AUSTRIAN JOURNAL OF STATISTICS
Volume 33 (2004), Number 3, 267-279

Kernel Density Estimation: Theory and Application in Discriminant Analysis

Thomas Ledl
Department of Statistics and Decision Support Systems,
University of Vienna, Austria

Abstract: Nowadays, one can find a huge set of methods to estimate the density function of a random variable nonparametrically. Since the first version of the most elementary nonparametric density estimator (the histogram) researchers produced a vast amount of ideas especially corresponding to the issue of choosing the bandwidth parameter in a kernel density estimator model. To focus not only on a descriptive application, the model seems to be quite suitable for application in discriminant analysis, where (multivariate) class densities are the basis for the assignment of a vector to a given class. This article gives insight to most popular bandwidth parameter selectors as well as to the performance of the kernel density estimator as a classification method compared to the classical linear and quadratic discriminant analysis, respectively. Both a direct estimation in a multivariate space as well as an application of the concept to marginal *normalizations* of the single variables will be taken into consideration. From this report the gap between theory and application is going to be pointed out.

Zusammenfassung: Heutzutage existieren zahlreiche Methoden die Dichtefunktion einer Zufallsvariablen nichtparametrisch zu schätzen. Seit dem ersten der elementarsten nichtparametrischen Dichteschätzer, dem Histogramm, wurden viele interessante Konzepte besonders für die Wahl der Bandbreite im Modell des Kerndichteschätzers ausgearbeitet. Neben einer deskriptiven Anwendung scheint sich dieses Modell a priori auch gut für eine Anwendung in der Diskriminanzanalyse zu eignen, wo die Schätzung von Klassendichten die Basis für die Zuordnung eines Objektes (Beobachtungsvektors) zu einer Gruppe ist. Neben der Diskussion dieser Probleme gibt dieser Artikel auch einen Einblick, wie sich dieses Konzept im Vergleich zur klassischen linearen bzw. quadratischen Diskriminanzanalyse, welche beide auf parametrischen Normalverteilungsschätzungen beruhen, profiliert. Es wird hierbei sowohl auf die Möglichkeit einer direkten multivariaten Kerndichteschätzung, als auch auf die Anwendung des Konzeptes für die *Normalisierung* der einzelnen Variablen eingegangen. Anhand dieser Darstellung soll auch die bei höherdimensionalen Daten auftretende große Kluft zwischen Theorie und Anwendung aufgezeigt werden.

Keywords: Multivariate Density Estimation, Bandwidth Parameter, Kernel Discriminant Analysis, Classification Performance.

1 Introduction

Since non-parametric smoothing methods provide an interesting alternative to classical parametric estimation methods, this paper is concerned with the genesis of kernel density estimation itself (for descriptive means) as well as at its application in discriminant analysis to serve as a competitor in estimating class densities with respect to the corresponding model-based discrimination rules, which are the linear and the quadratic discriminant analysis (LDA, QDA). Starting at the first description of the kernel density estimation concept by Rosenblatt (1956)

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) , \tag{1}$$

where $K(\cdot)$ denotes the kernel function (including some appropriate restrictions, e.g. $\int K(u)du = 1$, $K(u) \geq 0$ for all u, etc.) and h is the so called bandwidth parameter. Many proposals about how to select the bandwidth appropriate have been made. For the kernel function the most commonly used density is $N(\mu, \sigma^2)$, but also several other kernel densities with bounded support are used. Section 2 gives an overview about the improvement in techniques and strategies for optimizing K and h with respect to different optimization criteria during the last 20 to 25 years. Results are available for the univariate and the multivariate model. Besides, the issue of handling the data in high(er) dimensions to still keep the model as competitor to LDA and QDA, respectively for discriminatory purposes, is treated. Here, another problem of optimization arises as well as the so called *curse of dimensionality* which hinders the user to generalize the concept easily without having sufficient data. Finally, in Section 3 the reader will see *hard numbers*, where the theoretical ideas are applied to different types of datasets by means of a detailed simulation study, where altogether 14 estimators are used for 21 different datasets.

2 Ideas in Kernel Density Estimation and Techniques for Application in the Discrimination Context

2.1 The Univariate Setting

To classify a kernel density estimation $\hat{f}_h(\cdot)$ having specified kernel K and bandwidth h, as *well estimated* one has to create some kind of measure of deviation to the underlying original density $f(.)$. A straightforward idea is to take the integrated difference

$$L_p = \left(\int |\hat{f}_h(x) - f(x)|^p \right)^{1/p} \tag{2}$$

(usually $p = 1$, $p = 2$, $p \to \infty$) into consideration. Setting $p = 2$ here leads to the easiest to calculate results. For $p = 1$, Devroye and Györfi (1985) offer several optimality results for bandwidths concerning different kernel shapes.

Nevertheless, in most cases the easier results for $p = 2$ are used. As the construction in (2) is a random variable (the so called integrated squared error ISE), it is useful to take

the expectation

$$MISE(\hat{f}_h) = \int E\{\hat{f}_h(x) - f(x)\}^2 dx \tag{3}$$

into account. Jones (1991) provides an overview of what is better to use. Again, to circumvent difficulties, a Taylor approximation of (3) (called AMISE) is often used. Marron and Wand (1992) show, that those step-by-step adaptations (AMISE instead of MISE with $p = 2$ instead of $p = 1$ in (2) for reasons of easier calculation) can cause essential changes in the resulting estimated model parameter (in particular the optimal bandwidth). One has to be aware of the fact, that the extent of estimating the tails of the distribution well, decreases as p increases, which leads to the fact, that $p \to \infty$ stresses a good *overall fit* and does not care about a bad fit in the tails. Besides, one can think of other criteria like comparing the number (and location) of the estimated modes to the numbers (and location) of the modes of the original density (see Park and Turlach, 1992), but one cannot carry out useful calculations without knowing the original density in application. Marron and Tsybakov (1995) go even further and suggest to include a kind of horizontal distance between the curves for reasons of a more intuitive fit, as well. Nevertheless, the only useful possibility to derive automatically data-driven parameters in application is to handle a compact formula similar as in (2). Stressing now on the AMISE, Wand and Jones (1995) derive the following formula which is a decomposition of a bias and variance term:

$$AMISE(\hat{f}_h) = \frac{1}{nh} R(K) + \frac{1}{4} h^4 \mu_2(K)^2 R(f'') \,, \tag{4}$$

where $\mu_2(K) = \int z^2 K(z) dz$ denotes the variance of the kernel and $R(K) = \int K(x)^2 dx$. After separating K as in (4), the *Epanechnikow kernel* (see, e.g. Silverman, 1986) minimizes the AMISE, but the choice of the kernel is not that crucial, because even for the triangular and the uniform kernel, there are less than 10% additional data points necessary to get the same AMISE. Small improvements are possible by leaving the restriction $K(x) \geq 0$ for all x, which amounts in higher order kernels (Wand and Jones, 1995) leading to values $\hat{f}_h(x) < 0$ for some x, which are not adequate in the estimation of class densities in discriminant analysis. The more important bandwidth choice is carried out by minimizing (4) with respect to h, which leads to

$$h_{AMISE} = \left[\frac{R(K)}{\mu_2(K)^2 R(f'') n} \right]^{1/5} . \tag{5}$$

Before considering ideas and techniques for a concrete estimation of h_{AMISE} (the unknown density f has to be substituted) we discuss two further ways of improving the concept. First it is possible to smooth the density with different bandwidths, depending on a distance $d_{j,k}$ of the datapoint x_j to its k-th nearest neighbor (Breiman et al., 1977)

$$\hat{f}_h(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h d_{j,k}} \,, \tag{6}$$

where the curve is more radically smoothed in sparse regions, which is useful in highly skewed distributions (see Sheather, 1992, p. 246, for an example of failure of the standard model). The problem of choosing another parameter k, which is, again more complicated

in high dimensions does not qualify this model as fully data-driven and is therefore not suitable for unexperienced users. Another concept is the so-called *transformed kernel density estimator*. Its aim is to transform the observations into others, which results in easier-to-estimate densities (densities, which are much more effective to estimate with respect to the observation number). Wand and Jones (1995) give a parametric concept of a transformation rule, which again amounts in at least one parameter to be estimated by the unexperienced user, while Ruppert and Cline (1994) provide a calculation-intensive non-parametric approach that was used in the simulation study described below.

They refer to the fact, that if F and G are the cumulative distribution functions for the densities f and g, then $Y = G^{-1}(F(X))$ has density g. One is now able to choose any distribution, but the user will probably choose a normal distribution, because normalization is wanted. In application $t(x) = G^{-1}(\hat{F}_h)$ is taken and $\hat{F}_h(x)$ is a (pilot-) kernel estimate of $F(x)$. The issue of bandwidth selection in the univariate case is discussed detailed in literature. Authors offer many proposals to apply (5) in practice, starting by using a Gaussian kernel $N(0, \sigma^2)$ for K and replacing f'' in (5) by the respective functional of $N(0, \sigma^2)$ as well. This *rule of thumb* or *normal rule* was suggested by Silverman (1986) and amounts in the formula for the AMISE-optimal bandwidth

$$h_{opt} \approx 1.06 \sigma n^{-1/5} \,,$$

which is easy to calculate and therefore often used. σ can be estimated by the sample standard deviation or by a more robust estimate as the inter-quartile-range R (and a corresponding adaptation of the coefficient). However, it often oversmooths the density. Another idea, which is well-known in statistics is the concept of cross-validation. Bowman (1984) states the optimization problem by giving unbiased estimators for the minimization of the $ISE(\hat{f}_h)$ itself. In Section 3 the corresponding formula for the multivariate case is given. This estimator is known as *least-square cross-validation*. *Biased cross-validation* is similar, but minimizes $MISE(\hat{f}_h$. Both, *biased-* and *least-square cross-validation* have the drawback of high variances of the estimators and sometimes the occurrence of more than one local minimum. Sheather (1992) and Marron in his discussion of Sheather (1992) give different explanations about which one is the best to take. Finally, the *likelihood cross-validation* method (Silverman, 1986) leads to problems, if kernels having bounded support are used.

An interesting set of methods are the *plug-in methods* (see Sheather and Jones, 1991, or Park and Marron, 1990), which represent the *state of the art* in selection of the bandwidth parameter as it seems, because many simulation studies identify them as the best or at least one of the best estimators with respect to an intuitive fit, the variance of the estimator and the performance in the estimation of harder-to estimate densities (Jones et al., 1996; Sheather, 1992; Park and Turlach, 1992; Cao et al., 1994). Common to all plug-in methods is, that they include an estimate of the unknown density functional $R(f'')$in (5), which is performed by a kernel estimate, but in general with another kernel and smoothing parameter as for estimating f. This makes it different to the Biased cross-validation concept. Besides, regarding the asymptotic analysis of the bandwidth selectors, the plug-in approach has a faster convergence rate for h than the cross-validation methods.

2.2 The Multivariate Setting

When generalizing the univariate model up to d dimensions one gets

$$\hat{f_{\mathbf{H}}}(\mathbf{x}) = \frac{1}{n}|\mathbf{H}|^{-1/2} \sum_{i=1}^{n} K\left(\mathbf{H}^{-1/2}(\mathbf{x} - \mathbf{x}_i)\right) ,$$

where $\mathbf{H}$ is a symmetric positive definite $d \times d$-matrix, the bandwidth matrix, and K is a multivariate kernel satisfying $\int K(\mathbf{x})dx = 1$ (Wand and Jones, 1995). Regarding the problem of choosing one single parameter in the univariate model the problem of estimating not only d parameters (for d dimensions), but d^2 parameters (indicating the direction of smoothing) arises. Here, one can think of either one parameter h controlling $\mathbf{H} = h^2\mathbf{I}$ or parameterizations with d or d^2 parameters. At this point it should be mentioned, that there is much calculation time required for the selection of more than one parameter, regarding algorithms, which are more complicated than the *normal rule*. Therefore the easiest parametrization was chosen in the simulation study. For reasons of large variances in models having more parameters this choice is a good compromise. Essentially, the same ideas as in the univariate setting can be generalized, but the application suffers from exhaustive computation time and mathematical tractability (e.g. for difficulties in the generalization of the plug-in estimator see Wand and Jones, 1994). However, the crucial problem is the so-called *curse of dimensionality* (see Silverman, 1986, or Scott, 1992), which is based the fact, that there is, roughly speaking, *much more space* in high dimensions and hence (strongly) increasing numbers of observations are required to satisfy a constant estimation accuracy (see again Silverman, 1986). Since the calculation time for estimators in high dimensions is even higher (d times) for *equal* observation numbers, the problem becomes transparent.

2.3 The Context to Discriminant Analysis

An important fact is, that concerning the discrimination context the minimization of error rates in classification has not necessarily the same aim as the minimization of error criteria discussed in section 2.1. Since the theoretical *misclassification rate* is an L_1-based measure, a MISE-optimal bandwidth selector weights the fit in the tails of the distributions to a smaller extent. Actually, in higher dimensions an increasing part of the data *disappears* in the tails of the distribution. For that reason, a fit in the tails is demanded in classification tasks and Ripley (1996) underlines, that the estimation of differences of log-densities is crucial in classification. Hall and Wand (1988) treat the topic of at least estimating differences in densities (without logarithm), however, this is ignored in any other source. They use kernel functions having negative values and the multivariate generalization cannot be derived easily.

3 Simulation Study and Results

3.1 Preliminaries

The existing results for using kernel density estimation as a classification method go back to the late 1970s and provide a very limited view of the problem. Ness and Simpson (1976), Ness (1980), Habbema et al. (1978), and Remme et al. (1980) essentially treat uncorrelated multivariate normal distributions in high dimensions, in which groups were only separated by one variable. From the viewpoint of a *curse of dimensionality* improvements can only happen accidently and the essential issue of applying the concept to non-normal data is not taken into account. Besides, the parameters have been estimated with the *likelihood cross-validation* method, which is somewhat out-dated and often results in oversmoothing. All Studies used two classes and there was also no reason to change this in our simulation study, because the basic problem will probably be the same in a setting with more than two classes.

3.2 Construction of the Datasets

The first effect to study is the behavior of the model when the densities gradually move away from the normal distribution. In addition, skewed and bimodal distributions are interesting. Table 1 and Figure 1 show the univariate prototype distributions used and exhibit different types of deviations from the normal case. Table 2 lists the construction principles for the dataset based on univariate prototype distributions. Each dataset has dimension 1400×10 and consists of two classes, each with 700 observations, 600 for estimating the class density and 100 for calculating the classification criteria. Table 1 and Figure 1 show the distributions of class 1. For class 2 the parameters of the exponential distribution change from $\lambda = 1$ to $\lambda = 2$ and the normals and bimodals are shifted by 0.5 to the right.

After this *linking* step, these ten datasets (1-10) have been transformed linearly by ten 10×10-matrices, which are the roots of ten self-produced correlation-matrices. The datasets 11-20 have been produced exactly in the same way as datasets 1-10, but population 1 and population 2 have been transformed by different transformation-matrices. Concerning Table 2 and Figures 2-5 the datasets having equal covariance matrices have *1* as their last digit, the others have *2*. For example, the dataset *Bi42* consists originally of eight skewed distributions and two bimodals (whose bumps are strongly separated), and the transformation happened for both groups with unequal transformation matrices. The 30 ($10 + 2 \times 10$) correlation matrices have been produced by assuming a common factor in the ten variables having a regression coefficient, whose absolute value is uniformly distributed between 0.3 and 1. Finally an insurance dataset having the same dimensions like the synthetic data was used (dataset 21).

Table 1: Prototype distributions of the synthetic datasets.

Name	**Construction**	**Means**
Normal	$N(0,1)$	
Normal with "small noise"	$0.8N(0,1) + 0.2\frac{\sum_{i=1}^{25}\sqrt{\phi(\mu_i)}N(\mu_i, 0.1^2)}{25\sum_{i=1}^{25}\sqrt{\phi(\mu_i)}}$	$\mu_1 = -3$; μ_i, $i = 2,\ldots,25$, is created by adding stepwise uniform [0, 1/2] random variables
Normal with "medium noise"	$0.7N(0,1) + 0.3\frac{\sum_{i=1}^{13}\sqrt[4]{\phi(\mu_i)}N(\mu_i, 0.2^2)}{13\sum_{i=1}^{13}\sqrt[4]{\phi(\mu_i)}}$	$\mu_1 = -3$; μ_i, $i = 2,\ldots,13$, is created by adding stepwise uniform [0, 1/2] random variables
Normal with "large noise"	$0.5N(0,1)+0.5\frac{\sum_{i=1}^{7}\sqrt[4]{\phi(\mu_i)}N(\mu_i, 0.35^2)}{7\sum_{i=1}^{7}\sqrt[4]{\phi(\mu_i)}}$	$\mu_1 = -3$; μ_i, $i = 2,\ldots,7$, is created by adding stepwise uniform [0, 2] random variables
Exponential	Exp(1)	
Bimodal "close"	$0.5N(0,1) + 0.5N(2.5,1)$	
Bimodal "far"	$0.5N(0,1) + 0.5N(5,1)$	

Table 2: Description of the datasets used.

Dataset	**Abbrev.**	**Contents**
1	NN1	10 normal distributions with "small noise"
2	NN2	10 normal distributions with "medium noise"
3	NN3	10 normal distributions with "large noise"
4	SkN1	2 skewed (exp-)distributions and 8 normals
5	SkN2	5 skewed (exp-)distributions and 5 normals
6	SkN3	7 skewed (exp-)distributions and 3 normals
7	Bi1	4 normals, 4 skewed and 2 bimodal (close)-dist.
8	Bi2	4 normals, 4 skewed and 2 bimodal (far)-dist.
9	Bi3	8 skewed and 2 bimodal (close)-dist.
10	Bi4	8 skewed and 2 bimodal (far)-dist.

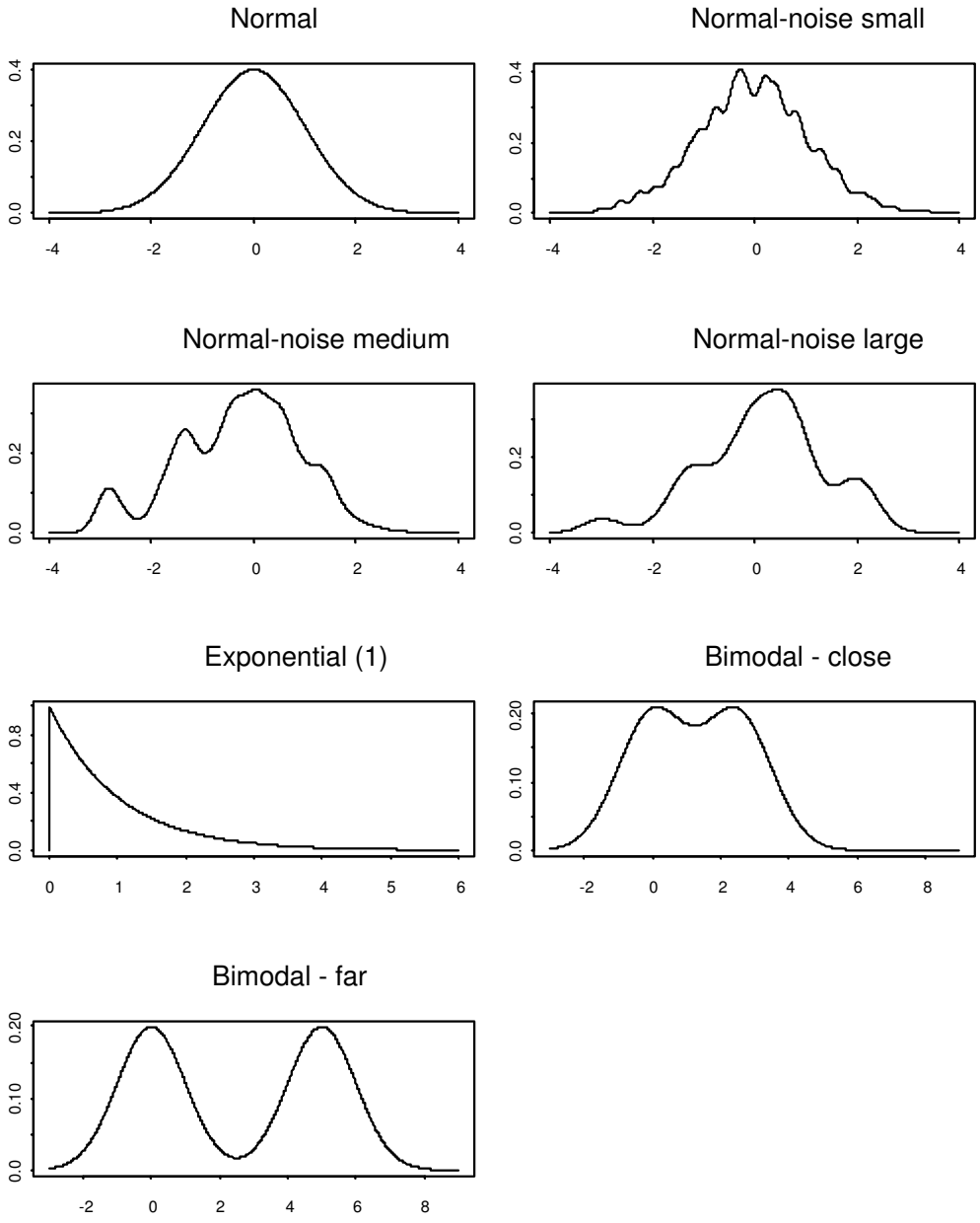

Figure 1: Prototype distributions of the synthetic datasets.

3.3 Construction of the Estimators

As mentioned above, the simulation study uses 14 estimators. First of all the LDA and QDA serve as *conservative* competitors. The multivariate densities were estimated by the multivariate *normal rule* (Scott, 1992) and the multivariate *least-squares cross-validation* (LSCV) selector (Bowman, 1984), which is given by the minimum of $LSCV(\mathbf{H})$ in (7)

with respect to h ($\mathbf{H} = h^2\mathbf{I}$).

$$LSCV(\mathbf{H}) = \frac{1}{n-1}N(0, 2\mathbf{H}) + \frac{n-2}{n(n-1)^2}\sum_{i\neq j} N(x_i - x_j, 2\mathbf{H}) - \frac{2}{n(n-1)}\sum_{i\neq j} N(x_i - x_j, \mathbf{H})\,. \tag{7}$$

For this reason, the original datasets have been projected classwise by a principal component analysis (PCA) onto a subspace consisting of 2 to 5 dimensions, respectively and were estimated by both selectors, normal rule and LSCV leading to estimators 3-10. By doing that a trade-off between the information loss caused by the PCA and the accuracy loss for the kernel density estimators caused by additional dimensions can be confronted with each other.

The issue of using marginal normalizations amount in the estimators 11-14. They are constructed as described in Subsection 2.1 for each of the ten variables in each dataset. Here, the univariate plug-in bandwidth selector of Sheather and Jones (1991) and the *normal rule* (Silverman, 1986) was applied to normalize the datasets and in a subsequent step, the LDA and QDA were used for the transformed ten dimensional distributions, resulting in the $2 \times 2 = 4$ last estimation procedures.

3.4 The Performance Measure

To evaluate the goodness of classification, the classical *error rate* (percentage of wrong classified elements) and the *Brier score* (Hand, 1997) were considered. The later is a similar, but less robust measure and is given by

$$BS = \frac{2}{n}\sum_{i=1}^{n}\Big(\hat{p}(\text{Group 2}|\mathbf{x}_i) - c_i\Big)^2\,,$$

where $c_i = 0$ if $\mathbf{x}_i$ is from group 1 and $c_i = 1$ otherwise. Good discrimination is indicated when both measures are small. The following results refer to the Brier score only.

3.5 Results

One of the most important results is that the bandwidth selection procedure was not crucial. As the kernel choice for descriptive applications is not that important, the bandwidth parameter is not for discrimination purposes. The *Sheather-Jones selector* performed similar to the *normal rule* in the univariate normalizations and the *LSCV selector* resembled the *normal rule* in the multivariate setting, and so using the normal rule, which extremely saves computation time is completely sufficient.

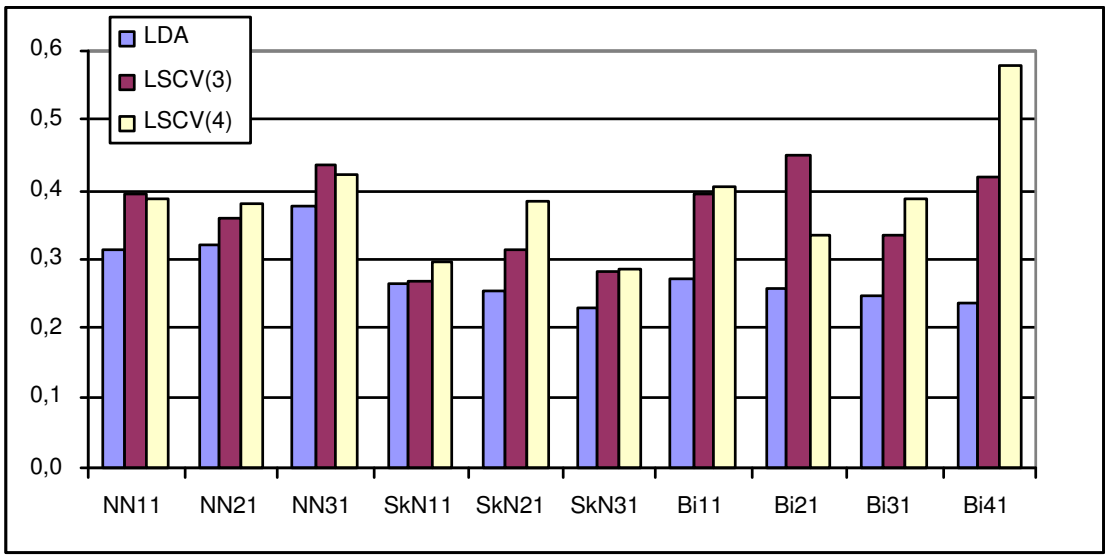

Figure 2: Brier-score of the datasets having equal correlation matrices. Comparison between the LDA and the Bayes-rule-kernel-methods constructed by the LSCV-selector.

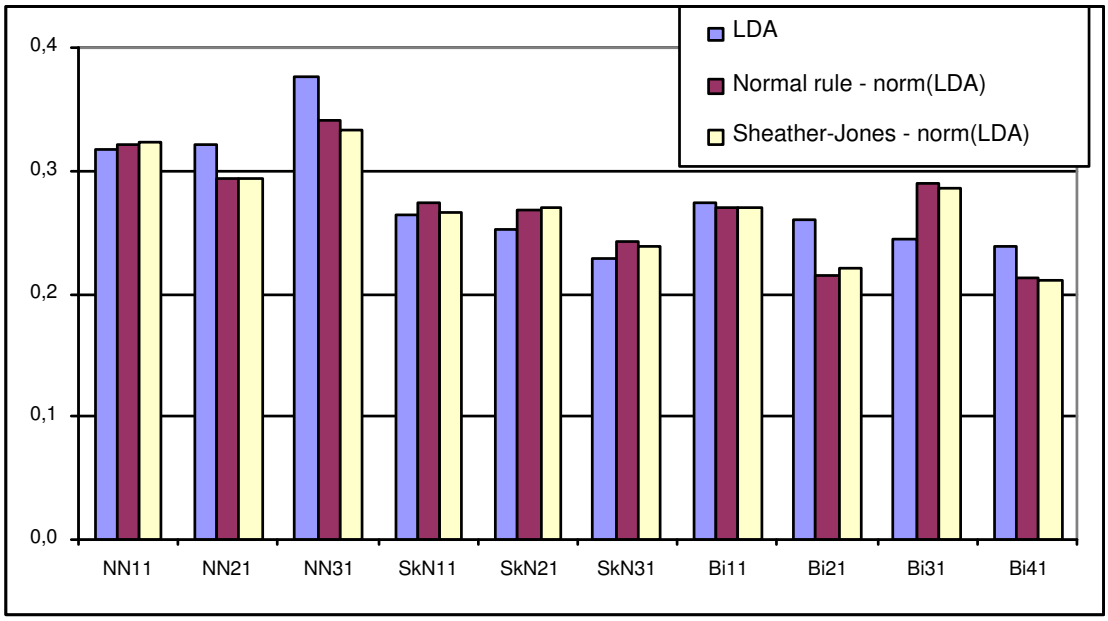

Figure 3: Brier-score of the datasets having equal correlation matrices. Comparison between the LDA before and after univariate normalizations derived by kernel estimates.

The better performance of LDA compared to QDA in the case of equal covariance matrices was also obvious after marginal normalizations by the kernel method and vice versa in case of unequal covariance matrices. The most interesting result appeared in the comparison of the two main estimation techniques.

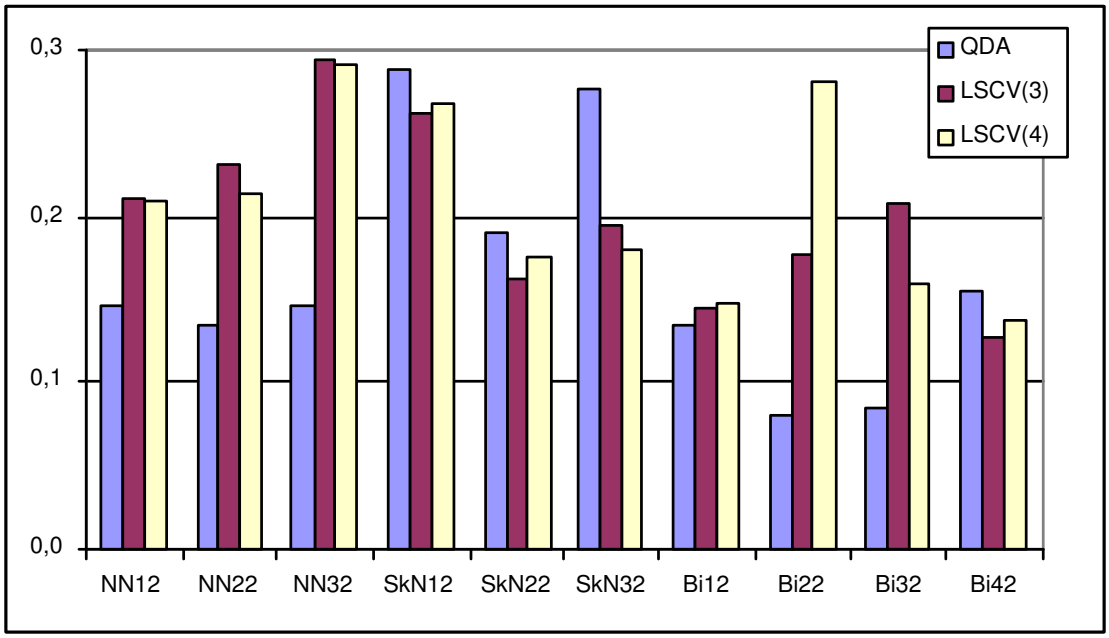

Figure 4: Brier-score of the datasets having unequal correlation matrices. Comparison between the QDA and the Bayes-rule-kernel-methods constructed by the LSCV-selector.

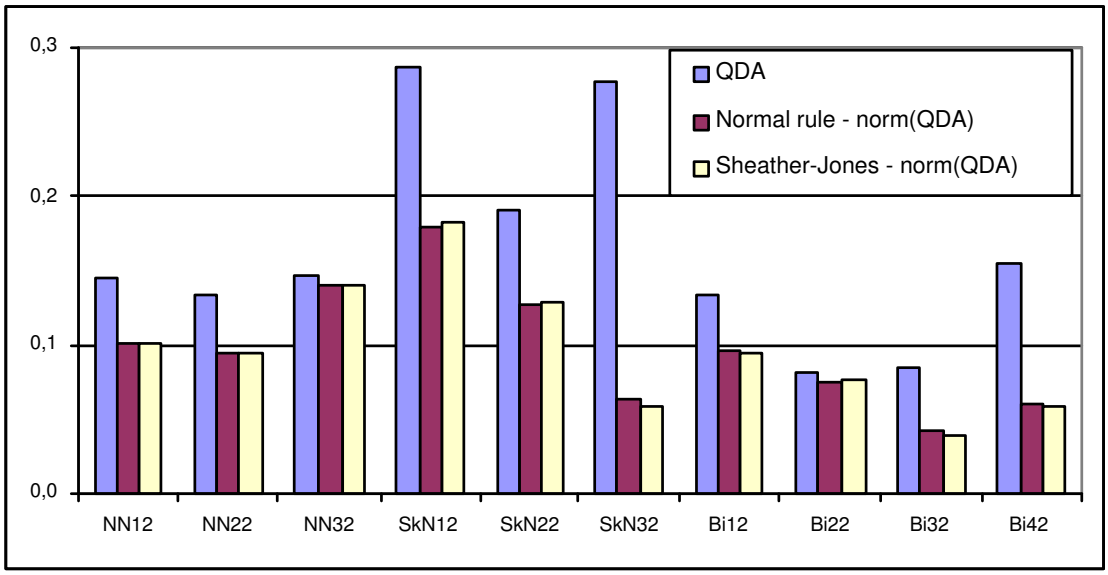

Figure 5: Brier-score of the datasets having unequal correlation matrices. Comparison between the QDA before and after univariate normalizations derived by kernel estimates.

The rates of the multivariate kernel estimators compared to the classical methods, LDA (in Figure 2 for equal correlation matrices) and QDA (in Figure 4 for unequal correlation matrices), are quite disappointing and the euphoria of the simulation studies in the past (e.g., Remme et al., 1980) are from this point of view not comprehensible. The direct density estimation in 2 to 5 dimensions had a poor performance compared to their parametric counterparts (LDA and QDA). Within those non-parametric estimators, the projection to 3 and 4 dimensions performed better than those to 2 and 5, but all in all this trial failed.

The LDA is *the best* in all cases and the kernel concepts are quite bad for datasets, where the assumption of multivariate normal distribution has to be rejected. This is really interesting, since the (normal-distribution-based) LDA should actually lose its advantage for those datasets. Figure 4 does not qualify the kernel based LSCV-method as a superior competitor to QDA case of non-equal covariance matrices, either. The univariate normalizations which are, however calculation intensive, led especially in the case of non-equal covariance matrices to considerable improvements (Figure 5). For equal covariance matrices (see Figure 3) the strain of calculating kernel smoothers appears to be not necessary, since (possible) improvements seem to happen accidently.

The classification for the insurance dataset failed for all selectors, since the classes did not differ much and some of the distributions were highly skewed.

References

Bowman, A. W. (1984). An alternative method of cross-validation for the smoothing of density estimates. *Biometrika*, *71*, 353-360.

Breiman, L., Meisel, W., and Purcell, E. (1977). Variable kernel estimates of multivariate densities. *Technometrics*, *19*, 135-144.

Cao, R., Cuevas, A., and Manteiga, W. (1994). A comparative study of several smoothing methods in density estimation. *Computational Statistics and Data Analalysis*, *17*, 153-176.

Devroye, L., and Györfi, L. (1985). *Nonparametric Density Estimation: The L_1 View*. New York: John Wiley.

Habbema, J. D. F., Hermans, J., and Remme, J. (1978). Variable kernel density estimation in discriminant analysis. In *Proceedings in Computational Statistics* (p. 178-185). Physica Verlag Wien.

Hall, P., and Wand, M. P. (1988). The plug-in bandwidth selection. *Biometrika*, *75*, 541-547.

Hand, D. J. (1997). *Construction and Assessment of Classification Rules*. Chicester: John Wiley & Sons.

Jones, M. C. (1991). The roles of ISE and MISE in density estimation. *Statistical Probability Letters*, *12*, 51-56.

Jones, M. C., Marron, J. S., and Sheather, S. J. (1996). A brief survey of bandwidth selection for density estimation. *Journal of the American Statistical Association*, *91*, 401-407.

Marron, J. S., and Tsybakov, A. B. (1995). Visual error criteria for qualitative smoothing. *Journal of the American Statistical Association*, *90*, 499-507.

Marron, J. S., and Wand, M. P. (1992). Exact mean integrated squared error. *The Annals of Statistics*, *20*, 712-736.

Ness, J. V. (1980). On the dominance of non-parametric Bayes rule discriminant algorithms in high dimensions. *Pattern Recognition*, *12*, 355-368.

Ness, J. W. V., and Simpson, C. (1976). On the effects of dimension in discriminant analysis. *Technometrics*, *18*, 175-187.

Park, B. U., and Marron, J. S. (1990). Comparison of data-driven bandwidth selectors. *Journal of the American Statistical Association, 85*, 66-72.

Park, B. U., and Turlach, B. (1992). Practical performance of several data-driven bandwidth selectors (with discussion). *Computational Statistics, 7*, 251-285.

Remme, J., Habbema, J. D. F., and Hermans, J. (1980). A simulative comparison of linear, quadratic and kernel discrimination. *Journal of Statistical Computation and Simulation, 11*, 87-106.

Ripley, B. D. (1996). *Pattern Recognition and Neural Networks*. Cambridge: Cambridge University Press.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics, 27*, 832-837.

Ruppert, D., and Cline, D. B. H. (1994). Transformation kernel density estimation – bias reduction by empirical transformations. *Annals of Statistics, 22*, 185-210.

Scott, D. W. (1992). *Multivariate Density Estimation: Theory, Practice and Visualization*. New York: Wiley.

Sheather, S. J. (1992). The performance of six popular bandwidth selection methods on some real datasets (with discussion). *Computational Statistics, 7*, 225-281.

Sheather, S. J., and Jones, M. C. (1991). A reliable data-based bandwidth-selection method for kernel density estimation. *Journal of the Royal Statistical Society, Series B, 53*, 683-690.

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

Wand, M. P., and Jones, M. C. (1994). Multivariate plug-in bandwidth selection. *Computational Statistics, 9*, 97-117.

Wand, M. P., and Jones, M. C. (1995). *Kernel Smoothing*. London: Chapman and Hall.

Author's address:

Mag. Thomas Ledl
Department of Statistics and Decision Support Systems
University of Vienna
Universitätsstr. 5/9
A-1010 Vienna
Austria

E-mail: thomas.ledl@univie.ac.at
http://www.univie.ac.at/statistics/