works presented in this thesis contributed to multiple research fields, among them mathematics, artificial intelligence, robotics, and cognitive science. This is not only reflected in the different conference proceedings and journals our works are published in, i.e., *METRON* journal for mathematics, *AISTATS* conference for artificial intelligence, *IROS* conference, *Robotics and Automation Letters* journal, and *RO-MAN* conference for robotics, and *Judgment and Decision Making* journal for cognitive science. The implications of our work for the different research fields are outlined below.

10.2.1 Mathematics

Within mathematics, a large number of different probability distributions have been developed to quantify uncertainty over probability estimates that might be correlated. In addition to bivariate generalizations of the Kumaraswamy distribution (Arnold & Ghosh, 2017), bivariate beta-generated distributions (Samanthi & Sepanski, 2019), or multivariate Gaussian distributions with logistic transformations (Bordley, 1982; Clemen & Winkler, 1987; French, 1980; Pirs & Strumbelj, 2019), many previous approaches deal with different kinds of bivariate beta distributions (Arnold & Ghosh, 2017; Arnold & Ng, 2011; Bran-Cardona et al., 2011; David Sam Jayakumar et al., 2019; El-Bassiouny & Jones, 2009; Gupta et al., 2011; Gupta & Wong, 1985; Jones, 2002; Koutoumanou et al., 2017; Libby & Novick, 1982; Magnussen, 2004; Nadarajah & Kotz, 2005; Nadarajah et al., 2017; Olkin & Liu, 2003; Olkin & Trikalinos, 2015; Orozco-Castañeda et al., 2012; Samanthi & Sepanski, 2019; Sarabia & Castillo, 2006; Ting Lee, 1996). In the present thesis, we have contributed an algorithm for exact parameter estimation for one of these bivariate beta distributions, which can model random variables with arbitrary beta marginals and positive correlation (Magnussen, 2004).

Furthermore, we have extended this bivariate beta distribution to a new probability distribution for representing the uncertainty over correlated probability estimates with higher dimensionality, i.e., over correlated categorical probability distributions. The new correlated Dirichlet distribution models positive correlations between random vectors $\boldsymbol{x^1} = (x_1^1, \ldots, x_J^1)$ and $\boldsymbol{x^2} = (x_1^2, \ldots, x_J^2)$ with arbitrary marginal Dirichlet distributions. Previous generalizations of the Dirichlet distribution only focused on more flexible correlations between random vector entries $x_1, \ldots, x_J$ of a Dirichlet variable $\boldsymbol{x}$ (Connor & Mosimann, 1969; Linderman et al., 2015; Wong, 1998).

The bivariate beta and the correlated Dirichlet distribution are not only interesting from a theoretical point of view but can also be applied to solve problems besides modeling probabilistic forecasts. For instance, the correlated Dirichlet distribution can generate stochastic matrices with different rows or columns being Dirichlet-distributed and correlated, which can be useful for Markov processes, optimal control, or reinforcement learning.

10.2.2 Artificial Intelligence

Classification is fundamental in Artificial Intelligence (AI) (Sharma & Singh, 2023). Ensembles of classifiers, which fuse individual classifiers in order to improve classification performance (Bishop, 2006; Dietterich, 2000; Hamed & Akbari, 2018; Kittler et al., 1998; Mohandes et al., 2018; Pirs & Strumbelj, 2019), are known to be the most successful classification systems (Kuncheva, 2014; Pirs & Strumbelj, 2019).

Due to partial observability and nondeterminism, uncertainty is also a fundamental challenge in AI (Abdar et al., 2021; Bhatt et al., 2021; Kompa et al., 2021; Russell & Norvig, 2010). In particular, classifier fusion methods have to fuse uncertain outputs of probabilistic classifiers while avoiding overconfidence caused by the usual correlation between individual classifiers to be fused (Jacobs, 1995; Kim & Ghahramani, 2012; Wilson, 2017).

In addition to a variety of different algorithms for fusing correlated probabilistic classifiers (e.g., Petrakos et al., 2000; Prabhakar & Jain, 2002; Safont et al., 2019; Srinivas et al., 2009; Ulaş et al., 2012), previous work proposed Bayesian generative models for classifier fusion, which can potentially formalize how to combine classifier outputs in a normative or Bayes optimal way (Kim & Ghahramani, 2012; Nazabal et al., 2016; Pirs & Strumbelj, 2019). In particular, Pirs and Strumbelj (2019) propose a Bayesian model for fusing probabilistic classifiers that models their correlation. However, the model requires transformations and mixture distributions to represent the correlated categorical output distributions. Most importantly, a normative description of how to combine correlated probabilistic classifiers and how to reduce uncertainty depending on the correlation is not provided in their approach. The present thesis closed this gap with the proposed normative Bayesian model of correlated probabilistic classifier fusion, which represents the classifiers' categorical output distributions with the correlated Dirichlet distribution, a conjugate prior of the categorical distribution that allows representing correlations between categorical distributions. This model explicitly quantifies how probabilistic classifiers should be fused Bayes optimally depending on their correlation as well as their individual uncertainty, bias, and variance and additionally outperforms the related fusion model by Pirs and Strumbelj (2019).

The proposed normative model of classifier fusion can be applied to any conceivable classification problem. Furthermore, it can serve as an orientation for future fusion methods on how uncertainty should normatively be considered for fusion, in particular if the individual classifiers are correlated.

In this way, it can affect a multitude of potential applications of classification, among them image classification, text classification, email spam filtering, object recognition, face detection, speaker identification, handwriting recognition, and medical diagnosis (Almazroi et al., 2020). Classifier fusion is of particular interest for critical applications of classification where a misclassification can be very costly, such as medical diagnosis or person recognition in autonomous driving (Mohandes et al., 2018). Especially in such critical applications, knowing the correct uncertainty of a classifier fusion system is crucial since overconfidence resulting from incorrect uncertainty reduction could cause wrong decisions and actions that might harm people (Guo et al., 2017; Kompa et al., 2021).

10.2.3 Robotics

The research field of Human-Robot Interaction (HRI) is a subdiscipline of robotics that aims to understand, design, and evaluate robotic systems in interactions and communications with humans (Goodrich & Schultz, 2008). Successful human-robot interaction requires robots with the ability to recognize human intentions (Hofmann & Williams, 2007). Since humans communicate their intentions using multiple modalities, such as speech and gestures (Barthelmess et al., 2006; Chandrasekaran et al., 2009; De Ruiter et al., 2012; So et al., 2009; Todisco et al., 2021), an intuitive and natural interaction between humans and robots requires also robots to recognize intentions from multimodal data (Goodrich & Schultz, 2008; Stiefelhagen et al., 2004).

Several approaches for multimodal intention recognition in HRI have been proposed (Cruz et al., 2016; Cruz et al., 2018; Foster et al., 2017; Mollaret et al., 2015; Mollaret et al., 2016; Rodomagoulakis et al., 2016; Vaufreydaz et al., 2016; W. Xu et al., 2015; Yu et al., 2015; Zlatintsi et al., 2018). Among these approaches, some even perform decision fusion by combining the outputs of individual classifiers (Cruz et al., 2016; Cruz et al., 2018; Rodomagoulakis et al., 2016; Yu et al., 2015; Zlatintsi et al., 2018). However, they either do not consider uncertainty reduction at all (Rodomagoulakis et al., 2016; Yu et al., 2015; Zlatintsi et al., 2018) or apply heuristic rules in order to reduce uncertainty through multimodal fusion (Cruz et al., 2016; Cruz et al., 2018). In human-robot interaction, disregarding uncertainty about the human's intentions and reducing it incorrectly could lead to inappropriate robot reactions, which in the worst case can cause dangerous situations.

In order to reduce uncertainty correctly, in this thesis, we have applied Bayesian classifier fusion to multimodal human-robot interaction. We have shown that Bayesian fusion with the method Independent Opinion Pool improves the recognition of human intentions from multimodal data in human-robot interaction and increases the learning speed in interactive reinforcement learning with multimodal detection of human action advice. Furthermore, a newly introduced data set for multimodal recognition of the intention for interaction from natural human behavior can serve for future investigations in intention recognition.

The proposed approaches for multimodal human-robot interaction lower the barriers for people to interact with robots by allowing them to interact multimodally as they are accustomed to doing (Barthelmess et al., 2006; Chandrasekaran et al., 2009; De Ruiter et al., 2012; Ernst & Bülthoff, 2004; So et al., 2009; Todisco et al., 2021). Moreover, correctly considering uncertainty in this multimodal interaction leads to more safe human-robot interaction (Baek & Kröger, 2023; Baek et al., 2023).

As outlined in Chapter 5, a promising application of Bayesian fusion in multimodal human-robot interaction is elderly assistance. Globally, societies are aging (Christoforou et al., 2020; Grinin et al., 2023; C. Johnston, 2022). This severely challenges traditional health-care and elderly care (Dino et al., 2022). One possible solution to this problem is the use of assistive robots. On the one hand, they can support elderly people in their homes in order to help them remain independent and stay in their own homes for a longer time (Graf et al., 2004; Martinez-Martin & Costa, 2021). In addition to improving the quality of life of affected elderly people who do not want to move to an elderly home, assistive robots at home also save money since a costly treatment in an elderly home can be avoided (Graf et al., 2004). Potential support tasks for such assistive robots are, among others, fetching and carrying things for the elderly users, serving drinks, managing their daytime, or supporting them in standing up or walking (Graf et al., 2004). On the other hand, assistant robots can potentially support caregivers in elderly homes (C. Johnston, 2022; Niemelä & Melkas, 2019), e.g., by fetching and bringing required items (Niemelä & Melkas, 2019). In the best case, this relieves the human caregivers, gives them more time to devote to the elderly people, and improves their working conditions.

In the caregiving context, intuitive and natural communication as enabled by multimodal interaction is of particular interest since potential users, i.e., the caregivers and in particular the elderly people themselves, cannot be assumed to be technically minded. The explicit consideration and reduction of uncertainty is even more important because high uncertainty can lead to wrong actions, which can irritate and in the worst case harm the possibly vulnerable elderly users.

However, note that in addition to elderly assistance, the proposed methods for multimodal human-robot interaction using Bayesian fusion are also applicable to industry settings, where humans and robots collaborate as coworkers (Vojić, 2020), as well as to all other conceivable applications of human-robot interaction.

10.2.4 Cognitive Science

Cognitive and psychological research aims to understand what is happening inside the mind of individuals in judgment and decision making tasks and how optimal performance of the collective intelligence of a group of individuals can be achieved (Steyvers & Miller, 2015). In particular, the aggregation of judgments can be treated as a cognitive modeling problem (M. D. Lee & Danileiko, 2014).

When aggregating probabilistic judgments, the calibration of individual forecasters is of particular importance: Most people are miscalibrated (Morgan, 2014), but their miscalibration should not affect the combination of the combination of their judgments (M. D. Lee & Danileiko, 2014). In addition, similar to classifiers, human forecasts are usually correlated (Berger, 1985; Hogarth, 1978; Lichtendahl Jr et al., 2022; Wilson & Farrow, 2018; Winkler et al., 2019; Wiper & French, 1995), which can lead to overconfidence if it is ignored when combining the forecasts (Wilson, 2017). The correlation of forecasts is one of the major challenges in forecast aggregation (McAndrew et al., 2021).

While there are many ad-hoc rules for the aggregation of probability estimates, e.g., weighted linear opinion pools (Budescu & Chen, 2015; Budescu & Rantilla, 2000; Cooke, 1991; Hanea et al., 2021; Karvetski et al., 2013; Ranjan & Gneiting, 2010) and multiplicative pooling methods (Berger, 1985; Dietrich & List, 2016), also a multitude of Bayesian models for combining subjective probability estimates have been proposed (Babic et al., 2022; Bordley, 1982; Clemen & Winkler, 1987; Di Bacco et al., 2003; French, 1980; Hanea et al., 2021; M. D. Lee & Danileiko, 2014; Lichtendahl Jr et al., 2022; Lindley, 1985; Satopää, 2022; Satopää et al., 2016; Turner et al., 2014; J. Wang et al., 2021). Among these models, there are models that do not consider a potential correlation between forecasts (Hanea et al., 2021; M. D. Lee & Danileiko, 2014; Lindley, 1985; Turner et al., 2014) and others that explicitly model this correlation (Babic et al., 2022; Bordley, 1982; Clemen & Winkler, 1987; Di Bacco et al., 2003; French, 1980; Lichtendahl Jr et al., 2022; Satopää, 2022; Satopää et al., 2016; J. Wang et al., 2021).

In particular, Turner et al. (2014) proposed some supervised Bayesian models for combining probabilistic forecasts with discrete truth values that combine the forecast by averaging them and calibrate them using the Linear-in-Log-Odds (LLO) calibration function. However, these fusion models are not motivated normatively, i.e., with a generative model of how the truth value generated the human forecasts. Also note that this model does not represent correlations between forecasts. An example of a normative generative model of forecast aggregation is a model proposed by Lindley (1985), which transforms the probability estimates to log-odds and represents them with a Gaussian distribution given the truth value. This approach can also be extended to model the log-odds with multivariate Gaussian distributions, which enables modeling the correlation between forecasts (Bordley, 1982; Clemen & Winkler, 1987; French, 1980).

In the present thesis, we have contributed a normative model for combining independent subjective probability estimates that directly models the probabilities with a beta distribution conditioned on their truth value without the need for any transformation. We show that this model implicitly calibrates the forecasts with the beta calibration function, which accommodates the LLO calibration function as a special case. In a second step, we have extended this normative model of forecast combination in order to also model the correlation between forecasts. The proposed model explicitly represents the skills of individual forecasters in terms of their bias, variance, and uncertainty as well as the difficulty of the queries for which forecasts are provided. In this way, correlations between forecasts provided by different forecasters can be modeled and considered for fusion respectively.

The forecast aggregation models introduced in this thesis allow drawing conclusions about human forecasting behavior, about the considered forecasters' calibration, their skill in terms of bias, variance, and uncertainty, and the correlation between different human forecasters. In particular, the skills and calibrations of individual forecasters can provide useful insights about the cognitive characteristics of individuals (M. D. Lee & Danileiko, 2014). Combining new unseen human forecasts using the proposed models is Bayes optimal given the assumptions of the models and can consider the forecasters' properties, such as their skill or correlation, for fusion. In addition, our model for Bayesian combination of correlated forecasts outperforms alternative models on a data set consisting of human subjective probability estimates.

We provided this forecasting data set as another contribution of this thesis. It can serve for future investigations of human forecasting behavior and for the evaluation of other fusion models that might be proposed in the future. In contrast to previous forecasting data sets (Graefe, 2018; Hanea et al., 2021; Karvetski et al., 2013; Prelec et al., 2017; Turner et al., 2014), which are limited in the number of forecasts per forecaster, the new data set includes the forecasts of 85 forecasters on 180 queries.

Since human probabilistic forecasts provided by domain experts are crucial in many different fields, among them finance, business, marketing, engineering, meteorology, environmental science, public health, and politics (McAndrew et al., 2021), the proposed fusion models are also applicable in all these domains. Combining the experts' forecasts Bayes optimally while explicitly considering their individual properties as well as their correlation can improve investing, planning, risk management and safety, and decision making. In addition, the fused forecasts can be used to build rule-based AI systems (Masri et al., 2019). Another interesting potential future application of forecast aggregation might be to combine peoples' confidence ratings on statements on social media platforms (N. Kriegeskorte & P. Stinson, personal communication, August 10, 2023). Social media enables users to connect with each other and to share and receive information (X. Zhang & Ghorbani, 2020). However, a significant portion of the information shared on social media is fake news (X. Zhang & Ghorbani, 2020), which is a substantial threat to our society (Qu et al., 2024; Soga et al., 2024; X. Zhang & Ghorbani, 2020). On the one hand, fake news can be automatically detected using classifiers (e.g., Qu et al., 2024; Soga et al., 2024), which are, however, challenged by the large amount of news and the high diversity of information (Qu et al., 2024). On the other hand, peoples' confidence ratings on the correctness of statements could easily be collected in large numbers on social media platforms, while fact-checking systems could provide corresponding truth values. Based on these data our models could learn the users' calibration, skills, and correlations and exploit them to normatively fuse their confidence ratings. The resulting fused confidence rating might then serve as an orientation for other users to detect fake news.

10.3 Limitations and Future Work

The methods and findings presented in this thesis, along with their limitations, suggest multiple interesting ideas for future research. In the following, we will discuss the main limitations of the present thesis and explore promising directions for future work respectively.

The proposed Bayesian fusion models in Chapters 3, 8, and 9 explicitly model the behavior of individual classifiers or humans in terms of their bias, variance, and uncertainty. By this, as we showed in Chapter 8, they implicitly calibrate the forecasts when fusing them. Thus, given correct model assumptions, the fused forecast according to these Bayesian models should be correct, i.e., calibrated (Satopää, 2022). However, for the ad-hoc Bayesian fusion method Independent Opinion Pool in Chapter 3 and the robotic applications in Chapters 5-7, which rely on Independent Opinion Pool for fusion, we do not learn a model of the base classifiers' output distributions and thus cannot calibrate them. Thus, for these works it has to be taken in mind that the fused uncertainty is only correct given the output distributions provided by the individual modalities' base classifiers. Future work could explicitly calibrate the individual base classifiers for different modalities in Chapters 5-7 before fusing them or directly model them in order to implicitly calibrate them while fusing. The latter could be realized with the Independent Fusion Model discussed in Chapter 3, which models independent classifier outputs with Dirichlet distributions. It remains an interesting open question for future investigations how the use of this fusion model could further improve the performance of fusion in human-robot interaction applications. However, note that modeling the classifiers with the Independent Fusion Model requires training data to learn the model. Such training data are not always available, particularly in the context of human-robot interaction.

The classifiers for different modalities in multimodal human-robot interaction in Chapters 5-7 are assumed to be conditionally independent given the truth value, e.g., the true intention to be recognized. While this independence assumption for different modalities was claimed to be plausible in the context of human perception (Ernst & Bülthoff, 2004; Oruç et al., 2003), it is not entirely certain if this also holds for the classifiers for different modalities used in our human-robot interaction studies. Falsely assumed independence of actually correlated classifiers could lead to an overconfident fused forecast (Wilson, 2017), which could have dangerous consequences in human-robot interaction, in particular in the context of robotic elderly assistance. Therefore, in future work, the independence assumption of the individual modalities' classifiers in Chapters 5-7 should be checked carefully. Another promising line for future work is to apply a fusion model that explicitly considers potential correlations between the individual modalities' classifiers, e.g., the Correlated Fusion Model introduced in Chapter 3.

However, a limitation of the Correlated Fusion Model impedes its application to human-robot interaction: We did not find an analytical solution for the probability density function of the correlated Dirichlet distribution and thus could not derive the fused forecast in closed form. Therefore, for fusion we have to rely on slow Gibbs sampling, which does not allow applying the fusion model in applications where fusion has to be performed in real-time, as for example in multimodal human-robot interaction. Future work should investigate alternative inference methods that allow faster inference of the fused forecast. This can equivalently be done for the fusion model in Chapter 9, which fuses correlated probabilistic forecasts provided by humans, since this model also relies on slow Gibbs sampling for fusion. However, fusing human forecasts, e.g., provided by domain experts, is usually less time-critical than classifier fusion.

In this thesis, we fused probabilistic forecasts provided by AI systems in the form of classifiers (Chapters 3, 5, 6, 7) as well as probabilistic forecasts provided by humans (Chapters 8, 9). However, the models introduced in this thesis can also potentially be applied to combine probabilistic forecasts provided by both humans and AI systems. Thus, an interesting line of future work would be to combine forecasts provided by humans and AI systems, as done, e.g., in the work of Steyvers et al. (2022). While our correlated fusion models in Chapters 3 and 9 can only model positive correlations, in this case, we should potentially also consider negative correlations since humans and AI systems might be more likely to be negatively correlated. In fact, it is an open question to be investigated in the future how negative correlations of forecasts should impact the resulting fused forecast.

Finally, another important question that should be addressed in future work is what to do with the posterior fused forecast, regardless of whether it is obtained from fusing classifiers' forecasts or human forecasts, or even both. On the one hand, we can provide the fused forecast to a human, who can use it as a support for decision making. While in this case it is on the human to exploit the fused forecast's uncertainty, future work should investigate how humans can make best use of the provided uncertainty for making good decisions. An example of such an attempt for AI support in decision making is, e.g., the work of Benz and Rodriguez (2023). If, on the other hand, the fused forecast is used to automatically execute actions, it has to be investigated how to choose an action based on the fused forecasts while explicitly considering its uncertainty. In decision theory, every decision for an action implies some cost function that assigns costs to actions that are executed in a specific state. Given this cost function and a posterior distribution over the state, the optimal action can be computed (see Section 2.4). For example, if a robot should correctly estimate a discrete human intention from multimodal data, the appropriate cost function is 0-1 loss, and the resulting optimal intention to choose is the one with the highest posterior probability given the data, i.e., the highest probability in the fused categorical distribution. If the robot should also react to the recognized intention with a predefined action, such as a handover of a specific object, this can be realized by choosing the most likely intention from the fused forecast and reacting with the action assigned to this intention, e.g., by handing over the intended object to the human. In this case, the robot's reaction is the same for a fused forecast, in which the probabilities for all intentions are approximately the same, but one has a slightly higher probability, and for a fused forecast, in which the probability for the chosen intention is 1 and all other probabilities are 0. However, due to their different uncertainties these two prototypical fused forecasts indicate completely different situations, which should have an effect on action execution. Several options for action execution considering uncertainty are conceivable. On the one hand, other cost functions can be defined, which can, however, be difficult for robot actions and does not necessarily lead to an effect of the robot's uncertainty on its reaction. Likewise, a threshold can be set that only allows a reaction above a predefined uncertainty level, as done e.g., in the works of Cruz et al. (2016) and Cruz et al. (2018). However, such a threshold discretizes uncertainty in an arbitrary way, and it is not clear how to best set it. Another option to exploit uncertainty for action selection is sampling according to the fused forecast, as we did in Chapter 6 for action selection in interactive reinforcement learning. By sampling, the actions are selected according to the uncertainty of the fused forecast. However, this can easily lead to wrong actions because also low probability actions could be executed. In future work, it should be investigated if there is a better possibility to exploit uncertainty in action selection, for example by modifying the action execution depending on the uncertainty (Kanazawa et al., 2019). This investigation of how to exploit the uncertainty of the fused forecast is beyond the topic of this thesis, which focused on how to obtain a correct, i.e., Bayes optimal, fused forecast. However, it is a fundamental research topic for the continuation of the work presented in this thesis because a correctly reduced uncertainty should also be correctly exploited.

BIBLIOGRAPHY

Abbasi, B., Monaikul, N., Rysbek, Z., Eugenio, B. D., & Žefran, M. (2019). A multimodal human-robot interaction manager for assistive robots, In *2019 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE.

Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U. R., Makarenkov, V., & Nahavandi, S. (2021). A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, *76*, 243–297.

Admoni, H., & Srinivasa, S. (2016). Predicting user intent through eye gaze for shared autonomy, In *AAAI fall symposia*.

Alais, D., & Burr, D. (2004). The ventriloquist effect results from near-optimal bimodal integration. *Current Biology*, *14*(3), 257–262.

Almazroi, A. A., Mohamed, O. A., Shamim, A., & Qureshi, M. A. (2020). Evaluation of state-of-the-art classifiers: A comparative study. *Researchpedia Journal of Computing*, *1*(1), 22–29.

Andriamahefa, T. R. (2017). *Integer occupancy grids: A probabilistic multi-sensor fusion framework for embedded perception* (Doctoral dissertation). Université Grenoble Alpes (ComUE).

Arnold, B. C., & Ghosh, I. (2017). Bivariate beta and Kumaraswamy models developed using the Arnold-Ng bivariate beta distribution. *REVSTAT–Statistical Journal*, *15*(2), 223–250.

Arnold, B. C., & Ng, H. K. T. (2011). Flexible bivariate beta distributions. *Journal of Multivariate Analysis*, *102*(8), 1194–1202.

Babic, B., Gaba, A., Tsetlin, I., & Winkler, R. L. (2022). Resolute and correlated Bayesians. *INSEAD Working Paper*, (2022/20/DSC).

Bach, D. R., & Dolan, R. J. (2012). Knowing how much you don't know: A neural organization of uncertainty estimates. *Nature Reviews Neuroscience*, *13*(8), 572–586.

Bach, P., Nicholson, T., & Hudson, M. (2014). The affordance-matching hypothesis: How objects guide action understanding and prediction. *Frontiers in Human Neuroscience*, *8*(254).

Baek, W.-J., & Kröger, T. (2023). Safety evaluation of robot systems via uncertainty quantification. *arXiv preprint arXiv:2302.10644*.

Baek, W.-J., Ledermann, C., & Kröger, T. (2023). Uncertainty estimation for safe human-robot collaboration using conservation measures, In *International conference on intelligent autonomous systems*. Springer.

Baertlein, B. A., Liao, W.-J., & Chen, D.-H. (2001). Predicting sensor fusion performance using theoretical models, In *Detection and remediation technologies for mines and minelike targets VI*. International Society for Optics and Photonics.

Bannat, A., Gast, J., Rehrl, T., Rösel, W., Rigoll, G., & Wallhoff, F. (2009). A multimodal human-robot-interaction scenario: Working together with an industrial robot, In *Human-computer interaction. Novel interaction methods and techniques*, Springer.

Bansal, G., Nushi, B., Kamar, E., Horvitz, E., & Weld, D. S. (2021). Is the most accurate AI the best teammate? Optimizing AI for teamwork, In *Proceedings of the AAAI conference on artificial intelligence*.

Baron, J., Mellers, B. A., Tetlock, P. E., Stone, E., & Ungar, L. H. (2014). Two reasons to make aggregated probability forecasts more extreme. *Decision Analysis*, *11*(2), 133–145.

Barthelmess, P., Kaiser, E., Lunsford, R., McGee, D., Cohen, P., & Oviatt, S. (2006). Human-centered collaborative interaction, In *Proceedings of the 1st ACM international workshop on human-centered multimedia*. Association for Computing Machinery.

Bennett, S. T., Benjamin, A. S., Mistry, P. K., & Steyvers, M. (2018). Making a wiser crowd: Benefits of individual metacognitive control on crowd performance. *Computational Brain & Behavior*, *1*, 90–99.

Benz, N. L. C., & Rodriguez, M. G. (2023). Human-aligned calibration for AI-assisted decision making. *arXiv preprint arXiv:2306.00074*.

Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis*. Springer.

Bertsekas, D., & Tsitsiklis, J. N. (2008). *Introduction to probability* (Vol. 1). Athena Scientific.

Bhatt, U., Antorán, J., Zhang, Y., Liao, Q. V., Sattigeri, P., Fogliato, R., Melançon, G., Krishnan, R., Stanley, J., Tickoo, O., Et al. (2021). Uncertainty as a form of transparency: Measuring, communicating, and using uncertainty, In *Proceedings of the 2021 AAAI/ACM conference on AI, ethics, and society*. Association for Computing Machinery.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Blakemore, S.-J., & Decety, J. (2001). From the perception of action to the understanding of intention. *Nature Reviews Neuroscience*, *2*(8), 561–567.

Blumberg, B., Downie, M., Ivanov, Y., Berlin, M., Johnson, M. P., & Tomlinson, B. (2002). Integrated learning for interactive synthetic characters, In *Proceedings of the 29th annual conference on computer graphics and interactive techniques*. Association for Computing Machinery.

Bohus, D., & Horvitz, E. (2009). Learning to predict engagement with a spoken dialog system in open-world settings, In *Sigdial 2009 conference*. Association for Computational Linguistics.

Bordley, R. F. (1982). A multiplicative formula for aggregating probability assessments. *Management Science*, *28*(10), 1137–1148.

Bran-Cardona, P. A., Orozco-Castañeda, J., & Nagar, D. K. (2011). Bivariate generalization of the Kummer-beta distribution. *Revista Colombiana de Estadística*, *34*(3), 497–512.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*, 123–140.

Brenner, L. A., Koehler, D. J., Liberman, V., & Tversky, A. (1996). Overconfidence in probability and frequency judgments: A critical examination. *Organizational Behavior and Human Decision Processes*, *65*(3), 212–219.

Brier, G. W. Et al. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, *78*(1), 1–3.

Britten, G. L., Mohajerani, Y., Primeau, L., Aydin, M., Garcia, C., Wang, W.-L., Pasquier, B., Cael, B. B., & Primeau, F. W. (2021). Evaluating the benefits of Bayesian hierarchical methods for analyzing heterogeneous environmental datasets: A case study of marine organic carbon fluxes. *Frontiers in Environmental Science*, *9*.

Bromiley, P. (2003). Products and convolutions of gaussian probability density functions. *Tina-Vision Memo*, *3*(4), 1.

Bruine de Bruin, W., Fischbeck, P. S., Stiber, N. A., & Fischhoff, B. (2002). What number is "fifty-fifty"?: Redistributing excessive 50% responses in elicited probabilities. *Risk Analysis: An International Journal*, *22*(4), 713–723.

Budescu, D. V., & Chen, E. (2015). Identifying expertise to extract the wisdom of crowds. *Management Science*, *61*(2), 267–280.

Budescu, D. V., & Rantilla, A. K. (2000). Confidence in aggregation of expert opinions. *Acta Psychologica*, *104*(3), 371–398.

Buja, A., Stuetzle, W., & Shen, Y. (2005). Loss functions for binary class probability estimation and classification: Structure and applications. *Working Draft*. `https://sites.stat.washington.edu/wxs/Learning-papers/paper-proper-scoring.pdf`

Bunt, H., Beun, R.-J., & Borghuis, T. (1998). *Multimodal human-computer communication: Systems, techniques, and experiments* (Vol. 1374). Springer Science & Business Media.

Burger, B., Ferrané, I., Lerasle, F., & Infantes, G. (2012). Two-handed gesture recognition and fusion with speech to command a robot. *Autonomous Robots*, *32*, 129–147.

Canal, G., Angulo, C., & Escalera, S. (2015). Gesture based human multi-robot interaction, In *International joint conference on neural networks*. IEEE.

Canbek, G., Temizel, T. T., & Sagiroglu, S. (2022). PToPI: A comprehensive review, analysis, and knowledge representation of binary classification performance measures/metrics. *SN Computer Science*, *4*(1), 13.

Cao, Z., Hidalgo Martinez, G., Simon, T., Wei, S., & Sheikh, Y. A. (2021). Openpose: Realtime multi-person 2D pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *43*(1), 172–186.

Cesta, A., Cortellessa, G., Giuliani, V., Pecora, F., Rasconi, R., Scopelliti, M., & Tiberio, L. (2007). Proactive assistive technology: An empirical study, In *Human-computer interaction – INTERACT 2007*. Springer.

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., & Ghazanfar, A. A. (2009). The natural statistics of audiovisual speech. *PLoS Computational Biology*, *5*(7), e1000436.

Christoforou, E. G., Panayides, A. S., Avgousti, S., Masouras, P., & Pattichis, C. S. (2020). An overview of assistive robotics and technologies for elderly care, In *XV mediterranean conference on medical and biological engineering and computing (MEDICON)*. Springer.

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, *5*(4), 559–583.

Clemen, R. T., & Winkler, R. L. (1987). Calibrating and combining precipitation probability forecasts. In R. Viertl (Ed.), *Probability and Bayesian statistics* (pp. 97–110). Springer.

Clemen, R. T., & Winkler, R. L. (1999). Combining probability distributions from experts in risk analysis. *Risk Analysis*, *19*(2), 187–203.

Cochran, W. G. (1937). Problems arising in the analysis of a series of similar experiments. *Supplement to the Journal of the Royal Statistical Society*, *4*(1), 102–118.

Connor, R. J., & Mosimann, J. E. (1969). Concepts of independence for proportions with a generalization of the Dirichlet distribution. *Journal of the American Statistical Association*, *64*(325), 194–206.

Cooke, R. (1991). *Experts in uncertainty: Opinion and subjective probability in science*. Oxford University Press.

Crackel, R., & Flegal, J. (2017). Bayesian inference for a flexible class of bivariate beta distributions. *Journal of Statistical Computation and Simulation*, *87*(2), 295–312.

Cruz, F., Parisi, G. I., Twiefel, J., & Wermter, S. (2016). Multi-modal integration of dynamic audiovisual patterns for an interactive reinforcement learning scenario, In *2016 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE.

Cruz, F., Parisi, G. I., & Wermter, S. (2018). Multi-modal feedback for affordance-driven interactive reinforcement learning, In *2018 international joint conference on neural networks (IJCNN)*. IEEE.

Cruz, F., Twiefel, J., Magg, S., Weber, C., & Wermter, S. (2015). Interactive reinforcement learning through speech guidance in a domestic scenario, In *2015 international joint conference on neural networks (IJCNN)*. IEEE.

David Sam Jayakumar, G., Sulthan, A., & Samuel, W. (2019). A new bivariate beta distribution of Kind-1 of Type-A. *Journal of Statistics and Management Systems*, *22*(1), 141–158.

De Ruiter, J. P., Bangerter, A., & Dings, P. (2012). The interplay between gesture and speech in the production of referring expressions: Investigating the tradeoff hypothesis. *Topics in Cognitive Science*, *4*(2), 232–248.

Der Kiureghian, A., & Ditlevsen, O. (2009). Aleatory or epistemic? Does it matter? *Structural Safety*, *31*(2), 105–112.

Di Bacco, M., Frederic, P., Lad, F., Et al. (2003). Learning from the probability assertions of experts. *University of Bologna Research Report, Dipartimento di Statistiche, Bologna*.

Dietrich, F., & List, C. (2016). Probabilistic opinion pooling. In *The Oxford handbook of probability and philosophy* (pp. 519–542). Oxford University Press.

Dietterich, T. G. (2000). Ensemble methods in machine learning, In *International workshop on multiple classifier systems*. Springer.

Dino, M. J. S., Davidson, P. M., Dion, K. W., Szanton, S. L., & Ong, I. L. (2022). Nursing and human-computer interaction in healthcare robots for older people: An integrative review. *International Journal of Nursing Studies Advances*, *4*, 100072.

Dong, X., & Hayes, C. C. (2012). Uncertainty visualizations: Helping decision makers become more aware of uncertainty and its implications. *Journal of Cognitive Engineering and Decision Making*, *6*(1), 30–56.

Drakopoulos, E., & Lee, C. C. (1988). Optimum fusion of correlated local decisions, In *Proceedings of the 27th IEEE conference on decision and control*. IEEE.

Dua, D., & Graff, C. (2017). UCI machine learning repository. `http://archive.ics.uci.edu/ml`

Dutta, V., & Zielinska, T. (2018). Predicting human actions taking into account object affordances. *Journal of Intelligent & Robotic Systems*, 1–17.

El-Bassiouny, A., & Jones, M. (2009). A bivariate F distribution with marginals on arbitrary numerator and denominator degrees of freedom, and related bivariate beta and t distributions. *Statistical Methods and Applications*, *18*(4), 465.

Elmannai, H., Saleh, H., Algarni, A. D., Mashal, I., Kwak, K. S., El-Sappagh, S., & Mostafa, S. (2022). Diagnosis myocardial infarction based on stacking ensemble of convolutional neural network. *Electronics*, *11*(23), 3976.

Elsaesser, D. (2007). Sensor data fusion using a probability density grid, In *International conference on information fusion*. IEEE.

Englert, P., Vien, N. A., & Toussaint, M. (2017). Inverse KKT: Learning cost functions of manipulation tasks from demonstrations. *The International Journal of Robotics Research*, *36*(13–14), 1474–1488.

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, *415*(6870), 429–433.

Ernst, M. O., & Bülthoff, H. H. (2004). Merging the senses into a robust percept. *Trends in Cognitive Sciences*, *8*(4), 162–169.

Ewerton, M., Neumann, G., Lioutikov, R., Amor, H. B., Peters, J., & Maeda, G. (2015). Learning multiple collaborative tasks with a mixture of interaction primitives, In *IEEE international conference on robotics and automation (ICRA)*. IEEE.

Faisal, A. A., Selen, L. P., & Wolpert, D. M. (2008). Noise in the nervous system. *Nature Reviews Neuroscience*, *9*(4), 292–303.

Faria, F. A., dos Santos, J. A., Sarkar, S., Rocha, A., & Torres, R. d. S. (2013). Classifier selection based on the correlation of diversity measures: When fewer is more, In *2013 XXVI conference on graphics, patterns and images*. IEEE.

Ferri, C., Hernández-Orallo, J., & Modroiu, R. (2009). An experimental comparison of performance measures for classification. *Pattern Recognition Letters*, *30*(1), 27–38.

Fischhoff, B., & Bruine De Bruin, W. (1999). Fifty-fifty= 50%? *Journal of Behavioral Decision Making*, *12*(2), 149–163.

Fiser, J., Berkes, P., Orbán, G., & Lengyel, M. (2010). Statistically optimal perception and learning: From behavior to neural representations. *Trends in Cognitive Sciences*, *14*(3), 119–130.

Foster, M. E., Gaschler, A., & Giuliani, M. (2017). Automatically classifying user engagement for dynamic multi-party human-robot interaction. *International Journal of Social Robotics*, *9*(5), 659–674.

French, S. (1980). Updating of belief in the light of someone else's opinion. *Journal of the Royal Statistical Society: Series A (General)*, *143*(1), 43–48.

Freund, Y., & Schapire, R. E. (1996). Experiments with a new boosting algorithm, In *International conference on machine learning (ICML)*. Morgan Kaufmann Publishers Inc.

Frydrychowicz, S., & Matejczuk, J. (2006). The role of intention in the process of interpersonal communication. *Psychology of Language and Communication*, *10*(2), 89–107.

Gaschler, A., Jentzsch, S., Giuliani, M., Huth, K., de Ruiter, J., & Knoll, A. (2012). Social behavior recognition using body posture and head pose for human-robot interaction, In *2012 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., & Rubin, D. (2013). *Bayesian data analysis (3rd ed.)* Chapman; Hall/CRC.

Genest, C., Zidek, J. V. Et al. (1986). Combining probability distributions: A critique and an annotated bibliography. *Statistical Science*, *1*(1), 114–135.

Gepshtein, S., & Banks, M. S. (2003). Viewing geometry determines how vision and haptics combine in size perception. *Current Biology*, *13*(6), 483–488.

Gepshtein, S., Burge, J., Ernst, M. O., & Banks, M. S. (2005). The combination of vision and touch depends on spatial proximity. *Journal of Vision*, *5*(11), 7–7.

Girshick, A. R., & Banks, M. S. (2009). Probabilistic combination of slant information: Weighted averaging and robustness as optimal percepts. *Journal of Vision*, *9*(9), 8–8.

Goebel, K., & Yan, W. (2004). Choosing classifiers for decision fusion, In *Proceedings of the 7th international conference on information fusion.*

Goodrich, M. A., & Schultz, A. C. (2008). Human-robot interaction: A survey. *Foundations and Trends® in Human-Computer Interaction*, *1*(3), 203–275.

Gordon, G., Spaulding, S., Westlund, J. K., Lee, J. J., Plummer, L., Martinez, M., Das, M., & Breazeal, C. (2016). Affective personalization of a social robot tutor for children's second language skills, In *Proceedings of the AAAI conference on artificial intelligence.*

Graefe, A. (2018). Predicting elections: Experts, polls, and fundamentals. *Judgment and Decision Making*, *13*(4), 334.

Graf, B., Hans, M., & Schraft, R. D. (2004). Care-O-bot II–Development of a next generation robotic home assistant. *Autonomous Robots*, *16*(2), 193–205.

Graham, J. R. (1996). Is a group of economists better than one? Than none? *Journal of Business*, *69*(2), 193–232.

Griffiths, T. L., & Tenenbaum, J. B. (2006). Optimal predictions in everyday cognition. *Psychological Science*, *17*(9), 767–773.

Grinin, L., Grinin, A., & Korotayev, A. (2023). Global aging: An integral problem of the future. How to turn a problem into a development driver? In V. Sadovnichy, A. Akaev, I. Ilyin, S. Malkov, L. Grinin, & A. Korotayev (Eds.), *Reconsidering the limits to growth: A report to the Russian association of the club of Rome* (pp. 117–135). Springer.

Grushka-Cockayne, Y., Jose, V. R. R., & Lichtendahl Jr, K. C. (2017). Ensembles of overfit and overconfident forecasts. *Management Science*, *63*(4), 1110–1130.

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks, In *International conference on machine learning (ICML).* PMLR.

Gupta, A. K., Orozco-Castañeda, J. M., & Nagar, D. K. (2011). Non-central bivariate beta distribution. *Statistical Papers*, *52*(1), 139–152.

Gupta, A. K., & Wong, C. (1985). On three and five parameter bivariate beta distributions. *Metrika*, *32*(1), 85–91.

Gurban, M., & Thiran, J. P. (2008). Using entropy as a stream reliability estimate for audio-visual speech recognition, In *European signal processing conference.* IEEE.

Hamed, M. G., & Akbari, A. (2018). Hierarchical Bayesian classifier combination, In *International conference on machine learning and data mining in pattern recognition.* Springer.

Hanea, A., Wilkinson, D. P., McBride, M., Lyon, A., van Ravenzwaaij, D., Singleton Thorn, F., Gray, C., Mandel, D. R., Willcox, A., Gould, E., Et al. (2021). Mathematically aggregating experts' predictions of possible futures. *PLOS ONE*, *16*(9), 1–24.

Harris, C. M., & Wolpert, D. M. (1998). Signal-dependent noise determines motor planning. *Nature*, *394*(6695), 780–784.

Hayman, E., & Eklundh, J.-O. (2002). Probabilistic and voting approaches to cue integration for figure-ground segmentation, In *Computer vision — ECCV 2002*, Springer.

Helbig, H. B., & Ernst, M. O. (2007). Optimal integration of shape information from vision and touch. *Experimental Brain Research*, *179*, 595–606.

Helmholtz, H. v. (1924). Treatise on physiological optics, 3 vols.

Hillis, J. M., Ernst, M. O., Banks, M. S., & Landy, M. S. (2002). Combining sensory information: Mandatory fusion within, but not between, senses. *Science*, *298*(5598), 1627–1630.

Hillis, J. M., Watt, S. J., Landy, M. S., & Banks, M. S. (2004). Slant from texture and disparity cues: Optimal cue combination. *Journal of Vision*, *4*(12), 1–1.

Hofmann, A. G., & Williams, B. C. (2007). Intent recognition for human-robot interaction., In *2007 AAAI spring symposium: Interaction challenges for intelligent assistants*.

Hogarth, R. M. (1978). A note on aggregating opinions. *Organizational behavior and human performance*, *21*(1), 40–46.

Hora, S. C. (1996). Aleatory and epistemic uncertainty in probability elicitation with an example from hazardous waste management. *Reliability Engineering & System Safety*, *54*(2–3), 217–223.

Huang, C.-M., Andrist, S., Sauppé, A., & Mutlu, B. (2015). Using gaze patterns to predict task intent in collaboration. *Frontiers in Psychology*, *6*.

Huang, J. (2005). Maximum likelihood estimation of Dirichlet distribution parameters. *CMU Technique Report*.

Huang, Z., Lam, H., & Zhang, H. (2021). Quantifying epistemic uncertainty in deep learning. *arXiv preprint arXiv:2110.12122*.

Hüllermeier, E., & Waegeman, W. (2021). Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, *110*, 457–506.

Ishii, R., Nakano, Y., & Nishida, T. (2013). Gaze awareness in conversational agents: Estimating a user's conversational engagement from eye gaze. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *3*(2), 1–25.

Ito, K., Kong, Q., Horiguchi, S., Sumiyoshi, T., & Nagamatsu, K. (2020). Anticipating the start of user interaction for service robot in the wild, In *2020 IEEE international conference on robotics and automation (ICRA)*. IEEE.

Jacobs, R. A. (1995). Methods for combining experts' probability assessments. *Neural Computation*, *7*(5), 867–888.

Jaimes, A., & Sebe, N. (2007). Multimodal human-computer interaction: A survey. *Computer Vision and Image Understanding*, *108*(1–2), 116–134.

Jeon, H. J., Milli, S., & Dragan, A. (2020). Reward-rational (implicit) choice: A unifying formalism for reward learning, In *Advances in neural information processing systems (NeurIPS)*. Curran Associates Inc.

Ji, D., Smyth, P., & Steyvers, M. (2020). Can I trust my fairness metric? Assessing fairness with unlabeled data and Bayesian inference, In *Advances in neural information processing systems (NeurIPS)*, Curran Associates, Inc.

Johansson, F. Et al. (2013). *Mpmath: A Python library for arbitrary-precision floating-point arithmetic (version 0.18)*. `http://mpmath.org/`

Johnson, N. L., Kotz, S., & Balakrishnan, N. (1995). *Continuous univariate distributions, volume 2* (Vol. 289). John Wiley & Sons.

Johnston, C. (2022). Ethical design and use of robotic care of the elderly. *Journal of Bioethical Inquiry*, *19*(1), 11–14.

Johnston, E. B., Cumming, B. G., & Parker, A. J. (1993). Integration of depth modules: Stereopsis and texture. *Vision Research*, *33*(5–6), 813–826.

Jones, M. (2002). Multivariate t and beta distributions associated with the multivariate F distribution. *Metrika*, *54*(3), 215–231.

Joslyn, S., & LeClerc, J. (2013). Decisions with uncertainty: The glass half full. *Current Directions in Psychological Science*, *22*(4), 308–315.

Jouini, M. N., & Clemen, R. T. (1996). Copula models for aggregating expert opinions. *Operations Research*, *44*(3), 444–457.

Kam, M., Zhu, Q., & Gray, W. S. (1991). On distributed detection with correlated local detectors, In *1991 American control conference*. IEEE.

Kanazawa, A., Kinugawa, J., & Kosuge, K. (2019). Adaptive motion planning for a collaborative robot based on prediction uncertainty to enhance human safety and work efficiency. *IEEE Transactions on Robotics*, *35*(4), 817–832.

Kaplan, F., Oudeyer, P.-Y., Kubinyi, E., & Miklósi, A. (2002). Robotic clicker training. *Robotics and Autonomous Systems*, *38*(3–4), 197–206.

Karvetski, C. W., Olson, K. C., Mandel, D. R., & Twardy, C. R. (2013). Probabilistic coherence weighting for optimizing expert forecasts. *Decision Analysis*, *10*(4), 305–326.

Kelley, R., Browne, K., Wigand, L., Nicolescu, M., Hamilton, B., & Nicolescu, M. (2012). Deep networks for predicting human intent with respect to objects, In *ACM/IEEE international conference on human-robot interaction*. Association for Computing Machinery.

Kelley, R., Tavakkoli, A., King, C., Ambardekar, A., Nicolescu, M., & Nicolescu, M. (2012). Context-based Bayesian intent recognition. *IEEE Transactions on Autonomous Mental Development*, *4*(3), 215–225.

Kersten, D., & Yuille, A. (2003). Bayesian models of object perception. *Current Opinion in Neurobiology*, *13*(2), 150–158.

Kim, H.-C., & Ghahramani, Z. (2012). Bayesian classifier combination, In *Proceedings of the international conference on artificial intelligence and statistics (AISTATS)*. PMLR.

Kittler, J., Hatef, M., Duin, R. P., & Matas, J. (1998). On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(3), 226–239.

Knill, D. C., & Pouget, A. (2004). The Bayesian brain: The role of uncertainty in neural coding and computation. *TRENDS in Neurosciences*, *27*(12), 712–719.

Knill, D. C., & Saunders, J. A. (2003). Do humans optimally integrate stereo and texture information for judgments of surface slant? *Vision Research*, *43*(24), 2539–2558.

Knox, W. B., & Stone, P. (2008). TAMER: Training an agent manually via evaluative reinforcement, In *2008 7th IEEE international conference on development and learning*. IEEE.

Knox, W. B., Stone, P., & Breazeal, C. (2013). Training a robot via human feedback: A case study, In *International conference on social robotics*. Springer.

Koblinger, Á., Fiser, J., & Lengyel, M. (2021). Representations of uncertainty: Where art thou? *Current Opinion in Behavioral Sciences*, *38*, 150–162.

Kochenderfer, M. J. (2015). *Decision making under uncertainty: Theory and application*. MIT Press.

Kochskaemper, S. (2018). IW-Report 33/18: Die Entwicklung der Pflegefallzahlen in den Bundeslaendern. Eine Simulation bis 2035. *IW – Wirtschaftliche Untersuchungen, Berichte und Sachverhalte.*

Koert, D., Kircher, M., Salikutluk, V., D'Eramo, C., & Peters, J. (2020). Multi-channel interactive reinforcement learning for sequential tasks. *Frontiers in Robotics and AI*, *7*, 97.

Kompa, B., Snoek, J., & Beam, A. L. (2021). Second opinion needed: Communicating uncertainty in medical machine learning. *NPJ Digital Medicine*, *4*(1), 4.

Kormushev, P., Calinon, S., & Caldwell, D. G. (2013). Reinforcement learning in robotics: Applications and real-world challenges. *Robotics*, *2*(3), 122–148.

Koutoumanou, E., Wade, A., & Cortina-Borja, M. (2017). Local dependence in bivariate copulae with beta marginals. *Revista Colombiana de Estadística*, *40*(2), 281–296.

Krahmer, E., & Swerts, M. (2005). How children and adults produce and perceive uncertainty in audiovisual speech. *Language and Speech*, *48*(1), 29–53.

Kuhlmann, G., Stone, P., Mooney, R., & Shavlik, J. (2004). Guiding a reinforcement learner with natural language advice: Initial results in robocup soccer, In *The AAAI-2004 workshop on supervisory control of learning and adaptive systems.*

Kulic, D., & Croft, E. A. (2003). Estimating intent for human-robot interaction, In *IEEE international conference on advanced robotics.*

Kull, M., Silva Filho, T., & Flach, P. (2017). Beta calibration: A well-founded and easily implemented improvement on logistic calibration for binary classifiers, In *Proceedings of the international conference on artificial intelligence and statistics (AISTATS).* PMLR.

Kuncheva, L. I. (2014). *Combining pattern classifiers: Methods and algorithms, second edition.* John Wiley & Sons.

Kuncheva, L. I., Bezdek, J. C., & Duin, R. P. (2001). Decision templates for multiple classifier fusion: An experimental comparison. *Pattern Recognition*, *34*(2), 299–314.

Kuncheva, L. I., & Jain, L. C. (2000). Designing classifier fusion systems by genetic algorithms. *IEEE Transactions on Evolutionary Computation*, *4*(4), 327–336.

Lacoste, A., Marchand, M., Laviolette, F., & Larochelle, H. (2014). Agnostic Bayesian learning of ensembles, In *International conference on machine learning (ICML).* PMLR.

Landy, M. S., Banks, M. S., & Knill, D. C. (2011). Ideal-observer models of cue integration. *Sensory Cue Integration*, 5–29.

Landy, M. S., & Kojima, H. (2001). Ideal cue combination for localizing texture-defined edges. *Journal of the Optical Society of America A*, *18*(9), 2307–2320.

Lee, D.-S., & Srihari, S. N. (1995). A theory of classifier combination: The neural network approach, In *Proceedings of 3rd international conference on document analysis and recognition.* IEEE.

Lee, M. D. (2018a). Bayesian methods in cognitive modeling. In *Stevens' handbook of experimental psychology and cognitive neuroscience* (pp. 1–48). John Wiley & Sons, Ltd.

Lee, M. D. (2018b). In vivo: Multiple approaches to hierarchical modeling. In S. Farrell & S. Lewandowsky (Eds.), *Computational modeling of cognition and behavior.* Cambridge University Press. https://webfiles.uci.edu/mdlee/LeeInVivo.pdf?uniq=fe8jrx

Lee, M. D., & Danileiko, I. (2014). Using cognitive models to combine probability estimates. *Judgment and Decision Making*, *9*(3), 258–272.

Lee, M. D., & Wagenmakers, E.-J. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge University Press.

Leite, I., Pereira, A., Castellano, G., Mascarenhas, S., Martinho, C., & Paiva, A. (2011). Modelling empathy in social robotic companions, In *International conference on user modeling, adaptation, and personalization*. Springer.

Li, L., Xu, Q., & Tan, Y. K. (2012). Attention-based addressee selection for service and social robots to interact with multiple persons, In *Proceedings of the workshop at SIGGRAPH Asia*. Association for Computing Machinery.

Libby, D. L., & Novick, M. R. (1982). Multivariate generalized beta distributions with applications to utility assessment. *Journal of Educational Statistics*, *7*(4), 271–294.

Lichtendahl Jr, K. C., Grushka-Cockayne, Y., Jose, V. R., & Winkler, R. L. (2022). Extremizing and antiextremizing in Bayesian ensembles of binary-event forecasts. *Operations Research*, *70*(5), 2998–3014.

Lin, J., Ma, Z., Gomez, R., Nakamura, K., He, B., & Li, G. (2020). A review on interactive reinforcement learning from human social feedback. *IEEE Access*, *8*, 120757–120765.

Linderman, S., Johnson, M. J., & Adams, R. P. (2015). Dependent multinomial models made easy: Stick-breaking with the polya-gamma augmentation, In *Advances in neural information processing systems (NeurIPS)*. Curran Associates, Inc.

Lindley, D. V. (1985). Reconciliation of discrete probability distributions. *Bayesian Statistics*, *2*, 375–390.

Lindley, D. V. (1987). The probability approach to the treatment of uncertainty in artificial intelligence and expert systems. *Statistical Science*, 17–24.

Lindley, D. V. (2013). *Understanding uncertainty*. John Wiley & Sons.

Liu, H., Fan, T., & Wu, P. (2014). Audio-visual keyword spotting based on adaptive decision fusion under noisy conditions for human-robot interaction, In *IEEE international conference on robotics and automation (ICRA)*. IEEE.

Liu, X., Ge, S. S., Jiang, R., & Goh, C. H. (2016). Intelligent speech control system for human-robot interaction, In *Chinese control conference (CCC)*. IEEE.

Luo, X., & Xu, J. (2016). Object-based representation for scene classification, In *Canadian conference on artificial intelligence*. Springer.

Ma, A. J., Yuen, P. C., & Lai, J.-H. (2013). Linear dependency modeling for classifier fusion and feature combination. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(5), 1135–1148.

Ma, W. J. (2019). Bayesian decision models: A primer. *Neuron*, *104*(1), 164–175.

Ma, W. J., Beck, J. M., Latham, P. E., & Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nature Neuroscience*, *9*(11), 1432–1438.

Ma, W. J., & Jazayeri, M. (2014). Neural coding of uncertainty and probability. *Annual Review of Neuroscience*, *37*, 205–220.

Ma, W. J., Zhou, X., Ross, L. A., Foxe, J. J., & Parra, L. C. (2009). Lip-reading aids word recognition most in moderate noise: A Bayesian explanation using high-dimensional feature space. *PloS one*, *4*(3), e4638.

Maclin, R., & Shavlik, J. W. (1996). Creating advice-taking reinforcement learners. *Machine Learning*, *22*(1), 251–281.

Magnussen, S. (2004). An algorithm for generating positively correlated beta-distributed random variables with known marginal distributions and a specified correlation. *Computational Statistics & Data Analysis*, *46*(2), 397–406.

Martinez-Martin, E., & Costa, A. (2021). Assistive technology for elderly care: An overview. *IEEE Access*, *9*, 92420–92430.

Masri, N., Sultan, Y. A., Akkila, A. N., Almasri, A., Ahmed, A., Mahmoud, A. Y., Zaqout, I., & Abu-Naser, S. S. (2019). Survey of rule-based systems. *International Journal of Academic Information Systems Research (IJAISR)*, *3*(7), 1–23.

McAndrew, T., Wattanachit, N., Gibson, G. C., & Reich, N. G. (2021). Aggregating predictions from experts: A review of statistical methods, experiments, and applications. *Wiley Interdisciplinary Reviews: Computational Statistics*, *13*(2), e1514.

Mennie, N., Hayhoe, M., & Sullivan, B. (2007). Look-ahead fixations: Anticipatory eye movements in natural tasks. *Experimental Brain Research*, *179*(3), 427–442.

Mi, A., Wang, L., & Qi, J. (2016). A multiple classifier fusion algorithm using weighted decision templates. *Scientific Programming*, *2016*.

Minson, J. A., Mueller, J. S., & Larrick, R. P. (2018). The contingent wisdom of dyads: When discussion enhances vs. undermines the accuracy of collaborative judgments. *Management Science*, *64*(9), 4177–4192.

Mohammed, M., Mwambi, H., Mboya, I. B., Elbashir, M. K., & Omolo, B. (2021). A stacking ensemble deep learning approach to cancer type classification based on TCGA data. *Scientific Reports*, *11*(1), 15626.

Mohandes, M., Deriche, M., & Aliyu, S. O. (2018). Classifiers combination techniques: A comprehensive review. *IEEE Access*, *6*, 19626–19639.

Mollaret, C., Mekonnen, A. A., Ferrané, I., Pinquier, J., & Lerasle, F. (2015). Perceiving user's intention-for-interaction: A probabilistic multimodal data fusion scheme, In *2015 IEEE international conference on multimedia and expo (ICME)*. IEEE.

Mollaret, C., Mekonnen, A. A., Lerasle, F., Ferrané, I., Pinquier, J., Boudet, B., & Rumeau, P. (2016). A multi-modal perception based assistive robotic system for the elderly. *Computer Vision and Image Understanding*, *149*, 78–97.

Morgan, M. G. (2014). Use (and abuse) of expert elicitation in support of decision making for public policy. *Proceedings of the National Academy of Sciences*, *111*(20), 7176–7184.

Morris, P. A. (1986). [Combining Probability Distributions: A Critique and an Annotated Bibliography]: Comment. *Statistical Science*, *1*(1), 141–144.

Moschen, L. M., & Carvalho, L. M. (2023). Bivariate beta distribution: Parameter inference and diagnostics. *arXiv preprint arXiv:2303.01271*.

Murphy, A. H. (1973). Hedging and skill scores for probability forecasts. *Journal of Applied Meteorology and Climatology*, *12*(1), 215–223.

Murphy, K. P. (2012). *Machine learning: A probabilistic perspective*. MIT Press.

Murphy, K. P. (2022). *Probabilistic machine learning: An introduction*. MIT Press.

Nadarajah, S., & Kotz, S. (2005). Some bivariate beta distributions. *Statistics*, *39*(5), 457–466.

Nadarajah, S., Shih, S. H., & Nagar, D. K. (2017). A new bivariate beta distribution. *Statistics*, *51*(2), 455–474.

Nadav-Greenberg, L., & Joslyn, S. L. (2009). Uncertainty forecasts improve decision making among nonexperts. *Journal of Cognitive Engineering and Decision Making*, *3*(3), 209–227.

Nakano, Y., & Ishii, R. (2010). Estimating user's engagement from eye-gaze behaviors in human-agent conversations, In *15th international conference on intelligent user interfaces*. Association for Computing Machinery.

Nazabal, A., Garcia-Moreno, P., Artes-Rodriguez, A., & Ghahramani, Z. (2016). Human activity recognition by combining a small number of classifiers. *IEEE Journal of Biomedical and Health Informatics*, *20*(5), 1342–1351.

Niemelä, M., & Melkas, H. (2019). Robots as social and physical assistants in elderly care. *Human-Centered Digitalization and Services*, 177–197.

Nweke, H. F., Teh, Y. W., Mujtaba, G., & Al-Garadi, M. A. (2019). Data fusion and multiple classifier systems for human activity detection and health monitoring: Review and open research directions. *Information Fusion*, *46*, 147–170.

Olkin, I., & Liu, R. (2003). A bivariate beta distribution. *Statistics & Probability Letters*, *62*(4), 407–412.

Olkin, I., & Trikalinos, T. A. (2015). Constructions for a bivariate beta distribution. *Statistics & Probability Letters*, *96*, 54–60.

Ooko, R., Ishii, R., & Nakano, Y. (2011). Estimating a user's conversational engagement based on head pose information, In *Intelligent virtual agents: 10th international conference*. Springer.

Orozco-Castañeda, J. M., Nagar, D. K., & Gupta, A. K. (2012). Generalized bivariate beta distributions involving Appell's hypergeometric function of the second kind. *Computers & Mathematics with Applications*, *64*(8), 2507–2519.

Oruç, I., Maloney, L. T., & Landy, M. S. (2003). Weighted linear cue combination with possibly correlated error. *Vision Research*, *43*(23), 2451–2468.

Paraschos, A., Daniel, C., Peters, J. R., & Neumann, G. (2013). Probabilistic movement primitives, In *Advances in neural information processing systems (NeurIPS)*. Curran Associates, Inc.

Petrakos, M., Kannelopoulos, I., Benediktsson, J. A., & Pesaresi, M. (2000). The effect of correlation on the accuracy of the combined classifier in decision level fusion, In *IEEE 2000 international geoscience and remote sensing symposium (IGARSS). Taking the pulse of the planet: The role of remote sensing in managing the environment*. IEEE.

Pirs, G., & Strumbelj, E. (2019). Bayesian combination of probabilistic classifiers using multivariate normal mixtures. *Journal of Machine Learning Research*, *20*(1), 1892–1909.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling, In *Proceedings of the 3rd international workshop on distributed statistical computing*.

Potamianos, G., & Neti, C. (2000). Stream confidence estimation for audio-visual speech recognition, In *International conference on spoken language processing*.

Prabhakar, S., & Jain, A. K. (2002). Decision-level fusion in fingerprint verification. *Pattern Recognition*, *35*(4), 861–874.

Prelec, D., Seung, H. S., & McCoy, J. (2017). A solution to the single-question crowd wisdom problem. *Nature*, *541*(7638), 532–535.

Qian, L. (2016). Pupil ROS plugin: Connecting pupil eye-tracking platform and robot operation system (ROS) platform. `https://github.com/qian256/pupil_ros_plugin`

Qu, Z., Meng, Y., Muhammad, G., & Tiwari, P. (2024). QMFND: A quantum multimodal fusion-based fake news detection model for social media. *Information Fusion*, *104*, 102172.

Quek, F., McNeill, D., Bryll, R., Duncan, S., Ma, X.-F., Kirbas, C., McCullough, K. E., & Ansari, R. (2002). Multimodal human discourse: Gesture and speech. *ACM Transactions on Computer-Human Interaction (TOCHI)*, *9*(3), 171–193.

Qureshi, A. H., Nakamura, Y., Yoshikawa, Y., & Ishiguro, H. (2016). Robot gains social intelligence through multimodal deep reinforcement learning, In *2016 IEEE-RAS 16th international conference on humanoid robots (Humanoids)*. IEEE.

Ranjan, R., & Gneiting, T. (2010). Combining probability forecasts. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *72*(1), 71–91.

Rodomagoulakis, I., Kardaris, N., Pitsikalis, V., Mavroudi, E., Katsamanis, A., Tsiami, A., & Maragos, P. (2016). Multimodal human action recognition in assistive human-robot interaction, In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE.

Rossi, A., Raiano, M., & Rossi, S. (2021). Affective, cognitive and behavioural engagement detection for human-robot interaction in a bartending scenario, In *2021 30th IEEE international conference on robot & human interactive communication (RO-MAN)*. IEEE.

Rothkopf, C. A., Ballard, D. H., & Hayhoe, M. M. (2007). Task and context determine where you look. *Journal of Vision*, *7*(14), 16–20.

Rothkopf, C. A., & Dimitrakakis, C. (2011). Preference elicitation and inverse reinforcement learning, In *Machine learning and knowledge discovery in databases*. Springer.

Rothkopf, C. A., Weisswange, T., & Triesch, J. (2010). Computational modeling of multisensory object perception. In J. Kaiser & M. J. Naumer (Eds.), *Multisensory object perception in the primate brain* (pp. 21–53). Springer.

Roulston, M. S., Bolton, G. E., Kleit, A. N., & Sears-Collins, A. L. (2006). A laboratory study of the benefits of including uncertainty information in weather forecasts. *Weather and Forecasting*, *21*(1), 116–122.

Russell, S., & Norvig, P. (2010). *Artificial intelligence: A modern approach, third edition*. Pearson Education.

Safont, G., Salazar, A., & Vergara, L. (2019). Multiclass alpha integration of scores from multiple classifiers. *Neural Computation*, *31*(4), 806–825.

Sainath, T. N., & Parada, C. (2015). Convolutional neural networks for small-footprint keyword spotting, In *Proceedings of Interspeech 2015*.

Samanthi, R. G., & Sepanski, J. (2019). A bivariate extension of the beta generated distribution derived from copulas. *Communications in Statistics–Theory and Methods*, *48*(5), 1043–1059.

Santhanaraj, K. K., & MM, R. (2021). A survey of assistive robots and systems for elderly care. *Journal of Enabling Technologies*, *15*(1), 66–72.

Sarabia, J. M., & Castillo, E. (2006). Bivariate distributions based on the generalized three-parameter beta distribution. In N. Balakrishnan, J. M. Sarabia, & E. Castillo (Eds.), *Advances in distribution theory, order statistics, and inference* (pp. 85–110). Birkhäuser Boston.

Satake, S., Kanda, T., Glas, D. F., Imai, M., Ishiguro, H., & Hagita, N. (2009). How to approach humans? Strategies for social robots to initiate interaction, In *4th*

ACM/IEEE international conference on human robot interaction. Association for Computing Machinery.

Satopää, V. A. (2022). Regularized aggregation of one-off probability predictions. *Operations Research*, *70*(6), 3558–3580.

Satopää, V. A., Baron, J., Foster, D. P., Mellers, B. A., Tetlock, P. E., & Ungar, L. H. (2014). Combining multiple probability predictions using a simple logit model. *International Journal of Forecasting*, *30*(2), 344–356.

Satopää, V. A., Pemantle, R., & Ungar, L. H. (2016). Modeling probability forecasts via information diversity. *Journal of the American Statistical Association*, *111*(516), 1623–1633.

Satopää, V. A., Salikhov, M., Tetlock, P. E., & Mellers, B. (2023). Decomposing the effects of crowd-wisdom aggregators: The bias-information-noise (BIN) model. *International Journal of Forecasting*, *39*(1), 470–485.

Saunders, J. A., & Chen, Z. (2015). Perceptual biases and cue weighting in perception of 3D slant from texture and stereo information. *Journal of Vision*, *15*(2), 14–14.

Saunders, J. A., & Knill, D. C. (2001). Perception of 3D surface orientation from skew symmetry. *Vision Research*, *41*(24), 3163–3183.

Scarfe, P. (2022). Experimentally disambiguating models of sensory cue integration. *Journal of Vision*, *22*(1), 5–5.

Scarfe, P., & Hibbard, P. B. (2011). Statistically optimal integration of biased sensory estimates. *Journal of Vision*, *11*(7), 12–12.

Schrempf, O. C., Albrecht, D., & Hanebeck, U. D. (2007). Tractable probabilistic models for intention recognition based on expert knowledge, In *IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE.

Schrempf, O. C., & Hanebeck, U. D. (2005). A generic model for estimating user intentions in human-robot cooperation, In *Proceedings of the 2nd international conference on informatics in control, automation, and robotics (ICINCO) – volume 4*. INSTICC.

Schubert, C. M., Leap, N. J., Oxley, M. E., & Bauer Jr, K. W. (2004). Quantifying the correlation effects of fused classifiers, In *Signal processing, sensor fusion, and target recognition XIII*. International Society for Optics and Photonics.

Schuldhaus, D., Leutheuser, H., & Eskofier, B. M. (2013). Classification of daily life activities by decision level fusion of inertial sensor data, In *Proceedings of the 8th international conference on body area networks*. ICST.

Schultheis, M., Straub, D., & Rothkopf, C. A. (2021). Inverse optimal control adapted to the noise characteristics of the human sensorimotor system, In *Advances in neural information processing systems (NeurIPS)*, Curran Associates, Inc.

Shafer, G. (1976). *A mathematical theory of evidence* (Vol. 42). Princeton University Press.

Shapiro, A. D. (1987). *Structured induction in expert systems*. Addison-Wesley.

Sharma, K. G., & Singh, Y. (2023). Predicting intrusion in a network traffic using variance of neighboring object's distance. *International Journal of Computer Network and Information Security*, *13*(2), 73.

Shi, C., Shiomi, M., Kanda, T., Ishiguro, H., & Hagita, N. (2015). Measuring communication participation to initiate conversation in human-robot interaction. *International Journal of Social Robotics*, *7*, 889–910.

Shi, L., Kodagoda, S., & Dissanayake, G. (2010). Multi-class classification for semantic labeling of places, In *International conference on control, automation, robotics, & vision*. IEEE.

Sidner, C. L., Lee, C., Kidd, C. D., Lesh, N., & Rich, C. (2005). Explorations in engagement for humans and robots. *Artificial Intelligence*, *166*(1–2), 140–164.

Silver, I., Mellers, B. A., & Tetlock, P. E. (2021). Wise teamwork: Collective confidence calibration predicts the effectiveness of group discussion. *Journal of Experimental Social Psychology*, *96*, 104157.

Simpson, E., Roberts, S., Psorakis, I., & Smith, A. (2013). Dynamic Bayesian combination of multiple imperfect classifiers. In T. V. Guy, M. Karny, & D. Wolpert (Eds.), *Decision making and imperfection* (pp. 1–35). Springer.

Singh, P. K., Sarkar, R., & Nasipuri, M. (2018). Correlation-based classifier combination in the field of pattern recognition. *Computational Intelligence*, *34*(3), 839–874.

So, W. C., Kita, S., & Goldin-Meadow, S. (2009). Using the hands to identify who does what to whom: Gesture and speech go hand-in-hand. *Cognitive Science*, *33*(1), 115–125.

Soga, K., Yoshida, S., & Muneyasu, M. (2024). Exploiting stance similarity and graph neural networks for fake news detection. *Pattern Recognition Letters*, *177*, 26–32.

Song, Y., Morency, L.-P., & Davis, R. (2012). Multimodal human behavior analysis: Learning correlation and interaction across modalities, In *Proceedings of the 14th ACM international conference on multimodal interaction*. Association for Computing Machinery.

Srinivas, N., Veeramachaneni, K., & Osadciw, L. A. (2009). Fusing correlated data from multiple classifiers for improved biometric verification, In *2009 12th international conference on information fusion*. IEEE.

Stepan, P., Kulich, M., & Preucil, L. (2005). Robust data fusion with occupancy grid. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *35*(1), 106–115.

Steyvers, M., & Miller, B. (2015). Cognition and collective intelligence. In T. W. Malone & M. S. Bernstein (Eds.), *Handbook of collective intelligence* (pp. 119–137). MIT Press.

Steyvers, M., Tejeda, H., Kerrigan, G., & Smyth, P. (2022). Bayesian modeling of human-AI complementarity. *Proceedings of the National Academy of Sciences*, *119*(11), e2111547119.

Steyvers, M., Wallsten, T. S., Merkle, E. C., & Turner, B. M. (2014). Evaluating probabilistic forecasts with Bayesian signal detection models. *Risk Analysis*, *34*(3), 435–452.

Stiefelhagen, R., Fugen, C., Gieselmann, R., Holzapfel, H., Nickel, K., & Waibel, A. (2004). Natural human-robot interaction using speech, head pose and gestures, In *2004 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE.

Straub, D., Schultheis, M., Koeppl, H., & Rothkopf, C. A. (2023). Probabilistic inverse optimal control for non-linear partially observable systems disentangles perceptual uncertainty and behavioral costs, In *Advances in neural information processing systems (NeurIPS)*, Curran Associates, Inc.

Sundaresan, A., Varshney, P. K., & Rao, N. S. (2011). Copula-based fusion of correlated decisions. *IEEE Transactions on Aerospace and Electronic Systems*, *47*(1), 454–471.

Surowiecki, J. (2005). *The wisdom of crowds*. Anchor.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT Press.

Svarverud, E., Gilson, S. J., & Glennerster, A. (2010). Cue combination for 3D location judgements. *Journal of Vision*, *10*(1), 5–5.

Tang, R., & Lin, J. (2017). Honk: A pytorch reimplementation of convolutional neural networks for keyword spotting. *arXiv preprint arXiv:1710.06554*.

Terrades, O. R., Valveny, E., & Tabbone, S. (2009). Optimal classifier fusion in a non-Bayesian probabilistic framework. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*(9), 1630–1644.

Thomaz, A. L., Hoffman, G., & Breazeal, C. (2006). Reinforcement learning with human teachers: Understanding how people want to teach robots, In *ROMAN 2006–the 15th IEEE international symposium on robot and human interactive communication*. IEEE.

Thomaz, A. L., Hoffman, G., & Breazeal, C. (2005). Real-time interactive reinforcement learning for robots, In *AAAI 2005 workshop on human comprehensible machine learning*.

Ting, K. M., & Witten, I. H. (1999). Issues in stacked generalization. *Journal of Artificial Intelligence Research*, *10*, 271–289.

Ting Lee, M.-L. (1996). Properties and applications of the Sarmanov family of bivariate distributions. *Communications in Statistics–Theory and Methods*, *25*(6), 1207–1222.

Todisco, E., Guijarro-Fuentes, P., Collier, J., & Coventry, K. R. (2021). The temporal dynamics of deictic communication. *First Language*, *41*(2), 154–178.

Trick, S., Herbert, F., Rothkopf, C. A., & Koert, D. (2022). Interactive reinforcement learning with Bayesian fusion of multimodal advice. *IEEE Robotics and Automation Letters*, *7*(3), 7558–7565.

Trick, S., Koert, D., Peters, J., & Rothkopf, C. A. (2019). Multimodal uncertainty reduction for intention recognition in human-robot interaction, In *Proceedings of the 2019 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE.

Trick, S., Lott, V., Scherf, L., Rothkopf, C. A., & Koert, D. (2023). What can I help you with: Towards task-independent detection of intentions for interaction in a human-robot environment, In *Proceedings of the 2023 32nd IEEE international conference on robot and human interactive communication (RO-MAN)*. IEEE.

Trick, S., & Rothkopf, C. A. (2022). Bayesian classifier fusion with an explicit model of correlation, In *Proceedings of the 25th international conference on artificial intelligence and statistics (AISTATS)*. PMLR.

Trick, S., Rothkopf, C. A., & Jäkel, F. (2023a). A normative model for Bayesian combination of subjective probability estimates. *Judgment and Decision Making*, *18*, e40.

Trick, S., Rothkopf, C. A., & Jäkel, F. (2023b). Parameter estimation for a bivariate beta distribution with arbitrary beta marginals and positive correlation. *METRON*, 1–18.

Tulyakov, S., Jaeger, S., Govindaraju, V., & Doermann, D. (2008). Review of classifier combination methods. *Machine learning in document analysis and recognition*, 361–386.

Tumer, K., & Ghosh, J. (1995). Theoretical foundations of linear and order statistics combiners for neural pattern classifiers. *Technical Report TR-95-02-98, Computer and Vision Research Center, University of Texas, Austin.*

Turk, M. (2014). Multimodal interaction: A review. *Pattern Recognition Letters*, *36*, 189–195.

Turner, B. M., Steyvers, M., Merkle, E. C., Budescu, D. V., & Wallsten, T. S. (2014). Forecast aggregation via recalibration. *Machine Learning*, *95*(3), 261–289.

Ueda, N., & Nakano, R. (1996). Generalization error of ensemble estimators, In *Proceedings of international conference on neural networks (ICNN)*. IEEE.

Ueda, N., Tanaka, Y., & Fujino, A. (2014). Robust naive Bayes combination of multiple classifications, In *The impact of applications on mathematics*, Springer.

Ulaş, A., Yıldız, O. T., & Alpaydın, E. (2012). Eigenclassifiers for combining correlated classifiers. *Information Sciences*, *187*, 109–120.

Van Ravenzwaaij, D., Cassey, P., & Brown, S. D. (2018). A simple introduction to Markov Chain Monte-Carlo sampling. *Psychonomic Bulletin & Review*, *25*(1), 143–154.

Vaufreydaz, D., Johal, W., & Combe, C. (2016). Starting engagement detection towards a companion robot using multimodal features. *Robotics and Autonomous Systems*, *75*, 4–16.

Veeramachaneni, K., Osadciw, L., Ross, A., & Srinivas, N. (2008). Decision-level fusion strategies for correlated biometric classifiers, In *2008 IEEE computer society conference on computer vision and pattern recognition workshops*. IEEE.

Veeriah, V., Pilarski, P. M., & Sutton, R. S. (2016). Face valuing: Training user interfaces with facial expressions and reinforcement learning. *arXiv preprint arXiv:1606.02807*.

Vilares, I., & Kording, K. (2011). Bayesian models: The structure of the world, uncertainty, behavior, and the brain. *Annals of the New York Academy of Sciences*, *1224*(1), 22–39.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., . . . SciPy 1.0 Contributors. (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, *17*, 261–272.

Vojić, S. (2020). Applications of collaborative industrial robots. *Machines. Technologies. Materials.*, *14*(3), 96–99.

Vovk, V. (2015). The fundamental nature of the log loss function. In L. D. Beklemishev, A. Blass, N. Dershowitz, B. Finkbeiner, & W. Schulte (Eds.), *Fields of logic and computation II* (pp. 307–318). Springer.

Wang, J., Liu, Y., & Chen, Y. (2021). Forecast aggregation via peer prediction, In *Proceedings of the AAAI conference on human computation and crowdsourcing*.

Wang, L., Liu, S., Liu, H., & Wang, X. V. (2020). Overview of human-robot collaboration in manufacturing, In *Proceedings of 5th international conference on the industry 4.0 model for advanced manufacturing*. Springer.

Watt, S. J., Akeley, K., Ernst, M. O., & Banks, M. S. (2005). Focus cues affect perceived depth. *Journal of Vision*, *5*(10), 7–7.

Weber, K., Ritschel, H., Aslan, I., Lingenfelser, F., & André, E. (2018). How to shape the humor of a robot-social behavior adaptation based on reinforcement learning, In *Proceedings of the 20th ACM international conference on multimodal interaction*. Association for Computing Machinery.

Weber, K., Ritschel, H., Lingenfelser, F., & André, E. (2018). Real-time adaptation of a robotic joke teller based on human social signals, In *Proceedings of the 17th inter-*

national conference on autonomous agents and multiagent systems. International Foundation for Autonomous Agents and Multiagent Systems.

Wilson, K. J. (2017). An investigation of dependence in expert judgement studies with multiple experts. *International Journal of Forecasting*, *33*(1), 325–336.

Wilson, K. J., & Farrow, M. (2018). Combining judgements from correlated experts. *Elicitation: The Science and Art of Structuring Judgement*, 211–240.

Winkler, R. L. (1981). Combining probability distributions from dependent information sources. *Management Science*, *27*(4), 479–488.

Winkler, R. L., Grushka-Cockayne, Y., Lichtendahl Jr, K. C., & Jose, V. R. R. (2019). Probability forecasts and their combination: A research perspective. *Decision Analysis*, *16*(4), 239–260.

Wiper, M. P., & French, S. (1995). Combining experts' opinions using a Normal-Wishart model. *Journal of Forecasting*, *14*(1), 25–34.

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, *5*(2), 241–259.

Wong, T.-T. (1998). Generalized Dirichlet distribution in Bayesian analysis. *Applied Mathematics and Computation*, *97*(2–3), 165–181.

Xu, L., & Amari, S.-i. (2009). Combining classifiers and learning mixture-of-experts. In J. R. R. Dopico, J. Dorado, & A. Pazos (Eds.), *Encyclopedia of artificial intelligence* (pp. 318–326). IGI Global.

Xu, Q., Li, L., & Wang, G. (2013). Designing engagement-aware agents for multiparty conversations, In *SIGCHI conference on human factors in computing systems*. Association for Computing Machinery.

Xu, W., Huang, J., & Yan, Q. (2015). Multi-sensor based human motion intention recognition algorithm for walking-aid robot, In *IEEE international conference on robotics and biomimetics*. IEEE.

Yildirim, I. (2012). Bayesian inference: Gibbs sampling. *Technical Note, University of Rochester*.

Yu, Z., Kim, S., Mallipeddi, R., & Lee, M. (2015). Human intention understanding based on object affordance and action classification, In *International joint conference on neural networks (IJCNN)*. IEEE.

Zadeh, L. A. (1983). The role of fuzzy logic in the management of uncertainty in expert systems. *Fuzzy Sets and Systems*, *11*(1–3), 199–227.

Zhang, X., & Ghorbani, A. A. (2020). An overview of online fake news: Characterization, detection, and discussion. *Information Processing & Management*, *57*(2), 102025.

Zhang, Z., Zheng, J., & Magnenat Thalmann, N. (2022). Engagement estimation of the elderly from wild multiparty human-robot interaction. *Computer Animation and Virtual Worlds*, e2120.

Zhang, Z., Zheng, J., & Thalmann, N. M. (2021). Engagement intention estimation in multiparty human-robot interaction, In *2021 30th IEEE international conference on robot & human interactive communication (RO-MAN)*. IEEE.

Zhou, Z.-H. (2012). *Ensemble methods: Foundations and algorithms*. CRC Press.

Zlatintsi, A., Rodomagoulakis, I., Koutras, P., Dometios, A., Pitsikalis, V., Tzafestas, C. S., & Maragos, P. (2018). Multimodal signal processing and learning aspects of human-robot interaction for an assistive bathing robot, In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE.

DECLARATION

I declare that I have developed and written the enclosed doctoral thesis entitled *Bayesian Fusion of Probabilistic Forecasts* completely by myself, and have not used sources or means without declaration in the text. Any thoughts from others or literal quotations are clearly marked. This thesis was not used in the same or in a similar version to achieve an academic grading or is being published elsewhere.

Darmstadt, March 7, 2024

Susanne Gabriele Trick