FEATURE SELECTION WHEN THERE ARE MANY INFLUENTIAL FEATURES

Peter Hall[12] Jiashun Jin[3] Hugh Miller[1]

SUMMARY. Recent discussion of the success of feature selection methods has argued that focusing on a relatively small number of features has been counterproductive. Instead, it is suggested, the number of significant features can be in the thousands or tens of thousands, rather than (as is commonly supposed at present) approximately in the range from five to fifty. This change, in orders of magnitude, in the number of influential features, necessitates alterations to the way in which we choose features and to the manner in which the success of feature selection is assessed. In this paper we suggest a general approach that is suited to cases where the number of relevant features is very large, and we consider particular versions of the approach in detail. We propose ways of measuring performance, and we study both theoretical and numerical properties of the proposed methodology.

KEYWORDS. Change-point analysis, classification, dimension reduction, logit model, maximum likelihood, ranking, thresholding, variable selection.

SUBJECT CLASSIFICATION. Primary 62G99, Secondary 62H30.

SHORT TITLE. Feature selection.

1 INTRODUCTION

In this paper we develop statistical methods for determining features that enable effective discrimination between two populations of very high dimensional data, when the number of component-wise differences that provide leverage for discrimination is relatively large but the sizes of those differences are potentially small. By way of contrast, conventional approaches to solving this problem tend to rely on relatively large differences and relatively small numbers of components where differences occur.

[1]Department of Mathematics and Statistics, The University of Melbourne, VIC 3010, Australia
[2]Department of Statistics, University of California, Davis, CA 95616, USA
[3]Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213, USA

In such problems it is generally going to be of substantial practical interest to identify, with reasonable accuracy, the components that have greatest leverage for correct discrimination. Simply constructing a classifier, which might depend in a difficult-to-determine way on differences between two populations, is generally not going to provide all the information that is sought. However, particularly when the number of such components is large, we may not be able to identify the components without error. How accurate can we be, and in what circumstances is accuracy high? In this paper we shall endeavour to answer these questions.

Achieving reasonable accuracy can involve relatively computer-intensive methods, for example algorithms that need $O(p^2)$ rather than $O(p)$ time if the problem is p-dimensional. However, if we use an initial, deterministic dimension reduction step, which decreases dimension to q where $q \ll p$, then $O(p^2)$ calculations can be reduced to $O(p \log p + q^2)$, where $p \log p$ is the computational cost of ordering the initial p components. In many cases we expect q to be a rather crude upper bound to the true number, r say, of components that impact on performance of the classifier. The four-stage algorithm that we shall introduce in section 2 enables us to reduce computational expense from $O(p \log p + q^2)$ to $O(p \log p + r^2)$. (These order-of-magnitude calculations ignore the effects of training sample size, n say, since in the problems we are considering n is typically much less than p, q or r and so has relatively little impact on the final result.)

Support for the conjecture that r can be quite large, for example in genomics problems, has been given by Goldstein (2009), who, in the words of J.N. Hirschhorn in the same issue of the *New England Journal of Medicine*, "builds a speculative mathematical model and infers that there will be tens of thousands of common variants influencing each disease and trait" (Hirschhorn, 2009). Goldstein's (2009) calculations are also consistent with r being in the thousands, not just the tens of thousands:

> ... the genetic burden of common diseases must be mostly carried by large numbers of rare variants. In this theory, schizophrenia, say, would be caused by combinations of 1,000 rare genetic variants, not of 10 common genetic variants.

(See Wade, 2009.) Kraft and Hunter (2009) argue that "many, rather than few, variant risk alleles are responsible for the majority of the inherited risk of each common disease." Again, r is large rather than small.

In the discussion above one might interpret p and r as representing numbers of single nucleotide polymorphisms (SNPs), alleles, or perhaps genes. There are believed to be between 10 and 30 million SNPs on a human chromosome, and some 25,000 genes. However, genomic analyses based on decoding the full DNA of individuals who suffer from specific conditions (Wade, 2009) increase the values of both p and r by orders of magnitude. (In practice r will be chosen empirically, and so will actually be a function of the data, but at the level of the discussion in the present section there is little to be gained by making this distinction.)

Methods for feature selection based on the linear model are generally considered only in cases where the number of features is relatively small. Otherwise, the value of the response variable can be unreasonably insensitive to changes in a single feature. Examples of approaches founded on the linear model include the nonnegative garrotte (e.g. Breiman, 1995, 1996; Gao, 1998), the lasso (Tibshirani, 1996), the Dantzig selector (Candes and Tao, 2007), and related techniques (e.g. Donoho and Huo, 2001; Fan and Li, 2001, 2006; Donoho and Elad, 2003; Tropp, 2005; Donoho, 2006a, 2006b; Fan and Ren, 2006; Fan and Fan, 2008). The feature-ranking approach that we consider is more closely related to correlation-based approaches of Fan and Lv (2008) and Hall and Miller (2008), but it is does not assume the existence of a response variable. Instead it utilises class labels via a logistic model. Monograph-length treatments of classifiers and related methodology include those of Duda *et al.* (2001), Hastie *et al.* (2001) and Shakhnarovich *et al.* (2005).

Section 2.1, immediately below, proposes a general algorithm for determining features that appear to have significant influence on whether a data vector comes from one population or another. Sections 2.2 and 2.3 discuss particular approaches to implementing the algorithm, and section 2.4 addresses computational labour. Section 3 develops theoretical properties of the component ranking stage in the algorithm, under the assumption that there is a large number of relatively small differences among component means. Section 4 explores properties of the adaptive dimension reduction stage, section 5 discusses numerical properties, and section 6 outlines technical arguments.

2 METHODOLOGY

2.1. Data and algorithm. Denote the two populations of interest by Π_0 and Π_1. Training data from each are acquired as p-vectors $X_i = (X_{i1}, \ldots, X_{ip})$, for $1 \le i \le n$.

We also record the value of a label I_i, for each i; it equals 0 or 1, indicating the index of the population from which X_i came.

One potential algorithm for identifying indices j, of vector components X_{ij} which capture differences between Π_0 and Π_1, has four stages, (1)–(4) below. The goal of the algorithm is to determine empirically a set $\{\hat\jmath_1, \ldots, \hat\jmath_r\}$ of indices, a subset of $\{1, \ldots, p\}$, such that the features with indices $\hat\jmath_1, \ldots, \hat\jmath_r$ have significant influence on whether X_i comes from Π_0 or Π_1. Those features can then be combined into a classifier, for example the support vector machine or a centroid-based method, to effect discrimination.

(1) Component ranking. Using a method such as that suggested in section 2.2, rank all components in terms of their individual influence on I_i, interpreted as a zero-one response variable. This stage takes $O(p \log p)$ time to run, and produces a permutation $\hat\jmath_1, \ldots, \hat\jmath_p$, say, of $1, \ldots, p$, where the order of the sequence $\hat\jmath_1, \ldots, \hat\jmath_p$ is of major importance and signifies that, for each k, the component with rank $\hat\jmath_k$ has greater leverage than the component with rank $\hat\jmath_{k+1}$ on a measure of our ability to predict I_i from X_i.

(2) Deterministic dimension reduction. Truncate at q (where $1 \le q \le p$) the sequence we derived in step (1). From this point we work only with q-vectors comprised of the components with indices $\hat\jmath_1, \ldots, \hat\jmath_q$. The value of q is determined largely by our computational resources, bearing in mind that the computational expense of constructing the classifier could be as high as $O(q^2)$.

(3) Adaptive dimension reduction. In this stage we use an empirical method to reduce dimension from q, chosen in stage (3), to r, so that the final choice of feature indices is $\hat\jmath_1, \ldots, \hat\jmath_r$. Potential approaches are discussed in section 2.3, and include methods based on: (3a) thresholding, (3b) change-point methods, or (3c) application of classifiers to blocks of components.

(4) Backing and filling. In practice it can be advantageous to rerun stage (3) of the algorithm using several of the values of j chosen early in stage (2), or early in the implementation of stage (3), bearing in mind that there is potential for noise in the choice of $\hat\jmath_1$, for example, to throw the algorithm off course for a period. At this point we could, for example, experiment with different choices of block size in method (3c).

2.2. Method for ranking components. Given an index j between 1 and p, and scalar parameters α and β, we capture the relationship between I_i and X_{ij} by assuming a logit model:

$$P(I_i = 0 \,|\, X_{ij}) = \{1 + \exp(\alpha + \beta\, X_{ij})\}^{-1}\,, \quad P(I_i = 1 \,|\, X_{ij}) = 1 - P(I_i = 0 \,|\, X_{ij})\,.$$

The likelihood of I_i, given X_{ij}, is

$$L_{ij}(I_i \,|\, \alpha, \beta \,; X_{ij}) = \Big(\frac{t_{ij}}{1+t_{ij}}\Big)^{I_i} \Big(\frac{1}{1+t_{ij}}\Big)^{1-I_i} = \frac{t_{ij}^{I_i}}{1+t_{ij}} \,,$$

where $t_{ij} = \exp(\alpha + \beta\, X_{ij})$. Therefore the negative log-likelihood is

$$\ell_{ij}(\alpha, \beta) = -\log L_{ij}(I_i \,|\, \alpha, \beta \,; X_{ij}) = -I_i\,(\alpha + \beta\, X_{ij}) + \log\big\{1 + \exp(\alpha + \beta\, X_{ij})\big\}\,, \quad (2.1)$$

and its counterpart for $X_{1j}, \ldots, X_{nj}$ is

$$\ell_j(\alpha, \beta) = \frac{1}{n} \sum_{i=1}^{n} \ell_{ij}(\alpha, \beta)\,. \quad (2.2)$$

Define $(\hat\alpha_j, \hat\beta_j)$ to be the value of (α, β) that minimises $\ell_j(\alpha, \beta)$, and put

$$\hat\ell_j = \ell_j(\hat\alpha_j, \hat\beta_j)\,. \quad (2.3)$$

The ordering $\hat\jmath_1, \ldots, \hat\jmath_p$ mentioned in step (1) of the algorithm in section 2.1 is determined by the values of $\hat\ell_j$. Specifically, $\hat\ell_{\hat\jmath_1} \le \ldots \le \hat\ell_{\hat\jmath_p}$.

2.3. Methods for adaptive dimension reduction. Several approaches are feasible, including: (3a) *Thresholding.* Here we compute, from the data, a subsidiary criterion $\tilde\ell_j$, for $1 \le j \le p$ (we might simply choose $\tilde\ell_j \equiv 0$) and a threshold t; we take $k_0 \ge 0$ to be an integer; and we define $r \in [k_0 + 1, q]$, a function of the data, to be the least integer in that range such that $\hat\ell_{\hat\jmath_{r+k}} - \tilde\ell_{\hat\jmath_{r+k}} > t$ for $1 \le k \le k_0$. See section 4 for an example. (3b) *Change-point methods.* Here we look for a change-point in the sequence $\hat\ell_{\hat\jmath_1}, \ldots, \hat\ell_{\hat\jmath_p}$, and we take $\hat\jmath_r$ to be the location of that point. (There is a vast literature on methodology and theory for change-point detection. It includes book-length accounts by Carlstein *et al.* (1994), Csörgő and Horváth (1997), Chen and Gupta (2000) and Wu (2005).) (3c) *Application of classifiers.* For $k \ge 1$, let $\mathcal{B}_k = \{\hat\jmath_{(k-1)b+1}, \ldots, \hat\jmath_{kb}\}$ denote the kth block of feature indices; here, b denotes block length. (Theoretical considerations suggest that taking $b \sim \text{const.}\, n$ is appropriate.) In step s of stage (3c) we construct the classifier that is based on the training data vectors where all but the components with indices in $\cup_{1 \le k \le s} \mathcal{B}_k$ have been stripped away. We use cross-validation to measure classifier performance, and in this way we determine whether progressing from step s to step $s+1$ gives an improvement. If it does not then, subject to the "jiggling" suggested in stage (4), we stop at step s. If performance is improved by passing to the $(s+1)$st block then we proceed to step $s+1$, where we again assess performance.

However, this approach can be biased in favour of low apparent error rate, without having the same impact on actual error rate; see section 5.3 for discussion.

2.4. Duration of algorithm. Stage (3) of the algorithm takes $O(r^2)$ time to complete, where r, in the range $1 \le r \le q$, denotes the final number of components on which we determine that the classifier should depend. In particular, the algorithm concludes with a list of r components, say $\hat{j}_{\ell_1}, \ldots, \hat{j}_{\ell_r}$ where $\{\hat{j}_{\ell_1}, \ldots, \hat{j}_{\ell_r}\} = \cup_{k \le \hat{s}} \mathcal{B}_k \subseteq \{\hat{j}_1, \ldots, \hat{j}_q\}$, on which the final classifier is based. The $O(r^2)$ figure is derived as follows: Constructing the classifier from s batches takes $O(sb)$ time, so the total time needed is $O(\sum_{1 \le s \le \hat{s}} sb) = O(\hat{s}^2 b) = O(r^2)$. Here, since b is generally of order n (see section #) and $\hat{s}^2 b = O(r^2 b)$, we have replaced $r^2 b$ by r^2 since, in the problems we are treating, n is generally so much less than r that it can be treated as fixed. Provided the number of reruns in stage (4) is only $O(1)$, the order of magnitude of the time taken to run the algorithm to completion is $O(p \log p + r^2)$.

2.5. Discussion. The procedures above are intended to reflect methodologies already used in practice. Our main aim is to show that such techniques can be used to address not just contemporary problems where there is believed to be considerable sparsity and only a small number of significant features (for example, five or ten genes out of thousands or tens of thousands), but also reduced sparsity and a larger total number of features (for instance, thousands or tens of thousands of DNA sequences out of tens or hundreds of thousands of possibilities). Additionally we show that the methods continue to work well under minimal distributional assumptions (for example, normality is not needed), and minimal conditions about the correlation structure among features. In all these senses, procedures such as those described above are particularly versatile.

3 PROPERTIES OF FEATURE RANKING

3.1. Main result on ranking. Let $\pi = E(I_i)$ denote the proportion of data that come from population Π_1. We assume below that $0 < \pi < 1$, and that when the training data are drawn randomly from the union of Π_0 and Π_1, the prior probability that any given datum X_i is from Π_1 equals π. Therefore the corresponding probability for Π_0 is $1 - \pi$. We take n, representing the total size of the training sample, to be the key asymptotic parameter, and interpret the dimension, p, of X_i as a function of n.

Next we describe our model. Write $X_i = (X_{i1}, \ldots, X_{ip})$, and, when X_i comes from

Π_1, take

$$X_{ij} = \mu_j + Z_{ij} \tag{3.1}$$

where $\mu_j \geq 0$ are constants and, for simplicity, we assume that the variables Z_{ij} are identically distributed as Z, say. (This condition can be relaxed; see (2.4) below.) Take $X_{ij} = Z_{ij}$ when X_i is drawn from Π_0. The vectors $(Z_{i1}, \ldots, Z_{in})$ and variables I_i, for $1 \leq i \leq n$, are assumed to be totally independent. We allow the μ_js to be functions of n.

Write $\hat{\alpha}_j$ and $\hat{\beta}_j$ for the values of α and β that jointly minimise $\hat{\ell}_j(\alpha, \beta)$ within radius n^{-c} of $(\alpha_0, 0)$, where $\alpha_0 = \log\{\pi/(1-\pi)\}$, for any given $c \in (0, \frac{1}{2})$. (A separate argument can be used to prove that if $\hat{\alpha}_j$ and $\hat{\beta}_j$ are chosen without constraint then, under the conditions of Theorem 1 below, they satisfy $\hat{\alpha}_j = \alpha_0 + O_p(n^{-c})$ and $\hat{\beta}_j = O_p(n^{-c})$ uniformly in $1 \leq j \leq p$, for some $c \in (0, \frac{1}{2})$.) Define

$$S_j = \frac{\sum_i (I_i - \pi)\, X_{ij}}{n\, \pi\, (1-\pi)}\,, \tag{3.2}$$

and note that $E(S_j) = 0$. Put $\lambda = (n^{-1} \log n)^{1/2}$.

Theorem 1. *Assume that for each n, $|\mu_j| \leq$ const. λ for $1 \leq j \leq p$; that $p = p(n) \to \infty$ and, for constants $B_1 > 0$ and $B_2 > 2\max(B_1 + 3, 2B_1)$, $p = O(n^{B_1})$ and $0 < E|Z|^{B_2} < \infty$; and that $E(Z) = 0$. Then, uniformly in $1 \leq j \leq p$,*

$$\hat{\ell}_j = \ell_j(\hat{\alpha}_j, \hat{\beta}_j) = R - \tfrac{1}{2}\,\pi\,(3 - 2\,\pi)\,\big(EZ^2\big)^{-1}\,(S_j + \mu_j)^2 + O_p\big(\lambda^3\big)\,, \tag{3.3}$$

where the random variable $R = R(n)$ does not depend on j. More particularly, the $O_p(\lambda^3)$ term in (3.3) can be written as $\Theta_j\,\lambda^3$, where, for a constant B_3 depending on B_1 and B_2, and with $B_4 = \frac{1}{4}\,\{B_2 - 2\max(B_1 + 3, 2B_1)\} - \epsilon$ for any $\epsilon > 0$, the random variables Θ_j, for $1 \leq j \leq p$, satisfy:

$$\sum_{j=1}^{p} P(|\Theta_j| > B_3) = O\big(n^{-B_4}\big) \tag{3.4}$$

as $n \to \infty$.

The statistic S_j is, up to normalisation, the well-known z-score statistic for testing whether the jth feature is significant or not; see for example Donoho and Jin (2008, 2009) and Jin (2009). In (3.3), since the first term, R, does not depend on j then the second term is the one that reflects the strengths of individual features. As a result,

ranking features according to $\hat{\ell}_j$ gives close, but not necessarily the same, results as ranking features according to $|S_j + \mu_j|$.

The assumption that each Z_{ij} has the same distribution is made to simplify discussion and notation, and is readily relaxed. For example, it suffices to assume that each Z_{ij}, for $1 \le i \le n$, is distributed as Z_j, say, where, instead of the assumptions imposed on Z in the theorem, we ask that:

$$\begin{aligned} &E(Z_j) = 0,\ E(Z_j^2) \text{ is bounded away from zero and infinity uniformly} \\ &\text{in } j, \text{ and } P(|Z_j| > x) \le P(|Z| > x) \text{ for all } x > 0 \text{ and some random} \\ &\text{variable } Z, \text{ where } E|Z|^{B_2} < \infty. \end{aligned} \tag{3.5}$$

In this case the moment $E(Z^2)$ in (3.3) would be replaced by $E(Z_j^2)$. The conclusions that we draw, below, from Theorem 1 are unchanged, provided we interpret μ_j as $\mu_j/(EZ_j^2)^{1/2}$ during discussion.

Although we ask that the vectors X_i be independent, we make no assumption about the relationships among their components. For example, the values of $Z_{i1}, \ldots, Z_{ip}$ can be highly correlated (indeed, in an extreme case, equal to one another) or completely independent. The latter instance is actually the most difficult, in terms of rigorously establishing that (3.3) and (3.4) hold. At the other end of the spectrum, the case where $Z_{i1} = \ldots = Z_{ip}$ with probability 1 is trivial, since there effectively only a single component index, with different candidate values for the mean, has to be treated.

3.2. Expected number of misrankings. Assume that some of the μ_js are zero and all the others are strictly positive. Ideally, we would like the criterion $\hat{\ell}_j$ to be a good indicator of the positivity of μ_j, in particular to take a lesser (or larger negative) value if μ_j is positive than it does when $\mu_j = 0$. Reflecting this aspiration, if there exist component indices j_1 and j_2 such that $\mu_{j_1} > 0$ and $\mu_{j_2} = 0$, but $\hat{\ell}_{j_2} < \hat{\ell}_{j_1}$, then we shall say that a misranking has occurred. The expected total number of misrankings,

$$\nu_{\text{misrank}} = \sum_{j_1\,:\,\mu_{j_1}>0} \sum_{j_2\,:\,\mu_{j_2}=0} P(\hat{\ell}_{j_2} < \hat{\ell}_{j_1})\,,$$

is a measure of the performance of $\hat{\ell}_j$ as a criterion for distinguishing between positive and zero values of μ_j; lower values of ν_{misrank} correspond to higher performance.

Since the random variable S_j in (3.2) and (3.3) has standard deviation of size $n^{-1/2}$ then, if the positive μ_js are of smaller order than $n^{-1/2}$, with probability converging to $\frac{1}{2}$ any attempt to rank any pair of means μ_j using the values of $\hat{\ell}_j$ will produce the wrong result about half the time. The following theorem makes this clear. Let p_1

denote the number of indices j for which $\mu_j > 0$, and put

$$\sigma^2_{j_1 j_2,\pm} = \pi\,(1-\pi)\,\mathrm{var}\,(Z_{1j_1} \pm Z_{1j_2}) \tag{3.6}$$

for respective choices of the plus and minus signs.

Theorem 2. *Assume the conditions of Theorem 1, that* $\sup_{j\le p} \mu_j = o(n^{-1/2})$*, and that* $\sigma^2_{j_1,j_2,\pm}$ *is bounded away from zero uniformly in* j_1*,* j_2 *and choices of the* $\pm$ *signs. Then* $P(\hat\ell_{j_2} < \hat\ell_{j_1}) \to \frac12$ *uniformly in* $1 \le j_1, j_2 \le p$*, in particular uniformly in pairs* j_1, j_2 *such that* $\mu_{j_1} > 0$ *and* $\mu_{j_2} = 0$*. Moreover,* $\nu_{\mathrm{misrank}} = \frac12\, p_1\,(p-p_1) + o\{p_1\,(p-p_1)\}$ *as* $n \to \infty$*.*

Likewise, if the positive μ_js are of size $n^{-1/2}$ then the probability of incorrectly ranking the j_1th component lower than the j_2th component, even though $\mu_{j_1} > 0$ and $\mu_{j_2} = 0$, does not converge to zero. The next theorem quantifies this property. There we define Φ to be the standard normal distribution function.

Theorem 3. *Assume the conditions of Theorem 1, and that each nonzero* μ_j *equals* $c\,n^{-1/2}$ *where* $c > 0$*. Then*

$$P\big(\hat\ell_{j_2} < \hat\ell_{j_1}\big) = \Phi(-c/\sigma_{j_1 j_2,+})\,\Phi(c/\sigma_{j_1 j_2,-}) + \Phi(c/\sigma_{j_1 j_2,+})\,\Phi(-c/\sigma_{j_1 j_2,-}) + o(1)\,,$$

uniformly in j_1, j_2 *such that* $\mu_{j_1} > 0$ *and* $\mu_{j_2} = 0$*. Furthermore,*

$$\begin{aligned}\nu_{\mathrm{misrank}} \;=\; \sum_{j_1\,:\,\mu_{j_1}>0} \;\sum_{j_2\,:\,\mu_{j_2}=0} \Big\{ &\Phi(-c/\sigma_{j_1 j_2,+})\,\Phi(c/\sigma_{j_1 j_2,-}) \\ &+\Phi(c/\sigma_{j_1 j_2,+})\,\Phi(-c/\sigma_{j_1 j_2,-})\Big\} + o\{p_1\,(p-p_1)\}\end{aligned}$$

as $n \to \infty$*.*

Here, by default, $\Phi(-\infty) = 0$ and $\Phi(\infty) = 1$. This is relevant when $\sigma_{j_1 j_2,\pm} = 0$.

If the number of components where the mean is positive is large, for example if it equals a non-negligible proportion of the total number, p, of components, then the number of misrankings can generally not be reduced to low levels unless we take the nonzero means to be a little larger than $n^{-1/2}$ in order of magnitude terms. It is enough to take the positive mean to be a logarithmic factor larger; specifically, the mean should equal $c\,\lambda$ where, as before, $c > 0$ and $\lambda = (n^{-1}\,\log n)^{1/2}$. Theorem 4, below, shows that in this case the expected number of misrankings can be reduced to a quantity of

smaller order than p, or even to a number that converges to zero polynomially fast, depending on how large we choose c.

As a prelude to stating Theorem 4 we introduce an assumption which asks that the random variables Z_j satisfy a pairwise Cramér continuity condition. Here it is convenient to assume that there exists an infinite stochastic process $Z_1, Z_2, \ldots$ such that:

(a) Each Z_j has the distribution of Z (in this sense the process $Z_1, Z_2, \ldots$ is weakly stationary), (b) if $X_i = (X_{i1}, \ldots, X_{ip})$ is drawn from Π_0 then $X_{i1}, \ldots, X_{ip}$ has the same joint distribution as $Z_1, \ldots, Z_p$, and (c) if X_i is drawn from Π_1 then $X_{i1}, \ldots, X_{ip}$ has the same joint distribution as $Z_1 + \mu_1, \ldots, Z_p + \mu_p$, where the μ_js are the nonnegative constants introduced prior to Theorem 1. (3.7)

The Cramér continuity condition we impose is the following:

$$\limsup_{t \to \infty} \sup_{|t_1|+|t_2|>t} \sup_{1 \le j_1 < j_2 < \infty} |E\{\exp(it_1 Z_{j_1} + it_2 Z_{j_2})\}| < 1\,, \tag{3.8}$$

where on this occasion $i = \sqrt{-1}$. For example, (3.8) would hold if the process Z_j were strictly stationary and each pair (Z_{j_1}, Z_{j_2}) had a joint density $f_{j_1 j_2}$ that satisfied $\sup_{j_1, j_2} \iint |\ddot{f}_{j_1 j_2}| < \infty$, where $\ddot{f}_{j_1 j_2}(x_1, x_2) = (\partial^2 / \partial x_1 \partial x_2)\, f_{j_1 j_2}(x_1, x_2)$. It would also hold if the variables Z_j were independent with a common nonsingular distribution.

Recall that p_1 equals the number of indices j such that $\mu_j > 0$, and that $B_4 = B_4(\epsilon) = \frac{1}{4}\{B_2 - 2\max(B_1 + 3, 2B_1)\} - \epsilon$ where $\epsilon > 0$. Define $\sigma_{j_1 j_2, \pm}$ by (3.6) and put $\kappa_n = (\log n)^{1/2}$.

Theorem 4. *Assume the conditions of Theorem 1, that* (3.7) *and* (3.8) *hold, and that B_2, in the moment condition $E|Z|^{B_2} < \infty$, is so large that for some $\epsilon > 0$, $p_1\,(p - p_1) = o(n^{B_4})$. Take each nonzero μ_j to equal $c\,(n^{-1}\,\log n)^{1/2}$, where $c > 0$. Then*

$$\begin{aligned} \nu_{\text{misrank}} \;= \{1 + o(1)\} \sum_{j_1\,:\,\mu_{j_1} > 0} \; \sum_{j_2\,:\,\mu_{j_2} = 0} \; & \{\Phi(-c\,\kappa_n / \sigma_{j_1 j_2, +}) \\ & + \Phi(-c\,\kappa_n / \sigma_{j_1 j_2, -})\} + o(1) \end{aligned} \tag{3.9}$$

as $n \to \infty$.

Elucidation of (3.9) requires information about the covariance of the process Z_j, in (3.7). For simplicity let us assume that the variables Z_j are uncorrelated. Then by

(3.6), $\sigma^2_{j_1 j_2, \pm} = 2\,\pi\,(1-\pi)\,E(Z^2) \equiv (2c_0)^{-1}$, say, for either choice of the $\pm$ signs. Hence (3.9) implies that

$$\nu_{\rm misrank} = \{1 + o(1)\}\, p_1\,(p - p_1)\,\{2\,\pi\,c\,(2\,c_0\,\log n)^{1/2}\}^{-1}\, n^{-c_0 c^2} + o(1)\,.$$

Therefore, if c is chosen so large that $p_1\,(p-p_1)\,(\log n)^{-1/2}\,n^{-c_0 c^2} \to 0$ then the expected number of misranks will converge to zero. For smaller positive values of c the expected number will be of smaller order than the potential number of misranks, $p_1\,(p-p_1)$, but it will not necessarily be negligible itself.

The results in Theorems 2–4 have benefited from a simplification afforded by the assumption that the variables Z_{ij} all have the same distribution, and in particular have the same variance. As noted below Theorem 1, that condition can be relaxed and the assumption (3.5) imposed instead. In practice, however, one could standardise, in a componentwise fashion, the values of X_{ij} for scale, and in that case it is possible to state versions of Theorems 2–4 in settings where $E(Z_{ij}^2)$ varies with j. The model that we have been using, i.e. $X_{ij} = Z_{ij} + c\,(n^{-1}\,\log n)^{1/2}\,I_i$ where the Z_{ij}s are independent, is (for moderate n) a good approximation to the standardised form $X'_{ij} = Z'_{ij} + c\,(n^{-1}\,\log n)^{1/2}\,I_i$, where the Z'_{ij}s satisfy (3.5). Detailed arguments here are similar to those given by Hall and Wang (2008).

3.3. Effects of dependence of the process Z_j on interpretations of (3.3). The expected value of the number of misrankings, which we treated in section 3.3, is not as much affected by dependence among components of the Z_j process as are other aspects of the distribution of the number of misrankings. For example, if the Z_js (in the stochastic process $Z_1, Z_2, \ldots$ introduced in (3.7)) are all independent then the quantities S_j, defined at (3.2) and on which the values of $\hat\ell_j$ predominantly depend (see (3.3)), are also independent, and so decisions based on the respective values of $\hat\ell_j$ are made virtually independently of one another. In this case the variance of the total number of misrankings is relatively low. However, if the Z_js are highly correlated then the variance can be higher, although it depends on how the positive means are distributed among the components of X_i. In the present section we briefly discuss these issues.

Let the process $Z_1, Z_2, \ldots$ in (3.7) be ζ-dependent, meaning that any subsequence $Z_{j_1}, \ldots, Z_{j_k}$ such that $j_{\ell+1} - j_\ell > \zeta$, for each ℓ, is comprised entirely of independent random variables. We permit ζ to diverge with n, and we suppose that p_1, the number of nonzero values of μ_j, can also increase with n and that $\limsup_{n\to\infty} p_1/p < 1$. One approach to arranging the nonzero means is to distribute them randomly, for example taking $\mu_j = J_j\,\mu$ for $1 \le j \le p$, where $J_1, \ldots, J_p$ is a random permutation of p_1 ones

and $p - p_1$ zeros and is independent of the Z_js in (3.7). In this setting the clustering that arises through dependence is often negligible, even if the dependence in the process $Z_1, \ldots, Z_p$ is quite strong.

To appreciate why, note that the expected number of nonzero means in each string of ζ consecutive components of X_i, when X_i is drawn from Π_1, equals $\zeta \times p^{-1} p_1$; and that if $(\zeta p_1)^2 = o(p)$ then the probability that none of the approximately p/ζ strings (placed end to end) of ζ consecutive components that contain one or more nonzero means are adjacent, and the probability that none of the strings contains more than one component, both converge to zero. Therefore, in view of the assumption of ζ-dependence, if we treat the Z_js as independent and identically distributed when making a statement about properties of rankings deduced from (3.3), the probability that we commit an error in the statement converges to zero as $n \to \infty$. It can then be deduced that, in cases where the positive means are randomly distributed and $(\zeta p_1)^2 = o(p)$, the variance of the number of misrankings is relatively low.

Alternatively, rather than scatter the nonzero means μ_j randomly throughout the vector $(Z_1, \ldots, Z_p)$, we could place them all down one end. This makes the distribution of those quantities just about as "clumpy" as possible, by exploiting the ζ-dependence property. For example, if $p_1 \le \zeta$ then all of the nonzero means are attributed to the first p_1 variables in the sequence $X_{i1}, \ldots, X_{ip}$, when X_i is drawn from population Π_1. The assumption of ζ-dependence permits $X_{i1} = \ldots = X_{ip_1}$ with probability 1, whenever X_i comes from either Π_0 or Π_1. This reduces the amount of available information, since all the components that contain information for discriminating between Π_0 and Π_1 are simply copies of one another; there are no independent sources of corroborating information. Moreover, the values of S_j in (3.3) are identical for $1 \le j \le p_1$, and so the values of $\hat{\ell}_j$ are the same too, up to remainders of order λ^3, which implies that the total number of misranks is approximately equal to p_1 times the number of times that a specific component with a positive mean is misranked.

Of course, this increases the variance of the number of misranks. The setting $p_1 \le \zeta$ can encompass instances where $(\zeta p_1)^2 = o(p)$, which was shown two paragraphs above to result in a relatively high amount of information about the differences between Π_0 and Π_1 when the nonzero means are scattered randomly in the data vector. These examples illustrate the more general rule that, in cases where the positive means are distributed consecutively in relatively long-range dependent vectors X_i, the variance of the number of misranks tends to be higher than in cases where those means are

distributed at random.

We conclude this section by comparing the notion of misranking with that of (feature) False Discovery Rate (FDR); for the latter, see for example Benjamini and Hochberg (1995) and Abramovich et al. (2006). For any set of selected features, FDR equals the fraction of falsely selected features. This concept and that of misranking both provide informative measures of how well important features are ranked, but they are nevertheless different in important respects. To appreciate why, let us focus on a specific feature. If we adhere to the notion of FDR then all that matters is whether the feature is selected or not. If we instead we use the concept of misranking then the order or rank of the feature being selected also matters. Technically there are also important differences. For instance, misrankings are defined quite simply in terms of pairwise comparisons of individual features, while FDR can involve higher-order relationships among different features. If we consider the influence, on these measures, of the correlation structure among features, then misranking depends only on pairwise correlations, but FDR may depend on high-order correlations. As a result, FDR can be significantly more difficult to characterise than misranking, and requires much more heavily constrained assumptions about dependence than are necessary using the misranking measure. Therefore, since the central problem is how well important features are ranked, it is more appropriate to assess performance here using misranking, rather than FDR.

The number of misrankings bears a close relationship to both the Wilcoxon rank-sum test and the area under curve (AUC) of the ROC plot (see Hanley and McNeil, 1982). In this case the ROC is constructed with respect to whether each variable is correctly classified as having nonzero mean on not. In fact, it is possible to show that

$$\text{AUC} = 1 - \{p_1(p - p_1)\}^{-1} \, (\#\text{ misrankings}) \, . \tag{3.10}$$

Thus we may interpret properties of ν_{misrank} in Theorems 2–4 in the context of AUC. For instance, under the assumptions of Theorem 2 the AUC score decays to 0.5, this being the score for the random guessing model.

4 THRESHOLDING FOR ADAPTIVE DIMENSION REDUCTION

Recall that in Theorem 1 we showed that $\hat{\ell}_j$ equals $-U_{j1}$, where

$$U_{j1} = \tfrac{1}{2} \, \pi \, (3 - 2 \, \pi) \, \left(E Z^2 \right)^{-1} \, (S_j + \mu_j)^2 \, , \tag{4.1}$$

plus a quantity that does not depend on j, plus a remainder term that is uniformly smaller in size. The feature-ranking step (stage (1) in the algorithm in section 2.1) aspires to re-order the indices j such that the indices for which $\mu_j \neq 0$ are ranked first, and those for which $\mu_j = 0$ are listed together at the end of the sequence. If this objective is largely achieved then the main task that remains is to choose the point in the ranking where the change occurs; this is stage (3) of the algorithm in section 2.1. In the present section we explore method (3a), based on thresholding; see section 2.3.

Observe that if U_{j1} is as at (4.1) then $U_{j1} = U_{j2} + U_{j3}$, where

$$U_{j2} = \tfrac{1}{2}\,\pi\,(3 - 2\,\pi)\,\bigl(EZ^2\bigr)^{-1}\,S_j^2\,, \quad U_{j3} = \tfrac{1}{2}\,\pi\,(3 - 2\,\pi)\,\bigl(EZ^2\bigr)^{-1}\,\bigl(\mu_j^2 + 2\,\mu_j\,S_j\bigr)\,. \tag{4.2}$$

If we can construct a good approximation, $\widehat{U}_{j2}$ say, to U_{j2} then we can subtract it from $\hat{\ell}_j$, leaving only U_{j3} plus a small remainder. The value of U_{j3} is exactly zero if $\mu_j = 0$, and is strictly positive with high probability if $\mu_j > 0$. The quantity $\tilde{\ell}_j = -\widehat{U}_{j2}$ is referred to in that notation in method (3a) in section 2.3. If we choose an appropriate threshold, t say, then we can implement (3a) as follows:

$$\begin{aligned}&\text{define } r \in [k_0 + 1, q], \text{ a random variable, to be the least integer in that}\\ &\text{range such that } \hat{\ell}_{\hat{\jmath}_{r+k}} - \tilde{\ell}_{\hat{\jmath}_{r+k}} > t \text{ for } 1 \le k \le k_0,\end{aligned} \tag{4.3}$$

where $k_0 \ge 0$ is a fixed integer. Then, subject to the jiggling step in stage (4) of the algorithm, we determine that the features with indices $\hat{\jmath}_1, \ldots, \hat{\jmath}_r$ are the ones that have greatest influence on whether a data value X_i came from Π_0 or Π_1.

To define $\widehat{U}_{j2}$, put $\hat{\pi} = n^{-1}\,\sum_i I_i$ and $\bar{X}_j = n^{-1}\,\sum_i X_{ij}$, define our estimator of $\tau^2 = E(Z^2)$ by $\hat{\tau}^2 = (np)^{-1}\,\sum_i \sum_j (X_{ij} - \bar{X}_j)^2$, and let $\widehat{S}_j = \{n\,\hat{\pi}\,(1 - \hat{\pi})\}^{-1}\,\sum_i (I_i - \hat{\pi})\,(X_{ij} - \bar{X}_j)$; compare the definition of S_j at (3.2). Then, motivated by the definition of U_{j2} at (4.2), put

$$-\tilde{\ell}_j = \widehat{U}_{j2} \equiv \tfrac{1}{2}\,\hat{\pi}\,(3 - 2\,\hat{\pi})\,\hat{\tau}^{-1/2}\,\widehat{S}_j^2\,.$$

Theorem 5, below, shows in effect that this is a good approximation to U_{j2}. Let B_4 be as in Theorem 1.

Theorem 5. *Under the conditions of Theorem 1 we can write*

$$\hat{\ell}_j - \tilde{\ell}_j = R - U_{j3} + \Omega_j\,\lambda^3\,, \tag{4.4}$$

where R is as in (3.3) *and, for a constant $B > 0$, the random variables Ω_j satisfy*

$$\sum_{j=1}^{p} P(|\Omega_j| > B) = O\bigl(n^{-B_4}\bigr)\,. \tag{4.5}$$

Finally we describe implementation of method (3a) in stage (3) in section 2.3. We shall assume we are in the context of Theorem 4, where the nonzero μ_js equal $c\,(n^{-1}\,\log n)^{1/2}$. Here it is appropriate to take the threshold $t = t(n)$ to be a sequence of negative numbers such that $|t|$ is of strictly larger order than λ^3 and of strictly smaller order than λ^2; that is, such that

$$|t| = o\big(\lambda^2\big) \quad \text{and} \quad \lambda^3 = o(|t|)\,. \tag{4.6}$$

Let k_1 denote the number of nonzero means added to the components of X_i when X_i is drawn from Π_1. To simplify discussion we assume that k_1 is nonrandom, although of course it may depend on n. Below Theorem 4 we discussed a case where the expected number of misrankings converged to zero, and hence the probability of a misranking occurring also tended to zero. In this setting,

$$\text{with probability converging to 1, } \mu_{\hat{\jmath}_k} > 0 \text{ for } k \le k_1 \text{ and } \mu_{\hat{\jmath}_k} = 0 \text{ for } k > k_1, \tag{4.7}$$

and it can be shown that if $t < 0$ satisfies (4.6) then the definition of r at (4.3) produces a random variable which, with probability converging to 1 as $n \to \infty$, equals $k_0 + k_1$. Therefore, taking $k_0 = 0$ in the rule at (4.3) ensures that, with probability converging to 1, r is exactly equal to k_1. Cases where the nonzero means are of size $c\,(n^{-1}\,\log n)^{1/2}$, but c is not sufficiently large to ensure that (4.7) holds, can be treated satisfactorily by choosing $t < 0$ to satisfy (4.6) but taking $k_0 \ge 1$. Depending on the strength of correlation between components of the vector X it is possible to choose a fixed k_0 such that, with probability converging to 1, r is within $C\rho k_1$ of k_1, where $C > 0$ and ρ equals the expected proportion of feature indices that have positive means when X_i is drawn from Π_1, but are incorrectly ranked at a low level.

5 NUMERICAL PROPERTIES

5.1. Stability of misranking totals. Here we simulate under the model at (3.1), where the signals are represented by μ_j and the noise by Z_{ij}. If we measure performance in terms of the total number of misrankings, or equivalently in terms of AUC (see (3.10)), and if the μ_js decrease at rate $n^{-1/2}$, then Theorem 3 implies that performance should be stable as a function of sample size, n. That is, it should depend very little on n. To explore this property numerically we consider the cases $n = 20$, 50, 100 and 200, with

$p = 0.4 \times n^2$. We take 90% of the μ_js to equal zero and the others to equal $c_0\,(20/n)^{1/2}$, where $c_0 = 0.4, 0.8, \ldots, 0.16$. The noise variables Z_{ij} are independent and identically distributed as $\mathrm{N}(0,1)$.

Figure 1 shows how the expected value of AUC varies with n. The dashed lines on either side of each curve are 95% pointwise confidence bands for the AUC estimate, and quantify the uncertainty of the simulation study. The key feature is that, as predicted by Theorem 3, AUC changes very little with n, even when n is small. As expected, and as predicted by Theorems 2–4, AUC increases with increasing c_0.

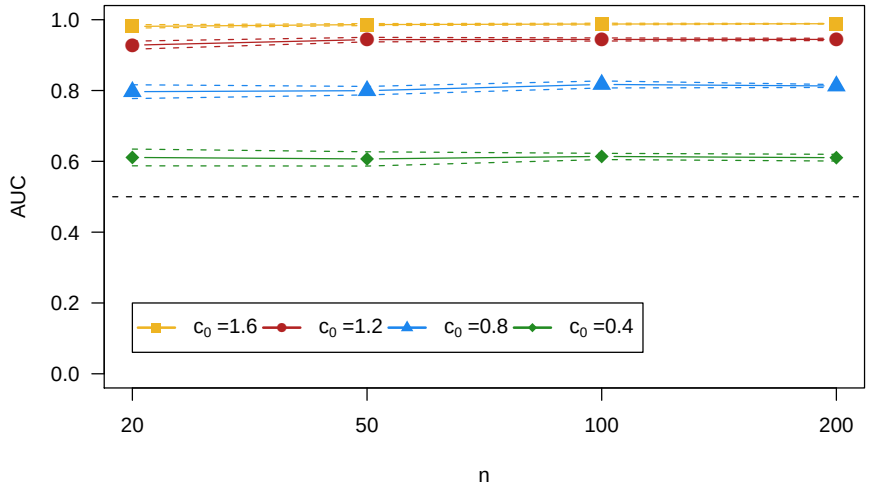

Figure 1: AUC scores in simulation study with $\mu_j = c_0\,(20/n)^{1/2}$, in the example of section 5.1.

We also explored cases where the nonzero μ_js took random values, in particular where they were drawn randomly and uniformly from the interval $[0, c_0\,(20/n)^{1/2}]$. This case is more challenging, since the genuine signals are now strictly smaller than in the previous situation. Therefore it comes as no surprise to learn that the AUC levels for each c_0 are reduced. However, the overall pattern of stability with respect to n is still evident, with very slightly more variation than in the case of fixed μ_js.

5.2. Influence of correlation on misranking performance. Next we discuss the effects of correlated noise Z_{ij} in the model at (3.1). We take the noise to be a moving average of order 1, i.e. $Z_{ij} = \rho\, Z_{i,j-1} + (1-\rho^2)^{1/2}\,\epsilon_{i,j}$, where the $\epsilon_{i,j}$s are independent and normal $\mathrm{N}(0,1)$. Thus, Z_{ij} and Z_{ik} are correlated for all pairs (j,k), with the coefficient of correlation decaying exponentially fast in $|j-k|$. The value of c_0 is fixed at 1.2, and nonzero μ_js are chosen uniformly in $[0, c_0\,(20/n)^{1/2}]$. The values of n, p and the number of true signals are as in section 5.1.

Table 1 gives values of Monte Carlo approximations to the means and standard deviations of AUC scores when the variables, or features, for which μ_j is nonzero are grouped together among the lowest values of j, as in the discussion following Theorem 4. The main observation is that while mean AUC remains stable across the table, the variability of AUC is much greater when strong correlation exists. For instance, if $n = 200$ then when $\rho = 0.99$ the standard deviation of AUC scores is ten times larger than when $\rho = 0$. The presence of correlation makes the problem significantly more difficult; it effectively inserts an element of randomness into the process of correctly ranking important features.

Table 2 shows results in the same setting, except that the indices of variables where μ_j is nonzero are distributed randomly between 1 and p. There is again a high degree of stability, but variability is comparatively less than that in Table 1, consistent with the discussion in section 3.3. In particular, by randomly distributing the indices of the nonzero μ_js we effectively reduce dependence among the variables that are important, and so, reflecting the results in Table 1, the problem becomes less statistically challenging.

	AUC means					AUC std dev.			
ρ	$n = 20$	$n = 50$	$n = 100$	$n = 200$		$n = 20$	$n = 50$	$n = 100$	$n = 200$
-0.99	0.725	0.698	0.706	0.709		0.136	0.078	0.039	0.018
-0.75	0.704	0.717	0.715	0.712		0.064	0.022	0.012	0.006
-0.50	0.650	0.705	0.718	0.711		0.070	0.024	0.011	0.006
-0.25	0.702	0.725	0.706	0.716		0.062	0.025	0.012	0.007
0.00	0.699	0.707	0.711	0.714		0.070	0.027	0.012	0.006
0.25	0.687	0.714	0.707	0.712		0.080	0.032	0.015	0.007
0.50	0.666	0.682	0.713	0.713		0.094	0.037	0.017	0.009
0.75	0.715	0.710	0.718	0.719		0.101	0.047	0.025	0.013
0.99	0.662	0.725	0.708	0.704		0.259	0.157	0.107	0.065

Table 1: Mean and standard deviation of AUC *scores for simulation with correlated noise and grouped effects.*

5.3. Prediction in a large simulated problem. Here we present the analysis of a single simulated dataset, demonstrating how our approach performs when the centroid classifier is used. We take $p = 10{,}000$ and $n = 100$ (50 for each class). Ten percent of the variables X_{ij} (in the model at (3.1)) include a nonzero signal. These μ_js are drawn from the uniform distribution on $[0, 0.35]$, and the noise variables Z_{ij} are independent $\mathrm{N}(0, 1)$. This is a particularly difficult problem, since the signals are very

	AUC means					AUC std dev.			
ρ	$n = 20$	$n = 50$	$n = 100$	$n = 200$		$n = 20$	$n = 50$	$n = 100$	$n = 200$
-0.99	0.712	0.712	0.713	0.715		0.139	0.058	0.032	0.016
-0.75	0.702	0.710	0.713	0.714		0.078	0.030	0.014	0.008
-0.50	0.705	0.710	0.711	0.714		0.076	0.032	0.015	0.008
-0.25	0.706	0.705	0.713	0.714		0.078	0.032	0.015	0.007
0.00	0.696	0.709	0.714	0.712		0.077	0.030	0.015	0.007
0.25	0.698	0.712	0.710	0.714		0.078	0.034	0.016	0.008
0.50	0.703	0.713	0.714	0.714		0.081	0.030	0.016	0.007
0.75	0.698	0.712	0.711	0.714		0.088	0.034	0.016	0.008
0.99	0.741	0.721	0.712	0.710		0.195	0.091	0.053	0.024

Table 2: Mean and standard deviation of AUC *scores for simulation with correlated noise and randomised effects.*

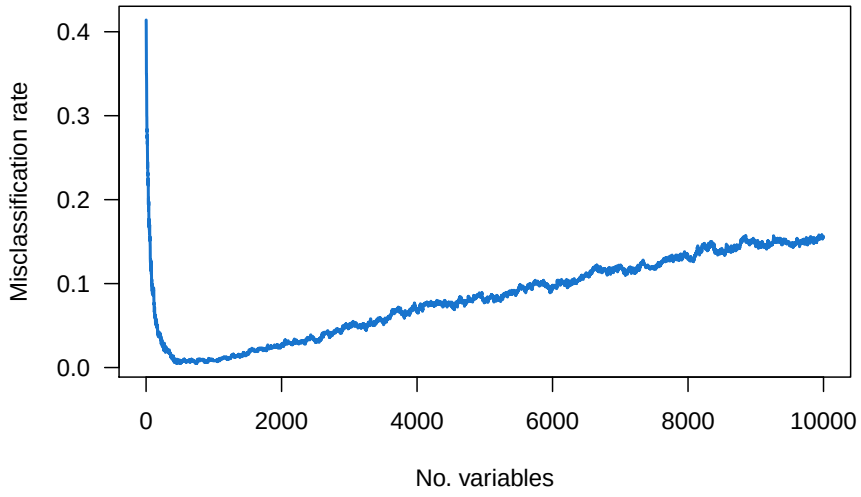

Figure 2: Ideal prediction success for the example of section 5.3.

weak compared to the noise, and sample size is quite small.

Figure 2 shows the prediction performance of the centroid classifier on a test set of 1,000 replicates in the case of "ideal variable selection," where the 1,000 variables with nonzero signals are selected first, in decreasing order of signal strength, followed by the 9,000 variables where the signal is not present. In particular, the order is not chosen empirically. The minimum of the graph occurs at 449 variables (out of a maximum of 1,000), and corresponds to a misclassification rate of only 0.5%. The decrease in predictive performance caused by less useful, or redundant, variables is apparent from the figure; the weaker genuine variables actually hurt prediction performance because they contain more noise than signal. Also of note is the fact that a large number of

variables is needed to obtain good prediction. For example, if attention is confined only to the strongest 50 variables then the misclassification rate increases to 16%.

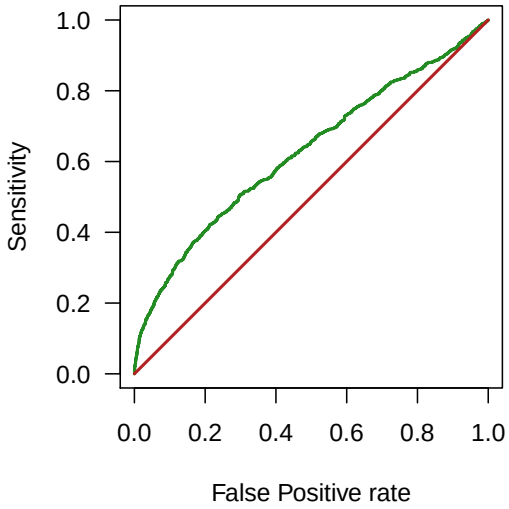

Figure 3: ROC plot for feature ranking in the example of section 5.3.

For the same dataset we undertook variable ranking based on the values of $\hat{\ell}_j$, defined at (2.3). The result was the ROC chart in Figure 3. There the value of k in the ranking $\hat{\jmath}_1, \ldots, \hat{\jmath}_k, \hat{\jmath}_{k+1}, \ldots, \hat{\jmath}_p$, defined in the sentence below (2.3), is represented as k/p on the horizontal axis, and the vertical axis depicts the value of $\hat{\ell}_{\hat{\jmath}_k}/\hat{\ell}_{\hat{\jmath}_1}$, a ratio of two negative numbers. The area under the empirical curve, i.e. AUC, equals 0.626, meaning that a fraction $1 - 0.626 \approx 37\%$ of the paired scores correspond to a misranking. The ROC curve is indexed by model size, with bottom left denoting an empty model and the top right a full model. For a given model size, we can read off the chart the corresponding sensitivity, or proportion of true variables included, and the false positive rate, or proportion of redundant variables in the model. Ideally a model should have high sensitivity and low false positive rate, and the chart indicates the tradeoff between the two for various model sizes.

Figure 4 shows how prediction accuracy varies with model size. Performance is now clearly a long way from that represented in Figure 2, where the 1,000 variables with nonzero signals were listed first in decreasing order of strength. The minimum misclassified rate is now 13.7%, and requires the use of 3,258 of the 10,000 features. As discussed in the previous paragraph, every model size corresponds to a position on the ROC plot in Figure 3, in this case (0.31,0.51). Hence the optimal model found here contains 51% of the genuine variables, and 31% of the redundant ones.

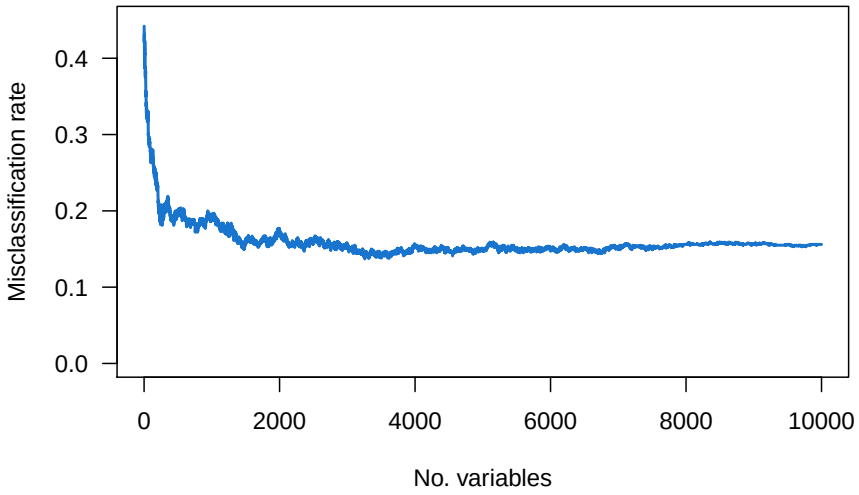

Figure 4: Performance using variable ranking and centroid classifier for the example of section 5.3.

We next explore stage (3) of the four-stage algorithm suggested in section 2.1, addressing in turn each of the approaches (3a)–(3c) discussed in section 2.3. We implement the threshold method, (3a), by comparison with a randomised model where the observed classes I_i are scrambled and likelihoods recalculated, and employ the approach suggested in the second paragraph of section 4; see (4.3). In particular, the threshold is chosen by computing the 100αth percentile of scores for the scrambled data. Doing this for $\alpha = 0.2$ corresponds to seeking a false positive rate of 0.2. In the numerical example that we are considering here, this recovers a model with 2,159 predictors and produces a test set misclassification rate of 16.8%. A model this size corresponds to the point (0.196, 0.40) on the ROC chart. Notice that we have effectively targeted the false positive rate of $\alpha = 0.2$ via this approach.

To provide an example of the change-point method, (3b), suggested in section 2.3 for choosing model size, we consider the ratio of the sorted likelihoods from the original and scrambled rankings. These are plotted in Figure 5, along with a 45° line. Starting with the weakest variables, we expect the ratio to remain near 1 until a sizeable number of variables that genuinely contain a positive signal cause the ratio to shrink. For this purpose we can use the simple change-point statistic for detecting a change in the mean (see Chapter 2 of Csörgő and Horváth, 1997),

$$T(t) = n^{-1/2} \{S(nt) - t\, S(n)\},$$

where $S(k)$ equals the cumulative sum of the first k ratios, and $t \in (0, 1)$ denotes the

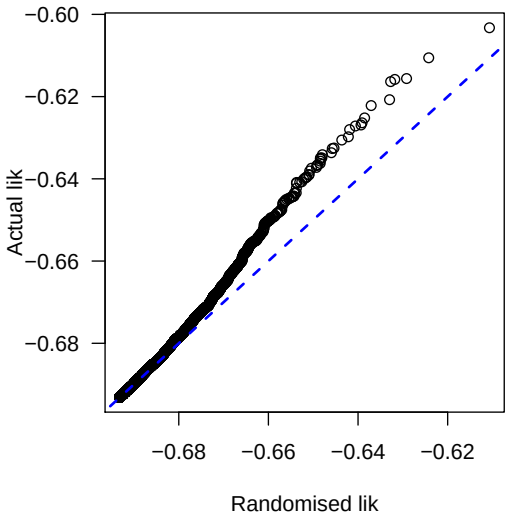

Figure 5: Comparison of actual and randomised log-likelihoods in the example of section 5.3.

examined proportion of the dataset. This leads to a model with 1,760 variables and a misclassification rate of 16.2%, comparable to that when using method (3a). This model size corresponds to the point (0.16, 0.36) on the ROC plot.

Finally, in reference to the classifier-based approach (3c) suggested in section 2.3, we note that the apparent error rate can be driven quickly to zero without the actual error rate being reduced as much as it is if we employ methods (3a) or (3b). For example, when using (3c) in conjunction with the centroid classifier the "best" model, with apparent error rate equal to zero, occurs when just 39 variables are selected; but the misclassification rate on the test set is 32.5%, almost twice that obtained for either of methods (3a) and (3b).

5.4. Results for real data examples. A challenge when using our methodology to analyse previously considered real datasets is that the latter were possibly considered because they illustrate cases where only a very small number of variables determine the class label. In particular, contrary to the concerns raised by Goldstein (2009), the number of influential components is quite small. To simplify matters, we demonstrate here that likelihood based ranking is a powerful tool for improving a wide variety of classifiers. We make use of three well-known sets of microarray data. These relate respectively to leukemia (Golub *et al.*, 1999), colon cancer (Alon *et al.*, 1999) and prostate cancer (Singh *et al.*, 2002) and have 7,129, 2,000 and 6,033 components respectively. Dettling (2004) and Donoho and Jin (2008) discuss the performance of a variety of classifiers on these datasets, using a two-thirds/one-third split of the data into training and test

samples. Their results are reported in Table 3. Readers may refer to the above papers for details on specific methods.

Method	Leukemia	Colon	Prostate
Bagboo	4.08	16.10	7.53
Boost	5.67	19.14	8.71
RanFor	1.92	14.86	9.00
SVM	1.83	15.05	7.88
DLDA	2.92	12.86	14.18
KNN	3.83	16.38	10.59
PAM	3.55	13.53	8.87
HCT	2.86	13.77	9.47

Table 3: Percentage misclassification rate of different methods on microarray datasets.

To test the effectiveness of likelihood-based ranking we chose the best classification method and the random forest classifier (a consistent performer) for each of the datasets. An extra step was added to each cross-validation fold; the two-thirds training data was used to rank variables based on the likelihood score, and then only a proportion of the top-ranked variables were used to estimate the final model. The results are presented in Table 4. The last row of the table shows results for the full dataset; they should in theory match those in Table 3, with differences attributable to tuning approaches. We could not reproduce the accuracy reported for DLDA on the colon dataset, and so used the next best method (PAM).

In each case accuracy can be improved by reducing the model size. For the best classifiers on each dataset, this effect was small but noticeable; for the leukemia data, dimension was reduced by 25% and error by 5%; for the colon dataset, dimension was reduced by 62.5% and error by 1%; and for the prostate dataset, dimension was reduced by 62.5% and error by 3%. For the random forest models the results were even more pronounced, with marked imporvement in prediction and significant dimension reduction. For the prostate dataset, the error was reduced by 25%, using just 0.005 of the available variables in each fold. This suggests that the likelihood based ranking method can effectively control the sparsity of a model and potentially improve model performance.

While firm conclusions are difficult here, we argue that this analysis presents evidence for a large number of relatively weak effects contributing to a model. Indeed, in all but one case we would prefer a model size larger than the dozens, or fewer, used in

Prop.	Leukemia		Colon		Prostate	
	SVM	RanFor	RDA	RanFor		RanFor
0.0025	34.72	4.58	14.29	16.58	8.82	8
0.005	34.61	3.78	13.00	15.71	8.37	7.67
0.01	6.86	3.36	13.13	15.77	7.84	7.73
0.025	1.89	2.97	13.10	15.74	7.16	8.20
0.05	1.67	2.58	13.26	15.16	7.29	8.55
0.10	1.58	2.50	13.19	14.94	7.02	8.90
0.15	1.67	2.22	13.16	14.94	7.10	9.14
0.25	1.64	2.36	12.90	15.03	7.06	9.49
0.375	1.61	2.11	12.87	15.52	6.82	9.67
0.50	1.58	2.22	13.00	15.58	6.82	9.90
0.75	1.53	2.36	12.97	16.13	7.02	9.80
1.00	1.61	2.31	13.00	16.32	7.04	10.24

Table 4: Performance of best methods on reduced datasets, using likelihood based ranking.

many conventional approaches to variable selection. Furthermore, our variable ranking appears to be a useful means of determining the effective model size.

6 TECHNICAL ARGUMENTS

6.1. Proof of Theorem 1. Define $\ell_{ij}(\alpha, \beta)$ and $\ell_j(\alpha, \beta)$ as at (2.1) and (2.2), respectively, and put $\ell(\alpha, \beta) = E\{\ell_{ij}(\alpha, \beta)\}$. To simplify notation, and since for the most part we shall work with one j at a time, we omit mention of j in the notation $\ell(\alpha, \beta)$, and likewise we drop the subscript j on μ_j.

The event that X_i comes from Π_1 occurs with probability π if X_i is sampled from the union of Π_0 and Π_1. Thus, conditional on the sampling operation, $X_{ij} = \mu + Z_{ij}$ with probability π, and $X_{ij} = Z_{ij}$ with probability $1 - \pi$, where each Z_{ij} is distributed

as Z. Therefore,

$$\begin{aligned}
\ell(\alpha,\beta) &= -\pi\,(\alpha+\beta\,\mu)+\pi\,E[\log\{1+\exp(\alpha+\beta\,\mu+\beta\,Z)\}] \\
&\qquad +(1-\pi)\,E[\log\{1+\exp(\alpha+\beta\,Z)\}]\,, \\
d_{10}(\alpha,\beta) &\equiv \frac{\partial}{\partial\alpha}\,\ell(\alpha,\beta) = 1-\pi-\pi\,E\big[\{1+\exp(\alpha+\beta\,\mu+\beta\,Z)\}^{-1}\big] \\
&\qquad -(1-\pi)\,E\big[\{1+\exp(\alpha+\beta\,Z)\}^{-1}\big]\,, \qquad (6.1) \\
d_{01}(\alpha,\beta) &\equiv \frac{\partial}{\partial\beta}\,\ell(\alpha,\beta) = -\pi\,E\big[(\mu+Z)\,\{1+\exp(\alpha+\beta\,\mu+\beta\,Z)\}^{-1}\big] \\
&\qquad -(1-\pi)\,E\big[Z\,\{1+\exp(\alpha+\beta\,Z)\}^{-1}\big]\,, \qquad (6.2)
\end{aligned}$$

where we have used the fact that $E(Z)=0$.

Let $\rho = e^\alpha/(1+e^\alpha)$. Since $E|Z|^4<\infty$ then Taylor expansion shows that, uniformly in $|\alpha|,|\mu|\le C$ for a fixed constant $C>0$, and as $\beta\to 0$,

$$\begin{aligned}
&E\big[\{1+\exp(\alpha+\beta\,\mu+\beta\,Z)\}^{-1}\big] \\
&\quad = \big(1+e^\alpha\big)^{-1}\,E\bigg(\Big[1+\rho\,\big\{\beta\,(\mu+Z)+\tfrac12\,\beta^2\,(\mu+Z)^2+\ldots\big\}\Big]^{-1}\bigg) \\
&\quad = (1-\rho)\,E\Big[1-\rho\,\big\{\beta\,(\mu+Z)+\tfrac12\,\beta^2\,(\mu+Z)^2\big\}+\rho^2\,\beta^2\,(\mu+Z)^2\Big]+O\big(|\beta|^3\big) \\
&\quad = (1-\rho)\,\Big\{1-\rho\,\beta\,\mu+\rho\,\big(\rho-\tfrac12\big)\,\beta^2\,\big(\mu^2+EZ^2\big)\Big\}+O\big(|\beta|^3\big)\,, \qquad (6.3) \\
&E\big[(\mu+Z)\,\{1+\exp(\alpha+\beta\,\mu+\beta\,Z)\}^{-1}\big] \\
&\quad = (1-\rho)\,E\Big((\mu+Z)\,\Big[1-\rho\,\big\{\beta\,(\mu+Z)+\tfrac12\,\beta^2\,(\mu+Z)^2\big\}+\rho^2\,\beta^2\,(\mu+Z)^2\Big]\Big) \\
&\qquad\qquad +O\big(|\beta|^3\big) \\
&\quad = (1-\rho)\,\Big\{\mu-\rho\,\beta\,\big(\mu^2+EZ^2\big)+\rho\,\big(\rho-\tfrac12\big)\,\beta^2\,\big(\mu^3+3\,\mu\,EZ^2+EZ^3\big)\Big\} \\
&\qquad\qquad +O\big(|\beta|^3\big)\,. \qquad (6.4)
\end{aligned}$$

Combining (6.1)–(6.4) we deduce that, uniformly in $|\alpha|,|\mu|\le C$ and as $\beta\to 0$,

$$\begin{aligned}
d_{10}(\alpha,\beta) &= 1-\pi-(1-\rho)\,\big\{1-\rho\,\beta\,\pi\,\mu+\rho\,\big(\rho-\tfrac12\big)\,\beta^2\,\big(\pi\,\mu^2+EZ^2\big)\big\} \\
&\qquad +O\big(|\beta|^3\big)\,, \qquad (6.5) \\
d_{01}(\alpha,\beta) &= -(1-\rho)\,\Big\{\pi\,\mu-\rho\,\beta\,\big(\pi\,\mu^2+EZ^2\big) \\
&\qquad +\rho\,\big(\rho-\tfrac12\big)\,\beta^2\,\big(\pi\,\mu^3+3\,\pi\,\mu\,EZ^2+EZ^3\big)\Big\}+O\big(|\beta|^3\big)\,. \qquad (6.6)
\end{aligned}$$

Put $\rho = \pi + \eta$. By Taylor expansion from (6.5) and (6.6),

$$\begin{aligned} d_{10}(\alpha,\beta) &= \eta + \pi\,(1-\pi)\left\{\beta\,\pi\,\mu - \left(\pi - \tfrac{1}{2}\right)\beta^2\,EZ^2\right\} + O\big(\delta^3\big)\,, \qquad (6.7)\\ d_{01}(\alpha,\beta) &= -(1-\pi-\eta)\left\{\pi\,\mu - (\pi+\eta)\,\beta\,EZ^2 + \pi\left(\pi - \tfrac{1}{2}\right)\beta^2\,EZ^3\right\} + O\big(\delta^3\big)\\ &= -\pi\,(1-\pi)\,\mu + \eta\,\pi\,\mu + \{\pi\,(1-\pi) + (1-2\pi)\,\eta\}\,\beta\,E\big(Z^2\big)\\ &\qquad -\pi\,(1-\pi)\left(\pi - \tfrac{1}{2}\right)\beta^2\,E\big(Z^3\big) + O\big(\delta^3\big)\,, \qquad (6.8) \end{aligned}$$

uniformly in $|\beta| \le \delta$ and α such that $|\rho - \pi| \le \delta$, as $\delta \to 0$.

Recall that $\lambda = (n^{-1}\,\log n)^{1/2}$. By Taylor expansion from formula (2.2) for $\ell_j(\alpha,\beta)$, we have:

$$\begin{aligned} &\frac{\partial}{\partial\alpha}\,(1-E)\,\ell_j(\alpha,\beta)\\ &\quad= \frac{1}{n}\sum_{i=1}^n (1-E)\,(1-I_i) - \frac{1}{n}\sum_{i=1}^n \{1+\exp(\alpha+\beta\,X_{ij})\}^{-1}\\ &\quad= \frac{1}{n}\sum_{i=1}^n (1-E)\,(1-I_i) + \rho\,(1-\rho)\,\beta\,\frac{1}{n}\sum_{i=1}^n (1-E)\,X_{ij} + O_p\big(\delta^2\,\lambda\big) \qquad (6.9)\\ &\quad= \frac{1}{n}\sum_{i=1}^n (1-E)\,(1-I_i) + \pi\,(1-\pi)\,\beta\,\frac{1}{n}\sum_{i=1}^n (1-E)\,X_{ij} + O_p\big(\delta^2\,\lambda\big)\,, \qquad (6.10)\\ &\frac{\partial}{\partial\beta}\,(1-E)\,\ell_j(\alpha,\beta)\\ &\quad= \frac{1}{n}\sum_{i=1}^n (1-E)\,(1-I_i)\,X_{ij} - \frac{1}{n}\sum_{i=1}^n X_{ij}\,\{1+\exp(\alpha+\beta\,X_{ij})\}^{-1}\\ &\quad= \frac{1}{n}\sum_{i=1}^n (1-E)\,(1-I_i)\,X_{ij} - (1-\rho)\,\frac{1}{n}\sum_{i=1}^n (1-E)\,X_{ij} + O_p(\delta\lambda) \qquad (6.11)\\ &\quad= \frac{1}{n}\sum_{i=1}^n (1-E)\,(1-I_i)\,X_{ij} - (1-\pi)\,\frac{1}{n}\sum_{i=1}^n (1-E)\,X_{ij} + O_p(\delta\lambda)\,, \qquad (6.12) \end{aligned}$$

uniformly in $1 \le j \le p$, in $|\alpha| \le C$ such that $|\rho - \pi| \le \delta$, and in $|\beta| \le \delta$, as $\delta \to 0$. Here we continue to take $\rho = e^\alpha/(e^\alpha + 1)$ and $\eta = \rho - \pi$. Deriving (6.10) and (6.12) for each fixed j, α and β is straightforward. In the Appendix we shall show that (6.10) and (6.12) hold uniformly in those quantities, and more particularly that the remainders $O_p(\delta^2\,\lambda)$ and $O_p(\delta\lambda)$ there can be written as $\Theta_j\,\delta^2\,\lambda$ and $\Theta_j\,\delta\,\lambda$, respectively, where in both cases Θ_j satisfies (3.4).

Define

$$\xi_j^{(1)} = n^{-1}\sum_{i=1}^n (1-E)\,X_{ij}\,, \quad \xi_j^{(2)} = n^{-1}\sum_{i=1}^n (1-E)\,(1-I_i)\,X_{ij}$$

and put

$$\xi = -n^{-1} \sum_{i=1}^{n} (1-E)\, I_i = n^{-1} \sum_{i=1}^{n} (1-E)\, (a - I_i) \tag{6.13}$$

for any constant a. Note that $(\partial/\partial\alpha)\, \ell_j(\alpha,\beta) = d_{10}(\alpha,\beta) + (\partial/\partial\alpha)\, (1-E)\, \ell_j(\alpha,\beta)$, and that a similar formula applies in the case of $(\partial/\partial\beta)\, \ell_j(\alpha,\beta)$. Combining these properties with (6.7), (6.8), (6.10) and (6.12) we deduce that

$$\begin{aligned} \frac{\partial}{\partial\alpha}\, \ell_j(\alpha,\beta) &= \xi + \pi\,(1-\pi)\,\beta\,\xi_j^{(1)} + \eta \\ &\qquad + \pi\,(1-\pi)\left\{\beta\,\pi\,\mu - \left(\pi - \tfrac{1}{2}\right)\beta^2\, EZ^2\right\} + O_p\big(\delta^3\big)\,, \qquad (6.14) \\ \frac{\partial}{\partial\beta}\, \ell_j(\alpha,\beta) &= \xi_j^{(2)} - (1-\pi-\eta)\,\xi_j^{(1)} \\ &\qquad - \pi\,(1-\pi)\,\mu + \pi\,(1-\pi)\,\beta\, E\big(Z^2\big) + O_p\big(\delta^2\big)\,, \qquad (6.15) \end{aligned}$$

uniformly in $1 \le j \le p$, $|\alpha| \le C$ and $|\beta|, |\mu| \le \delta$, where we assume $\delta \ge \lambda$. The fact that $\xi_j^{(1)}$ and $\xi_j^{(2)}$ equal $O_p(\lambda)$ uniformly in $1 \le j \le p$ can be proved as in the Appendix.

Equating the right-hand side of (6.14) to zero, and solving for η, we deduce that

$$\eta = \eta_j \equiv -\Big[\xi + \pi\,(1-\pi)\,\beta\,\xi_j^{(1)} \pi\,(1-\pi)\left\{\beta\,\pi\,\mu - \left(\pi - \tfrac{1}{2}\right)\beta^2\, EZ^2\right\}\Big] + O_p\big(\delta^3\big)\,, \tag{6.16}$$

uniformly in $1 \le j \le p$ and solutions (α,β) of the equations

$$\frac{\partial}{\partial\alpha}\, \ell_j(\alpha,\beta) = 0\,, \quad \frac{\partial}{\partial\beta}\, \ell_j(\alpha,\beta) = 0 \tag{6.17}$$

that satisfy $|\alpha| \le C$ and $|\beta| \le \delta$, where it is assumed that $|\mu| \le \delta$ and $\delta \ge \lambda$. Write $(\hat\alpha_j, \hat\beta_j)$ for any such solution.

Substituting the expression (6.16) for η into the right-hand side of (6.15), and equating to zero, we deduce that

$$\begin{aligned} \xi_j^{(2)} - (1-\pi+\xi)\,\xi_j^{(1)} &- \pi\,(1-\pi)\,\mu + \pi\,(1-\pi)\,\beta\, E\big(Z^2\big) \\ &- \pi\,(1-\pi)\left(\pi - \tfrac{1}{2}\right)\beta^2\, E\big(Z^3\big) = O_p\big(\delta^2\big)\,, \qquad (6.18) \end{aligned}$$

uniformly in $1 \le j \le p$ and solutions (α,β) of (6.17) for which $|\alpha| \le C$ and $|\beta| \le \delta$. Solving (6.18) for β we obtain:

$$\beta\,\pi\,(1-\pi)\left(EZ^2\right)\{1 + O_p(\delta)\} = (1-\pi)\,\xi_j^{(1)} - \xi_j^{(2)} + \pi\,(1-\pi)\,\mu + O_p\big(\delta^2\big)\,,$$

from which it follows that, uniformly in the sense of:

$$\text{all } 1 \le j \le p \text{ and all } (\hat\alpha_j, \hat\beta_j) \text{ satisfying (6.17), } |\hat\alpha_j| \le C \text{ and } |\hat\beta_j| \le \delta, \tag{6.19}$$

we have:

$$\hat{\beta}_j = \frac{(1-\pi)\,\xi_j^{(1)} - \xi_j^{(2)} + \pi\,(1-\pi)\,\mu}{\pi\,(1-\pi)\,EZ^2} + O_p\big(\delta^2\big)\,. \tag{6.20}$$

Recall that $\alpha_0 = \log\{\pi/(1-\pi)\}$. Uniformly in the sense of (6.19),

$$\begin{aligned}\hat{\alpha}_j - \alpha_0 &= \log\Big(\frac{\hat{\rho}_j}{1-\hat{\rho}_j}\Big) - \log\Big(\frac{\pi}{1-\pi}\Big) \\ &= \log\Big\{1 + \frac{\eta_j}{\pi\,(1-\pi)}\Big\} + O_p\big(\eta_j^2\big) = \frac{\eta_j}{\pi\,(1-\pi)} + O_p\big(\eta_j^2\big)\,,\end{aligned}$$

where $\hat{\rho}_j = \exp(\hat{\alpha}_j)/\{1+\exp(\hat{\alpha}_j)\}$ and η_j is given by (6.16). Therefore, in view of (6.16) and (6.20),

$$\hat{\alpha}_j - \alpha_0 = \frac{\eta_j}{\pi\,(1-\pi)} + O_p\big(\eta_j^2\big) = -\frac{\xi}{\pi\,(1-\pi)} + O_p\big(\delta^2\big)\,, \tag{6.21}$$

uniformly in the sense of (6.19).

By (6.5) and (6.6), $d_{10}(\alpha_0, 0) = 0$ and $d_{01}(\alpha_0, 0) = -\pi\,(1-\pi)\,\mu$. It can be deduced from (6.1) and (6.2) that

$$\begin{aligned}d_{20}(\alpha,\beta) &= -\pi\,E\Big\{\frac{\exp(\alpha+\beta\,\mu+\beta\,Z)}{1+\exp(\alpha+\beta\,\mu+\beta\,Z)}\Big\} \\ &\qquad -(1-\pi)\,E\Big\{\frac{\exp(\alpha+\beta\,Z)}{1+\exp(\alpha+\beta\,Z)}\Big\}\,, \\ d_{02}(\alpha,\beta) &= -\pi\,E\Big\{(\mu+Z)^2\,\frac{\exp(\alpha+\beta\,\mu+\beta\,Z)}{1+\exp(\alpha+\beta\,\mu+\beta\,Z)}\Big\} \\ &\qquad -(1-\pi)\,E\Big\{Z^2\,\frac{\exp(\alpha+\beta\,Z)}{1+\exp(\alpha+\beta\,Z)}\Big\}\,, \\ d_{11}(\alpha,\beta) &= -\pi\,E\Big\{(\mu+Z)\,\frac{\exp(\alpha+\beta\,\mu+\beta\,Z)}{1+\exp(\alpha+\beta\,\mu+\beta\,Z)}\Big\} \\ &\qquad -(1-\pi)\,E\Big\{Z\,\frac{\exp(\alpha+\beta\,Z)}{1+\exp(\alpha+\beta\,Z)}\Big\}\,.\end{aligned}$$

Therefore, $d_{20}(\alpha_0, 0) = -\pi$, $d_{02} = -\pi^2\,\mu^2 - \pi\,E\big(Z^2\big)$ and $d_{11}(\alpha_0, 0) = -\pi^2\,\mu$.

Combining (6.21) and the results in the previous paragraph, and using Taylor expansion, we deduce that:

$$\begin{aligned}&\ell(\hat{\alpha}_j, \hat{\beta}_j)\ -\ell(\alpha_0, 0) \\ &\quad= (\hat{\alpha}_j - \alpha_0)\,d_{10}(\alpha_0,0) + \hat{\beta}_j\,d_{01}(\alpha_0,0) + \tfrac{1}{2}\,(\hat{\alpha}_j-\alpha_0)^2\,d_{20}(\alpha_0,0) \\ &\qquad +\tfrac{1}{2}\,\hat{\beta}_j^2\,d_{02}(\alpha_0,0) + (\hat{\alpha}_j-\alpha_0)\,\hat{\beta}_j\,d_{11}(\alpha_0,0) + O_p\big(|\hat{\alpha}_j-\alpha_0|^3 + |\hat{\beta}_j|^3\big) \\ &\quad= -\pi\,(1-\pi)\,\hat{\beta}_j\,\mu - \tfrac{1}{2}\,\xi^2\,\big\{\pi\,(1-\pi)^2\big\}^{-1} - \tfrac{1}{2}\,\hat{\beta}_j^2\,\pi\,EZ^2 + O_p\big(\delta^3\big)\,,\end{aligned} \tag{6.22}$$

uniformly in the sense of (6.19). Define $R_1 = \ell(\alpha_0, 0) - \frac{1}{2}\,\xi^2\,\{\pi\,(1-\pi)^2\}^{-1}$ and $S_{1j} = \{(1-\pi)\,\xi_j^{(1)} - \xi_j^{(2)}\}\,\{\pi\,(1-\pi)\,EZ^2\}^{-1}$. (Here and below, R_1, R_2 and R_3 will denote quantities not depending on j, and in particular do not depend on $\mu = \mu_j$.) In this notation it follows from (6.20) that

$$\hat\beta_j = S_{1j} + \mu\,(EZ^2)^{-1} + O_p\big(\delta^2\big)\,, \tag{6.23}$$

uniformly in the sense of (6.19), and then from (6.22) that

$$\begin{aligned}
\ell(\hat\alpha_j, \hat\beta_j) &= R_1 - \pi\,(1-\pi)\,\Big(S_{1j} + \frac{\mu}{EZ^2}\Big)\,\mu - \tfrac{1}{2}\,\pi\,\Big(S_{1j} + \frac{\mu}{EZ^2}\Big)^2\,EZ^2 + O_p\big(\delta^3\big) \\
&= R_1 - \mu\,S_{1j}\,\{\pi\,(1-\pi) + \pi\} \\
&\qquad -\mu^2\,\big(EZ^2\big)^{-1}\,\big\{\pi\,(1-\pi) + \tfrac{1}{2}\,\pi\big\} - \tfrac{1}{2}\,\pi\,S_{1j}^2\,EZ^2 + O_p\big(\delta^3\big)\,. \qquad (6.24)
\end{aligned}$$

It can be deduced from (2.1) and (2.2), by Taylor expansion, that $D_j(\alpha,\beta) \equiv (1-E)\,\ell_j(\alpha,\beta)$ satisfies

$$\begin{aligned}
D_j(\alpha,\beta) &= \frac{1}{n}\sum_{i=1}^n (1-E)\,\{\pi\,\beta\,X_{ij} - I_i\,(\alpha + \beta\,X_{ij})\} + O_p\big(\delta^3\big) \\
&= \beta\,\frac{1}{n}\sum_{i=1}^n (1-E)\,(\pi - I_i)\,X_{ij} - \alpha\,\frac{1}{n}\sum_{i=1}^n (1-E)\,I_i + O_p\big(\delta^3\big) \\
&= \beta\,\big\{(\pi-1)\,\xi_j^{(1)} + \xi_j^{(2)}\big\} + \alpha\,\xi + O_p\big(\delta^3\big)\,, \qquad (6.25)
\end{aligned}$$

uniformly in $1 \le j \le p$, $|\alpha| \le C$ and $|\beta| \le \delta$, where ξ is as at (6.13). Replacing (α,β) here by $(\hat\alpha_j, \hat\beta_j)$, and noting from (6.21) that $\hat\alpha_j - \alpha_0 = -\xi\,\{\pi\,(1-\pi)\}^{-1} + O_p(\delta^2)$ and also that $\hat\beta_j$ satisfies (6.23); and defining $R_2 = [\alpha_0 - \xi\,\{\pi\,(1-\pi)\}^{-1}]\,\xi$; we deduce from (6.25) that

$$D_j(\hat\alpha_j, \hat\beta_j) = S_{1j}\,S_{2j} + \mu\,\big(EZ^2\big)^{-1}\,S_{2j} + R_2 + O_p\big(\delta^3\big)\,, \tag{6.26}$$

uniformly in the sense of (6.19), where

$$S_{2j} = (\pi-1)\,\xi_j^{(1)} + \xi_j^{(2)} = -\pi\,(1-\pi)\,(EZ^2)\,S_{1j}\,.$$

Combining (6.24) and (6.26), and observing that $\ell_j(\alpha,\beta) = \ell(\alpha,\beta) + D_j(\alpha,\beta)$, we find

that

$$\begin{aligned}
\ell_j(\hat\alpha_j,\hat\beta_j) &= R_3 + S_{1j}\,S_{2j} + \mu\,\bigl(EZ^2\bigr)^{-1}\,S_{2j} - \mu\,S_{1j}\,\pi\,(2-\pi) \\
&\qquad -\mu^2\,\bigl(EZ^2\bigr)^{-1}\,\tfrac12\,\pi\,(3-2\,\pi) - \tfrac12\,\pi\,\bigl(EZ^2\bigr)\,S_{1j}^2 + O_p\bigl(\delta^3\bigr) \\
&= R_3 - \pi\,(1-\pi)\,\bigl(EZ^2\bigr)\,S_{1j}^2 - \tfrac12\,\pi\,\bigl(EZ^2\bigr)\,S_{1j}^2 - \pi\,(1-\pi)\,\mu\,S_{1j} \\
&\qquad -\pi\,(2-\pi)\,\mu\,S_{1j} - \tfrac12\,\pi\,(3-2\,\pi)\bigl(EZ^2\bigr)^{-1}\,\mu^2 + O_p\bigl(\delta^3\bigr) \\
&= R_3 - \tfrac12\,\pi\,(3-2\,\pi)\,\bigl(EZ^2\bigr)\,S_{1j}^2 - \pi\,(3-2\,\pi)\,\mu\,S_{1j} \\
&\qquad -\tfrac12\,\pi\,(3-2\,\pi)\,\bigl(EZ^2\bigr)^{-1}\,\mu^2 + O_p\bigl(\delta^3\bigr) \\
&= R_3 - \tfrac12\,\pi\,(3-2\,\pi)\,\bigl(EZ^2\bigr)\,\Bigl(S_{1j} + \frac{\mu}{EZ^2}\Bigr)^2 + O_p\bigl(\delta^3\bigr)\,, \qquad (6.27)
\end{aligned}$$

where $R_3 = R_1 + R_2$ and does not depend on j.

Since we assumed that $(\hat\alpha_j,\hat\beta_j)$ minimises $\ell_j(\alpha,\beta)$ within n^{-c} of $(\alpha_0,0)$, where $0<c<1$, then we may take $\delta = n^{-c}$. Then, iterating (6.27), we conclude that (6.27) holds with $\delta = \lambda$. In that case (6.27) is identical to (3.3).

6.2. Proofs of Theorems 2 and 3. For brevity we treat only pairs j_1, j_2 such that $\mu_{j_1} > 0$ and $\mu_{j_2} = 0$. Assume that the conditions of either Theorem 2 or Theorem 3 hold. Put $W_j = n^{1/2}\,S_j$ and $\delta_j = n^{1/2}\,\mu_j$. Then

$$P\bigl(\hat\ell_{j_2} < \hat\ell_{j_1}\bigr) = P\Bigl\{(W_{j_1} - W_{j_2} + \delta_{j_1})\,(W_{j_1} + W_{j_2} + \delta_{j_1}) \le \Theta_{j_1 j_2}\,n^{-1/2}\,(\log n)^{3/2}\Bigr\}\,, \quad (6.28)$$

where, by (3.3), the random variable $\Theta_{j_1 j_2}$ satisfies

$$\lim_{C\to\infty}\,\limsup_{n\to\infty} P\biggl(\max_{j_1\,:\,\mu_{j_1}>0}\;\max_{j_2\,:\,\mu_{j_2}=0}\,|\Theta_{j_1 j_2}| > C\biggr) = 0 \qquad (6.29)$$

for a constant $C > 0$. Also, defining $\sigma^2 = \pi\,(1-\pi)\,EZ^2$ to denote the asymptotic variance of $n^{1/2}\,S_j$, we have:

$$\operatorname{cov}(W_{j_1} + W_{j_2}, W_{j_1} - W_{j_2}) = E\bigl(W_{j_1}^2\bigr) - E\bigl(W_{j_2}^2\bigr) = \sigma^2 - \sigma^2 + o(1) \to 0\,.$$

From this property, the definitions of W_{j_1} and W_{j_2}, and the Berry-Esseen bound for sums of independent random vectors, we find that

$$P\Bigl(W_{j_1} + W_{j_2} \le x_1, W_{j_1} - W_{j_2} \le x_2\Bigr) = \Phi(x_1/\sigma_{j_1 j_2,+})\,\Phi\bigl(x_1/\sigma_{j_1 j_2,-}\bigr) + o(1)\,, \quad (6.30)$$

uniformly in real numbers x_1, x_2 and in j_1, j_2 such that $\mu_{j_1} > 0$ and $\mu_{j_2} = 0$, as $n\to\infty$. Combining (6.28)–(6.30) we deduce that if $\mu_{j_1} = n^{-1/2}\,c_{nj_1}$ then

$$\begin{aligned}
P(\hat\ell_{j_2} < \hat\ell_{j_1}) &= \Bigl\{\Phi(-c_{nj_1}/\sigma_{j_1 j_2,+})\,\Phi\bigl(c_{nj_1}/\sigma_{j_1 j_2,-}\bigr) \\
&\qquad +\Phi(c_{nj_1}/\sigma_{j_1 j_2,+})\,\Phi\bigl(-c_{nj_1}/\sigma_{j_1 j_2,-}\bigr)\Bigr\} + o(1)\,, \quad (6.31)
\end{aligned}$$

uniformly in j_1, j_2 such that $\mu_{j_1} > 0$ and $\mu_{j_2} = 0$ as $n \to \infty$. Theorems 2 and 3 follow from (6.31). Note that in the context of Theorem 2, $c_{nj} \to 0$ uniformly in $1 \le j \le p$.

6.3. Proof of Theorem 4. Result (6.28) continues to hold, with $\delta_j = n^{1/2}\,\mu_j = c\,(\log n)^{1/2}$ when $\mu_j > 0$. In place of (6.29) we use the following result, a consequence of (3.4): for a constant $C > 0$,

$$\max_{j_1\,:\,\mu_{j_1}>0}\ \max_{j_2\,:\,\mu_{j_2}=0} P(|\Theta_{j_1 j_2}| > C) = O\big(n^{-B_4}\big)\,,$$

where $B_4 = \frac{1}{4}\,\{B_2 - 2\max(B_1 + 3, 2B_1)\} - \epsilon$ for any $\epsilon > 0$. Therefore, since $n^{-1/6} \times n^{-1/6} \gg n^{-1/2}$,

$$\begin{aligned} P\big(\hat\ell_{j_2} < \hat\ell_{j_1}\big) \;&=\; P\big(W_{j_1} + W_{j_2} + \delta_{j_1} \le -\tfrac{1}{2}\,n^{-1/6}\,,\, W_{j_1} - W_{j_2} + \delta_{j_1} > \tfrac{1}{2}\,n^{-1/6}\big) \\ &\quad + P\big(W_{j_1} + W_{j_2} + \delta_{j_1} > \tfrac{1}{2}\,n^{-1/6}\,,\, W_{j_1} - W_{j_2} + \delta_{j_1} \le -\tfrac{1}{2}\,n^{-1/6}\big) \\ &\quad + \theta_{j_1 j_2} + O\big(n^{-B_4}\big)\,, \end{aligned} \tag{6.32}$$

uniformly in j_1, j_2 such that $\mu_{j_1} > 0$ and $\mu_{j_2} = 0$, where

$$\begin{aligned} |\theta_{j_1 j_2}| \le \phi_{j_1 j_2} \;&=\; P\big(W_{j_1} - W_{j_2} + \delta_{j_1} \in \big[-\tfrac{1}{2}\,n^{-1/6}, \tfrac{1}{2}\,n^{-1/6}\big]\big) \\ &\quad + P\big(W_{j_1} + W_{j_2} + \delta_{j_1} \in \big[-\tfrac{1}{2}\,n^{-1/6}, \tfrac{1}{2}\,n^{-1/6}\big]\big)\,. \end{aligned} \tag{6.33}$$

Given a random variable V, write $(1 - E)\,V$ to denote $V - E(V)$. Define $V_j = \{\pi\,(1-\pi)\}^{-1}\,n^{-1/2}\,\sum_i\,(I_i - \pi)\,Z_{ij}$, which has zero mean, and note that $\pi\,(1-\pi)\,(W_j - V_j) = c\,(\log n)^{1/2}\,\Delta_j$, where, defining I_{ij} to be the indicator of the event that $\mu_j > 0$ (conditional on X_i is from Π_1, or equivalently on $I_i = 1$), we put

$$\Delta_j = n^{-1}\,\sum_{i=1}^{n}\,(1 - E)\,(I_i - \pi)\,I_i\,I_{ij} = (1 - \pi)\,n^{-1}\,\sum_{i=1}^{n}\,(1 - E)\,I_i\,I_{ij}\,.$$

Bernstein's inequality can therefore be used to prove that for all $B_5, B_6 > 0$,

$$\sum_{j=1}^{p}\,P\big(|\Delta_j| > B_5\,n^{-1/6}\big) = O\big(n^{-B_6}\big)\,.$$

It follows that (6.32) and (6.33) continue to hold if we replace W_j by V_j, and $\frac{1}{2}\,n^{-1/6}$ by $n^{-1/6}$, at each appearance:

$$\begin{aligned} P\big(\hat\ell_{j_2} < \hat\ell_{j_1}\big) \;&=\; P\big(V_{j_1} + V_{j_2} + \delta_{j_1} \le -n^{-1/6}\,,\, V_{j_1} - V_{j_2} + \delta_{j_1} > n^{-1/6}\big) \\ &\quad + P\big(V_{j_1} + V_{j_2} + \delta_{j_1} > n^{-1/6}\,,\, V_{j_1} - V_{j_2} + \delta_{j_1} \le -n^{-1/6}\big) \\ &\quad + \phi_{j_1 j_2} + O\big(n^{-B_4}\big)\,, \end{aligned} \tag{6.34}$$

uniformly in j_1, j_2 such that $\mu_{j_1} > 0$ and $\mu_{j_2} = 0$, where

$$\begin{aligned} |\theta_{j_1 j_2}| \le \phi_{j_1 j_2} &= P\bigl(V_{j_1} - V_{j_2} + \delta_{j_1} \in \bigl[-n^{-1/6}, n^{-1/6}\bigr]\bigr) \\ &\quad + P\bigl(V_{j_1} + V_{j_2} + \delta_{j_1} \in \bigl[-n^{-1/6}, n^{-1/6}\bigr]\bigr)\,. \end{aligned} \tag{6.35}$$

Property (3.8) implies that the random variables $(I_1 - \pi)\, Z_{1j}$ satisfy a uniform pairwise Cramér continuity condition: (3.8) holds if Z_{j_1} and Z_{j_2} there are replaced by $(I_1 - \pi)\, Z_{1j_1}$ and $(I_1 - \pi)\, Z_{1j_2}$, respectively. This property, and standard methods for deriving Edgeworth expansions of distributions of random vectors (Bhattacharya and Rao, 1976), then allow us to prove that, uniformly in j_1, j_2 such that $\mu_{j_1} > 0$ and $\mu_{j_2} = 0$, and in real numbers x_1 and x_2,

$$\begin{aligned} P(V_{j_1} + V_{j_2} \le x_1\,,\, V_{j_1} - V_{j_2} \le x_2) &= P(N_{j_1 j_2,+} \le x_1, N_{j_1 j_2,-} \le x_2) \\ &\quad + \sum_{k=1}^{k_0} n^{-k/2}\, P_k(x_1, x_2)\, \phi(x_1, x_2) + O\bigl(n^{-B_2}\bigr)\,, \end{aligned} \tag{6.36}$$

where $(N_{j_1 j_2,+}, N_{j_1 j_2,-})$ denotes a normally distributed random vector having zero mean and the same covariance matrix as $(V_{j_1} + V_{j_2}, V_{j_1} - V_{j_2})$, ϕ is the density of the distribution of $(N_{j_1 j_2,+}, N_{j_1 j_2,-})$, the quantities P_k are polynomials of degree $3k - 1$ with the same parity as $k + 1$ and uniformly bounded coefficients (for $k \le k_0$), and k_0 equals the largest integer strictly less than B_2 (recall that in Theorem 1, and hence also in Theorem 4, we assumed that $E|Z|^{B_2} < \infty$). Here we have used the fact that, in view of (3.8),

$$\begin{aligned} &E(V_{j_1} \pm V_{j_2})^2 \text{ is bounded away from zero uniformly in values of } j_1, j_2 \\ &\text{satisfying } 1 \le j_1 \le j_2 \text{ and } j_1 \ne j_2, \text{ and for both choices of the } \pm \text{ signs.} \end{aligned} \tag{6.37}$$

Now, $E\{(V_{j_1} + V_{j_2})\,(V_{j_1} - V_{j_2})\} = E(V_{j_1}^2) - E(V_{j_2}^2) = 0$, and so $N_{j_1 j_2,+}$ and $N_{j_1 j_2,-}$ are independent. Therefore (6.36) implies that, uniformly in the sense there,

$$\begin{aligned} P(V_{j_1} + V_{j_2} \le x_1\,,\, V_{j_1} - V_{j_2} \le x_2) &= P(N_{j_1 j_2,+} \le x_1)\, P(N_{j_1 j_2,-} \le x_2) \\ &\quad + \sum_{k=1}^{k_0} n^{-k/2}\, P_k(x_1, x_2)\, \phi_+(x_1)\, \phi_-(x_2) + O\bigl(n^{-B_2}\bigr)\,, \end{aligned} \tag{6.38}$$

where $\phi_\pm$ denotes the density of $N_{j_1 j_2,\pm}$ for respective values of the signs.

Put $\kappa_n = (\log n)^{1/2}$and recall that $\sigma_{j_1 j_2,\pm}$ is defined at (3.6). Using (6.37) and

(6.38) it can be deduced that

$$\begin{aligned}\sum_{j_1\,:\,\mu_{j_1}>0}\ \sum_{j_2\,:\,\mu_{j_2}=0} & P\bigl(V_{j_1}+V_{j_2}+\delta_{j_1}\le -n^{-1/6}\,,\,V_{j_1}-V_{j_2}+\delta_{j_1}>n^{-1/6}\bigr)\\ &\sim \sum_{j_1\,:\,\mu_{j_1}>0}\ \sum_{j_2\,:\,\mu_{j_2}=0} P\bigl(N_{j_1j_2,+}\le -n^{-1/6}-c\,\kappa_n\bigr)\,P\bigl(N_{j_1j_2,-}>n^{-1/6}-c\,\kappa_n\bigr)\\ &\sim \sum_{j_1\,:\,\mu_{j_1}>0}\ \sum_{j_2\,:\,\mu_{j_2}=0} P\bigl(N_{j_1j_2,+}\le -n^{-1/6}-c\,\kappa_n\bigr)\\ &\sim \sum_{j_1\,:\,\mu_{j_1}>0}\ \sum_{j_2\,:\,\mu_{j_2}=0} \Phi(-c\,\kappa_n/\sigma_{j_1j_2,+})\,, \end{aligned} \tag{6.39}$$

and similarly,

$$\begin{aligned}\sum_{j_1\,:\,\mu_{j_1}>0}\ \sum_{j_2\,:\,\mu_{j_2}=0} & P\bigl(V_{j_1}+V_{j_2}+\delta_{j_1}>-n^{-1/6}\,,\,V_{j_1}-V_{j_2}+\delta_{j_1}\le n^{-1/6}\bigr)\\ &\sim \sum_{j_1\,:\,\mu_{j_1}>0}\ \sum_{j_2\,:\,\mu_{j_2}=0} \Phi(-c\,\kappa_n/\sigma_{j_1j_2,-})\,, \end{aligned} \tag{6.40}$$

and that both of these quantities are of a strictly larger order of magnitude than

$$\sum_{j_1\,:\,\mu_{j_1}>0}\ \sum_{j_2\,:\,\mu_{j_2}=0} P\bigl(V_{j_1}\pm V_{j_2}+\delta_{j_1}\in\bigl[-n^{-1/6},n^{-1/6}\bigr]\bigr)\,,$$

for either choice of the $\pm$ sign. Write (P) to denote the latter property. (Note that if N is a standard normal random variable then $P(N>t)\sim P(N>t+\delta_t)$ as $t\to\infty$, for any quantity δ_t that satisfies $t\,\delta_t\to 0$, and that if in addition $\delta_t>0$, $P(N\in[t-\delta_t,t+\delta_t])\sim(2/\pi)^{1/2}\,\delta_t\,\exp(-t^2/2)$.) Combining (P), (6.34), (6.35), (6.39) and (6.40), and noting that by assumption $p_1\,(p-p_1)=o(n^{B_4})$, we deduce that (3.9) holds.

6.4. Proof of Theorem 5. Result (4.4), with the remainder term satisfying (4.5), will follow from Theorem 1 if we show that for a constant $B_7>0$,

$$\sum_{j=1}^{p} P\bigl(\bigl|\widehat{U}_{j2}-U_{j2}\bigr|>B_7\,\lambda^3\bigr)=O\bigl(n^{-B_4}\bigr)\,. \tag{6.41}$$

Bernstein's inequality can be used to prove that, for any given $B_8>0$, we can choose $B_9>0$, depending on B_8, such that

$$P(|\hat\pi-\pi|>B_9\,\lambda)=O\bigl(n^{-B_8}\bigr)\,. \tag{6.42}$$

Put $\bar{Z}_j = n^{-1} \sum_i Z_{ij}$ and $\bar{Z}_j^{(2)} = n^{-1} \sum_i (Z_{ij}^2 - EZ^2)$. Then, since $E|Z|^{B_2} < \infty$, we have for constants $B_{10}, B_{11} > 0$ and all $\epsilon > 0$:

$$\sum_{j=1}^{p} \{P(|\bar{Z}_j| > B_{10}\,\lambda) + P(|S_j| > B_{10}\,\lambda)\} = O\big(n^{B_1+\epsilon-(B_2-1)/2}\big)\,, \tag{6.43}$$

$$P\big(\big|\bar{Z}^{(2)}\big| > B_{11}\,\lambda\big) = O\big(n^{1+\epsilon-(B_2/4)}\big)\,. \tag{6.44}$$

(The method of proof is that leading to (A.1) in the Appendix.) Define

$$T_j = \frac{1}{n} \sum_{i=1}^{n} (I_i - \pi)\, X_{ij} = \pi\,(1-\pi)\, S_j\,, \quad \widehat{T}_j = \frac{1}{n} \sum_{i=1}^{n} (I_i - \hat{\pi})\,(X_{ij} - \bar{X}_j)\,,$$

and note that

$$\widehat{T}_j = T_j - \{\bar{Z}_j + \hat{\pi}\,\mu_j + (\hat{\pi} - \pi)\,\mu_j\}\,(\hat{\pi} - \pi)\,. \tag{6.45}$$

Result (6.41) follows from (6.42)–(6.45), the bound $|\mu_j| \le c\,\lambda$, and the property

$$\widehat{U}_{j2} = \frac{\hat{\pi}\,(3 - 2\,\hat{\pi})\,\widehat{T}_j^2}{2\,\hat{\tau}^{1/2}\,\{\hat{\pi}\,(1-\hat{\pi})\}^2}\,, \quad U_{j2} = \frac{\pi\,(3-2\,\pi)\,T_j^2}{2\,\tau^{1/2}\,\{\pi\,(1-\pi)\}^2}\,.$$

Here we used the fact that, by the definition of B_4, $B_4 < \frac{1}{4}\,B_2 - 1$.

A APPENDIX: Proof that (6.10) and (6.12) hold uniformly in $1 \le j \le p$, $|\alpha| \le C$ such that $|\rho - \pi| \le \delta$, and $|\beta| \le \delta$

We shall give proofs of (6.11) and (6.12). Derivations of (6.9) and (6.10) are almost identical, and in fact passing from (6.9) to (6.10), and from (6.11) to (6.12), requires only the properties $|\rho - \pi| \le \delta$, which is an assumption, and

$$\sup_{1 \le j \le p} \left| \frac{1}{n} \sum_{i=1}^{n} (1 - E)\, X_{ij} \right| = O_p(\lambda)\,,$$

which follows from standard moderate-deviation results (see Rubin and Sethuraman, 1965; Amosova, 1972). Indeed, those results imply the following stronger property: for a constant $B_3 > 0$, depending on B_1 and B_2 in Theorem 1,

$$\sum_{j=1}^{p} P\left\{ \left| \frac{1}{n} \sum_{i=1}^{n} (1 - E)\, X_{ij} \right| > B_3\,\lambda \right\} = O\big(n^{B_1+\epsilon-(B_2-1)/2}\big) \tag{A.1}$$

for each $\epsilon > 0$.

The claim that (6.11) holds uniformly in $1 \le j \le p$, in $|\alpha| \le C$ and in $|\beta| \le \delta$ will follow if we show that, with $\Delta_{ij}(\alpha, \beta) = \{1 + \exp(\alpha + \beta\, X_{ij})\}^{-1} - (1 - \rho)$, we have:

$$\sup_{1 \le j \le p,\, |\alpha| \le C,\, |\beta| \le \delta} \left| \frac{1}{n} \sum_{i=1}^{n} (1 - E)\, X_{ij}\, \Delta_{ij}(\alpha, \beta) \right| = O_p(\delta\lambda)\,, \tag{A.2}$$

uniformly in the same sense. In fact, it follows from (A.8) and (A.9) below that a sharper form of (A.2),

$$\sum_{j=1}^{p} P\left\{ \sup_{|\alpha| \le C,\, |\beta| \le \delta} \left| \frac{1}{n} \sum_{i=1}^{n} (1 - E)\, X_{ij}\, \Delta_{ij}(\alpha, \beta) \right| > \text{const.}\, \delta\lambda \right\} = O\big(n^{-B_4}\big)\,,$$

where $B_4 = \frac{1}{4}\{B_2 - 2\max(B_1 + 3, 2B_1)\} - \epsilon$ for any $\epsilon > 0$, holds. This stronger bound, together with (A.1), implies that (6.11) and (6.12) hold in the stronger sense that the remainder $O_p(\delta\lambda)$ there can be replaced by $\Theta_j\, \delta\lambda$, where $\sum_j P(|\Theta_j| > \text{const.}) = O(n^{-B_4})$ and $B_4 = \frac{1}{4}\{B_2 - 2\max(B_1 + 3, 2B_1)\} - \epsilon$. That result gives the sharper version, implied by (3.4), of (3.3) in Theorem 1.

With probability 1,

$$|\Delta_{ij}(\alpha, \beta)| \le K_1\, \min(1, |\beta\, X_{ij}|)\,, \quad \frac{|\Delta_{ij}(\alpha_1, \beta_1) - \Delta_{ij}(\alpha_2, \beta_2)|}{|\alpha_1 - \alpha_2| + |\beta_1 - \beta_2|} \le K_1\, (1 + |X_{ij}|)\,, \tag{A.3}$$

uniformly in $|\alpha|, |\alpha_1|, |\alpha_2|, |\beta|, |\beta_1|, |\beta_2| \le C$, where, here and below, $K_1, K_2, \ldots$ are fixed positive constants. Using the first part of (A.3), and standard arguments for proving moderate-deviation results for sums of independent random variables (Rubin and Sethuraman, 1965; Amosova, 1972); and assuming that $E|Z|^{2(B+1+\epsilon)} < \infty$, where $B, \epsilon > 0$; it can be shown that

$$\sup_{1 \le j \le p,\, |\alpha| \le C,\, |\beta| \le \delta} P\left\{ \left| \frac{1}{n} \sum_{i=1}^{n} (1 - E)\; X_{ij}\, \Delta_{ij}(\alpha, \beta) \right| > K_2\, \delta\, \big(B\, n^{-1}\, \log n\big)^{1/2} \right\} \le K_3\, n^{-B}\,. \tag{A.4}$$

Partition $[-C, C]$ and $[-\delta, \delta]$ into lattices of equal edge width d_n, where $0 < d_n \le \delta$, and let $\mathcal{V}_1$ and $\mathcal{V}_2$ denote the respective sets of lattice vertices. The number of vertices in each lattice equals $O(d_n^{-1})$, and so by (A.4),

$$\sup_{1 \le j \le p} P\left\{ \sup_{\alpha \in \mathcal{V}_1,\, \beta \in \mathcal{V}_2} \left| \frac{1}{n} \sum_{i=1}^{n} (1 - E)\; X_{ij}\, \Delta_{ij}(\alpha, \beta) \right| > K_2\, \delta\, \big(B\, n^{-1}\, \log n\big)^{1/2} \right\} \le K_3\, d_n^{-2}\, n^{-B}\,. \tag{A.5}$$

Given $\alpha_1 \in [-C, C]$ and $\beta_1 \in [-\delta, \delta]$, let α_2 and β_2 be the lattice vertices nearest to α_1 and β_1, respectively; ties can be broken arbitrarily. In view of the second part of (A.3), for all $\alpha_1 \in [-C, C]$ and all $\beta_1 \in [-\delta, \delta]$, and with probability 1,

$$\begin{aligned}\Big|\frac{1}{n}\sum_{i=1}^{n}(1-E)\,X_{ij}\,\Delta_{ij}(\alpha_1,\beta_1)\Big| \\ \le \Big|\frac{1}{n}\sum_{i=1}^{n}(1-E)\,X_{ij}\,\Delta_{ij}(\alpha_2,\beta_2)\Big| + 4\,K_1\,d_n\,\frac{1}{n}\sum_{i=1}^{n}(1+|X_{ij}|)^2\,. \quad \text{(A.6)}\end{aligned}$$

Since $E|Z|^{2(B+1+\epsilon)} < \infty$ then, by Markov's inequality, for each $K_4 \ge E(\mu^2 + Z^2) + 1$,

$$\sup_{1\le j\le p} P\Big\{n^{-1}\sum_{i=1}^{n}(1+|X_{ij}|)^2 > K_4\Big\} \le K_5\,n^{-(B+1+\epsilon)/2}\,. \quad \text{(A.7)}$$

Combining (A.5), (A.6) and (A.7) we deduce that

$$\begin{aligned}\equiv P\Big\{\sup_{1\le j\le p,\,|\alpha|\le C,\,|\beta|\le\delta}\Big|\frac{1}{n}\sum_{i=1}^{n}(1-E)\,X_{ij}\,\Delta_{ij}(\alpha,\beta)\Big| > K_6\left(\delta\,\lambda + d_n\right)\Big\} \\ = O\big\{p\left(d_n^{-2}\,n^{-B} + n^{-(B+1+\epsilon)/2}\right)\big\}\,. \quad \text{(A.8)}\end{aligned}$$

Take $d_n = \delta\lambda$ and $\delta \ge \lambda$. Then (A.2) follows from (A.8) provided that $p\,(\lambda^{-2}\,n^{-B} + n^{-(B+1+\epsilon)/2}) \to 0$. Since $p = O(n^{B_1})$ then it is sufficient that $B > \max(B_1+1, 2B_1-1)$. That is, we should ensure that $E|Z|^{B_2} < \infty$ where $B_2 > 2\max(B_1+3, 2B_1)$, which assumption is imposed in the theorem. In this case it follows from (A.8) that

> the left-hand side of (A.8) equals $O(n^{-B_4})$ where $B_4 = \frac{1}{4}\{B_2 - 2\max(B_1+3, 2B_1)\} - \epsilon$. (A.9)

REFERENCES

ABRAMOVICH, F., BENJAMINI, Y., DONOHO, D. AND JOHNSTONE, I. (2006). Adapting to unknown sparsity by controlling the false discovery rate. *Ann. Statist.* **34**, 584–653.

ALON, U., BARKAI, N., NOTTERMAN, D.A., GISH, K., YBARRA, S., MACK, D. AND LEVINE, A.J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci.* **96**, 6745–6750.

AMOSOVA, N.N. (1972). Limit theorems for the probabilities of moderate deviations. *Vestnik Leningrad. Univ.* No. 13, Mat. Meh. Astronom. Vyp. 3 (1972), 5–14, 148.

BENJAMINI, Y. AND HOCHBERG, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc.* Ser. B **57**, 289–300.

BHATTACHARYA, R.N. AND RAO, R.R. (1976). *Normal Approximation and Asymptotic Expansions*. Wiley, New York.

BREIMAN, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics* **37**, 373–384.

BREIMAN, L. (1996). Heuristics of instability and stabilization in model selection. *Ann. Statist.* **24**, 2350–2383.

BREIMAN, L. (2001). Random forests. *Machine Learning* **45**, 5-32.

CANDES, E. AND TAO, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.* **35**, 2313–2351.

CARLSTEIN, E., MÜLLER, H.-G. AND SIEGMUND, D. (Eds) (1994). *Change-Point Problems*. Papers from the AMS-IMS-SIAM Summer Research Conference held at Mt. Holyoke College, South Hadley, MA, July 11–16, 1992. Institute of Mathematical Statistics Lecture Notes—Monograph Series, 23. Institute of Mathematical Statistics, Hayward, CA.

CHEN, J. AND GUPTA, A.K. (2000). *Parametric Statistical Change Point Analysis*. Birkhäuser Boston, Boston, MA.

CSÖRGŐ, M. AND HORVÁTH, L. (1997). *Limit Theorems in Change-Point Analysis*. Wiley, New York.

DETTLING, M. (2004). BagBoosting for tumor classification with gene expression data. *Bioinformatics* **20**, 3583-3593.

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

DONOHO, D.L. (2006b). For most large underdetermined systems of equations, the minimal ℓ_1-norm near-solution approximates the sparsest near-solution. *Comm. Pure Appl. Math* **59**, 907–934.

DONOHO AND ELAD, M. (2003). Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization. *Proc. Nat. Acad. Sci. USA* **100**, 2197–2202.

DONOHO, D.L. AND HUO, X. (2001). Uncertainty principles and ideal atomic decomposition. *IEEE Trans. Inform. Th.* **47**, 2845–2862.

DONOHO, D.L. AND JIN, J. (2008). Higher criticism thresholding: optimal feature selection when useful features and rare and weak. *Proc. Nat. Acad. Sci.* **105**, 14790–14795.

DONOHO, D.L. AND JIN, J. (2009). Feature selection by higher criticism thresholding achieves optimal phase diagram. *Phil. Trans. R. Soc.* Ser. A **367**, 4449–4470.

DUDA, R.O., HART, P.E. AND STORK, D.G. (2001). *Pattern Classification*, 2nd edition. Wiley, New York.

FAN, J. AND FAN, Y. (2008). High-dimensional classification using features annealed independence rules. *Ann. Statist.* **36**, 2605–2637.

FAN, J. AND LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96**, 1348–1360.

FAN, J. AND LI, R. (2006). Statistical challenges with high dimensionality: Feature selection in knowledge discovery. *Proceedings of the International Congress of Mathematicians*, eds. M. Sanz-Sole, J. Soria, J.L. Varona and J. Verdera, eds.), Vol. III, 595–622.

FAN, J. AND LV, J. (2008). Sure independence screening for ultra-high dimensional feature space. *J. Roy. Statist. Soc.* Ser. B **70**, 849–911.

FAN, J. AND REN, Y. (2006). Statistical analysis of DNA microarray data. *Clinical Cancer Research* **12**, 4469–4473.

FRIEDMAN, J., HASTIE, T. AND TIBSHIRANI, R. (2008). Regularization paths for generalized linear models via coordinate descent. *Manuscript, Department of Statistics, Stanford University.*

GAO, H.-Y. (1998). Wavelet shrinkage denoising using the non-negative garrote. *J. Comput. Graph. Statist.* **7**, 469–488.

GOLDSTEIN, D.B. (2009). Common genetic variation and human traits. *New England J. Med.* **360**, 1696–1698.

GOLUB, T.R., SLONIM, D.K, TAMAYO, P., HUARD, C., GAASENBEEK, M., MESIROV, J.P., COLLER, H., LOH, M.L., DOWNING, J.R., CALIGIURI, M.A., BLOOMFIELD, C.D. AND LANDER, E.S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**, 531–537.

GUO, Y., HASTIE, T. AND Tibshirani, R. (2007). Regularized linear discriminant analysis and its application in microarrays. *Biostatistics*, **8**, 86-100.

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

HALL, P. AND WANG, Q. (2008). Strong approximations of level exceedences related to multiple hypothesis testing. Manuscript.

HANLEY, J.A., MCNEIL, B.J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **143**, 29-36.

HASTIE, T., TIBSHIRANI, R. AND FRIEDMAN, J. (2001). *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*. Springer, New York.

HIRSCHHORN, J.N. (2009). Genomewide association studies—illuminating biologic pathways. *New England J. Med.* **360**, 1699–1701.

JIN, J. (2009). Impossibility of successful classication when useful features are rare and weak *Proc. Nat. Acad. Sci.* **106**, 8859–8864.

KRAFT, P. AND HUNTER, D.J. (2009). Genetic risk prediction—Are we there yet? *New England J. Med.* **360**, 1701–1703.

SHAKHNAROVICH, G., DARRELL, T. AND INDYK, P. (2005). *Nearest-Neighbor Methods in Learning and Vision. Theory and Practice*. MIT Press, Cambridge, Mass.

SINGH, D. FEBBO, P. G. ROSS, K. JACKSON, D. G. MANOLA, J. LADD, C. TAMAYO, P. RENSHAW, A. A. D AMICO, A. V. AND RICHIE, J. P. (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell* **1**, 203-209.

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

TIBSHIRANI, R., HASTIE, T., NARASIMHAN, B. AND CHU, G. (2002). *Diagnosis of multiple cancer types by shrunken centroids of gene expression. P Natl Acad. Sci. USA*, **99**, 6567-6572.

TROPP, J.A. (2005). Recovery of short, complex linear combinations via ℓ_1 minimization. *IEEE Trans. Inform. Th.* **51**, 1568–1570.

RUBIN, H. AND SETHURAMAN, J. (1965). Probabilities of moderate deviations. *Sankhyā* Ser. A **27**, 325–346.

WADE, N. (2009). Genes show limited value in predicting diseases. *New York Times* April 15. `www.nytimes.com/2009/04/16/health/research/16 gene.html`

WU, Y. (2005). *Inference for Change-Point and Post-Change Means After a CUSUM Test. Lecture Notes in Statistics* **180**. Springer, New York.