Pairwise local Fisher and Naive Bayes: Improving two standard discriminants

Håkon Otneim, Martin Jullum and Dag Tjøstheim

19.12.2018

You are reading the pre-print version of this paper. Publisher's version is available here: https://doi.org/10.1016/j.jeconom.2020.01.019

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 modelling 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 give significant improvements.

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 (Fisher, 1936) and the Naive Bayes discriminant (Hastie et al., 2009, p. 210-211). These are easy to understand and to apply and have been much used in practice(Hastie et al., 2009). 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.

In this paper we present novel discrimination procedures generalizing the Fisher and the Naive Bayes, respectively. 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. 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 modelled conditionally on the continuous variables with a logistic regression type procedure.

1.1 Background

Let us first provide some backround for the classification[1] problem. The K-class discrimination problem consists in assigning the d-dimensional data vector $X = (X_1, \dots, X_d)$ to one of K classes. Examples ranges from fraud detection, authorship and text analysis, spam-email detection, credit rating, bankruptcy prediction and seismic discrimination (see e.g. Phua et al. (2010), Jullum et al. (2018), Zheng et al. (2006), Aggarwal and Zhai (2012), Satabdi (2018), Min and Jeong (2009), Blanzieri and Bryl (2008), and Tjøstheim (1978)). 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

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

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 2 - 5 f_k will be a density function, whereas we look at the discrete and mixed cases in Sections 6 - 8. 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 probabilties:

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

The classification rule D_B, which is Bayes with respect to R_f, then follows by minimizing R_f, 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 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 needs 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 training data having class k. The estimation of $f_k, k = 1, \ldots, K$, may typically be estimated in a number of different ways, and it is this choice of estimation method that essentially distinguishes different classification methods from each other. The remaining part of the paper shall therefore, to a large extent, concern methods for estimating $f_k, k = 1, \ldots, K$, and the comparison of these, in the discrimination context of (4). 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 examples 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 Fisher (1936), 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.,

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

where $|\cdot|$ denotes the determinant and T the transposed. If $\Sigma_k = \Sigma$ for all k, the Bayes rule in (4) takes the following form (Johnson and Wichern, 2007, 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 hyperplanes 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 (5), 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 simplcity, QDA and LDA have a proven track record in many situations where the class distributions are clearly non-normal (Hastie et al., 2009).

It is, however, crucially important to note here that the QDA and LDA discriminant 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 cannot perform classification based on multivariate normal approximations. Instead, we may consider another reference discriminant, the Naive Bayes. This method, on the other hand, resorts to an approximation that simply ignores any dependence between the variables X_j and X_l, taking the form

$$P_{f(x)}(D = k|X = x) = \prod_{j=1}^{d} P_{f_j(x_j)}(D = k|X_j = x_j). \tag{6}$$

This approximation may work surprisingly well even in situations where property (6) is not satisfied. The marginal distributions in (6) may be estimated parametrically (for instance with a Gaussian distribution) as well as non-parametrically (with e.g. a kernel density estimator), in both cases avoiding the curse of dimensionality.

We will in this paper 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

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

where $\{X^{(k)}(i), i = 1, \dots, n\}$ 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 nonsingular 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 does not work well, see e.g. Silverman (1986, 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 Hall et al. (2004) and Li and Racine (2007). Other nonparametric approaches are nearest neighbour classifiers, see e.g. Samworth (2012) and classification using data depth (Li et al., 2012), 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 litterature provides various other approaches to density estimation, such as the use of mixtures of a parametric and nonparametric approach that may reduce the concequences of the curse of dimensionality, see e.g. Hjort and Glad (1995). To a lesser degree this has also been the case in discrimination, see Chaudhuri et al. (2009), who basically choose a parametric approach, but allows a nonparametric perturbation similar to that of Hjort and Glad (1995). Another such method is the local likelihood estimator proposed by Hjort and Jones (1996) and by Loader (1996), 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 (5). 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 (4). We will pursue both approaches. The local Gaussian approach has been recently used with success in a number of different contexts, see Berentsen and Tjøstheim (2014), Berentsen et al. (2014b), Lacal and Tjøstheim (2017), Lacal and Tjøstheim (2018), Otneim and Tjøstheim (2017), Otneim and Tjøstheim (2018) and Tjøstheim and Hufthammer (2013). R-packages for computing local Gaussian quantities exist, as described by Berentsen et al. (2014a) and Otneim (2018). We will in particular use the local Gaussian density estimation technique as presented by Otneim and Tjøstheim (2017), 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. We extend the idea of describing dependence by means of pairwise relationships to discrete variables in the second part of the paper, whereas mixture of continuous and discrete variables is sought described by a link function and a logistic regression or a GAM type procedure. The discrete case is handled by a successive conditioning argument, in a sense simular to the pair-copula construction described in Aas et al. (2009). We will come back to this in Sections 6 - 8.

The rest of the paper is organized as follows: In Section 2 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 3 and 4. A number of examples in the continuous case are given in Section 5. Section 6 and 7 deal with the purely discrete case and the mixed discrete-continuous case, respectively, with corresponding examples in Section 8. 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 within the neighbourhood N_x of x. 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 $\theta_k(x) = (\mu_{k1}(x), \mu_{k2}(x), \sigma_{k1}^2(x), \sigma_{k2}^2(x), \rho_k(x))$, we write (7) as

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

Moving 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 (7) with $\mu_k(x) \equiv \mu_k$ and $\Sigma_k(x) \equiv \Sigma_k$. Tjøstheim and Hufthammer (2013) discuss non-trivial questions of existence and uniqueness. As the local parameter functions $\mu_k(x)$ and $\Sigma_k(x)$ takes 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 (5) by simply replacing μ_k and Σ_k by $\mu_k(x)$ and $\Sigma_k(x)$ for $k = 1, \ldots, K$. This gives the *local Fisher discriminant*

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

To practically apply this procedure, we need estimates of the involved parameter functions for all class distributions $k = 1, \ldots, K$. Following Hjort and Jones (1996) we estimate the parameters $\mu_k(x)$ and $\Sigma_k(x)$ given data $X(1), \ldots, X(n)$ with class label k, by maximizing the local log likelihood

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

where K_{B_k} is a kernel function depending on a bandwidth paramater (matrix) B_k. We refer to Tjøstheim and Hufthammer (2013) and Otneim and Tjøstheim (2017) for details on parameter estimation.

From the description of the local Gaussian likelihood above, the two discriminants in (8) and (12) below appear to be highly affected by the curse of dimensionality. Otneim and Tjøstheim (2017) 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

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

leading to a pairwise local dependence structure, which can in some way be likened to the additivity assumption in additive regression. 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. 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 (8), and 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 1.

In the left panel of Figure 1 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 (8) essentially reduces to the the global QDA in (5), and we achieve precisely this by choosing a large bandwidth in the estimation of the local parameters in (10) using the local likelihood function in (9). 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 example is completely data driven by means of a cross-validation procedure that we describe in Section 4.

In the second panel of Figure 1 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 seeks to minimize classification error in a certain way that will become clear in Section 4, produces a small smoothing parameter, allowing the local Fisher discriminant (8) to become very local, non-linear and non-quadratic. In both these constructed illustrations the LGDE-based discriminator in (12) is essentially identical to the local Fisher discriminant. This is not always the case, though.

As a by-product of the local likelihood setup and estimation procedure in (9), 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 $\widehat{\Sigma}_k(x)$:

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

These *locally Gaussian density estimates (LGDE)* (Otneim and Tjøstheim, 2017) of the class distributions $f_k(x)$ gives 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 (1) and (4), respectively. This gives the following discriminant:

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

Following the standard recipe of Otneim and Tjøstheim (2017), with the pairwise simplification described above, the estimate $\widehat{f}_{\text{LGDE}}$ involves a further simplification resulting from transforming each variable to approximate standard normality and then fixing $\mu_{j,k}(x) \equiv 0$ and $\sigma_{j,k}(x) \equiv 1$, $j = 1, \ldots, d$. The procedure is especially attractive if the data contain extreme outliers. As it is not quaranteed 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 $\widehat{f}_{\text{LGDE}}$ departs from unit integral mostly depends on the number of variables, and will thus not significatly affect the ratio $\widehat{f}_{\text{LGDE},k}/\widehat{f}_{\text{LGDE},j}$ for two classes k and j. Furthermore, as noted in Section 4, 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.

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 the bandwidth (matrix) $B_k \to 0$. Otneim and Tjøstheim (2017, 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 $\int \widehat{f}_k(x)dx \to 1$.

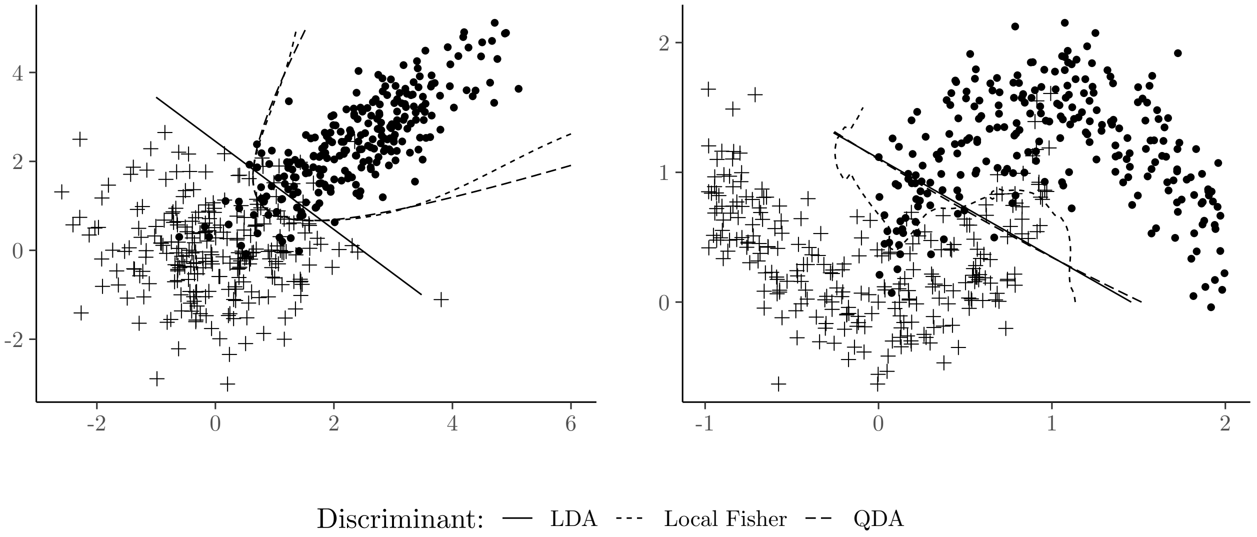

Figure 1: The two-class discrimination problem in two different cases.

3 Some asymptotics of Bayes risk

The Bayes risk, as we have already seen, depends on density functions which may be estimated parametrically or nonparametrically. In the former case 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, $B_k = \text{diag}\{b_{j,k}\}$ 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 by Otneim and Tjøstheim (2017).

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 Marron (1983), which shows that a broad class of nonparametric density estimates (not restricted to kernel density estimates) achieve 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 (3) 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

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

of the Bayes rule, and we are interested in the asymptotic behaviour 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 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 $\sum_i \pi_i = 1$. Marron (1983) studies the mode of convergence of

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

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

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

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 $k = 1, \dots, K$,

(i) $f_k \leq M_k$ on $\mathbb{R}^p$
(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 Marron (1983, Theorem 3), there is a constant $c > 0$ and a density estimator $\widehat{f}_{n,k}$ so that when $r = 2q/(2q + d)$,

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

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 Marron (1983, Theorem 1), then there is a constant $c > 0$ and a classification rule $\widehat{D}_n(x, \pi, T^n)$ so that

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

This 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 Marron (1983) 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 $\widehat{f}_k$ of densities in $\mathcal{F}_k$.

It is easy to find density estimates that satisfy (14). 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. Jones and Signorini (1997) 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 $r = 2q/(2q + d)$. 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 Jones and Signorini, the practical usefulness of higher order kernels is debatable, and a realistic 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 (15) only uses the general convergence property in (14), 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 (14). 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 Otneim and Tjøstheim (2017) and the additional mild conditions (13) 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 Otneim and Tjøstheim (2017) it follows that the variance of the LGDE is of order $(nb^2)^{-1}$. From the log likelihood expression in (9) 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

$$0 = \frac{\partial L_n(\widehat{\theta}, x)}{\partial \theta_j} \overset{P}{\to} \int K_b(y-x) u_j(y, \theta_{b,K}) \Big\{ f(y) - \psi(y, \theta_{b,K}(y)) \Big\} \, \mathrm{d}y$$

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 (9). 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. 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 have 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, condition (iv) of Theorem 4 in Otneim and Tjøstheim (2017) requires $n^{1/2}b^2 \to 0$ which means that $b = n^{-1/6}$ is not a valid choice of bandwidth. The bandwidth b must be of smaller order than $n^{-1/4}$, and this means that the best rate r that can be obtained with the present proof of Theorem 4 in Otneim and Tjøstheim (2017) is $r = n^{-1/2+\varepsilon}$, where $\varepsilon > 0$ can be taken to be arbitrarily small. This leads to the same rate for the Bayes risk because the proof of Theorem 1 in Marron (1983) do not depend on the form of the estimator as soon as the mild conditions for that theorem is fulfilled. It is tempting to conjecture that Theorem 4 of Otneim and Tjøstheim (2017) can be proved without the bound on b used in condition (iv) of that theorem, but such a conjecture remains to be verified. We still have that the rate $n^{-1/2+\varepsilon}$ is independent of d which is a huge advantage compared to the corresponding rate for the kernel estimator.

4 Choice of bandwidth

The preceding section concerns asymptotic results as the size of the training sets grow 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 hyperparameters, and 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. Hart and Vieu (1990) 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.

Berentsen and Tjøstheim (2014) suggest cross-validation as a viable strategy, that Berentsen et al. (2014b), Otneim and Tjøstheim (2017) and Otneim and Tjøstheim (2018) 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 2. 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 Ghosh and Chaudhuri (2004). They examine the missclassification 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 discriminator, 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 (Ghosh and Hall, 2008). Below we present such an optimization algorithm for the case with two classes only.

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 (Fawcett, 2006). The AUC is constructed for two-class classification, but generalisations to $K > 2$ classes exist (Fawcett, 2006, 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 $= 1$ the classifier, all true 1's have higher probabilitiets than all true 0's, enabling perfect classification for some threshold values. 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 5 separate sets, i.e. 5-fold cross validation (Kohavi, 1995). To reduce the search space for the cross validation procedure we limit the bandwidth matrix B_n to diagonal ones 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 of the five folds separately To summarise, we tune the c parameter in b_n for the locally Gaussian discriminants according to the following cross validation procedure:

1. Divide the training set into 5 folds at random.
2. For each proportionality constants c on a specified grid:
    - (a) For each fold $j = 1, \dots, 5$:
        - i. For each class $k = 1, \dots, K$:
            - A. Extract the covariates 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 covariate 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 examples 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 5 folds.

5 Examples

5.1 Simulations

Let us demonstrate some properties of the local Fisher discriminant (8) from a two-class simulation perspective by generating data in increasing dimension d from three different multivariate classification problems that pose increasingly difficult conditions for the traditional discriminants. These problems are:

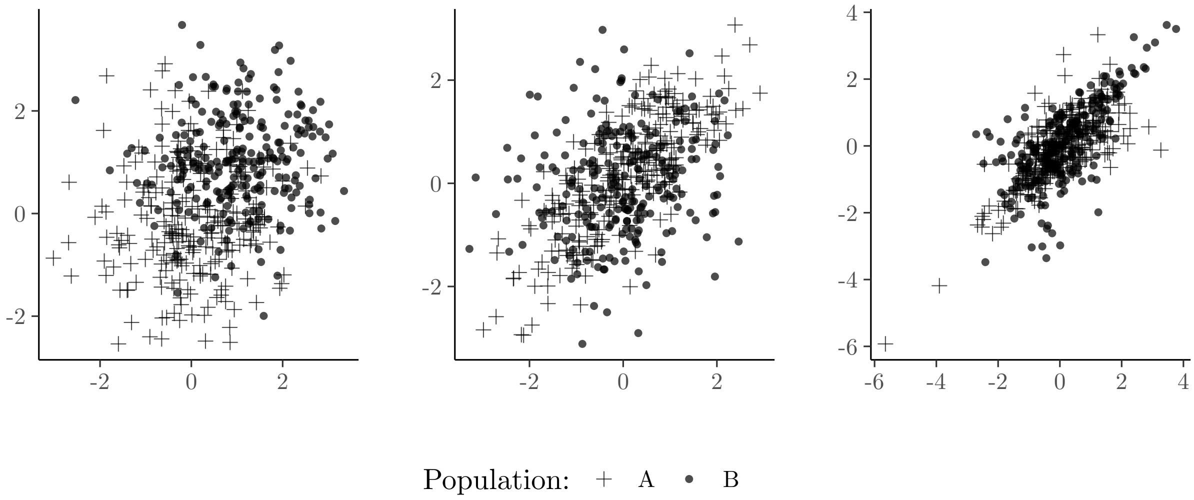

Figure 2: Data from the bivariate versions of the three simulated classification problems.

- **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 $(1, \ldots, 1)^T$.
- **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 (Nelsen, 2007) 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 2.

In all simulated examples 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 4. In addition to the AUC we will also measure the *Brier score* of our predictions (Brier, 1950). The Brier score is essentially the mean squared error of a $0 - 1$-loss classifier, defined for a test data set of size N in the two-class problem with class labels $D = 0, 1$ as

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

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

In Figure 3 we see results for the first example, 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 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. (12) gives very similar results to the local Fisher discriminant in these examples). For the latter three discriminants we choose one bandwidth for each realization based on a 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. The plots report the average AUC and Brier scores for the various discriminants as a function of the number of variables, as well as the standard deviation over the 100 repetitions which we plot as error bars. In each experiment, we evaluate the discrimination using a test data set of size $N = 500$.

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

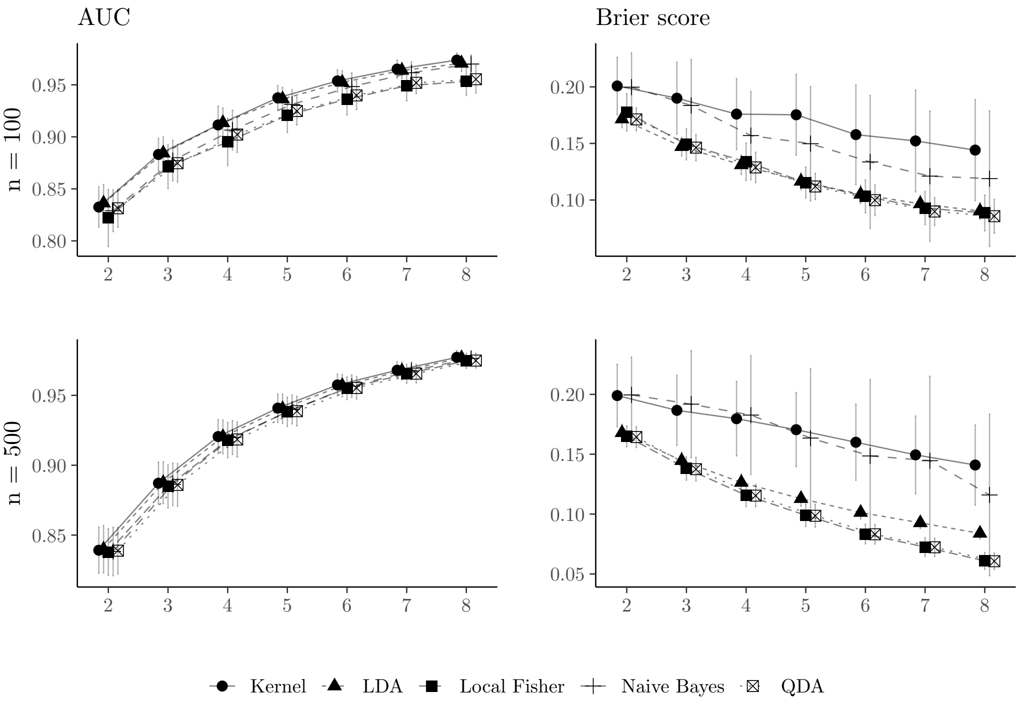

Figure 3: Simulation results for the first example: Two multinormal distributions with different means but equal covariance matrices. Error measured as a function of dimension.

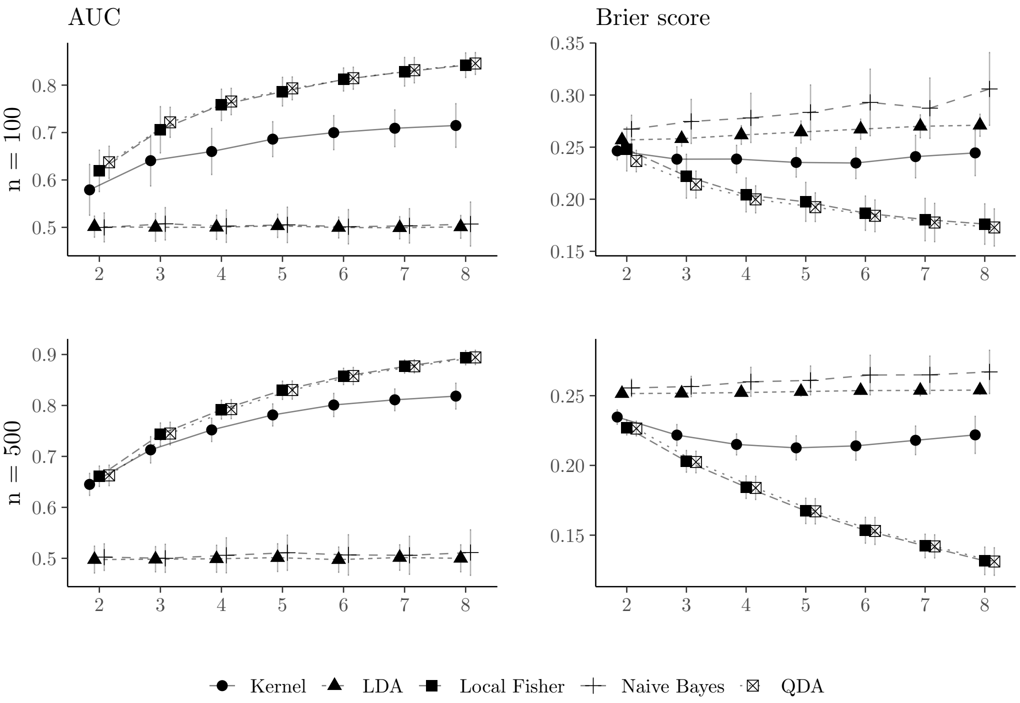

Figure 4: Simulation results for the second example: Two multinormal distributions with different covariance matrices. Error measured as a function of dimension.

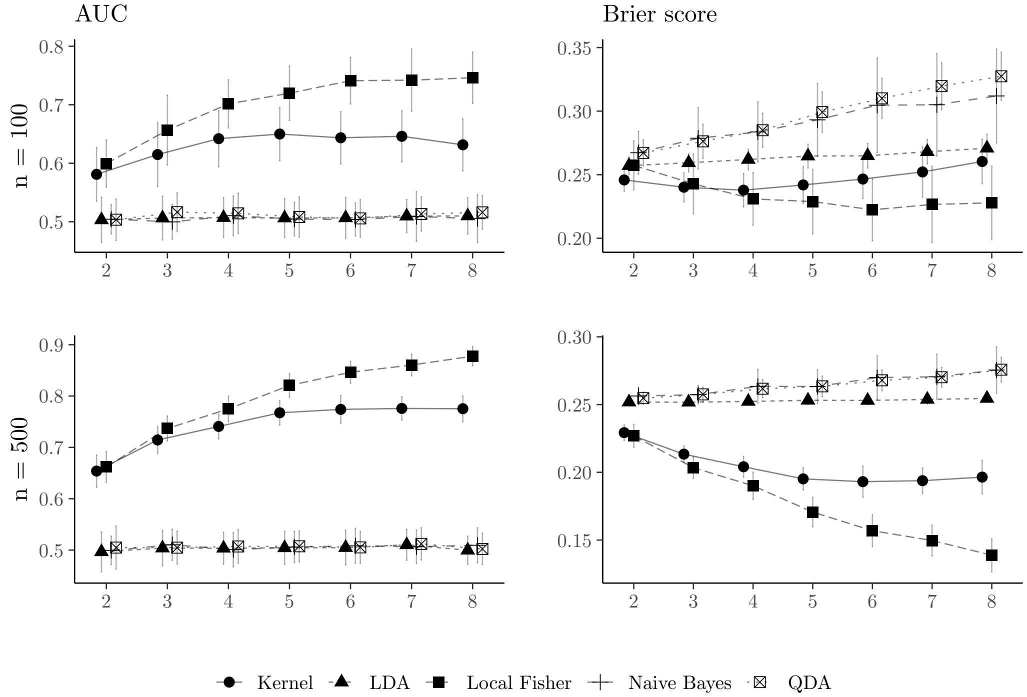

Figure 5: Simulation results for the third example: Two multinormal distributions with different covariance matrices. Error measured as a function of dimension.

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.

We see results from the second example in Figure 4, 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 (a classifier that assigns the posterior probability of 0.5 to all test data in a problem with $\pi_1 = \pi_2 = 0.5$ has a Brier score of 0.25). 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 disciminator 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 example 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 5 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.

5.2 Fraud detection example

Due to the enormous amounts involved, financial crimes such as money laundering is considered a serious threat to societies and economies across the world (Schott, 2006). 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

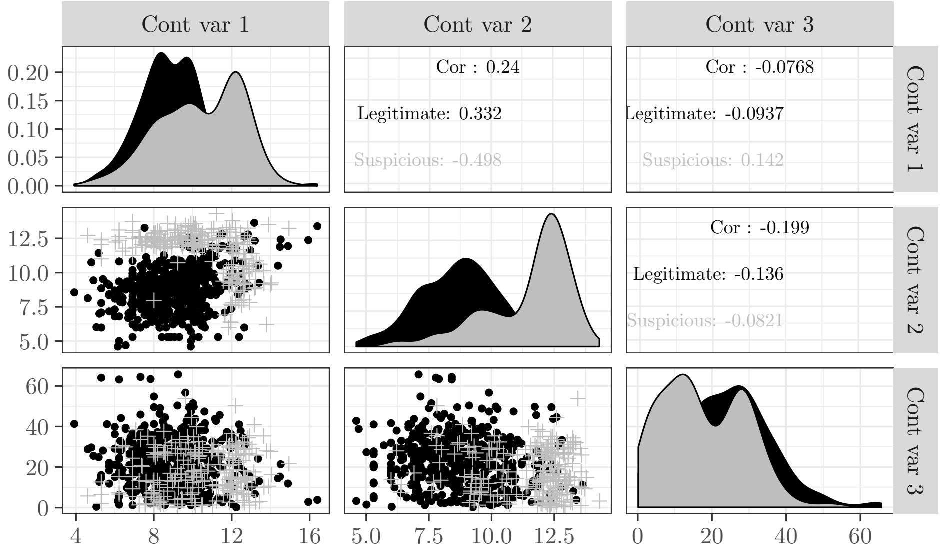

Figure 6: Summary plots for the three continuous variable training data for the money laundering example. Grey (crosses) marks suspicious transactions, while black (dots) marks legitimate ones.

	LGDE	LDA	QDA	Naive Bayes	Kernel	Local Fisher
AUC	0.970	0.917	0.960	0.960	0.960	0.968
Brier	0.058	0.103	0.074	0.063	0.071	0.064

Table 1: Results using the three continuous variables in the money laundering example.

making the source legally liable. In a money laundering setting with a large Norwegian bank, Jullum et al. (2018) develop and train a machine learning model for filtering out suspicious transactions from the legitimate ones. Working with a simplified subset of the their data, both in terms of the transactions we use, and the variables we use for discrimination, we illustrate the use of our local Fisher (and $\widehat{D}_{\mathrm{LGDE}}$) discriminant and compare it to the classical discriminants from the above simulation experiments.

We have used a total of 785 transactions to train the model, of which roughly 28% are marked as suspicious. To check how well our discriminants perform, we use a test set of 226 transactions, and rely on he AUC and Brier scores as in the simulations experiments.[2] To mimic a realistic scenario, all test transactions occur *after* all the transactions in the training set. To simplify this example and illustration of the data, we have restricted ourselves to three continuous variables only. In Section 8.3 we will add more discrete variables to this examples. Due to data restrictions, we can unfortunatly not reveal what the variables we allow the discriminant to use actually are. The data are plotted in Figure 6. 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.

The AUC and Brier score obtained by the various methods are presented in Table 1. As we can see in terms of the AUC, all methods are able to distinguish between the two classes fairly well. The best model in terms of both the AUC and Brier score is the LGDE model. Note however, that with the exception of the LDA model, the differences are not very large between any of the models.

[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 Jullum et al. (2018) for details.

6 Discrete variables: Extending Naive Bayes

We now move from continuous class distributions to discrete class distribution, which is highly relevant in discrimination settings. Discrete variable is a broad term that 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 (1), and carry out the discrimination according to (3). Thus, the rest of this section concerns methods for estimation of such an 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

$$p(x_{1j_1}, \ldots, x_{dj_d}) = P_p(X_1 = x_{1j_1}, \ldots, X_d = x_{dj_d}), \tag{16}$$

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 estimate $p(x_{1j_1}, \ldots, x_{dj_d})$ by the straight forward frequency estimator

$$\widehat{p}_{\text{Frequency}}(x_{1j_1}, \ldots, x_{dj_d}) = \frac{1}{n} \sum_{i=1}^{n} 1(X_1(i) = x_{1j_1}, \ldots, X_d(i) = x_{dj_d}) = n_{1j_1, \ldots, dj_d}/n, \tag{17}$$

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 influential work of Li and Racine (2007; 2008) tackle this problem by a discrete-value smoothing algorithm based on earlier work by Aitchison and Aitken (1976). 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

$$\widehat{p}_{\text{NP}}(x) = \frac{1}{n} \sum_{i=1}^{n} L(X(i), x, \lambda) \tag{18}$$

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, not a change in the values themselves, of these same discrete variables. 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 Hall et al. (2004). The algorithm is implemented in the R-package `np` (Hayfield and Racine, 2008).

6.1 Pairwise Naive Bayes

Contrasting the frequency approach in (17), an obvious and much more radical solution to the problem is to use the Naive Bayes approach where dependence between components is ignored and $p(x_{1j_1}, \dots, x_{dj_d})$ is estmated by

$$\widehat{p}_{\text{Naive Bayes}}(x_{1j_1}, \dots, x_{dj_d}) = \Pi_{r=1}^{d} \widehat{p}_{\text{Frequency}}(x_{rj_r}) = \Pi_{r=1}^{d} n_{rj_r}/n. \tag{19}$$

Except for certain very rare cases, this apporach automatically avoids the problem of empty cells. As this approach has the obvious drawback that all dependence between the variables are 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 (2017; 2018) we will try to achive this by deriving an estimator for (16) using solely marginals $p_{rj_r} = P(X_r = x_{rj_r})$ with $j_r = 1, \dots, k_r$ and bivariate probabilities $p_{rj_r,sj_s} = P(X_r = x_{rj_r}, X_s = x_{sj_s})$ with $j_r = 1, \dots, k_r$ and $j_s = 1, \dots, k_s$. 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 pairwise Naive Bayes approach uses a construction which in some sense is similar to the pair-copula construction, see e.g. Aas et al. (2009). More precisely, when a pair of variables is conditioned on a 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 $p(x_{lj_l}, \dots, x_{dj_d})$ and $p_{m|l\cdots d}$ instead of $p(x_{mj_m}|x_{lj_l}, \dots, x_{dj_d}) = P(X_m = x_{mj_m}|X_l = x_{lj_l}, \dots, X_d = x_{dj_d})$. 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 $p(x_{lj_l}, x_{mj_m}|x_{uj_u}, \dots, x_{dj_d}) = P(X_l = x_{lj_l}, X_m = x_{mj_m}|X_u = x_{uj_u}, \dots, X_d = x_{dj_d})$, and assuming that the dimension d is an even number:

$$p_{1\cdots d} = p_{12|3\cdots d}p_{3\cdots d} = p_{12|3\cdots d}p_{34|5\cdots d} \cdots p_{d-1,d} \tag{20}$$

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 a different result. (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 (20) can be done in 6 different ways

$$p_{1234} = \begin{cases} p_{12|34}p_{34} \approx p_{12}p_{34} \\ p_{13|24}p_{24} \approx p_{13}p_{24} \\ p_{14|23}p_{23} \approx p_{14}p_{13} \\ p_{23|14}p_{14} \approx p_{23}p_{14} \\ p_{24|13}p_{13} \approx p_{24}p_{13} \\ p_{34|12}p_{12} \approx p_{34}p_{12}. \end{cases} \tag{21}$$

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

$$\widehat{p}_{1234} = \left(\widehat{p}_{12} \cdot \widehat{p}_{34} \cdot \widehat{p}_{13} \cdot \widehat{p}_{24} \cdot \widehat{p}_{14} \cdot \widehat{p}_{13} \cdot \widehat{p}_{23} \cdot \widehat{p}_{14} \cdot \widehat{p}_{24} \cdot \widehat{p}_{13} \cdot \widehat{p}_{34} \cdot \widehat{p}_{12}\right)^{1/6}, \tag{22}$$

where $\widehat{p}_{lm}$ is used as a shorthand notation for $\widehat{p}_{\text{Frequency}}(x_{lj_l}, x_{mj_m})$. The factors in (22) are identical in pairs, and taking this into account, (22) reduces to

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

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 (20), 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

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

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

Denote each decomposition by $g_1, \dots, g_R$. In the general version of (21) - (23), there are $R \cdot (d/2)$ factors in total, but there are only $d(d-1)/2$ pairs and thus unique factors after the approximtion (after we drop the conditioning). The number of times each factor occurs, then, is equal to

$$S = \frac{\text{No. of lines like those in eq. (21)} \times \text{No. of factors in each}}{\text{No. of unique factors}} = \frac{\frac{d!}{2^{d/2}} \cdot \frac{d}{2}}{\frac{d(d-1)}{2}} = \frac{d(d-2)!}{2^{d/2}}.$$

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

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

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

$$\begin{aligned}\widehat{p}_{\text{Pairwise Naive Bayes, even}}(x_{1j_1}, \dots, x_{dj_d}) &= \left(\prod_{j=1}^{R} g_j\right)^{1/R} = \left(\prod_{l<j\leq d} \widehat{p}_{jl}^{S}\right)^{1/R} \\ &= \left(\prod_{l<j\leq d} \widehat{p}_{jl}^{\frac{d(d-2)!}{2^{d/2}}}\right)^{\frac{2^{d/2}}{d!}} = \left(\prod_{l<j\leq d} \widehat{p}_{jl}\right)^{\frac{1}{d-1}}.\end{aligned} \tag{24}$$

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 (23) 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

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

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 contain 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, the Naive Bayes formula.

It is important to realize that, unlike the Naive Bayes, the pairwise approximations in (24) and (25) need not be proper probability distributions, i.e. they may not sum to 1. To arrive at proper probability estimators, one must normalize:

$$
\begin{aligned}
\widehat{p}_{\text{Pairwise Naive Bayes, even}}(x_{1j_1},\dots,x_{dj_d}) &= \frac{\widehat{p}'_{\text{Pairwise Naive Bayes, even}}(x_{1j_1},\dots,x_{dj_d})}{\sum_{l_1=1}^{k_1}+\cdots+\sum_{l_d=1}^{k_d}\widehat{p}'_{\text{Pairwise Naive Bayes, even}}(x_{1l_1},\dots,x_{dl_d})}, \\
\widehat{p}_{\text{Pairwise Naive Bayes, odd}}(x_{1j_1},\dots,x_{dj_d}) &= \frac{\widehat{p}'_{\text{Pairwise Naive Bayes, odd}}(x_{1j_1},\dots,x_{dj_d})}{\sum_{l_1=1}^{k_1}+\cdots+\sum_{l_d=1}^{k_d}\widehat{p}'_{\text{Pairwise Naive Bayes, odd}}(x_{1l_1},\dots,x_{dl_d})},
\end{aligned}
\tag{26}
$$

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 (26) achieves the usual root-n asymptotic normality property when compared to respectively $p_{\text{Pairwise, odd}}$ 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. Note that the comparison quantities $p_{\text{Pairwise, odd}}$ and $p_{\text{Pairwise, odd}}$. 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 obervations» 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 favourite procedure for modelling 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, in many situations, this independence assumption could be considered to be too drastic.

Our take on this is to take dependence between continuous and discrete variables into consideration by first modelling the continuous variables with the LGDE approach in Section 2, and then conditioned on the continuous variables set up a logistic, log linear or even generalized additive model (GAM), cf. Hastie and Tibshirani (1990). 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

$$\phi(p_{ir_r,sj_s}) = \beta_0^{rj_r,sj_s} + \sum_{j=1}^{d_c} \beta_j^{rj_r,sj_s} x_j^c \tag{27}$$

or additively as

$$\phi(p_{rj_r,sj_s}) = h_0^{rj_r,sj_s} + \sum_{j=1}^{d_c} h_j^{rj_r,sj_s}(x_j^c). \tag{28}$$

The unknown β parameters can be estimated by maximum likelihood using a GLM software package, and in the additive case the h_i-s can be estimated by a GAM software package [**HO: need proper referencing here. I also want to cite all other R-packages that we used in a separate section before the reference list. Agree? If so: Håkon will fix**]. Note that if the dimension of x^c is large, which is likely in e.g. fraud applications, then one may considered (ridge or lasso type of) regularized logistic regression (Hastie et al., 2009, 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 modelled directly in the same manner as $\phi(p_{rj_r,sj_s})$. In the training phase this should be done separately for the 2 (or 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 (unnormalised) probability $\widehat{p}'_{\text{Pairwise Naive Bayes, k}}(x^d_{1j_1}, \ldots, x^d_{dj_d}|x^c)$ using the procedure of Section 6.1. The pairwise estimator of the class distributions for mixed data is finally completed by multiplying with the estimate of the continuous density, i.e.:

$$\widehat{f}_{\text{Pairwise, mixed}, k} = \widehat{p}'_{\text{Pairwise Naive Bayes, k}}(x^d_{1j_1}, \ldots, x^d_{dj_d}|x^c)\widehat{f}_{\text{LGDE},k}(x^c). \tag{29}$$

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 equation (4).

8 Examples 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 5.1, 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:

- **Problem 1:** We define define two continuous populations both being marginally standard normal, but having two different dependence structures defined by the Clayton copula (Nelsen, 2007) 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.
- **Problem 2:** We complicate the discrimination task between the populations in 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 examples, 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 discriminators:

1. Estimate $p_{1\ldots,d}$ using empirical frequencies as in (17), proceed via Bayes formula (1), and carry out the discrimination according to eq. (3).
2. Calculate conditional class probabilities directly using the smoothing algorithm in (18), implemented in the `np`-package.
3. Estimate $p_{1\ldots,d}$ using our pairwise probability approximation in (24) and (25), proceed via Bayes formula (1), and carry out the discrimination according to (3).

We evaluate the discriminators using the AUC and Brier scores as we did in Section 5.1.

Consult Figure 7 for the results of Problem 1. We have allowed the dimension of the problem to range from 2-12 because the calculations needed in the discrete case is much lighter than the continuous case, computationally speaking. 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.

We present the results of Problem 2 in Figure 8, 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 **Problem 1** in the preceding subsection as follows:

- **Mixed Problem**: We generate continuous variables in the same way as in **Problem 1** in Section 5.1. 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.

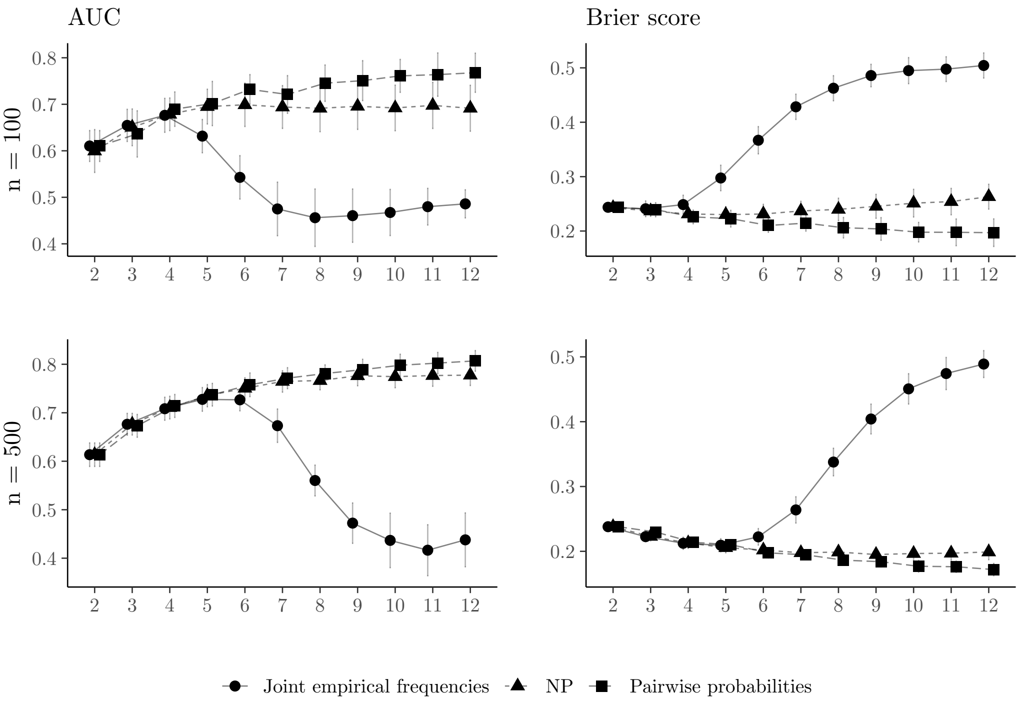

Figure 7: Simulation results for discrete example, problem 2: Two discretisized Clayton populations with having weak and strong dependence, respectively.

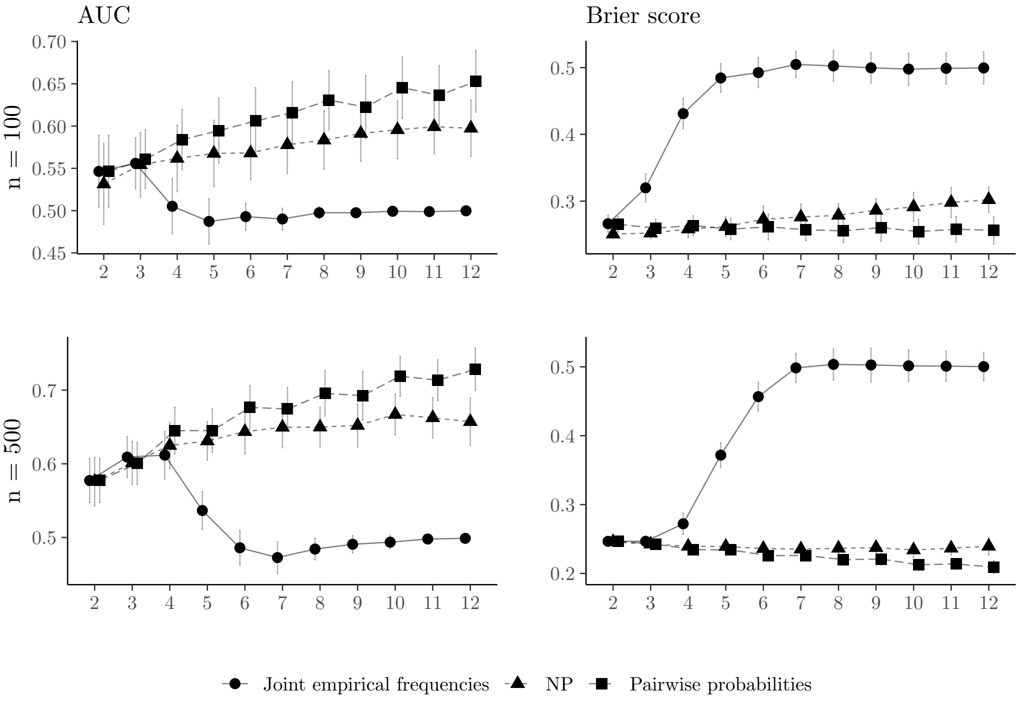

Figure 8: Simulation results for discrete example, problem 2: Two three-category discretized Clayton populations with having weak and not-as-strong dependence, respectively.

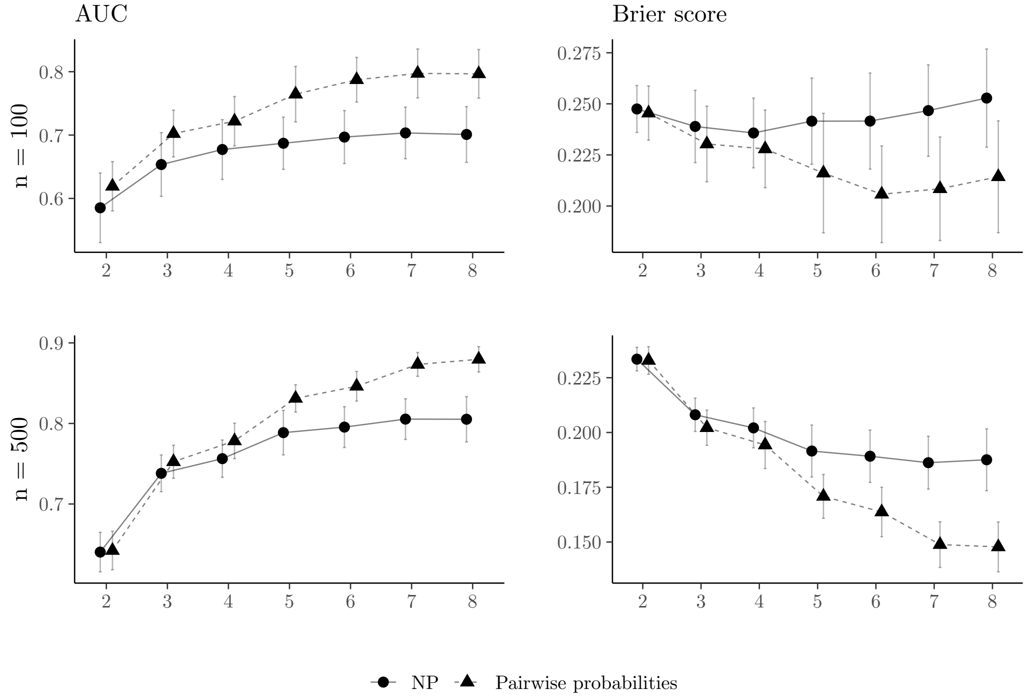

Figure 9: Simulation results for the mixed example: Two Clayton populations where every other variable is discretized.

We 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, but given our findings in earlier examples, 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 Li and Racine (2008) method for computing conditional probabilities directly.
2. Our pairwise procedure for combining locally Gaussian density estimates with the pairwise frequency approach as in (29), using the logistic regression (27) or the generalized additive model (28).

The results in the mixed case are presented in Figure 9. We have used the logistic regression 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. We must note here though that the Li and Racine (2008) method for estimating conditional probabilities is not tuned specifically towards discrimination.

8.3 Fraud detection example

In this section we build further on the money laundering example in Section 5.2, 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 2. 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 5.2. 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 rely on the AUC and Brier score here as well.

Disc var 1	# Suspicious	# Legitimate
Category 1	146	30
Category 2	1	28
Category 3	25	335
Category 4	8	8
Category 5	9	59
Category 6	14	46
Category 7	23	53

Disc var 4	# Suspicious	# Legitimate
Category 1	64	179
Category 2	154	367
Category 3	0	6
Category 4	5	3
Category 5	3	4

Disc var 2	# Suspicious	# Legitimate
Category 1	203	556
Category 2	23	3

Disc var 5	# Suspicious	# Legitimate
Category 1	5	19
Category 2	221	540

Disc var 3	# Suspicious	# Legitimate
Category 1	29	70
Category 2	197	489

Disc var 6	# Suspicious	# Legitimate
Category 1	7	10
Category 2	12	4
Category 3	202	535
Category 4	5	10

Disc var 7	# Suspicious	# Legitimate
Category 1	13	67
Category 2	213	492

Table 2: Share of training observations in the different categories for the seven discrete variables in the money laundering example.

	Joint empirical frequencies	NP	Pairwise probabilities
AUC	0.876	0.866	0.882
Brier	0.102	0.097	0.097

Table 3: Results using the seven discrete variables in the money laundering example.

8.3.1 Discrete variables only

For illustrational purposes, we first allow the discriminators to use the seven discrete variables only. The performance results from the various discriminators on the test set are in Table 3. As seen from the table, our pairwise probability based approach is the best model in terms of AUC, tying the first place with the NP approach for the Brier score.

8.3.2 Mixed variables

Now we allow discriminators to use both the three continuous variables and the seven discrete variables. The performance results for the three discriminators accessible in this setting are shown in Table 4. As seen from the table, the NP approach is the best in terms of AUC. In terms of the Brier score, the NP model shares the first place with the pairwise probability based approach using GLM to model dependence between the discrete and continuous variables. One possible reason that the GAM based version of the pairwise

	NP	Pairwise probabilities (GLM)	Pairwise probabilities (GAM)
AUC	0.987	0.982	0.938
Brier	0.044	0.044	0.079

Table 4: Results using two continuous variables and four discrete variables in the money laundering example.

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 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 a nonparametric discriminant based on the kernel density estimator in the continuous case, and NP-filtered probability estimator considered by Li and Racine (2008) in the discrete and mixed distribution case. Our experiments show significant improvements compared to the two classic discriminants, and also good performance results compared to the nonparametric alternatives.

There is a substatial potential for further research and modifications. For instance, we have ignored the normalization issue in computing 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 possibly is a variant of the NP-filtering of Li and Racine (2008) applied to the initial pairwise probabilities. Bagging and boosting (Hastie et al., 2009) 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 discriminaor in every situation, but merely to naturally 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 algoritmic methods in the machine learning society, 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 utilise the data differently than those methods, and therefore bring something new to the table. If this is indeed the case, combining different flavoured discriminants, for instance by methods as in Ranjan and Gneiting (2010), seems like a promising approach.

10 Software

- References to the R-packages that we have used.
- Note on code. The easiest way out might be to zip the `simulation_experiments` folder together with the `.Rmd`-file and ship it along with the paper. Not outstandingly user-friendly though as it stands today.

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.

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.

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

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.

Jullum, M., Løland, A., Huseby, R. B., Ånonsen, G., and Lorentzen, J. P. (2018). Detecting money laundering transactions – which transactions should we learn from? Submitted.

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.

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*.

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.

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.

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.