Bayesian Classifier Fusion with an Explicit Model of Correlation

Susanne Trick
Centre for Cognitive Science &
Institute of Psychology,
Technical University of Darmstadt
susanne.trick@tu-darmstadt.de

Constantin A. Rothkopf
Centre for Cognitive Science &
Institute of Psychology,
Technical University of Darmstadt
constantin.rothkopf@cogsci.tu-darmstadt.de

Abstract

Combining the outputs of multiple classifiers or experts into a single probabilistic classification is a fundamental task in machine learning with broad applications from classifier fusion to expert opinion pooling. Here we present a hierarchical Bayesian model of probabilistic classifier fusion based on a new correlated Dirichlet distribution. This distribution explicitly models positive correlations between marginally Dirichlet-distributed random vectors thereby allowing explicit modeling of correlations between base classifiers or experts. The proposed model naturally accommodates the classic Independent Opinion Pool and other independent fusion algorithms as special cases. It is evaluated by uncertainty reduction and correctness of fusion on synthetic and real-world data sets. We show that a change in performance of the fused classifier due to uncertainty reduction can be Bayes optimal even for highly correlated base classifiers.

1 INTRODUCTION

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 (Kittler et al., 1998; Dietterich, 2000; Mohandes et al., 2018; Pirs and 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 (Schubert et al., 2004; Mohandes et al., 2018), although in practice, classifiers trained on the same target as well as human experts are highly correlated (Jacobs, 1995).

Different strategies for coping with correlated classifiers have been proposed, such as selecting only those classifiers with the lowest correlation (Petrakos et al., 2000; Prabhakar and Jain, 2002; Goebel and Yan, 2004; Faria et al., 2013; Singh et al., 2018), explicitly decorrelating the classifiers before fusion (Ulaş et al., 2012), or weighting them according to their correlation (Srinivas et al., 2009; Terrades et al., 2009; Lacoste et al., 2014; Safont et al., 2019). While there are several non-Bayesian models of improved fusion of correlated classifiers (Drakopoulos and Lee, 1988; Kam et al., 1991; Baertlein et al., 2001; Veeramachaneni et al., 2008; Sundaresan et al., 2011; Ma et al., 2013), 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 (Drakopoulos and Lee, 1988; Tumer and Ghosh, 1996; Kuncheva and Jain, 2000; Petrakos et al., 2000; Baertlein et al., 2001; Zhou, 2012), this has not been shown for probabilistic classifiers. Here,

Proceedings of the 25$^{\text{th}}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151.

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 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.

2 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 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 (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 BAYESIAN MODELS OF CLASSIFIER FUSION

Throughout this work, we assume K base classifiers $C_k, k = 1, ..., K$ 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 $t_i = 1, ..., J$. 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 $P(t_i|o_i^1, ..., o_i^K)$. Accordingly, in the following we investigate Bayes optimal fusion methods with successively more general assumptions. In Section 3.1 we start with assuming independent classifiers whose behavior is not known. In Section 3.2 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 3.3, which explicitly models classifiers' correlations. Our implementation of the proposed fusion methods is publicly available at `https://github.com/RothkopfLab/Bayesian_Correlated_Classifier_Fusion`.

3.1 Independent Opinion Pool

If we assume that the outputs of all base classifiers are conditionally independent given t_i with an uninformed prior, by applying Bayes' rule we can transform the sought $P(t_i|o_i^1, ..., o_i^K)$ to:

$$P(t_i|o_i^1, ..., o_i^K) \propto \prod_{k=1}^{K} P(t_i|o_i^k), \tag{1}$$

which needs to be renormalized to sum to 1 (Proof in SM). This fusion rule, which is known as Independent Opinion Pool (IOP) (Berger, 1985), is therefore Bayes optimal given the stated assumptions. Also, it 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), and the more uncertain a base distribution, the less it affects the resulting fused distribution (Hayman and Eklundh, 2002).

3.2 Independent Fusion Model

Although IOP is Bayes optimal given conditionally independent base classifiers and an uninformed prior, 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, while still assuming conditional independence of classifiers and an uninformed prior over classes. 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.

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. Thus, if t_i can take J different values, each base classifier's outputs are modeled by J Dirichlet distributions, $P(\boldsymbol{x_i^k}|t_i = 1), ..., P(\boldsymbol{x_i^k}|t_i = J)$. The graphical model of the proposed IFM is shown in Figure 1(a). 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 $\boldsymbol{p} = (\frac{1}{J}, ..., \frac{1}{J})$. $\boldsymbol{\alpha}$ holds the parameters of the Dirichlet distributions that model the classifiers' outputs. $\boldsymbol{\alpha_j^k}$ with $\alpha_{j\,l}^k > 0$ for $l = 1, ..., J$ 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.

3.2.1 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 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 scale set to 10^{-3}. Of course, one could choose any other prior given additional domain knowledge about the data. In the following, we take the expectations of inferred posterior distributions as point estimates for $\boldsymbol{\alpha_j^k}$.

3.2.2 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 $\boldsymbol{\alpha}$, $P(t_i|\boldsymbol{x_i^1}, ..., \boldsymbol{x_i^K}, \boldsymbol{\alpha})$, 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 pos-

terior fused distribution can be derived analytically:

$$p(t_i = j|\boldsymbol{x_i}, \boldsymbol{\alpha_j}) \propto \prod_{k=1}^{K} \frac{1}{\mathrm{B}(\boldsymbol{\alpha_j^k})} \prod_{l=1}^{J} (x_{i\,l}^k)^{\alpha_{j\,l}^k - 1}. \quad (2)$$

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

As (2) shows, 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 potential learned biases. Moreover, also the variances and uncertainties of the base classifiers can be considered for fusion.

This 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 simplify (2) to:

$$p(t_i = j|\boldsymbol{x_i}, \boldsymbol{\alpha_j}) \propto (x_{i\,j}^1)^n (x_{i\,j}^2)^m \quad (3)$$

for $j = 1, 2, 3$. 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. 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 expectation, which is a categorical distribution. The precision of a Dirichlet distribution with parameter $\boldsymbol{\alpha}$, defined as $\sum_{j=1}^{J} \alpha_j$, is higher, the more concentrated the distribution is around the Dirichlet's expectation (Huang, 2005). Thus, a Dirichlet distribution with a high precision models a classifier with a low variance. On the other hand, the entropy of a Dirichlet's expectation quantifies the average uncertainty of the modeled classifier. If we keep a fixed and increase n, the precision of the corresponding Dirichlet distribution increases. Also, it can be shown that its expectation uncertainty decreases (Proof in SM). Thus, the lower classifier C_1's variance and uncertainty, the higher is its fusion impact and uncertainty reduction through fusion. If we instead increase a while keeping n fixed, this again increases precision and reduces C_1's variance, but also increases its mean uncertainty (Proof in SM). Hence, a classifier with low variance and high uncertainty has the same fusion impact as a classifier with high variance and low uncertainty.

Note that if we set $K = 1$ in (2), 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.

3.3 Correlated Fusion Model

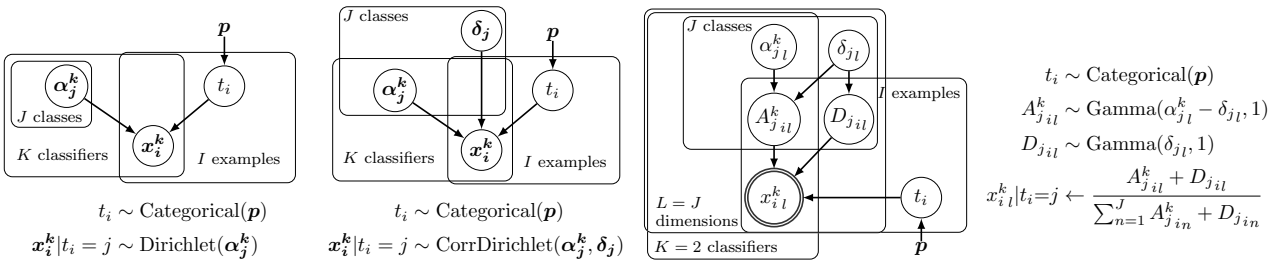
The IFM introduced in Section 3.2 enables optimal fusion of categorical output distributions of conditionally independent base classifiers. However, in practice most classifiers trained on the same target are highly correlated (Jacobs, 1995). Therefore, we extend the IFM to a Correlated Fusion Model (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 $P(\boldsymbol{x_i^1}, ..., \boldsymbol{x_i^K}|t_i)$ with a new correlated Dirichlet distribution that can express correlations between the classifiers' outputs.

3.3.1 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* $x_1, ..., x_J$ of a Dirichlet variate $\boldsymbol{x}$ (Connor and Mosimann, 1969; Wong, 1998; Linderman et al., 2015), here we introduce a correlated Dirichlet distribution that models *correlations between two 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.

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

$$x_l^k = \frac{A_l^k + D_l}{\sum_{n=1}^{J} A_n^k + D_n}, \quad l = 1, ..., J, k = 1, 2, \quad (4)$$



(a) IFM (b) CFM (c) CFM with latent variables of correlated Dirichlet

Figure 1: Graphical models of the IFM (a), CFM (b) and a detailed CFM for $K = 2$ classifiers with all latent variables (c). The full CFM model for $K > 2$ can be found in the SM.

are marginally Dirichlet-distributed with $\text{Dirichlet}(\boldsymbol{x^1}; \alpha^1_1, ..., \alpha^1_J)$ and $\text{Dirichlet}(\boldsymbol{x^2}; \alpha^2_1, ..., \alpha^2_J)$. Their positive correlation, i.e. positive correlations between x^1_l and x^2_l for $l = 1, ..., J$, is generated by the shared variables $D_1, ..., D_J$ with the correlation parameters $\delta_1, ..., \delta_J$. If δ_l tends to zero for $l = 1, ..., J$, $\boldsymbol{x^1}$ 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. While no closed-form solution for the distribution is available, sampling from it is straightforward so that it can be applied to the CFM.

Figure 1(b) shows the CFM's graphical model. The only difference to the IFM in Figure 1(a) is that classifier outputs $\boldsymbol{x_i^1}, ..., \boldsymbol{x_i^K}$ 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 $\alpha^k_{j\,l} > 0$ 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 1(c) 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 $\delta_{j_l} < \alpha^k_{j\,l}$ for $l = 1, ..., J, k = 1, ..., K$.

3.3.2 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 scale 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 3.2.1 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. As for the IFM, we use the expectation of the inferred posterior distributions as point estimates for $\boldsymbol{\alpha_j^k}$ and $\boldsymbol{\delta_j}$.

3.3.3 Normative Fusion Behavior

The fusion of K categorical base distributions $\boldsymbol{x_i^1}, ..., \boldsymbol{x_i^K}$ 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 $\boldsymbol{\delta}$, $P(t_i|\boldsymbol{x_i^1}, ..., \boldsymbol{x_i^K}, \boldsymbol{\alpha}, \boldsymbol{\delta})$. 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 $\boldsymbol{x_i^1}, ..., \boldsymbol{x_i^K}$ to be observed, inference of latent t_i can be performed with Gibbs Sampling. 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 in Section 3.2.2. Thus, bias, variance, and uncertainty 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. We implement inference using JAGS (Plummer et al., 2003). The blue bars in Figure 2 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)). 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 2 show the fusion results given different correlation levels for marginal parameters that imply increased uncertainty reduction compared to IOP ($n = m = 2$ in (3)) 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 Section 3.2.2). 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.

4 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.

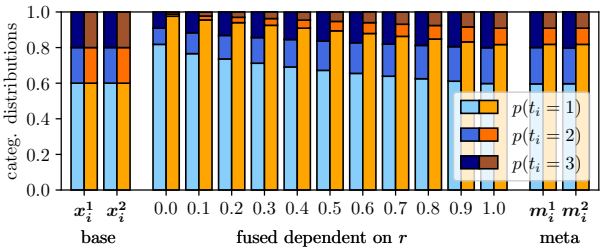
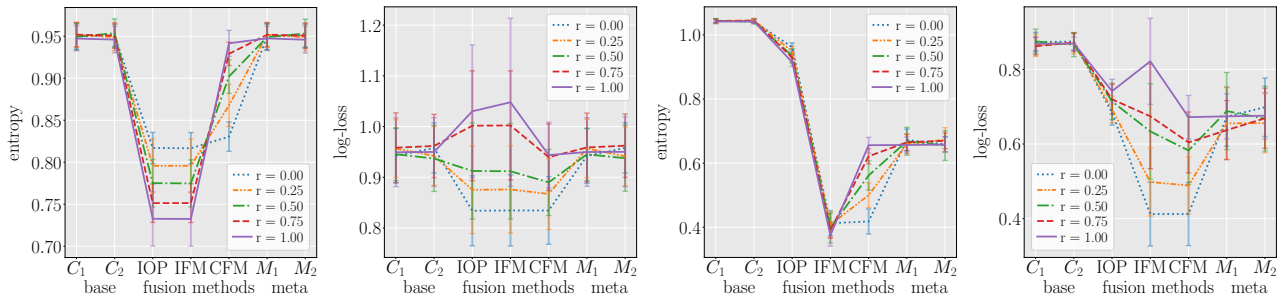


Figure 2: The base distributions $\boldsymbol{x_i^1}$=$\boldsymbol{x_i^2}$=(0.6, 0.2, 0.2) are fused using the CFM assuming different marginal parameters and correlations. Each bar represents a categorical distribution consisting of the probabilities for $p(t_i = 1)$, $p(t_i = 2)$, $p(t_i = 3)$. For 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 $\boldsymbol{m_i^2}$.

4.1 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 1(b)). To show the normative fusion behavior depending on the base classifiers' correlation, for two 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 two 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, one represents two classifiers with decreased variance and uncertainty.

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(a), therefore, the results of IOP and the IFM are equal regarding entropy and log-loss. The shown entropies reveal



(a) SIM 1: $\boldsymbol{\alpha^1} = \boldsymbol{\alpha^2} = ((3,2,2),(2,3,2),(2,2,3))$ (b) SIM 2: $\boldsymbol{\alpha^1} = \boldsymbol{\alpha^2} = ((12,8,8),(8,12,8),(8,8,12))$

Figure 3: Fusion performances on simulated data in terms of mean entropy and log-loss. We compare the performance 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) and higher reinforcement due to decreased classifier variance and uncertainty (b). Standard deviations are shown as error bars.

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) 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 and uncertainty is decreased. Accordingly, Figure 3(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(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.

4.2 Real Data Sets

In addition to simulated data sets, we also evaluated the CFM on 5 real data sets, Bookies A, Bookies B, DNA A, DNA B, DNA C. Both Bookies data sets are composed of $K = 2$ bookmakers' odds for football matches of the English Premier League[1] (Bookies A) and the German Bundesliga[2] (Bookies B). The target variable has $J = 3$ possible outcomes, and for each match example, 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 Bookies data sets is composed of two categorical distributions and a true class label. The correlation between the bookmakers' predictions is approximately 1 in both data sets; 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] with a target variable with $J = 3$ possible outcomes was used to construct three more data sets for evaluating the CFM. For each, 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 same training data but different hyperparameters ($k = 120$

[1] https://www.football-data.co.uk/englandm.php
[2] https://www.football-data.co.uk/germanym.php
[3] https://archive.ics.uci.edu/ml/datasets/Molecular+Biology+(Splice-junction+Gene+Sequences)

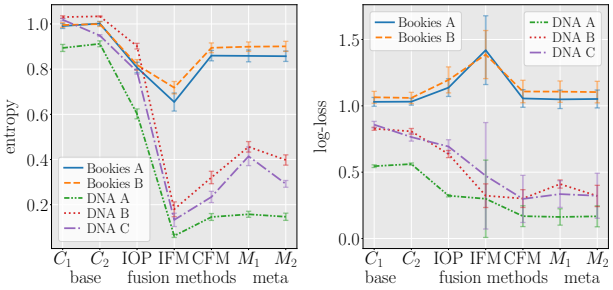


Figure 4: Fusion performances on real data in terms of mean entropy and log-loss. We compare the performance of base classifiers C_1, C_2, the fusion methods IOP, IFM, and CFM, and the meta classifiers M_1, M_2. Standard deviations are shown as error bars.

and $k = 150$). The correlation between both classifiers is approximately 1; it ranges from 0.962 to 0.986 for different dimensions and values of t_i. For DNA B, we trained two classifiers using the same classification method (kNN, $k = 50$) but different training data. Their correlation ranges from 0.463 to 0.709 for different dimensions and values for t_i. DNA C was created by training two different classifiers, a kNN classifier ($k = 50$) and a Random Forest classifier, on the same training set. The correlation between their output distributions ranges from 0.5 to 0.693 in different dimensions and for different values of t_i. More detailed information on the data sets can be found in the SM.

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 random seeds, and expectations and standard deviations of the resulting performance measures were computed, which are shown in Figure 4.

For the three highly correlated data sets, Bookies A, B, 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, as also noticed by Pirs and Strumbelj (2019).

Table 1: Comparison of log-losses of the CFM and Pirs' method (Pirs and Strumbelj, 2019) on simulated and real data.

data set	CFM ($\mu \pm \sigma$)	Pirs ($\mu \pm \sigma$)
SIM 1 $_{r=0.0}$	0.834 ± 0.067	0.915 ± 0.03
SIM 1 $_{r=0.5}$	0.89 ± 0.065	0.938 ± 0.039
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 $_{r=0.5}$	0.583 ± 0.092	0.66 ± 0.065
SIM 2 $_{r=1.0}$	0.672 ± 0.058	0.717 ± 0.041
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

For the less correlated data sets, 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.

4.3 Comparison to Pirs and Strumbelj

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 1, the CFM outperforms on all tested data sets. The simulated data sets not displayed in the table showed similar results but are left out for brevity. In addition to the real data sets discussed in Section 4.2 we also compared the CFM and Pirs' model on an additional data set equivalent to Bookies A but with $K = 3$ bookmakers, Bookies C. Also on this data set, the CFM outperform Pirs' method.

5 CONCLUSION

In this work, we introduced a Bayesian model of classifier fusion based on a new correlated Dirichlet distribution. We derived Bayes optimal fusion behavior for probabilistic classifiers that output categorical distributions, which considers the classifiers' bias, variance, uncertainty, and correlation. We showed that uncer-

tainty 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 paper, 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.

Acknowledgments

This work has been funded by the German Federal Ministry of Education and Research (BMBF) (projects Kobo34 [16SV7984] and IKIDA [01IS20045]). Additionally, we acknowledge support by the Hessian Ministry of Higher Education, Research, Science, and the Arts (HMWK) (projects "The Third Wave of AI" and "The Adaptive Mind") and the Hessian research priority program LOEWE within the project "WhiteBox". We thank the anonymous reviewers for their constructive comments, which helped to improve our work.

References

Tiana Rakotovao Andriamahefa. *Integer Occupancy Grids: a probabilistic multi-sensor fusion framework for embedded perception*. PhD thesis, 2017.

Brian A Baertlein, Wen-Jiao Liao, and De-Hui Chen. Predicting sensor fusion performance using theoretical models. In *Detection and Remediation Technologies for Mines and Minelike Targets VI*, volume 4394, pages 1035–1046. International Society for Optics and Photonics, 2001.

James O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer, London, 1985.

Robert J Connor and James E Mosimann. Concepts of independence for proportions with a generalization of the Dirichlet distribution. *Journal of the American Statistical Association*, 64(325):194–206, 1969.

Thomas G Dietterich. Ensemble methods in machine learning. In *International Workshop on Multiple Classifier Systems*, pages 1–15. Springer, 2000.

E Drakopoulos and Chung Chieh Lee. Optimum fusion of correlated local decisions. In *Proceedings of the 27th IEEE Conference on Decision and Control*, pages 2489–2494. IEEE, 1988.

Fabio A Faria, Jefersson A dos Santos, Sudeep Sarkar, Anderson Rocha, and Ricardo da S Torres. Classifier selection based on the correlation of diversity measures: When fewer is more. In *2013 XXVI Conference on Graphics, Patterns and Images*, pages 16–23. IEEE, 2013.

Simon French. Updating of belief in the light of someone else's opinion. *Journal of the Royal Statistical Society: Series A (General)*, 143(1):43–48, 1980.

Christian Genest, James V Zidek, et al. Combining probability distributions: A critique and an annotated bibliography. *Statistical Science*, 1(1):114–135, 1986.

Kai Goebel and Weizhong Yan. Choosing classifiers for decision fusion. In *Proceedings of the 7th International Conference on Information Fusion*, volume 1, pages 563–568, 2004.

Mohammad Ghasemi Hamed and Ahmad Akbari. Hierarchical Bayesian classifier combination. In *International Conference on Machine Learning and Data Mining in Pattern Recognition*, pages 113–125. Springer, 2018.

Eric Hayman and Jan-Olof Eklundh. Probabilistic and voting approaches to cue integration for figure-ground segmentation. In Anders Heyden, Gunnar Sparr, Mads Nielsen, and Peter Johansen, editors, *Computer Vision — ECCV 2002*, pages 469–486, Berlin, Heidelberg, 2002. Springer Berlin Heidelberg. ISBN 978-3-540-47977-2.

Jonathan Huang. Maximum likelihood estimation of dirichlet distribution parameters. *CMU Technique Report*, 2005.

Robert A Jacobs. Methods for combining experts' probability assessments. *Neural Computation*, 7(5): 867–888, 1995.

Mohamed N Jouini and Robert T Clemen. Copula models for aggregating expert opinions. *Operations Research*, 44(3):444–457, 1996.

Moshe Kam, Qiang Zhu, and W Steven Gray. On distributed detection with correlated local detectors. In *1991 American Control Conference*, pages 2174–2175. IEEE, 1991.

Hyun-Chul Kim and Zoubin Ghahramani. Bayesian classifier combination. In *Artificial Intelligence and Statistics*, pages 619–627, 2012.

Josef Kittler, Mohamad Hatef, Robert PW Duin, and Jiri Matas. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–239, 1998.

Ludmila I Kuncheva and Lakhmi C Jain. Designing classifier fusion systems by genetic algorithms. *IEEE Transactions on Evolutionary Computation*, 4 (4):327–336, 2000.

Alexandre Lacoste, Mario Marchand, François Laviolette, and Hugo Larochelle. Agnostic Bayesian learning of ensembles. In *International Conference on Machine Learning*, pages 611–619, 2014.

Scott Linderman, Matthew J Johnson, and Ryan P Adams. Dependent multinomial models made easy: Stick-breaking with the polya-gamma augmentation. In *Advances in Neural Information Processing Systems*, pages 3456–3464, 2015.

Dennis V Lindley. Reconciliation of discrete probability distributions. *Bayesian Statistics*, 2:375–390, 1985.

Andy Jinhua Ma, Pong C Yuen, and Jian-Huang Lai. Linear dependency modeling for classifier fusion and feature combination. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(5):1135–1148, 2013.

Mohamed Mohandes, Mohamed Deriche, and Salihu O Aliyu. Classifiers combination techniques: A comprehensive review. *IEEE Access*, 6:19626–19639, 2018.

Alfredo Nazabal, Pablo Garcia-Moreno, Antonio Artes-Rodriguez, and Zoubin Ghahramani. Human activity recognition by combining a small number of classifiers. *IEEE Journal of Biomedical and Health Informatics*, 20(5):1342–1351, 2016.

Michalis Petrakos, Ioannis Kannelopoulos, Jon Atli Benediktsson, and Martino Pesaresi. The effect of correlation on the accuracy of the combined classifier in decision level fusion. In *IGARSS 2000. IEEE 2000 International Geoscience and Remote Sensing Symposium. Taking the Pulse of the Planet: The Role of Remote Sensing in Managing the Environment. Proceedings*, volume 6, pages 2623–2625. IEEE, 2000.

Gregor Pirs and Erik Strumbelj. Bayesian combination of probabilistic classifiers using multivariate normal mixtures. *Journal of Machine Learning Research*, 20:51–1, 2019.

Martyn Plummer et al. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*, volume 124, pages 1–10. Vienna, Austria., 2003.

Salil Prabhakar and Anil K Jain. Decision-level fusion in fingerprint verification. *Pattern Recognition*, 35 (4):861–874, 2002.

Gonzalo Safont, Addisson Salazar, and Luis Vergara. Multiclass alpha integration of scores from multiple classifiers. *Neural Computation*, 31(4):806–825, 2019.

Christine M Schubert, Nathan J Leap, Mark E Oxley, and Kenneth W Bauer Jr. Quantifying the correlation effects of fused classifiers. In *Signal Processing, Sensor Fusion, and Target Recognition XIII*, volume 5429, pages 373–383. International Society for Optics and Photonics, 2004.

Edwin Simpson, Stephen Roberts, Ioannis Psorakis, and Arfon Smith. Dynamic Bayesian combination of multiple imperfect classifiers. In *Decision Making and Imperfection*, pages 1–35. Springer, 2013.

Pawan Kumar Singh, Ram Sarkar, and Mita Nasipuri. Correlation-based classifier combination in the field of pattern recognition. *Computational Intelligence*, 34(3):839–874, 2018.

Nisha Srinivas, Kalyan Veeramachaneni, and Lisa Ann Osadciw. Fusing correlated data from multiple classifiers for improved biometric verification. In *2009 12th International Conference on Information Fusion*, pages 1504–1511. IEEE, 2009.

Ashok Sundaresan, Pramod K Varshney, and Nageswara SV Rao. Copula-based fusion of correlated decisions. *IEEE Transactions on Aerospace and Electronic Systems*, 47(1):454–471, 2011.

Oriol Ramos Terrades, Ernest Valveny, and Salvatore Tabbone. Optimal classifier fusion in a non-Bayesian probabilistic framework. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(9):1630–1644, 2009.

Kagan Tumer and Joydeep Ghosh. Theoretical foundations of linear and order statistics combiners for neural pattern classifiers. Technical report, IEEE Transactions on Neural Networks, 1996.

Naonori Ueda, Yusuke Tanaka, and Akinori Fujino. Robust naive Bayes combination of multiple classifications. In *The Impact of Applications on Mathematics*, pages 141–155. Springer, 2014.

Aydın Ulaş, Olcay Taner Yıldız, and Ethem Alpaydın. Eigenclassifiers for combining correlated classifiers. *Information Sciences*, 187:109–120, 2012.

Kalyan Veeramachaneni, Lisa Osadciw, Arun Ross, and Nisha Srinivas. Decision-level fusion strategies for correlated biometric classifiers. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–6. IEEE, 2008.

Vladimir Vovk. The fundamental nature of the log loss function. In *Fields of Logic and Computation II*, pages 307–318. Springer, 2015.

Robert L Winkler. Combining probability distributions from dependent information sources. *Management Science*, 27(4):479–488, 1981.

Tzu-Tsung Wong. Generalized Dirichlet distribution in Bayesian analysis. *Applied Mathematics and Computation*, 97(2-3):165–181, 1998.

Zhi-Hua Zhou. *Ensemble methods: foundations and algorithms*. CRC press, 2012.

Supplementary Material: Bayesian Classifier Fusion with an Explicit Model of Correlation

A DERIVATION OF THE FUSION RULE OF INDEPENDENT OPINION POOL

In the following, we show in detail how the fusion rule (1) known as Independent Opinion Pool can be derived using Bayes' rule if we assume conditional independence between the fused classifiers given the true class label t_i and an uninformed prior over t_i. Particularly, this derivation proves that Independent Opinion Pool is the Bayes optimal fusion rule given these assumptions.

By applying Bayes' rule, we can transform

$$P(t_i|o_i^1, ..., o_i^K) = \frac{P(o_i^1, ..., o_i^K|t_i)P(t_i)}{P(o_i^1, ..., o_i^K)}. \tag{A.5}$$

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

$$P(t_i|o_i^1, ..., o_i^K) = \frac{\prod_{k=1}^K P(o_i^k|t_i)P(t_i)^K}{P(o_i^1, ..., o_i^K)P(t_i)^{K-1}}. \tag{A.6}$$

By again applying Bayes' rule and commutativity we get

$$P(t_i|o_i^1, ..., o_i^K) = \frac{\prod_{k=1}^K \left(\frac{P(t_i|o_i^k)P(o_i^k)}{\cancel{P(t_i)}} \cancel{P(t_i)}\right)}{P(o_i^1, ..., o_i^K)P(t_i)^{K-1}} \tag{A.7}$$

$$= \frac{\prod_{k=1}^K P(t_i|o_i^k)P(o_i^k)}{P(o_i^1, ..., o_i^K)P(t_i)^{K-1}} \tag{A.8}$$

$$= \frac{\prod_{k=1}^K P(t_i|o_i^k) \prod_{k=1}^K P(o_i^k)}{P(o_i^1, ..., o_i^K)P(t_i)^{K-1}} \tag{A.9}$$

$$= \underbrace{\frac{\prod_{k=1}^K P(o_i^k)}{P(o_i^1, ..., o_i^K)}}_{\text{constant}} \frac{\prod_{k=1}^K P(t_i|o_i^k)}{P(t_i)^{K-1}}, \tag{A.10}$$

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

$$P(t_i|o_i^1, ..., o_i^K) \propto \frac{\prod_{k=1}^K P(t_i|o_i^k)}{P(t_i)^{K-1}}. \tag{A.11}$$

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

$$P(t_i|o_i^1, ..., o_i^K) \propto \prod_{k=1}^K P(t_i|o_i^k), \tag{A.12}$$

which needs to be renormalized to sum to 1.

B DERIVATION OF THE FUSION BEHAVIOR OF THE INDEPENDENT FUSION MODEL

The normative fusion behavior of the Independent Fusion Model introduced in Section 3.2 can be derived analytically. We show the derivation in detail, as well for the most general case (Section B.1) as for two examples

demonstrating the influence of classifier variance and uncertainty (Section B.2) and bias (Section B.3) on the normative fusion behavior.

B.1 Derivation of the Normative Fusion Behavior

The joint distribution of the Independent Fusion Model shown in Figure 1(a) is

$$P(\boldsymbol{x_i}, \boldsymbol{\alpha}, t_i) = P(t_i|\boldsymbol{p})P(\boldsymbol{\alpha}) \prod_{k=1}^{K} P(\boldsymbol{x_i^k}|\boldsymbol{\alpha_j^k}, t_i). \tag{B.13}$$

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

$$P(\boldsymbol{x_i}, \boldsymbol{\alpha}, t_i) \propto \prod_{k=1}^{K} P(\boldsymbol{x_i^k}|\boldsymbol{\alpha_j^k}, t_i). \tag{B.14}$$

The fusion rule (2) can be obtained by computing the posterior probability of $t_i = j$ for $j = 1, ..., J$ given the categorical distributions $\boldsymbol{x_i}$ and the respective $\boldsymbol{\alpha}$ learned before,

$$\begin{aligned} p(t_i = j|\boldsymbol{x_i}, \boldsymbol{\alpha_j}) &\propto p(t_i = j, \boldsymbol{x_i}, \boldsymbol{\alpha_j}) && \text{(B.15)}\\ &\propto \prod_{k=1}^{K} p(\boldsymbol{x_i^k}|\boldsymbol{\alpha_j^k}, t_i = j) && \text{(B.16)}\\ &= \prod_{k=1}^{K} \text{Dirichlet}(\boldsymbol{x_i^k}; \boldsymbol{\alpha_j^k}) && \text{(B.17)}\\ &= \prod_{k=1}^{K} \frac{1}{\text{B}(\boldsymbol{\alpha_j^k})} \prod_{l=1}^{J} (x_{i\,l}^k)^{\alpha_{j_l}^k - 1}. && \text{(B.18)} \end{aligned}$$

This unnormalized posterior probability can now be computed for all $t_i = j$ for $j = 1, ..., J$. Normalizing the respective values in order to make them sum to 1 gives the posterior fused categorical distribution.

B.2 Influence of Classifier Variance and Uncertainty on the Fusion Behavior

In order to demonstrate how a classifier's variance and uncertainty affect the normative fusion behavior, we derive the fusion rule for fusing two exemplary classifiers C_1 and C_2, modeled with parameters $\boldsymbol{\alpha_1^1} = (a+n, a, a)$ for $t_i = 1$, $\boldsymbol{\alpha_2^1} = (a, a+n, a)$ for $t_i = 2$, and $\boldsymbol{\alpha_3^1} = (a, a, a+n)$ for $t_i = 3$ for C_1 and $\boldsymbol{\alpha_1^2} = (b+m, b, b)$, $\boldsymbol{\alpha_2^2} = (b, b+m, b)$, and $\boldsymbol{\alpha_3^2} = (b, b, b+m)$ for C_2 with $a, b, n, m > 0$.

The detailed derivation of the resulting fusion rule (3) is shown in the following.

If we set $K = 2$ and $J = 3$ according to the chosen 2 classifiers which differentiate between 3 classes, (B.18) can be simplified to

$$\begin{aligned} p(t_i = j|\boldsymbol{x_i}, \boldsymbol{\alpha_j}) &\propto \prod_{k=1}^{2} \frac{1}{\text{B}(\boldsymbol{\alpha_j^k})} \prod_{l=1}^{3} (x_{i\,l}^k)^{\alpha_{j_l}^k - 1} && \text{(B.19)}\\ &\propto \prod_{k=1}^{2} \prod_{l=1}^{3} (x_{i\,l}^k)^{\alpha_{j_l}^k - 1} && \text{(B.20)}\\ &= (x_{i\,1}^1)^{\alpha_{j_1}^1 - 1} (x_{i\,2}^1)^{\alpha_{j_2}^1 - 1} (x_{i\,3}^1)^{\alpha_{j_3}^1 - 1} (x_{i\,1}^2)^{\alpha_{j_1}^2 - 1} (x_{i\,2}^2)^{\alpha_{j_2}^2 - 1} (x_{i\,3}^2)^{\alpha_{j_3}^2 - 1}. && \text{(B.21)} \end{aligned}$$

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

$$p(t_i = 1|\boldsymbol{x_i}, \boldsymbol{\alpha_1}) = (x^1_{i\,1})^{a+n-1}(x^1_{i\,2})^{a-1}(x^1_{i\,3})^{a-1}(x^2_{i\,1})^{b+m-1}(x^2_{i\,2})^{b-1}(x^2_{i\,3})^{b-1} \tag{B.22}$$

$$= (x^1_{i\,1})^n(x^2_{i\,1})^m \underbrace{(x^1_{i\,1}x^1_{i\,2}x^1_{i\,3})^{a-1}(x^2_{i\,1}x^2_{i\,2}x^2_{i\,3})^{b-1}}_{\text{constant}} \tag{B.23}$$

$$\propto (x^1_{i\,1})^n(x^2_{i\,1})^m. \tag{B.24}$$

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

$$p(t_i = 2|\boldsymbol{x_i}, \boldsymbol{\alpha_2}) \propto (x^1_{i\,2})^n(x^2_{i\,2})^m \tag{B.25}$$

$$p(t_i = 3|\boldsymbol{x_i}, \boldsymbol{\alpha_3}) \propto (x^1_{i\,3})^n(x^2_{i\,3})^m \tag{B.26}$$

and can thus formulate the general fusion rule (3)

$$p(t_i = j|\boldsymbol{\alpha_j}, \boldsymbol{x_i}) \propto (x^1_{i\,j})^n(x^2_{i\,j})^m, \tag{B.27}$$

while, again, the resulting values must be normalized to sum to 1.

The shown example demonstrates the relation between variance and uncertainty of a classifier and the fusion behavior. A higher value of n results in a higher fusion impact of classifier C_1 and more uncertainty reduction of the fused distribution; equivalently this applies to m and classifier C_2.

In the following, we will show how n corresponds to the variance and uncertainty of classifier C_1.

Classifier variance can be quantified with the precision s^k_j of the corresponding Dirichlet distributions, which is the sum of all elements in $\boldsymbol{\alpha^k_j}$ for each $j = 1, ..., 3$. 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 $j = 1, ..., 3$ and therefore can be regarded as a measure for the classifier's variance. The precision, sometimes also termed as concentration parameter, determines how concentrated the categorical samples are around the Dirichlet distribution's expectation $\boldsymbol{m^k_j}$. The higher the precision, the closer the categorical samples, i.e. the classifier outputs, are to the expectation and thus the lower is the classifier's variance.

Classifier uncertainty can be described by the entropies of the modeling Dirichlet distributions' expectations $\boldsymbol{m^k_j}$ for $j = 1, ..., 3$. Again, in general the entropies of the expectations can be different for different values of j, but due to the chosen example parameters the expectations' entropies are equal for $j = 1, ..., 3$. Therefore, we can regard this mean entropy 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 obviously increases, implying a lower variance of classifier C_1. In addition, its mean entropy decreases, which we show in the following.

The expectation of classifier C_1 for $j = 1$ is $\boldsymbol{m^1_1} = (\frac{a+n}{3a+n}, \frac{a}{3a+n}, \frac{a}{3a+n})$. Thus, its entropy is

$$\mathrm{H}_{\boldsymbol{m}_1^1} = -\left(\frac{a+n}{3a+n}\cdot\log\left(\frac{a+n}{3a+n}\right)+\frac{a}{3a+n}\cdot\log\left(\frac{a}{3a+n}\right)+\frac{a}{3a+n}\cdot\log\left(\frac{a}{3a+n}\right)\right) \tag{B.28}$$

$$= -\frac{1}{3a+n}((a+n)\cdot\log(a+n)-(a+n)\cdot\log(3a+n)+a\cdot\log(a)-a\cdot\log(3a+n) \tag{B.29}$$
$$+a\cdot\log(a)-a\cdot\log(3a+n))$$

$$= -\frac{1}{3a+n}(a\cdot\log(a+n)+n\cdot\log(a+n)-a\cdot\log(3a+n)-n\cdot\log(3a+n) \tag{B.30}$$
$$+a\cdot\log(a)-a\cdot\log(3a+n)+a\cdot\log(a)-a\cdot\log(3a+n))$$

$$= -\frac{1}{3a+n}(a\cdot(\log(a+n)-\log(3a+n)+\log(a)-\log(3a+n)+\log(a)-\log(3a+n)) \tag{B.31}$$
$$+n\cdot(\log(a+n)-\log(3a+n)))$$

$$= -\frac{1}{3a+n}\left(a\cdot\log\left(\frac{a^2(a+n)}{(3a+n)^3}\right)+n\cdot\log\left(\frac{a+n}{3a+n}\right)\right). \tag{B.32}$$

Differentiating $\mathrm{H}_{\boldsymbol{m}_1^1}$ w.r.t. n yields

$$\mathrm{H}'_{\boldsymbol{m}_1^1}(n) = \frac{a\left(\log\left(\frac{a^2(a+n)}{(3a+n)^3}\right)-3\log\left(\frac{a+n}{3a+n}\right)\right)}{(3a+n)^2}. \tag{B.33}$$

The derivative $\mathrm{H}'_{\boldsymbol{m}_1^1}(n)$ is negative for all $a, n > 0$.

Proof:

$$\mathrm{H}'_{\boldsymbol{m}_1^1}(n) = \frac{a\left(\log\left(\frac{a^2(a+n)}{(3a+n)^3}\right)-3\log\left(\frac{a+n}{3a+n}\right)\right)}{(3a+n)^2} \quad < 0 \tag{B.34}$$

$$\Leftrightarrow \log\left(\frac{a^2(a+n)}{(3a+n)^3}\right)-3\log\left(\frac{a+n}{3a+n}\right) \quad < 0 \tag{B.35}$$

$$\Leftrightarrow \log\left(\frac{a^2(a+n)}{(3a+n)^3}\right)-\log\left(\frac{(a+n)^3}{(3a+n)^3}\right) \quad < 0 \tag{B.36}$$

$$\Leftrightarrow \log\left(\frac{a^2(a+n)}{(3a+n)^3}\cdot\frac{(3a+n)^3}{(a+n)^3}\right) \quad < 0 \tag{B.37}$$

$$\Leftrightarrow \log\left(\frac{a^2}{(a+n)^2}\right) \quad < 0 \tag{B.38}$$

$$\Leftrightarrow \frac{a^2}{(a+n)^2} \quad < 1 \tag{B.39}$$

$$\Leftrightarrow a^2 \quad < (a+n)^2 \tag{B.40}$$

$$\Leftrightarrow a \quad < a+n \tag{B.41}$$

$$\Leftrightarrow 0 \quad < n \tag{B.42}$$

□

Hence, $\mathrm{H}_{\boldsymbol{m}_1^1}$ 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,

$$\mathrm{H}'_{\boldsymbol{m}_1^1}(a) = \frac{-n\left(\log\left(\frac{a^2(a+n)}{(3a+n)^3}\right) - 3\log\left(\frac{a+n}{3a+n}\right)\right)}{(3a+n)^2} \tag{B.43}$$

The derivative $\mathrm{H}'_{\boldsymbol{m}_1^1}(a)$ is positive for all $a, n > 0$.

Proof:

$$\begin{aligned}
\mathrm{H}'_{\boldsymbol{m}_1^1}(a) = \frac{-n\left(\log\left(\frac{a^2(a+n)}{(3a+n)^3}\right) - 3\log\left(\frac{a+n}{3a+n}\right)\right)}{(3a+n)^2} &> 0 && \text{(B.44)}\\
\Leftrightarrow \log\left(\frac{a^2(a+n)}{(3a+n)^3}\right) - 3\log\left(\frac{a+n}{3a+n}\right) &< 0 && \text{(B.45)}\\
\Leftrightarrow \log\left(\frac{a^2(a+n)}{(3a+n)^3}\right) - \log\left(\frac{(a+n)^3}{(3a+n)^3}\right) &< 0 && \text{(B.46)}\\
\Leftrightarrow \log\left(\frac{a^2(a+n)}{(3a+n)^3} \cdot \frac{(3a+n)^3}{(a+n)^3}\right) &< 0 && \text{(B.47)}\\
\Leftrightarrow \log\left(\frac{a^2}{(a+n)^2}\right) &< 0 && \text{(B.48)}\\
\Leftrightarrow \frac{a^2}{(a+n)^2} &< 1 && \text{(B.49)}\\
\Leftrightarrow a^2 &< (a+n)^2 && \text{(B.50)}\\
\Leftrightarrow a &< a+n && \text{(B.51)}\\
\Leftrightarrow 0 &< n && \text{(B.52)}
\end{aligned}$$

□

Consequently, in addition to a decreased variance, the mean entropy $\mathrm{H}_{\boldsymbol{m}_1^1}$ 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 (B.27) 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.

B.3 Influence of Classifier Bias on the Fusion Behavior

The Independent Fusion Model 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 $\operatorname{argmax}(\boldsymbol{\alpha}_j^k) \neq j$ for some class j. As a consequence, also for the Dirichlet's categorical expectation $\boldsymbol{m}_j^k$ it applies that $\operatorname{argmax}(\boldsymbol{m}_j^k) \neq j$. Hence, on average, the classifier would misclassify class j as another class.

The example classifiers introduced in Section B.2 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 $\boldsymbol{\alpha}_1^1 = (a+n, a, a)$ for $t_i = 1$, $\boldsymbol{\alpha}_2^1 = (a, a+n, a)$ for $t_i = 2$, and $\boldsymbol{\alpha}_3^1 = (a, a, a+n)$ for $t_i = 3$ for C_1 and $\boldsymbol{\alpha}_1^2 = (b, b+m, b)$, $\boldsymbol{\alpha}_2^2 = (b+m, b, b)$, and $\boldsymbol{\alpha}_3^2 = (b, b, b+m)$ 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 $t_i = 1$ (B.21) can be transformed to:

$$\begin{aligned}
p(t_i = 1|\boldsymbol{x_i}, \boldsymbol{\alpha_1}) &= (x_{i\,1}^1)^{a+n-1}(x_{i\,2}^1)^{a-1}(x_{i\,3}^1)^{a-1}(x_{i\,1}^2)^{b-1}(x_{i\,2}^2)^{b+m-1}(x_{i\,3}^2)^{b-1} && \text{(B.53)}\\
&= (x_{i\,1}^1)^n(x_{i\,2}^2)^m \underbrace{(x_{i\,1}^1 x_{i\,2}^1 x_{i\,3}^1)^{a-1}(x_{i\,1}^2 x_{i\,2}^2 x_{i\,3}^2)^{b-1}}_{\text{constant}} && \text{(B.54)}\\
&\propto (x_{i\,1}^1)^n(x_{i\,2}^2)^m && \text{(B.55)}
\end{aligned}$$

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

$$p(t_i = 2|\boldsymbol{x_i}, \boldsymbol{\alpha_2}) \propto (x^1_{i\,2})^n (x^2_{i\,1})^m \tag{B.56}$$
$$p(t_i = 3|\boldsymbol{x_i}, \boldsymbol{\alpha_3}) \propto (x^1_{i\,3})^n (x^2_{i\,3})^m \tag{B.57}$$

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 was used.

C MORE DETAILS ON THE CORRELATED FUSION MODEL AND THE CORRELATED DIRICHLET DISTRIBUTION

In the proposed Correlated Fusion Model we jointly model all classifiers' output distributions with a new correlated Dirichlet distribution, which we introduce in Section 3.3.1. Section C.1 shows some examples of the proposed distribution, which were omitted in the manuscript for brevity. Section C.2 presents a more intuitive interpretation of the correlated Dirichlet distribution as a pairwise combination of 3 Dirichlet distributions. In Section C.3 we additionally show a detailed version of the CFM's graphical model including all latent variables of the correlated Dirichlet distribution also for $K > 2$ classifiers. In Section C.4 we give additional information on Gibbs Sampling for parameter inference and fusion with the CFM, including the derivation of the conditional distributions required in Gibbs Sampling for parameter inference with the CFM.

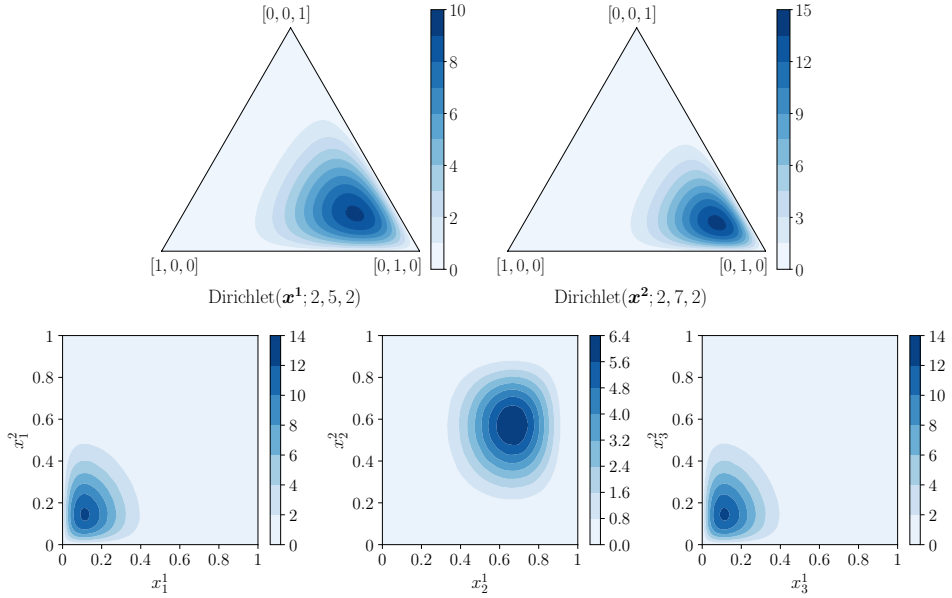
C.1 Examples of the Correlated Dirichlet Distribution

In Section 3.3.1 we introduce the correlated Dirichlet distribution for jointly modeling multiple classifiers' categorical output distributions. While previous generalizations of the Dirichlet distribution such as the generalized Dirichlet distribution (Connor and Mosimann, 1969; Wong, 1998) or the work of Linderman et al. (2015) focused on more flexible correlations between individual random vector entries $x_1, ..., x_J$ of a Dirichlet variate $\boldsymbol{x}$, the correlated Dirichlet distribution can model 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. Figures C.5 and C.6 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 C.5(a)) and 1 (Figure C.6(b)). In the figures, r_{jj} names the correlation in the j-th dimension of the correlated Dirichlet distribution, thus the correlation between x^1_j and x^2_j. As Figure C.6(a) shows, the correlation can also differ for different dimensions of the correlated Dirichlet distribution.

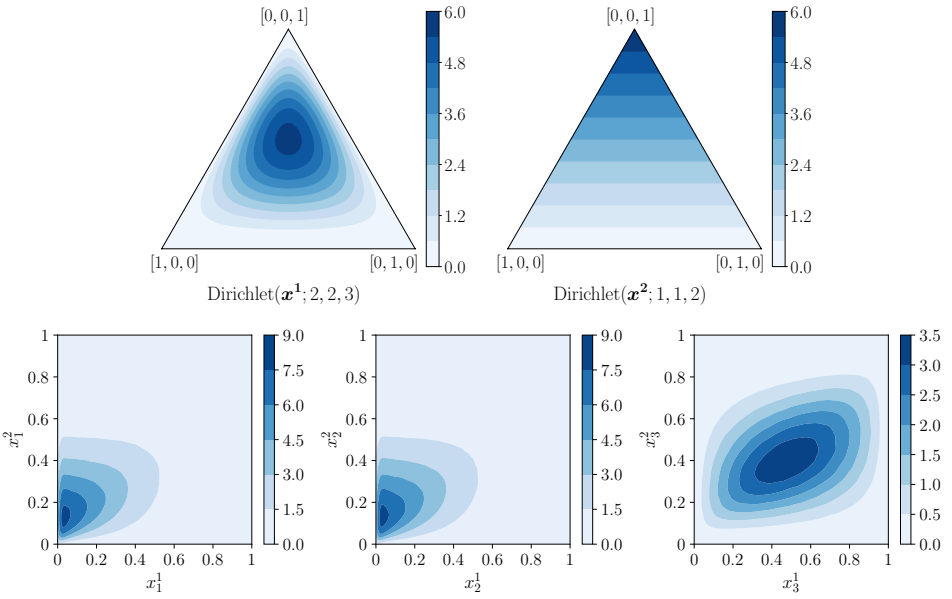
C.2 The Correlated Dirichlet Distribution as a Pairwise Combination of 3 Dirichlet Distributions

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.

To show this we transform the $3J$ independent gamma-distributed random variables $A^1_1, \ldots, A^1_J$, $A^2_1, \ldots, A^2_J$, $D_1, \ldots, D_J$ into three independent gamma- and three independent Dirichlet-distributed random variables

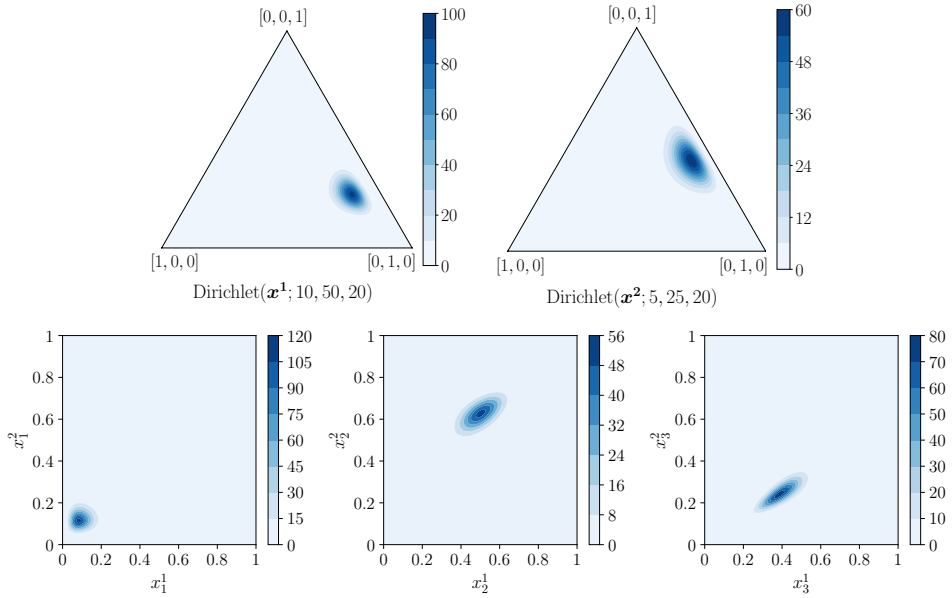


(a) $\boldsymbol{\alpha^1} = (2, 5, 2)$, $\boldsymbol{\alpha^2} = (2, 7, 2)$, $\boldsymbol{\delta} = (0.01, 0.01, 0.01) \rightarrow r_{11} = 0.0, r_{22} = 0.0, r_{33} = 0.0$

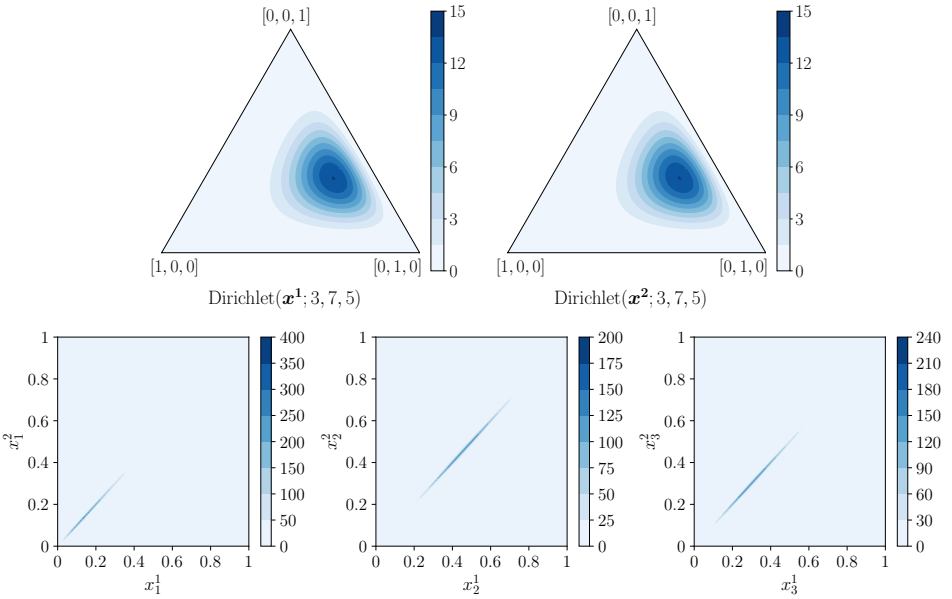


(b) $\boldsymbol{\alpha^1} = (2, 2, 3)$, $\boldsymbol{\alpha^2} = (1, 1, 2)$, $\boldsymbol{\delta} = (0.4, 0.4, 0.9) \rightarrow r_{11} = 0.26, r_{22} = 0.24, r_{33} = 0.29$

Figure C.5: Marginal and joint densities of correlated Dirichlet distributions with selected parameter values leading to low correlations. The simplexes display the marginal Dirichlet distributions of $\boldsymbol{x^1}$ and $\boldsymbol{x^2}$, while the joint densities of x_i^1 and x_i^2, $i = 1, \ldots, 3$, are shown for each dimension of the correlated Dirichlet distribution. In (a) $\boldsymbol{x^1}$ and $\boldsymbol{x^2}$ are independent with dimension-wise correlations $r_{11} = r_{22} = r_{33} = 0.0$ between x_i^1 and x_i^2 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) $\rightarrow r_{11}$=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 C.6: Marginal and joint densities of correlated Dirichlet distributions with selected parameter values. The simplexes display the marginal Dirichlet distributions of $\boldsymbol{x^1}$ and $\boldsymbol{x^2}$, while the joint densities of x_i^1 and x_i^2, $i = 1, \ldots, 3$, 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_i^1 and x_i^2. (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.

$U_1, U_2, U_3, \boldsymbol{W_1}, \boldsymbol{W_2}, \boldsymbol{W_3}$ with

$$\begin{aligned}
U_1 &= \sum_{i=1}^{J} A_i^1, && U_1 \sim \text{Gamma}(\upsilon_1, 1) \\
U_2 &= \sum_{i=1}^{J} A_i^2, && U_2 \sim \text{Gamma}(\upsilon_2, 1) \\
U_3 &= \sum_{i=1}^{J} D_i, && U_3 \sim \text{Gamma}(\upsilon_3, 1) \\
W_{1i} &= \frac{A_i^1}{\sum_{j=1}^{J} A_j^1}, \quad i = 1, \ldots, J, && \boldsymbol{W_1} \sim \text{Dirichlet}(\alpha_1^1 - \delta_1, \ldots, \alpha_J^1 - \delta_J) \\
W_{2i} &= \frac{A_i^2}{\sum_{j=1}^{J} A_j^2}, \quad i = 1, \ldots, J, && \boldsymbol{W_2} \sim \text{Dirichlet}(\alpha_1^2 - \delta_1, \ldots, \alpha_J^2 - \delta_J) \\
W_{3i} &= \frac{D_i}{\sum_{j=1}^{J} D_j}, \quad i = 1, \ldots, J, && \boldsymbol{W_3} \sim \text{Dirichlet}(\delta_1, \ldots, \delta_J)
\end{aligned} \tag{C.58}$$

with

$$\upsilon_1 = \sum_{i=1}^{J} \alpha_i^1 - \delta_i, \qquad \upsilon_2 = \sum_{i=1}^{J} \alpha_i^2 - \delta_i \qquad \upsilon_3 = \sum_{i=1}^{J} \delta_i. \tag{C.59}$$

With these definitions we can then rewrite construction (4) as

$$\begin{aligned}
\boldsymbol{x^1} &= \frac{U_1}{U_1 + U_3} \cdot \boldsymbol{W_1} + \frac{U_3}{U_1 + U_3} \cdot \boldsymbol{W_3} = X' \boldsymbol{W_1} + (1 - X') \boldsymbol{W_3} \\
\boldsymbol{x^2} &= \frac{U_2}{U_2 + U_3} \cdot \boldsymbol{W_2} + \frac{U_3}{U_2 + U_3} \cdot \boldsymbol{W_3} = Y' \boldsymbol{W_2} + (1 - Y') \boldsymbol{W_3}.
\end{aligned} \tag{C.60}$$

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.

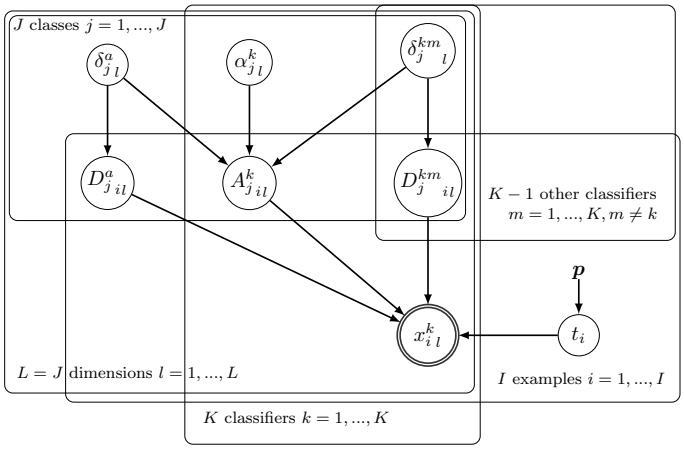
C.3 Detailed Graphical Model of the Correlated Fusion Model for $K > 2$ Classifiers

Figure C.7 shows the graphical model of the CFM given in Figure 1(c) 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, $m = 1, ..., K, m \neq k$ 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 $C_1, ..., C_k$ if $t_i = j$. Thus, note that for the special case of $K = 2$ classifiers $\boldsymbol{\delta_j}$ only consists of $\boldsymbol{\delta_j^a}$.

C.4 Gibbs Sampling

The Correlated Fusion Model requires Gibbs Sampling for parameter inference (Section 3.3.2), i.e. inference of the posterior distribution over parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\delta}$ given observed classifier outputs $\boldsymbol{x}$ and their true labels $\boldsymbol{t}$, $P(\boldsymbol{\alpha}, \boldsymbol{\delta} | \boldsymbol{x}, \boldsymbol{t})$, and fusion, i.e. inference of 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 $\boldsymbol{\delta}$, $P(t_i | \boldsymbol{x_i^1}, ..., \boldsymbol{x_i^K}, \boldsymbol{\alpha}, \boldsymbol{\delta})$ (Section 3.3.3).

Inferring the parameters of the CFM requires inferring the parameters of J J-dimensional correlated Dirichlet distributions, where the j-th correlated Dirichlet distribution jointly models all I_j categorical distributions $\boldsymbol{x_i^1}, ..., \boldsymbol{x_i^K}$ with $t_i = j$.



$$t_i \sim \text{Categorical}(\boldsymbol{p})$$

$$A^k_{j\,il} \sim \text{Gamma}(\alpha^k_{j\,l} - \delta^a_{j\,l} - \sum_{\substack{m=1\\m\neq k}}^{K} \delta^{km}_{j\ \ l}, 1)$$

$$D^{km}_{j\ \ il} \sim \text{Gamma}(\delta^{km}_{j\ \ l}, 1)$$

$$D^a_{j\,il} \sim \text{Gamma}(\delta^a_{j\,l}, 1)$$

$$x^k_{i\,l}|t_i = j \leftarrow \frac{A^k_{j\,il} + D^a_{j\,il} + \sum_{\substack{m=1\\m\neq k}}^{K} D^{km}_{j\ \ il}}{\sum_{n=1}^{J} A^k_{j\,in} + \sum_{n=1}^{J} D^a_{j\,in} + \sum_{\substack{m=1\\m\neq k}}^{K} D^{km}_{j\ \ in}}$$

Figure C.7: The graphical model of the proposed Correlated Fusion Model for $K > 2$ classifiers. Note that $\delta^{km}_{j\ \ l} = \delta^{mk}_{j\ \ l}$ and equivalently $D^{km}_{j\ \ il} = D^{mk}_{j\ \ il}$.

The conditional distributions required for Gibbs Sampling for a single correlated Dirichlet distribution over $K = 2$ J-dimensional random variables $\boldsymbol{x^1}$ and $\boldsymbol{x^2}$ can be obtained with a change of variables with $\boldsymbol{x^k_l}$ defined as in (4), $\boldsymbol{a^k_j} = \boldsymbol{\alpha^k_j} - \boldsymbol{\delta_j}, \boldsymbol{c^k} = \sum_{l=1}^{J} \boldsymbol{A^k_{j\,l}}, \boldsymbol{z_l} = \boldsymbol{D_l}$ for $l = 1, ..., J, k = 1, 2$:

$$\begin{aligned}\log P(a^k_{j\,l}|\Theta_{-a^k_{j\,l}}) \propto\ & (\xi - 1)\log(a^k_{j\,l}) - \xi a^k_{j\,l} - I_j \log(\Gamma(a^k_{j\,l})) \\ & + a^k_{j\,l} \sum_{i=1}^{I_j} \log\left(x^k_{i\,l}\left(c^k_i + \sum_{n=1}^{J} z_{ni}\right) - z_{li}\right), \quad k = 1, 2, l = 1, ..., J\end{aligned} \tag{C.61}$$

$$\begin{aligned}\log P(\delta_{jl}|\Theta_{-\delta_{jl}}) \propto\ & (\xi - 1)\log(\delta_{jl}) - \xi\delta_{jl} - I_j \log(\Gamma(\delta_{jl})) \\ & + \delta_{jl} \sum_{i=1}^{I_j} \log(z_{li}), \quad l = 1, ..., J\end{aligned} \tag{C.62}$$

$$\begin{aligned}\log P(\boldsymbol{c^k}|\Theta_{-\boldsymbol{c^k}}) \propto\ & \sum_{i=1}^{I_j}\Bigg[(J-1)\log\left(c^k_i + \sum_{n=1}^{J} z_{ni}\right) \\ & + \sum_{n=1}^{J}\left[(a^k_{j\,n} - 1)\log\left(x^k_{i\,n}\left(c^k_i + \sum_{m=1}^{J} z_{mi}\right) - z_{ni}\right)\right] - c^k_i\Bigg] \quad k = 1, 2\end{aligned} \tag{C.63}$$

$$\begin{aligned}\log P(\boldsymbol{z_l}|\Theta_{-\boldsymbol{z_l}}) \propto\ & \sum_{i=1}^{I_j}\Bigg[(J-1)\sum_{k=1}^{2}\log\left(c^k_i + \sum_{n=1}^{J} z_{ni}\right) \\ & + \sum_{k=1}^{2}\left[\sum_{n=1}^{J}\left[(a^k_{j\,n} - 1)\log\left(x^k_{i\,n}\left(c^k_i + \sum_{m=1}^{J} z_{mi}\right) - z_{ni}\right)\right]\right] \\ & + (\delta_{jl} - 1)\log(z_{li}) - z_{li}\Bigg], \quad l = 1, ..., J\end{aligned} \tag{C.64}$$

with $\Theta = \{\boldsymbol{x^1}, \boldsymbol{x^2}, a^1_{j\,1}, ..., a^1_{j\,J}, a^2_{j\,1}, ..., a^2_{j\,J}, \delta_{j\,1}, ..., \delta_{j\,J}, \boldsymbol{c^1}, \boldsymbol{c^2}, \boldsymbol{z_1}, ..., \boldsymbol{z_J}\}$. Note that for mathematical simplicity we sample from $\boldsymbol{a^k_j} = \boldsymbol{\alpha^k_j} - \boldsymbol{\delta_j}$ instead of directly sampling from $\boldsymbol{\alpha^k_j}$. Therefore, after sampling, the marginal parameters $\boldsymbol{\alpha^k_j}$ must be calculated from $\boldsymbol{a^k_j}$.

With the conditional distributions derived above a Gibbs Sampler can be implemented that can be used to infer the parameters of J correlated Dirichlet distributions in the CFM if we consider $K = 2$ classifiers. Again, note that for each correlated Dirichlet distribution to be inferred we only consider the I_j categorical classifier outputs $\boldsymbol{x^1_i}, \boldsymbol{x^2_i}$ of all examples for which $t_i = j$.

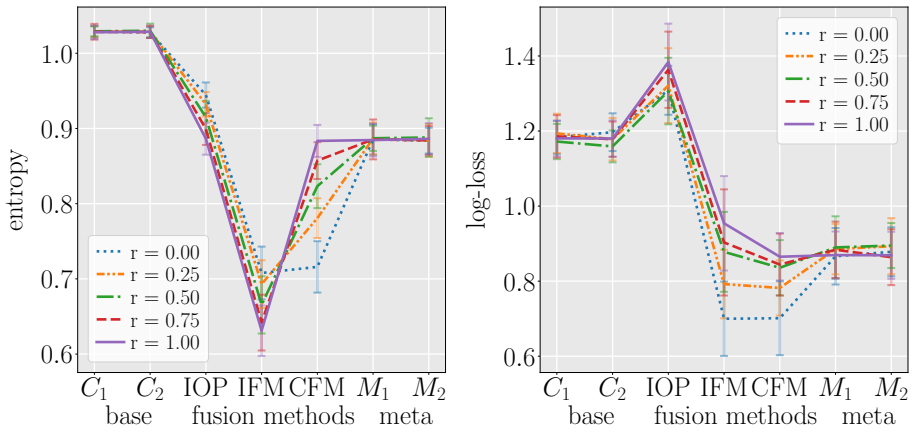


Figure D.8: Fusion performances on simulated data of two biased classifiers in terms of mean entropy and log-loss. We compare the performance of base classifiers C_1, C_2, the three fusion methods IOP, IFM, and CFM, and the meta classifiers M_1, M_2 for five levels of correlation between the base classifiers. The marginal parameters of the two biased classifiers are $\boldsymbol{\alpha^1} = \boldsymbol{\alpha^2} = ((7,5,5),(5,5,7),(5,7,5))$. Standard deviations are shown as error bars.

However, it is much more efficient to implement inference using standard inference tools such as JAGS (Plummer et al., 2003), which allow sampling given a definition of the generative model of the CFM (as shown in Figure 1(c) and C.7). Thus, for efficiency reasons, as well for parameter inference as for inferring the fused distribution $P(t_i|\boldsymbol{x_i^1},...,\boldsymbol{x_i^K},\boldsymbol{\alpha},\boldsymbol{\delta})$ we resort to Gibbs Sampling using JAGS. Since $x_{i\,l}^k$ 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 $x_{i\,l}^{k\,*}$ is normally distributed with $x_{i\,l}^{k\,*} \sim \mathcal{N}(x_{i\,l}^k, \epsilon)$ and $\epsilon = 10^{-4}$.

D MORE DETAILS ON EVALUATION

Here, we provide additional information on the evaluations presented in Section 4 of the paper and show further evaluations on additional data sets. In Section D.1, we evaluate the CFM on an additional simulated data set consisting of biased base classifiers. Section D.2 provides detailed information on the real data sets in Section 4.2. In Section D.3 we additionally evaluate the CFM on a real data set consisting of the output distributions of $K = 3$ classifiers, and in Section D.4 we provide details on the comparison of the CFM and the approach of Pirs and Strumbelj (2019) and additionally compare them in terms of required time for fusion. Finally, in Section D.5, we provide the CFM's parameters used for generating the simulated data sets and inferred for the real data sets.

D.1 Additional Evaluations on Simulated Biased Classifiers

In Section 4.1, we show the normative fusion behavior of the CFM on two simulated data sets. The first data set was generated with marginal parameters leading to IOP fusion for zero correlation, the second one leads to higher uncertainty reduction through fusion due to decreased classifier variance and uncertainty. Since the CFM also considers potential biases of classifiers for fusion, here we additionally show the respective fusion performances in terms of entropy and log-loss for two biased base classifiers, which on average predict class 3 if $t_i = 2$ and vice versa. For the resulting third simulated data set (SIM 3), in Figure D.8 we see similar results as for the other simulated data sets in Figure 3(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, 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.

D.2 Detailed Information on the Real Data Sets in Section 4.2

As stated in Section 4.2, we evaluated the CFM on 5 real data sets, Bookies A, Bookies B, DNA A, DNA B, DNA C. In the following we give detailed information on these data sets.

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[4] 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.

Bookies B consists of the predictions of two bookmakers (B365 and BW) for matches of the German Bundesliga[5] 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[6], 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 two 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.

D.3 Additional Evaluations on a Real Data Set Consisting of $K = 3$ Classifiers' Outputs

In Section 4.2 we evaluated the CFM on five real data sets. Since all of these data sets consist of the output distributions of only $K = 2$ classifiers, here we additionally evaluate the CFM on Bookies C, a data set consisting of $K = 3$ classifiers. 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.

Figure D.9 shows the same behavior as for the data sets Bookies A and B in Figure 4. The CFM's performance is equal to the performance of all three meta classifiers. Thus, also when fusing three highly correlated classifiers

[4] https://www.football-data.co.uk/englandm.php
[5] https://www.football-data.co.uk/germanym.php
[6] https://archive.ics.uci.edu/ml/datasets/Molecular+Biology+(Splice-junction+Gene+Sequences)

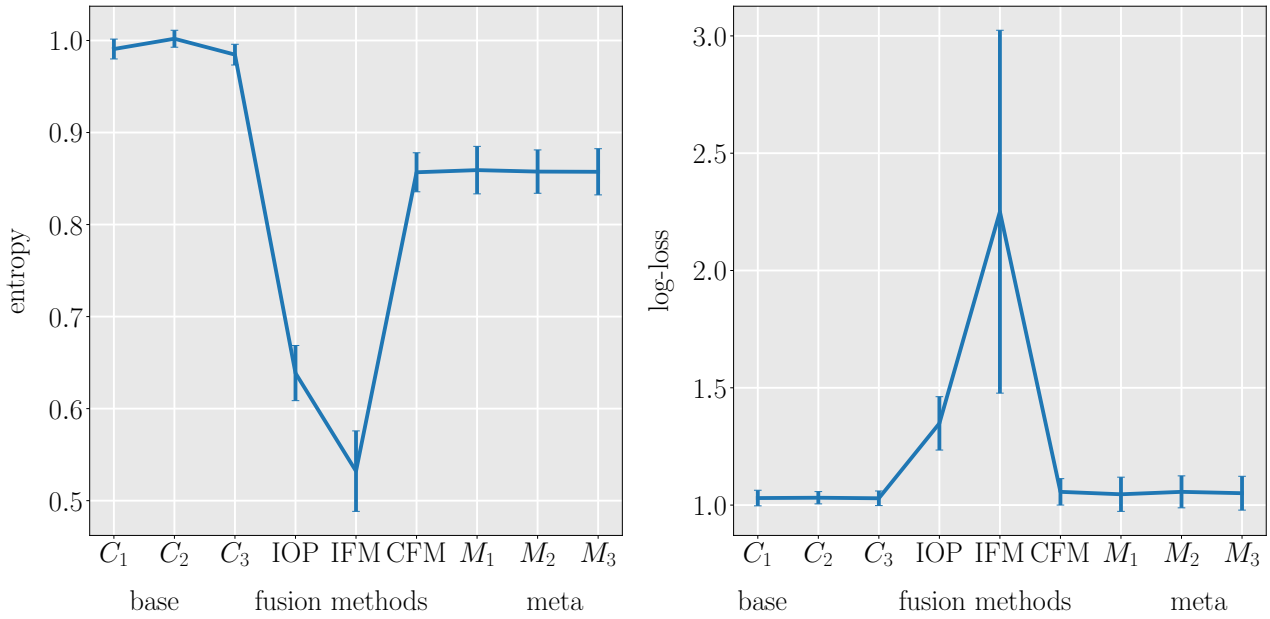


Figure D.9: Fusion performances on the additional real data set Bookies C in terms of mean entropy and log-loss. We compare the performance of base classifiers C_1, C_2, C_3, the three fusion methods IOP, IFM, and CFM, and the meta classifiers M_1, M_2, and M_3. Standard deviations are shown as error bars.

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.

D.4 Comparison to the Approach of Pirs and Strumbelj

In Section 4.3 we compare the performance in terms of log-loss of the proposed CFM to the performance of the only comparable model introduced by Pirs and Strumbelj (2019). For this, we use the code[7] provided by Pirs and Strumbelj (2019) and apply it to our data sets.

For the simulated data sets, we compare the Bayes optimal fusion performance of the CFM to the performance of the model proposed by Pirs and Strumbelj (2019). For each simulated data set and correlation level, we generated training sets consisting of 1500 examples (500 per class label) according to the generative model of the CFM for learning the model parameters for Pirs' model, fused the categorical distributions of the simulated data sets described in Section 4.1, and compared the fusion performances to the ones of the CFM shown in Figure 3. The results in Table D.2 show that the CFM outperforms Pirs' model on all simulated data sets.

Note that Table D.2 not only shows the results given in Table 1 in the paper, but additionally shows the model performances on SIM 1 $_{r=0.25}$, SIM 1 $_{r=0.75}$, SIM 2 $_{r=0.25}$, SIM 2 $_{r=0.75}$, which were left out for brevity in the paper, and the model performances on the SIM 3 data set, we additionally included in the Supplementary Material in Section D.1.

Also for the five real data sets in Section 4.2, Bookies A, Bookies B, DNA A, DNA B, DNA C, and the additional data set Bookies C, which consists of the outputs of three instead of two classifiers (Section D.3), we compared the performances of the CFM and Pirs' model in terms of log-loss. We trained Pirs' model on the same training sets we used for inferring the CFM's parameters and fused the distributions in the respective test sets accordingly. As can be seen in Table D.2, also on all tested real data sets, the CFM outperforms Pirs and Strumbelj (2019).

[7] https://github.com/gregorp90/MM

Table D.2: 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 and the six real data sets Bookies A, Bookies B, DNA A, DNA B, DNA C, Bookies C.

data set	CFM ($\mu \pm \sigma$)	Pirs ($\mu \pm \sigma$)
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

D.4.1 Comparison of Required Time for Fusion

As we discuss in Section 5, one 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 to 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 samples each. Fusing the same test split with the model by Pirs and Strumbelj 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 is not in the scope of this paper but left for future work. 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 can be selected with the risk of incorrect fusion and performance losses.

D.5 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 D.5.1) and that were inferred for the real data sets (Section D.5.2).

D.5.1 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 D.3 for the first simulated data set (SIM 1), Table D.4 for the second simulated data set (SIM 2), and Table D.5 for the third simulated data set (SIM 3), which was not shown in the paper but additionally in Section D.1. Note that the shown correlations can only be generated approximately with the presented parameters.

D.5.2 Parameters for the Real Data Sets

The parameters of the Correlated Fusion Model that we inferred for the five real data sets Bookies A, Bookies B, DNA A, DNA B, DNA C are presented in Table D.6.

For the three data sets Bookies A, Bookies B, 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 D.7 shows the parameters of the additional real data set Bookies C, which consists of the predictions of three 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 D.3: 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 D.4: 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 $r \approx 1$. $\boldsymbol{\alpha^1}$ 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	$\boldsymbol{\alpha^1}$	$\boldsymbol{\alpha^2}$	$\boldsymbol{\delta}$
$r \approx 0.0$	$\begin{bmatrix} 12 & 8 & 8 \\ 8 & 12 & 8 \\ 8 & 8 & 12 \end{bmatrix}$	$\begin{bmatrix} 12 & 8 & 8 \\ 8 & 12 & 8 \\ 8 & 8 & 12 \end{bmatrix}$	$\begin{bmatrix} 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 \end{bmatrix}$
$r \approx 0.25$	$\begin{bmatrix} 12 & 8 & 8 \\ 8 & 12 & 8 \\ 8 & 8 & 12 \end{bmatrix}$	$\begin{bmatrix} 12 & 8 & 8 \\ 8 & 12 & 8 \\ 8 & 8 & 12 \end{bmatrix}$	$\begin{bmatrix} 3 & 2 & 2 \\ 2 & 3 & 2 \\ 2 & 2 & 3 \end{bmatrix}$
$r \approx 0.5$	$\begin{bmatrix} 12 & 8 & 8 \\ 8 & 12 & 8 \\ 8 & 8 & 12 \end{bmatrix}$	$\begin{bmatrix} 12 & 8 & 8 \\ 8 & 12 & 8 \\ 8 & 8 & 12 \end{bmatrix}$	$\begin{bmatrix} 6 & 4 & 4 \\ 4 & 6 & 4 \\ 4 & 4 & 6 \end{bmatrix}$
$r \approx 0.75$	$\begin{bmatrix} 12 & 8 & 8 \\ 8 & 12 & 8 \\ 8 & 8 & 12 \end{bmatrix}$	$\begin{bmatrix} 12 & 8 & 8 \\ 8 & 12 & 8 \\ 8 & 8 & 12 \end{bmatrix}$	$\begin{bmatrix} 9 & 6 & 6 \\ 6 & 9 & 6 \\ 6 & 6 & 9 \end{bmatrix}$
$r \approx 1.0$	$\begin{bmatrix} 12 & 8 & 8 \\ 8 & 12 & 8 \\ 8 & 8 & 12 \end{bmatrix}$	$\begin{bmatrix} 12 & 8 & 8 \\ 8 & 12 & 8 \\ 8 & 8 & 12 \end{bmatrix}$	$\begin{bmatrix} 11.9 & 7.9 & 7.9 \\ 7.9 & 11.9 & 7.9 \\ 7.9 & 7.9 & 11.9 \end{bmatrix}$

Table D.5: Model parameters of the Correlated Fusion Model that we used to generate the third simulated data set (SIM 3) presented in D.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	$\boldsymbol{\alpha^1}$	$\boldsymbol{\alpha^2}$	$\boldsymbol{\delta}$
$r \approx 0.0$	$\begin{bmatrix} 7 & 5 & 5 \\ 5 & 5 & 7 \\ 5 & 7 & 5 \end{bmatrix}$	$\begin{bmatrix} 7 & 5 & 5 \\ 5 & 5 & 7 \\ 5 & 7 & 5 \end{bmatrix}$	$\begin{bmatrix} 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 \end{bmatrix}$
$r \approx 0.25$	$\begin{bmatrix} 7 & 5 & 5 \\ 5 & 5 & 7 \\ 5 & 7 & 5 \end{bmatrix}$	$\begin{bmatrix} 7 & 5 & 5 \\ 5 & 5 & 7 \\ 5 & 7 & 5 \end{bmatrix}$	$\begin{bmatrix} 1.75 & 1.25 & 1.25 \\ 1.25 & 1.25 & 1.75 \\ 1.25 & 1.75 & 1.25 \end{bmatrix}$
$r \approx 0.5$	$\begin{bmatrix} 7 & 5 & 5 \\ 5 & 5 & 7 \\ 5 & 7 & 5 \end{bmatrix}$	$\begin{bmatrix} 7 & 5 & 5 \\ 5 & 5 & 7 \\ 5 & 7 & 5 \end{bmatrix}$	$\begin{bmatrix} 3.5 & 2.5 & 2.5 \\ 2.5 & 2.5 & 3.5 \\ 2.5 & 3.5 & 2.5 \end{bmatrix}$
$r \approx 0.75$	$\begin{bmatrix} 7 & 5 & 5 \\ 5 & 5 & 7 \\ 5 & 7 & 5 \end{bmatrix}$	$\begin{bmatrix} 7 & 5 & 5 \\ 5 & 5 & 7 \\ 5 & 7 & 5 \end{bmatrix}$	$\begin{bmatrix} 5.25 & 3.75 & 3.75 \\ 3.75 & 3.75 & 5.25 \\ 3.75 & 5.25 & 3.75 \end{bmatrix}$
$r \approx 1.0$	$\begin{bmatrix} 7 & 5 & 5 \\ 5 & 5 & 7 \\ 5 & 7 & 5 \end{bmatrix}$	$\begin{bmatrix} 7 & 5 & 5 \\ 5 & 5 & 7 \\ 5 & 7 & 5 \end{bmatrix}$	$\begin{bmatrix} 6.9 & 4.9 & 4.9 \\ 4.9 & 4.9 & 6.9 \\ 4.9 & 6.9 & 4.9 \end{bmatrix}$

Table D.6: Model parameters of the Correlated Fusion Model that we inferred for the real data sets in Section 4.2. $\boldsymbol{\alpha^1}$ 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 $t_i = j$. 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 data sets.

data set	$\boldsymbol{\alpha^1}$	$\boldsymbol{\alpha^2}$	$\boldsymbol{\delta}$
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 D.7: Model parameters of the Correlated Fusion Model that we inferred for the additional real data set 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 for all data sets.

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