On selecting interacting features from high-dimensional data

Peter Hall[a], Jing-Hao Xue[b,*]

[a]*Department of Mathematics and Statistics, The University of Melbourne, VIC 3010, Australia*
[b]*Department of Statistical Science, University College London, London WC1E 6BT, UK*

Abstract

For high-dimensional data most feature-selection methods, such as SIS and the lasso, involve ranking and selecting features individually. These methods do not require many computational resources, but they ignore feature interactions. A simple recursive approach which, without requiring many more computational resources, also allows identification of interactions, is investigated. This approach can lead to substantial improvements in the performance of classifiers, and provide insight into the way in which features work together in a given population. It also enjoys attractive statistical properties.

Keywords: Classification, correlation, generalised correlation, feature ranking

1. Introduction

Feature selection is an important technique for high-dimensional data analysis. Excellent reviews and comparisons have been given by Guyon and Elisseeff (2003), Saeys et al. (2007), Hua et al. (2009) and Fan and Lv (2010), for example. The method is often implemented through ranking features in terms of certain criteria. Among feature-ranking approaches, correlation-based methods are technically simple, theoretically elegant and widely used in practice (Guyon and Elisseeff, 2003; Saeys et al., 2007). Correlation-based

*Corresponding author. Tel.: +44-20-7679-1863; Fax: +44-20-3108-3105
Email addresses: halpstat@ms.unimelb.edu.au (Peter Hall), jinghao.xue@ucl.ac.uk (Jing-Hao Xue)

ranking can be implemented either in a univariate paradigm, ranking single features independently (see e.g. Fan and Lv (2008), Hall and Miller (2009)), or in a multivariate paradigm, ranking subsets of single features iteratively (see e.g. Hall (2000)). For high-dimensional data most feature-selection methods, such as "sure independence screening" (SIS) (Fan and Lv, 2008) and the lasso (Tibshirani, 1996), reflect the univariate paradigm. In this setting, practical and theoretical justifications have been provided by Saeys et al. (2007), Fan and Lv (2008) and Fan et al. (2009), among others.

These feature-wise methods do not require many computational resources, but they ignore feature interactions. To overcome this problem a simple recursive approach is investigated, also requiring relatively few computational resources but nevertheless allowing identification of interactions. Those interactions are of didactic interest, in terms of the information that they convey about how features work together in a particular population. The new approach also enables remarkable improvements in the performance of simple, classical classifiers, for example the centroid-based classifier and the nearest-neighbour (1-NN) classifier. It also enjoys attractive statistical properties for additive models.

2. Methodology

2.1. Ranking based on (generalised) correlation

Suppose independent and identically distributed data pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$ are observed, where the explanatory variables X_i are p-vectors and the response variables Y_i are scalars.

Traditionally, correlation-based feature-ranking methods use conventional linear correlation to measure the association between features X_{ij} and responses Y_i. Linear correlation can be used in cases where the Y_is take values in the continuum and, via a logit model, in cases where they are class labels, taking for example the values 0 and 1. Treating nonlinear relationships between X_{ij} and Y_i, Hall and Miller (2009) adopts generalised correlation, as follows.

Let $\mathcal{H}$ denote a vector space of functions, which is assumed to include all linear functions. If Y_i takes a continuum of values then theoretical and empirical measures of generalised correlation between Y_i and the jth component, X_{ij}, of X_i are given respectively by

$$\rho_j = \sup_{h \in \mathcal{H}} \frac{\mathrm{cov}\{h(X_{1j}), Y_1\}}{\sqrt{\mathrm{var}\{h(X_{1j})\}\,\mathrm{var}(Y_1)}} \tag{1}$$

and

$$\hat{\rho}_j = \sup_{h \in \mathcal{H}} \frac{\sum_i \{h(X_{ij}) - \bar{h}_j\} (Y_i - \bar{Y})}{\sqrt{\sum_i \{h(X_{ij})^2 - \bar{h}_j^2\} \cdot \sum_i (Y_i - \bar{Y})^2}} \,, \tag{2}$$

where $\bar{h}_j = n^{-1} \sum_i h(X_{ij})$ and $\bar{Y} = n^{-1} \sum_i Y_i$. Of course, the factors $\mathrm{var}(Y_1)$ and $\sum_i (Y_i - \bar{Y})^2$, in the denominators at (1) and (2), do not depend on j, and so they could be dropped without affecting ranking-based methodology. To rank the feature indices $1, \ldots, p$ in decreasing order of the apparent influence that the explanatory variables have on the response, the $\hat{\rho}_j$s are first ranked in decreasing order, obtaining

$$\hat{\rho}_{\hat{\jmath}_1} \geq \hat{\rho}_{\hat{\jmath}_2} \geq \ldots \geq \hat{\rho}_{\hat{\jmath}_p} \,, \tag{3}$$

say. In this notation the ranking that is sought is

$$\hat{\jmath}_1 \succeq \ldots \succeq \hat{\jmath}_p \,. \tag{4}$$

If $\mathcal{H}$ were restricted to just its linear elements then the quantities in (1) and (2) would be the absolute values of conventional correlation coefficients.

Next, cases where Y_i is a categorical variable are considered. They include instances where $Y_i = I_i$, equal to either 0 or 1. If the index j lies between 1 and p then the relationship between I_i and X_{ij} is captured in a "functional logistic model", where

$$P(I_i = 0 \,|\, h, X_{ij}) = \{1 + \exp h(X_{ij})\}^{-1}$$

and

$$P(I_i = 1 \,|\, h, X_{ij}) = 1 - P(I_i = 0 \,|\, h, X_{ij}) \,,$$

with $h \in \mathcal{H}$. The likelihood of I_i, given X_{ij}, is

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

where $t_{ij} = \exp h(X_{ij})$. Therefore the negative log-likelihood, $\ell_{ij}(h) = -\log L_{ij}(I_i \,|\, h, X_{ij})$, is

$$\ell_{ij}(h) = -I_i \, h(X_{ij}) + \log\{1 + \exp h(X_{ij})\} \,,$$

and its analogue for $X_{1j}, \ldots, X_{nj}$ is $\ell_j(h) = n^{-1} \sum_i \ell_{ij}(h)$.

Put $\hat{\ell}_j = \inf_{h \in \mathcal{H}} \ell_j(h)$. In the present setting, $\hat{\ell}_j$ plays the role of $\hat{\rho}_j$, at (1), and in particular the ordering at (4) is now determined by

$$\hat{\ell}_{\hat{\jmath}_1} \le \ldots \le \hat{\ell}_{\hat{\jmath}_p} \,, \tag{5}$$

rather than (3). The conventional linear logistic model is obtained by taking $\mathcal{H}$ to be the vector space of linear functions only. Applications of maximum marginal likelihood to rank features have been discussed in Hall and Miller (2009) and Fan et al. (2009), for example.

2.2. Recursive selection of interactions

The suggested approach proceeds in a sequence of steps, as follows.

2.2.1. Step 1: Ranking single features

Using the approach at (3), in the case of a continuous response, or at (5), for a categorical response; or employing a related method; compute a ranking (4) of $p_1 = p$ vector-component indices. On this occasion it is helpful to write (4) as

$$\hat{\jmath}_{11} \succeq \ldots \succeq \hat{\jmath}_{1p} \,, \tag{6}$$

where the first subscript "1" signifies "first step".

2.2.2. Step 2: Ranking single features and pairwise interactions together

Let indices $\hat{\jmath}_{11}, \ldots, \hat{\jmath}_{1p}$ be as in (6). Consider lengthening the vector $X_i = X_i(1)$ to $X_i(2)$, say, by adjoining to the existing features all the "cross features" formed from pairwise interactions, i.e. $X_{i\hat{\jmath}_{1k_1}} X_{i\hat{\jmath}_{1k_2}}$ for $1 \le k_1 < k_2 \le p$. The data vectors $X_i(2)$ are of length $p + \frac{1}{2}\, p\,(p-1) = \frac{1}{2}\, p\,(p+1)$, and, for the sake of definiteness, the adjoined components are listed after the components of $X_i(1)$. However, $\frac{1}{2}\, p\,(p+1)$ usually will be too large for the available computational resources, and so length is reduced to p_2, say, principally by reducing the number of cross features, and to a lesser extent reducing the number of single components.

The most appropriate tradeoff is not specified, since it can vary from one problem to another. However, if computational resources were sufficient to allow ranking of all p features in Step 1, they would likely be adequate to allow ranking of p single features together with $\frac{1}{2}\, \sqrt{p}\,(\sqrt{p}-1)$ pairwise interactions, resulting in a ranking of $p_2 = p + \frac{1}{2}\, \sqrt{p}\,(\sqrt{p}-1)$ numbers. Pairwise interactions are computed from the $\sqrt{p}$ features that were ranked most highly in Step 1.

More generally, when reducing dimension to p_2, attention continues to be confined to single and cross features whose indices are the most highly ranked. In the case of the cross feature $X_{i\hat{\jmath}_{1k_1}} X_{i\hat{\jmath}_{1k_2}}$, dimension can be reduced by placing an upper bound on the size of either $\max(k_1, k_2)$ or $k_1 + k_2$. In a slight abuse of notation, write $X_i(2)$ for the resulting p_2-vector.

Next, Step 1 is repeated for the data $X_i(2)$, obtaining in place of (6) a ranking

$$\hat{\jmath}_{21} \succeq \ldots \succeq \hat{\jmath}_{2p_2} \,, \tag{7}$$

based on the empirical measure of association between Y_i and the components of $X_i(2)$. This ranking indicates the relative strengths of relationships among single-variable contributions and some of the pairwise interactions, taken together.

An attractive alternative approach might be to group features into ranked classes whose sizes are computationally manageable, and use the top ranked features within a class to construct interactions among the classes. However, it is not clear what criteria should be used to define the classes, or to rank features within a class. Conventional correlation may not be appropriate here, since, in a highly correlated pair of features, it is often the case that one of the features is redundant if the other is included; and clearly low correlation is not appropriate either. Searching for high correlation is certainly more attractive than looking for low correlation, but it is not really an appropriate method for identifying features that complement, rather than substitute for, features that have already been determined. For these reasons the grouping idea will not be explored further here.

3. Theoretical Properties

Assume that the response variable, Y_i, depends on only the first $r \leq p$ variables in the explanatory vector, $X_i = (X_{i1}, \ldots, X_{ip})$, and that there is interaction between X_{ij_1} and X_{ij_2} but not among any of the other components of X_i. These relationships can be expressed in an additive model with interaction term:

$$Y_i = \sum_{j=1}^{r} g_j(X_{ij}) + \gamma_1(X_{ij_1})\,\gamma_2(X_{ij_2}) + \epsilon_i,\ 1 \leq i \leq n, \tag{8}$$

where $r = r(n) \leq p$ and can diverge as $n \to \infty$; $p = p(n) \to \infty$ as $n \to \infty$; and $g_1, \ldots, g_r$, γ_1 and γ_2 are functions.

Specific regularity conditions $(A.1)$ and $(A.4)$ will be given in Appendix A. Among other properties, those conditions permit r to diverge with n; they take $g_1, \ldots, g_{r_0}$ to be fixed, where $r_0 \leq r - 2$ is a fixed integer; they allow $g_{r_0+1}, \ldots, g_r$ to depend on n; they take $j_1 = r_0 + 1$ and $j_2 = r_0 + 2$ (for notational convenience) and ask that correlations involving $g_j(X_{ij_1})$ and $g_j(X_{ij_2})$ do not decrease too quickly; and they take p', of size $\sqrt{p}$, to be the number of top-ranked features, in the first step, that are used to construct interactions. Against this background, the theorem below argues that the components of the interaction term in (8) can be well down in the first ranking of features, at (6), but that nevertheless the interaction term typically enjoys a high position in the second ranking, at (7). Similar results can be derived when using logistic-based correlation analysis in instances where the response Y_i is a zero-one class label.

Theorem 1. *Assume the model represented by* (8)*, that conditions* $(A.1)$ *and* $(A.4)$ *hold, and that at each step in the algorithm in section 2.2 generalised correlations, described in section 2.1, were used to effect the rankings. Then, with probability converging to 1 as* $n \to \infty$*,*

(a) in the ranking at (6)*, the sequence* $\hat{\jmath}_{11}, \ldots, \hat{\jmath}_{1,r_0}$ *is a permutation of* $1, \ldots, r_0$*, and* $r_0 + 1$ *and* $r_0 + 2$ *are included among* $\hat{\jmath}_{1,r_0+1}, \ldots, \hat{\jmath}_{1p'}$*;*

(b) in the ranking at (7)*, the sequence* $\hat{\jmath}_{21}, \ldots, \hat{\jmath}_{2,r_0+1}$ *is a permutation of* $1, \ldots, r_0$*, together with whatever index corresponds to the pair* (j_1, j_2)*.*

A proof of Theorem 1 is given in Appendix B.

4. Numerical Studies

4.1. Classification of benchmark datasets

In microarray studies the response Y_i is often binary and the explanatory variables X_i are ultra-high dimensional. An example is the identification of differentially expressed genes among the components X_{ij} of X_i. In such a classification problem, conventional and generalised correlations have both been used to select single features; see Fan and Lv (2008) for the correlation-based SIS method and Hall and Miller (2009) for the generalised-correlation-based "functional logistic model", for example. Computationally, fitting logistic regression is much slower than calculating correlation.

Section 4.1.1 will give results obtained using linear correlation for two ultra-high-dimensional microarray datasets, and section 4.1.2 will present results for generalised correlation applied to two moderate-dimensional datasets, relating to cardiac imaging and ionospheric structure, respectively.

It is desired to know whether the suggested recursive approach, which selects both top-ranked single and cross features, can provide better classification performance than SIS and the "functional logistic model", both of which use top-ranked single features only. For illustrative purposes the selected features are used in two popular, simple classifiers, the centroid-based and 1-NN classifiers. The former classifies a new data vector, X, into the class for which the empirical centroid is closest to X; the latter classifies X into the class that contains the data vector closest to X.

4.1.1. Classification of microarray datasets

The two microarray datasets used here are for hepatocellular carcinoma (Iizuka et al., 2003) and prostate cancer (Singh et al., 2002), respectively. They have been used often in literature to evaluate classification methods for high-dimensional data (Pochet et al., 2004; Fan and Fan, 2008; Hall et al., 2009). The hepatocellular carcinoma dataset (denoted by "Hepatocellular" below) involves 12 patients with early intrahepatic recurrence, and 21 patients without recurrence in the predetermined training set; and 8 patients with recurrence and 19 without in the test set. Each patient is represented by a data vector of $p = 7129$ genes. The prostate cancer dataset ("Prostate") involves $p = 12600$ genes for 50 normal and 52 tumour vectors in the training set, and 9 normal and 25 tumour vectors in the test set.

In each analysis, as in other studies such as those of Hall et al. (2009) and Dudoit et al. (2002), a four-step procedure is used to preprocess all the data: truncating intensities to make them positive; removing genes having little variation in intensity; transforming intensities to base 10 logarithms; and standardising each data vector to have zero mean and unit variance. This procedure keeps $p = 1627$ genes for the "Hepatocellular" dataset, and $p = 3239$ genes for the "Prostate" dataset. The new data are then used for feature selection and classification.

To evaluate the performance of feature selection and classification, the standard n-out-of-n bootstrap technique (i.e. random sampling with replacement) is applied to the training set, to construct a new training set of size equal to the original one. The new training set is used for feature selection and for classifier training; the predetermined test set is then used to calculate misclassification error rates. Such a procedure is repeated 200 times. The average error rates (and their standard errors) over these 200 replicates are plotted in Fig. 1 (and Fig. 2), versus selecting 1 to 50 top-ranked single features using SIS, or single plus cross features using the suggested approach.

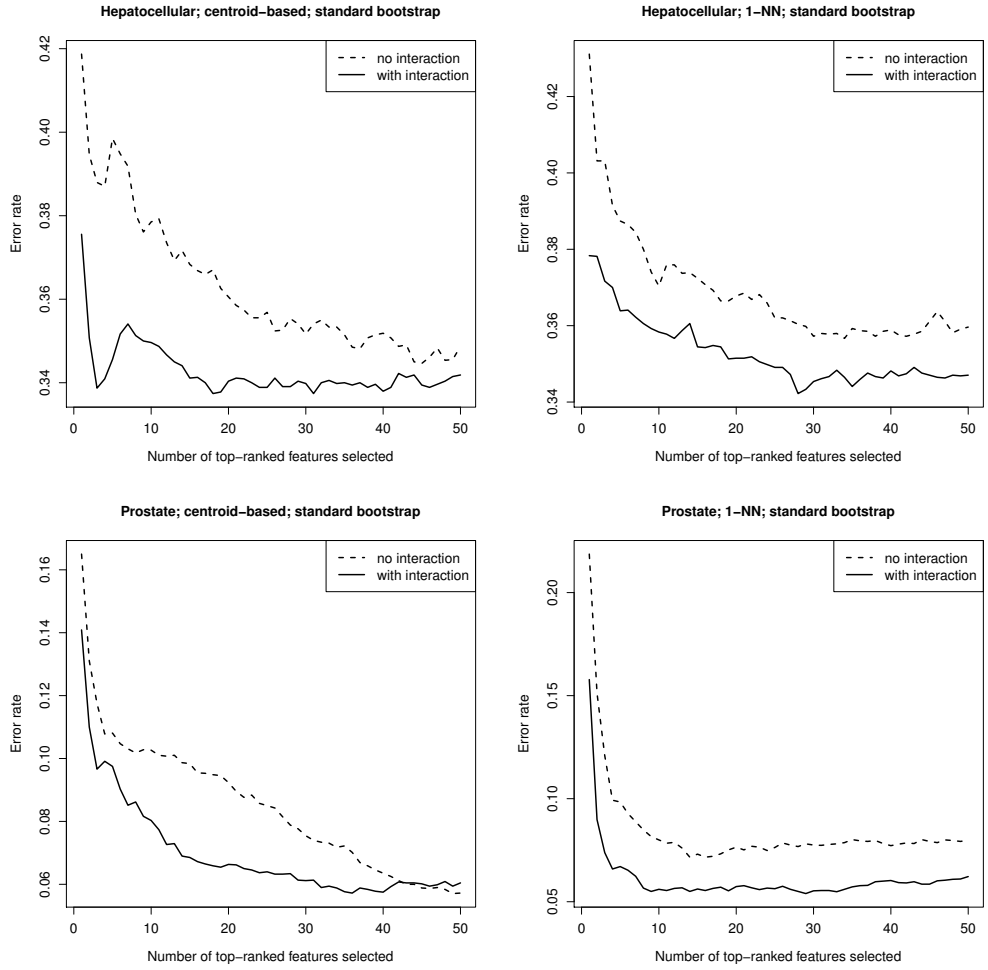

Figure 1: Misclassification error rates (ER) for the "Hepatocellular" and "Prostate" datasets, obtained from applying the centroid-based and 1-NN classifiers to the predetermined test sets. The vertical axis gives the average ER over classification results corresponding to 200 bootstrap replicates of the training set. The horizontal axis gives the number of top-ranked single features, or single features plus interactions, selected. (Dashed line: ranking single features only; solid line: recursively ranking both single features and pairwise interactions. Left-hand column: for the centroid-based classifier; right-hand column: for 1-NN.)

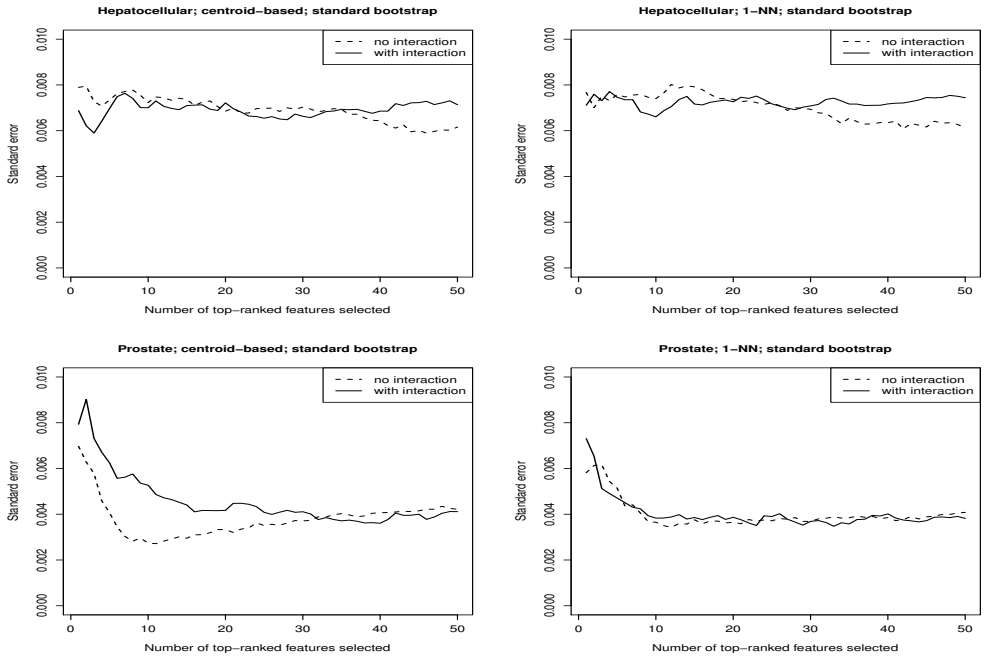

Figure 2: Standard errors of the average ER in Fig. 1.

When using the suggested approach, all $p + \frac{1}{2}\sqrt{p}\,(\sqrt{p} - 1)$ single and cross features were ranked.

From Fig. 1 it can be seen that, for the "Hepatocellular" and "Prostate" datasets, compared with SIS which selects only influential single features, the suggested approach, which chooses both influential single and cross features, enables both the 1-NN and centroid-based classifiers to achieve better out-of-sample classification. The improvement is consistent over varying numbers of selected features. From Fig. 2 it can be observed that the standard errors are of similar magnitudes for SIS and the suggested approach, and are relatively stable when different numbers of features were selected, except for the "Prostate" dataset when only a small number of genes were selected.

4.1.2. Classification of UCI datasets

In this section a comparison is made of the performance of the suggested approach and a single-feature-only method for classifying two widely-adopted benchmark datasets from the the University of California at Irvine (UCI) machine learning repository, available at archive.ics.uci.edu/ml. The first dataset ("Heart") comprises 267 cardiac SPECT images including 55 normal cases and 212 abnormal cases, with each image represented by a data vector of $p = 44$ features (Kurgan et al., 2001). The second dataset ("Ionosphere") consists of 351 radar returns used to detect evidence of structure in the ionosphere, including 225 "good" returns and 126 "bad" returns, for each of $p = 33$ effective features (Sigillito et al., 1989). Compared with the microarray datasets in section 4.1.1, these two UCI datasets can be viewed as illustrating moderate-dimensional properties.

The "Heart" dataset has predetermined training and test sets but the "Ionosphere" dataset has not, hence each of the full original datasets was randomly split equally into a new training set and a new test set with the original proportion of the two classes preserved. Fig. 3 shows the average misclassification error rates over 400 random splits, and the corresponding standard errors are displayed in Fig. 4.

From Fig. 3 it can be seen that, for the two datasets, taking interactions into consideration for feature selection generally improves classification performance for both the 1-NN and centroid-based classifiers. The pattern is similar to that in Fig. 1. This reflects the fact that the suggested approach produces pairwise interactions that have more influence on the response than some important single features. That can be particularly insightful to a practitioner. The standard errors shown in Fig. 4 demonstrate that the perfor-

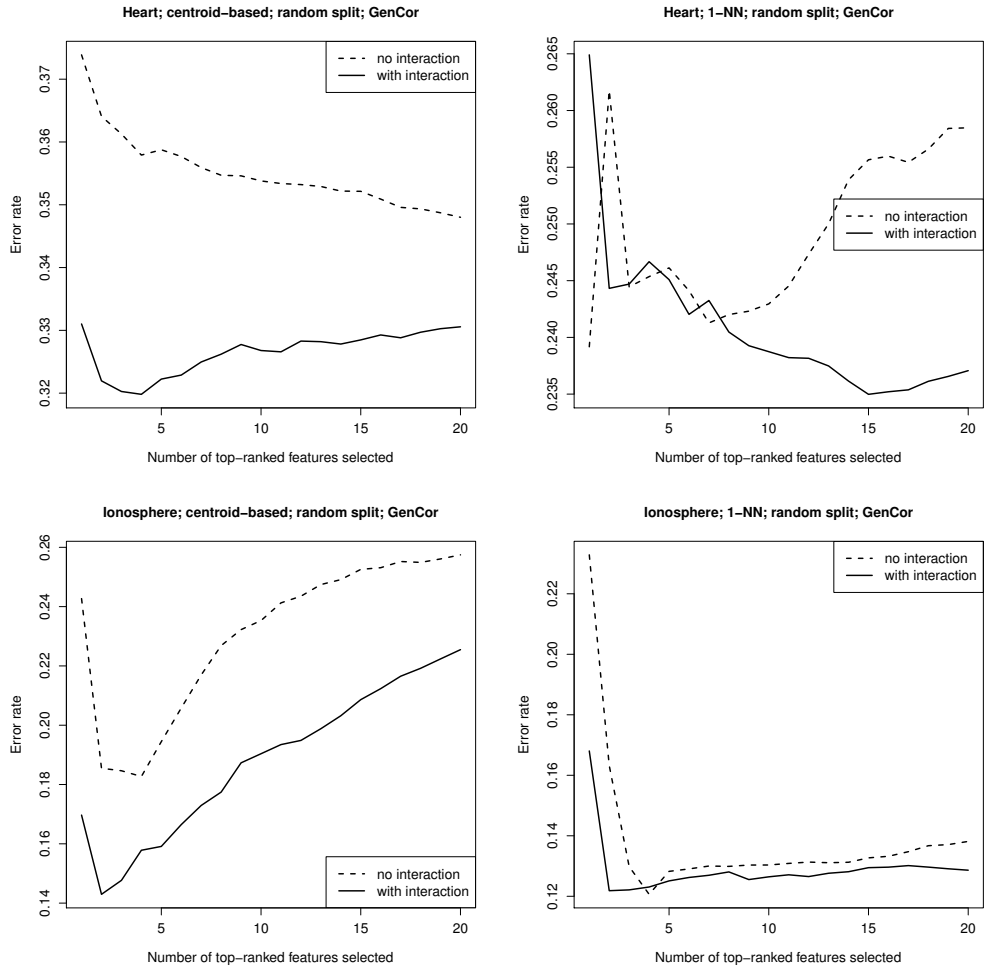

Figure 3: Misclassification error rates (ER) for the "Heart" and "Ionosphere" datasets, obtained from applying the centroid-based and 1-NN classifiers to test sets obtained by randomly splitting 400 times. Feature selection is based on generalised correlation (indicated by "GenCor" in each panel). Other notation is as for Fig. 1.

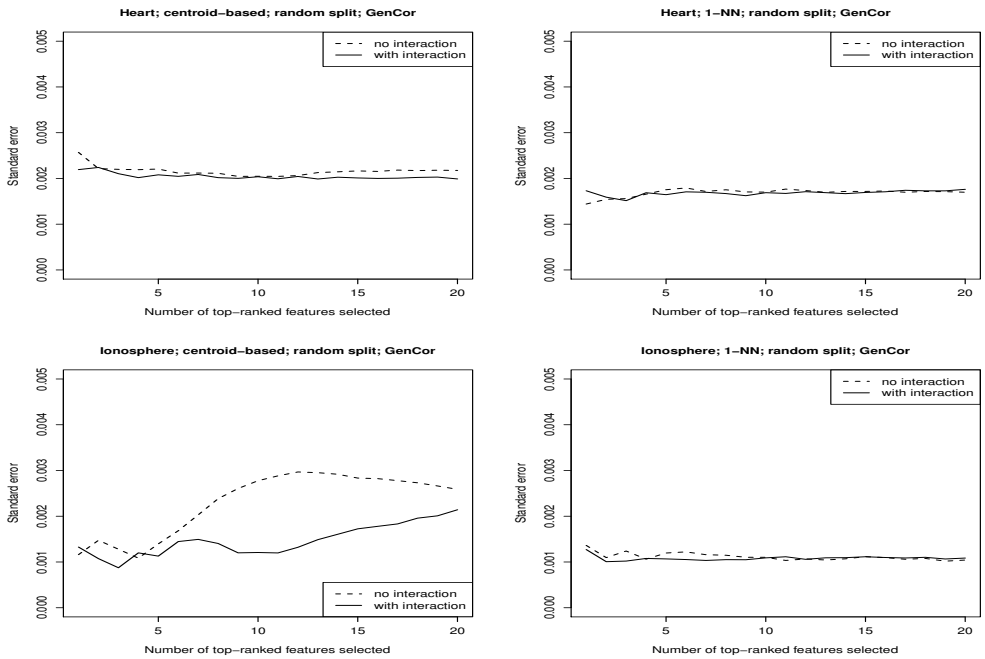

Figure 4: Standard errors of the average ER in Fig. 3.

mance of the suggested approach is more stable with the number of selected features, compared with that of the method selecting no interactions.

4.2. Two extensions of the suggested approach

The suggested approach can be generalised to ranking all single features and all $\frac{1}{2}\,p\,(p-1)$ pairwise interactions together, if computational feasible, and to ranking features using p-values instead of conventional and generalised correlations. In sections 4.2.1 and 4.2.2, respectively, empirical comparisons will be made of the originally suggested approach with these two extensions.

4.2.1. Extension-1: ranking all interactions

In this section an investigation is undertaken into the effects of limiting the search for pairwise interactions to the top $\frac{1}{2}\,\sqrt{p}\,(\sqrt{p}-1)$ pairs among the initially top-ranked $\sqrt{p}$ features, instead of searching exhaustively the $\frac{1}{2}\,p\,(p-1)$ pairwise interactions among all p features, which is infeasible for very large values of p. For illustration the "Heart" and "Ionosphere" datasets are used; Fig. 5 plots the average misclassification error rates obtained by applying the centroid-based and 1-NN classifiers to the data selected by SIS, and using the suggested approach and its extension. The averages were calculated over 400 test sets obtained from random splits of the data.

From Fig. 5 it can be seen that, for the two datasets, ranking only the $\frac{1}{2}\,\sqrt{p}\,(\sqrt{p}-1)$ interactions does not necessarily degrade classification performance. On the contrary, its performance is, for the numbers of top-ranked features shown, much the same as, or even a little better than, ranking all the $\frac{1}{2}\,p\,(p-1)$ pairwise interactions, for both the centroid-based classifier and the 1-NN classifier.

4.2.2. Extension-2: using p-values for ranking

The suggested approach adopts a two- or multiple-stage strategy. There exist in the literature alternative two-stage approaches to detecting influential interactions (Marchini et al., 2005; Ionita and Man, 2006; Evans et al., 2006). They usually first identify significant single features and then detect significant interactions among the initially identified features, often based on p-values obtained from statistical tests feature by feature. Unlike the methodology suggested here, they rely on at least one feature in an interaction being significant, and so they do not lead readily to the discovery of (for example) two new features that are not individually significant but have significant impact when working together.

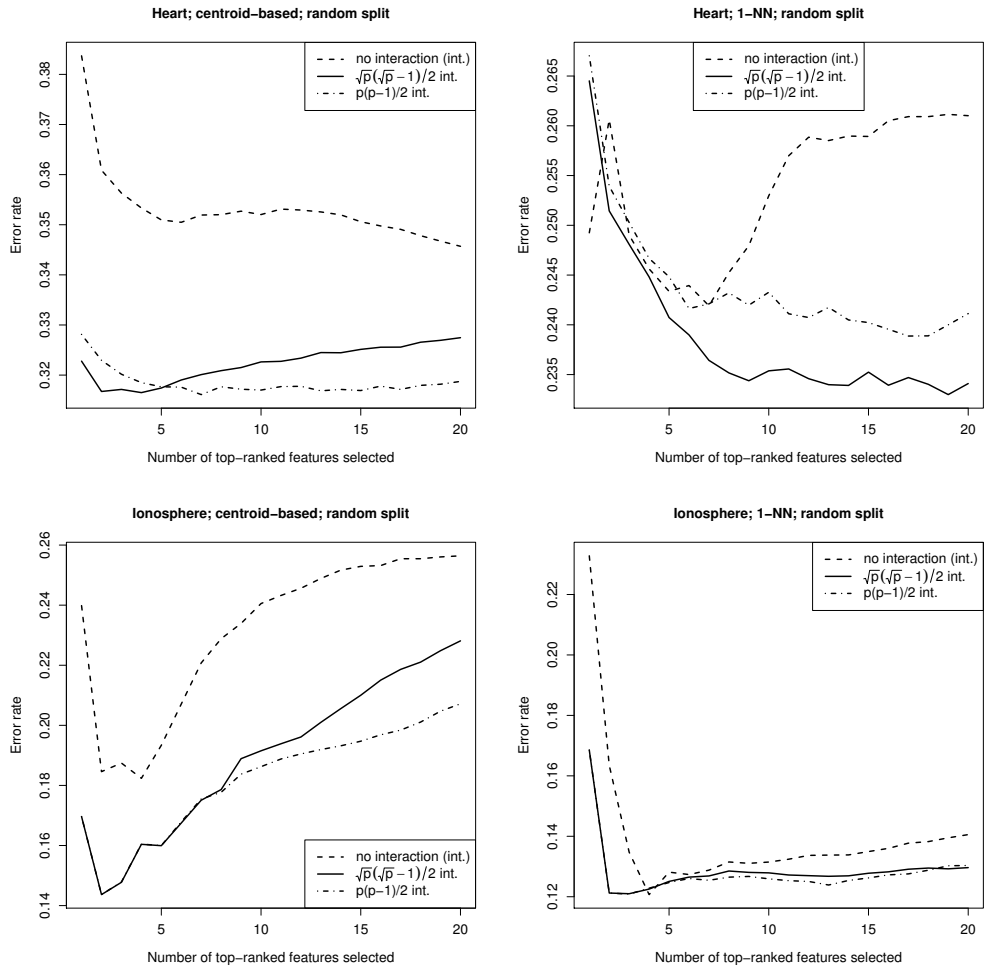

Figure 5: Comparison between methods ranking $\frac{1}{2}\,p\,(p-1)$ and $\frac{1}{2}\sqrt{p}\,(\sqrt{p}-1)$ pairwise interactions. Notation is as for Fig. 3.

Nevertheless it is desirable to investigate, in the context of classification, first, whether the suggested approach using p-values is better than using (generalised) correlation, and, secondly, whether adding interactions can also improve performance when p-values are used. Logistic regression is used to obtain, for each feature, the p-value.

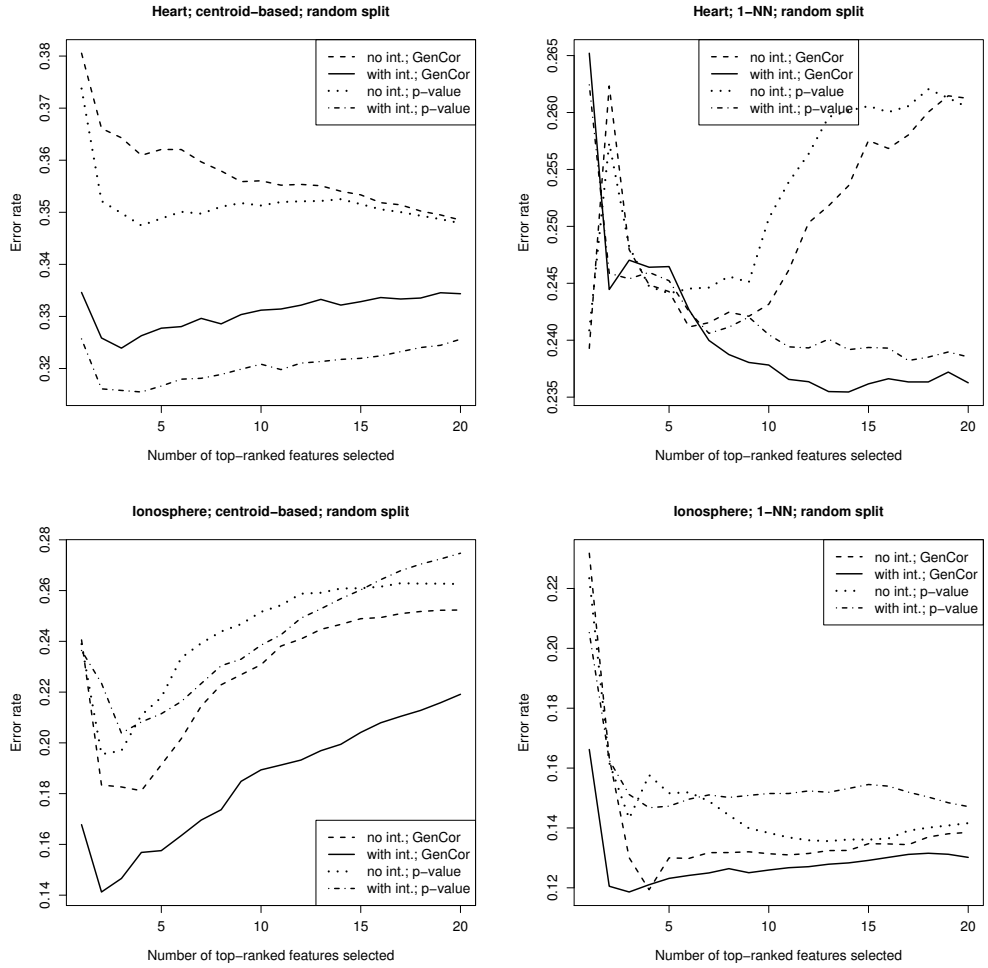

Figure 6: Comparison between methods using generalised correlation (GenCor) or the p-value for ranking. Notation is as for Fig. 3.

Fig. 6 presents the average misclassification error rates obtained by applying centroid-based and 1-NN classifiers to the features selected by four approaches: the suggested approach (indicated by solid lines), its p-value

extension (dot-dashed lines) and the corresponding methods ignoring interactions (dashed lines and dotted lines, respectively).

From Fig. 6 the following can be seen. First, for the suggested approach, which considers interactions, using generalised correlation leads to better performance than using p-values, for the "Ionosphere" data classified by both the centroid-based and 1-NN classifiers and for the "Heart" data classified by the 1-NN classifier. Secondly, as observed in Fig. 3 already, when generalised correlation is used, classification performance is improved after considering interactions for both datasets. However, when p-values are used, such an improvement from adding interactions is gained for the "Heart" dataset but not consistently for the "Ionosphere" dataset (see the improvement from the dotted lines to the dot-dashed lines).

4.3. Reliability of the suggested ranking

The reliability of a ranking is of significant practical importance. To investigate it for the suggested approach, data were simulated from two logistic regression models for classification, so that truly influential features and interactions, which should be ranked highly, would be known.

The models are given by

$$\log\frac{P(Y_i=1\,|\,X_i)}{P(Y_i=0\,|\,X_i)}=\beta_0+\sum_{j=1}^{3}\frac{4-j}{3}\,(X_{ij}+X_{i,j+3}+X_{ij}\,X_{i,j+3}) \qquad (9)$$

and

$$\log\frac{P(Y_i=1\,|\,X_i)}{P(Y_i=0\,|\,X_i)}=\beta_0+\sum_{j=1}^{3}\frac{4-j}{3}\,\left\{X_{ij}+X_{i,j+3}+\sin(X_{ij})\,e^{X_{i,j+3}}\right\}\ , \qquad (10)$$

where, for the ith data pair (X_i, Y_i), the response $Y_i \in \{0, 1\}$ follows a Bernoulli distribution with success probability $P(Y_i = 1 \,|\, X_i)$. For $1 \le j \le 3$, each truly influential feature pair $(X_{ij}, X_{i,j+3})$ contains two highly-correlated $\mathrm{N}(0, 1)$ random variables with their correlation equal to 0.85, as in the work of Hall and Miller (2009). Other single features, $X_{i7}, \ldots, X_{ip}$, all follow the standard $\mathrm{N}(0, 1)$ distribution, where $p = 1000$; and the intercept $\beta_0 = -2.5$.

To validate the reliability and authority of the ranking provided by the suggested approach, 100 random samples were simulated. Each sample consisted of n data pairs (X_i, Y_i), with each data vector X_i comprised of $p = 1000$ dimensions. In the models at (9) and (10) there are six single features,

$X_{i1}, \ldots, X_{i6}$, and three cross features, $\{X_{ij}\, X_{i,j+3}\}_{j=1}^{3}$, with decreasing influence on Y_i. An investigation was made of the ability of the suggested approach to identify these nine truly influential features in their ideal ranking: X_{i1}, X_{i4} or $X_{i1}\, X_{i4}$; X_{i2}, X_{i5} or $X_{i2}\, X_{i5}$; and X_{i3}, X_{i6} or $X_{i3}\, X_{i6}$.

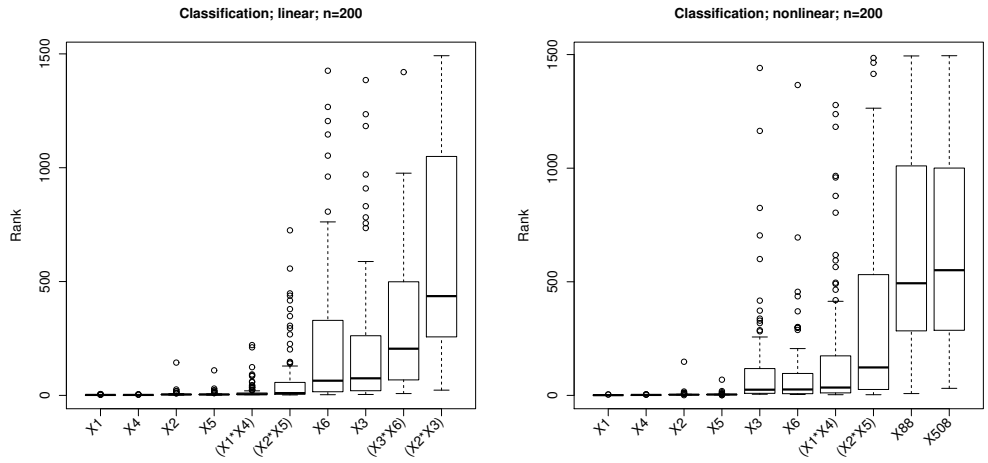

Figure 7: Reliability of rankings obtained using the suggested approach. The top 10 features, ordered by their median ranks, are shown for $n = 200$ in the cases of model (9) in the left-hand panel and model (10) in the right-hand panel.

The top 10 features with lowest median ranks are depicted in Fig. 7 for the models at (9) and (10), respectively, with $n = 200$ for each sample. To enhance the reliability of the results, attention was confined to interactions that were ranked at least one third of the time (i.e. 33 out of the 100 samples).

From Fig. 7 it can be seen that the suggested approach is powerful and reliable in identifying truly influential features, by ranking them highly and in a relatively accurate order. In particular, the top eight features, i.e. all truly influential features except for one of the weakest among them, $X_{i3}\, X_{i6}$, have remarkably lower medians and smaller spreads of their ranks than those of other features. Therefore, if the true models (9) and (10) were unknown, the top eight or nine truly influential features could still be identified, and the model recovered accurately.

4.4. How many features to be selected?

From Fig. 7 it can be seen that the number of features to be selected is eight or nine for models (9) and (10), and indeed the true models are

made up of the nine features suggested in the figure. In practice, since the true model is unknown, cross-validation can be employed to determine an "optimal" number of features (denoted by p^* below), by comparing a goal-oriented performance measure over a set of potential choices of the number. Misclassification error rate (ER) for centroid-based classification is adopted as the measure.

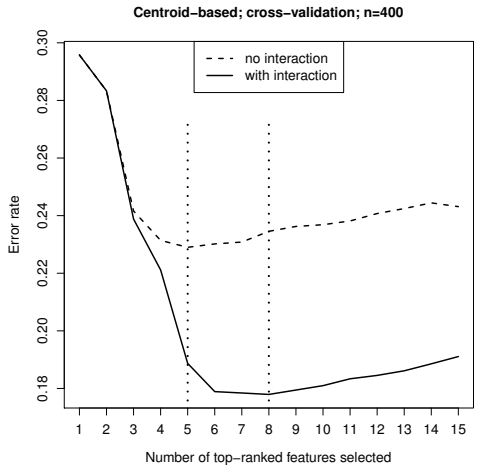

Figure 8: Determination of p^*, the "optimal" number of features to be selected. The average misclassification error rates (ER) were computed by 10-fold cross-validation from 100 random samples in the case of model (9). ER was estimated by applying centroid-based classifiers to the validation data. For each curve, the value of p^* that minimises ER is indicated by a vertical dotted line.

For the classification model (9), 10-fold cross-validation was implemented for each of 100 random samples, enabling calculation of ER for validation data. The average values of ER, over 100 10-fold cross-validation procedures, are plotted in Fig. 8. There, two vertical dotted lines indicate the values of p^* for which minimal ER is achieved, in the cases of SIS and the suggested approach, respectively.

From Fig. 8 it can be seen that, as indicated by the vertical dotted lines in Fig. 8, the suggested approach can select eight features for model (9), which is close to the true value, nine, of p^*. In fact, it appears almost equally reasonable to choose any value between six and nine, because the differences in ER for these numbers of features are small. In contrast, since SIS does not consider interactions, the (dashed) curves depicting its ER have

their minima obtained from selecting around five single features; and the ER there are noticeably higher than for the suggested approach.

Furthermore, the value of p^* determined by using cross-validated ER, as indicated in Fig. 8, matches well the p^* that can be "determined" visually based on the plots for ranking reliability in Fig. 7.

5. Limitations and discussions

Here two limitations of the suggested approach will be discussed. First, mainly because it is driven by insufficient computational resources for ultra-high-dimensional data (i.e. for very large p), the suggested approach considers a pairwise interaction only if both of its single features rank within the top $\sqrt{p}$ in Step 1. Therefore it cannot identify interactions of any features ranking lower than $\sqrt{p}$.

If computational feasible, the approach can be extended to ranking all the $\frac{1}{2}\,p\,(p-1)$ pairwise interactions. However, as shown in section 4.2.1, the extension may not improve classification performance. Moreover, after selecting top-ranked single and cross features, the extension increases the chances of keeping an interaction while discarding its constituents (i.e. main effects), which generally is not favoured in statistical modelling. Alternatively, if in certain scenarios practitioners are interested only in influential interactions of non-influential single features, the suggested approach can be extended by considering interactions of the lowest $\sqrt{p}$-ranked single features.

Secondly, as with all other feature-selection methods, improvements in classification cannot always be obtained by using the suggested approach. A reason is that, for some datasets, using many single features and interactions with higher correlation with the response Y_i may increase the risk of multi-collinearity and degrade classification performance.

However, this is generally not a serious issue in the context of classification problems. As pointed out by Guyon and Elisseeff (2003), very highly (but not perfectly) correlated features may not be redundant in some cases. Additionally, given two competing methods, for example an approach selecting and an approach ignoring interactions, cross-validation can be used to estimate the error rate of each approach for any given dataset, and the one with higher error rate can be discarded.

In practice Steps 1 and 2 are often sufficient for a noticeable improvement, and our numerical results were given only for a two-step algorithm. However, more steps can in principle be used recursively, as follows.

Step ℓ+1: Ranking features and multiple interactions. Suppose that in Step ℓ we achieved a ranking

$$\hat{\jmath}_{\ell 1} \succeq \ldots \succeq \hat{\jmath}_{\ell p_\ell} \tag{11}$$

of the components of the n p_ℓ-vectors $X_i(\ell)$. See, for example, (6) and (7). The components whose indices are ranked at (11) will be referred to below as "generalised features," and the pairwise interactions to which they lead will be said to be "generalised interactions."

In Step $\ell + 1$ we nominally replace $X_i(\ell)$ by a vector $X_i(\ell + 1)$ of length $p_{\ell+1} = p_\ell + \frac{1}{2}\, p_\ell\, (p_\ell - 1)$, consisting of all the generalised features in $X_i(\ell)$ and their generalised interactions. However, vectors as long as this will generally be too computationally challenging to handle, and so we again reduce length, principally by reducing the number, $\frac{1}{2}\, p_\ell\, (p_\ell - 1)$, of generalised interactions to $\frac{1}{2}\, \sqrt{p_\ell}\, (\sqrt{p_\ell} - 1)$. Again we adhere to the principle of keeping the most highly ranked generalised features from Step ℓ, both when they are included in isolation in $X_i(\ell)$ and when they are included in pairs as generalised interactions.

The average misclassification error rates obtained by applying to the microarray and UCI datasets a three-step version of the suggested method, which adopts generalised interactions up to order 4, are depicted in Fig. 9 by dot-dashed lines. Compared with the two-step method with single features and pairwise interactions (solid lines), the three-step method performs better for the "Hepatocellular" dataset and for the centroid-classified "Heart" data, worse for the "Prostate" dataset, and inconsistently with the number of selected top generalised features in other cases.

In order to tune the number of steps, cross-validation on the training data can be employed. Nevertheless, the increase in computation due to the involvement of high-order interactions and the tuning of the number of steps has to be taken into account in practice. Furthermore, the inclusion of higher-order interactions can lead to more correlated generalised features, and the variability of the effect of such an inclusion on classification can become greater.

Acknowledgments

This work was partly supported by an International Travel Grant to J.-H.X. from the Royal Society of London.

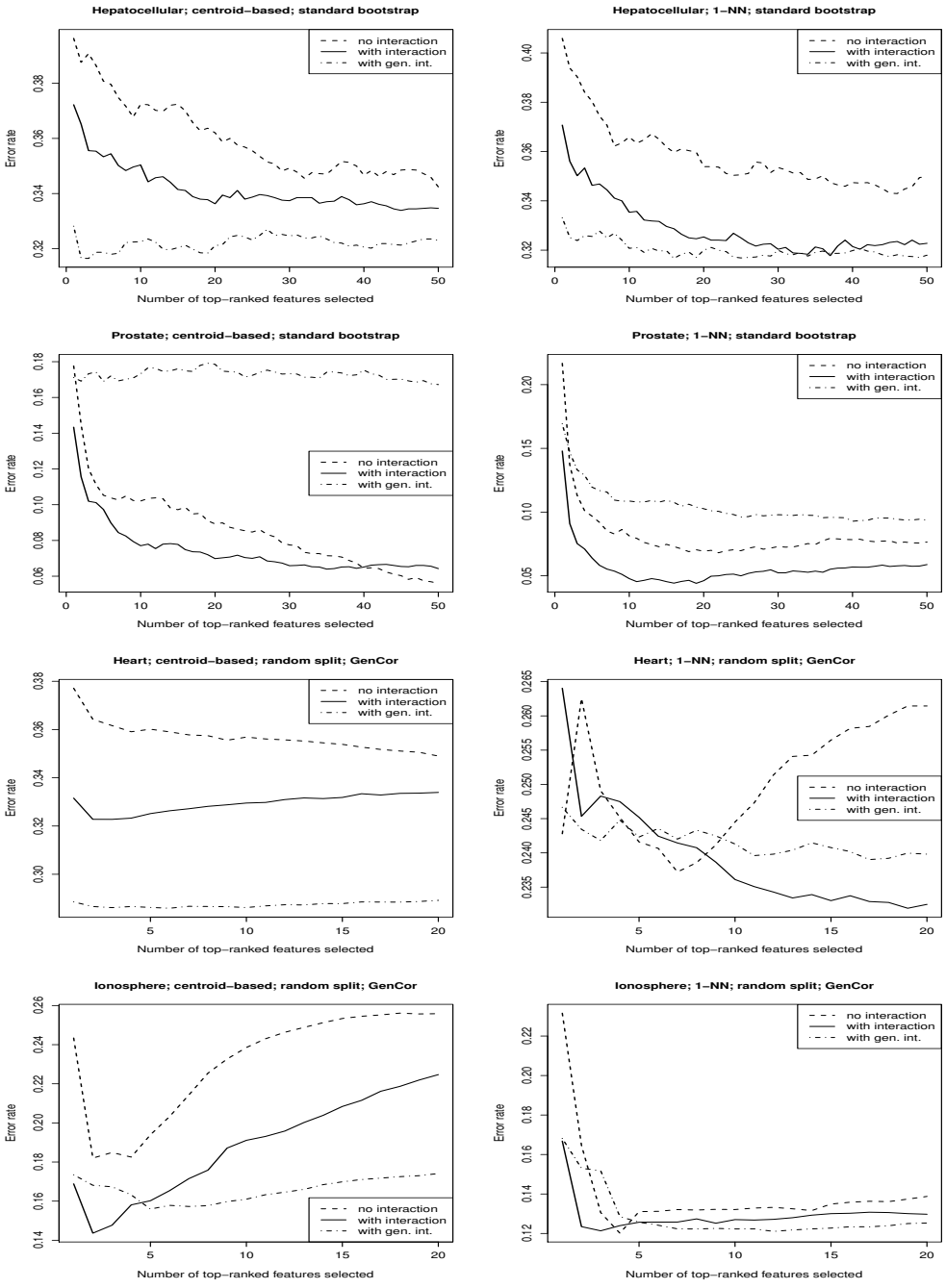

Figure 9: Misclassification error rates (ER) for the "Hepatocellular", "Prostate", "Heart" and "Ionosphere" datasets. (Dashed line: ranking single features only; solid line: recursively ranking both single features and pairwise interactions; dot-dashed line: recursively ranking single features, pairwise interactions, and generalised interactions up to order 4.) Other notation is as for Figs. 1 and 3.

Appendix A. Conditions for Theorem 1

To make the model (8) more explicit, it is assumed that:

(a) for a fixed integer r_0, such that $1 \le r_0 \le r-2$ for all n, the functions $g_1, \ldots, g_{r_0}$, γ_1 and γ_2 are non-degenerate and do not depend on n; (b) $j_1 = r_0 + 1$ and $j_2 = r_0 + 2$; (c) the functions g_j, for $r_0 + 3 \le j \le r$, depend on n and satisfy $\sup |g_j| \le C_1\, n^{-a}$ where $a, C_1 > 0$; (d) the X_{ij}s and ϵ_is are all independent; and (e) $r = O(n^a)$. (A.1)

Far from making the problem less challenging, the independence assumption in (A.1)(d) makes it more difficult. In particular, it implies that the main influence of X_{ij_1} and X_{ij_2}, which is through the term $\gamma_1(X_{ij_1})\,\gamma_2(X_{ij_2})$, is not detectable if variables are addressed one at a time. Methods based on correlation, whether standard linear correlation or its more general form suggested in section 2.1, will miss that term, as too will techniques based on fitting a linear model. Therefore the suggested methodology relies on picking up a trace of the impact of X_{ij_1} and X_{ij_2} through their impact on the additive part of (8). The condition $\sup |g_j| \le C_1\, n^{-a}$, in (A.1)(c), means that this impact is difficult to detect and so makes the feature-selection problem even more challenging. The hope is that the small impact will nevertheless be sufficient to give the feature indices j_1 and j_2 a sufficiently high rank in the first-step ranking (see (6)) for the pair of indices (j_1, j_2) to be ranked highly in the second step (see (7)).

More than one interaction could have been included on the right-hand side of (8). However, since results in that setting are straightforward extensions of the case of only a single interaction, then notational complexity is reduced by assuming that there is only a single term. Assumptions (A.1)(c) and (A.1)(e) together ensure that the series at (8) is uniformly bounded. It follows from (A.1)(a) and (A.1)(c) that the variables with greatest influence on the response, Y_i, in an additive, as distinct from interactive, sense are generally $X_{i1}, \ldots, X_{i,r_0+2}$; the other explanatory variables, $X_{i,r_0+3}, \ldots, X_{ip}$, have relatively minor individual impact.

To make the problem still more challenging the distributions of the variables X_{ij}, and the functions g_j, are allowed to vary increasingly erratically as n increases. To formalise this property, the function space $\mathcal{H}$, introduced in section 2.1, is taken to be the set of all polynomials of a given degree $d \ge 1$, although since the correlations at (1) and(2) are invariant under changes of

location and scale of $h \in \mathcal{H}$ then attention can be confined to $h \in \mathcal{H}_0$, denoting the set of $h \in \mathcal{H}$ for which $E\{h(N)^2\} = 1$, where N is distributed as normal $\mathrm{N}(0,1)$. Given an integer $K \geq 1$, $\mathcal{F}_K$ and $\mathcal{G}_K$ are defined to be classes of distributions F of a random variable V, and functions g respectively, such that $E_F|h(V)|^K$ and $E_F|g(V)|^K$ are bounded uniformly in $F \in \mathcal{F}_K$, $g \in \mathcal{G}_K$ and $h \in \mathcal{H}_0$, and, for a constant $C(K) > 0$ and for all $k = 1, \ldots, K$,

$$\frac{E_F|h(V) - E_F h(V)|^{2k}}{\{\mathrm{var}_F\, h(V)\}^k} \leq C(K)\,, \text{ for all } F \in \mathcal{F}_K,\, h \in \mathcal{H}_0\,, \tag{A.2}$$

and

$$\frac{E_F|g(V) - E_F g(V)|^{2k}}{\{\mathrm{var}_F\, g(V)\}^k} \leq C(K)\,, \text{ for all } F \in \mathcal{F}_K,\, g \in \mathcal{G}_K\,. \tag{A.3}$$

(Here E_F and var_F denote expectation and variance when V has distribution F.) For example, (A.3) holds if $\mathcal{F}_K$ is the class of all distributions that have a density bounded above C_1, say, on a given non-degenerate, compact interval $\mathcal{I}$, and if $\mathcal{G}_K = \mathcal{G}_K(C_2, C_3)$ is the set of all constant multiples of continuous functions g that satisfy $|g| \leq C_2$ and $\mathrm{var}\, g(U) \geq C_3$, where $C_1, C_2, C_3 > 0$ and U is uniformly distributed on $\mathcal{I}$.

Let V_j have the distribution of X_{ij}, interpret the notation $a_n \asymp b_n$ to mean that a_n/b_n is bounded away from zero and infinity as the positive numbers a_n and b_n diverge, and take p and p' to be functions of n. The following assumptions are made:

> (a) in Step 2 of the algorithm in section 2.1 the methodology adjoins (to the $p_1 \equiv p$ original features) all distinct pairs of the top p' features among those ranked in (6), where $p' \asymp \sqrt{p}$; and ranks the resulting $p_2 \equiv p + \frac{1}{2}\,p'\,(p'-1)$ features; (b) each distribution of V_j is in $\mathcal{F}_K$ and each function g_j and γ_j is in $\mathcal{G}_K$; (c) $\sup_{h \in \mathcal{H}_0} \mathrm{corr}\{g_j(V_j), h(V_j)\} \geq C_4$ for $j = 1, \ldots, r_0$, $\sup_{h \in \mathcal{H}_0} \mathrm{corr}\{\gamma_1(V_{j_1})\,\gamma_2(V_{j_2}), h(V_{j_1}\,V_{j_2})\} \geq C_4$, $\sup_{h \in \mathcal{H}_0} \mathrm{corr}\{g_j(V_j), h(V_j)\} \leq n^{-a_2}$ for at least $p - C_5\,p'$ values of $j \in \{1, \ldots, p\}$, and $\sup_{h \in \mathcal{H}_0} \mathrm{corr}\{g_j(V_j), h(V_j)\} \geq n^{-a_1}$ for $j = r_0+1, r_0+2$, where $a \leq a_1 < a_2 < \frac{1}{2}$, a is as in (A.3)(c), $C_4 > 0$ and $0 < C_5 < 1$; (d) $p = O(n^{C_6})$; (e) $E|\epsilon|^{C_7} < \infty$, and both C_7 and K, in the definitions of $\mathcal{F}_K$ and $\mathcal{G}_K$, are chosen sufficiently large, depending on C_6. (A.4)

Condition (A.4)(a) is appropriate if computing resources restrict the experimenter to $O(p)$ calculations; (A.4)(c) ensures that the features that are

excluded by the decision, in (A.4)(a), to confine attention to the top ranked $O(\sqrt{p})$ features will, with high probability, not exclude the key features that comprise the interaction term in (8); (A.4)(d) asks that p be no more than polynomially large as a function of n; and (A.3)(e) asserts that the error distribution has polynomially many moments. If that distribution and the distributions of the X_{ij}s were to have uniformly, exponentially light tails, for example if they all had normal distributions with bounded variances, then p could be taken exponentially large as a function of n.

Assumption (A.4)(c) is the most important part of (A.4); it ensures that, in most cases, the features with indices $r_0 + 1$ and $r_0 + 2$ endure from Step 1 to Step 2. As Theorem 1 implies, once that has happened the interaction is (with high probability) ranked very highly in Step 2, in fact among the top r_0 contributions among the p_2 features and interactions that are ranked in Step 2.

Appendix B. Proof of Theorem 1

Given $h \in \mathcal{H}_0$, define $\bar{g}_j = n^{-1} \sum_{i=1}^n g(X_{ij})$, $\bar{h}_j = n^{-1} \sum_{i=1}^n h(X_{ij})$, $\overline{\gamma\gamma} = n^{-1} \sum_{i=1}^n \gamma_1(X_{ij_1})\, \gamma_2(X_{ij_2})$ and $\bar{\epsilon} = n^{-1} \sum_i \epsilon_i$. Put $\xi_k(h) = \mathrm{cov}\{h(X_{ik}), Y_i\}$, of which an estimator is given by

$$\hat{\xi}_k(h) = \frac{1}{n} \sum_{i=1}^n \{h(X_{ik}) - \bar{h}_k\} (Y_i - \bar{Y}) = \sum_{j=1}^r S_{jk}(h) + T_k(h) + U_k(h)\,,$$

where (8) was used to obtain the second identity, and

$$S_{jk}(h) = \frac{1}{n} \sum_{i=1}^n \{h(X_{ik}) - \bar{h}_k\} \{g_j(X_{ij}) - \bar{g}_j\}\,,$$

$$T_k(h) = \frac{1}{n} \sum_{i=1}^n \{h(X_{ik}) - \bar{h}_k\} \{\gamma_1(X_{ij_1})\, \gamma_2(X_{ij_2}) - \overline{\gamma\gamma}\}\,,$$

and

$$U_k(h) = \frac{1}{n} \sum_{i=1}^n \{h(X_{ik}) - \bar{h}_k\} (\epsilon_i - \bar{\epsilon})\,.$$

The quantities $S_{jk}(h)$ and $T_k(h)$ here estimate $s_{jk}(h) = \mathrm{cov}\{h(X_{ik}), g_j(X_{ij})\}$ and $t_k(h) = \mathrm{cov}\{h(X_{ik}), \gamma_1(X_{ij_1})\, \gamma_2(X_{ij_2})\}$, respectively.

Define $\xi_{k_1k_2}(h) = \mathrm{cov}\{Y_i, h(X_{ik_1}X_{ik_2})\}$, which is estimated by

$$\hat{\xi}_{k_1k_2}(h) = \frac{1}{n}\sum_{i=1}^{n}\{h(X_{ik_1}\,X_{ik_2}) - \bar{h}_{k_1,k_2}\}\,(Y_i - \bar{Y})\,,$$

where $\bar{h}_{k_1,k_2} = n^{-1}\sum_i h(X_{ik_1}\,X_{ik_2})$. Let $\rho_{k_1k_2} = \sup_{h\in\mathcal{H}_0}\mathrm{corr}\{Y_i, h(X_{ik_1}\,X_{ik_2})\}$, of which an estimator is

$$\hat{\rho}_{k_1k_2} = \sup_{h\in\mathcal{H}_0}\frac{\hat{\xi}_{k_1k_2}(h)}{\{\hat{\zeta}_{k_1k_2}(h)\,W\}^{1/2}}\,,$$

where $\hat{\zeta}_{k_1k_2}(h)^2 = n^{-1}\sum_i\{h(X_{ik_1}\,X_{ik_2}) - \bar{h}_{k_1k_2}\}^2$ estimates $\zeta_{k_1k_2}(h)^2 = \mathrm{var}\{h(X_{ik_1}\,X_{ik_2})\}$.

The estimators $\hat{\xi}_k(h)$, $\hat{\xi}_{k_1k_2}(h)$, $S_{jk}(h)$ and $T_k(h)$ are root-n consistent, and, using (A.4)(b), (A.4)(d), and (A.4)(e), can be shown to enjoy the following properties, valid for each $\eta > 0$:

$$\max_{1\le k\le p}\sup_{h\in\mathcal{H}_0} P\big\{|\hat{\xi}_k(h) - \xi_k(h)| > n^{\eta-\frac{1}{2}}\big\} = O\big(n^{-C\eta}\big)\,, \tag{B.1}$$

$$\max_{1\le k_1<k_2\le p}\sup_{h\in\mathcal{H}_0} P\big\{|\hat{\xi}_{k_1k_2}(h) - \xi_{k_1k_2}(h)| > n^{\eta-\frac{1}{2}}\big\} = O\big(n^{-C\eta}\big)\,, \tag{B.2}$$

$$\max_{1\le j,k\le p}\sup_{h\in\mathcal{H}_0} P\big\{|S_{jk}(h) - s_{jk}(h)| > n^{\eta-\frac{1}{2}}\big\} = O\big(n^{-C\eta}\big)\,, \tag{B.3}$$

and

$$\max_{1\le k\le p}\sup_{h\in\mathcal{H}_0} P\big\{|T_k(h) - t_k(h)| > n^{\eta-\frac{1}{2}}\big\} = O\big(n^{-C\eta}\big)\,, \tag{B.4}$$

where the constant C can be made arbitrarily large by choosing K and C_7 in (A.4) sufficiently large. For any given $B_1 > 0$, a subset $\mathcal{H}_0(n)$ of $\mathcal{H}_0$, containing $O(n^{B_2})$ functions for some $B_2 > 0$, can be constructed such that, for each $h \in \mathcal{H}_0$, there exists $h_n \in \mathcal{H}_0(n)$ satisfying $|h(x) - h_n(x)| \le n^{-B_1}$ for all $x \in [-n^{B_1}, n^{B_1}]$. By approximating to $h \in \mathcal{H}_0$ by h_n, and noting that p is only polynomially large as a function of n; and choosing C sufficiently large (or equivalently, K and C_7 sufficiently large); it can be proved from (B.1)–(B.4) that the versions of those results with "max" and "sup" inside the probability statements are valid:

$$P\Big\{\max_{1\le k\le p}\sup_{h\in\mathcal{H}_0}|\hat{\xi}_k(h) - \xi_k(h)| > n^{\eta-\frac{1}{2}}\Big\} = O\big(n^{-C\eta}\big)\,, \tag{B.5}$$

$$P\Big\{\max_{1\le k_1<k_2\le p}\sup_{h\in\mathcal{H}_0}|\hat{\xi}_{k_1k_2}(h) - \xi_{k_1k_2}(h)| > n^{\eta-\frac{1}{2}}\Big\} = O\big(n^{-C\eta}\big)\,, \tag{B.6}$$

$$P\Big\{\max_{1\le j,k\le p}\sup_{h\in\mathcal{H}_0}|S_{jk}(h) - s_{jk}(h)| > n^{\eta-\frac{1}{2}}\Big\} = O\big(n^{-C\eta}\big)\,, \tag{B.7}$$

and

$$P\Big\{ \max_{1\le k\le p} \sup_{h\in\mathcal{H}_0} |T_k(h) - t_k(h)| > n^{\eta-\frac{1}{2}} \Big\} = O\big(n^{-C\eta}\big)\,. \tag{B.8}$$

In view of (A.4)(b) it can be assumed, without loss of generality, that the scales of vector components X_{ij} have been adjusted so that $v_j(h) = \mathrm{var}\{h(X_{ij})\}$ is bounded away from zero and infinity, uniformly in n, in $1 \le j \le p$ and in $h \in \mathcal{H}_0$. Then, defining

$$V_k(h) = \frac{1}{n}\sum_{i=1}^{n} \{h(X_{ik}) - \bar{h}_k\}^2\,, \quad W = \frac{1}{n}\sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

and $w = E(W)$, and using the arguments leading from (B.1)–(B.4) to (B.5)–(B.8), it can be proved that, for each $\eta > 0$,

$$P\Big\{ \max_{1\le k\le p} \sup_{h\in\mathcal{H}_0} |V_k(h) - v_k(h)| > n^{\eta-\frac{1}{2}} \Big\} = O\big(n^{-C\eta}\big)\,,$$

and

$$P\big\{|W - w| > n^{\eta-\frac{1}{2}}\big\} = O\big(n^{-C\eta}\big)\,,$$

whence it can be shown that, for constants B_3, B_4 satisfying $0 < B_3 < B_4 < \infty$,

$$P\Big\{B_3 \le V_k(h) \le B_4, \text{ for all } k = 1,\ldots,p,\ h \in \mathcal{H}_0\Big\} = 1 - O\big(n^{-C\eta}\big) \tag{B.9}$$

and

$$B_3 \le w \le B_4 \,\text{for all sufficiently large } n\,. \tag{B.10}$$

Define $W = n^{-1}\sum_i (Y_i - \bar{Y})^2$ and $w = E(W)$, and note that it can be shown from (1) and (2) that for each $\eta > 0$,

$$\hat{\rho}_j = \sup_{h\in\mathcal{H}_0} \frac{\hat{\xi}_j(h)}{\{V_j(h)\,W\}^{1/2}} = \sup_{h\in\mathcal{H}_0} \frac{S_{jj}(h) + \{\hat{\xi}_j(h) - S_{jj}(h)\}}{\{V_j(h)\,W\}^{1/2}} \tag{B.11}$$

and

$$\rho_j = \sup_{h\in\mathcal{H}_0} \frac{\xi_j(h)}{\{v_j(h)\,w\}^{1/2}} = \sup_{h\in\mathcal{H}_0} \frac{s_{jj}(h) + \{\xi_j(h) - s_{jj}(h)\}}{\{v_j(h)\,w\}^{1/2}}\,, \tag{B.12}$$

with analogous formulae holding for $\hat{\rho}_{k_1k_2}$ and $\rho_{k_1k_2}$. Combining (B.11), (B.12), their just-mentioned analogues, (A.1)(d), (A.4)(c) and (B.5)–(B.10), it can be deduced that for each $\eta > 0$,

$$P\Big\{\hat{\rho}_j \geq (1-\eta)\, C_4 \text{ for } j = 1, \ldots, r_0 - 2\Big\} \to 1\,,$$
$$P\Big\{\hat{\rho}_j \leq (1+\eta)\, n^{-a_2} \text{ for at least } p - C_5\, p' \text{ values of } j \in \{1, \ldots, p\}\Big\} \to 1\,,$$
$$P\Big\{(1-\eta)\, n^{-a_1} \leq \hat{\rho}_j \leq (1+\eta)\, n^{-a} \text{ for } j = r_0 + 1, r_0\Big\} \to 1\,,$$
$$P\big\{\hat{\rho}_{j_1 j_2} \geq (1-\eta)\, C_4\big\} \to 1\,,$$

and

$$P\Big\{\hat{\rho}_{k_1k_2} \leq n^{\eta - a}, 2 \leq k_1 < k_2 \leq p, (k_1, k_2) \neq (j_1, j_2)\Big\} \to 1\,.$$

The first three of these properties imply part (a) of Theorem 1, and that result and the next two properties above give part (b) of Theorem 1.

References

Dudoit, S., Fridlyand, J., Speed, T. P., 2002. Comparison of discrimination methods for the classification of tumors using gene expression data. Journal of the American Statistical Association 97 (457), 77–87.

Evans, D. M., Marchini, J., Morris, A. P., Cardon, L. R., 2006. Two-stage two-locus models in genome-wide association. PLoS Genetics 2 (9), e157.

Fan, J., Fan, Y., 2008. High-dimensional classification using features annealed independence rules. The Annals of Statistics 36 (6), 2605–2637.

Fan, J., Lv, J., 2008. Sure independence screening for ultra-high dimensional feature space (with discussion). Journal of the Royal Statistical Society B 70 (5), 849–911.

Fan, J., Lv, J., 2010. A selective overview of variable selection in high dimensional feature space. Statistica Sinica 20 (1), 101–148.

Fan, J., Samworth, R., Wu, Y., 2009. Ultra high dimensional feature selection: beyond the linear model. Journal of Machine Learning Research 10, 2013–2038.

Guyon, I., Elisseeff, A., 2003. An introduction to variable and feature selection. Journal of Machine Learning Research 3, 1157–1182.

Hall, M. A., 2000. Correlation-based feature selection for discrete and numeric class machine learning. In: International Conference on Machine Learning. pp. 359–366.

Hall, P., Miller, H., 2009. Using generalised correlation to effect variable selection in very high dimensional problems. Journal of Computational and Graphical Statistics 18 (3), 533–550.

Hall, P., Titterington, D. M., Xue, J.-H., 2009. Tilting methods for assessing the influence of components in a classifier. Journal of the Royal Statistical Society: Series B 71 (4), 783–803.

Hua, J., Tembe, W. D., Dougherty, E. R., 2009. Performance of feature-selection methods in the classification of high-dimension data. Pattern Recognition 42 (3), 409–424.

Iizuka, N., Oka, M., Yamada-Okabe, H., Nishida, M., Maeda, Y., Mori, N., Takao, T., Tamesa, T., Tangoku, A., Tabuchi, H., Hamada, K., Nakayama, H., Ishitsuka, H., Miyamoto, T., Hirabayashi, A., Uchimura, S., Hamamoto, Y., 2003. Oligonucleotide microarray for prediction of early intrahepatic recurrence of hepatocellular carcinoma after curative resection. Lancet 361, 923–929.

Ionita, I., Man, M., 2006. Optimal two-stage strategy for detecting interacting genes in complex diseases. BMC Genetics 7 (1), 39.

Kurgan, L. A., Cios, K. J., Tadeusiewicz, R., Ogiela, M. R., Goodenday, L. S., 2001. Knowledge discovery approach to automated cardiac SPECT diagnosis. Artificial Intelligence in Medicine 23 (2), 149–169.

Marchini, J., Donnelly, P., Cardon, L. R., 2005. Genome-wide strategies for detecting multiple loci that influence complex diseases. Nature Genetics 37 (4), 413–417.

Pochet, N., De Smet, F., Suykens, J. A. K., De Moor, B. L. R., 2004. Systematic benchmarking of microarray data classification: assessing the role of non-linearity and dimensionality reduction. Bioinformatics 20 (17), 3185–3195.

Saeys, Y., Iñza, I., Larrañaga, P., 2007. A review of feature selection techniques in bioinformatics. Bioinformatics 23 (19), 2507–2517.

Sigillito, V. G., Wing, S. P., Hutton, L. V., Baker, K. B., 1989. Classification of radar returns from the ionosphere using neural networks. Johns Hopkins APL Technical Digest 10, 262–266.

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., Sellers, W. R., 2002. Gene expression correlates of clinical prostate cancer behavior. Cancer Cell 1 (2), 203–209.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Series B 58 (1), 267–288.