Supplementary materials for this article are available at http://pubs.amstat.org/toc/jasa/104/488.

Median-Based Classifiers for High-Dimensional Data

Peter HALL, D. M. TITTERINGTON, and Jing-Hao XUE

Conventional distance-based classifiers use standard Euclidean distance, and so can suffer from excessive volatility if vector components have heavy-tailed distributions. This difficulty can be alleviated by replacing the L_2 distance by its L_1 counterpart. For example, the L_1 version of the popular centroid classifier would allocate a new data value to the population to whose centroid it was closest in L_1 terms. However, this approach can lead to inconsistency, because the centroid is defined using L_2, rather than L_1, distance. In particular, by mixing L_1 and L_2 approaches, we produce a classifier that can seriously misidentify data in cases where the means and medians of marginal distributions take different values. These difficulties motivate replacing centroids by medians. However, in the very-high-dimensional settings commonly encountered today, this can be problematic if we attempt to work with a conventional spatial median. Therefore, we suggest using componentwise medians to construct a robust classifier that is relatively insensitive to the difficulties caused by heavy-tailed data and entails straightforward computation. We also consider generalizations and extensions of this approach based on, for example, using data truncation to achieve additional robustness. Using both empirical and theoretical arguments, we explore the properties of these methods, and show that the resulting classifiers can be particularly effective. Supplementary materials are available online.

KEY WORDS: Centroid classifier; Componentwise median; Data depth; Distance-based classifier; High-dimensional data; L_1 method; Robust method; Sample median; Spatial median; Strength of dependence.

1. INTRODUCTION

Standard distance-based classifiers, including those popularly used for high-dimensional data, can have problems caused by excessive volatility (see, e.g., Jörnsten 2004; Ghosh and Chaudhuri 2005; Ding et al. 2007). In particular, if some of the components of data vectors have high variability, then, regardless of whether or not those components convey information about the populations under study, classification accuracy can be poor.

One approach to tackling this problem is to use classifiers based on L_1, rather than L_2, distance. But implementing multivariate L_1 methods can be awkward, especially in the very-high-dimensional, small-sample size contexts required in contemporary classification problems. In particular, multivariate medians can be cumbersome to compute in high-dimensional settings (see, e.g., Bose, Maheshwari, and Morin 2003), and in some instances are conceptually difficult to interpret (e.g., Kowalski and Powell 2004). In addition, there are several numerically different ways of defining a multivariate median (e.g., Gentleman 1965; Oja 1983; Liu 1990; Donoho and Gasko 1992). These difficulties make componentwise medians attractive.

With this motivation in mind, in this article we develop high-dimensional data classifiers based on componentwise medians. We show that it is inappropriate to mix L_1 and L_2 distances when defining classifiers, for example, by working with L_1 distance to a centroid, because a centroid is defined in L_2 terms. A classifier of this hybrid type can produce chronically inconsistent results, for example, in cases where the standard centroid method works well.

Median-based classifiers, such as those discussed by Jörnsten (2004), Ghosh and Chaudhuri (2005), and Ding et al. (2007), have several unusual properties, enhancing their interest and making the exploration of their properties more challenging. For example, the fact that the theoretical median of the sample median is not necessarily equal to the median of the population from which the data were drawn, whereas the expected value of a sample mean always equals the population mean, means that theoretical properties of median-based classifiers can be quite different from those of their mean-based counterparts.

In this work we explore theoretical properties, and use numerical methods to elucidate finite-sample aspects of median-based classifier performance. We give particular attention to the effects of information sparsity in data vectors and of the strength of dependence among data components. We focus much of our attention on median-based classifiers that are analogs of the more conventional mean-based, centroid-method approach, but we also consider other approaches. These include techniques based on truncating median-based methods, for example, replacing median differences by 0–1 random variables or other bounded functions.

The main technical requirements for a componentwise, median-based classifier are a finite first moment and the existence and positivity of the density in a neighborhood of the median. These assumptions should hold in a sense that is uniform in the vector components. In justification of these assumptions, we mention that L_1 distance is generally used to define a median, and the condition of nonzero density ensures both uniqueness of the median and a reasonable convergence rate. These assumptions hold for components that have a normal, lognormal, double-exponential, or exponential distribution, to name just a few possibilities. The Cauchy distribution is an exception, because its mean is undefined.

In recent years, numerous classifiers have been applied to very-high-dimensional data, particularly gene expression data. Among these classifiers, centroid-based methods have enjoyed considerable popularity (see, e.g., Tibshirani et al. 2002; Yeung

Peter Hall is Professor, Department of Mathematics and Statistics, University of Melbourne, Melbourne, VIC 3010, Australia (E-mail: *halpstat@ms.unimelb.edu.au*). D. M. Titterington is Professor, Department of Statistics, University of Glasgow, Glasgow G12 8QQ, U.K. (E-mail: *m.titterington@stats.gla.ac.uk*). Jing-Hao Xue is Lecturer, Department of Statistical Science, University College London, London WC1E 6BT, U.K. (E-mail: *jinghao@stats.ucl.ac.uk*). This work was supported in part by the award of a Carnegie Centenary Professorship to P. H. and a Hutchison Whampoa-EPSRC Dorothy Hodgkin Postgraduate Award to J.-H. X. The work benefited from the participation of D. M. T. and J.-H. X. in the research program on Statistical Theory and Methods for Complex, High-Dimensional Data at the Isaac Newton Institute for Mathematical Sciences, Cambridge. The authors thank the referees and the associate editor for their extensive, constructive comments.

Journal of the American Statistical Association
December 2009, Vol. 104, No. 488, Theory and Methods
DOI: 10.1198/jasa.2009.tm08107

and Bumgarner 2003; Dabney 2005; Dabney and Storey 2005, 2007; Wang and Zhu 2007; Winter et al. 2007; Fan and Fan 2008; Hall, Titterington, and Xue 2009). Tree-based classifiers have been investigated by Moon et al. (2006), Ahn et al. (2007), and Tibshirani and Hastie (2007), and support vector machines have been used by Nahar et al. (2007) and Wang, Zhu, and Zou (2008). To the best of our knowledge, componentwise median-based classifiers have not yet been discussed in the context of gene expression-level data, however.

2. DEFINITIONS OF CLASSIFIERS

2.1 Examples of Classifiers

Many statistical classification problems can be posed in the following way: Given sets $\mathcal{X} = \{X_1, \ldots, X_m\}$ and $\mathcal{Y} = \{Y_1, \ldots, Y_n\}$ of p-variate data, from populations Π_X and Π_Y, and a new data value Z, from one of the two populations, we wish to classify Z to either Π_X or Π_Y. Distance-based classifiers achieve this by assigning Z to Π_X if a measure of the "distance" of Z from $\mathcal{X}$ is less than the same measure of the distance of Z from $\mathcal{Y}$.

Often this decision can be phrased in the following way: assign Z to Π_X if

$$\sum_{k=1}^{p} \{d(Z_k, \mathcal{Y}_k) - d(Z_k, \mathcal{X}_k)\} > 0, \tag{2.1}$$

and to Π_Y otherwise, where $d(Z_k, \mathcal{X}_k)$ denotes a measure of the distance from the kth component of $Z = (Z_1, \ldots, Z_p)$ to the set

$$\mathcal{X}_k = \{X_{1k}, \ldots, X_{mk}\} \tag{2.2}$$

of the kth components of the data vectors $X_i = (X_{i1}, \ldots, X_{ip})$, and $\mathcal{Y}_k$ and $d(Z_k, \mathcal{Y}_k)$ are interpreted analogously. For example, the centroid classifier and the average-distance classifier assign Z to Π_X if

$$\|Z - \bar{Y}\| > \|Z - \bar{X}\|, \tag{2.3}$$

or if $n^{-1} \sum_j \|Z - Y_j\|^2 > m^{-1} \sum_i \|Z - X_i\|^2$, respectively. (Here $\|\cdot\|$ denotes conventional Euclidean distance.) In these settings, (2.1) gives an equivalent classification decision if we define $d(Z_k, \mathcal{X}_k) = (Z_k - \bar{X}_k)^2$ in the centroid-method case, or if $d(Z_k, \mathcal{X}_k) = m^{-1} \sum_i (Z_k - X_{ik})^2$ for the average distance classifier. (Here $\bar{X}_k$ denotes the kth component of the sample mean, $\bar{X} = m^{-1} \sum_i X_i$.)

In particular, in the case of the centroid classifier, the rule defined by (2.1) implies that we should classify Z as coming from Π_X if

$$\sum_{k=1}^{p} (\bar{Y}_k - \bar{X}_k)(\bar{X}_k + \bar{Y}_k - 2Z_k) > 0, \tag{2.4}$$

and as coming from Π_Y otherwise.

One attractive feature of distance-based classifiers is that they readily lend themselves to multiple-class extensions. Indeed, the extended classifier involves measuring the distance from the new data value, Z, to each of the training samples, and assigning Z to the population corresponding to the nearest sample. Thus, although in this article we restrict our discussion to two-class classification problems, multiple-class extensions are straightforward.

2.2 Classifier Robustness and L_1 Methods

Especially when the sample sizes m and n are small, but also in other cases, formula (2.4) indicates that the classifier's performance may be sensitive to variability. To appreciate this point, note that, if the distributions of X_k and Y_k have finite variance, then the expected value of the kth summand in (2.4) equals

$$s(Z)(\mu_{Xk} - \mu_{Yk})^2 + m^{-1}\sigma_{Yk}^2 - n^{-1}\sigma_{Xk}^2, \tag{2.5}$$

where $\mu_{Xk} = E(X_k)$, $\sigma_{Xk}^2 = \mathrm{var}(X_k)$, μ_{Yk} and σ_{Yk}^2 are defined analogously, and $s(Z) = 1$ or -1 according to whether Z is from Π_X or Π_Y. But if the distribution of X_k or Y_k is sufficiently heavy-tailed (e.g., if it has infinite variance), then the expected value of the kth summand on the left side of (2.4) is not well defined. In such cases, the stochastic fluctuations of that quantity can heavily outweigh the contributions of the terms in μ_{Xk} and μ_{Yk}. This property is significant, because the differences between component means often contribute substantially to effective classification.

Use of an approach based on the L_1, rather than L_2, distance can alleviate this problem. One L_1 version of the centroid classifier is given by (2.3) but with $\|\cdot\|$ there interpreted as the L_1 distance. Equivalently, the classifier has the following form: Assign Z to Π_X if

$$\sum_{k=1}^{p} (|\bar{Y}_k - Z_k| - |\bar{X}_k - Z_k|) > 0, \tag{2.6}$$

and to Π_Y otherwise. Another L_1-based method is founded on the same interpretation of (2.3) but replaces the means $\bar{X}_k$ and $\bar{Y}_k$ by the componentwise medians, $\mathrm{med}\,\mathcal{X}_k$ and $\mathrm{med}\,\mathcal{Y}_k$, where $\mathrm{med}\,\mathcal{X}_k$ equals the empirical median of the data set $\mathcal{X}_k$, defined at (2.2), and $\mathrm{med}\,\mathcal{Y}_k$ is defined analogously. In particular, we assign Z to Π_X if

$$\sum_{k=1}^{p} (|\,\mathrm{med}\,\mathcal{Y}_k - Z_k| - |\,\mathrm{med}\,\mathcal{X}_k - Z_k|) > 0, \tag{2.7}$$

and to Π_Y otherwise.

2.3 Comparison of Classifiers Determined by (2.6) and (2.7)

Of these two classifiers, the one based on (2.6) arguably has greater fidelity to the popular centroid method. But the fact that its construction involves a mismatch between L_1 and L_2 metrics, with the latter used implicitly to define the means $\bar{X}_k$ and $\bar{Y}_k$ and the former used in the norm at (2.6), signals potential problems. In fact, unlike the standard centroid classifier defined by (2.4), this classifier can suffer from significant inconsistency problems, even in the case of light-tailed data.

The intuition behind this property is that it is the median, not the mean, that minimizes L_1 distance. Consequently, even if Z comes from Π_X, Z can be strictly closer, in an L_1 sense, to the mean of Π_Y than to the mean of Π_X. This has the potential to produce erroneous classification decisions when those decisions are based on L_1 distance and on the means of the training samples.

To provide more detail, assume, for simplicity, that m and n diverge as p increases, that the distributions of $X_{ik} - \mu_{Xk}$ and

$Y_{jk} - \mu_{Yk}$ are all identical [to the distribution of Δ, say, with $E(\Delta^2) < \infty$], and that $\mu_{Xk} - \mu_{Yk} = \delta \neq 0$, the same value for each k. Then, using a law of large numbers for weakly dependent variables, we see that the left side of (2.4), divided by p, converges to $s(Z)\delta^2$, where $s(Z)$ is at (2.5). It follows that the probability that the classifier based on L_2 distance makes the correct decision converges to 1; however, the left side of (2.6), divided by p, converges to

$$t(Z) = \begin{cases} E|\Delta + \delta| - E|\Delta| & \text{if } Z \in \Pi_X \\ E|\Delta| - E|\Delta - \delta| & \text{if } Z \in \Pi_Y. \end{cases} \tag{2.8}$$

If $|\delta|$ is sufficiently large, or if the distribution of Δ is symmetric, then the sign of $t(Z)$ is identical to that of $s(Z)$, and thus asymptotically correct classification results. But if the median of the distribution of Δ is different from its mean, then, for an appropriate choice of δ, $s(Z)$ and $t(Z)$ can have opposite signs. This is the case if, for example, we take δ equal to the median of the distribution of Δ. Here, even if the sample sizes m and n are both infinite, the probability that the classifier determined by (2.6) assigns to the X population a random vector drawn from Π_Y converges to 1 as p increases.

These difficulties do not arise if we use the median-based classifier determined by (2.7). In that case δ equals the difference between the medians, ν_{Xk} and ν_{Yk} say, of the distributions of X_{ik} and Y_{jk}, respectively, where δ is assumed to not depend on k, and Δ now has the distribution of $X_{ik} - \nu_{Xk}$ and of $Y_{jk} - \nu_{Yk}$. It follows that if with this new interpretation of δ and Δ, $t(Z)$ is again given by (2.8), then $t(Z) \geq 0$ if $Z \in \Pi_X$ and $t(Z) \leq 0$ if $Z \in \Pi_Y$. As a result, the classifier enjoys consistency for large values of m, n, and p. This property can be readily extended to more general settings; see, for example, Theorems 1 and 2 in Section 4.

The message conveyed by these examples is that the purely L_1 classifier, which assigns Z to Π_X if and only if (2.7) holds, can be effective. On the other hand, the "hybrid" classifier, defined by (2.6), can lead to serious consistency problems.

If the components of data vectors were independent and identically normally distributed, then the standard centroid classifier determined by (2.3) or, equivalently, by (2.4), would enjoy optimality properties; it would be Fisher's quadratic discriminator. Likewise, the L_1-based classifier determined by (2.7) would be optimal if the data components were independent and double-exponentially distributed.

Another feature of median-based classifiers distinguishes them from their competitors, such as the centroid-method classifier determined by the rule at (2.4), that depend on means. To provide some background for this point, we note that in general the expected value and theoretical median of the sample median both depend on sample size, and neither necessarily equals the median of the population from which the data were drawn. As a result, if the training sample sizes m and n are kept fixed as p diverges, then a median-based classifier does not necessarily give asymptotically correct classification, even if all vector components are independent.

This issue is unimportant for most practical purposes, because the difference between the population median of the sample median and the median of the population is generally small, even for small sample sizes. Thus the inconsistency noted in the previous paragraph is of little practical relevance. Nevertheless, it mandates that in the theory discussed in Section 4, we must insist that m and n diverge with p, although not necessarily at the same rate as p.

In contrast to properties of the sample median, the expected value of the sample mean is always equal to the population mean. Thus it comes as no surprise to learn that under second-moment conditions, the standard centroid classifier generally gives asymptotically correct results even if m and n are held fixed as p increases.

The problem of calculating the componentwise median of an n-sample of p-dimensional data is of comparable difficulty to computing the mean vector; in both cases the computations can be done in $O(np)$ time, using the selection algorithm in the case of the median. In particular, the $O(np)$ figure represents the computational complexity of the classifiers in formulas (2.6), (2.7), (2.9), and (2.10). On the other hand, computation for methods based on data depth generally is much more laborious, typically requiring $O(n^{p-1} \log n)$ time when $p \geq 2$ (see, e.g., Baggerly and Scott 1999; Liu and Singh 1999). These issues make it easy and attractive to use the componentwise median as a supplement to other classifiers without first using diagnostics to assess its feasibility.

2.4 Alternative Classifiers

An analog of the median-based classifier, defined by assigning Z to Π_X if and only if (2.7) holds, can be constructed in an average-distance setting. We classify Z as coming from Π_X if and only if

$$\sum_{k=1}^{p} \left(\frac{1}{n} \sum_{j=1}^{n} |Z_k - Y_{jk}| - \frac{1}{m} \sum_{i=1}^{m} |Z_k - X_{ik}| \right) > 0, \tag{2.9}$$

where X_{ik} and Y_{jk} denote the kth components of the p-vectors X_i and Y_j. The classifier determined by (2.7) generally has better performance, however.

Any one of the classifiers determined by (2.6), (2.7), and (2.9) can be made still more robust by replacing the summands in the sums over k in those formulas by their respective signs, that is, by $+1$ if they are positive and -1 otherwise. More generally, the summand could be replaced by a bounded, symmetric function of the summand. Specifically, if ψ were a uniformly bounded function for which $\psi(u) = -\psi(-u)$ for all u and $\psi(u) > 0$ for $u > 0$, then in place of the rule at (2.7), we could assign the new data value Z to Π_X if

$$\sum_{k=1}^{p} \psi(|\operatorname{med} \mathcal{Y}_k - Z_k| - |\operatorname{med} \mathcal{X}_k - Z_k|) > 0 \tag{2.10}$$

and to Π_Y otherwise.

A disadvantage of this "truncation-based" classifier is that it is relatively insensitive to gradations in the sizes of the differences between medians. As a result, unless the function $\psi(u)$ is constructed so as to give high weight to relatively large values of u, truncation-based methods can perform poorly in cases where the only nonzero median differences are small in number but large in size. But truncation-based classifiers can enjoy superior performance when the marginal distributions are very heavy-tailed, for example, when each vector component distribution has infinite mean and the sample sizes m and n are relatively small. In such cases they may detect an accumulation of small median differences that the classifier at (2.7) will miss because it is swamped by stochastic fluctuations.

3. NUMERICAL RESULTS

3.1 Overview of Data Sets, Classifiers, and Computational Algorithms

We compared the performances of the componentwise median-based classifier defined at (2.7), as well as its truncated form in (2.10), with those of another 11 classifiers, using 16 high-dimensional data sets or simulation settings, for which $p > m + n$. These data sets included 3 widely used gene expression data sets, 1 text-mining data set, and 12 data sets simulated from lognormal and Student's t distributions. The four real-world data sets were judged to have a reasonable proportion of heavy-tailed components. The lognormal distribution has one heavy tail, whereas Student's t distribution has two heavy tails.

The 11 classifiers that we compared were the hybrid L_1 and L_2 classifier defined at (2.6), the componentwise centroid method given in (2.4), the naive Bayes classifier, Fisher's linear discriminant analysis (LDA) classifier, the L_1 data depth-based classifier DDclass (Jörnsten 2004), the nearest-shrunken centroid method (Tibshirani et al. 2002, 2003), penalized logistic regression (Park and Hastie 2008), the support vector machine (SVM), the componentwise trimmed centroid method, the nearest-neighbor classifier (1-NN), and recursive partitioning and regression trees (rpart; Breiman et al. 1984).

The naive Bayes classifier is a widely used componentwise generative method (Hand and Yu 2001), based on assumptions of independence between components within each class; each continuous component X_k or Y_k is generally assumed to be Gaussian. Fisher's LDA classifier, not necessarily assuming Gaussian distributions for components, uses a linear combination of components to maximise the separation between Π_X and Π_Y in terms of between-to-within variance. This idea was extended by Ghosh and Chaudhuri (2005), based on two concepts of data depth for linear classification of low-dimensional, heavy-tailed data. Classification can be based on other definitions of data depth, such as L_1 depth (Jörnsten 2004) and spatial depth (Ding et al. 2007), which are founded on definitions of multivariate median.

If Gaussian distributions and equal within-class covariances are assumed for components, then Fisher's LDA becomes the normal- or Gaussian-based LDA. This method can be simplified into a naive Bayes classifier if a diagonal within-class covariance matrix is assumed. For high-dimensional data, the nearest-shrunken centroid method (Tibshirani et al. 2002, 2003) enhances performance by shrinking standardized centroids.

In contrast to generative methods, which model and estimate the joint distribution of components and class labels for classification, discriminative classifiers, such as linear logistic regression and SVM, model the discriminant boundary directly. To provide reasonable estimates in high-dimensional ($p > m + n$) settings, Park and Hastie (2008) used logistic regression with a quadratic penalization on the coefficients. The SVM classifier was used by Nahar, Ali, and Chen (2007) and Wang, Zhu, and Zou (2008), among others, for classification of gene expression data.

Nearest-neighbor classifiers and classification trees have been applied to high-dimensional data, such as gene expression data (Dudoit, Fridlyand, and Speed 2002) and text documents (Sebastiani 2002). In this work, we used rpart and 1-NN to represent such methods. At the suggestion of a reviewer, we also included a componentwise trimmed centroid.

In our implementation, we used the R package `e1071` for the naive Bayes classifier and SVM, the package `MASS` for Fisher's LDA, the package `pamr` for the nearest-shrunken centroid method, the package `stepPlr` for penalized logistic regression, the source code of DDclass by Jörnsten (2004), the package `tm` for text mining data and facilities (Feinerer, Hornik, and Meyer 2008), the package `class` for 1-NN, and the package `rpart` for rpart. We mainly used the default settings (e.g., Gaussian RBF kernel for SVM) of these classifiers for computational and comparative reasons, except in the following cases. For the nearest-shrunken centroid method, we chose the largest value among the thresholds that gave the lowest training errors as the threshold for classification of the test set, and for penalized logistic regression we chose regularization parameter $\lambda = 1$. For the trimmed centroid classifier, a total of 25% of the data vectors were trimmed symmetrically, that is, 12.5% at either end.

3.2 Real Data Examples

The three gene expression data sets were "Leukemia" (Golub et al. 1999), "Lung Cancer" (Gordon et al. 2002), and "Prostate Cancer" (Singh et al. 2002). All came with predetermined, separate training and test sets of data vectors. The Leukemia data set comprised $p = 7{,}129$ genes for $m = 27$ acute lymphoblastic leukemia (ALL) and $n = 11$ acute myeloid leukemia (AML) vectors in the training set. The test set included 20 ALL and 14 AML vectors. The Lung Cancer data set comprised $p = 12{,}533$ genes for $m = 16$ adenocarcinoma (ADCA) and $n = 16$ mesothelioma training vectors, along with 134 ADCA and 15 mesothelioma test vectors. In the Prostate Cancer data set, there were $p = 12{,}600$ genes for $m = 50$ normal and $n = 52$ prostate tumor vectors in the training set and 9 normal and 25 tumor vectors in the test set.

Following Dudoit, Fridlyand, and Speed (2002) and Jörnsten (2004), we first preprocessed the data in the following steps: We truncated intensities to ensure positivity, removed genes that showed little variation in intensity across all the vectors, transformed intensities to base-10 logarithms, and then standardized each vector to have 0 mean and unit variance. This preprocessing procedure kept $p = 3934$ genes for the Leukemia data set, $p = 2959$ genes for the Lung Cancer data set, and $p = 3239$ genes for the Prostate Cancer data set. We then selected the $\ell = 50$ and 1000 most important genes, in the sense of highest significance of two-sample t-tests, a criterion that was also used by Alon et al. (1999), Gordon et al. (2002), and Fan and Fan (2008) and is related to the between-to-within sum of squares used by Dudoit, Fridlyand, and Speed (2002) and Jörnsten (2004) for gene expression data. Such a feature-selection step is commonly used in practice for classification of high-dimensional data, particularly gene expression data, because there are many genes making no good contribution to the classification. This step generally leads to such advantages as improved classification performance, better understanding of influential genes, and more efficient computation (see, e.g., Guyon and Elisseeff 2003; Fan and Fan 2008; Zucknick, Richardson, and Stronach

2008; Hua, Tembe, and Dougherty 2009 for empirical and theoretical arguments). Some closely related work involves a step of selecting from among the $\ell = 50$ most important genes (Golub et al. 1999; Dudoit, Fridlyand, and Speed 2002; Gordon et al. 2002; Tibshirani et al. 2002; Fan and Fan 2008; Fan and Lv 2008; Hall, Titterington, and Xue 2009). To reduce the potential bias toward the classifiers that the feature-selection step may generate, we also used $\ell = 1000$.

The Reuters-21578 test collection (Lewis 1997) is a benchmark data set for text categorization (Sebastiani 2002). A subset of two topics, "acq" and "crude," was used via the R package `tm` for demonstration (Feinerer, Hornik, and Meyer 2008), with $m = 50$ articles for "acq" and $n = 20$ articles for "crude." The articles were preprocessed to remove digits, convert characters to lower case and eliminate stopwords and stem words, and then were used to construct a 70×1506 term document matrix in which the number of terms (or columns)—that is, p—was equal to 1506. This matrix was very sparse, as often occurs in term document matrices in text-mining applications; sometimes variable selection, or "term selection," is carried out before text categorization (Sebastiani 2002). Because no predetermined test set was presented by the package `tm` for these two topics, we used 10-fold cross-validation to assess the performance of the classifiers. Within each fold of the algorithm, we used Fisher's exact test to select ℓ terms having the greatest leverage for discrimination of topics.

Misclassification error rates are listed in Table 1 for the predetermined test data in the Lung Cancer, Leukemia and Prostate Cancer data sets, and in Table 2 for the 10-fold cross-validation applied to the Reuters-21578 subset. In the tables the componentwise median-based classifier is denoted by "c-median," its truncated form by "c-truncated," the componentwise "hybrid L_1 and L_2" classifier by "c-hybrid," the standard componentwise centroid method by "c-centroid," the naive Bayes classifier by "n-Bayes," Fisher's LDA classifier by "LDA," the L_1 data depth-based classifier by "DDclass," the nearest-shrunken centroid method by "NSC," penalized logistic regression by "PLR," and the componentwise trimmed-centroid method by "c-trim-cent."

Table 1. Misclassification error rates for the pre-determined test set of the Leukemia, Lung Cancer, and Prostate Cancer data sets

	Leukemia		Lung Cancer		Prostate Cancer	
Methods	$\ell = 50$	1000	$\ell = 50$	1000	$\ell = 50$	1000
c-median	0.03	0.09	0	0.01	0	0.09
c-truncated	0.06	0.09	0.01	0.01	0	0.15
c-hybrid	0.03	0.09	0	0.01	0	0.24
c-centroid	0.03	0.06	0	0.01	0	0.15
n-Bayes	0.26	0.32	0.01	0.01	0	0.21
LDA	0.21	0.03	0.03	0.01	0.15	0.15
DDclass	0.32	0.24	0	0.01	0	0
NSC	0	0.06	0.01	0.01	0.09	0.15
PLR	0.03	0.09	0.01	0.01	0.09	0.09
SVM	0.21	0.26	0.01	0.03	0	0.03
c-trim-cent	0.03	0.06	0	0.01	0	0.12
1-NN	0.03	0	0.01	0.01	0.12	0
rpart	0.21	0.21	0.06	0.06	0.24	0.24

Table 2. Mean misclassification error rates, and their standard errors in parentheses, from a 10-fold cross-validation for a Reuters-21578 subset

Methods	$\ell = 50$	$\ell = 1000$
c-median	0.06 (0.02)	0.03 (0.02)
c-truncated	0.09 (0.03)	0.11 (0.05)
c-hybrid	0.21 (0.06)	0.27 (0.05)
c-centroid	0.07 (0.02)	0.09 (0.02)
n-Bayes	0.13 (0.04)	0.11 (0.06)
LDA	0.09 (0.02)	0.24 (0.05)
NSC	0.07 (0.02)	0.07 (0.02)
SVM	0.10 (0.04)	0.30 (0.07)
c-trim-cent	0.06 (0.02)	0.04 (0.02)
1-NN	0.06 (0.02)	0.14 (0.03)
rpart	0.03 (0.02)	0.03 (0.02)

NOTE: For technical reasons, no results here for DDclass and PLR.

It can be seen from these tables that for these four data sets, the componentwise median-based classifier and its truncated form performed better than or comparably to the other 11 classifiers. Moreover, the two componentwise median-based approaches were relatively robust and were always among the best few classifiers; the trimmed-centroid method generally was a reasonable alternative. Moreover, the componentwise median-based classifier, its truncated form, and the componentwise centroid- and trimmed-centroid-based classifiers are simple and fast to calculate. In summary, our data show that the componentwise median-based classifier and its truncated form are efficient and effective for classifying high-dimensional, heavy-tailed data. This reflects an earlier observation (Hand 2006) that simple classifiers tend to have comparable performance to more complicated classifiers.

The numbers, h_X and h_Y, of heavy-tailed genes (with kurtosis exceeding 2) out of ℓ genes, or terms, in populations Π_X and Π_Y are given in Table 3. The proportion of heavy-tailed components was not high for the three gene expression data sets.

Tables 1–3 also show that for many classifiers applied to these four data sets, using more features does not necessarily lead to lower misclassification errors, as also has been pointed out by Hua et al. (2009) and Zucknick et al. (2008).

3.3 Simulation Experiments

We simulated p-vectors $X_1, \ldots, X_m$, identically distributed as $(\nu_{X1} + U_1, \ldots, \nu_{Xp} + U_p)$, and $Y_1, \ldots, Y_n$, identically distributed as $(\nu_{Y1} + U_1, \ldots, \nu_{Yp} + U_p)$, where U_k was either log-normally distributed or Student's t–distributed, and thus heavy-tailed. Here $\nu_{X1} = \cdots = \nu_{Xp} = 0$, while $(\nu_{Y1}, \ldots, \nu_{Yp})$ had η nonzero components. We chose $\eta = p/2$ such that for some $\epsilon >$

Table 3. Numbers of heavy-tailed genes, h_X and h_Y, out of the selected ℓ genes in populations Π_X and Π_Y for three gene-expression data sets and a Reuters-21578 subset

Data set	$\ell\ (h_X, h_Y)$		
Leukemia	50 (4, 2)	1000 (170, 41)	3934 (895, 357)
Lung Cancer	50 (3, 3)	1000 (61, 81)	2959 (229, 356)
Prostate Cancer	50 (7, 2)	1000 (151, 262)	3239 (458, 649)
Reuters-21578	50 (18, 19)	1000 (582, 598)	1506 (1088, 627)

0, the proportion of values $k \in [1, p]$ for which $|\nu_{Xk} - \nu_{Yk}| > \epsilon$ was of strictly larger order than $\{\min(m, n)\}^{-1/2}$ as $p \to \infty$, where $\min(m, n) > 4$.

The $p/2$ nonzero components ν_{Yk} were randomly distributed among indexes in the interval $[1, p]$, and we chose $\nu_{Yk}^2 = (2 \log p)\zeta$, where $\zeta = \{\sigma_{U_k}^2 (1/m + 1/n)\}$, in which $\sigma_{U_k}^2$ was the variance of U_k and could be derived analytically in many cases. In the setting of lognormal data, $\log U_k$ was distributed as normal N(0, 1). For the Student's t data, U_k was t-distributed with 2 degrees of freedom; we used the variance of the Student's t distribution with 3 degrees of freedom as $\sigma_{U_k}^2$. We chose $p = 200$.

To assess the impact on classifier performance of dependence among components, we generated the data as long-range dependent moving averages. In particular, X_{ik} [or $\log X_{ik}$ in the case of lognormal data] was taken equal to $w\epsilon_{k-r+1} + \cdots + w\epsilon_k$, and $Y_{jk} - \nu_{Yk}$ [or $\log(Y_{jk} - \nu_{Yk})$ for lognormal data] equalled $w\varepsilon_{k-r+1} + \cdots + w\varepsilon_k$ for $k = 1, \ldots, p$, where $\epsilon_{2-r}, \ldots, \epsilon_p$, $\varepsilon_{2-r}, \ldots, \varepsilon_p$ were independent and t-distributed [or normal N(0, 1) in the case of lognormal data], and $w = r^{-1/2}$.

We took $r = p/4, p/8, \ldots, p/128$. Therefore, in a strongly correlated data set, each component was correlated with a quarter of the other components, whereas in a weakly correlated data set, each component was correlated with only another $p/128$ components. Thus in total we simulated six lognormally distributed and six t-distributed data sets.

Each data set consisted of 200 vectors with 100 in each class, and each vector had $p = 200$ components. Each data set was randomly, equally partitioned into a training set and a test set. The training set contained $m = 50$ vectors from Π_X and $n = 50$ vectors from Π_Y. In addition, each such partition was repeated 40 times, resulting in 40 pairs of training test sets. The mean misclassification error rate and the standard error of the mean were estimated from the classification results for the 40 test sets.

The standard error of mean misclassification error rate, defined in terms of sample variance computed over replicates and obtained either from repeated random partitions into training and test sets or by cross-validation, often is reported as an estimate of (the square root of) the variance of the mean error rate estimate (McLachlan 1972, 1976; Nadeau and Bengio 2003; Markatou et al. 2005).

The mean error rates and their standard errors (vs. proportional dependence length, r/p) are shown in Figures 1 and 2 for lognormally distributed data and in Figures 3 and 4 for t-distributed data. Each panel in each figure compares the c-median with three other classifiers. It can be seen from these figures that the c-median and its truncated form (c-truncated) performed among the best, and with low standard errors, for the simulated heavy-tailed data, particularly when the dependence among components was weak. Misclassification rates increased with the strength of dependence. In contrast, Fisher's LDA and PLR performed the best when the dependence was strong, indicating that a strong dependence among components might favor a method seeking a linear combination of some components influential to the discrimination. In addition, PLR and SVM performed very well and were relatively insensitive to the strength of component dependence.

4. THEORETICAL PROPERTIES

4.1 Introduction

Theoretical performance of the median-based classifier, determined by (2.7), depends in large measure on two properties of the distributions of the marginals X_{1k} and Y_{1k}: the strength of dependence among the marginals and the smoothness of the marginal distributions in the neighborhood of the respective medians. Strength of dependence affects performance by determining the extent to which noise cancels, through a law of large numbers effect, from the sum on the left side of (2.7). Smoothness governs the rate of convergence of the empirical medians, med $\mathcal{X}_k$ and med $\mathcal{Y}_k$, to the true medians. For example, a root-n rate of convergence of an empirical median requires a condition similar to the existence of a density in a neighborhood of the true median.

Bearing these issues in mind, here we present results for two types. The first, Theorem 1 in Section 4.2, imposes a very mild condition on strength of dependence [see (4.4)] and no assumption about smoothness, but requires a relatively large number of nonzero median differences to guarantee consistency [see (4.6)]. The second, Theorem 2 in Section 4.3, requires stronger conditions on dependence and smoothness but permits a much greater degree of sparsity among median differences [see (4.10)].

4.2 Properties Under Weak Dependence Assumption

In line with the previous notation, let $\vec{U} = (U_1, U_2, \ldots)$ denote an infinite sequence of random variables each with a uniquely defined median, equal to 0 if necessary after a shift in location, and satisfying moment, tightness, and mixing conditions,

$$\lim_{\lambda \to \infty} \sup_{k \ge 1} E\{|U_k| I(|U_k| > \lambda)\} = 0, \tag{4.1}$$

$$\text{for each } c > 0, \quad \inf_{k \ge 1} \inf_{|x| \ge c} (E|U_k + x| - E|U_k|) > 0, \tag{4.2}$$

$$\text{for each } \epsilon > 0, \quad \inf_{k \ge 1} \left[\min\left\{\tfrac{1}{2} - P(U_k \le -\epsilon), \tfrac{1}{2} - P(U_k \ge \epsilon)\right\}\right] > 0, \tag{4.3}$$

$$\lim_{k \to \infty} \sup_{k_1, k_2 : |k_1 - k_2| \ge k} \sup_{B_1, B_2 \in \mathcal{B}} \left| P\big(U_{k_1} \in B_1, U_{k_2} \in B_2\big) - P\big(U_{k_1} \in B_1\big) P\big(U_{k_2} \in B_2\big) \right| = 0, \tag{4.4}$$

where $\mathcal{B}$ denotes the class of Borel subsets of the real line. Let $(\nu_{X1}, \nu_{X2}, \ldots)$ and $(\nu_{Y1}, \nu_{Y2}, \ldots)$ be infinite sequences of constants. Assume that for each p, the p-vectors $X_1, \ldots, X_m$ are identically distributed as $(\nu_{X1} + U_1, \ldots, \nu_{Xp} + U_p)$, $Y_1, \ldots, Y_n$ are identically distributed as $(\nu_{Y1} + U_1, \ldots, \nu_{Yp} + U_p)$, Z is distributed as either X_1 or Y_1, and $X_1, \ldots, X_m$, $Y_1, \ldots, Y_n$, and Z are totally independent. Of the median differences $\nu_{Xk} - \nu_{Yk}$, we assume that

$$\text{the differences } |\nu_{Xk} - \nu_{Yk}| \text{ are uniformly bounded,} \tag{4.5}$$

$$\text{for each sufficiently small } \epsilon > 0, \text{ the proportion of values } k \in [1, p] \text{ for which } |\nu_{Xk} - \nu_{Yk}| > \epsilon \text{ is bounded away from zero as } p \text{ diverges.} \tag{4.6}$$

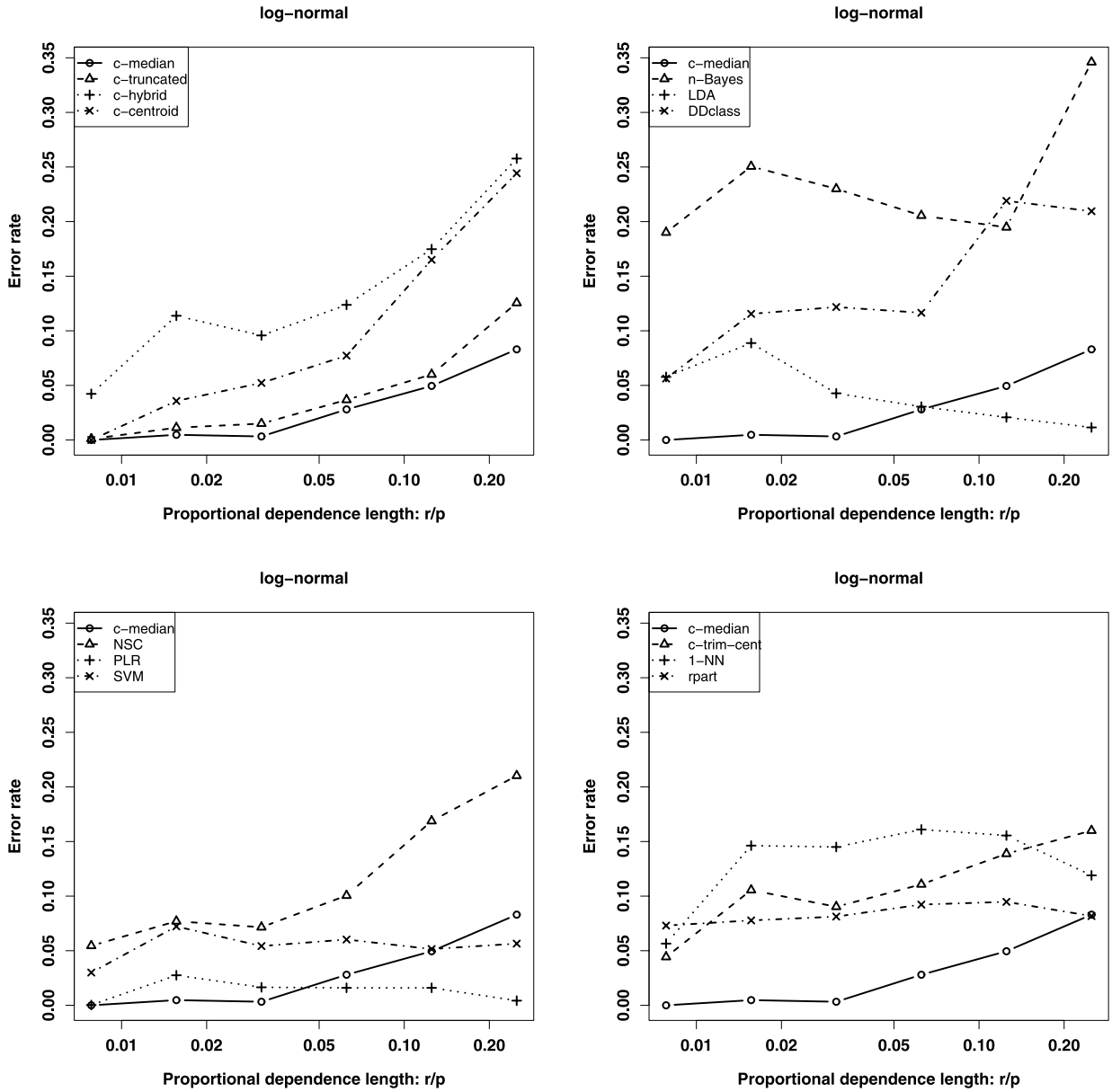

Figure 1. Mean misclassification-error rate vs. proportional dependence length r/p for the simulated lognormally-distributed data sets.

We next interpret assumptions (4.1)–(4.6). Assumption (4.1) mandates that the first moments of the variables U_k be uniformly bounded in a strong sense; for example, (4.1) holds if the U_k's are identically distributed with finite mean. Because the U_k's have uniquely defined median, equal to 0, then for all $x > 0$, $E|U_k \pm x| - E|U_k| > 0$, $\frac{1}{2} - P(U_k \le -x) > 0$, and $\frac{1}{2} - P(U_k \ge x) > 0$. Assumptions (4.2) and (4.3) require that these inequalities hold uniformly in k; if the U_k's are identically distributed, then the conditions hold under the basic assumption of uniquely defined median. Assumption (4.4) is a standard α-mixing condition that holds when, for example, the U_k's compose a stationary Gaussian process with autocovariance γ, where $\gamma(j) \to 0$ as $j \to \infty$. Constraint (4.6) asks that a nonnegligible proportion of the componentwise differences of medians be bounded away from 0, and (4.5) imposes the condition that the differences be bounded.

For each fixed $p \ge 1$, let Z denote a random variable drawn from either the X or the Y population. These events are indicated by $Z \in \Pi_X$ or $Z \in \Pi_Y$, respectively. Let $\mathcal{C}(Z)$ denote the median-based classifier determined by (2.7); it assigns Z to the X population if (2.7) holds and to the Y population otherwise. We express these decisions as $\mathcal{C}(Z) = X$ and $\mathcal{C}(Z) = Y$, respectively. Write P_X and P_Y for probability measures under the assumptions that $Z \in \Pi_X$ and $Z \in \Pi_Y$.

In the following theorem, we treat the sample sizes m and n as functions of vector length, p. But we make no assumption about the rate at which m and n diverge as p increases. In particular, they can increase at a much slower rate than p.

Theorem 1. Assume that (4.1)–(4.6) hold, and that both m and n diverge as $p \to \infty$. Then, with probability converging to 1 as p increases, the classifier $\mathcal{C}$ makes the correct decision,

$$P_X\{\mathcal{C}(Z) = Y\} + P_Y\{\mathcal{C}(Z) = X\} \to 0. \tag{4.7}$$

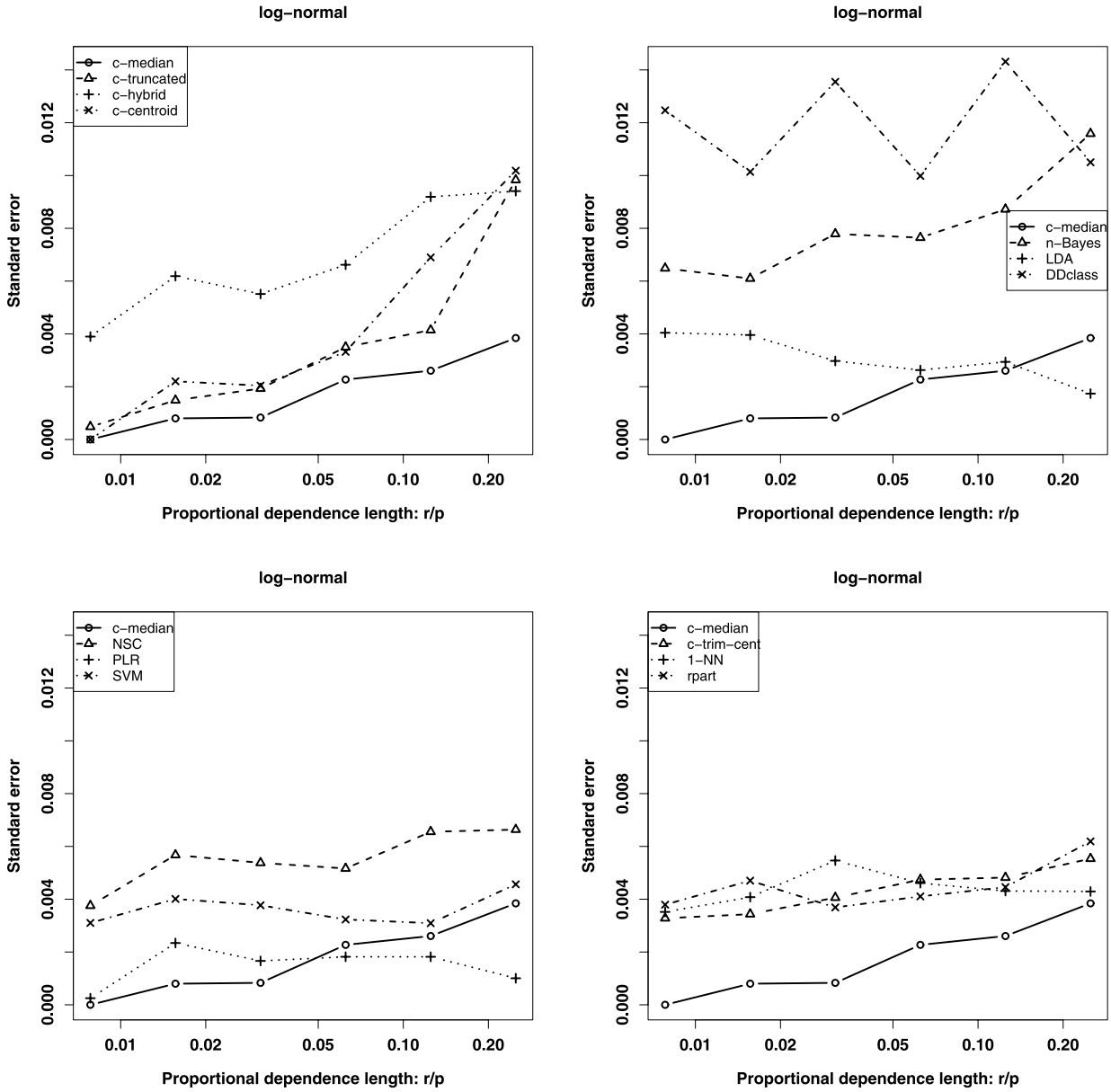

Figure 2. Standard error of mean misclassification-error rate vs. proportional dependence length r/p for the simulated lognormally-distributed data sets, corresponding to the results shown in Figure 1.

4.3 Properties Under Stronger Dependence and Smoothness Conditions

We assume that the data X_i and Y_j are generated in the same manner as described in Section 4.2; we alter only the regularity conditions on the distribution of $\vec{U}$ and on the median differences $\nu_{Xk} - \nu_{Yk}$. Suppose that the empirical medians computed from random samples drawn from the distribution of U_k converge to their asymptotic limits at the standard "root-n" rate, as n increases, in an L_1 sense and uniformly in k. Specifically, if $\mathcal{W}_k(\ell)$ denotes a sample consisting of ℓ independent random variables distributed as U_k, then, for a constant $C_1 > 0$ and for all sufficiently large ℓ,

$$\sup_{k \ge 1} E|\operatorname{med} \mathcal{W}_k(\ell)| \le C_1 \ell^{-1/2}. \tag{4.8}$$

For example, if the U_k's are identically distributed with 0 median, then this condition holds if the common distribution has a nonzero density at the origin. In place of (4.4), we impose a ψ-mixing condition,

$$\text{for a function } a \ge 0 \text{ satisfying } \textstyle\sum_{k \ge 1} a(k) < \infty, \text{ and for all } k \ge 1, \tag{4.9}$$

$$\sup_{k_1,k_2 \,:\, |k_1 - k_2| \ge k} \; \sup_{B_1, B_2 \in \mathcal{B}} \left| \frac{P(U_{k_1} \in B_1, U_{k_2} \in B_2)}{P(U_{k_1} \in B_1) P(U_{k_2} \in B_2)} - 1 \right| \le a(k).$$

In the ratio of probabilities in (4.9), $0/0$ is interpreted as 1. Mixing conditions and their implications have been discussed by Bradley (2005), who noted that, for example, (4.9) holds if the stochastic process $U_1, U_2, \ldots$ is Gaussian and r-dependent for some r.

Instead of (4.6), we impose the following weaker condition:

$$\text{for some } \epsilon > 0, \text{ the proportion of values } k \in [1, p] \text{ for which } |\nu_{Xk} - \nu_{Yk}| > \epsilon \text{ is of strictly larger order than } m^{-1/2} \text{ as } p \to \infty. \tag{4.10}$$

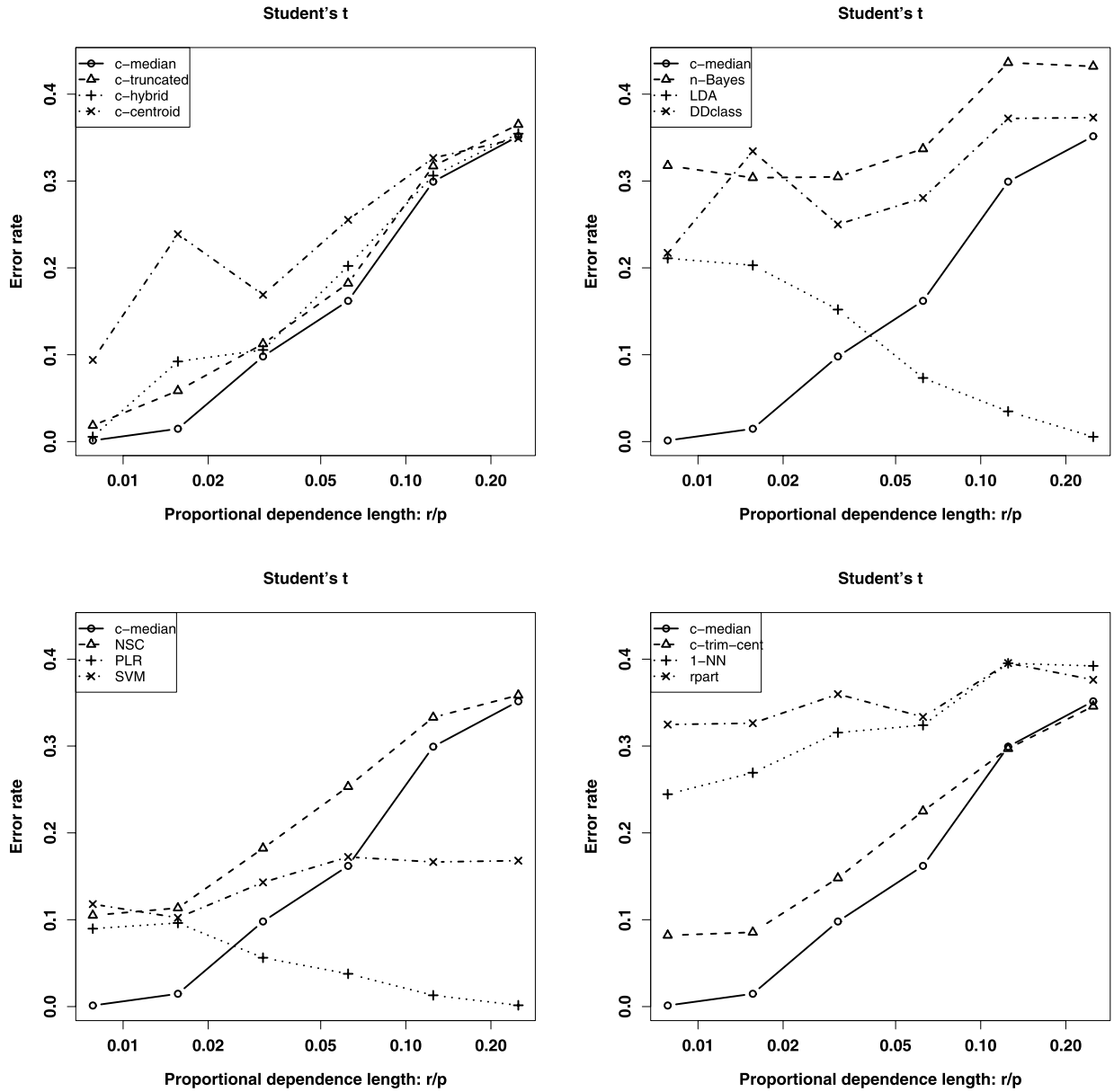

Figure 3. Mean misclassification-error rate vs. proportional dependence length r/p for the simulated Student's t-distributed data sets.

Finally, we ask that p be at least as large as a constant multiple of sample size, and that the sample sizes, m and n, be of the same order:

$$\text{for a constant } C_2 > 0,\ p \geq C_2 \max(m, n), \text{ and, additionally, } m/n \text{ is bounded away from 0 and infinity as } p \to \infty. \tag{4.11}$$

Condition (4.10) permits an order of magnitude greater degree of sparsity compared with (4.6). In particular, in (4.10) the proportion of nonzero signals can decrease to 0 as p increases, in fact, almost as fast as $p^{-1/2}$ if m and n are of the same size as p, without damaging the consistency of the classifier. Condition (4.11) asks that the training sample sizes be of the same order, and be bounded above by p in order of magnitude.

Theorem 2. If in Theorem 1 we replace conditions (4.1)–(4.6) by (4.2), (4.5), and (4.8)–(4.11), and if both m and n diverge as $p \to \infty$, then the conclusions of Theorem 1 continue to hold.

APPENDIX: TECHNICAL ARGUMENTS

A.1 Proof of Theorem 1

The argument here and in Section A.2 is abbreviated, and extended derivations are available online. Let ν_{Zk} denote the median of Z_k, and put $\nu_{XZk} = \nu_{Xk} - \nu_{Zk}$ and $\nu_{YZk} = \nu_{Yk} - \nu_{Zk}$. Write $\mathcal{U}_k$ and $\mathcal{V}_k$ for samples of sizes m and n of independent random variables each distributed as U_k and all independent of Z_k, and put $W_k = Z_k - \nu_{Zk}$, denoting another random variable distributed as U_k. Define

$$D_k = \big|\,|\operatorname{med} \mathcal{Y}_k - Z_k| - |\operatorname{med} \mathcal{X}_k - Z_k| - (|\nu_{YZk} - W_k| - |\nu_{XZk} - W_k|)\big|,$$

and take $\mathcal{E}_k(\epsilon)$ to be the set on which $D_k > \epsilon$. It can be proved that, under the conditions of Theorem 1,

for each $\epsilon > 0$,

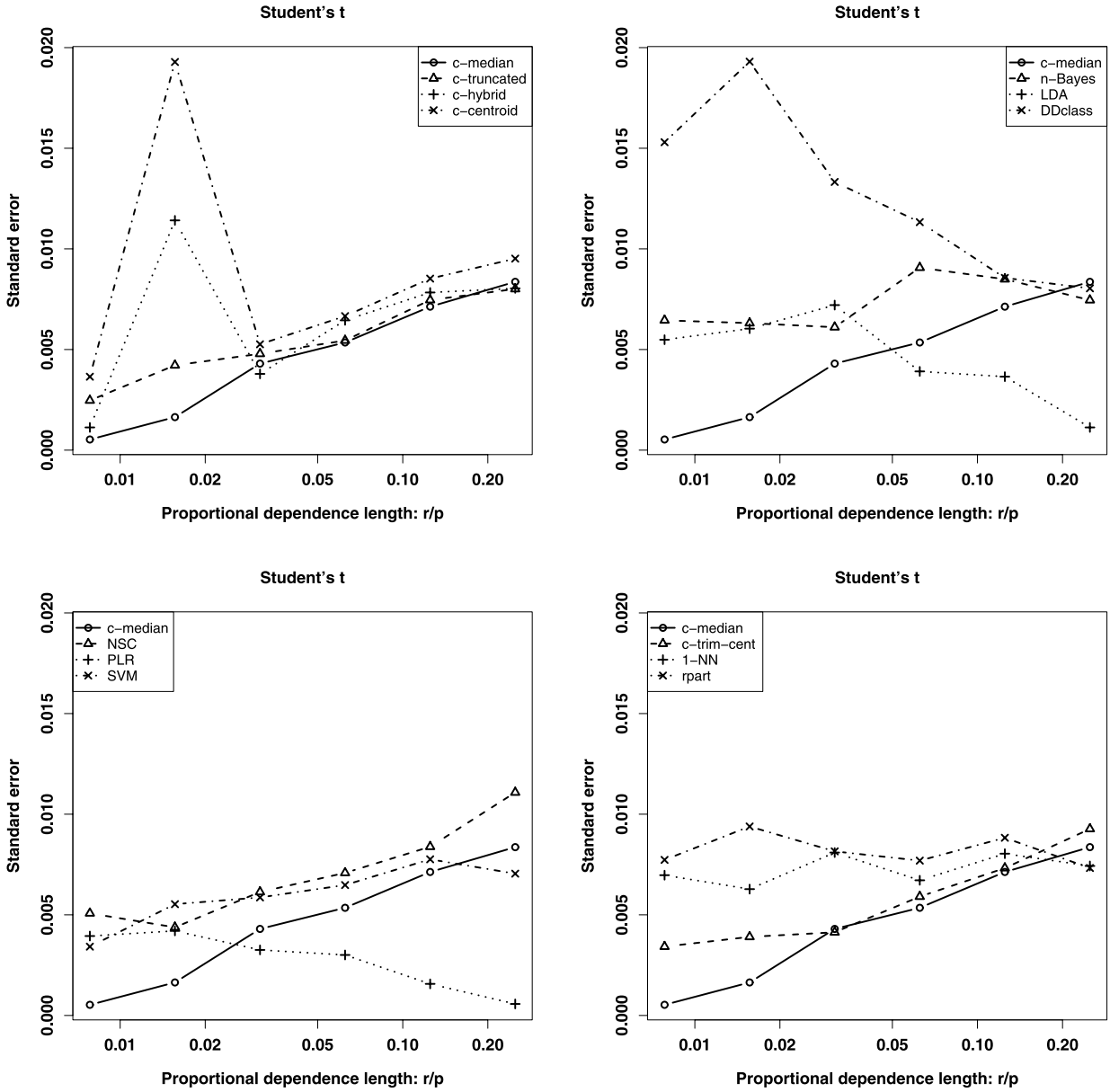

Figure 4. Standard error of mean misclassification-error rate vs. proportional dependence length r/p for the simulated Student's t-distributed data sets, corresponding to the results shown in Figure 3.

$$\lim_{p\to\infty} \sup_{1\le k\le p} \left\{P(|\operatorname{med}\mathcal{U}_k| > \epsilon) + P(|\operatorname{med}\mathcal{V}_k| > \epsilon)\right\} = 0, \tag{A.1}$$

$$\lim_{\lambda\to\infty} \sup_{k\ge 1} E\{D_k I(D_k > \lambda)\} = 0. \tag{A.2}$$

Result (A.1), and the fact that $\operatorname{med}\mathcal{X}_k - \nu_{Xk}$ and $\operatorname{med}\mathcal{Y}_k - \nu_{Yk}$ are distributed as $\operatorname{med}\mathcal{U}_k$ and $\operatorname{med}\mathcal{V}_k$, entail that for each $\epsilon > 0$, and with P denoting either P_X or P_Y, $\sup_{1\le k\le p} P\{\mathcal{E}_k(\epsilon)\} \to 0$ as $p \to \infty$, for each $\epsilon > 0$. This property and (A.2) imply that for each $\epsilon > 0$, $\sum_k E[D_k I\{\mathcal{E}_k(\epsilon)\}] = o(p)$. Thus, because

$$D(Z) \equiv \left| \sum_{k=1}^{p} (|\operatorname{med}\mathcal{Y}_k - Z_k| - |\operatorname{med}\mathcal{X}_k - Z_k|) - \sum_{k=1}^{p} (|\nu_{YZk} - W_k| - |\nu_{XZk} - W_k|) \right|$$

$$\le p\epsilon + \sum_{k=1}^{p} D_k I\{\mathcal{E}_k(\epsilon)\}$$

for each $\epsilon > 0$,

$$D(Z) = o_p(p). \tag{A.3}$$

Define $S_\lambda = \sum_k (|\nu_{YZk} - W_k| - |\nu_{XZk} - W_k|) I(|W_k| \le \lambda)$ and $d(Z) = S_\infty$. It can be deduced from (4.4) and (4.5) that for each fixed λ, $\operatorname{var}(S_\lambda) = o(p^2)$ as $p \to \infty$, and thus that $S_\lambda - E(S_\lambda) = o_p(p)$. This result and (4.5) can be used to prove that $d(Z) = E\{d(Z)\} + o_p(p)$, which, together with (A.3), implies that

$$\sum_{k=1}^{p} (|\operatorname{med}\mathcal{Y}_k - Z_k| - |\operatorname{med}\mathcal{X}_k - Z_k|) = E\{d(Z)\} + o_p(p). \tag{A.4}$$

Given $\epsilon > 0$, let $\mathcal{K}_\epsilon = \mathcal{K}_\epsilon(p)$ denote the set of indexes $k \in [1, p]$ such that $|\nu_{Xk} - \nu_{Yk}| \ge \epsilon$. If $k \in [1, p]$ but $k \notin \mathcal{K}_\epsilon$ then $||\nu_{YZk} - W_k| -$

$|\nu_{XZk} - W_k|| \le |\nu_{Xk} - \nu_{Yk}| \le \epsilon$. Thus, defining $d_k = E(|\nu_{YZk} - W_k| - |\nu_{XZk} - W_k|)$, we have

$$\left| E\{d(Z)\} - \sum_{k \in \mathcal{K}_\epsilon} d_k \right| \le p\epsilon. \tag{A.5}$$

Given $c > 0$, let $e(c)$ equal the infimum of $E|U_k + x| - E|U_k|$ over all $|x| > c$ and all $k \ge 1$. In view of (4.2), $e(c) > 0$. If Z is from the X population, then $\nu_{XZk} = 0$ and $\nu_{YZk} = \nu_{Yk} - \nu_{Xk} \equiv -\delta_k$, say, where, for $k \in \mathcal{K}_c$, $d_k = E_X|U_k + \delta_k| - E_X|U_k| \ge e(c)I(|\delta_k| > c)$. (Here and later, E_X denotes expectation under the assumption that Z is drawn from Π_X.) Therefore, provided that $c \ge \epsilon$,

$$\sum_{k \in \mathcal{K}_\epsilon} d_k \ge e(c)(\#\mathcal{K}_c). \tag{A.6}$$

We know from (4.6) that if c is taken sufficiently small, then, for a constant $b = b(c) > 0$, $(\#\mathcal{K}_c) \ge bp$ for all sufficiently large p. From this property and the fact that for arbitrarily small $\epsilon > 0$, $E_X\{d(Z)\} \ge e(c)(\#\mathcal{K}_c) - p\epsilon$ for all sufficiently large p [use (A.5) and (A.6)], it can be shown that $p^{-1}E_X\{d(Z)\}$ is bounded above 0 as $p \to \infty$. This result and (A.4) imply that there exists a constant $C > 0$ such that

$$\liminf_{p \to \infty} P_X\left\{ \sum_{k=1}^{p} (|\operatorname{med} \mathcal{Y}_k - Z_k| - |\operatorname{med} \mathcal{X}_k - Z_k|) \ge Cp \right\} = 1.$$

If follows that if Z is from Π_X, then the probability that the left side of (2.7) is positive or, equivalently, that the classifier assigns Z to the X population, converges to 1. Similarly, it can be proved that if Z is from Π_Y, then the probability that the classifier assigns Z to the Y population converges to 1. Together, these results establish (4.7).

A.2 Proof of Theorem 2

It can be proved that

$$\begin{aligned} T_1 &\equiv \sum_{k=1}^{p} (|\operatorname{med} \mathcal{Y}_k - Z_k| - |\operatorname{med} \mathcal{X}_k - Z_k|) \\ &= T_2 + \Theta_1 R_1 + \Theta_2 R_2, \end{aligned} \tag{A.7}$$

where Θ_1 and Θ_2 denote random variables satisfying $|\Theta_1|, |\Theta_2| \le 1$, $T_2 = \sum_k (|W_k - \nu_{YZk}| - |W_k - \nu_{XZk}|)$, $R_1 = \sum_k |\operatorname{med} \mathcal{Y}_k - \nu_{Yk}|$, and $R_2 = \sum_k |\operatorname{med} \mathcal{X}_k - \nu_{Xk}|$. Property (4.8) (with C_1 as defined there), (A.7), and Markov's inequality can be used to prove that for each choice of $c_1, c_2 > 0$, both of which may depend on p,

$$P_X\big(T_1 > c_1 - c_2 m^{-1/2} - c_2 n^{-1/2}\big) P_X(T_2 > c_1) - 2pc_2^{-1}C_1. \tag{A.8}$$

Let A_1 denote the upper bound to $|\mu_{Xk} - \nu_{Yk}|$ asserted in (4.5), and put $A_2 = 8A_1^2\{1 + \sum_{k \ge 1} a(k)\}$. Then, by (4.9),

$$E_X(T_2 - E_X T_2)^2 \le 8A_1^2 \left\{ p + \sum\sum_{1 \le k_1 < k_2 \le p} a(k_2 - k_1) \right\} \le A_2 p.$$

Thus, provided that $c_1 < \frac{1}{2} E_X(T_2)$,

$$\begin{aligned} P_X(T_2 > c_1) &\ge I(E_X T_2 > 2c_1) - P_X(|T_2 - E_X T_2| > c_1) \\ &\ge 1 - A_2 c_1^{-2} p. \end{aligned} \tag{A.9}$$

Also, if we define δ_k, $e(c)$, and $\mathcal{K}_c$ as in the proof of Theorem 1, from (4.2) and (A.6), it can be shown that for each $c > 0$, $E_X(T_2) = \sum_k (E|W_k + \delta_k| - E|W_k|) \ge e(c)(\#\mathcal{K}_c)$.

Define $\ell_{\min} = \min(m, n)$ and $\ell_{\max} = \max(m, n)$, and note that

$$p \ge C_2 \ell_{\max} \quad \text{and,} \quad \text{for some } c > 0, \quad \big(m^{1/2} \#\mathcal{K}_c\big)^{-1} p \to 0. \tag{A.10}$$

Given $\epsilon \in (0, 1)$, choose $A_3 > 0$ so large that

$$A_2 A_3^{-2} C_2^{-1} \le \tfrac{1}{2}\epsilon, \tag{A.11}$$

and put $A_4 = 4C_1\epsilon^{-1}$, $c_1 = A_3 p \ell_{\max}^{-1/2}$ and $c_2 = A_4 p$. Then $2pc_2^{-1} \times C_1 = \frac{1}{2}\epsilon$, and, in view of the first part of (A.10), $A_2 c_1^{-2} p = A_2 A_3^{-2} \times \ell_{\max} p^{-1} \le A_2 A_3^{-2} C_2^{-1} \le \frac{1}{2}\epsilon$. From these results, (A.8), and (A.9), it can be proved that

$$\begin{aligned} P_X\big(T_1 > c_1 - 2c_2 \ell_{\min}^{-1/2}\big) &\ge 1 - A_2 c_1^{-2} p - 2pc_2^{-1} C_1 \\ &\ge 1 - \epsilon. \end{aligned} \tag{A.12}$$

The condition $c_1 < \frac{1}{2} E(T_2)$, which is needed for (A.9) and thus for (A.12), will follow provided that $A_3 p m^{-1/2} < \frac{1}{2} e(c) \#\mathcal{K}_c$, which, due to the last part of (A.10), holds for all sufficiently large p. Thus, by (A.12), and provided that (A.10) and (A.11) hold, we have that

for sufficiently large p,

$$P_X\big(T_1 > A_3 p \ell_{\max}^{-1/2} - 8C_1\epsilon^{-1} p \ell_{\min}^{-1/2}\big) \ge 1 - \epsilon. \tag{A.13}$$

We wish (A.13) to imply that $P_X(T_1 > 0) \ge 1 - \epsilon$. For this, we need $A_3 p \ell_{\max}^{-1/2} - 8C_1\epsilon^{-1} p \ell_{\min}^{-1/2} > 0$, or, equivalently, $\ell_{\max}/\ell_{\min} < (A_3\epsilon/8C_1)^2$. Now C_1 is fixed by (4.8), and (A.11) is the only constraint imposed so far on A_3, so we may take $A_3 = \lambda(2A_2/C_2\epsilon)^{1/2}$ for any $\lambda \ge 1$. For this choice, $(A_3\epsilon/8C_1)^2 = A_2^2\lambda^2\epsilon/(32C_1^2C_2)$, which can be made as large as we wish by choosing λ sufficiently large.

Thus, for any given $\epsilon > 0$ and any given upper bound to $\ell_{\max}/\ell_{\min}$, we can choose $A_3 > 0$ such that (A.11) holds and $(A_3\epsilon/8C_1)^2$ strictly exceeds that upper bound or, equivalently, such that $A_3 p \ell_{\max}^{-1/2} - 8C_1\epsilon^{-1} p \ell_{\min}^{-1/2} > 0$. Therefore, by (A.13), $P_X(T_1 > 0) \ge 1 - \epsilon$ for all sufficiently large p. Because this is true for each $\epsilon > 0$, we have $P_X(T_1 > 0) \to 1$ and, similarly, $P_Y(T_1 < 0) \to 1$. This proves Theorem 2.

SUPPLEMENTAL MATERIALS

Extended Derivations: Extended derivations of theoretical properties. (hall-titerington-xue-jasa-suppl.pdf)

[Received February 2008. Revised April 2009.]

REFERENCES

Ahn, H., Moon, H., Fazzari, M. J., Lim, N., Chen, J. J., and Kodell, R. L. (2007), "Classification by Ensembles From Random Partitions of High-Dimensional Data," *Computational Statistics & Data Analysis*, 51, 6166–6179.

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," *Proceedings of the National Academy of Sciences of the United States of America*, 96, 6745–6750.

Baggerly, K. A., and Scott, D. W. (1999), Discussion of "Multivariate Analysis by Data Depth: Descriptive Statistics, Graphics and Inference," by R. Y. Liu, J. M. Parelius, and K. Singh, *The Annals of Statistics*, 27, 841–843.

Bose, P., Maheshwari, A., and Morin, P. (2003), "Fast Approximations for Sums of Distances, Clustering and the Fermat–Weber Problem," *Computational Geometry: Theory and Applications*, 24, 135–146.

Bradley, R. C. (2005), "Basic Properties of Strong Mixing Conditions. A Survey and Some Open Questions," *Probability Surveys*, 2, 107–144.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984), *Classification and Regression Trees*, Belmont, CA: Wadsworth.

Dabney, A. R. (2005), "Classification of Microarrays to Nearest Centroids," *Bioinformatics*, 21, 4148–4154.

Dabney, A. R., and Storey, J. D. (2005), "Optimal Feature Selection for Nearest Centroid Classifiers, With Applications to Gene Expression Microarrays," Working Paper 267, University of Washington.

——— (2007), "Optimality Driven Nearest Centroid Classification From Genomic Data," *PLoS One*, 2, e1002.

Ding, Y., Dang, X., Peng, H., and Wilkins, D. (2007), "Robust Clustering in High Dimensional Data Using Statistical Depths," *BMC Bioinformatics*, 8, S8.

Donoho, D. L., and Gasko, M. (1992), "Breakdown Properties of Location Estimates Based on Halfspace Depth and Projected Outlyingness," *The Annals of Statistics*, 20, 1803–1827.

Dudoit, S., Fridlyand, J., and Speed, T. P. (2002), "Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data," *Journal of the American Statistical Association*, 97, 77–87.

Fan, J., and Fan, Y. (2008), "High Dimensional Classification Using Features Annealed Independence Rules," *The Annals of Statistics*, 36, 2605–2637.

Fan, J., and Lv, J. (2008), "Sure Independence Screening for Ultra-High Dimensional Feature Space" (with discussion), *Journal of the Royal Statistical Society, Ser. B*, 70, 849–911.

Feinerer, I., Hornik, K., and Meyer, D. (2008), "Text Mining Infrastructure in R," *Journal of Statistical Software*, 25, 1–54.

Gentleman, W. M. (1965), "Robust Estimation of Multivariate Location by Minimizing *p*th Power Deviations," unpublished Ph.D. thesis, Princeton University.

Ghosh, A. K., and Chaudhuri, P. (2005), "On Data Depth and Distribution-Free Discriminant Analysis Using Separating Surfaces," *Bernoulli*, 11, 1–27.

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.

Gordon, G. J., Jensen, R. V., Hsiao, L. L., Gullans, S. R., Blumenstock, J. E., Ramaswamy, S., Richards, W. G., Sugarbaker, D. J., and Bueno, R. (2002), "Translation of Microarray Data Into Clinically Relevant Cancer Diagnostic Tests Using Gene Expression Ratios in Lung Cancer and Mesothelioma," *Cancer Research*, 62, 4963–4967.

Guyon, I., and Elisseeff, A. (2003), "An Introduction to Variable and Feature Selection," *Journal of Machine Learning Research*, 3, 1157–1182.

Hall, P., Titterington, D. M., and Xue, J.-H. (2009), "Tilting Methods for Assessing the Influence of Components in a Classifier," *Journal of the Royal Statistical Society, Ser. B*, 71, 783–803.

Hand, D. J. (2006), "Classifier Technology and the Illusion of Progress" (with discussion), *Statistical Science*, 21, 1–34.

Hand, D. J., and Yu, K. (2001), "Idiot's Bayes—Not so Stupid After All?" *International Statistical Review*, 69, 385–398.

Hua, J., Tembe, W. D., and Dougherty, E. R. (2009), "Performance of Feature-Selection Methods in the Classification of High-Dimension Data," *Pattern Recognition*, 42, 409–424.

Jörnsten, R. (2004), "Clustering and Classification Based on the L_1 Data Depth," *Journal of Multivariate Analysis*, 90, 67–89.

Kowalski, J., and Powell, J. (2004), "Nonparametric Inference for Stochastic Linear Hypotheses: Application to High-Dimensional Data," *Biometrika*, 91, 393–408.

Lewis, D. (1997), "Reuters-21578 Text Categorization Collection Distribution 1.0," available at *http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html*.

Liu, R. Y. (1990), "On a Notion of Data Depth Based on Random Simplices," *The Annals of Statistics*, 18, 405–414.

Liu, R. Y., and Singh, K. (1999), "Rejoinder: Multivariate Analysis by Data Depth: Descriptive Statistics, Graphics and inference," by R. Y. Liu, J. M. Parelius, and K. Singh, *The Annals of Statistics*, 27, 854–858.

Markatou, M., Tian, H., Biswas, S., and Hripcsak, G. (2005), "Analysis of Variance of Cross-Validation Estimators of the Generalization Error," *Journal of Machine Learning Research*, 6, 1127–1168.

McLachlan, G. J. (1972), "An Asymptotic Expansion for the Variance of the Errors of Misclassification of the Linear Discriminant Function," *Australian Journal of Statistics*, 14, 68–72.

——— (1976), "The Bias of the Apparent Error Rate in Discriminant Analysis," *Biometrika*, 63, 239–244.

Moon, H., Ahn, H., Kodell, R. L., Lin, C. J., Baek, S., and Chen, J. J. (2006), "Classification Methods for the Development of Genomic Signatures From High-Dimensional Data," *Genome Biology*, 7, R121.1–R121.7.

Nadeau, C., and Bengio, Y. (2003), "Inference for the Generalization Error," *Machine Learning*, 52, 239–281.

Nahar, J., Ali, S., and Chen, Y. P. (2007), "Microarray Data Classification Using Automatic SVM Kernel Selection," *DNA and Cell Biology*, 26, 707–712.

Oja, H. (1983), "Descriptive Statistics for Multivariate Distributions," *Statistics & Probability Letters*, 1, 327–333.

Park, M. Y., and Hastie, T. (2008), "Penalized Logistic Regression for Detecting Gene Interactions," *Biostatistics*, 9, 30–50.

Sebastiani, F. (2002), "Machine Learning in Automated Text Categorization," *ACM Computing Surveys*, 34, 1–47.

Singh, D., Febbo, P. G., Ross, K., Jackson, D. G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A. A., D'Amico, A. V., Richie, J. P., Lander, E. S., Loda, M., Kantoff, P. W., Golub, T. R., and Sellers, W. R. (2002), "Gene Expression Correlates of Clinical Prostate Cancer Behavior," *Cancer Cell*, 1, 203–209.

Tibshirani, R., and Hastie, T. (2007), "Margin Trees for High-Dimensional Classification," *Journal of Machine Learning Research*, 8, 637–652.

Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. (2002), "Diagnosis of Multiple Cancer Types by Shrunken Centroids of Gene Expression," *Proceedings of the National Academy of Sciences of the United States of America*, 99, 6567–6572.

——— (2003), "Class Prediction by Nearest Shrunken Centroids, With Applications to DNA Microarrays," *Statistical Science*, 18, 104–117.

Wang, L., Zhu, J., and Zou, H. (2008), "Hybrid Huberized Support Vector Machines for Microarray Classification and Gene Selection," *Bioinformatics*, 24, 412–419.

Wang, S., and Zhu, J. (2007), "Improved Centroids Estimation for the Nearest Shrunken Centroid Classifier," *Bioinformatics* 23, 972–979.

Winter, S. S., Jiang, Z., Khawaja, H. M., Griffin, T., Devidas, M., Asselin, B. L., and Larson, R. S. (2007), "Identification of Genomic Classifiers that Distinguish Induction Failure in T-Lineage Acute Lymphoblastic Leukemia: A Report From the Children's Oncology Group," *Blood*, 110, 1429–1438.

Yeung, K. Y., and Bumgarner, R. E. (2003), "Multiclass Classification of Microarray Data With Repeated Measurements: Application to Cancer," *Genome Biology*, 4, R83.

Zucknick, M., Richardson, S., and Stronach, E. A. (2008), "Comparing the Characteristics of Gene Expression Profiles Derived by Univariate and Multivariate Classification Methods," *Statistical Applications in Genetics and Molecular Biology*, 7, Article 7.