Pairwise local Fisher and naive Bayes: Improving two standard discriminants*

Håkon Otneim
Norwegian School of Economics
Martin Jullum
Norwegian Computing Center
Dag Tjøstheim†
University of Bergen and Norwegian Computing Center

Abstract

The Fisher discriminant is probably the best known likelihood discriminant for continuous data. Another benchmark discriminant is the naive Bayes, which is based on marginals only. In this paper we extend both discriminants by modeling dependence between pairs of variables. In the continuous case this is done by local Gaussian versions of the Fisher discriminant. In the discrete case the naive Bayes is extended by taking geometric averages of pairwise joint probabilities. We also indicate how the two approaches can be combined for mixed continuous and discrete data. The new discriminants show promising results in a number of simulation experiments and real data illustrations.

Classification codes: C14, C44, C65
Keywords: Kernel estimator, local Fisher discriminant, naive Bayes discriminant, pairwise dependence, local Gaussian density

1 Introduction

The statistical classification problem consists in allocating observed data samples to one of several possible classes based on information obtained from a set of observations having known class membership. Two standard classifiers are the Fisher discriminant (**?**) and the naive Bayes discriminant (**?**, p. 210-211). These are easy to understand and to apply, and have been much used in practice. The Fisher discriminant assumes that each class is multivariate normally distributed, while the naive Bayes is based on the assumption of independent variables, so that multivariate class distributions are replaced by the product of its marginal distributions. The Fisher discriminant requires continuous data, whereas the naive Bayes works both for continuous and discrete data. For both methods, Bayes' formula is typically used to obtain class probabilities.

The Fisher and Naive Bayes classification rules have some obvious problems though: They can not separate between classes that differ in their dependence structure beyond independence and the second moments, respectively. In this paper we seek to rectify this by presenting novel discrimination procedures generalizing these basic classification methods. For continuous data we replace the standard Fisher classifier by a local Fisher discriminant, that uses locally normal approximations of the class distributions. The local approximation has a pairwise dependence structure and is constructed such that, in the limit experiment, our discriminant coincides with the standard Fisher discriminant if the class distributions are, in fact, multinormal. For discrete data, we generalize the naive Bayes classifier by replacing the product of marginal distributions within each class by a type of geometric mean of pairwise distributions, which again reduces to the naive Bayes in case of independence. We believe that this pairwise representation of a joint discrete probability is both novel and useful. It is derived by arguments that are very different from the continuous local Gaussian representation, and we think that its applicability is not limited to classification.

*This work was supported by The Norwegian Research Council through the Big Insight Center for research-driven innovation (grant number 237718).

†Corresponding author. Department of Mathematics, University of Bergen, P.B. 7803, 5020 Bergen, Norway. dag.tjostheim@uib.no

For situations with both continuous and discrete data present, we incorporate the dependence between the data types by first modeling the continuous variable with the local Gaussian distributions. Then the pairs of discrete variables are modeled conditionally on the continuous variables with a logistic regression type procedure. Thus, our paper aims at generalizing the Fisher and Naive Bayes classifiers in all these three, equally important, data situations.

1.1 Background

Let us first provide some background for the classification[1] problem. The K-class discrimination problem consists in assigning the d-dimensional data vector $X = (X_1, \ldots, X_d)$ to one of K classes. Examples ranges from fraud detection, authorship and text analysis, spam-email detection, credit rating, bankruptcy prediction and even seismic discrimination (see e.g. **?**, **?**, **?**, **?**, **?**, **?**, **?**, and **?**). Usually (in supervised learning) a training data set is available. Each training set consists of data X from a known class that we use to get an idea of the stochastic features within each class, and that we again describe by the class-wise probability distribution functions f_k, $k = 1, \ldots, K$, hereafter referred to as class distributions. These distributions may be continuous, discrete or mixed. In Sections **??** - **??**, f_k will be a density function, whereas we look at the discrete and mixed cases in Sections **??** - **??**. We may also have available an (unconditional) prior probability $\pi_k = P(\text{class}(X) = k)$ for each class, or at least such a probability can be estimated from the training data.

Let D be a decision variable that takes the values $1, \ldots, K$. Let us also write $f = (f_1, \ldots, f_K)$, and $\pi = (\pi_1, \ldots, \pi_K)$. On the basis of a new sample X and the available training data, one must determine the value of D in an optimal way. Optimality is usually obtained by minimizing the so-called Bayes risk. Assuming that f_k and $\pi_k = P(D = k)$ are known for all k, we obtain the posterior probability of having $D = k$ using Bayes' Theorem:

$$P_f(D = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{j=1}^{K} \pi_j f_j(x)}. \tag{1}$$

Now assign a loss function $L(k, j)$ which gives the loss of assigning x to k, when in fact $D = j$. The Bayes risk is defined as the expected loss with respect to the posterior probabilities:

$$R_f(k, x, \pi) = \sum_j L(k, j) P_f(D = j|X = x). \tag{2}$$

The classification rule D_B, which is Bayes optimal with respect to R_f, then follows by minimizing R_f over k, or in other words, D_B is given by

$$D_B(x, \pi) = \underset{k=1,\ldots,K}{\arg\min}\, R_f(k, x, \pi). \tag{3}$$

In the particular case of a 0-1 loss ($L(k, j) = 1(k \neq j)$ where $1(\cdot)$ is the indicator function), it is easy to compute the Bayes rule, since the decision rule takes the simple form

$$D_B(x, \pi) = \underset{k=1,\ldots,K}{\arg\max}\, P_f(D = k|X = x) = \underset{k=1,\ldots,K}{\arg\max}\, \pi_k f_k(x). \tag{4}$$

This is a small but vital part of Bayesian decision theory and Bayesian inference whose foundations are explored in the classic text of **?**. The expression (**??**) forms the «intuitive» solution to the classification problem, and we shall rely on this decision rule throughout the paper. Note, however, that the methodology we develop and the comparisons we perform, are equally valid with decision rules originating from other loss functions. In the practical situation when f (and π) are not known, these need to be estimated from data in order to reach a decision. When π is unknown it may typically be estimated by the relative class-wise frequencies observed in the training data: $\widehat{\pi}_k = n_k/n$, where n is the total number of observations, and n_k is the number of observations belonging to class k. The estimation of $f_k, k = 1, \ldots, K$, may typically be done in a number of different ways, and it is this choice of estimation method that essentially distinguishes

[1] We will use the terms discrimination and classification interchangeably throughout this paper, referring to the same concept.

different classification methods from each other. The remaining part of the paper shall therefore, to a large extent, be concerned with methods for estimating $f_k, k = 1, \ldots, K$ in the continuous, discrete and mixed continuous/discrete cases, and the comparison of these, in the discrimination context of (**??**). In many situations there are only two classes, $K = 2$. Although all presented methodology works for general K, we will for simplicity concentrate on the $K = 2$ case in the illustrations considered in the present paper.

1.2 Estimating discriminants

If the f_ks are continuous, one may assume that they belong to a particular parametric family of densities. The estimation problem then consists in estimating the parameters of that parametric density. The classic Fisher discriminant originates from the work by **?**, who assumes that the d-variate data from each class k are normally distributed, written $\mathcal{N}(\mu_k, \Sigma_k)$, where the μ_k and Σ_k are class-wise mean vectors and covariance matrices, respectively; i.e.,

$$f_k(x) = \frac{1}{(2\pi)^{d/2}|\Sigma_k|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k)\right),$$

where $|\cdot|$ denotes the determinant and T the transposed. If we assume $\Sigma_k = \Sigma$ for all k, the Bayes rule in (**??**) takes the form (**?**, Chapter 11.3)

$$\widehat{D}_{\text{LDA}}(x) = \underset{k=1,\ldots,K}{\arg\max}\, x^T\widehat{\Sigma}^{-1}\widehat{\mu}_k - \frac{1}{2}\widehat{\mu}_k^T\widehat{\Sigma}^{-1}\widehat{\mu}_k + \log\widehat{\pi}_k,$$

where the $\widehat{\mu}_k$ are the class-wise empirical mean vectors and $\widehat{\Sigma}$ is the common empirical covariance matrix, respectively, that we calculate using training data. This particular classification rule is called *linear discriminant analysis* (LDA) because the estimated decision boundaries between classes are linear in x and thus forms hyper planes in the d-dimensional Euclidean space. The general case where we allow the covariance matrices Σ_k to be different within each class, leads to the classification rule

$$\widehat{D}_{\text{QDA}}(x) = \underset{k=1,\ldots,K}{\arg\max} -\frac{1}{2}x^T\widehat{\Sigma}_k^{-1}x + x^T\widehat{\Sigma}_k^{-1}\widehat{\mu}_k - \frac{1}{2}\widehat{\mu}_k^T\widehat{\Sigma}_k^{-1}\widehat{\mu}_k - \frac{1}{2}\log|\widehat{\Sigma}_k| + \log\widehat{\pi}_k, \tag{5}$$

which is termed *quadratic discriminant analysis* (QDA) due to the quadratic term in (**??**), causing a second order (quadratic) decision boundary.

One advantage of the Fisher discriminant is that f_k is easy to estimate also for quite a large d, since for each k the estimation reduces to *marginal* estimates of means $\mu_{j,k}$, $j = 1, \ldots, d$ and *pairwise* estimates of covariances $\Sigma_{jl,k}$, $j, l = 1, \ldots, d$. This corresponds to pairwise dependencies between components. A general d-dimensional density does not have this property, such that dependence between any two variables may not be so easily extracted from the joint distribution. Despite, or perhaps due to, their simplicity, QDA and LDA have a proven track record in many situations where the class distributions are clearly non-normal (**?**).

It is, however, crucially important to note here that the QDA and LDA discriminants do not see any difference between populations having equal mean vectors and covariance matrices, even though the populations may be radically different in terms of nonlinear dependence. In that case, we should rather resort to a method that does allow for non-linear dependence or more flexibility in terms of the distributional form of $f_k(x)$. While it might be most natural to handle such situations by a method that has both these properties, one may also consider a method with only one of the properties. Naive Bayes is a well known references discriminant of this type, allowing for more flexibility for the marginal distributions, but completely ignoring any dependence between the variables X_j and X_l. Naive Bayes, which works both for discrete and continuous data, takes the form

$$P_f(D = k|X = x) = \prod_{j=1}^{d} P_{f_{(j)}}(D = k|X_j = x_j), \tag{6}$$

where $f_{(j)}$ denotes the marginal distribution of X_j. This approximation may work surprisingly well even in situations where property (**??**) is not satisfied. The marginal distributions in (**??**) may be estimated

parametrically (for instance with a Gaussian distribution) as well as non-parametrically (for instance with a kernel density estimator), in both cases avoiding the curse of dimensionality. Note that naive Bayes may actually work in cases where the means and covariances of the populations are identical, i.e. cases where the Fisher discriminant cannot work. This is because it is possible to have different non-Gaussian marginal distributions where variances and means are the same. A simple example is when class 1 has Uniform$[-3, 3]$ marginals, while class 2 has $\mathcal{N}(0, 3^2)$ marginals. These distributions have the same mean and variances, but the distributions are still very different.

In the first part of this paper, where we focus on the continuous case, we construct generalizations of the QDA and naive Bayes that take general pairwise dependencies between pairs (X_j, X_l), into account, not just correlations (linear dependence), but also having the important property that they collapse to simpler forms if that indeed is optimal.

One alternative to choosing between the approximations described above is to pursue a fully nonparametric approach. Then f_k can be estimated for example using the kernel density estimator

$$\widehat{f}_{\text{kernel},\,k}(x) = \frac{1}{n_k} \sum_{i=1}^{n_k} K_{B_k}(X^{(k)}(i) - x),$$

where $\{X^{(k)}(i), i = 1, \ldots, n_k\}$ are observations in the training set of class k, and where $K_{B_k}(\cdot) = B_k^{-1} K(B_k^{-1}\cdot)$, with K being a kernel function, and B_k is a non-singular bandwidth parameter (matrix) for class k. When $n_k \to \infty$, then $\widehat{f}_k \to f_k$ under weak regularity conditions, but a considerable disadvantage is the curse of dimensionality. For d moderate or large, bigger than 3 or 4 say, the kernel estimator may not work well, see e.g. **?**, Chapter 4.5. This limits the potential usefulness of the kernel estimator in discrimination problems, where d may be quite big. In these situations the problem may be alleviated to some extent by a judicious choice of bandwidth. See in particular the work by **?** and **?**. Other nonparametric approaches are nearest neighbor classifiers, see e.g. **?** and classification using data depth (**?**), but the basic problem of the curse of dimensionality remains unless we accept the radical simplification provided by the naive Bayes with nonparametric margins.

The literature provides various other approaches to density estimation, such as the use of mixtures of a parametric and nonparametric approach that may reduce the consequences of the curse of dimensionality, see e.g. **?**. To a lesser degree this has also been the case in discrimination, see **?**, who basically choose a parametric approach, but allows a nonparametric perturbation similar to that of **?**. Another such method is the local likelihood estimator proposed by **?** and by **?**, who estimate $f_k(x)$ by fitting a whole family of parametric distributions, such that the parameter vector $\theta = \theta(x)$ is allowed to vary locally with x. We will pursue this idea in the first part of this paper by choosing the multivariate normal as the local approximant. This makes it possible to replace the pairwise correlations used by the Fisher discriminant with locally pairwise dependence functions directly in (**??**). An alternative, non-equivalent option, which we shall also visit, is to perform classification by inserting the class distributions obtained with the local (Gaussian) likelihood approach into (**??**). We will pursue both approaches. The local Gaussian approach has been recently used with success in a number of different contexts, see **?**, **?**, **?**, **?**, **?**, **?** and **?**. `R`-packages for computing local Gaussian quantities also exist, as described by **?** and **?**. We will in particular use the local Gaussian density estimation technique as presented by **?**, who show that the curse of dimensionality can be avoided, at least to a certain degree, by restricting the local correlations to pairwise dependence.

The local Gaussian discriminant is limited to the continuous case, but discrimination problems often involve discrete variables, or even mixtures of continuous and discrete variables. In the second, and equally important, part of the paper we consider discrete variables and mixtures of continuous and discrete variables. In that part we extend the idea of describing dependence by means of pairwise relationships to discrete variables, relying on geometric means of pairwise probabilities and successive conditioning, in a sense similar to the pair-copula construction described in **?**. For the case of mixed continuous and discrete variables, we first model the continuous variables with methodology described in the succeeding section. Then, conditioning on the continuous variables, the discrete variables is sought described by a link function and a logistic regression or a GAM type procedure. We will come back to this in Section **??** - **??**.

The rest of the paper is organized as follows: In Section **??** some aspects of local Gaussian density estimation are introduced. Asymptotics of the Bayes risk and bandwidth choice are presented, in particular in the context of local Gaussian discrimination, in Sections **??** and **??**. A number of illustrations in the continuous case are given in Section **??**. Section **??** and **??** deal with the purely discrete case and the mixed discrete-continuous case, respectively, with corresponding illustrations in Section **??**. Finally, in Section **??**, we present some conclusions and a brief discussion.

2 A local Gaussian Fisher discriminant

Considering the case with a continuous class distribution, let us now derive a local Fisher discriminant. We will start by introducing the local Gaussian approximation for a class distribution of a single class k. The idea of the local Gaussian approximation is to approximate $f_k(x)$ in a neighborhood N_x around x by a Gaussian density

$$\psi\Big(v, \mu_k(x), \Sigma_k(x)\Big) = (2\pi)^{-d/2}|\Sigma_k(x)|^{-1/2} \exp\Big\{(v-\mu_k(x))^T\Sigma_k^{-1}(x)(v-\mu_k(x))\Big\}, \tag{7}$$

where v is the running variable. The size of N_x is determined by a bandwidth parameter (matrix). In the bivariate case ($d=2$) with $x=(x_1,x_2)$ and with parameters $\theta_k(x) = (\mu_{k1}(x), \mu_{k2}(x), \sigma_{k1}(x), \sigma_{k2}(x), \rho_k(x))$, we write (**??**) as

$$\begin{aligned}\psi(v, \mu_{k1}(x), \mu_{k2}(x), \sigma_{k1}(x), \sigma_{k2}(x), \rho_k(x)) &= \frac{1}{2\pi\sigma_{k1}(x)\sigma_{k2}(x)\sqrt{1-\rho_k^2(x)}} \\ \times \exp\Bigg[-\frac{1}{2(1-\rho_k^2(x))}&\left(\frac{(v_1-\mu_{k1}(x))^2}{\sigma_{k1}^2(x)} - 2\rho_k(x)\frac{(v_1-\mu_{k1}(x))(v_2-\mu_{k2}(x))}{\sigma_{k1}(x)\sigma_{k2}(x)} + \frac{(v_2-\mu_{k2}(x))^2}{\sigma_{k2}^2(x)}\right)\Bigg].\end{aligned}$$

Moving from x to another point y, we use a possibly different Gaussian approximation $\psi(v, \mu_k(y), \Sigma_k(y)), v \in N_y$. The family of Gaussian distributions is especially attractive in practical use because of its exceptionally simple mathematical properties, which truly stands out in the theory of multivariate analysis. Our intention in this work is to exploit these properties *locally*. Note that the multivariate normal $\mathcal{N}(\mu_k, \Sigma_k)$ is a special case of the family of locally Gaussian distributions (**??**) with $\mu_k(x) \equiv \mu_k$ and $\Sigma_k(x) \equiv \Sigma_k$. **?** discuss non-trivial questions of existence and uniqueness. As the local parameter functions $\mu_k(x)$ and $\Sigma_k(x)$ take the place of the fixed parameters μ_k and Σ_k for each class distribution k in the Gaussian case, it is natural to extend the QDA of (**??**) by simply replacing μ_k and Σ_k by $\mu_k(x)$ and $\Sigma_k(x)$ for $k=1,\ldots,K$. This gives what we term the *local Fisher discriminant*

$$\begin{aligned}\widehat{D}_{\text{Local Fisher}}(x) = \arg\max_{k=1,\ldots,K} -\frac{1}{2}x^T\widehat{\Sigma}_k^{-1}(x)x + x^T\widehat{\Sigma}_k^{-1}(x)\widehat{\mu}_k(x) - \frac{1}{2}\widehat{\mu}_k(x)^T\widehat{\Sigma}_k^{-1}(x)\widehat{\mu}_k(x) \\ - \frac{1}{2}\log|\widehat{\Sigma}_k(x)| + \log\widehat{\pi}_k.\end{aligned} \tag{8}$$

To practically apply this procedure, we need estimates of the involved parameter functions for all class distributions $k=1,\ldots,K$. Following **?**, we estimate the parameters $\mu_k(x)$ and $\Sigma_k(x)$ given data $X^{(k)}(1),\ldots,X^{(k)}(n_k)$ with class label k, by maximizing the local log likelihood

$$L(X^{(k)}(1),\ldots,X^{(k)}(n_k),\theta_k(x)) = {n_k}^{-1}\sum_{i=1}^{n_k} K_{B_k}(X^{(k)}(i)-x)\log\psi(X^{(k)}(i),\theta_k(x)) - \int K_{B_k}(v-x)\psi(v,\theta_k(x))\,\mathrm{d}v, \tag{9}$$

where K_{B_k} is a kernel function depending on a bandwidth parameter (matrix) B_k. We refer to **?** and **?** for details on parameter estimation.

From the description of the local Gaussian likelihood above, the two discriminants in (**??**) and in (**??**) below appear to be highly affected by the curse of dimensionality. **?** suggest a particular simplification in order to relieve this effect, which we will adopt throughout the paper. The solution is to apply the following simplification

$$\mu_{j,k}(x) = \mu_{j,k}(x_j) \quad \text{and} \quad \Sigma_{jl,k}(x) = \Sigma_{jl,k}(x_j, x_l), \tag{10}$$

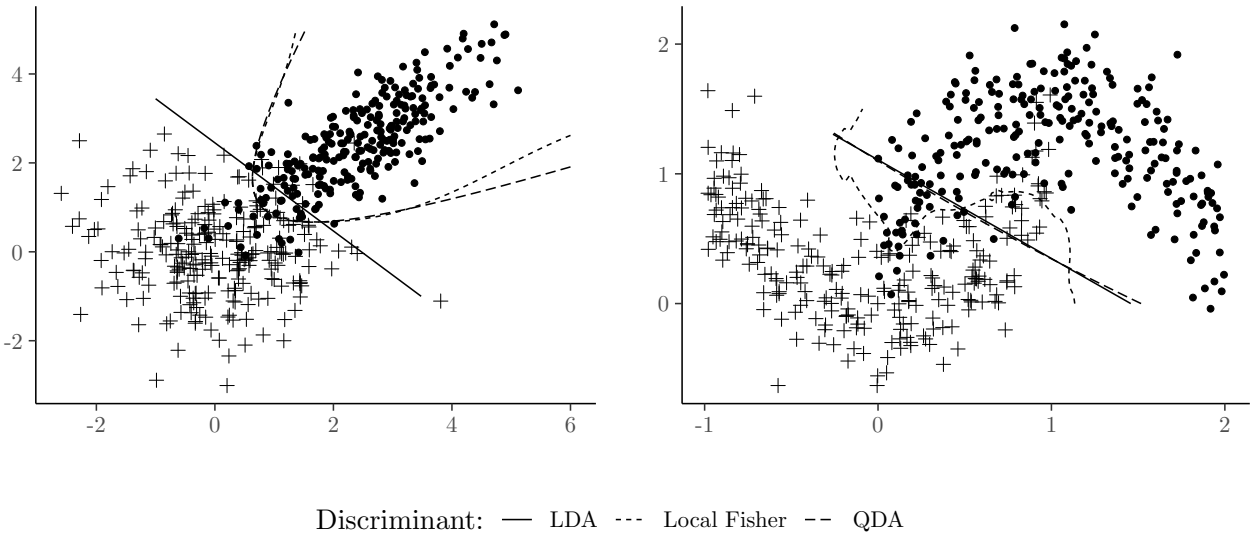

Figure 1: The two-class discrimination problem in two different cases.

leading to a pairwise local dependence structure. Examples can be found where this approximation is not at all valid, but the experience so far indicates that it covers a fairly wide set of circumstances; see e.g. **?**, **?** and the references therein. With this simplification it is possible to do a *pairwise* local dependence analysis in a multivariate non-Gaussian and nonlinear context, such as the local Fisher discriminant (**??**). This can be done such that, as $n_k \to \infty$, it reduces to the familiar pairwise correlation case if the true class distributions are indeed Gaussian. We illustrate this point graphically in Figure **??**.

In the left panel of Figure **??** we have plotted observations from two bivariate Gaussian populations, signified by "•" and "+", that have different mean vectors as well as different covariance matrices. In this case the LDA, being derived from the assumption of equal covariance matrices, is not optimal, as we appreciate from the plot where we have drawn the linear decision boundary as a solid line. The QDA, on the other hand, is in fact optimal because the parametric assumption of binormal populations having unequal covariance matrices is correct. The quadratic decision boundary is indicated by a dashed line. Furthermore, in this particular case, we observe that the local Fisher discriminant (**??**) essentially reduces to the the global QDA in (**??**), and we achieve precisely this by choosing a large bandwidth in the estimation of the local parameters in (**??**) using the local likelihood function in (**??**). The resulting decision boundary is displayed in the figure as a dotted line that for the most part coincides with the QDA boundary. It is important to note that the bandwidth selection in this illustration is completely data driven by means of a cross-validation procedure that we describe in Section **??**.

In the second panel of Figure **??** we have a different situation. The two populations are clearly not normally distributed, but their covariance matrices are equal (indeed, they are diagonal). This means that the QDA in practice collapses to the LDA, producing a near straight line. In this constructed example, though, we see immediately from the plot that a linear decision boundary is sub-optimal. In this case, our bandwidth selection algorithm that we present in Section **??** produces a small smoothing parameter, allowing the local Fisher discriminant (**??**) to become very local, non-linear and non-quadratic. This appears to work well for this discrimination problem.

As a by-product of the local likelihood setup and estimation procedure in (**??**), we approximate $f_k(x)$ by a family $\{\psi(v, \mu_k(x), \Sigma_k(x))\}$ of multivariate Gaussians, with estimates of the parameter functions $\widehat{\mu}_k(x)$ and $\widehat{\Sigma}_k(x)$:

$$\widehat{f}_{\mathrm{LGDE},\,k}(x) = \psi(x, \widehat{\mu}_k(x), \widehat{\Sigma}_k(x)). \tag{11}$$

These *locally Gaussian density estimates (LGDE)* (**?**) of the class distributions $f_k(x)$ give rise to a second option for utilizing the local Gaussian likelihood method in the discrimination setting. This option is to use $f_k(x), k = 1, \ldots, K$ directly to compute posterior probabilities and perform classification via (**??**) and (**??**),

respectively. This gives the following discriminant:

$$\widehat{D}_{\mathrm{LGDE}}(x) = \arg\max_{k=1,\ldots,K} \pi_k \widehat{f}_{\mathrm{LGDE},\,k}(x). \tag{12}$$

With the pairwise simplification described above, the estimate $\widehat{f}_{\mathrm{LGDE}}$ involves a further simplification resulting from transforming each variable to approximate standard normality, i.e., one can use the transformation $\widehat{Z}^{(k)}(j) = \Phi^{-1}\left(\widehat{F}_{n_k}(X^{(k)}(j))\right)$, $j = 1, \ldots, n_k$, of the marginals $X^{(k)}(j)$ in each population, where $\widehat{F}_{n_k}$ is the empirical distribution function and Φ the cumulative distribution function of the standard normal. Then, as a further simplification, we fix $\mu_k(z) \equiv 0$ and $\sigma_k(z) \equiv 1$, $k = 1, \ldots, d$. Alternatively, one could estimate $\mu_k(z)$ and $\sigma_k(z)$ by local likelihood, as has been done by **?** and **?**. This leads to larger flexibility and accuracy in the estimation, but at the cost of more complicated asymptotic analysis. The transformation procedure is especially attractive if the data contain extreme outliers.

Transforming back one obtains estimates $\widehat{f}_k(x)$. As it is not guaranteed that $\int \widehat{f}_k(x)dx = 1$ for a fixed n_k and bandwidth (matrix) B_k, the recipe also involves normalization of the f_k by a simple Monte Carlo procedure in the end. We do not normalize the locally Gaussian density estimates in this paper. Our experience is that the factor by which the density estimate $\widehat{f}_{\mathrm{LGDE}}$ departs from the unit integral mostly depends on the number of variables, and will thus not significantly affect the ratio $\widehat{f}_{\mathrm{LGDE},k}/\widehat{f}_{\mathrm{LGDE},j}$ for two classes k and j. Furthermore, as noted in Section **??**, we do not pursue precise density estimates as such in this paper, but rather tune our bandwidths to optimize discrimination performance. This can, in principle and in practice, be done regardless of whether the class-wise probability density estimates exactly integrate to one. In both constructed examples shown in Figure **??**, the LGDE based discriminant (**??**) is essentially identical to the local Fisher discriminant. This is not always the case though.

Asymptotic theory has been developed for the estimate $\widehat{f}_k(x) = \psi(x, \widehat{\mu}_k(x), \widehat{\Sigma}_k(x))$ as $n_k \to \infty$ and as the bandwidth (matrix) $B_k \to 0$. **?**, Theorems 3 and 4 demonstrate asymptotic normality and consistency under certain regularity conditions. In particular, $\widehat{f}_k(x) = \psi(x, \widehat{\mu}_k(x), \widehat{\Sigma}_k(x)) \to f_k(x)$ implies that $\int \widehat{f}_k(x)\,\mathrm{d}x \to 1$, which is relevant also for the asymptotic behavior of the Bayes risk.

3 Some asymptotics of Bayes risk

The Bayes risk in (**??**), as we have already seen, depends on density functions which may be estimated parametrically or nonparametrically. In the former case, assuming for simplicity $n_k = n$, $k = 1, \ldots, K$, this typically gives an asymptotic standard error of order $n^{-1/2}$, where n is the size of the training set. In the latter case, using kernel density estimation, assume the bandwidth matrix B_k is diagonal, $B_k = \mathrm{diag}\{b_{j,k}\}$ with $b_{j,k} = b_k$ for $j = 1, \ldots, d$. A kernel estimate of f_k has asymptotic standard error of order $(nb_k^d)^{-1/2}$, which is large if d is large. Due to the reduction to a pairwise structure, the locally Gaussian parameters discussed above, and thus the corresponding density estimate, has error of order $(nb_k^2)^{-1/2}$ irrespective of the dimension d. The full asymptotic distribution is given in Theorem 4 of **?**.

In discrimination, the asymptotics of the density estimates do not hold the main interest, but rather the asymptotics of the related Bayes risk. The purpose of the present section is to show that the local Gaussian discriminant has an asymptotic Bayes risk independent of d under weak regularity conditions. To do this we will base ourselves on **?**, who shows that a broad class of nonparametric density estimates (not restricted to kernel density estimates) achieves a mean square convergence rate of n^{-r} for some $0 < r < 1$.

To indicate how these results can be applied to locally Gaussian estimation, assume first that the class densities $f_1, \ldots, f_K$ are known. Recall from (**??**) that the Bayes rule takes the form $D_B = \arg\min_{k\in(1,\ldots,K)} R_f(k, x, \pi)$ for each x and π. However, in practice f is unknown, and has to be estimated. Estimating f by $\widehat{f} = (\widehat{f}_1, \ldots, \widehat{f}_K)$ leads to an estimate

$$\widehat{D}_n = \arg\min_{k\in(1,\ldots,K)} R_{\widehat{f}}(k, x, \pi)$$

of the Bayes rule, and we are interested in the asymptotic behavior of $\widehat{D}_n$ relative to D_B as n increases, both in terms of consistency as well as its rate of convergence. To this end we need some assumptions on the loss

function L introduced in (**??**) and the smoothness of f. The loss function L must satisfy

$$\max_k L(k,k) \leq \min_{k \neq j} L(k,j). \tag{13}$$

To define the mode of convergence, let C be a compact set $C \subset \mathbb{R}^d$, and let S_K be the simplex defined by $\sum_i \pi_i = 1$. **?** studies the mode of convergence of

$$\int_{S_K} \int_C \left| R_f(\widehat{D}_n, x, \pi) - R_f(D_B, x, \pi) \right| \mathrm{d}x \, \mathrm{d}\pi,$$

where we in fact do not need to take absolute value of the integrand since by definition, for every $x \in \mathbb{R}^d, k \in (1, \ldots, K)$,

$$R_f(k, x, \pi) \geq R_f\left(D_B, x, \pi\right).$$

Let further $\nabla_\alpha = \partial^{|\alpha|}/(\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d})$, $||x|| = (x_1^2 + \cdots + x_d^2)^{1/2}$ and $|\alpha| = \sum_{j=1}^d \alpha_j$. Then the following boundedness and smoothness assumptions are imposed on f. Let M_k be a constant $M_k > 1$, let m be a non-negative integer and $\beta \in (0, 1]$, and let $q = m + \beta$. We denote by $\mathcal{F}_k$ the class of probability densities f_k on $\mathbb{R}^d$ such that for all $k = 1, \ldots, K$,

(i) $f_k \leq M_k$ on $\mathbb{R}^d$.
(ii) $f_k \geq M_k^{-1}$ on C.
(iii) For all $x, y \in \mathbb{R}^d$, and all $|\alpha| = m$, we have

$$|\nabla_\alpha f_k(x) - \nabla_\alpha f_k(y)| \leq M_k ||x - y||^\beta.$$

As is well known, the smoothness of f_k determines the rate of convergence of $\widehat{f}_{n,k}$. More specifically, let $f_k \in \mathcal{F}_k$, then, according to **?**, Theorem 3, there is a constant $c > 0$ and a density estimator $\widehat{f}_{n,k}$ so that when $r = 2q/(2q + d)$,

$$\lim_{n \to \infty} \sup_{f_k \in \mathcal{F}_k} P_f \left[\int_C \left(\widehat{f}_{n,k}(x) - f_k(x) \right)^2 \mathrm{d}x > c n^{-r} \right] = 0. \tag{14}$$

Moreover, let $\mathcal{F}$ denote the K-fold Cartesian product of the $\mathcal{F}_k$, and T^n the set of training samples, each of size n. From **?**, Theorem 1, then there is a constant $c > 0$ and a classification rule $\widehat{D}_n(x, \pi, T^n)$ so that

$$\lim_{n \to \infty} \sup_{f \in \mathcal{F}} P_f \left[\int_{S_k} \int_C \left[R_f\left(\widehat{D}_n, x, \pi\right) - R_f\left(D_B, x, \pi\right) \right] \mathrm{d}x \, \mathrm{d}\pi > c n^{-r} \right] = 0. \tag{15}$$

The rate r in (**??**) describes the speed at which $\widehat{D}_n$ approaches the Bayes rule D_B. The rate turns out to be the same as for the density estimation rate for the class of densities in $\mathcal{F}_k$. In Theorem 2 of **?** it is shown that this rate is optimal in the sense that no better rate can be obtained for any classification rule $\widehat{D}_n$ based on density estimates $\widehat{f}_{n,k}$ of densities in $\mathcal{F}_k$.

It is easy to find density estimates that satisfy (**??**). If X is d-dimensional, and assuming existence of a bounded second derivative of f_k, the traditional kernel estimate has a variance of order $(nb_k^d)^{-1}$ and a bias of order b_k^2. Balancing the order of variance and bias squared; i.e., putting $(nb_k^d)^{-1} = b_k^4$ leads to $r = 4/(4 + d)$. Assuming existence of a bounded q-th order derivative of f_k and using higher order kernels; as in e.g. **?** leads to a bias of order h^q, whereas the order of the variance is unchanged. Again, equating the order of the variance and the bias squared leads to $r = 2q/(2q + d)$. By increasing q, it may seem like one may in the limit obtain the parametric rate of n^{-1} for the mean square error, but this is illusory as extremely large sample sizes would be required for the higher order asymptotics to kick in. In fact, as demonstrated by **?**, the practical usefulness of higher order kernels is debatable, and a realistic mean square convergence rate in practice is $n^{-4/(4+d)}$, which is a slow rate for d greater than 4, say.

The key of Marron's paper is that the derivation of (**??**) only uses the general convergence property in (**??**), the definition of R_f, and the general assumptions on L and f stated earlier in this section. This means

that it is not limited to kernel estimation, but can be applied to any density estimate that satisfies these requirements and has a rate as determined by (**??**). In turn this means that it can be applied to the locally Gaussian density estimator (LGDE, described in the preceding section) satisfying the regularity conditions of Theorem 4 of **?** and the additional mild conditions (**??**) and (i) - (iii) in this section. Note that the pairwise LGDE is defined irrespective of whether there actually is such a structure. In general it can serve as a computational approximation in the same way as an additive computational model can serve such a purpose in nonlinear regression.

Under the regularity assumptions stated in Theorem 4 by **?** it follows that the variance of the LGDE is of order $(nb^2)^{-1}$. From the log likelihood expression in (**??**) it is seen that by taking derivatives and using the weak law of large numbers, a local likelihood estimate of θ; would have to satisfy

$$0 = \frac{\partial L_n(\widehat{\theta}, x)}{\partial \theta_j} \xrightarrow{P} \int K_b(y-x) u_j(y, \theta_b) \Big\{ f(y) - \psi(y, \theta_b(y)) \Big\} \, \mathrm{d}y \tag{16}$$

where $u_j(\cdot, \theta) = \partial / \partial \theta_j \log \psi(\cdot, \theta)$. By Taylor expanding this integral we see that the difference between between $f(y)$ and $\psi(y, \theta_b)$ is of order b^2 as $b \to 0$. This means that $\psi(\theta_b)$ approximates f at this rate, and it is in fact the reason for including the last term in the log likelihood in (**??**). Contemplating that we obtain the estimates of $\widehat{\theta}$ by setting the log likelihood equal to zero, it is not difficult to see that the bias of the LGDE is of order b^2, see also **?**. Combining this with the expression for the order of the variance of the LGDE and equating bias squared and variance, this leads to $b = n^{-1/6}$ and $r = -2/3$, and this would lead to a rate of the mean square risk of $n^{-2/3}$ which is much better than the risk rate for the kernel estimator as d increases.

However, **?** used the log-spline approach to density estimation, see **?** and **?**. Its convergence rate as applied to local Gaussian density estimation is explained in detail in Appendix A1 in **?**. The density estimate requires the added restriction on the bandwidth that $n^{1/2+\epsilon} b^2 \to 0$, where $\epsilon \in (0, 1/2)$ is a design parameter having to do with the density of knots in the spline approximation (ϵ close to 0 means that new knots are added very fast, whereas ϵ close to $1/2$ means a slower rate). If the limit theorem should be valid over the entire ϵ-range, this implies the added condition $n^{1/2} b^2 \to 0$ (see condition (iv) of Theorem 4 in **?**), leading to a non-sharp convergence rate of the Bayes risk with $b = n^{-(1/2+\epsilon)}$, where $\epsilon \in (0, 1/2)$. However, by taking the design parameter, which is user-controlled, to be in the range $1/6 < \epsilon < 1/2$, it is seen that no extra restriction on the bandwidth is required, leading to the mean square convergence rate of $n^{-r} = n^{-2/3}$ irrespective of the dimension d. We also remark that condition (iii) of Theorem 2 in **?** implies a mild tail behavior condition on f.

An alternative to the log-spline approach is obtained by taking as an estimate of the marginal cumulative distribution function the integral of the kernel density estimate, see e.g. **?** and **?**. The problem of the design parameter ϵ is then avoided. The marginal density must be sufficiently smooth to guarantee the existence of the derivative of $F^{-1}(x)$, and again in the pairwise local Gaussian case a mean square convergence rate of $n^{-2/3}$ is obtained.

To summarize, all this means that using pairwise local Gaussian density estimation (with Bayes risk convergence rate $n^{-2/3}$) instead of kernel estimation (with Bayes risk convergence rate $n^{-4/(4+d)}$) leads to improvements as d increases. We confirm this in the simulation experiments in Section **??** with d in the range $2 \leq d \leq 8$.

It is not difficult to check that a higher order kernel applied to (**??**) will reduce the bias in the same way as for ordinary kernel estimation. Moreover, since no moments of the kernel function enter into the calculation of the variance of the local Gaussian density estimate, the convergence rate of the variance is not influenced by this, and higher order kernels will lead to a convergence rate of $n^{-\frac{2q}{2q+2}}$ of the Bayes risk. We remain relatively skeptical to the practical significance of this result, though.

In practical error estimation in discrimination, the empirical error frequencies are used via the AUC and Brier measures, see Section **??**. If the population densities are known, error estimates and upper bounds can be obtained by integrating over the tails of the densities. This is related to the evaluation of Value-at-Risk (VaR) in finance, and it is well known that it is sensitive to misspecification of densities. Especially, if Gaussians are used when true densities are thick-tailed, very serious underestimation may occur. This is illustrated in Table

1 in **?** in a comparison between Gaussian, kernel, so-called NP-estimates (**?**) and local Gaussian estimates, the latter being a clear winner in this particular example.

4 Choice of bandwidth

The preceding section concerns asymptotic results as the size of the training sets grows to infinity. We proceed now to establish rules for selecting bandwidths in finite-sample situations, which is clearly a problem of greater practical interest.

Nonparametric and semiparametric density estimators must as a general rule be *tuned* in one way or the other, usually by fixing a set of hyper parameters. The development of optimal strategies to do just that has been a topic of great interest in nonparametric analysis over the last couple of decades. The kernel density estimator, in particular, is associated with many bandwidth selection algorithms, and results on optimal choice of bandwidth have been known for some time, see e.g. **?** for a fairly general cross-validation case. The locally Gaussian density estimator is much more recent, and has seen but a few results on bandwidth selection.

? suggest cross-validation as a viable strategy, that **?**, **?** and **?** apply with reasonable results. It clearly works best on data that has been transformed towards marginal standard normality, which is a strategy that was mentioned in Section **??**. The method is time consuming, however, and the plug-in estimator $b_n = cn^{-1/6}$ has been used as well, for which the value of c may be determined empirically. No optimality theory of bandwidth selection exists for local likelihood density estimation.

The purpose of most bandwidth routines is to obtain good estimates of a density function f. We must here ask the following basic question, however: Is it true that an optimal bandwidth algorithm developed for density estimation is still optimal in a discrimination context? In the discrimination problem one is more concerned with the local properties of f_k where these densities overlap rather than the overall quality of the estimate of f_k. There are in fact several indications that a density-optimal bandwidth may not be discrimination-optimal.

This issue has been examined in some special cases by **?**. They examine the misclassification probability as a function of the bandwidth in the case of two multivariate Gaussian populations of dimensions 1 - 6 , and they found that the density-optimal bandwidth performed much worse than a bandwidth optimized with respect to the discrimination error in the case of equal a priori probabilities $\pi_1 = \pi_2 = 0.5$. The latter bandwidth was much larger, and in fact the classification error was largely insensitive to the choice of the bandwidth when it exceeded a certain threshold, whereas the density-optimal bandwidth was far below this threshold. For unequal prior probabilities, $\pi_1 = 0.4$, $\pi_2 = 0.6$ they reported less clear results.

We are interested in obtaining the best possible discriminant, rather than the best possible density estimators for the different classes. We therefore rely on a cross-validation scheme which optimizes the bandwidth parameter (matrix) in terms of discrimination performance (**?**).

The area under the receiver operating characteristic (ROC) curve, or simply AUC, is a widely used *ranking-based* metric for measuring the quality of a probability based discrimination procedure (**?**). The AUC is constructed for two-class classification, but generalizations to $K > 2$ classes exist (**?**, Section 10), and may replace the AUC in the description below when $K > 2$. A classifier that has an AUC value equal to 0.5 in a balanced classification problem is equivalent to pure guesswork, while if AUC = 1, this enables perfect classification.

We have chosen to optimize the bandwidth parameter in terms of this metric in our cross-validation scheme. As a reasonable trade-off between stability and computational expense, we perform cross-validation with a single split into m separate sets, i.e. m-fold cross-validation (**?**). To reduce the search space for the cross-validation procedure we require the bandwidth matrix B_n to be diagonal, with all diagonal entries on the form $b_n = cn^{-1/6}$, as mentioned above. The precise metric we optimize over is the average of the AUCs computed for each fold separately. To summarize, we tune the c parameter in b_n for the locally Gaussian discriminants according to the following cross-validation procedure:

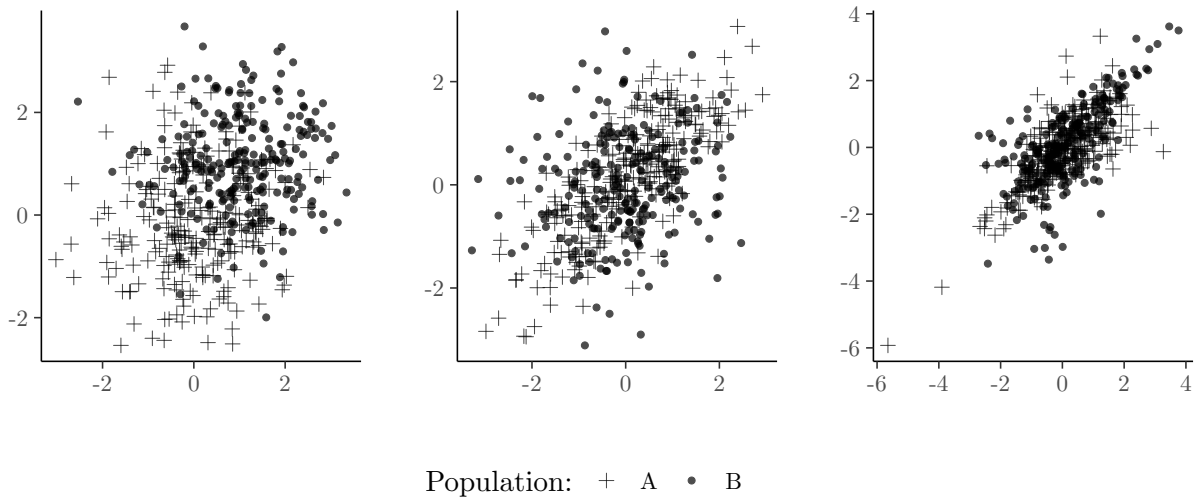

Figure 2: Data from the bivariate versions of the three simulated classification problems.

1. Divide the training set into m folds at random. We have used $m = 5$ in our experiments below.
2. For each proportionality constant c on a specified grid:
   (a) For each fold $j = 1, \ldots, m$:
      i. For each class $k = 1, \ldots, K$:
         A. Extract the variables corresponding to class k from all folds except fold j, and fit a local Gaussian density estimators with bandwidth matrix $B_n = \mathrm{diag}(cn^{-1/6})$.
         B. Use the fitted density to compute the out-of-fold estimated posterior probabilities $P_f(D = k|X = x)$ for all variable combinations x in fold j.
      ii. Compute the AUC in fold j using all the out-of-fold estimated $P_f(D = k|X = x)$'s and corresponding true classes, and denote it by $\mathrm{AUC}_j(B_n)$.
   (b) Compute the averaged AUC over all folds: $\overline{\mathrm{AUC}}(B_n) = (1/5)\sum_{j=1}^{5} \mathrm{AUC}_j(B_n)$
3. Choose the bandwidth matrix B_n with the largest $\overline{\mathrm{AUC}}(B_n)$.

In our illustrations in the following sections we also tune the non-parametric kernel estimators in the same way. Note further that, if there is a high degree of class imbalance in the training set, one may consider stratification when splitting the data into the m folds.

5 Illustrations

5.1 Simulations

Let us demonstrate some properties of the local Fisher discriminant (**??**) from a two-class simulation perspective. We generate data in increasing dimension d from three different multivariate classification problems that pose increasingly difficult conditions for the traditional discriminants:

- **Problem 1:** Two multivariate normal distributions, both having all correlations equal to zero and all standard deviations equal to one (so their covariance matrices are equal), but the first population has mean vector equal to $(0, \ldots, 0)^T$, while the second population has mean vector equal to $(1, \ldots, 1)^T$.

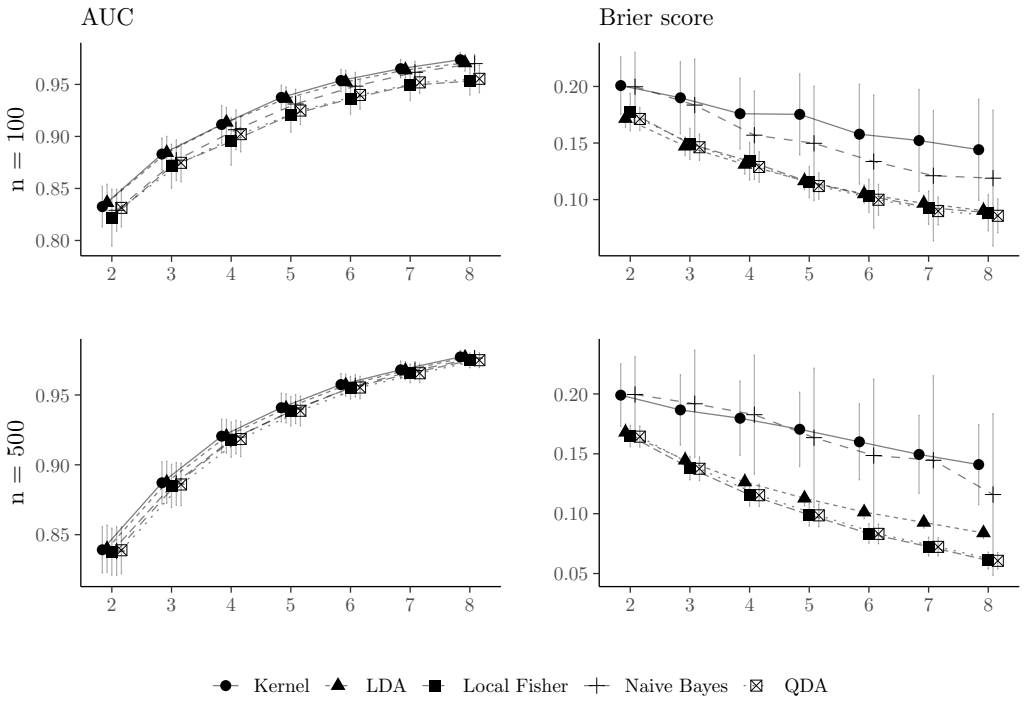

Figure 3: Simulation results for the first illustration: Two multinormal distributions with different means but equal covariance matrices. Error measured as a function of dimension.

- **Problem 2:** Two multivariate normal distributions having means and standard deviations equal to zero and one, respectively (so their marginal distributions are equal), but the first population has all correlations equal to 0.7 and the second population has all correlations equal to 0.2.
- **Problem 3:** The first population consists of observations on the stochastic vector X having $t(10)$-distributed marginals and a Clayton copula (**?**) with parameter $\theta = 2$. The second population consists of observations on $-X$.

We have plotted realizations with $n = 500$ of the bivariate versions of these problems in Figure **??**.

In all simulations we let $\pi_1 = \pi_2 = 0.5$. We measure classification performance in two standard ways. First, we use the AUC, as briefly introduced in Section **??**. In addition to the AUC we will also measure the *Brier score* of our predictions (**?**). The Brier score is essentially the mean squared error of a $0-1$-loss classifier. For a test data set of size N in the two-class problem with class labels $D = 0, 1$, it takes the form

$$\text{Brier score} = \frac{1}{N}\sum_{i=1}^{N}\Big(P_{\widehat{f}}(D = 1|X = x) - D\Big)^2.$$

As such, *smaller* Brier scores translate to better classification.

In Figure **??** we see results for the first illustration, where we try to classify previously unseen test data into one of two multinormal populations that differ only in their means. In particular, we generate training data of total size $n = 100$ and $n = 500$ (that is, on average 50 and 250 in each class) and try *five* separate discrimination methods: the parametric LDA and QDA, the multivariate kernel density estimator, the naive Bayes with marginal kernel density estimates, as well as the new local Fisher discriminant (The $\widehat{D}_{\text{LGDE}}$ of eq. (**??**) gives very similar results to the local Fisher discriminant in these illustrations). For the latter three discriminants we choose one bandwidth for each population based on the cross-validation routine that seeks to maximize the AUC as described in the preceding section. We repeat the experiment 100 times for each combination of sample size and dimension. In each experiment, we evaluate the discrimination using a test data set of size $N = 500$. The plots report the average AUC and Brier scores for the various discriminants as

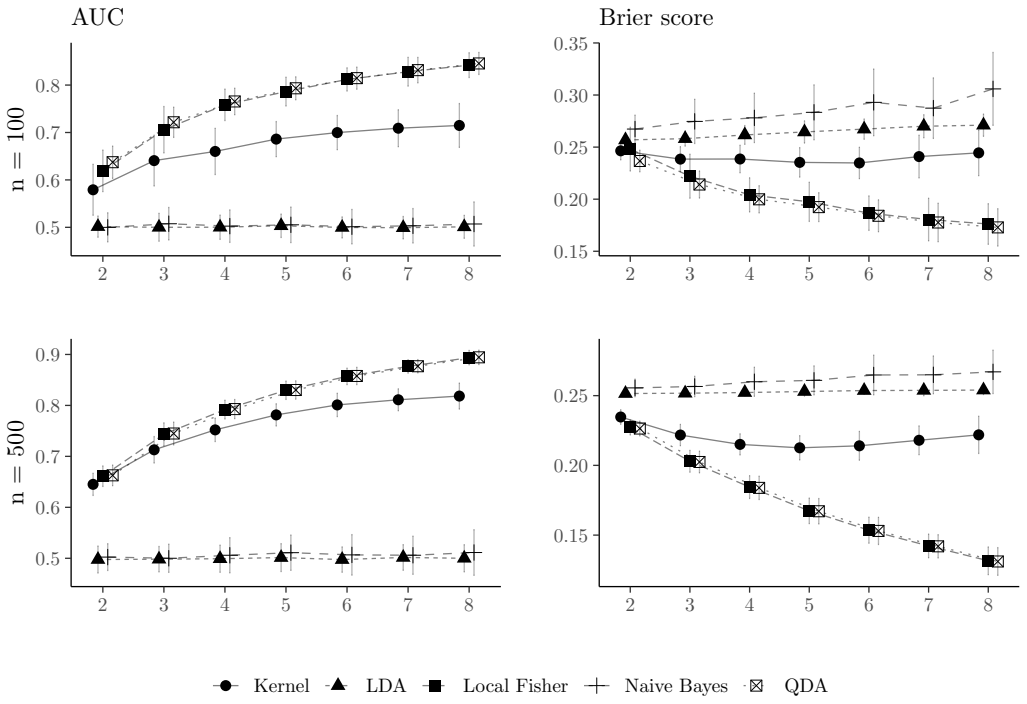

Figure 4: Simulation results for the second illustration: Two multinormal distributions with different covariance matrices. Error measured as a function of dimension.

a function of the number of variables, as well as the standard deviation over the 100 repetitions which we plot as error bars.

In terms of AUC, all methods perform similarly in this case, but in terms of the Brier score, the correctly specified LDA and QDA are clearly better than the two non-parametric methods, and we also see that the local Fisher discriminant performs on par with the QDA, which comes as no surprise because the QDA-rate is attainable for the local Fisher discriminant by choosing large bandwidths.

The results from the second illustration are shown in Figure **??**, and we see clearly that the various discrimination methods are more separated in this case. The two populations, while both being Gaussian, differ only in their covariance matrices which means that the LDA as well as the naive Bayes can simply not see any difference between them, and this emerges clearly in the plots. The kernel density estimator is able to discriminate in this case, but seems to struggle with the curse of dimensionality, especially from the Brier perspective. The QDA represents a correct parametric specification, and thus also the optimal discriminant in this case, but we also see that the local Fisher discriminant has no problems at all to match its performance. This is again due to our cross-validated choice of bandwidths, that seeks to maximize the AUC.

Finally, we look at the third illustration in which the two populations have both equal marginal distributions as well as equal covariance matrices. Since there is no discriminatory information at all in the marginals, nor in the second moments, we see in Figure **??** that also the QDA collapses. We are left with the purely nonparametric kernel estimator – that works, but clearly feels the curse of dimensionality – and the local Fisher discriminant that now must allow its bandwidths to shrink in order to reveal non-Gaussian structures. It does that very well, as we see in the plots, and the pairwise estimation structure for the local covariance matrices is seemingly able to detect clear differences between the two populations regardless of the number of variables.

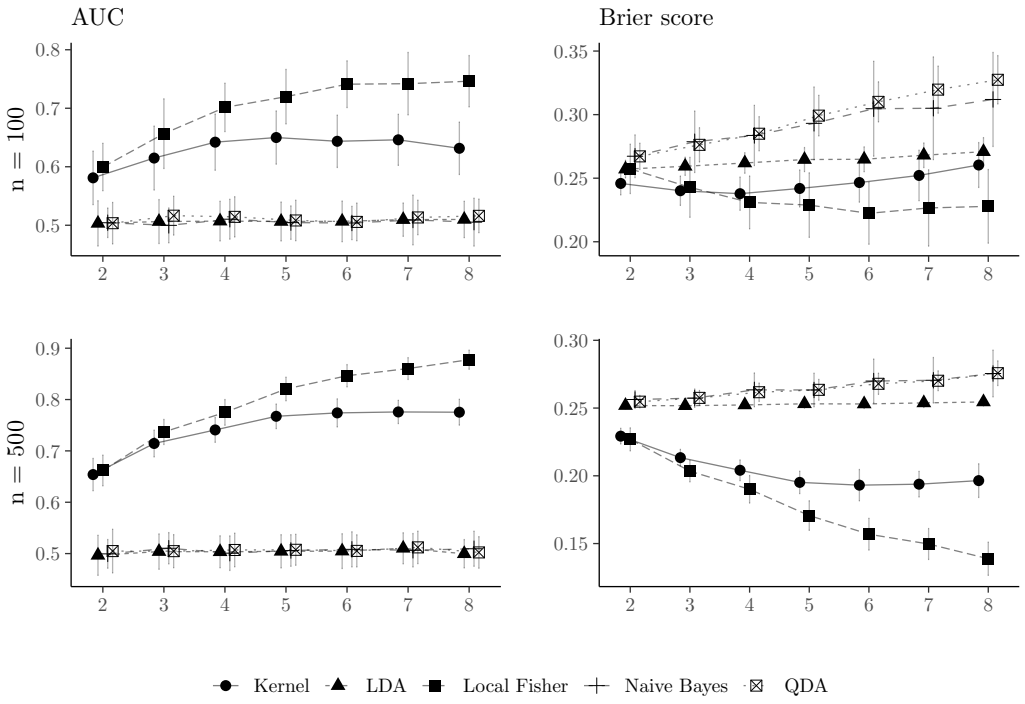

Figure 5: Simulation results for the third illustration: Two multinormal distributions with different covariance matrices. Error measured as a function of dimension.

5.2 Illustration: Fraud detection

Due to the enormous amounts involved, financial crimes such as money laundering is considered a serious threat to societies and economies across the world (**?**). It is therefore crucial that banks and other financial institutions report suspicious transactions and behavior to the authorities, such that thorough investigations and monitoring can be put into effect – ultimately leading to stopping the criminal activity and making the source legally liable. In a money laundering setting with a large Norwegian bank, **?** develop and train a machine learning model for filtering out suspicious transactions from the legitimate ones. Working with a simplified subset of their data, both in terms of the transactions we use and the variables used for discrimination, we illustrate the use of our local Fisher (and $\widehat{D}_{\text{LGDE}}$) discriminant and compare it to the classical discriminants from the above simulation experiments.

We have a data set consisting of 1011 transactions, of which roughly 28% are marked as suspicious. To check how well our discriminants perform, we randomly split this full data set into a training and test set, and rely on the AUC and Brier scores on the test set, as in the simulations experiments.[2] In order to minimize the randomness introduced when splitting the data in training and test sets, we repeat this process 100 times. This is typically referred to as Monte Carlo cross-validation or repeated learning-testing validation (**?**). The reported results are thus mean AUC and Brier scores over the 100 sets, accompanied with 95% confidence intervals for the means using a central limit theorem based normal distribution approximation. To simplify this illustration, we have restricted ourselves to three continuous variables only. In Section **??** we will add discrete variables to this illustration. Due to data restrictions, we can not provide further details about the variables in the data set. The data are plotted in Figure **??**. As seen from the plot, the *combination* of the two first variables, seems to distinguish the two classes fairly well. The third variable may also improve slightly upon their contribution.

[2]The subset of the data used in this illustration contains a small sample of regular customer transactions and transactions reported as suspicious. Transactions which are investigated, but ultimately not reported are not included in our data. This makes the discrimination task much easier than in practice. The true proportion of suspicious transactions is also much smaller. See **?** for details.

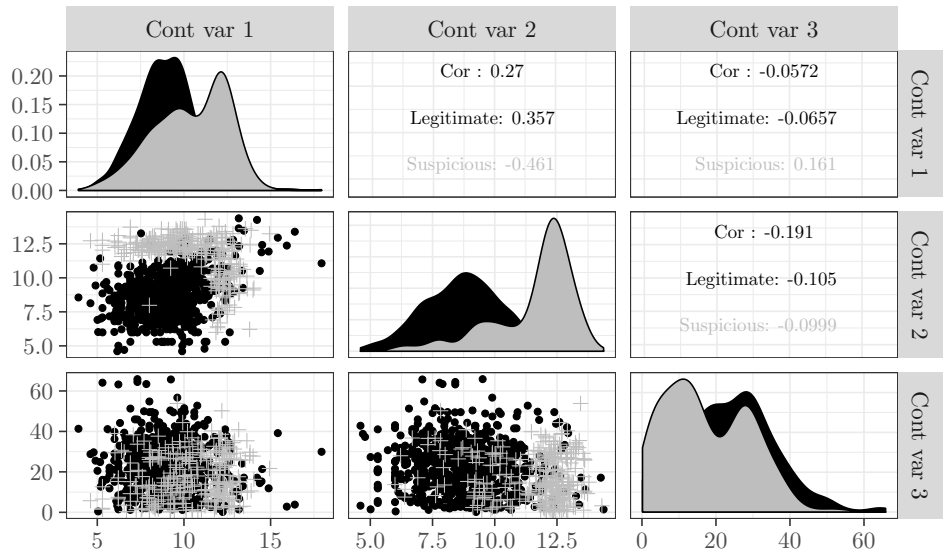

Figure 6: Summary plots for the three continuous variable in the fraud detection illustration. Grey (crosses) marks suspicious transactions, while black (dots) marks legitimate ones.

Table 1: Results using the three continuous variables in the fraud detection example. 95% confidence intervals are shown in brackets. (continued below)

	LGDE	LDA	QDA
AUC	0.964 [0.962, 0.966]	0.904 [0.900, 0.908]	0.949 [0.946, 0.951]
Brier	0.0649 [0.0635, 0.0664]	0.116 [0.115, 0.118]	0.0807 [0.079, 0.0824]

	Naive Bayes	Kernel	Local Fisher
AUC	0.947 [0.944, 0.949]	0.944 [0.941, 0.946]	0.953 [0.950, 0.955]
Brier	0.0794 [0.0774, 0.0813]	0.0847 [0.0828, 0.0866]	0.0768 [0.0748, 0.0787]

Table **??** shows the AUC and Brier scores obtained by the various methods, averaged over the 100 repeated training/test splits, with 95% confidence intervals. Generally speaking, all methods are able to distinguish between the two classes fairly well, as all methods have AUCs larger than 0.9. The LGDE model, however, is clearly the best model for this classification task, both in terms of the AUC and the Brier score, with small confidence intervals. The Local Fisher model is the second best model, with QDA, Kernel and Naive Bayes not too far behind. The LDA model appears to be the least appropriate of these models, with significantly smaller AUC and larger Brier score than the other methods.

6 Discrete variables: Extending naive Bayes

We now move from continuous to discrete class distributions, which is highly relevant in discrimination settings. The term «discrete variables» is broad, and may refer to interpretable numeric variables which can take only some specific values, to unordered categorical variables, or to ordered categorical variables. In this context we shall use the term as a replacement for unordered categorical variables.

As for the discrimination cases with continuous class distributions, the methods we consider are essentially based on estimating the class distributions f_k (which now are probability mass functions and not densities, and therefore will be referred to as p_k) for each class and applying Bayes formula (**??**), and carry out the discrimination according to (**??**). Thus, the rest of this section concerns methods for estimation of such a p_k for a single class k. As we shall only be concerned with the general k-th class distribution, we will throughout this section simplify notation by omitting the k subscript referring to the class.

Consider a sequence of discrete vector variables $X(i), i = 1, \ldots, n$ (from a common class distribution). Each vector variable has d components $X = (X_1, \ldots, X_d)$. Each of these components, X_r, can take k_r different values $\{x_{r1}, \ldots, x_{rk_r}\}$. Since the component X_r can take k_r different values, the vector X can take on $\prod_{r=1}^{d} k_r$ values. The question is then, how we can estimate

$$p(x_{1j_1}, \ldots, x_{dj_d}) = P(X_1 = x_{1j_1}, \ldots, X_d = x_{dj_d}), \tag{17}$$

where $j_1 = 1, \ldots k_1, \ldots, j_d = 1, \ldots k_d$.

There is a sort of curse of dimensionality for discrete variables as well, but it works in a different way than for the continuous case. In the general case there are $\Pi_{r=1}^{d} k_r$ different cells to consider. In the special case of binary variables, then $k_r = 2$ and the number of cells is 2^d. For d large, this will be a very large number. One can still in principle estimate $p(x_{1j_1}, \ldots, x_{dj_d})$ by the straight forward frequency estimator

$$\widehat{p}_{\text{Frequency}}(x_{1j_1}, \ldots, x_{dj_d}) = \frac{1}{n} \sum_{i=1}^{n} 1(X_1(i) = x_{1j_1}, \ldots, X_d(i) = x_{dj_d}) = n_{1j_1,\ldots.dj_d}/n, \tag{18}$$

where n is the total number of observations and $n_{1j_1,\ldots,dj_d}$ is the number of observations in the cell defined by $X_r = x_{rj_r}$, $r = 1, \ldots, d$. Unlike the continuous case there is no bandwidth involved, and $\widehat{p}(x_{1j_i}, \ldots, x_{dj_d})$ converges to $p(x_{1j_1}, \ldots, x_{dj_d})$ with the standard convergence rate of $n^{-1/2}$. However, the problem in practice is that many of the cells may be empty or contain very few observations if d is reasonably large, making it difficult in practice to estimate the probability (**??**).

The influential work of Li and Racine (**?**; **?**) tackle this problem by a discrete-value smoothing algorithm based on earlier work by **?**. The suggested smoothing for component r is

$$l(X_r, x_r, \lambda_r) = \begin{cases} 1 - \lambda_r & \text{if } X_r = x_r; \\ \lambda_r/(k_r - 1) & \text{if } X_r \neq x_r. \end{cases}$$

with $\lambda_r \in [0, (k_r - 1)/k_r]$. For $\lambda_r = 0$, one is back to the indicator function. For $\lambda_r = (k_r - 1)/k_r$, $l(X_r, x_r, \lambda_r) = 1/k_r$; i.e., all differences are in a sense smoothed out. Li and Racine then use the product kernel

$$L(X, x, \lambda) = \prod_{r=1}^{d} l(X_r, x_r, \lambda_r) = \prod_{r=1}^{d} \left(\frac{\lambda_r}{k_r - 1} \right)^{1(X_r \neq x_r)} (1 - \lambda_r)^{1(X_r = x_r)}.$$

The smoothed probability estimate is then given by

$$\widehat{p}_{\mathrm{NP}}(x) = \frac{1}{n}\sum_{i=1}^{n} L(X(i), x, \lambda). \tag{19}$$

Li and Racine find the optimal smoothing parameters $\lambda = (\lambda_1, \ldots, \lambda_d)$ by a cross-validation algorithm. Note that this is a «smoothing» of discrete variables that results in changed probabilities for the values of these same discrete variables, not a change in the values themselves. The cross-validation is done in a clever way to eliminate non-relevant variables in a conditional situation (such as the classification problem). See in particular **?**. The algorithm is implemented in the `R`-package `np` (**?**).

6.1 Pairwise naive Bayes

Contrasting the frequency approach in (**??**), an obvious and much more radical solution to the problem is to use the naive Bayes approach where dependence between components is ignored and $p(x_{1j_1}, \ldots, x_{dj_d})$ is estimated by

$$\widehat{p}_{\text{Naive Bayes}}(x_{1j_1}, \ldots, x_{dj_d}) = \prod_{r=1}^{d} \widehat{p}_{\text{Frequency}}(x_{rj_r}) = \prod_{r=1}^{d} \frac{n_{rj_r}}{n}. \tag{20}$$

Except for certain very rare cases, this approach automatically avoids the problem of empty cells. As this approach has the obvious drawback that all dependence between the variables is completely ignored, it is natural to ask whether one can extend the method in such a way that dependence is accounted for. Motivated in parts by the pairwise approximations of Otneim and Tjøstheim (**?**; **?**), the aim of this section is to derive a new estimator for (**??**) using solely marginals $p_{rj_r} = P(X_r = x_{rj_r})$ with $j_r = 1, \ldots, k_r$ and bivariate probabilities $p_{rj_r,sj_s} = P(X_r = x_{rj_r}, X_s = x_{sj_s})$ with $j_r = 1, \ldots, k_r$ and $j_s = 1, \ldots, k_s$. Note that $\sum_{j_r=1}^{k_r} p_{rj_r} = 1$ and $\sum_{j_r=1}^{k_r}\sum_{j_s=1}^{k_s} p_{rj_r,sj_s} = 1$. Our novel pairwise naive Bayes approach uses a construction which in some sense is similar to the pair-copula construction, see e.g. **?**. More precisely, when a pair of variables is conditioned on another set of variables in a successive conditional representation of a joint distribution, then the conditioning variables are ignored. To simplify notation, write $p_{l\cdots d}$ instead of $p(x_{lj_l}, \ldots, x_{dj_d})$ and $p_{m|l\cdots d}$ instead of $p(x_{mj_m}|x_{lj_l}, \ldots, x_{dj_d}) = P(X_m = x_{mj_m}|X_l = x_{lj_l}, \ldots, X_d = x_{dj_d})$. Consider

$$p_{1\cdots d} = p_{1|2\cdots d}p_{2\cdots d} = p_{1|2\cdots d}p_{2|3\cdots d}p_{3\cdots d}.$$

Continuing in this way, and ignoring the conditioning, results in the naive Bayes formula $p_{1\cdots d} = p_1 p_2 \cdots p_d$. We now try to do the same reasoning, but on pairwise probabilities. Writing p_{lm} instead of $p(x_{lj_l}, x_{mj_m})$ and $p_{lm|u\cdots d}$ instead of $p(x_{lj_l}, x_{mj_m}|x_{uj_u}, \ldots, x_{dj_d}) = P(X_l = x_{lj_l}, X_m = x_{mj_m}|X_u = x_{uj_u}, \ldots, X_d = x_{dj_d})$, and assuming that the dimension d is an even number:

$$p_{1\cdots d} = p_{12|3\cdots d}p_{3\cdots d} = p_{12|3\cdots d}p_{34|5\cdots d}\cdots p_{d-1,d}. \tag{21}$$

Omitting conditioning we approximate this expression by

$$p_{\text{Pairwise,even}} = p_{12}p_{34}\cdots p_{d-1,d},$$

with the similar expression $p_{\text{Pairwise,odd}} = p_{12}p_{34}p_{d-2,d-1}p_d$ in the case where d is odd. This approximation can be done in many ways, however, in general each giving different results. (The decomposition can of course be done in many ways in the naive Bayes case as well, but here they all give the same result $p_1 p_2 \cdots p_d$).

In the case of four variables the decomposition (**??**) can be done in 6 different ways

$$p_{1234} = \begin{cases} p_{12|34}p_{34} \approx p_{12}p_{34} \\ p_{13|24}p_{24} \approx p_{13}p_{24} \\ p_{14|23}p_{23} \approx p_{14}p_{13} \\ p_{23|14}p_{14} \approx p_{23}p_{14} \\ p_{24|13}p_{13} \approx p_{24}p_{13} \\ p_{34|12}p_{12} \approx p_{34}p_{12}. \end{cases} \tag{22}$$

Since the various $p_{ij}p_{kl}$ products generally give different answers, we suggest an estimate obtained by taking the geometric mean,

$$\widehat{p}_{1234} = (\widehat{p}_{12} \cdot \widehat{p}_{34} \cdot \widehat{p}_{13} \cdot \widehat{p}_{24} \cdot \widehat{p}_{14} \cdot \widehat{p}_{13} \cdot \widehat{p}_{23} \cdot \widehat{p}_{14} \cdot \widehat{p}_{24} \cdot \widehat{p}_{13} \cdot \widehat{p}_{34} \cdot \widehat{p}_{12})^{1/6}, \tag{23}$$

where $\widehat{p}_{lm}$ is used as a shorthand notation for $\widehat{p}_{\text{Frequency}}(x_{lj_l}, x_{mj_m})$. The factors in (**??**) are identical in pairs, and taking this into account, (**??**) reduces to

$$\widehat{p}_{1234} = (\widehat{p}_{12} \cdot \widehat{p}_{34} \cdot \widehat{p}_{13} \cdot \widehat{p}_{24} \cdot \widehat{p}_{14} \cdot \widehat{p}_{23})^{1/3}, \tag{24}$$

which we easily see reduces to the naive Bayes formula in case of independence between all variables.

Let us now turn to the general derivation when d is even. Corresponding to the expression (**??**), in the first position, there are $d(d-1)/2$ options. In the second position, we have used two variables, so there are $(d-2)(d-3)/2$ pairs left to choose from, and so on. This means that the number of decompositions consisting only of pairs of variables, is

$$S = \frac{d(d-1)}{2} \cdot \frac{(d-2)(d-3)}{2} \cdot \cdots \cdot \frac{2 \cdot 1}{2} = \frac{d!}{2^{d/2}},$$

because there are exactly $d/2$ factors in each such decomposition.

Denote each decomposition by $g_1, \ldots, g_S$. In the general version of (**??**) - (**??**), there are $S \cdot (d/2)$ factors in total, but there are only $d(d-1)/2$ pairs and thus unique factors after the approximation (after we drop the conditioning). The number of times each factor occurs, then, is equal to

$$T = \frac{\text{No. of lines like those in eq. (??)} \times \text{No. of factors in each line}}{\text{No. of unique factors}} = \frac{\frac{d!}{2^{d/2}} \cdot \frac{d}{2}}{\frac{d(d-1)}{2}} = \frac{d(d-2)!}{2^{d/2}}.$$

We approximate $p_{1\cdots p}$ by taking the geometric mean of all the approximations $g_1, \ldots, g_S$:

$$\left(\prod_{j=1}^{S} g_j\right)^{1/S} = \left(\prod_{j=1}^{d!/2^{d/2}} g_j\right)^{2^{d/2}/d!}.$$

This, in turn, simplifies because the individual pairwise probabilities comprising $g_1, \ldots, g_S$ are repeated S times each in the product above, so that we get the following estimator

$$\begin{aligned}\widehat{p}'_{\text{Pairwise Naive Bayes, even}}(x_{1j_1}, \ldots, x_{dj_d}) &= \left(\prod_{j=1}^{S} g_j\right)^{1/S} = \left(\prod_{l<j\le d} \widehat{p}_{jl}^{\,S}\right)^{1/S} \\ &= \left(\prod_{l<j\le d} \widehat{p}_{jl}^{\,\frac{d(d-2)!}{2^{d/2}}}\right)^{\frac{2^{d/2}}{d!}} = \left(\prod_{l<j\le d} \widehat{p}_{jl}\right)^{\frac{1}{d-1}}.\end{aligned} \tag{25}$$

This is not the geometric mean of the $d(d-1)/2$ pairwise probabilities, but their product raised to the $(d-1)^{-1}$st power, see (**??**) for the special case with $d = 4$. It is seen that this reduces to the product of marginal probabilities under independence; i.e., naive Bayes, because each variable will be represented in exactly $d-1$ pairs each. Moreover, in case $d = 2$, it reduces to p_{12}.

We now turn to the case when d is odd. This is very similar, but we have to include the marginal probabilities into the formula. It is not difficult to show that in this case one ends up with

$$\widehat{p}'_{\text{Pairwise Naive Bayes, odd}}(x_{1j_1},\ldots,x_{dj_d}) = \left(\prod_{j=1}^{S} g_j\right)^{1/S} = \left(\prod_{j=1}^{d!/2^{(d-1)/2}} g_{ij}\right)^{2^{(d-1)/2}/d!}$$
$$= \left(\prod_{j<l\leq d} \widehat{p}_{jl}^{\frac{(d-1)!}{2^{(d-1)/2}}} \prod_{j=1}^{d} \widehat{p}_j^{\frac{(d-1)!}{2^{(d-1)/2}}}\right)^{2^{(d-1)/2}/d!} = \left(\prod_{j<l\leq d} \widehat{p}_{jl} \prod_{j=1}^{d} \widehat{p}_j\right)^{1/d}. \tag{26}$$

The first product in the expression above is the same as in the even case, but with the exponent $1/d$ instead of $1/(d-1)$. The second product is in fact the geometric mean of the marginal probabilities. Under independence, the first product contains each marginal probability $d-1$ times (as before), and then each of them enter once more in the second product. The exponent then cancels, and we are left with just the product of the marginal probabilities, i.e. the naive Bayes formula.

It is important to realize that, unlike the naive Bayes, the pairwise approximations in (**??**) and (**??**) need not be proper probability distributions, i.e. they may not sum to 1. To arrive at proper probability estimators, one must normalize:

$$\begin{aligned}
\widehat{p}_{\text{Pairwise Naive Bayes, even}}(x_{1j_1},\ldots,x_{dj_d}) &= \frac{\widehat{p}'_{\text{Pairwise Naive Bayes, even}}(x_{1j_1},\ldots,x_{dj_d})}{\sum_{l_1=1}^{k_1}+\cdots+\sum_{l_d=1}^{k_d}\widehat{p}'_{\text{Pairwise Naive Bayes, even}}(x_{1l_1},\ldots,x_{dl_d})}, \\
\widehat{p}_{\text{Pairwise Naive Bayes, odd}}(x_{1j_1},\ldots,x_{dj_d}) &= \frac{\widehat{p}'_{\text{Pairwise Naive Bayes, odd}}(x_{1j_1},\ldots,x_{dj_d})}{\sum_{l_1=1}^{k_1}+\cdots+\sum_{l_d=1}^{k_d}\widehat{p}'_{\text{Pairwise Naive Bayes, odd}}(x_{1l_1},\ldots,x_{dl_d})},
\end{aligned} \tag{27}$$

but as in the continuous case we have used the non-normalized quantities in discrimination ratios.

This procedure can clearly be generalized to consider products of $\binom{d}{3}$ factors of trivariate probabilities for dimensions $d = 3d'$ for some integer d' (with some adjustments for $d = 3d' + j$, $j = 1, 2$) and then taking the $\binom{d-1}{2}$-root of this and normalize. Again this reduces to the right thing for $d = 3$ or in the independent case. This can be generalized to higher order interactions.

It is not difficult to show that the pairwise naive Bayes estimators in (**??**) achieve the usual root-n asymptotic normality property when compared to respectively $p_{\text{Pairwise, even}}$ and $p_{\text{Pairwise, odd}}$. Due to the notational complexity of their construction, their asymptotic variance is also quite complicated and notationally inconvenient to derive. We will therefore only sketch the derivation of the estimators' asymptotic normality. Since the estimators are both continuously differentiable functions (products and d-roots) of the various $\widehat{p}_j$ and $\widehat{p}_{jl}$, it suffices to show asymptotic normality for each of these, and applying the delta method. Since both $\widehat{p}_j$ and $\widehat{p}_{jl}$ are sums of independent variables, it follows from the ordinary central limit theorem for iid variables that $\sqrt{n}(\widehat{p}_j - p_j)$ and $\sqrt{n}(\widehat{p}_{jl} - p_{jl})$ converge in distribution to zero-mean normals with certain variances. Thus, zero-mean asymptotic normality of $\sqrt{n}(\widehat{p}_{\text{Pairwise Naive Bayes, even}} - p_{\text{Pairwise,even}})$ and $\sqrt{n}(\widehat{p}_{\text{Pairwise Naive Bayes, odd}} - p_{\text{Pairwise, odd}})$ follows by the delta method. In the general case, where the dependence between the variables takes a more complicated structure than the pairwise, these estimators will be biased.

One potential problem in this context is the possibility of empty pairwise cells. This phenomenon is likely to appear more often as the number of variables increases, and poses a particular problem in the discrimination setting because it may happen that the two posterior class probability estimates both equal zero because of this, regardless of the values of the other pairwise probability estimates. In order to avoid this we suggest to simply «add ϵ observations» to the empty variable pairs in the training data where $\epsilon \in (0, 1)$. At present we have used an ad hoc solution in choosing $\epsilon = 1/2$, resulting in replacing pairwise empirical frequencies of 0 with $\frac{1}{2}/n$.

7 The mixed continuous-discrete case

So far we have considered the situations where all variables that ought to be used for discrimination are either continuous or discrete. In the present section we discuss the situation where we have both variable types present at once.

The simplest solution to handle mixed data types is to treat the continuous and discrete variables separately. Within each class of the classification problem one could then choose ones favorite procedure for modeling the continuous variables, and vice versa for the discrete ones – for instance, respectively, via our pairwise Fisher and pairwise naive Bayes approaches. Assuming independence between the continuous and discrete set of variables allows multiplying estimated distributions together, giving an estimate which can be used for classification as described earlier. However, this independence assumption is too drastic in most situations.

Our take on this is to take dependence between continuous and discrete variables into consideration by first modeling the continuous variables with the LGDE approach in Section **??**, and then conditioned on the continuous variables set up a logistic, log-linear or even generalized additive model (GAM), cf. **?**. To clarify notation, let us use X^c and x^c for the d_c-dimensional continuous data, and similarly X^d and x^d for the d_d-dimensional discrete data. Assuming $d_d \geq 2$, if $\phi(u)$ is a link function; e.g. $\phi(u) = \log(u/(1-u))$, then for an observed continuous d_c-dimensional x^c with $u = p_{rj_r,sj_s}$ one can model $\phi(p_{rj_r,sj_s})$ linearly as

$$\phi(p_{ir_r,sj_s}) = \beta_0^{rj_r,sj_s} + \sum_{j=1}^{d_c} \beta_j^{rj_r,sj_s} x_j^c, \tag{28}$$

or additively as

$$\phi(p_{rj_r,sj_s}) = h_0^{rj_r,sj_s} + \sum_{j=1}^{d_c} h_j^{rj_r,sj_s}(x_j^c). \tag{29}$$

The unknown β parameters can be estimated by maximum likelihood using a GLM software package, and in the additive case the h_is can be estimated by e.g. the `mgcv`-package (**?**) in the `R` programming language (**?**). Note that if the dimension of x^c is large, which is likely in e.g. fraud applications, then one may consider (ridge or lasso type of) regularized logistic regression (**?**, Ch. 5). We obtain estimates of marginal probabilities $p_{rj_r}(x)$ by using $p_{rj_r}(x) = \sum_{j_s=1}^{k_s} p_{rj_r,sj_s}(x)$. If there is only a single discrete variable ($d_d = 1$), then $\phi(p_{rj_r})$ is modeled directly in the same manner as $\phi(p_{rj_r,sj_s})$ above. In the training phase this should be done separately for the K training sets. In case there is no dependence on continuous variables the estimate of the intercept β_0 or h_0 will be close to a ϕ-transformation of $\widehat{p}_{rj_r,sj_s} = n_{rj_r,sj_s}/n$.

Once we have estimated the x^c-dependent probabilities $p(x^d_{rj_r}|x^c)$ and $p(x^d_{rj_r,sj_s}|x^c)$, we compute the corresponding (unnormalized) probability $\widehat{p}'_{\text{Pairwise Naive Bayes, k}}(x^d_{1j_1},\ldots,x^d_{dj_d}|x^c)$ using the procedure of Section **??**. The pairwise estimator of the class distributions for mixed data is finally completed by multiplying with the estimate of the continuous density, i.e.:

$$\widehat{f}_{\text{Pairwise, mixed, }k} = \widehat{p}'_{\text{Pairwise naive Bayes, k}}(x^d_{1j_1},\ldots,x^d_{dj_d}|x^c)\widehat{f}_{\text{LGDE},k}(x^c). \tag{30}$$

By obtaining estimates of the a priori probabilities π_k, we may proceed to perform the classification task through a straightforward application of the Bayes rule as in (**??**).

8 Illustrations in the discrete and mixed case

8.1 Simulations in the discrete case

One way to explore the finite sample properties of the pairwise discrete probability estimator in a classification setting is to set up a simulation experiment in the same way as we did in the continuous case in Section **??**, where we gradually increase the number of variables. We shall consider two different types of problems, which have fundamental similarities to Problem 3 for the continuous case:

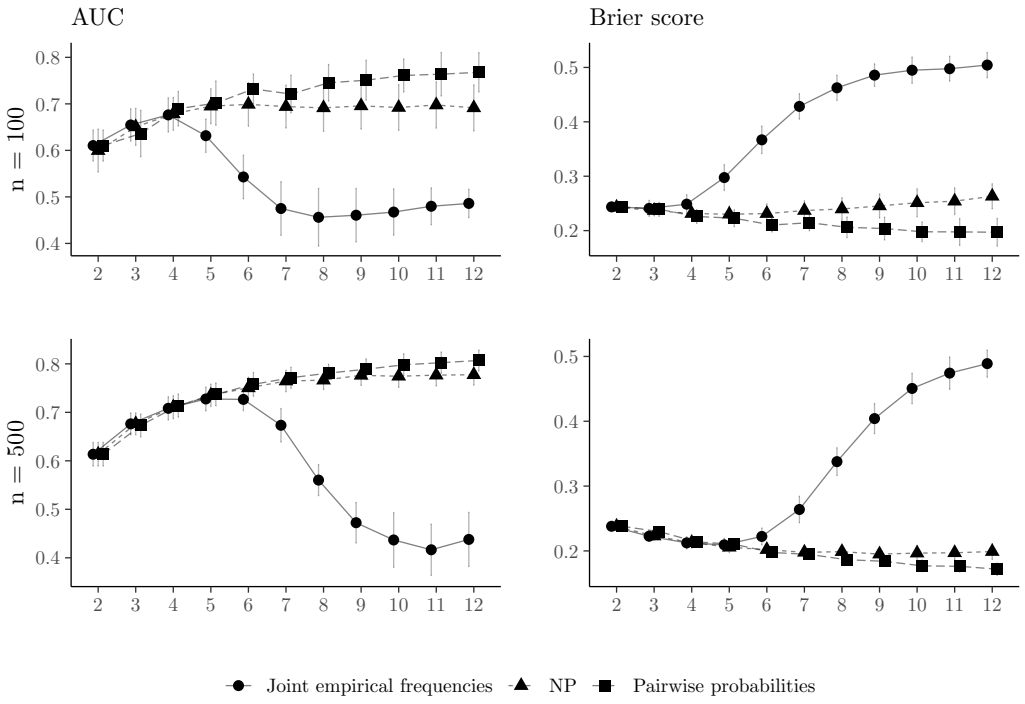

Figure 7: Simulation results for discrete problem 2: Two discretized Clayton populations with having weak and strong dependence, respectively.

- **Discrete problem 1:** We define two continuous populations both being marginally standard normal, but having two different dependence structures defined by the Clayton copula (**?**) with two different parameter values: $\theta = 0.1$ (weak dependence between the variables) and $\theta = 2$ (strong dependence between the variables). Then we discretize these observation by assuming that we only observe the *sign* of them: -1 or 1.
- **Discrete problem 2:** We complicate the discrimination task between the populations in discrete problem 1 in two ways: 1) We reduce the dependence between the variables in the second population by setting $\theta = 0.9$ (while keeping $\theta = 0.1$ in the first population.) 2) We discretize the continuous variables into three categories instead of two, -1, 0 and 1 by placing the boundaries between the categories in such a way that all marginal distributions in both populations are uniform.

Since the marginals for the two populations are equal in both illustrations, there is no point in trying to discriminate between the populations by looking only at marginal probabilities and using the naive Bayes. We must, one way or the other, extract discriminatory information from the dependence between variables. We shall compare the following three discriminants:

1. Estimate $p_{1\ldots,d}$ using empirical frequencies as in (**??**), proceed via Bayes formula (**??**), and carry out the discrimination according to (**??**).
2. Calculate conditional class probabilities directly using the smoothing algorithm in (**??**), implemented in the `np`-package.
3. Estimate $p_{1\ldots,d}$ using our pairwise probability approximation in (**??**) and (**??**), proceed via Bayes formula in (**??**), and carry out the discrimination according to (**??**).

We evaluate the discriminants using the AUC and Brier scores as we did in Section **??**.

Consult Figure **??** for the results of discrete problem 1. We have allowed the dimension of the problem to range from 2 to 12. We see clearly that the curse of dimensionality ruins the joint empirical frequencies from dimension 5 or 6, depending a little bit on the sample size. The NP-estimator as well as the pairwise probability estimates, on the other hand, perform much better, the latter of which having a slight advantage in this case.

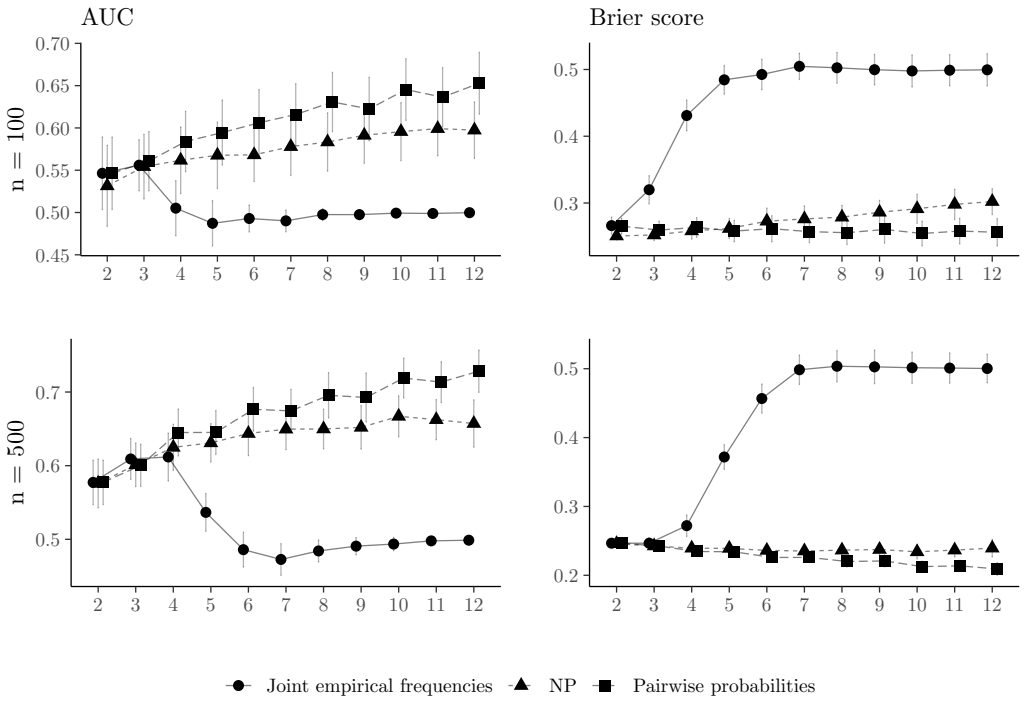

Figure 8: Simulation results for discrete problem 2: Two three-category discretized Clayton populations with having weak and not-as-strong dependence, respectively.

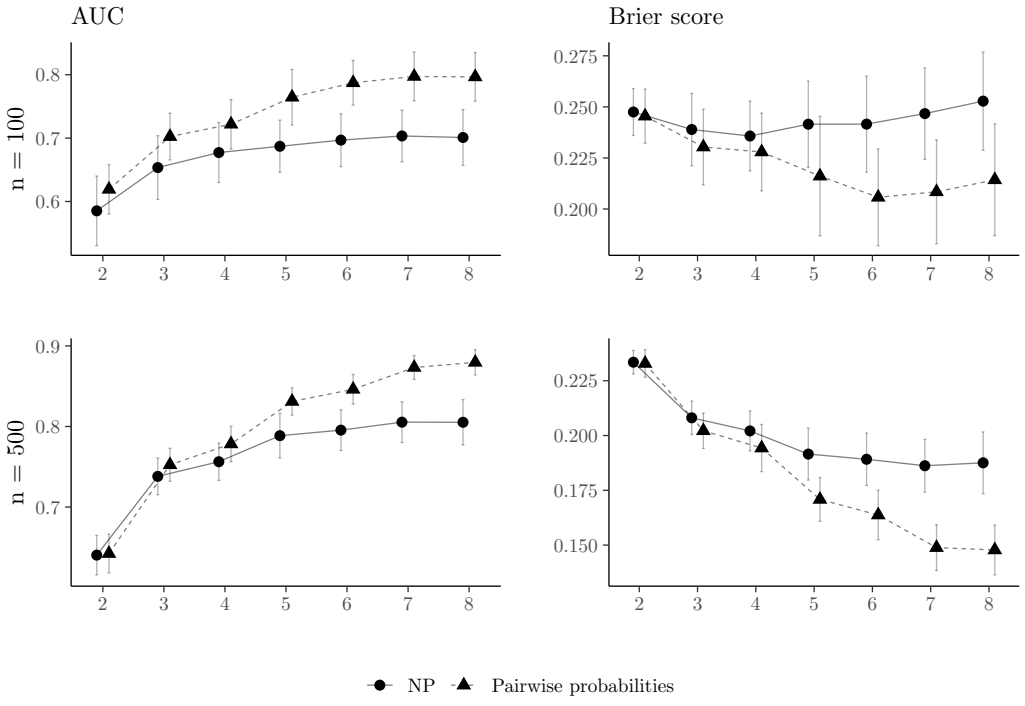

Figure 9: Simulation results for the mixed example: Two Clayton populations where every other variable is discretized.

We present the results of discrete problem 2 in Figure **??**, where we see the total collapse of the empirical frequencies, as well as evidence suggesting that the two alternatives are useful discriminants in all dimensions, again with an advantage given to the pairwise procedure.

8.2 Simulations in the mixed variable case

To accompany the simulations for the situation with purely discrete variables, let us consider a simulation experiment with mixed variables, where we again explore the performance while gradually increasing the dimension of the variables in the two class distributions. We do this by modifying **Discrete problem 1** in the preceding subsection as follows:

- **Mixed problem**: We generate continuous variables in the same way as in **Problem 1** in Section **??**. To create mixed variables, we discretize *every other* variable: Variables $2, 4, \ldots$ are converted to the categories -1 or 1 corresponding to their sign.

We can no longer use empirical frequencies directly when there are continuous variables present. We could, of course, construct a naive estimate of the posterior probabilities in the mixed case by multiplying the naive Bayes, or joint kernel estimates, with empirical frequencies. However, given our findings in earlier illustrations, we have little hope in producing good classification from such a procedure, so we choose not to implement it here. We are rather left with two options:

1. The **?** method for computing conditional probabilities and densities directly.
2. Our pairwise procedure for combining locally Gaussian density estimates with the pairwise frequency approach as in (**??**), using the logistic regression (**??**) or the generalized additive model (**??**).

The results in the mixed case are presented in Figure **??**. We have used the logistic regression approach for $n = 100$ in order to ensure numerical stability, but switch to the GAM when $n = 500$. The two methods perform comparably in terms of both error measures, but our new method is again slightly better with increasing dimension. We must note here though that the **?** method for estimating conditional probabilities is not tuned specifically towards discrimination.

8.3 Illustration: Fraud detection

In this section we build further on the fraud detection example in Section **??**, by including seven discrete variables in addition to the three continuous ones. The number of training observations in each of the categories are shown in Table **??**. Category 1 of discrete variable 1 seems to be a decent indicator of a suspicious transaction. Apart from that, there seems to be little information in the variables when looking at them one by one, but there may of course be crucial patterns appearing when combining them both with each other and with the continuous variables from Section **??**. We will check the performance of the discriminants used in the above simulation experiments on the test data, both when using only the discrete variables, and when combining the two data types. We use the same validation scheme as in Section **??**, validating the performance using AUC and Brier scores on 100 repeated training/test splits of the full data set.

8.3.1 Discrete variables only

For illustrative purposes, we first allow the discriminants to use the seven discrete variables only. The performance results from the various discriminants on the test set are shown in Table **??**. As seen from the table, our Pairwise probabilities approach and the NP approach perform essentially equally well, both for the AUC and the Brier score. By carefully looking at the confidence intervals, our pairwise probabilities method has slightly superior bounds, but this is highly uncertain. These two methods are anyway clearly superior to the joint empirical frequencies method.

Disc var 1	# Suspicious	# Legitimate
Category 1	188	42
Category 2	1	40
Category 3	32	437
Category 4	8	9
Category 5	9	76
Category 6	16	60
Category 7	27	66

Disc var 2	# Suspicious	# Legitimate
Category 1	255	724
Category 2	26	6

Disc var 3	# Suspicious	# Legitimate
Category 1	35	93
Category 2	246	637

Disc var 4	# Suspicious	# Legitimate
Category 1	94	259
Category 2	175	454
Category 3	0	7
Category 4	8	5
Category 5	4	5

Disc var 5	# Suspicious	# Legitimate
Category 1	5	27
Category 2	276	703

Disc var 6	# Suspicious	# Legitimate
Category 1	7	16
Category 2	17	6
Category 3	252	697
Category 4	5	11

Disc var 7	# Suspicious	# Legitimate
Category 1	14	88
Category 2	267	642

Table 3: Share of observations in the different categories for the seven discrete variables in the money laundering fraud detection example.

8.3.2 Mixed variables

In this section we allow the discriminants to use both the three continuous variables and the seven discrete variables. The performance results for the three discriminants accessible in this setting are shown in Table **??**. As seen from the table, the NP method seems to be the best performing method in terms of both AUC and the Brier score. The confidence interval for the AUC does, however, overlap with that of the GLM based pairwise probability approach, so the results are not fully conclusive in this manner. This is not the case for the Brier score. One possible reason that the GAM based version of the pairwise probability approach is not performing as well here, is that it might be overfitting the dependence between the discrete and continuous variables.

9 Summary remarks

We have demonstrated how the two standard discriminants, the Fisher and the naive Bayes, can be extended by a (pairwise) local Gaussian Fisher discriminant and by a geometric mean of pairwise probabilities, respectively, for continuous and discrete variables. For the mixed case, we merge the two approaches and handle dependence between the two variable types with a logistic regression type approach. The performance of the new discriminants have been compared to the ordinary Fisher and naive Bayes discriminant as well as to a nonparametric discriminant based on the kernel density estimator in the continuous case, and NP-filtered probability estimators considered by **?** in the discrete and mixed distribution case. Our experiments show

	Joint empirical frequencies	NP	Pairwise probabilities
AUC	0.833 [0.828, 0.838]	0.857 [0.853, 0.861]	0.858 [0.855, 0.862]
Brier	0.126 [0.123, 0.128]	0.112 [0.111, 0.114]	0.112 [0.110, 0.113]

Table 4: Results using the seven discrete variables in the money laundering fraud detection example. 95% confidence intervals are shown in brackets.

	NP	Pairwise probabilities (GLM)	Pairwise probabilities (GAM)
AUC	0.979 [0.977, 0.980]	0.969 [0.967, 0.972]	0.944 [0.940, 0.947]
Brier	0.048 [0.0464, 0.0496]	0.051 [0.0492, 0.0529]	0.0766 [0.0738, 0.0794]

Table 5: Results using three continuous variables and seven discrete variables in the money laundering fraud detection example. 95% confidence intervals are shown in brackets.

significant improvements compared to the two classic discriminants, and also good performance results compared to the nonparametric alternatives.

There is a substantial potential for further research and modifications. For instance, we have ignored the normalization issue in computing discrimination ratios. Further, in the discrete case, we have only worked with unordered categorical variables, while extensions to ordered categorical variables or numerically-valued discrete data would clearly also be of interest. The method for replacing zeros in the estimated discrete pairwise probabilities also warrant a more systematic investigation. One possibility is a variant of the NP-filtering of **?** applied to the initial pairwise probabilities. Bagging and boosting (**?**) being general methods for potential improvements of discriminants, may also represent a possible direction for improvement.

Finally, the purpose and motivation for the paper has not been to invent the ultimately best discriminant in every situation, but merely to extend two classical discriminants in a coherent way. This is also the reason for comparing our methods to the most natural statistically founded alternatives – as opposed to comparing them to top notch algorithmic methods in the machine learning literature, which often require specification of long lists of tuning parameters. It would, however, be interesting to see whether our approaches, being built on completely different grounds, can utilize the data differently than those methods, and therefore bring something new to the table. If this is indeed the case, combining the different flavored discriminants, for instance by an ensemble method from **?**, seems like a promising approach.

Acknowledgment: We are grateful to two anonymous referees for constructive comments.

References

Aas, K., Czado, C., Frigessi, A., and Bakken, H. (2009). Pair-copula constructions of multiple dependence. *Insurance: Mathematics and Economics*, 44(2):182–198.

Aggarwal, C. C. and Zhai, C. (2012). A survey of text classification algorithms. In *Mining text data*, pages 163–222. Springer.

Aitchison, J. and Aitken, C. G. (1976). Multivariate binary discrimination by the kernel method. *Biometrika*, 63(3):413–420.

Azzalini, A. (1981). A note on the estimation of a distribution function and quantiles by a kernel method. *Biometrika*, 68(1):326–328.

Berentsen, G. D., Kleppe, T. S., and Tjøstheim, D. (2014a). Introducing localgauss, an r-package for estimating and visualising local gaussian correlation. *Journal of Statistical Software*, 56(1):1–18.

Berentsen, G. D., Støve, B., Tjøstheim, D., and Nordbø, T. (2014b). Recognizing and visualizing copulas: an approach using local gaussian approximation. *Insurance: Mathematics and Economics*, 57:90–103.

Berentsen, G. D. and Tjøstheim, D. (2014). Recognizing and visualizing departures from independence in bivariate data using local gaussian correlation. *Statistics and Computing*, 24(5):785–801.

Blanzieri, E. and Bryl, A. (2008). A survey of learning-based techniques of email spam filtering. *Artificial Intelligence Review*, 29(1):63–92.

Box, G. E. P. and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis*. John Wiley & Sons.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthey Weather Review*, 78(1):1–3.

Burman, P. (1989). A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods. *Biometrika*, 76(3):503–514.

Chaudhuri, P., Ghosh, A. K., and Oja, H. (2009). Classification based on hybridization of parametric and nonparametric classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(7):1153–1164.

Fawcett, T. (2006). An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2):179–188.

Ghosh, A. and Hall, P. (2008). On error rate estimation in nonparametric classification. *Statistica Sinica*, 18:1081–1100.

Ghosh, A. K. and Chaudhuri, P. (2004). Optimal smoothing in kernel discriminant analysis. *Statistica Sinica*, 14:457–483.

Hall, P., Racine, J., and Li, Q. (2004). Cross-validation and the estimation of probability densities. *Journal of the American Statistical Association*, 99(99):1015–1026.

Hart, J. D. and Vieu, P. (1990). Data-driven bandwidth choice for density estimation based on dependent data. *Annals of Statistics*, 18:873–890.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*. Springer, New York. 2nd edition.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*. Chapman and Hall, London.

Hayfield, T. and Racine, J. S. (2008). Nonparametric econometrics: The np package. *Journal of Statistical Software*, 27(5):1–32.

Hjort, N. and Jones, M. (1996). Locally parametric nonparametric density estimation. *Annals of Statistics*, 24:1619–1647.

Hjort, N. L. and Glad, I. K. (1995). Nonparametric density estimation with a parametric start. *Annals of Statistics*, 23:882–904.

Johnson, R. A. and Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis, Sixth Edition*. Pearson Education Iternational.

Jones, M. C. and Signorini, D. (1997). A comparison of higher-order bias kernel density estimators. *Journal of the American Statistical Association*, 92(439):1063–1073.

Jordanger, L. A. and Tjøstheim, D. (2019). Nonlinear spectral analysis: A local gaussian approach. *Preprint arXiv: 1709.02166v2*.

Jullum, M., Løland, A., Huseby, R. B., Ånonsen, G., and Lorentzen, J. P. (2020). Detecting money laundering transactions – which transactions should we learn from? *Journal of Money Laundering Control, to appear*.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *International Joint Conference on Artificial Intelligence (IJCAI)*, volume 14, pages 1137–1145. Montreal, Canada.

Lacal, V. and Tjøstheim, D. (2017). Local gaussian autocorrelation and tests of serial independence. *Journal of Time Series Analysis*, 38(1):51–71.

Lacal, V. and Tjøstheim, D. (2018). Estimating and testing nonlinear local dependence between two time series. *Journal of Business and Economic Statistics*. in press.

Li, J., Cuesta-Albertos, J. A., and Liu, R. Y. (2012). Dd-classifier: Nonparametric classification procedure based dd-plot. *Journal of the American Statistical Association*, 107(498):737–753.

Li, Q. and Racine, J. S. (2007). *Nonparametric Econometrics: Theory and Practice*. Princeton University Press, Princeton.

Li, Q. and Racine, J. S. (2008). Nonparametric estimation of conditional cdf and quantile functions with mixed categorical and continuous data. *Journal of Business and Economic Statistics*, 26(4):423–434.

Loader, C. R. (1996). Local likelihood density estimation. *Annals of Statistics*, 34:1602–1618.

Marron, J. S. (1983). Optimal rates of convergence to bayes risk in nonparametric discrimination. *Annals of Statistics*, 11(4):1142–1155.

Min, J. H. and Jeong, C. (2009). A binary classification method for bankruptcy prediction. *Expert Systems with Applications*, 36(3):5256–5263.

Nadaraya, E. A. (1964). Some new estimates for distribution functions. *Theory of Probability & Its Applications*, 9(3):497–500.

Nelsen, R. B. (2007). *An introduction to copulas*. Springer Science & Business Media.

Otneim, H. (2018). *lg: Locally Gaussian Distributions: Estimation and Methods*. R package version 0.3.0.

Otneim, H. and Tjøstheim, D. (2017). The locally Gaussian density estimator for multivariate data. *Statistics and Computing*, 27(6):1595–1616.

Otneim, H. and Tjøstheim, D. (2018). Conditional density estimation using the local Gaussian correlation. *Statistics and Computing*, 28(2):303–321.

Phua, C., Lee, V., Smith, K., and Gayler, R. (2010). A comprehensive survey of data mining-based fraud detection research. *arXiv preprint arXiv:1009.6119*.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ranjan, R. and Gneiting, T. (2010). Combining probability forecasts. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(1):71–91.

Samworth, R. (2012). Optimal weighted nearest neighbour classifiers. *Annals of Statistics*, 40:2733–2763.

Satabdi, P. (2018). A svm approach for classification and prediction of credit rating in the indian market. Working paper.

Schott, P. A. (2006). *Reference guide to anti-money laundering and combating the financing of terrorism*. The World Bank.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London.

Stone, C. J. (1990). Large-sample inference for log-spline models. *The Annals of Statistics*, pages 717–741.

Stone, C. J., Hansen, M. H., Kooperberg, C., Truong, Y. K., et al. (1997). Polynomial splines and their tensor products in extended linear modeling: 1994 wald memorial lecture. *The Annals of Statistics*, 25(4):1371–1470.

Tjøstheim, D. (1978). Improved seismic discrimination using pattern recognition. *Physics of the Earth and Planetory Interiors*, 16:85–108.

Tjøstheim, D. and Hufthammer, K. O. (2013). Local gaussian correlation: A new measure of dependence. *Journal of Econometrics*, 172:33–48.

Wood, S. (2017). *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC, 2 edition.

Zheng, R., Li, J., Chen, H., and Huang, Z. (2006). A framework for authorship identificationof online messages: writing style features and classification techniques. *Journal of the American Society for Information Science and Technology*, 57(3):378–393.