High dimensional classification when useful information comes from many, perhaps all features

Arindam Chatterjee
Peter Hall

High dimensional classification when useful information comes from many, perhaps all features

Arindam Chatterjee[*]   Peter Hall[†]

Abstract

In the analysis of high-dimensional data it is common to reduce dimension from thousands or tens of thousands to a much smaller number, often between 5 and 20. One reason for such a substantial reduction is to reduce the conceptual difficulty of the problem. This difficulty highlights the need for models that permit a small number of features to provide the majority of information available for classification, but allow a much larger number, indeed potentially all features, to supply the remaining information that is needed for a higher level of performance. Inference in such cases is almost bound to involve significantly nonlinear aspects. In this paper we suggest approaches of this type, based on empirical approximations to Bayes rule classifiers and involving adaptive feature selection to optimise performance. This intrinsically nonlinear approach enables the methodology to exploit any interactions among features that might enhance classifier accuracy. The methodology is sequential, and involves steadily building a model of increasing complexity, stopping when an empirical measure of error indicates that further complexity would only degrade performance.

KEYWORDS. Bayes rule, density estimation, feature selection, kernel methods, likelihood ratio, variable selection.
SHORT TITLE. High dimensional classification.

---

[*]Stat-Math Unit, Indian Statistical Institute, New Delhi 110016, India. `cha@isid.ac.in`
[†]Department of Mathematics and Statistics, The University of Melbourne, VIC 3010, Australia. `halpstat@ms.unimelb.edu.au`

1 Introduction

Contemporary methods for feature selection in high-dimensional data analysis are generally based on assumptions of linearity. For example, if a response variable, Y, is observed in the continuum then it is argued that the dependence of Y on a p-vector X can be expressed through a linear predictive model, $Y = \alpha + \beta^{\mathrm{T}} X + \text{error}$, where α denotes a scalar and β is an unknown p-vector. A major step in reducing dimension involves setting almost all the values of β equal to zero, for example by imposing an L_1 penalty as in the case of the Lasso (Tibshirani (1996); Zou (2006); Meinshausen (2007)). A similar approach is used when the response variable is a zero-one class label, L. There the probability that the class label equals zero is taken to be expressed by a logit-linear model:

$$P(L=0) = \left\{ 1 + \exp\left(\alpha + \beta^{\mathrm{T}} X \right) \right\}^{-1}, \quad P(L=1) = 1 - P(L=0) , \tag{1.1}$$

and β can again be dimension-reduced by applying an L_1 penalty.

Typically, p is decreased from a value in the thousands or tens of thousands to one between about 5 and 20 by following this approach. However, the massive dimension reduction here reflects a degree of pragmatism, rather than a firm conclusion that a model such as that at (1.1) is the most appropriate. In particular, we generally do not believe that the distribution of the class label depends on only a small number of the components of X. More likely it depends at least partly on many more, but on pragmatic grounds we focus on a relatively small number. If we were to attempt to build a classifier where the dependence of the class label L, or the response variable Y, on X involved more than a few features, then the linear model would likely be inadequate, since in relatively high dimensional cases the manner of dependence can be quite complex.

In some problems of this type it is in theory possible to construct a classifier so that it enjoys especially strong performance. Suppose, for example, that datasets $\mathcal{X}_j = \{X_{j1}, \ldots, X_{jn_j}\}$ are drawn from respective populations Π_j, for $j = 1, 2$, where each data vector $X_{ji} = (X_{ji1}, \ldots, X_{jip})$ is of length p and has probability density f_j, say. Assume too, only for the sake of simple exposition, that the vector components $X_{ji1}, \ldots, X_{jip}$ are completely independent, and have a common, symmetric density ϕ when $j = 1$ but have the respective densities $\phi(\cdot + \nu_1), \ldots, \phi(\cdot + \nu_p)$ when $j = 2$, where $\nu_1, \ldots, \nu_p$ are constants. (Therefore Π_1 and Π_2 are distinguished from one another by component-wise translation.) Then, if $\sum_k \nu_k^2 = \infty$, the Bayes-rule classifier can asymptotically distinguish perfectly between Π_1 and Π_2, as $p \to \infty$, but not if $\sum_k \nu_k^2 < \infty$. Details will be given in section 2.2.

Of course, practical, empirical classifiers can be constructed to give the same high level of performance when $\sum_k \nu_k^2 = \infty$. In this case a great deal of information about the differences between Π_1 and Π_2 accumulates as dimension grows. However, in many practical problems this high degree of accuracy is unreasonable, and we expect to encounter much less than perfect classification. This is illustrated by the case $\sum_k \nu_k^2 < \infty$, which, in broader terms than those described by the above examples, is the subject of the present paper.

Motivated by examples such as these we develop methods for classification in problems where a small number of components is sufficient for reasonably effective classification, but a larger number is required for optimal or near-optimal performance. Our approach is based on a bottom-up, rather than

top-down, algorithm; that is, we look first for the single most effective component, and steadily adjoin new components until an estimator of classification error fails to decrease. Most competing algorithms operate in the reverse direction, steadily reducing the number of components.

Our approach is nonlinear, in that it avoids models such as that at (1.1) for the class-label distribution, and instead uses empirical approximations to Bayes-rule procedures, based on projections and nonparametric density estimators. This technique enables the methodology to capitalise on interactions among features that might enhance classification accuracy.

There is a vast literature on statistical methods for classification. As well as linear techniques, discussed above but also including contributions by, among many others, Donoho and Elad (2003), Donoho (2006) and Candes and Tao (2007), it involves approaches founded on thresholding (e.g. Fan and Li (2001), the elastic net (Zou and Hastie (2005)), covariance regularisation (e.g. Bickel and Levina (2008); Witten and Tibshirani (2009)), sure independence screening (Fan and Lv (2008)) and nonlinear covariance ranking (Hall and Miller (2009)). Cross-validation methods for estimating error rates, similar to approaches proposed in section 3, include those suggested by Efron (1983), Efron and Tibshirani (1997) and Ghosh and Hall (2008). Techniques based on projection pursuit and related ideas (see e.g. Friedman and Tukey (1974); Friedman (1987)) can be used to reduce the variability of classification methods, and comprise part of the methodology that we propose. Empirical approaches to classification via the Bayes rule include those discussed by Krzyżak (1991), Lapko (1993), Pawlak (1993), Devroye et al. (1996), Ancukiewicz (1998) and Hall and Kang (2005). There are several book-length treatments of classification and related problems, for example those given by Duda et al. (2001), Hastie et al. (2009) and Shakhnarovich et al. (2006). Dudoit et al. (2002) discuss the performance of a variety of classifiers.

2 Bayes rule motivation

2.1 Bayes classifier as a benchmark

We observe data in the form of independent p-vectors X_{ji}, for $1 \le i \le n_j$ and $j = 1, 2$, where X_{ji} comes from population Π_j. The subscript $j = L(X_{ji})$ denotes the (class) label of any given data vector X_{ji} in the sample $\mathcal{X}_j = \{X_{j1}, \ldots, X_{jn_j}\}$. We focus on cases where the n_j's are much less than p, and where the samples $\mathcal{X}_1$ and $\mathcal{X}_2$ are used to construct a classifier, Cl say. Given a new p-vector, Z say, the classifier determines that $\mathrm{Cl}(Z) = 1$ if Z was assessed by Cl to have come from Π_1, i.e. to have label $L(Z) = 1$.

The ideal Bayes classifier $\mathrm{Cl}_{\mathrm{Bayes}}$, which minimises error rate, concludes that Z comes from Π_1 if $g_{\mathrm{Bayes}}(Z) > \frac{1}{2}$, where

$$g_{\mathrm{Bayes}}(x) = P\{L(Z) = 1 \mid Z = z\} = \frac{\pi f_1(x)}{\pi f_1(x) + (1 - \pi) f_2(x)} \ , \tag{2.1}$$

f_j denotes the probability density of data drawn from Π_j, and π is the prior probability of Π_1. In particular, $g_{\mathrm{Bayes}}(z)$ equals the probability that Z comes from Π_1, given that $Z = z$. The Bayes risk, or error rate, is given by:

$$\mathrm{errate}_{\mathrm{Bayes}} \equiv P\{\mathrm{Cl}_{\mathrm{Bayes}}(Z) \ne L(Z)\} = \pi \int_{z: g_{\mathrm{Bayes}}(z) \le \frac{1}{2}} f_1(z) \ dz + (1 - \pi) \int_{z: g_{\mathrm{Bayes}}(z) > \frac{1}{2}} f_2(z) \ dz \ , \tag{2.2}$$

We formalise the definition of the distribution of Z by taking the distribution to be a mixture: Z is drawn from Π_1 with probability π and from Π_2 with probability $1 - \pi$.

The Bayes classifier is impractical for very high-dimensional data, because we neither know the densities f_j nor have much opportunity to estimate them from the relatively small samples that are typically available. However, this is not a serious drawback for the sorts of problems we have in mind, discussed in section 1 and in section 2.2 below.

2.2   Motivating examples

In this section we expand on the elementary example introduced in section 1, making the simplifying assumption that the prior probabilities π and $1-\pi$ equal $\frac{1}{2}$ and that the components X_{jik}, for $1 \le k \le p$, of the p-vector $X_{ji} = (X_{ji1}, \ldots, X_{jip})$, are independent with respective density functions ϕ_{jk}. Further, we take the corresponding distributions to differ only in terms of location, and define $\phi_{1k}(u) = \phi(u)$ and $\phi_{2k}(u) = \phi(u + \nu_k)$ for each k, where ϕ is a fixed density and the ν_k's are constants. This case is useful because its simplicity enables a particularly transparent illustration of the general context of our work.

If we knew the population densities f_1 and f_2 then we would assign a new data value Z to Π_1 if $g_{\mathrm{Bayes}}(Z) > \frac{1}{2}$, where g_{Bayes} is as defined at (2.1). Equivalently, we would assert that Z came from Π_1 if $\log\{f_1(Z)/f_2(Z)\} > 0$. Now, when $Z = (Z_1, \ldots, Z_p)$,

$$\log\{f_1(Z)/f_2(Z)\} = \sum_{k=1}^{p} \left\{\log \phi(Z_k) - \log \phi(Z_k + \nu_k)\right\} , \tag{2.3}$$

If the location differences ν_k are fixed and nonzero, or converge to a nonzero limit, then the right-hand side of (2.3) diverges in probability to $+\infty$ as p increases, if Z comes from Π_1, and diverges to $-\infty$ otherwise. In this case we achieve asymptotically perfect classification, and of course we can construct practical, empirical methods that do the same, without knowing f_1 and f_2. (For example, the empirical Bayes approach suggested in sections 3.1 and 3.3 achieves this outcome.)

Perfect classification also holds in cases where ν_k converges to zero but the series $\sum_k \nu_k^2$ diverges:

$$\sum_{k=1}^{\infty} \nu_k^2 = \infty , \tag{2.4}$$

Indeed, under mild conditions that are satisfied for most standard distributions, it is shown in Appendix A.1 that if Z is drawn from Π_1, and W has the distribution of a general component of Z, then

$$E\{\log \phi(W) - \log \phi(W + \nu)\} = \tfrac{1}{2}\, \nu^2\, E\{\phi'(W)/\phi(W)\}^2 + o(\nu^2) \tag{2.5}$$

as $\nu \to 0$. An argument based on the law of large numbers can now be used to show that if (2.4) holds and Z comes from Π_1 then the series on the right-hand side of (2.3) diverges to $+\infty$ with probability 1, and hence that $P\{g_{\mathrm{Bayes}}(Z) > \frac{1}{2}\} \to 1$. Similarly, if Z comes from Π_2 then $P\{g_{\mathrm{Bayes}}(Z) < \frac{1}{2}\} \to 1$. Therefore, asymptotically perfect classification again prevails, and a practical, empirical classifier can be constructed to achieve the same result without knowing the densities f_1 and f_2.

4

In the present paper we focus on broader, more challenging settings, exemplified by but of course not restricted to the independent component model where (2.4) fails, and where asymptotically perfect classification is not possible. In the case of independent components when Π_1 and Π_2 are distinguished only by location changes, the case we consider is characterised by the assumption

$$\sum_{k=1}^{\infty} \nu_k^2 < \infty \,. \tag{2.6}$$

Here, under conditions (A.1) and (A.2) in Appendix A.1, the series at (2.3) converges with probability 1 as $p \to \infty$, no matter whether Z comes from Π_1 or Π_2. It can be proved from this property that the limit of the probability that the Bayes rule classifier makes an error, when classifying data that are equally likely to come from Π_1 and Π_2, equals

$$\tfrac{1}{2} \left[P\{S_1(\infty) \leq 0\} + P\{S_2(\infty) \geq 0\} \right], \tag{2.7}$$

where

$$S_1(q) = \sum_{k=1}^{q} \{\log \phi(W_k) - \log \phi(W_k + \nu_k)\}, \quad S_2(q) = \sum_{k=1}^{q} \{\log \phi(W_k - \nu_k) - \log \phi(W_k)\}$$

and the random variables $W_1, W_2, \ldots$ are independent and identically distributed with density ϕ. (Recall that we are assuming that both prior probabilities equal $\tfrac{1}{2}$.) Of course, the Bayes rule minimises error rate over all possible classifiers, and so no other approach can beat the rate given by (2.7).

This example has at least two important implications. First, even if we know f_1 and f_2, the probability of misclassification does not necessarily converge to zero as p (and sample size) diverge. In particular, allowing both sample size and dimension to be very large does not necessarily result in a substantial reduction in error rate, since the amount of information that accumulates through increasing dimension can be relatively small, although strictly positive. Secondly, if we do not know f_1 and f_2, but can consistently estimate the joint q-variate densities of data from Π_1 and Π_2 where $q < p$, then we can construct a classifier for which the error rate is $\tfrac{1}{2} \left[\{P(S_1(q) \leq 0\} + P\{S_2(q) \geq 0\} \right]$ (compare (2.7)), and this can be made as close as we like to the minimum error rate at (2.7) by choosing q large. This setting and its generalisations comprise the context of this paper. We show that in particularly difficult problems we can get arbitrarily close to the minimum error rate by constructing approximations, of finite but increasingly high dimension, to the much higher dimensional, but unknowable, densities of the actual data vectors.

Although the independent-component case is very specialised, its considerable simplicity helps to convey intuition. We shall use it again in section 5.1, where we shall show in different respects that, for large q, the ordering of components in terms of decreasing values of $|\nu_k|$ minimises classification error for a given value of q.

Generalisations and extensions are also possible. For example, restricting attention to the independent component case for simplicity, the location change in the model can be altered to a scale change, where the definition $\phi_{2k}(u) = \phi(u + \nu_k)$ is replaced by $\phi_{2k}(u) = \phi(\sigma_k u) \, \sigma_k$, the constants σ_k satisfy $\sigma_k \to 1$ as $k \to \infty$, and assumption (2.6) is altered to the assertion that the series $\sum_k (\sigma_k - 1)$ converges.

3 Methodology

3.1 Empirical Bayes rule

In this section we suggest methodology based on conventional kernel density estimation. Section 3.2 will treat alternative approaches founded on ideas from projection pursuit and additive modelling. In both cases the techniques are based on a sequence of nonparametric density approximations, and particularly in section 3.1 it is simpler if we standardise each feature for scale, since this makes it feasible to use the same bandwidth for each feature. Note that, since we are working simultaneously with thousands or tens of thousands of features, the natural scales of those quantities can vary extensively.

These considerations lead us to replace the kth component, X_{jik}, of the p-vector X_{ji} by

$$\widehat{X}_{jik} = X_{jik}\big/\hat{\sigma}_{j.k}\,, \tag{3.1}$$

where $\hat{\sigma}_{j.k}^2 = n_j^{-1} \sum_i (X_{jik} - \bar{X}_{j.k})^2$ and $\bar{X}_{j.k} = n_j^{-1} \sum_i X_{jik}$. Alternatively, a robust estimator of scale could be used.

Our methodology is recursive, as follows. Assume that, at the previous step, we chose feature indices $\hat{k}_1, \ldots, \hat{k}_{q-1}$. It helps interpretation if we consider $\hat{k}_\ell$ to be an estimator of some specific feature index k_ℓ, say, where $k_1, \ldots, k_{q-1}$ are distinct. Write $f_j(\cdot \,|\, k_1, \ldots, k_{q-1})$ for the joint density of the features with indices $k_1, \ldots, k_{q-1}$, in a random vector drawn from Π_j.

At the next step, choose a new feature index, k say, from the set $\{1, \ldots, p\} \setminus \{\hat{k}_1, \ldots, \hat{k}_{q-1}\}$, and construct a nonparametric estimator $\hat{f}_j(\cdot \,|\, \hat{k}_1, \ldots, \hat{k}_{q-1}, k)$ of $f_j(\cdot \,|\, k_1, \ldots, k_{q-1}, k)$, for $j = 1, 2$, based on the features with indices $\hat{k}_1, \ldots, \hat{k}_{q-1}, k$ and using a bandwidth sequence $h_1, \ldots, h_q$. Specifically, define

$$\hat{f}_j(x \,|\, \hat{k}_1, \ldots, \hat{k}_{q-1}, k) = \frac{1}{n_j \prod_\ell h_\ell} \sum_{i=1}^{n_j} \prod_{\ell = \hat{k}_1, \ldots, \hat{k}_{q-1}, k} K\left(\frac{x_\ell - \widehat{X}_{ji\ell}}{h_\ell}\right), \tag{3.2}$$

where $x = (x_1, \ldots, x_p)$ and K is a univariate kernel. Thus, $\hat{f}_j$ is a conventional q-variate kernel density estimator, although based on the rescaled data $\widehat{X}_{ji}$ rather than the original data X_{ji}, and treating only the vector components with indices $\hat{k}_1, \ldots, \hat{k}_{q-1}, k$.

Since we standardise the vector components for scale (see (3.1)) then in applications we would likely take the bandwidths h_ℓ to be constant, and choose this value to minimise a leave-one-out estimator of error rate, as discussed two sections below.

Put

$$\begin{aligned} \hat{g}(x) &= \hat{g}(x \,|\, \hat{k}_1, \ldots, \hat{k}_{q-1}, k) \\ &= \frac{\pi\, \hat{f}_1(x \,|\, \hat{k}_1, \ldots, \hat{k}_{q-1}, k)}{\pi\, \hat{f}_1(x \,|\, \hat{k}_1, \ldots, \hat{k}_{q-1}, k) + (1 - \pi)\, \hat{f}_2(x \,|\, \hat{k}_1, \ldots, \hat{k}_{q-1}, k)}\,, \end{aligned} \tag{3.3}$$

and take the classifier $\mathrm{Cl}(\cdot \,|\, \hat{k}_1, \ldots, \hat{k}_{q-1}, k)$ to be the one that assigns Z to Π_1 if $\hat{g}(Z) > \frac{1}{2}$, and to Π_2 otherwise. Estimate the error rate of this classifier using leave-one-out methods (see section 3.3), and choose k to minimise the error rate. Ties for the minimising q can be broken by choosing, from among

6

the tied values, the one that minimises the average distance that $\hat{g}$ lies on the wrong side of $\frac{1}{2}$ when an incorrect decision is made. Note that $\hat{g}$ can equivalently be defined by $\hat{g} = \hat{g}_1/(\hat{g}_1 + \hat{g}_2)$, where

$$\hat{g}_j(x) = \pi_j\, n_j^{-1} \sum_{i=1}^{n_j} \prod_{\ell = \hat{k}_1, \ldots, \hat{k}_{q-1}, k} K\left(\frac{x_{k_\ell} - \widehat{X}_{ji\ell}}{h_\ell}\right), \tag{3.4}$$

$x = (x_1, \ldots, x_p)$, $\pi_1 = \pi$ and $\pi_2 = 1 - \pi$.

We continue iterating this process until no further improvements are made. If the same bandwidth is used for each feature, its value can be updated after each new feature selection step, using minimum error rate as the criterion. (Choosing a separate bandwidth for each feature introduces significant noise, and so is unattractive.) At termination the classifier is a relatively low-dimensional approximation to the Bayes classifier discussed in section 2.1. Therefore we refer to it as empirical Bayes.

This approach differs from more conventional ones in that it is bottom up, rather than top down. That is, rather than start with vectors of length p and successively knock out features, which (for example) the Lasso does as the penalty is increased, our approach starts with vectors of length 1 and successively adds features depending on the success of the classifier based on the features selected so far.

3.2   Projective density approximations

Alternatives to the density estimator at (3.2) include variants based on projection pursuit and related ideas. For example, suppose we have selected $q \geq 2$ components, $\hat{k}_1, \ldots, \hat{k}_q$ say, using the ideas in section 3.1. At this point we might take q no larger than 2 to 4. Let $\check{X}_{ji}$ denote the subvector of X_{ji} obtained by removing all but the components with indices $\hat{k}_1, \ldots, \hat{k}_q$. We can construct a univariate density approximation as follows:

$$\check{f}_j(u \,|\, \omega) = \frac{1}{nh} \sum_{i=1}^{n_j} K\left(\frac{u - \omega \cdot \check{X}_{ji}}{h}\right), \tag{3.5}$$

for $j = 1, 2$, where ω is a q-vector of unit length and, if $\omega^{(1)}$ and $\omega^{(2)}$ are q-vectors, $\omega^{(1)} \cdot \omega^{(2)}$ denotes their inner (or scalar) product. We can choose $\omega = \widehat{\omega}$ to minimise a leave-one-out estimator of the error rate of the classifier constructed as in section 3.1 but where $\hat{g}$, rather than being defined by (3.3), is taken to be $\hat{g} = \hat{g}_1/(\hat{g}_1 + \hat{g}_2)$, where $\hat{g}_j$, instead of having the definition at (3.4), is given by $\hat{g}_j = \pi_j\, \check{f}_j(\cdot \,|\, \widehat{\omega})$, and $\check{f}_j$ is given by (3.5). This approach is motivated by a desire to reduce the variance inherent in a regular multivariate density estimator by converting it to a one-dimensional form using a projection that minimises error rate.

3.3   Estimating error rate

The error rate for a general classifier Cl, for classifying a new data value at z, is $\mathrm{err}(z) = P\{\mathrm{Cl}(Z) \neq L(Z)\,|\,Z = z\}$, and the expected (overall) error rate is $\mathrm{errate} = E\{\mathrm{err}(X)\}$. A leave-one-out estimator of errate is given by

$$\widehat{\mathrm{errate}} = \sum_{j=1}^{2} \pi_j\, n_j^{-1} \sum_{i=1}^{n_j} I\{\mathrm{Cl}_{-ji}(X_{ji}) \neq I(X_{ji})\},$$

7

where Cl_{-ji} is the version of Cl when X_{ji} is omitted from the data $\mathcal{X}_1 \cup \mathcal{X}_2$.

In particular, if $\hat{g}_{-ji}$ denotes the version of $\hat{g}$, at (3.3), constructed when X_{ji} is omitted from $\mathcal{X}_j$ and the other sample is left unchanged, then

$$\widehat{\text{errate}} = \pi \, n_1^{-1} \sum_{i=1}^{n_1} I\{\hat{g}_{-1i}(X_{-1i}) \leq \tfrac{1}{2}\} + (1-\pi) \, n_2^{-1} \sum_{i=1}^{n_2} I\{\hat{g}_{-2i}(X_{-2i}) > \tfrac{1}{2}\}. \tag{3.6}$$

Pursuing the algorithm in section 3.1, and given the values of the previously chosen indices $\hat{k}_1, \ldots, \hat{k}_{q-1}$, we select $k = \hat{k}_q$ to minimise $\widehat{\text{errate}}$ when the latter is computed from data for which all features except those with the indices $(\hat{k}_1, \ldots, \hat{k}_{q-1}, k)$ have been deleted. Ties can be broken as suggested below (3.3).

4 Numerical results

In this section we compare the error rates in finite samples for some commonly used classification methods in high dimensional settings and compare it with our proposed methodology. The following alternative classification methods are considered: nearest shrunken centroid (NSC), elastic net (ENET) and sure independence screening (SIS). The error rates of all of the above classifiers will depend on the choice of a regularisation parameter and we employ two different methods for comparing the error rates. Firstly, we compare the minimum error rates over the selected range of regularisation values; the method providing the absolute minimum error in this sense could be termed the 'best' method. Secondly, we use cross-validation (CV) to select a regularisation parameter for each method and compare the error rates on the test sample at each of these selected regularisation parameters.

We now briefly describe the alternative classification methods. The nearest shrunken centroid (NSC) is a regularised version of the nearest centroid classifier. The NSC method shrinks the classwise mean towards the overall mean, for each feature separately. Details can be found in Hastie et al. (2009). The elastic net (ENET) procedure of Zou and Hastie (2005) employs a weighted combination of the ℓ_1 and ℓ_2 norms (in our case we used weight $\frac{1}{2}$) and treats the resulting weighted sum as the penalty function. The SIS procedure of Fan and Lv (2008) initially selects a set of d predictors out of p, where $d \approx \gamma n$ for some $\gamma \in (0, 1)$, and applies any suitably chosen estimation procedure on the smaller set of predictors. In our simulations we used $d = \lfloor n/2 \rfloor$ and the Lasso as the second stage estimation procedure.

4.1 Simulation results: samples selected from standard normal

We consider $\Pi_j \equiv N_p(\mu_j, \mathbf{I}_p)$, $j = 1, 2$, with $\mu_j = (\mu_{j,1}, \ldots, \mu_{j,p})$. Sixty observations were selected from each group for constructing the training sample and thirty observations from each group were used for the test sample and $p = 500$. The class probabilities are $\pi = 1 - \pi = \frac{1}{2}$. The following four cases describe different classification situations, corresponding to different choices of μ_1 and μ_2.

(**a**): Assign $\mu_{j,k} = 0$ for all j, k. Let $r = \lfloor p/5 \rfloor$. Select the following five components $\{r, \ldots, 5r\}$ and change the corresponding component values in μ_1 and μ_2 to 2 and -2, respectively. In this case there is a clear difference between the two means at a very small number of components.

(**b**): In this case μ_1 is a linear periodic function taking equispaced values in $[-2, 2]$ and μ_2 is a periodic function within the interval $[-2, 2]$, with a parabolic shape. μ_1 and μ_2 are very close at almost all components, there is a systematic difference between them.

(**c**): μ_1 is the same as Case (**b**) above. And $\mu_{2,r} = \mu_{1,r} + \delta_r$, for all $1 \leq r \leq p$, where the δ_r are independent Gaussian $(0, 1/10)$ random variables. In this case the two populations are very similar (individual components are very close) and yet over all the p components, a distinct pattern of difference emerges.

(**d**): This is same as Case (**c**) above, except δ_r are independent Cauchy $(0, 1/10)$ random variables.

Table 1 shows the relative performance of different methods in Cases (**a**)–(**d**). It can be seen that the method suggested in this paper, indicated throughout by NEW, is similar to the other competing methods. We conclude that in the case of light tails, NEW performs similarly to competing classification methods.

Table 1: Error rates for different classification methods with observations from a **Gaussian** distribution.

Case	Error rates for different classification methods[†]			
	NSC	ENET	SIS	NEW
(**a**)	0.4667	0.45	0.4667	0.4663
	0.4667	0.5167	0.5333	0.5667
(**b**)	0.35	0.40	0.3833	0.3667
	0.3833	0.4167	0.50	0.4883
(**c**)	0.4167	0.40	0.3667	0.3833
	0.5333	0.45	0.5167	0.5167
(**d**)	0.3333	0.40	0.4333	0.3833
	0.3833	0.4833	0.45	0.65

[†] For each case, first row shows the absolute minimum error rate over all regularisation values. Second row shows the error rate evaluated at the optimum (CV-based) regularisation value.

4.2 Simulation results: samples selected from a heavy tailed distribution

We now consider the situation where Π_j are Cauchy with location parameter μ_j, $j = 1, 2$ and scale parameter 1. All the components of the p-dimensional sample are assumed to be independent. As earlier, we consider the same classification settings, with μ_j being modified differently in each of the four cases. Table 2 show the relative performance of different methods in Cases (**a**)–(**d**).

The superiority of the NEW method in the Cauchy case is displayed clearly through the figures in

column one (under error rates). Due to the heavy tailed nature of the data the competing methods, whose classification procedure is based on the mean, fail to capture the difference between the two populations.

5 Theoretical properties

5.1 Minimising error rate

The error rate of the Bayes classifier is given by (2.2), and can be written equivalently as:

$$\text{errate}_{\text{Bayes}} = 1 - \pi \int_{f_2 - f_1 > 0} (f_2 - f_1) = 1 - (1 - \pi) \int_{f_1 - f_2 > 0} (f_1 - f_2).$$

Therefore, given a choice between two or more versions of the pair (f_1, f_2), the error rate is minimised by selecting the pair for which $\int_{f_1 - f_2 > 0}(f_1 - f_2)$, or equivalently, $\int_{f_2 - f_1 > 0}(f_2 - f_1)$, is greatest. In the methodology suggested in section 3.3 we are making this choice empirically and sequentially.

In particular, given feature indices $k_1, \ldots, k_{q-1}$ (in practice, estimators $\hat{k}_1, \ldots, \hat{k}_{q-1}$ of those indices), we wish to select the next index k_q which is such that, among all $p - q$ choices of k_q,

$$d(\vec{k}) \equiv \int_{x \,:\, f_1(x \,|\, \vec{k}) - f_2(x \,|\, \vec{k}) > 0} \{f_1(x \,|\, \vec{k}) - f_2(x \,|\, \vec{k})\} \, dx$$

is largest. Here, $\vec{k} = (k_1, \ldots, k_q)$ and $x = (x_1, \ldots, x_q)$. The quantity $\widehat{\text{errate}}$ in (3.6) is an estimator of

Table 2: Error rates for different classification methods with observations from a **Cauchy** distribution.

Case	Error rates for different classification methods[†]			
	NSC	ENET	SIS	NEW
(a)	0.3167	0.3167	0.4167	0.2667
	0.35	0.3833	0.5333	0.3167
(b)	0.50	0.50	0.40	0.2333
	0.5667	0.5667	0.4167	0.5883
(c)	0.50	0.50	0.3833	0.2167
	0.6333	0.6833	0.50	0.7167
(d)	0.50	0.50	0.50	0.30
	0.55	0.70	0.7167	0.4167

[†] For each case, first row shows the absolute minimum error rate over all regularisation values. Second row shows the error rate evaluated at the optimum (CV-based) regularisation value.

$1 - (1 - \pi) d(\vec{k})$, and choosing the qth index to minimise $\widehat{\text{errate}}$, for $\hat{k}_1, \ldots, \hat{k}_{q-1}$ fixed, is an empirical approximation to choosing k_q, for given $k_1, \ldots, k_{q-1}$, to maximise $d(\vec{k})$.

In the independent component model discussed in section 2.2 the difference $f_1 - f_2$, when we consider $k_q = k$ to be the index of the next added feature, is very nearly proportional to $|\nu_k|$, as can be seen by simple Taylor expansion. This suggests that, except possibly for the first few feature indices, the order in which new indices should be added is that of the values of $|\nu_k|$. Theoretical confirmation of this property will be given in section 5.2.

5.2 Formula for error rate

In the independent component model, if $j = 1$ then the p-variate distribution of X_{ji} has density $f_1(x) = \Pi_{k \leq p} \phi(x_k)$, where ϕ is a univariate density, and if $j = 2$ then the density is $f_2(x) = \Pi_{k \leq p} \phi(x_k + \nu_k)$, where $\nu_1, \ldots, \nu_p$ are real numbers. For simplicity we again take the prior probabilities π and $1 - \pi$ to equal $\frac{1}{2}$.

Suppose we have chosen specific component indices $k_1, \ldots, k_{q-1}$ (without loss of generality they equal $1, \ldots, q - 1$, respectively), and we wish to choose the next index so as to minimise error rate of the Bayes classifier. In Appendix A.2 we shall prove that, if component index $k \geq q$ is selected, then the error rate of the classifier based on components with indices $1, \ldots, q - 1, k$ is given by

$$\text{errate}(1, \ldots, q-1, k) = \text{errate}(1, \ldots, q-1) - \tfrac{1}{2}\, \nu_k^2\, E\left\{ \frac{\phi'(W)}{\phi(W)} \right\}^2 f_{q-1}(0) + o\big(\nu_k^2\big) \tag{5.1}$$

for small ν_k, where $\text{errate}(1, \ldots, q-1)$ denotes the error rate of the classifier based on the first $q - 1$ components, f_{q-1} equals the density of the distribution of

$$S_{q-1} = \sum_{\ell=1}^{q-1} \log\{\phi(W_\ell + \nu_\ell)/\phi(W_\ell)\}\,, \tag{5.2}$$

and $W, W_1, \ldots, W_{q-1}$ are independent and identically distributed with density ϕ. This makes it clear that:

$$\begin{array}{ll} \text{for large values of } q \text{ the error rate is reduced by the greatest amount by} \\ \text{choosing, as the next component, the one for which } |\nu_k| \text{ is largest.} \end{array} \tag{5.3}$$

We shall also show that this result is essentially unchanged if we replace $\phi(u)$ by the expected value, $\mu(u)$, of the corresponding kernel density estimator:

$$\mu(u) = \int K(v)\, \phi(u - hv)\, dv\,. \tag{5.4}$$

That is, in effect, what happens when we use kernel density estimators to compute the error densities f_1 and f_2; this connection will be drawn in detail in section 5.3. In particular, although the formulae for $\text{errate}(1, \ldots, q-1, k)$ and $\text{errate}(1, \ldots, q-1)$ in (5.1) alter when ϕ is replaced by μ, and $E\{\phi'(W)/\phi(W)\}^2$ changes to $E\{\mu'(W)/\mu(W)\}^2$, the conclusion (5.3) is unaltered, because:

$$\begin{array}{ll} \text{even though the effect of smoothing might be felt significantly on the value of error} \\ \text{rate, its influence on the accuracy with which we choose the order of the sequence of} \\ \text{components is relatively minor.} \end{array} \tag{5.5}$$

See Appendix A.2 for a proof of (5.5). This property ensures that empirical Bayes methods can be effective.

5.3 With probability converging to 1 as $n \to \infty$, the empirical Bayes rule deduces the correct order of features, and the rule is asymptotically equivalent to the Bayes rule

For the sake of simplicity and brevity we again confine attention to the independent component case, showing in the theorem below that the empirical Bayes algorithm suggested in sections 3.1 and 3.3 gives, with probability converging to 1, the correct ordering of components along a sequence that is unboundedly long as $n \to \infty$. High-dimensional settings other than the independent component one are also treatable, but can be so complex that they lose transparency. The theorem below depends on a lemma that applies in the general case, and is stated in Appendix A.3. It can be used to derive the more general results.

Next we state regularity conditions. We take the asymptotic parameter to be n, and assume that n_1, n_2 and p are all functions of n, with n_1 and n_2 being approximately the same size as n. Recall that if $Z = (Z_1, \ldots, Z_p)$ is drawn from Π_1 then the vector components are all distributed as W, and that the empirical Bayes algorithm is based on kernel density estimators with kernel K and bandwidth h_ℓ, depending on n, applied to the ℓth selected component. We assume of these quantities that:

(a) K is a symmetric, compactly supported probability density satisfying
$|K(x - y)| \le B_1 \, |x - y|$ for all x, y, where $B_1 > 0$;

(b) $h_\ell^{-1} \le B_2 \, h^{-1}$ for $\ell \ge 1$ and each n, where $h = h(n)$, $B_2 > 0$ and $\max h_\ell \to 0$;

(c) $n \asymp n_1 \asymp n_2$ and $p = O(n^{B_3})$ for some $B_3 > 0$;

(d) $E(|W|^{B_4}) < \infty$ where $B_4 > 2$ is sufficiently large, and the variance of each component equals 1.

(5.6)

The assumption in part (d) that all variances equal 1 is reasonable because we rescale all components; see (3.1). It means that the moment condition assumed earlier in (d) is imposed with respect to the same scale for each component.

Throughout we take the distribution of W, and the sequence $\nu_1, \nu_2, \ldots$ not to depend on n, and we ask that the sequence satisfy (2.6). As argued in section 2.2, (2.6) is equivalent to asserting that the classification problem is difficult; the problem would be much easier if the series in (2.6) diverged. Suppose too that the function $\lambda(u \mid h) = \mu(u)$ (the latter defined as at (5.4)) has a finite moment generating function in a given neighbourhood of the origin for all sufficiently small bandwidths, and that $\mu(\cdot \mid h)$ is sufficiently regular. Specifically: for some $r, t, h_0 > 0$ and $s > 1$,

$$\sup_{|\nu| \le \sup_k |\nu_k|} \ \sup_{h \in (0, h_0]} E\big[\exp\{r \, |\lambda(W + \nu \mid h)|\}\big] < \infty \,, \tag{5.7}$$

$$E\left|\frac{\mu'(W \mid h)}{\mu(W \mid h)}\right|^{2s} + E\left|\frac{\mu''(W \mid h)}{\mu(W \mid h)}\right|^{s} + E\left|\frac{\psi_t(W \mid h)}{\mu(W \mid h)}\right| < \infty \,, \tag{5.8}$$

where $\psi_t(u \mid h) = \sup_{v \, : \, |v| \le t} |\mu'''(u + v \mid h)|$. Let $k_1^0, k_2^0, \ldots$ denote the feature indices which, in that

12

order, minimise error rate for the Bayes classifier, and assume that:

for a sequence $\ell(1), \ldots, \ell(r)$ of distinct integers, the product over $1 \leq j \leq r$ of the characteristic function $\chi_{\pm}$ of $\lambda(W \pm \nu_{k^0_{\ell(j)}} \mid h) - \lambda(W \mid h)$ is absolutely integrable for both choices of the $\pm$ signs, and the integral is bounded uniformly in all $h \in (0, h_0]$, for some $h_0 > 0$. $\qquad$ (5.9)

(Assumptions (5.7)–(5.9) are readily verified for standard distributions of W, for example the Normal and Student's t.) We also assume that, for constants $B_5 > 0$ and $\xi \in (0, \frac{1}{2})$,

$$p = O(n^{B_5}), \quad h \leq 1, \quad hH(q_0) \geq n^{-\xi}, \qquad (5.10)$$

with $h = \min\{h_1, \ldots, h_{q_0}\}$ and $H = H(q) = \prod_{1 \leq h \leq q} h_\ell$. Let $k^0_1, \ldots, k^0_p$, a permutation of $1, \ldots, p$, denote the sequence of indices that give successive minima of the error rate, at (2.2), of the ideal Bayes classifier. That is, if we have constructed a classifier based on vector components with indices $k^0_1, \ldots, k^0_{q-1}$ where $q \leq p$, then, among all choices of the qth component index from $\{1, \ldots, p\} \setminus \{k^0_1, \ldots, k^0_{q-1}\}$ that could be used to increase dimension by one unit, the choice k^0_q reduces the error rate of the Bayes classifier by the greatest amount.

The last part of (5.10) imposes a condition on the sizes of the bandwidths, which we illustrate here in the practically important case where each h_ℓ takes the same value. There $hH(q_0) = h^{q_0+1}$, and so to ensure the last part of (5.10) we need $\log h^{-1} \leq \{\xi (\log n)/(q + 1)$. Since, by the first part of (5.9), $q = o(\log n)$, then taking $h = \exp\{-\xi (\log n)/(q + 1)\}$ will ensure that $h \to 0$ and the last part of (5.10) holds.

Let $\hat{k}_1, \hat{k}_2, \ldots$ denote the feature indices chosen empirically as suggested in section 3.3. Write $\widetilde{L}(Z)$ for the class label determined by the ideal Bayes rule (i.e. $\widetilde{L}(Z) = 1$ if $g_{\mathrm{Bayes}}(Z) > \frac{1}{2}$ and equals 2 otherwise, where g_{Bayes} is given by (2.1)), and let $\widehat{L}(Z)$ denote the class label determined by the classifier suggested in sections 3.1 and 3.3 (i.e. $\widehat{L}(Z) = 1$ if $\hat{g}(Z) > \frac{1}{2}$ and equals 2 otherwise, where $\hat{g} = \hat{g}_1/(\hat{g}_1 + \hat{g}_2)$ and $\hat{g}_j$ is given by (3.4) but with $q = q_0$ and the set of feature indices taken to be $\hat{k}_1, \ldots, \hat{k}_{q_0}$).

The following theorem shows that, with probability converging to 1 as n diverges, the empirically chosen indices $\hat{k}_1, \ldots, \hat{k}_{q_0}$ are identical to the ideal ones, $k^0_1, \ldots, k^0_q$, and the empirical Bayes classifier suggested in sections 3.1 and 3.3 gives the same results as the ideal Bayes classifier discussed in section 2.1.

Theorem 5.1. *Assume* (5.6)–(5.10), *and, given* $\eta > 0$, *let* $q_0 = q_0(n)$ *denote any sequence of integers diverging to infinity such that* $q_0 = o(\log n)$ *and the* q_0th *largest value of* ν_k *satisfies* $n^{-\eta} = o(\nu^2_{q_0})$. *Then, for some* $\eta, \eta' > 0$,

$$P(\hat{k}_q = k^0_q \quad for \quad 1 \leq q \leq q_0) = 1 - O(n^{-\eta'}), \quad P\{\widehat{L}(Z) = L(Z)\} = 1 - O(n^{-\eta'}) \qquad (5.11)$$

as $n \to \infty$.

Appendix A

A.1. *Proof of* (2.5): Let the random variable W have the distribution with density ϕ. Assume that ϕ is three times differentiable, and define $\psi_{1t}(u) = \sup_{v \,:\, |v| \leq t} |\phi'''(u+v)|$ and $\psi_{2t}(u) = \sup_{v \,:\, |v| \leq t} |\log \phi(u+$

$v)|$, where $t > 0$. We shall assume of ϕ that for for some $s > 1$ and some $t > 0$,

$$E\left|\frac{\phi'(W)}{\phi(W)}\right|^{2s} + E\left|\frac{\phi''(W)}{\phi(W)}\right|^{s} + E\left|\frac{\psi_{1t}(W)}{\phi(W)}\right| < \infty, \tag{A.1}$$

and for all $s > 0$ and some $t > 0$,

$$E\{\psi_{2t}(W)^s\} < \infty. \tag{A.2}$$

We shall prove that (A.1) and (A.2) imply (2.5). Conditions (A.1) and (A.2) are satisfied by standard distributions, such as the normal, Student's t distribution, the gamma distribution with shape parameter not less than 3, etc.

By Taylor expansion,

$$\frac{\phi(W+\nu)}{\phi(W)} = 1 + \nu\,\frac{\phi'(W)}{\phi(W)} + \tfrac{1}{2}\,\nu^2\,\frac{\phi''(W)}{\phi(W)} + \tfrac{1}{6}\,\nu^3\,\frac{\phi'''(W+\Theta\nu)}{\phi(W)},$$

where the random variable Θ satisfies $P(|\Theta| \le 1) = 1$. Let $\mathcal{E}$ denote the event that none of $|\nu\phi'(W)/\phi(W)|$, $|\nu^2\phi''(W)/\phi(W)|$ and $|\nu^3\psi_{1\nu}(W)/\phi(W)|$ exceeds $\tfrac{1}{2}$. By further Taylor expansion,

$$E\left[I(\mathcal{E})\log\left\{\frac{\phi(W+\nu)}{\phi(W)}\right\}\right] = E\left(I(\mathcal{E})\left[\nu\,\frac{\phi'(W)}{\phi(W)} + \tfrac{1}{2}\nu^2\,\frac{\phi''(W)}{\phi(W)} - \tfrac{1}{2}\left\{\nu\,\frac{\phi'(W)}{\phi(W)}\right\}^2\right]\right) + O(\nu^3). \tag{A.3}$$

Markov's inequality and (A.1) imply that $P(\tilde{\mathcal{E}}) = O(\nu^{\min(2s_1,3)})$, where $s_1 \ge 1$ is a value of s in (A.1). Moreover, if $u, v > 1$ and $u^{-1} + v^{-1} = 1$ then, in view of (A.2),

$$E\left[I(\tilde{\mathcal{E}})\left|\log\left\{\frac{\phi(W+\nu)}{\phi(W)}\right\}\right|\right] \le P(\tilde{\mathcal{E}})^{1/u}\,E\left[\left|\log\left\{\frac{\phi(W+\nu)}{\phi(W)}\right\}\right|^{v}\right]^{1/v}$$

$$= O\{P(\tilde{\mathcal{E}})^{1/u}\} = o(\nu^2), \tag{A.4}$$

where to obtain the two identities we took v so large that $(1 - v^{-1})\min(2s_1, 3) > 2$.

Similarly,

$$E\left[I(\tilde{\mathcal{E}})\left|\nu\,\frac{\phi'(W)}{\phi(W)} + \tfrac{1}{2}\nu^2\,\frac{\phi''(W)}{\phi(W)} - \tfrac{1}{2}\left\{\nu\,\frac{\phi'(W)}{\phi(W)}\right\}^2\right|\right]$$

$$\le E\left\{I(\tilde{\mathcal{E}})\left|\nu\,\frac{\phi'(W)}{\phi(W)}\right|\right\} + \nu^2\,E\left(I(\tilde{\mathcal{E}})\left[\left|\frac{\phi''(W)}{\phi(W)}\right| + \left\{\frac{\phi'(W)}{\phi(W)}\right\}^2\right]\right)$$

$$= O\{P(\tilde{\mathcal{E}})^{1/2}\,\nu + P(\tilde{\mathcal{E}})^{1/s}\,\nu^2\} = o(\nu^2). \tag{A.5}$$

Combining (A.3)–(A.5) we conclude that

$$E\left[\log\left\{\frac{\phi(W+\nu)}{\phi(W)}\right\}\right] = E\left[\nu\,\frac{\phi'(W)}{\phi(W)} + \tfrac{1}{2}\nu^2\,\frac{\phi''(W)}{\phi(W)} - \tfrac{1}{2}\left\{\nu\,\frac{\phi'(W)}{\phi(W)}\right\}^2\right] + o(\nu^2)$$

$$= -\tfrac{1}{2}\,\nu^2\,E\{\phi'(W)/\phi(W)\}^2 + o(\nu^2), \tag{A.6}$$

which is identical to (2.5). $\qquad\square$

A.2. *Proof of* (5.1): We assume (A.1) and (A.2) and note that a rigorous proof follows closely the lines in Appendix A.1. Therefore we give only an outline here. In the notation of (5.2), define $\Delta_\pm = \log\{\phi(W \pm \nu)/\phi(W)\}$, where the signs are taken respectively, $\nu = \nu_k$ and

$$S_{q-1} = \sum_{\ell=1}^{q-1} \log\{\phi(W_\ell + \nu_\ell)/\phi(W_\ell)\}, \quad T_{q-1} = \sum_{\ell=1}^{q-1} \log\{\phi(W_\ell)/\phi(W_\ell - \nu_\ell)\}.$$

By (A.6),

$$E(\Delta_\pm) = \mp \tfrac{1}{2}\,\nu^2\,E\{\phi'(W)/\phi(W)\}^2 + o(\nu^2)\,,$$

and therefore, writing F_{q-1} and G_{q-1} for the distribution functions of S_{q-1} and T_{q-1}, respectively,

$$
\begin{aligned}
P(S_{q-1} + \Delta_+ > 0) &= E\{P(S_{q-1} > -\Delta_+ \mid \Delta_+)\} \\
&= 1 - F_{q-1}(0) + E(\Delta_+)\,F'_{q-1}(0) + o(\nu^2) \\
&= 1 - F_{q-1}(0) - \tfrac{1}{2}\,\nu^2\,E\{\phi'(W)/\phi(W)\}^2\,f_{q-1}(0) + o(\nu^2)\,, \\
P(-T_q + \Delta_- < 0) &= E\{P(T_{q-1} > \Delta_- \mid \Delta_-)\} \\
&= 1 - G_{q-1}(0) + \tfrac{1}{2}\,\nu^2\,E\{\phi'(W)/\phi(W)\}^2\,f_{q-1}(0) + o(\nu^2)\,.
\end{aligned}
$$

(We assume that the first two derivatives of F_{q-1} and G_{q-1} are uniformly bounded, and converge to the corresponding derivatives of the limit of the distribution functions.) The error rate of the classifier based on the components with indices $1, \ldots, q-1, k$ is therefore

$$
\begin{aligned}
\mathrm{errate}(1, \ldots, q-1, k) &= \tfrac{1}{2}\left\{P(S_{q-1} + \Delta_+ > 0) + P(T_{q-1} - \Delta_- < 0)\right\} \\
&= \mathrm{errate}(1, \ldots, q-1) - \tfrac{1}{2}\,\nu^2\,E\{\phi'(W)/\phi(W)\}^2\,f_{q-1}(0) + o(\nu^2) \quad\quad \text{(A.7)}
\end{aligned}
$$

as $\nu \to 0$, uniformly in $k \in [q, p]$, where (as before) $\nu = \nu_k$ and $\mathrm{errate}(1, \ldots, q-1) = 1 - F_{q-1}(0) + G_{q-1}(0)$ is the error based on just the components with indices $1, \ldots, q-1$. Result (A.7) is equivalent to (5.1).

Under conditions (5.8) and (5.9), a proof in the case where ϕ is replaced by μ, the latter defined at (5.4), is similar to that given above, except that the ratio ϕ'/ϕ in (A.7) is replaced by μ'/μ, which converges to ϕ'/ϕ as the bandwidth, h, converges to zero. Also, F_{q-1} and G_{q-1} are similarly altered, to $F_{q-1,h}$ and $G_{q-1,h}$, say, again in a way that becomes negligibly small as $h \to 0$. For example, F_{q-1} changes to the distribution function $F_{q-1,h}$ of

$$S_{q-1,h} = \sum_{\ell=1}^{q-1} \log\{\mu(W_\ell + \nu_\ell)/\mu(W_\ell)\}\,.$$

Therefore, defining $\mathrm{errate}_{q-1,h} = 1 - F_{q-1,h}(0) + G_{q-1,h}(0)$, and recalling that $\nu = \nu_k$, we see that the analogue of (A.7) now holds in the form:

$$\mathrm{errate}_{q-1,h} = \mathrm{errate}_{q-1,h} - \tfrac{1}{2}\,\nu^2\,E\{\mu'(W)/\mu(W)\}^2\,f_{q-1}(0) + o(\nu^2)\,. \quad\quad \text{(A.8)}$$

Again it is clear that, for large q and provided that the values of $|\nu_k|$ are decreasing steadily, the greatest reduction in error rate (here $\mathrm{errate}_{q-1,h}$, rather than $\mathrm{errate}(1, \ldots, q-1)$) is achieved by adjoining the

component with index k for which $|\nu_k|$ is largest. Therefore, even though the effect of smoothing might be felt significantly on the value of error rate, and the quantities $\text{errate}_{q-1,h}$ and $\text{errate}(1,\ldots,q-1)$ might not be directly comparable, smoothing does not appreciably affect the accuracy with which we choose the order of the sequence of components. This establishes (5.5). $\qquad\square$

A.3. *Lemma for Appendix* A.*4:* First we state regularity conditions. As before we take the asymptotic parameter to be n, but we do not assume the independent component model. Therefore, when we refer below to (5.6) we interpret part (d) of that assumption as meaning that "the B_4th moments of all components of X are bounded by a constant that depends only on B_4, chosen sufficiently large..." rather than that "$E(|W|^{B_4}) < \infty$ where $B_4 > 2$ is sufficiently large...."

Recall that $\vec{k} = (k_1, \ldots, k_q)$ and that $f_j(\cdot \mid \vec{k})$ is the joint density of the components of X_{j1} whose indices comprise the vector $\vec{k}$. Defining $x = (x_1, \ldots, x_p)$, let

$$\mu_j(x \mid \vec{k}) = \int \ldots \int \left\{ \prod_{\ell=1}^{q} K(u_\ell) \right\} f_j(x_{k_1} - h_1 u_1, \ldots, x_{k_q} - h_q u_q \mid \vec{k}) \, du_1 \, \ldots \, du_q$$

denote the expected value of the following "naive" density estimator, based on the unscaled features X_{jik}:

$$\tilde{f}_j(x \mid \vec{k}) = \frac{1}{n_j \prod_{\ell \leq q} h_\ell} \sum_{i=1}^{n_j} \prod_{\ell = k_1, \ldots, k_q} K\left(\frac{x_{k_\ell} - X_{ji\ell}}{h_\ell} \right). \tag{A.9}$$

In the definition of $\mu_j(\cdot \mid \vec{k})$ we have suppressed dependence on $h_1, \ldots, h_q$.

Define $\rho(x \mid \vec{k}) = \pi_2 \, \mu_2(x \mid \vec{k}) / \pi_1 \, \mu_1(x \mid \vec{k})$, where $\pi_1 = \pi$ and $\pi_2 = 1 - \pi$. Put

$$\text{errate}_q(\vec{k}) = \sum_{j=1}^{2} \pi_j \, P\{ \rho(X_{j1} \mid \vec{k}) \bowtie_{2j} 1 \}, \tag{A.10}$$

where X_{j1} is a p-vector drawn from Π_j, $\bowtie_{1j}$ denotes $\leq$ or $>$ according as $j = 1$ or 2, and $\bowtie_{2j}$ means $\geq$ or $<$ according as $j = 1$ or $j = 2$. Recall that $f_j(\cdot \mid \vec{k})$ denotes the joint density of $(Z_{k_1}, \ldots, Z_{k_q})$, and let $g(x \mid \vec{k}) = \pi \, f_1(x \mid \vec{k}) / \{ \pi \, f_1(x \mid \vec{k}) + (1 - \pi) \, f_2(x \mid \vec{k}) \}$. In this notation, $\text{errate}_q(\vec{k})$ is the error rate that would arise if we were to replace the function $g(x \mid \vec{k})$ in the Bayes-rule estimator based on data components with indices in the sequence $\vec{k}$, by $\pi \, \mu_1(x \mid \vec{k}) / \{ \pi \, \mu_1(x \mid \vec{k}) + (1 - \pi) \, \mu_2(x \mid \vec{k}) \}$.

Let $\delta, \eta_1, \eta_2, \eta_3 > 0$ satisfy $\eta_1 < \eta_2$ and $\eta_1 + 2\eta_2 + \eta_3 < 1$. We impose the following conditions on $\mu_j(\cdot \mid \vec{k})$:

$$\inf_{k \,:\, 1 \leq k \leq p, \, k \neq k_q^0} \text{errate}_q(k_1^0, \ldots, k_{q-1}^0, k) - \text{errate}_q(k_1^0, \ldots, k_{q-1}^0, k_q^0) > 20 \, \epsilon, \quad \text{for } 1 \leq q \leq q_0, \tag{A.11}$$

where $\epsilon > 0$ may depend on n, and

$$\max_{1 \leq q \leq q_0} \max_{\vec{k} \subseteq \mathcal{K}(q)} \max \left[\max_{j_1, j_2 \in \{1,2\}} P\{ \mu_{j_1}(X_{j_2 1} \mid \vec{k}) \leq n^{-\eta_1} \}, \right.$$

$$\left. \max_{j=1,2} P\{ |\rho(X_{j1} \mid \vec{k}) - 1| \leq A_1 \, n^{-\eta_2} \} \right] \leq \epsilon, \tag{A.12}$$

$$n^{\eta_1 + \eta_2} \left(\kappa^{q-1} \, B_1 \, B_2 \, q \, h^{-1} \, n^{\delta - (1/2)} \, e + \kappa^q \, n^{-C_1} \right) \leq H, \tag{A.13}$$

$$3 \, (n_j - 1)^{-1} \, n^{\eta_1 + 2\eta_2 + \eta_3} \leq H, \tag{A.14}$$

16

where $H = \prod_{\ell \leq q} h_\ell$, $\kappa \sup K$ and $C_1 > 0$. Define

$$P(n) = 24pq_0 \left[(n_1 \vee n_2) \left\{ \exp\left(-n^{\eta_3} \right) + n^{-C_1} \right\} + \exp\left\{ -(n_1 \wedge n_2)\,\epsilon^2 \right\} \right], \tag{A.15}$$

and let Z denote a p-vector drawn from the mixture population $\pi\,\Pi_1 + (1-\pi)\,\Pi_2$. Let $k_1^0, k_2^0, \ldots$ be as defined below (5.8), and let $\hat{k}_1^0, \hat{k}_2^0, \ldots$ be as defined below (3.6).

The following result is used in Appendix A.4 and proved in a longer version of this paper (Chatterjee and Hall (2009)):

Lemma 1. *If* (5.6) *and* (A.11)–(A.14) *hold then*

$$P\big(\hat{k}_q = k_q^0 \quad for \quad 1 \leq q \leq q_0\big) \geq 1 - P(n)\,, \quad P\big\{\widehat{L}(Z) = L(Z)\big\} \geq 1 - P(n)\,. \tag{A.16}$$

$\square$

A.4. *Proof of Theorem 5.1:* The theorem will follow from the lemma above if we establish (A.11)–(A.14) in the case $\epsilon = n^{-\eta}$, for a sufficiently small $\eta > 0$. To this end, let $W, W_1, W_2, \ldots$ be independent and identically distributed with density ϕ. Formula (A.8) holds under conditions (5.8) and (5.9), and implies (A.11) provided that ϵ in (A.11) is of smaller order then the square of q_0th largest value of ν_k. This was stipulated in the theorem by asking that $n^{-\eta} = o(\nu_{q_0}^2)$.

Next we derive (A.12). Recall the definition of μ at (5.4), and put $\lambda = \log \mu$. Result (A.12) in Appendix A.3 holds if, for each choice of $\zeta, \zeta_1, \zeta_2 \in \{-1, 0, 1\}$ such that $\zeta_1 \neq \zeta_2$ and $\zeta_1 \zeta_2 = 0$, we have, uniformly in $\vec{k} \in \mathcal{K}(q)$ for $1 \leq q \leq q_0$ and for all sufficiently large n:

$$P\left\{ \sum_{\ell=1}^q \lambda(W_\ell + \zeta\,\nu_{k_\ell}) \leq -\eta_1 \log n \right\} \leq \epsilon\,, \tag{A.17}$$

$$P\left(\left| \sum_{\ell=1}^q \left[\lambda(W_\ell + \zeta_1\,\nu_{k_\ell}) - \lambda(W_\ell + \zeta_2\,\nu_{k_\ell}) - \log\{\pi/(1-\pi)\} \right] \right| \leq A_2\,n^{-\eta_2} \right) \leq \epsilon\,, \tag{A.18}$$

where $A_2 > 0$ is any constant. To derive (A.17) for $\epsilon = n^{-\eta}$, where $\eta > 0$, it is necessary only to note that, since the variables $\lambda(W + \nu)$ have a finite moment generating function uniformly in $|\nu| \leq \sup_k |\nu_k|$ and $0 < h \leq h_0$ (see (5.8)), and since $q = o(\log n)$ (see the statement of the theorem), then an exponential inequality for a sum of independent random variables (see e.g. de la Peña (1999)) implies that, for a constant $D_1 > 0$, the left-hand side of (A.17) is dominated by

$$\exp\left[-D_1 \min\left\{ q^{-1} (\log n)^2, \log n \right\} \right] = O\big(n^{-D_1}\big)\,,$$

uniformly in ζ and in $\vec{k} \in \mathcal{K}(q)$ for $1 \leq q \leq q_0$.

To simplify exposition we assume that (5.9) holds for $r = 1$ and $\ell(1) = 1$. Other cases can be treated similarly. Property (A.18) in the case $\epsilon = n^{-\eta}$ is equivalent to:

$$\max{}^{(1)} P\left\{ Q_q(\vec{k}, \zeta_1, \zeta_2) \in \left[c_0 - A_2\,n^{-\eta_2}, c_0 + A_2\,n^{-\eta_2} \right] \right\} \leq n^{-\eta}\,, \tag{A.19}$$

where $Q_q(\vec{k}, \zeta_1, \zeta_2) = \sum_{\ell \leq q} \{\lambda(W_\ell + \zeta_1\,\nu_{k_\ell}) - \lambda(W_\ell + \zeta_2\,\nu_{k_\ell})\}$ and $\max^{(1)}$ denotes the maximum over $\vec{k} \in \mathcal{K}(q)$, over $1 \leq q \leq q_0$, and over ζ_1, ζ_2 such that $\zeta_1 \neq \zeta_2$ and $\zeta_1 \zeta_2 = 0$. It is straightforward to show

17

that, for any fixed q_1 and for all $C > 0$, $P(\hat{k}_\ell = k_\ell^0$ for $1 \leq \ell \leq q_1) = 1 - O(n^{-C})$ as $n \to \infty$. In particular, this result holds for $q_1 = 1$, and so it suffices to derive the version of (A.19) when k_1 is held fixed at k_1^0. That is, in (A.19) we may take

$$Q_q(\vec{k}, \zeta_1, \zeta_2) = Q_0 + \sum_{\ell=2}^{q} \left\{ \lambda(W_\ell + \zeta_1 \, \nu_{k_\ell}) - \lambda(W_\ell + \zeta_2 \, \nu_{k_\ell}) \right\}, \tag{A.20}$$

where $Q_0 = \lambda(W_1 + \zeta_1 \, \nu_{k_1^0}) - \lambda(W_1 + \zeta_2 \, \nu_{k_1^0})$.

By assumption (see (5.9)), the characteristic function $\chi_\pm$ of $\lambda(W \pm \nu_{k_1^0}) - \lambda(W)$ is absolutely integrable for both choices of the $\pm$ signs, and so the characteristic function of Q_0 is too. Therefore, $\chi = \max(|\chi_+|, |\chi_-|)$ is integrable. The characteristic function of $Q_q(\vec{k}, \zeta_1, \zeta_2)$ (on the left-hand side of (A.20)), being the product of the characteristic functions of the independent summands on the right-hand side of (A.20), is bounded above in absolute value by χ. Therefore the density of $Q_q(\vec{k}, \zeta_1, \zeta_2)$ is bounded above by $D_2 = (2\pi)^{-1} \int \chi$. Hence, the left-hand side of (A.19), when we take $Q_q(\vec{k}, \zeta_1, \zeta_2)$ to be given by (A.20), is bounded above by $2D_2 A_2 n^{-\eta_2}$, and therefore (A.19) holds for all sufficiently large n if $\eta < \eta_2$. This proves (A.18).

By assumption, $hH \geq n^{-\xi}$ where $\xi \in (0, \frac{1}{2})$. Without loss of generality, $C_1 \geq \frac{1}{2}$ in (A.20), and $\eta_1 + \eta_2$ and δ are sufficiently small to allow us to write $\xi = \frac{1}{2} - \xi_1 - \eta_1 - \eta_2 - \delta$, where $\xi_1 > 0$. Then the left-hand side of (A.13), multiplied by H^{-1}, is dominated by a constant multiple of $n^{\eta_1 + \eta_2 + \delta - (1/2) + \xi} \kappa^q q = n^{-\xi_1} \kappa^q q = o(1)$, where the last identity holds since $q = o(\log n)$. This proves (A.13). Also, if η_1, η_2 and η_3 are so small that $\eta_1 + 2\eta_2 + \eta_3 < \frac{1}{2}$ then the left-hand side of (A.14), multiplied by H^{-1}, is bounded above by a constant multiple of $n^{-1 + \eta_1 + 2\eta_2 + \eta_3 + \xi} \leq n^{\xi - (1/2)} = o(1)$. This establishes (A.14).

This completes the derivation of (A.11)–(A.14); we have assumed that $q = o(\log n)$, that ϵ in (A.11)–(A.14) equals $n^{-\eta}$ for sufficiently small $\eta > 0$, and that q_0 has the property that the q_0th largest ν_k is of strictly smaller order than $n^{-\eta}$. By taking η sufficiently small we ensure that the right-hand side of (A.15) equals $O(n^{-\eta'})$ for some $\eta' > 0$. The theorem now follows from (A.16). $\qquad \square$

References

Ancukiewicz, M. (1998). An unsupervised and nonparametric classification procedure based on mixtures with known weights. *J. Classification*, 15(1):129–141.

Bickel, P. J. and Levina, E. (2008). Covariance regularization by thresholding. *Ann. Statist.*, 36(6):2577–2604.

Candes, E. and Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.*, 35(6):2313–2351.

Chatterjee, A. and Hall, P. (2009). High dimensional classification when useful information comes from many, perhaps all features. (long version).

de la Peña, V. H. (1999). A general class of exponential inequalities for martingales and ratios. *Ann. Probab.*, 27(1):537–564.

Devroye, L., Györfi, L., and Lugosi, G. (1996). *A probabilistic theory of pattern recognition*, volume 31 of *Applications of Mathematics (New York)*. Springer-Verlag, New York.

Donoho, D. L. (2006). For most large underdetermined systems of linear equations the minimal l_1-norm solution is also the sparsest solution. *Comm. Pure Appl. Math.*, 59(6):797–829.

Donoho, D. L. and Elad, M. (2003). Optimally sparse representation in general (nonorthogonal) dictionaries via l^1 minimization. *Proc. Natl. Acad. Sci. USA*, 100(5):2197–2202 (electronic).

Duda, R. O., Hart, P. E., and Stork, D. G. (2001). *Pattern classification*. Wiley-Interscience, New York, second edition.

Dudoit, S., Fridlyand, J., and Speed, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Amer. Statist. Assoc.*, 97(457):77–87.

Efron, B. (1983). Estimating the error rate of a prediction rule: improvement on cross-validation. *J. Amer. Statist. Assoc.*, 78(382):316–331.

Efron, B. and Tibshirani, R. (1997). Improvements on cross-validation: the .632+ bootstrap method. *J. Amer. Statist. Assoc.*, 92(438):548–560.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.*, 96(456):1348–1360.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, 70(5):849–911.

Friedman, J. H. (1987). Exploratory projection pursuit. *J. Amer. Statist. Assoc.*, 82(397):249–266.

Friedman, J. H. and Tukey, J. W. (1974). A projection pursuit algorithm for exploratory data analysis. *IEEE Trans. Comput.*, 23(9):881–890.

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

Hall, P. and Kang, K.-H. (2005). Bandwidth choice for nonparametric classification. *Ann. Statist.*, 33(1):284–306.

Hall, P. and Miller, H. (2009). Using generalized correlation to effect variable selection in very high dimensional problems. *J. Comput. Graph. Statist.*, (to appear).

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning*. Springer Series in Statistics. Springer-Verlag, New York. Data mining, inference, and prediction.

Krzyżak, A. (1991). On exponential bounds on the Bayes risk of the nonparametric classification rules. In *Nonparametric functional estimation and related topics (Spetses, 1990)*, volume 335 of *NATO Adv. Sci. Inst. Ser. C Math. Phys. Sci.*, pages 347–360. Kluwer Acad. Publ., Dordrecht.

Lapko, A. V. (1993). *Neparametricheskie metody klassifikatsii i ikh primenenie.* VO "Nauka", Novosibirsk.

Meinshausen, N. (2007). Relaxed Lasso. *Comput. Statist. Data Anal.*, 52(1):374–393.

Pawlak, M. (1993). Kernel classification rules from missing data. *IEEE Transactions on Information Theory*, 39(3):979–988.

Shakhnarovich, G., Darrell, T., and Indyk, P. (2006). *Nearest-Neighbor Methods in Learning and Vision: Theory and Practice (Neural Information Processing).* The MIT Press.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288.

Witten, D. and Tibshirani, R. (2009). Covariance-regularized regression and prediction for high-dimensional problems. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, (to appear).

Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.*, 101(476):1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 67(2):301–320.

Appendix B : Proof of Lemma 1 (not-for-publication)

We derive only the first inequality in (A.16); a proof of the second is similar.

Step 1: Bounding the effect of replacing $\widehat{X}_{ji}$ by X_{ji}. Let $1 \leq q \leq p$, and let $\vec{k} = (k_1, \ldots, k_q)$ denote any sequence of distinct values chosen from $\{1, \ldots, p\}$ and ordered arbitrarily. Given $\delta \in (0, \frac{1}{2})$, let $\mathcal{E}_1 = \mathcal{E}_1(\delta)$ denote the event that

$$\max_{j=1,2} \max_{1 \leq i \leq n_j} \max_{1 \leq k \leq p} \left| \widehat{X}_{jik} - X_{jik} \right| \leq n^{\delta - (1/2)} , \tag{B.1}$$

and write $\tilde{\mathcal{E}}_1$ for the complement of $\mathcal{E}_1$. Since (i) the sth moments of all components of X are bounded by a constant that depends only on s, (ii) the variances $\text{var}(X_{jik})$ all equal 1 (for (i) and (ii), see (5.6)(d), and (iii) p diverges at only a polynomial rate (see (5.6)(d)), then Markov's inequality can be used to prove that:

$$\text{for each given } C_1 > 0 \text{ and } \delta \in (0, 1), \ P\{\tilde{\mathcal{E}}_1(\delta)\} \leq n^{-C_1} \text{ for all sufficiently large } n, \tag{B.2}$$

where the required value of B_4 in (5.6)(d) increases as C_1 increases and as δ decreases. Let B_1 and B_2 be as in (5.6)(a) and (5.6)(b), respectively; define $\kappa = \sup K$; and assume that

$$\kappa^{-1} B_1 B_2 q h^{-1} n^{\delta - (1/2)} \leq 1 . \tag{B.3}$$

The first of the following identities can be proved by Taylor expansion (using (B.1) and the Hölder bound in (5.6)(a)), and the second by collecting terms from the product in the first (and using the bound on $\sup K$ in (5.6)(a)):

$$
\prod_{\ell=1}^{q} K\left(\frac{x_{k_\ell} - \widehat{X}_{jik_\ell}}{h_\ell}\right)
$$
$$
= \prod_{\ell=1}^{q} \left\{ K\left(\frac{x_{k_\ell} - X_{jik_\ell}}{h_\ell}\right) + \Theta_1(\ell, x)\, B_1\, B_2\, h^{-1} n^{\delta-(1/2)} \right\}
$$
$$
= \prod_{\ell=1}^{q} K\left(\frac{x_{k_\ell} - X_{jik_\ell}}{h_\ell}\right) + \Theta_1(x)\, \kappa^{q-1}\, B_1\, B_2\, q\, h^{-1}\, n^{\delta-(1/2)}\, e\,, \tag{B.4}
$$

where, with probability at least $P(\mathcal{E}_1)$, the random variables $\Theta_1(\ell, x)$ and $\Theta_1(x)$ satisfy $|\Theta_1(\ell, x)|, |\Theta_1(x)| \leq 1$ uniformly in the following sense:

> uniformly in all choices of q such that (B.3) holds, in $j = 1, 2$, in all $1 \leq i \leq n_j$, in all choices of $\vec{k}$, in all ℓ and in all x.

$$\tag{B.5}$$

Here we have used the fact that, if (B.3) is true,

$$
\left\{ \kappa + B_1\, B_2\, h^{-1}\, n^{\delta-(1/2)} \right\}^q - \kappa^q \leq \kappa^q \left[\exp\left\{ \kappa^{-1}\, B_1\, B_2\, q\, h^{-1}\, n^{\delta-(1/2)} \right\} - 1 \right]
$$
$$
\leq \kappa^{q-1}\, B_1\, B_2\, q\, h^{-1}\, n^{\delta-(1/2)}\, e\,,
$$

since $e^x - 1 \leq ex$ for $0 < x \leq 1$.

Taking the expected value of both sides of (B.4) we deduce that:

$$
\beta_j(\vec{k}) \equiv E\left\{ \prod_{\ell=1}^{q} K\left(\frac{x_{k_\ell} - \widehat{X}_{jik_\ell}}{h_\ell}\right) \right\} - E\left\{ \prod_{\ell=1}^{q} K\left(\frac{x_{k_\ell} - X_{jik_\ell}}{h_\ell}\right) \right\}
$$
$$
= \theta_1(x) \left\{ \kappa^{q-1}\, B_1\, B_2\, q\, h^{-1}\, n^{\delta-(1/2)}\, e + \kappa^q\, P(\tilde{\mathcal{E}}_1) \right\}, \tag{B.6}
$$

where $|\theta(x)| \leq 1$ and and a bound for $P(\tilde{\mathcal{E}}_1)$ is given at (B.2). Therefore, defining $(1 - E)\, R$ to equal $R - E(R)$ for any given random variable R, we have:

$$
(1 - E) \prod_{\ell=1}^{q} K\left(\frac{x_{k_\ell} - \widehat{X}_{jik_\ell}}{h_\ell}\right) - (1 - E) \prod_{\ell=1}^{q} K\left(\frac{x_{k_\ell} - X_{jik_\ell}}{h}\right)
$$
$$
= \Theta_2(x) \left[2\, \kappa^{q-1}\, B_1\, B_2\, q\, h^{-1}\, n^{\delta-(1/2)}\, e + \kappa^q\, \{ I(\tilde{\mathcal{E}}_1) + P(\tilde{\mathcal{E}}_1) \} \right], \tag{B.7}
$$

where $I(\tilde{\mathcal{E}}_1)$ denotes the indicator of the event $\tilde{\mathcal{E}}_1$, and the random variable $\Theta_2(x)$ satisfies $|\Theta_2(x)| \leq 1$ uniformly in the sense of (B.5).

Averaging (B.7) over all i in the range $1 \leq i \leq n_j$ we deduce that

$$
\widehat{\Delta}_j(\vec{k}, x) - \Delta_j(\vec{k}, x)
$$
$$
= \Theta_3(x) \left[2\, \kappa^{q-1}\, B_1\, B_2\, q\, h^{-1}\, n^{\delta-(1/2)}\, e + \kappa^q\, \{ I(\tilde{\mathcal{E}}_1) + P(\tilde{\mathcal{E}}_1) \} \right], \tag{B.8}
$$

where

$$\widehat{\Delta}_j(\vec{k}, x) = (1 - E)\frac{1}{n_j}\sum_{i=1}^{n_j}\prod_{\ell=1}^{q} K\left(\frac{x_{k_\ell} - \widehat{X}_{jik_\ell}}{h_\ell}\right), \tag{B.9}$$

$$\Delta_j(\vec{k}, x) = (1 - E)\frac{1}{n_j}\sum_{i=1}^{n_j}\prod_{\ell=1}^{q} K\left(\frac{x_{k_\ell} - X_{jik_\ell}}{h_\ell}\right), \tag{B.10}$$

and the random variable $\Theta_3(x)$ satisfies $|\Theta_3(x)| \le 1$ uniformly in the following sense:

uniformly in all choices of q such that (B.3) holds, in $j = 1, 2$, in all choices of $\vec{k}$ and in all x. (B.11)

Using the notation at (A.9), (B.6), (B.9) and (B.10), writing $H = \prod_{\ell \le q} h_\ell$, and recalling the definition of $\hat{f}_j$ at (3.2), we see from (B.8) that:

$$
\begin{aligned}
\hat{f}_j(x \mid \vec{k}) &= \frac{1}{n_j H}\sum_{i=1}^{n_j}\prod_{\ell=1}^{q} K\left(\frac{x_{k_\ell} - \widehat{X}_{jik_\ell}}{h_\ell}\right) \\
&= E\{\tilde{f}_j(x \mid \vec{k})\} + H^{-1}\left\{\widehat{\Delta}_j(\vec{k}, x) - \Delta_j(\vec{k}, x) + \Delta_j(\vec{k}, x) + \beta_j(\vec{k})\right\} \\
&= E\{\tilde{f}_j(x \mid \vec{k})\} + H^{-1}\Delta_j(\vec{k}, x) + \Theta_4(x)\, H^{-1}\Big[3\,\kappa^{q-1}\, B_1\, B_2\, q\, h^{-1}\, n^{\delta - (1/2)}\, e \\
&\qquad\qquad + \kappa^q\left\{2\, I(\tilde{\mathcal{E}}_1) + P(\tilde{\mathcal{E}}_1)\right\}\Big],
\end{aligned}
\tag{B.12}
$$

where $|\Theta_4(x)| \le 1$ uniformly in the sense of (B.11).

Step 2: Bounds to deviations of $\Delta_j(\vec{k}, x)$. Define

$$\gamma_j(x) = \sum_{i=1}^{n_j}\prod_{\ell=1}^{q} K\left(\frac{x_{k_\ell} - X_{jik_\ell}}{h_\ell}\right),$$

so that $(1 - E)\,\gamma_j(x) = n_j\,\Delta_j(\vec{k}, x)$. To simplify notation, assume for the present that $k_\ell = \ell$ for $1 \le \ell \le q$, and write f_{jq} for the joint density of $X_{j11}, \ldots, X_{j1q}$. In this notation,

$$\frac{E\{\gamma_j(x)\}}{n_j H} = \int \ldots \int \left\{\prod_{k=1}^{q} K(u_k)\right\} f_{jq}(x_1 - h_1 u_1, \ldots, x_q - h_q u_q)\, du_1\, \ldots\, du_q\,,$$

$$
\begin{aligned}
\frac{\operatorname{var}\{\gamma_j(x)\}}{n_j H} &\le \int \ldots \int \left\{\prod_{k=1}^{q} K(u_k)^2\right\} f_{jq}(x_1 - h_1 u_1, \ldots, x_q - h_q u_q)\, du_1\, \ldots\, du_q \\
&\le \frac{E\{\gamma_j(x)\}\,\kappa^q}{n_j H}\,.
\end{aligned}
$$

We can write $(1 - E)\,\gamma_j = \sum_{i \le n_j} \Delta_{ji}$ where the random variables Δ_{ji}, for $1 \le i \le n_j$, are independent and identically distributed with zero mean and satisfy $P(|\Delta_{ji}| \le \kappa^q) = 1$. Therefore by Bernstein's inequality, writing $\mu_j = E(\tilde{f}_j) = (n_j H)^{-1}\, E(\gamma_j)$, defining $\Delta_j(\vec{k}, x)$ as at (B.10) and taking $0 < \xi \le 1$, we have:

$$
\begin{aligned}
P\left\{H^{-1}\,|\Delta_j(\vec{k}, \cdot)| > \xi\,\mu_j\right\} &= P\{|(1 - E)\,\gamma_j| > \xi\, n_j\, H\, \mu_j\} \\
&\le 2\exp\left\{-\frac{\xi^2\, n_j\, H\, \mu_j/(2\kappa^q)}{1 + (\xi/3)}\right\} \le 2\exp\left\{-\xi^2\, n_j\, H\, \mu_j/(3\kappa^q)\right\}.
\end{aligned}
\tag{B.13}
$$

(The inequality is interpreted in a pointwise sense, for the functions γ_j and μ_j evaluated at arbitrary particular points.)

Step 3: Implications of (B.12) *and* (B.13). The calculations in Step 2 remain valid if we work with the notationally more complex case where $\vec{k}$ is a general vector with components equal to distinct integers chosen between 1 and p. Let $x = (x_1, \ldots, x_p)$ be a p-vector. Given $j = 1$ or 2, write $\mu_j(x \mid \vec{k})$ for the version of $\mu_j(x)$ in this setting, and continue to take $H = \prod_{\ell \leq q} h_\ell$. Assume that $\eta_1, \eta_2, \eta_3 > 0$ satisfy $\eta_1 < \eta_2$ and $\eta_1 + 2\eta_2 + \eta_3 < 1$, let $\mathcal{S}_j(\vec{k})$ denote the set of p-vectors x such that $\mu_j(x \mid \vec{k}) > n^{-\eta_1}$, and put $\xi = \xi(n) = n^{-\eta_2}$. Let $\Delta_j^{(1)}(x \mid \vec{k}) = H^{-1} \Delta_j(\vec{k}, x)$. Then, provided that (A.13) and (A.14) hold, it follows from (B.13) that

$$P\{|\Delta_j^{(1)}(x \mid \vec{k})| > n^{-\eta_2} \mu_j(x \mid \vec{k})\} \leq 2 \exp(-n^{\eta_3}) \text{ for all } q \geq 1, \text{ all } \vec{k} \in \mathcal{K}(q),$$
$$\text{all } h_1, \ldots, h_p \text{ such that (A.13) and (A.14) hold, and all } x \in \mathcal{S}_j(\vec{k}). \tag{B.14}$$

We introduce the following condition:

$$H^{-1}\left(\kappa^{q-1} B_1 B_2 q h^{-1} n^{\delta-(1/2)} e + \kappa^q n^{-C_1}\right) \leq n^{-\eta_2} \mu_j(x \mid \vec{k}), \tag{B.15}$$

where $C_1 > 0$ is the constant in (B.2). If $|\Delta_j^{(1)}(x \mid \vec{k})| \leq n^{-\eta_2} \mu_j(x \mid \vec{k})$ for $j = 1, 2$, and if (B.15) holds, then it follows from (B.2), (B.12) and (B.14) that, if $\vec{k} \in \mathcal{K}(q)$ and $x \in \mathcal{S}_3(\vec{k}) = \mathcal{S}_1(\vec{k}) \cap \mathcal{S}_2(\vec{k})$, we have: $\hat{f}_j(x \mid \vec{k}) = \mu_j(x \mid \vec{k})\{1 + 2\theta_{1j}(x \mid \vec{k}) n^{-\eta_2}\}$, where, in the case $r = 1$, and provided that (A.14) and (B.15) hold,

$$P\{|\theta_{rj}(x \mid \vec{k})| > 1\} \leq 2 \exp\left(-n^{\eta_3}\right) + n^{-C_1}. \tag{B.16}$$

Hence, for the same $\vec{k}$ and x,

$$\hat{g}(x \mid \vec{k}) = \frac{\pi_1 \hat{f}_1(x \mid \vec{k})}{\pi_1 \hat{f}_1(x \mid \vec{k}) + \pi_2 \hat{f}_2(x \mid \vec{k})}$$
$$= \{1 + 2\theta_{2j}(x \mid \vec{k}) n^{-\eta_2}\} \frac{\pi_1 \mu_1(x \mid \vec{k})}{\pi_1 \mu_1(x \mid \vec{k}) + \pi_2 \mu_2(x \mid \vec{k})}, \tag{B.17}$$

where, provided that (A.14) and (B.15) hold, $\theta_{2j}(x \mid \vec{k})$ satisfies (B.16) with $r = 2$.

Write $\mathcal{S}_4(\vec{k})$ for the set of x such that (B.15) holds for $j = 1, 2$. If (A.13) holds then $\mathcal{S}_4(\vec{k}) \subseteq \mathcal{S}_3(\vec{k})$. Recall that $\rho(x \mid \vec{k}) = \pi_2 \mu_2(x \mid \vec{k})/\pi_1 \mu_1(x \mid \vec{k})$, let $A_1 > 0$, and define $\mathcal{S}_5(\vec{k})$ to be the set of all p-vectors x such that $|\rho(x \mid \vec{k}) - 1| > A_1 n^{-\eta_2}$. Since (B.17) holds for all $\vec{k} \in \mathcal{K}(q)$ and all $x \in \mathcal{S}_3(\vec{k})$, with $\theta_{2j}(x \mid \vec{k})$ satisfying (B.16), then an absolute constant A_1 can be chosen such that, for all $\vec{k} \in \mathcal{K}(q)$ and all $x \in \mathcal{S}_6(\vec{k}) = \mathcal{S}_3(\vec{k}) \cap \mathcal{S}_5(\vec{k})$, and provided that (A.13) and (A.14) hold,

$$P\left[\{\hat{g}(x \mid \vec{k}) > \tfrac{1}{2}\} \triangle \{\rho(x \mid \vec{k}) < 1\}\right] \leq 2 \exp\left(-n^{\eta_3}\right) + n^{-C_1}. \tag{B.18}$$

Step 4: Approximating the error rate estimator. Let $\hat{g}_{-ji}$ denote the version of $\hat{g}$ that we obtain if we drop X_{ji} from $\mathcal{X}_j$. The arguments leading to (B.18) give: for all $\vec{k} \in \mathcal{K}(q)$ and all $x \in \mathcal{S}_6(\vec{k})$, and provided that (A.13) and (A.14) hold,

$$P\left[\{\hat{g}_{-1i}(x \mid \vec{k}) > \tfrac{1}{2}\} \triangle \{\rho(x \mid \vec{k}) < 1\}\right] \leq 2 \exp\left(-n^{\eta_3}\right) + n^{-C_1} \text{ for } 1 \leq i \leq n_1,$$
$$P\left[\{\hat{g}_{-2i}(x \mid \vec{k}) \leq \tfrac{1}{2}\} \triangle \{\rho(x \mid \vec{k}) \geq 1\}\right] \leq 2 \exp\left(-n^{\eta_3}\right) + n^{-C_1} \text{ for } 1 \leq i \leq n_2, \tag{B.19}$$

23

where $\mathcal{E}^{(1)} \triangle \mathcal{E}^{(2)}$ denotes the symmetric difference between events $\mathcal{E}^{(1)}$ and $\mathcal{E}^{(2)}$. Put $\vec{h} = (h_1, \ldots, h_q)$. The error rate estimator when the components chosen are those in the vector $\vec{k} = (k_1, \ldots, k_q)$, in that order, is:

$$\widehat{\text{errate}}_q = \widehat{\text{errate}}_q(\vec{k} \,|\, \vec{h}) = \sum_{j=1}^{2} \pi_j \, n_j^{-1} \sum_{i=1}^{n_j} I\{\hat{g}_{-ji}(X_{ji} \,|\, \vec{k}) \bowtie_{1j} \tfrac{1}{2}\}$$

$$= \sum_{j=1}^{2} \pi_j \, n_j^{-1} \sum_{i=1}^{n_j} I\{\rho(X_{ji} \,|\, \vec{k}) \bowtie_{2j} 1\} + \Theta_4(\vec{k}), \tag{B.20}$$

where the symbol $\bowtie_{1j}$ denotes "$\leq$" if $j = 1$ and "$>$" if $j = 2$, and $\bowtie_{2j}$ denotes "$\geq$" if $j = 1$ and "$<$" if $j = 2$;

$$\left|\Theta_4(\vec{k})\right| \leq \sum_{j=1}^{2} \pi_j \, n_j^{-1} \#\{i : X_{ji} \notin \mathcal{S}_6(\vec{k})\}; \tag{B.21}$$

the first identity in (B.20), being equivalent to the definition of $\widehat{\text{errate}}_q(\vec{k} \,|\, \vec{h})$, holds with probability 1; the second identity in (B.20) holds on the set $\mathcal{E}_2$, of which the complement is

$$\tilde{\mathcal{E}}_2 = \left(\bigcup_{i=1}^{n_1} \left[\{\hat{g}_{-1i}(X_{1i} \,|\, \vec{k}) > \tfrac{1}{2}\} \triangle \{\rho(X_{1i} \,|\, \vec{k}) < 1\} \right] \cap \{X_{1i} \in \mathcal{S}_6(\vec{k})\} \right)$$

$$\cup \left(\bigcup_{i=1}^{n_2} \left[\{\hat{g}_{-2i}(X_{2i} \,|\, \vec{k}) \leq \tfrac{1}{2}\} \triangle \{\rho(X_{2i} \,|\, \vec{k}) \geq 1\} \right] \cap \{X_{2i} \in \mathcal{S}_6(\vec{k})\} \right); \tag{B.22}$$

and, in view of (B.19) and (B.22),

$$P(\tilde{\mathcal{E}}_2) \leq 4 \, (n_1 + n_2) \, \exp\left(- n^{\eta_3} \right) + 2 \, (n_1 + n_2) \, n^{-C_1}. \tag{B.23}$$

Returning to the bound on $|\Theta_4(\vec{k})|$ at (B.21), we see that

$$\left|\Theta_4(\vec{k})\right| \leq \sum_{j_1=1}^{2} \sum_{j_2=1}^{2} n_{j_2}^{-1} \left(N_{j_1 j_2 1} + N_{j_1 j_2 2} \right) + \sum_{j=1}^{2} n_j^{-1} N_j, \tag{B.24}$$

where

$$N_{j_1 j_2 1} = \sum_{i=1}^{n_{j_2}} I\{\mu_{j_1}(X_{j_2 i} \,|\, \vec{k}) \leq n^{-\eta_1}\},$$

$$N_{j_1 j_2 2} = \sum_{i=1}^{n_{j_2}} I\left\{ H^{-1} \left(\kappa^{q-1} B_1 B_2 q \, h^{-1} n^{\delta - (1/2)} e + \kappa^q n^{-C_1} \right) \leq n^{-\eta_2} \mu_{j_1}(X_{j_2 i} \,|\, \vec{k}) \right\}$$

and $N_j = \sum_{i \leq n_j} I\{|\rho(X_{ji} \,|\, \vec{k}) - 1| \leq A_1 n^{-\eta_2}\}$ denote independent random variables having the binomial $\text{Bi}(n_{j_2}, r_{j_1 j_2})$, $\text{Bi}(n_{j_2}, s_{j_1 j_2})$ and $\text{Bi}(n_j, t_j)$ distributions, respectively, with $r_{j_1 j_2} = P\{\mu_{j_1}(X_{j_2 1} \,|\, \vec{k}) \leq n^{-\eta_1}\}$,

$$s_{j_1 j_2} = P\left\{ H^{-1} \left(\kappa^{q-1} B_1 B_2 q \, h^{-1} n^{\delta - (1/2)} e + \kappa^q n^{-C_1} \right) \leq n^{-\eta_2} \mu_{j_1}(X_{j_2 i} \,|\, \vec{k}) \right\}$$

and $t_j = P\{|\rho(X_{j1} \,|\, \vec{k}) - 1| \leq A_1 n^{-\eta_2}\}$. Bernstein's inequality implies that if N has the $\text{Bi}(m, r)$ distribution, and $0 \leq \epsilon \leq \tfrac{1}{2}$, then

$$P(N - mr > m\epsilon) \leq \exp\left(- m\epsilon^2 \right), \quad P(|N - mr| > m\epsilon) \leq 2 \exp\left(- m\epsilon^2 \right). \tag{B.25}$$

24

Using (B.24) and the first part of (B.25) we deduce that

$$P\left\{ \left| \Theta_4(\vec{k}) \right| > \sum_{j_1=1}^{2} \sum_{j_2=1}^{2} (r_{j_1 j_2} + s_{j_1 j_2} + 2\epsilon) + \sum_{j=1}^{2} (t_j + \epsilon) \right\} \leq 10 \sum_{j=1}^{2} \exp\left(- n_j \, \epsilon^2 \right), \qquad \text{(B.26)}$$

and using the second part of (B.25) we obtain:

$$P\left[\left| \sum_{i=1}^{n_j} (1 - E) \, I\left\{ \rho(X_{ji} \,|\, \vec{k}) \bowtie_{2j} 1 \right\} \right| > n_j \, \epsilon \right] \leq 2 \, \exp\left(- n_j \, \epsilon^2 \right). \qquad \text{(B.27)}$$

Recall that $\mathrm{errate}_q(\vec{k})$ is defined at (A.10). Combining the second identity in (B.20); the fact that that result holds on the set $\mathcal{E}_2$, the probability of the complement of which is bounded by (B.23); and (B.26) and (B.27); we see that, if

$$\max\left\{ \max_{j_1, j_2 \in \{1,2\}} \max_{j=1,2} \max(r_{j_1 j_2}, s_{j_1 j_2}), \max_{j=1,2} t_j \right\} \leq \epsilon \qquad \text{(B.28)}$$

then

$$P\left\{ \left| \widehat{\mathrm{errate}}_q(\vec{k} \,|\, \vec{h}) - \mathrm{errate}_q(\vec{k}) \right| > 20 \, \epsilon \right\}$$

$$\leq 4 \left\{ (n_1 + n_2) \exp\left(- n^{\eta_3} \right) + n^{-C_1} \right\} + 12 \sum_{j=1}^{2} \exp\left(- n_j \, \epsilon^2 \right)$$

$$\leq 24 \left[(n_1 \vee n_2) \left\{ \exp\left(- n^{\eta_3} \right) + n^{-C_1} \right\} + \exp\left\{ - (n_1 \wedge n_2) \, \epsilon^2 \right\} \right]. \qquad \text{(B.29)}$$

Step 5: Choosing successive features. Let k_1^0 denote the value of k that minimises $\mathrm{errate}(k)$ (here k is a scalar, not a vector, so $q = 1$), and, given $k_1^0, \ldots, k_{q-1}^0$, let $k = k_q^0$ be the minimiser of $\mathrm{errate}_q(k_1^0, \ldots, k_{q-1}^0, k)$. In this step we assume not only that $k_1^0, \ldots, k_{q_0}^0$ are uniquely defined, but also that (A.11) holds for $1 \leq q \leq q_0$, where $\epsilon > 0$ is as in (B.28) and (B.29). We suppose too that (A.12) and (A.13) hold; it is the analogue of (B.28) uniformly over values of q and $\vec{k}$.

In the first step of empirical error rate minimisation, where $q = 1$, we choose $k = \hat{k}_1$ to minimise $\widehat{\mathrm{errate}}(k)$. This involves searching among p different features. More generally, given values of $\hat{k}_1, \ldots, \hat{k}_{q-1}$, in step q we choose $\vec{k} = (\hat{k}_1, \ldots, \hat{k}_q)$ to minimise $\widehat{\mathrm{errate}}(\hat{k}_1, \ldots, \hat{k}_{q-1}, k)$ by searching among $p - q + 1$ distinct values of k, to determine $\hat{k}_q$. If $q \leq q_0$ then the total number of searches involved is less than pq_0. If (A.11) and (A.12) hold then, in view of (B.29), the probability of the event $\mathcal{F}$ that $\hat{k}_q = k_q^0$ for $1 \leq q \leq q_0$ satisfies:

$$1 - P(\mathcal{F}) \leq 24 p q_0 \left[(n_1 \vee n_2) \left\{ \exp\left(- n^{\eta_3} \right) + n^{-C_1} \right\} + \exp\left\{ - (n_1 \wedge n_2) \, \epsilon^2 \right\} \right].$$

This is equivalent to the first inequality in (A.16).