Using Generalized Correlation to Effect Variable Selection in Very High Dimensional Problems

2 authors:

Peter Hall
University of Melbourne
83 PUBLICATIONS **2,946** CITATIONS

SEE PROFILE

Hugh Miller
University of Melbourne
10 PUBLICATIONS **161** CITATIONS

SEE PROFILE

Using Generalized Correlation to Effect Variable Selection in Very High Dimensional Problems

Peter HALL and Hugh MILLER

Using the traditional linear model to implement variable selection can perform very effectively in some cases, provided the response to relevant components is approximately monotone and its gradient changes only slowly. In other circumstances, nonlinearity of response can result in significant vector components being overlooked. Even if good results are obtained by linear model fitting, they can sometimes be bettered by using a nonlinear approach. These circumstances can arise in practice, with real data, and they motivate alternative methodologies. We suggest an approach based on ranking generalized empirical correlations between the response variable and components of the explanatory vector. This technique is not prediction-based, and can identify variables that are influential but not explicitly part of a predictive model. We explore the method's performance for real and simulated data, and give a theoretical argument demonstrating its validity. The method can also be used in conjunction with, rather than as an alternative to, conventional prediction-based variable selections, by providing a preliminary "massive dimension reduction" step as a prelude to using alternative techniques (e.g., the adaptive lasso) that do not always cope well with very high dimensions. Supplemental materials relating to the numerical sections of this paper are available online.

Key Words: Bootstrap; Classification; Errors in variables; Generalized correlation; Hidden explanatory variables; Instrumental variables; Linear model; Measurement error; Regression.

1. INTRODUCTION

A variety of linear model-based methods have been proposed for variable selection. In this approach it is argued that a response variable, Y_i, might be expressible as a linear form in a long p-vector, X_i, of explanatory variables, plus error, that is,

$$Y_i = \alpha + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \text{error}, \tag{1.1}$$

Peter Hall is ARC Federation Fellow, Department of Mathematics and Statistics, University of Melbourne, Melbourne, VIC 3010, Australia (E-mail: *halpstat@ms.unimelb.edu.au*). Hugh Miller is a Ph.D. Student, Department of Mathematics and Statistics, University of Melbourne, Melbourne, VIC 3010, Australia (E-mail: *H.Miller@ms.unimelb.edu.au*).

and that variable selection could be effected by choosing many of the coefficients β_j to be zero. The lasso method, proposed for this purpose by Tibshirani (1996), has proved particularly popular and effective. See also Chen and Donoho (1998), Zou (2006), Candes and Tao (2007), and Bickel, Ritov, and Tsybakov (2009), among many other contributions. Typical applications include those where X_{ij} equals the expression level of the jth gene of person i, $Y_i = 1$ if the person has a particular medical condition and $Y_i = 0$ otherwise, p is in the thousands or tens of thousands, and $1 \leq i \leq n$ where n is in the tens or hundreds.

Many, indeed the majority, of applications of the model (1.1) represent cases where the response is unlikely to be an actual linear function of X_i, for example, where Y_i is a zero-one variable but the fitted response takes values that often lie outside the unit interval. However, inconsistency of prediction does not necessarily detract from the usefulness of such methods as devices for determining the components X_{ij} that most influence the value of Y_i. For example, inconsistency is often not a significant problem if the response of Y_i to an influential component X_{ij} is qualitatively linear, in particular if it is monotone and the gradient does not change rapidly.

In other settings, however, there is a risk that fitting an incorrect linear model will cause us to overlook some important components altogether. Theoretical examples of this type are identical to those used to show, by counterexample, that the existence of conventional correlation does not equate to absence of a relationship. In Section 2, Example 1 will discuss a practical instance of this difficulty, and Example 2 there will treat another real dataset where challenges of a different nature arise. More generally, using an ill-fitting model to solve a variable-selection problem can result in reduced performance.

A little more subtly, even if the linear model is perfectly correct, fitting it can conceal components that potentially influence linearly the value of Y_i. For instance, genes whose expression levels are strongly linearly associated with Y_i, and so would be of biological interest, can be confounded or not uniquely represented. In particular, if $X_{i1} = X_{i3} + X_{i4}$ and $X_{i2} = X_{i3} + X_{i5}$, then the linear models $Y_i = X_{i1} - X_{i2} + \text{error}$ and $Y_i = X_{i4} - X_{i5} + \text{error}$, and of course infinitely many others, are equally valid. This nonidentifiability issue arises because the variable-selection problem is posed as one of model fitting, or prediction, which in our view is not necessarily a good idea. Thus, even nonlinear extensions to variable-selection methods that focus on prediction, such as the group lasso or group LARS (Yuan and Lin 2006), may still be inadequate in detecting all influential variables. Example 4 in Section 4 will explore this type of behavior in greater detail. Also, Example 3 in Section 2 explores a real dataset where this masking interferes with variable selection.

These examples, and others that we shall give, argue in favor of methods for variable selection that focus specifically on that problem, without requiring a restrictive model such as that at (1.1). In this article we suggest techniques based on ranking generalized empirical correlations between components of X and the response Y. Section 2 discusses real-data examples which motivate our approach, Section 3 introduces our methodology, and Section 4 extends the discussion in Section 2 and also presents simulation studies which explore properties of the methodology. Section 5 provides theory that demonstrates the methodology's general properties.

There is a large literature on variable-selection methods relating to the linear model at (1.1). It includes, but is by no means restricted to, work on the nonnegative garotte (e.g.,

Breiman 1995; Gao 1998), on soft thresholding (e.g., Donoho et al. 1995), and related work (e.g., Donoho and Huo 2001; Fan and Li 2001; Donoho and Elad 2003; Tropp 2005; Donoho 2006a, 2006b).

2. MOTIVATING EXAMPLES

Here we discuss three real datasets which motivate the methodology we shall introduce in Section 3.

Example 1 (Cardiomyopathy microarray data)**:**   These data were used by Segal, Dahlquist, and Conklin (2003) to evaluate regression-based approaches to microarray analysis. The aim was to determine which genes were influential for overexpression of a G protein-coupled receptor, designated Ro1, in mice. The research related to understanding types of human heart disease. The Ro1 expression level, Y_i, was measured for $n = 30$ specimens, and genetic expression levels, X_i, were obtained for $p = 6{,}319$ genes.

Our analysis will be based on ranking, over j, the maximum over h of the correlation between $h(X_{ij})$ and Y_i, where the correlation is computed from all data pairs (X_i, Y_i) for $i = 1, \ldots, n$. Here h is confined to a class $\mathcal{H}$ of functions. Taking $\mathcal{H}$ to consist entirely of linear functions gives the (absolute value of the) conventional correlation coefficient, but using a larger class enables us to explore nonlinear relationships. We shall take $\mathcal{H}$ to be a set of cubic splines. See Example 1 in Section 4 for further technical detail.

This approach leads us to rank two genes, Msa.2877.0 and Msa.1166.0, first and second, respectively. The first of these genes was identified by the linear-regression approach adopted by Segal, Dahlquist, and Conklin (2003), but the second was not. Figure 1 indicates why this is the case, by showing the scatterplots and corresponding cubic-spline fits. Whereas Msa.2877.0 shows an essentially linear relationship, which is identified by many existing techniques, Msa.1166.0 exhibits clear nonlinear behavior, where the response "flatlines" once the expression reaches a certain threshold. Another factor is the strong correlation of -0.75 between the two variables. This "masking effect" confounds standard linear modeling approaches to variable selection, and was discussed in Section 1. See also Examples 4, 5, and 6 in Section 4.

Example 2 (Acute leukemia microarray data)**:**   These data come from a study by Golub et al. (1999), where the aim was to use microarray evidence to distinguish between two types of acute leukemia (ALL/AML). There were $p = 7{,}129$ genes and $n = 38$ observations in the training data (27 ALL and 11 AML). There were also 34 observations in a separate test dataset with 20 ALL and 14 AML.

Methods based on linear correlation, of which those proposed in this article are a generalization, are analogous to minimizing the deviance of a normal model with identity link
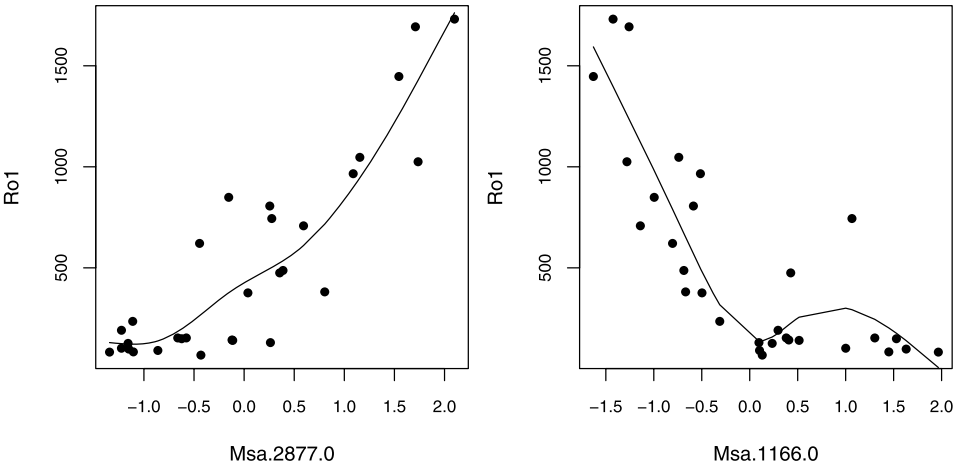
Figure 1.   Top two variables with cubic-spline fits for Example 1.

under the generalized linear model framework (McCullagh and Nelder 1989). This suggests that binary data could be treated by minimizing the deviance formula for Bernoulli data with a logistic link for each X_i, and using this to rank the values of

$$\inf_{h \in \mathcal{H}} \sum_{i=1}^{n} \left\{ -Y_i \log\left(e^{h(X_{ij})}\right) + \log\left(1 + e^{h(X_{ij})}\right) \right\}, \tag{2.1}$$

where each Y_i equals zero or 1 and $\mathcal{H}$ is a class of functions, for example, the class of polynomials of a given degree. In the analysis reported later we took $\mathcal{H}$ to be the set of all linear functions. For further detail see Example 2 in Section 4.

There is considerable overlap between the genes we found using this approach, and those discovered in other studies (Golub et al. 1999; Tibshirani et al. 2002; Hall, Titterington, and Xue 2007b; Fan and Fan 2008; Fan and Lv 2008). However, we argue that the set found in the present analysis represents an improvement over choices made by alternative methods. To address this point, a simple classifier was constructed. For the genes giving the five largest values of the quantity at (2.1), a classifier was chosen that minimized the misclassification rate on the training data, weighted so that the two classes had equal authority. These classifiers all had one decision value, above which the classification would be one class and below which it would be the other. Whichever class had the most "votes" out of the five would then be the overall predicted class. Although this was a very simple classifier, it performed perfectly on the training data and had only one misclassification on the test set. This means the classifier performed at least as well as other approaches in the literature and, in most cases, used considerably fewer genes. We again stress that our purpose was not to build a predictive model, but to identify influential variables. If the latter problem, rather than prediction, is the ultimate aim, and it generally is, then it can be advantageous to focus on it from the start.

Example 3 (Breast tumor X-ray data):   This dataset was used as the training dataset in the 2008 KDD Cup data mining competition. It consists of 102,294 observations, each

Table 1. Average number of variables detected under 5% sampling in Example 3.

	No. of top 12 effects detected
Group lasso	5.86 (0.08)
Generalized corr.	10.02 (0.08)
Random forest	9.44 (0.10)

corresponding to a potential malignant tumor spot on an X-ray. Each observation has 117 continuous variables identifying different attributes of the spot and a binary response identifying whether the spot is malignant, which is the case for 623 observations. For convenience here we disregard dependencies caused by spots resulting from the same patient. See *www.kddcup2008.com* for further details on the dataset.

This dataset actually forms a "large n, large p" problem and it is possible to build a fairly accurate classification model using the entire dataset. Suppose, however, that we had access to only 5% of the data. Then the roughly 30 positive responses would be insufficient to build a reasonable model, so detecting which variables are most important might be a more appropriate goal. With this in mind, consider the simulation experiment where we examine how effectively generalized correlation detects variables compared to a predictive method. The top 12 variables in the entire dataset were determined using a weighted random forest model (Breiman 2001). The random forest was chosen to be a reasonably "model-neutral" method for determining variable importance. We sampled 5% of the data and attempted to determine the 12 most influential variables using a given approach. Then we compared the results to the top 12 variables derived from the entire dataset and calculated the number in common. Table 1 shows the results of 100 simulations for a generalized correlation approach using (2.1) and the logistic group lasso (Meier, van de Geer, and Bühlmann 2008), each based on cubic splines with knots at the quartiles. The group lasso is a penalized regression method that allows for groups of variables, such as a collection of splines, and so is an appropriate candidate for comparison in the simulation study. The results for the top variables from a random forest applied to the sample are also included.

Generalized correlation performed better than both the random forest and group lasso models, the latter picking up less than half the variables on average. These results show that predictive methods are not necessarily the optimal way to approach the variable-selection problem, if variable selection is the ultimate aim. In this particular case it is possible to show that correlations among variables are hindering variable selection for the group lasso and random forest procedures.

3. METHODOLOGY

3.1 GENERALIZED CORRELATION

Let $\mathcal{H}$ denote a vector space of functions, which for simplicity we take to include all linear functions. By restricting $\mathcal{H}$ to just its linear elements we obtain, in (3.1) below,

the absolute values of conventional correlation coefficients, but more generally we could take $\mathcal{H}$ to be the vector space generated by any given set of functions h.

Assume that we observe independent and identically distributed pairs $(X_1, Y_1), \ldots,$ (X_n, Y_n) of p-vectors X_i and scalars Y_i. A generalized measure of correlation between Y_i and the jth component X_{ij} of X_i is given and estimated by

$$\sup_{h \in \mathcal{H}} \frac{\text{cov}\{h(X_{1j}), Y_1\}}{\sqrt{\text{var}\{h(X_{1j})\} \text{var}(Y_1)}}, \qquad \sup_{h \in \mathcal{H}} \frac{\sum_i \{h(X_{ij}) - \bar{h}_j\}(Y_i - \bar{Y})}{\sqrt{\sum_i \{h(X_{ij})^2 - \bar{h}_j^2\} \cdot \sum_i (Y_i - \bar{Y})^2}}, \qquad (3.1)$$

respectively, where $\bar{h}_j = n^{-1} \sum_i h(X_{ij})$. Because each of the factors $\text{var}(Y_1)$ and $\sum_i (Y_i - \bar{Y})^2$, in the denominators at (3.1), do not depend on j, either may be replaced by any constant without affecting our ranking-based methodology. Therefore we shall work instead with

$$\psi_j = \sup_{h \in \mathcal{H}} \frac{\text{cov}\{h(X_{1j}), Y_1\}}{\sqrt{\text{var}\{h(X_{1j})\}}}, \qquad \hat{\psi}_j = \sup_{h \in \mathcal{H}} \frac{\sum_i \{h(X_{ij}) - \bar{h}_j\}(Y_i - \bar{Y})}{\sqrt{n \sum_i \{h(X_{ij})^2 - \bar{h}_j^2\}}}. \qquad (3.2)$$

These measures of association reflect the approach suggested by Grindea and Postel-nicu (1977). However, a variety of alternative measures could be used. See, for example, Griffiths (1972), Csörgő and Hall (1982), and Schechtman and Yitzhaki (1987). At first it might appear that the challenge of computing $\hat{\psi}_j$ in (3.2), for large p, might be onerous, even by modern computing standards. However, the following theorem simplifies the problem in a wide range of cases.

Theorem 1. *Assume $\mathcal{H}$ is a finite-dimensional function space including the constant function, and that there exists $h \in \mathcal{H}$ that achieves $\hat{\psi}_j$ in the definition at (3.2). Then*

$$\operatorname*{argmin}_{h \in \mathcal{H}} \sum_{i=1}^{n} \{Y_i - h(X_{ij})\}^2 \subseteq \operatorname*{argmax}_{h \in \mathcal{H}} \hat{\psi}_j.$$

That is, the maximizer of $\hat{\psi}_j$ is the solution to the least squares problem in $\mathcal{H}$.

The proof is not difficult. For the case described in Theorem 1 (e.g., polynomials up to some degree d), the least squares problem has an explicit analytic solution. This avoids a potentially cumbersome optimization problem and allows "basis expansions" of X_{ij}. Global modeling techniques generally preclude basis expansions on the grounds that they create an even larger dimensionality problem and make it difficult to assess the influence of the underlying variables.

One implication of Theorem 1 is that the ranks of the $\hat{\psi}_j$'s are the same whether we consider $\hat{\psi}_j$ itself or the reduction in the size of squared error,

$$\hat{\varphi}_j = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 - \inf_{h \in \mathcal{H}} \sum_{i=1}^{n} \{Y_i - h(X_{ij})\}^2.$$

This is particularly useful when some of the components of X_i are categorical. In such a case the correlation (simple or generalized) cannot be easily defined, but $\hat{\varphi}_j$ can be measured by taking h to be the mean response of each category. Restricting $\mathcal{H}$ to a space of constant and linear functions recovers the ranking based on conventional correlations.

3.2 CORRELATION RANKING

We order the estimators $\hat{\psi}_j$ at (3.2) as $\hat{\psi}_{\hat{j}_1} \geq \cdots \geq \hat{\psi}_{\hat{j}_p}$, say, and take

$$\hat{j}_1 \succeq \cdots \succeq \hat{j}_p \tag{3.3}$$

to represent an empirical ranking of the component indices of X in order of their impact, expressed through a generalized coefficient of correlation. In (3.3), the notation $j \succeq j'$ means formally that $\hat{\psi}_j \geq \hat{\psi}_{j'}$, and informally that "our empirical assessment, based on correlation, suggests that the jth coefficient of X has at least as much influence on the value of Y as does the j'th coefficient." Using this criterion, the ranking $r = \hat{r}(j)$ of the jth component is defined to be the value of r for which $\hat{j}_r = j$.

The authority of the ranking at (3.3) can be assessed using bootstrap methods, as follows. For each j in the range $1 \leq j \leq p$, compute $\hat{\psi}_j^*$, being the bootstrap version of $\hat{\psi}_j$ and calculated from a resample $(X_1^*, Y_1^*), \ldots, (X_n^*, Y_n^*)$, drawn by sampling randomly, with replacement, from the original dataset $\mathcal{D} = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$. Compute the corresponding version of the ranking at (2.5), denoted by $\hat{j}_1^* \succeq \cdots \succeq \hat{j}_p^*$, and calculate too the corresponding bootstrap version, $r^*(j)$ say, of $r(j)$. Given a value α, such as 0.05, 0.10, or 0.20, compute a nominal $(1 - \alpha)$-level, two-sided, equal-tailed, percentile-method prediction interval for the ranking, that is, an interval $[\hat{r}_-(j), \hat{r}_+(j)]$ where

$$P\{r^*(j) \leq \hat{r}_-(j) \mid \mathcal{D}\} \approx P\{r^*(j) \geq \hat{r}_+(j) \mid \mathcal{D}\} \approx \frac{1}{2}\alpha.$$

We indicate approximations in these formulas because the discreteness of ranks restricts the smoothness of the bootstrap distribution.

Display these intervals as lines stacked one beside the other on the same figure, each plotted on the same scale and bearing a mark showing the respective value of $\hat{r}(j)$. Convenient orderings for the lines include the one indicated in (3.3), or the ordering in terms of increasing $\hat{r}_+(j)$. The second choice generally provides greater insight because it emphasizes variables that consistently rank strongly in the bootstrap simulations. Only lines for relatively low values of $\hat{r}$ or $\hat{r}_+$ would be depicted; see the next section for examples. If two prediction intervals (represented by the lines) failed to overlap, this would indicate empirical evidence that the more highly ranked component did indeed enjoy greater impact on Y than its competitor, at least in terms of the way we have measured impact.

An important consideration of the approach presented is determining at what level significance is drawn, that is, how to decide which variables are influential and which are not. One proposed criterion is to regard a variable as influential if $\hat{r}_+(j) < \frac{1}{2}p$. This rule assumes that the number of influential variables is considerably less than the total number p; if all components were genuinely related to Y, then the rule would reject at least half of them. There are many circumstances, such as genetic microarray data, where this assumption is reasonable. The rationale is if all the variables were independent of Y, then the rank of each would randomly fluctuate across 1 through p, with an average rank of $\frac{1}{2}p$. If the prediction interval of a variable's rank does not breach $\frac{1}{2}p$ for a given significance level, then it is unlikely to be independent of Y. However, there may be an undesirably high rate of false positives under this criterion, particularly for small n. A natural way to tune the

rule to alleviate this problem is to replace $\frac{1}{2}p$ by some smaller fraction of p. Rather than trying to predict a suitable level, it is generally easier to plot the results and allow the data to suggest a suitable level. This principle is further explored in Example 4 of the numerical work.

3.3 RANKING CONVENTIONAL CORRELATIONS

Because conventional correlation measures the strength of a linear relationship, then, in many cases, component ranking in terms of conventional correlation gives results not unlike those obtained by linear model fitting, for example using the lasso. In particular, if the linear model at (1.1) holds in a form where only a fixed number, q say, of the coefficients β_j are nonzero, and if the coefficients of correlation of Y_i with all other components of X_i are bounded away from ± 1, then under moment conditions on the components (see Section 5), the probability that the q special components appear first in a ranking of the absolute values of conventional correlation coefficients converges to 1 as n and p diverge.

However, linear methods such as the lasso can be challenged when it comes to identifying, purely empirically, the q special components. The conventional lasso can fail to correctly choose the components, even if p is kept fixed as n diverges. Component ranking, based on the absolute values of conventional correlation coefficients, can be used for an initial "massive dimension reduction" step, reducing dimension in one hit from p to a relatively low value, larger than q, from which dimension can be further reduced to q by implementation of an existing adaptively penalized form of the lasso.

Another potential advantage of ranking methods based on conventional correlation coefficients is that they overcome problems with errors in variables. For example, suppose that, in a generalization of (1.1),

$$Y_i = g(W_i) + \text{error}, \tag{3.4}$$

where g is a potentially nonlinear function, W_i denotes the p-vector of actual (but hidden) explanatory variables, and the error is independent of W_i. In errors-in-variables problems we observe only Y_i and $X_i = W_i + \delta_i$, where the p-vector δ_i is a second source of error with zero mean, independent of W_i and of the error in (3.4). There is a large literature on problems framed in this way, in cases where p is substantially smaller than n; usually, $p = 1$. This work can be accessed through the monograph by Carroll et al. (2006). The effect of the errors δ_i vanishes entirely from the correlation between X_{ij} and Y_i:

$$\text{cov}(X_{ij}, Y_i) = \text{cov}\{W_{ij} + \delta_{ij}, g(W_i)\} = \text{cov}\{W_{ij}, g(W_i)\} = \text{cov}(W_{ij}, Y_i).$$

In particular, the conventional correlation between X_{ij} and Y_i is exactly equal to the conventional correlation between W_{ij} and Y_i. Generalized correlations will not in general retain this property; if the distribution of δ_i were known, then the effect of the error could be at least partially reduced by "deconvolution," but this approach is not attractive when $p \gg n$.

Therefore, component ranking in terms of the absolute values of conventional correlations is an effective way of removing the effects of errors in variables, even if, as in (3.4), the response is a nonlinear function of the hidden explanatory variable. Example 5 in Section 4 will address problems of this type.

4. NUMERICAL PROPERTIES

Example 1 (Continuation of Example 1 from Section 2)**:** We used natural cubic splines, with three interior knots on the quartiles of the variable's observed values, because, unlike quadratic splines, such functions model both nonlinear monotone functions and multimodal functions. This gives them significant flexibility. To implement the bootstrap method described in Section 3.2 we used 400 resamples, $\alpha = 0.02$ and a $\frac{1}{4}p$ cutoff for $\hat{r}_+$. This resulted in the selection of 14 genes, of which two, the genes Msa.2877.0 and Msa.1166.0 discussed in Section 2, were particularly influential. This can be deduced from the marked jump in the length of the prediction intervals, represented by vertical lines in Figure 2, between the second and third most highly ranked genes. Examples 5 and 6, below, will summarize the results of simulation studies motivated by the findings above.

Example 2 (Continuation of Example 2 from Section 2)**:** When $\mathcal{H}$ is constrained to include linear functions of X_{ij}, as was the case in our treatment of this example in Section 2, the approach is analogous to ranking the absolute values of conventional correlation coefficients. Our bootstrap implementation used 200 resamples and $\alpha = 0.05$. All variables were standardized to have sample mean zero and sample variance 1. Figure 3 shows the influential genes using a $\frac{1}{8}p$ cutoff for $\hat{r}_+$. The first two or three genes are seen to stand out, in terms of influence, and then influence remains approximately constant until genes 9 or 10. From that point there is another noticeable drop in influence, to a point from which it tails off fairly steadily.

Example 4 (Variable masking)**:** Motivated by an example discussed in the Introduction, we look at a linear model where variables are highly correlated, and we compare the variable-selection performance of our method and the lasso (Tibshirani 1996).
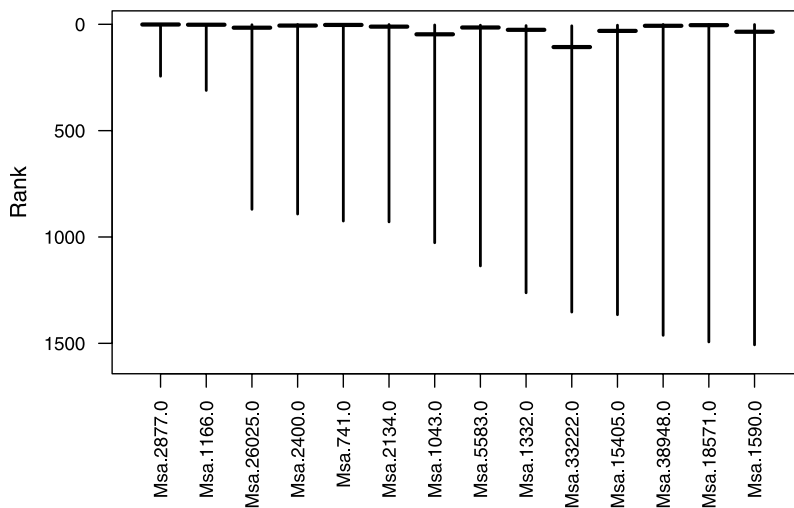

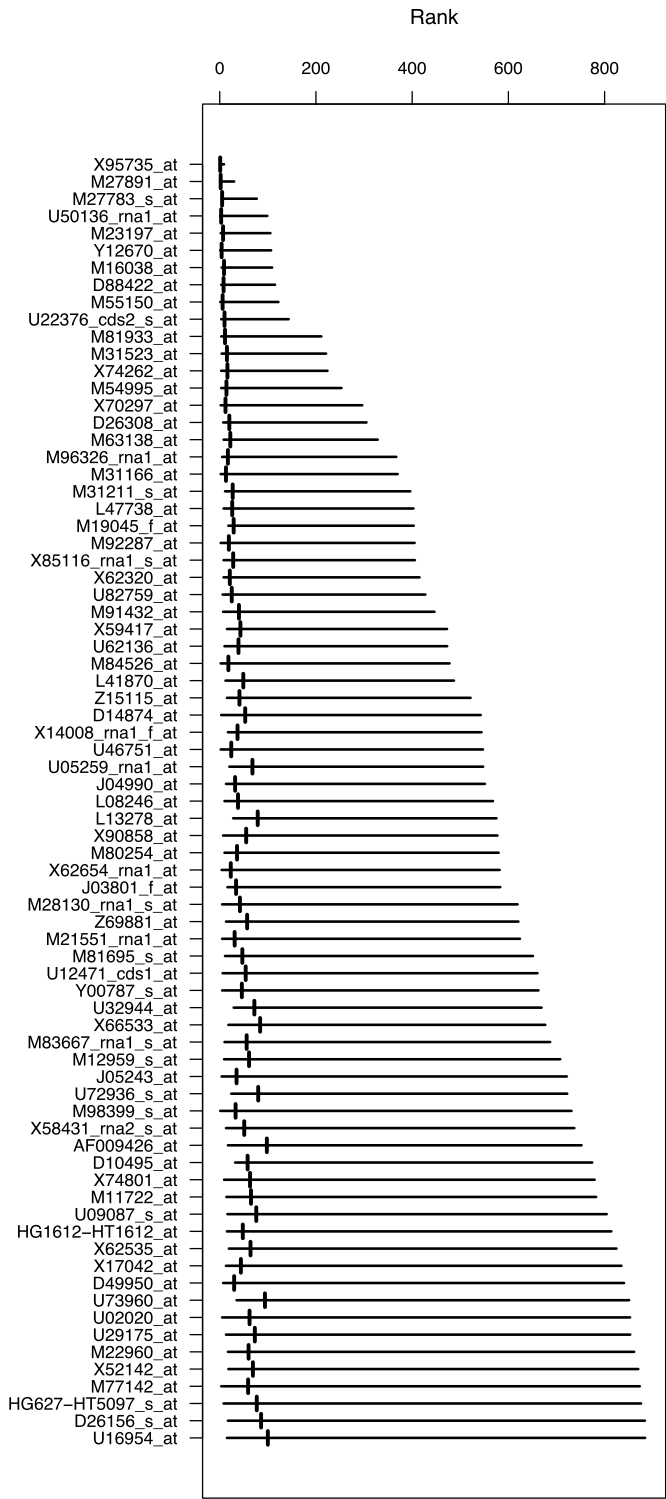Figure 2. Variables ordered by $\hat{r}_+$ for Example 1.

Figure 3.   Top 67 variables by $\hat{r}_+$ for Example 2.

First we describe the model generating the data. For $1 \leq j \leq 5$, let $\{X_{ij}, X_{i,j+5}\}$ be independent pairs of normal random variables with zero means, unit variances, and correlation equal to 0.85. Let

$$Y_i = \sum_{j=1}^{5} \frac{6-k}{5}(X_{ij} + X_{i,j+5}) + \epsilon_i,$$

where ϵ_i is a normal error with zero mean and standard deviation 5. Thus the pairs make a decreasing contribution to Y_i as j increase. Also, let X_{ij} be an independent standard normal random variable, for $11 \leq j \leq 5{,}000$. Thus, Y_i is a linear function of just the first 10 components in a vector of 5,000 $N(0, 1)$ components.

To apply the lasso we used the least angle regression (LARS) implementation (Efron et al. 2004), and to implement our method we ranked the absolute values of conventional correlation coefficients. These two approaches were compared by examining the top ten variables that each suggested. For the correlation-based approach this meant taking the ten variables with lowest $\hat{r}_+$, whereas for the lasso it involved gradually relaxing the penalization condition until just 10 variables were admitted (note that the cross-validated, lowest-error lasso model under the "one standard error rule" generally admitted fewer than ten variables). For each set we then counted how many main effects were detected (i.e., for how many $j \in [1, 5]$ did one of $\{X_{ij}, X_{i,j+5}\}$ appear in the set), as well as how many surrogate effects were detected (the number of $j \in [1, 5]$ for which both X_{ij} and $X_{i,j+5}$ were in the set). Even though the effects were linear, we have also included results for detections using generalized correlation and group LARS using cubic splines. Group LARS was used rather than the group lasso, for reasons of computational feasibility, but the two methods

Table 2. Average number of variables detected under simulation.

		No. of main effects detected	No. of surrogate effects detected
$n = 100$	lasso	1.76	0.39
	corr	1.58	1.11
	gLARS	1.06	0.49
	gcorr	0.98	0.49
$n = 200$	lasso	2.91	0.99
	corr	2.57	2.06
	gLARS	2.09	1.21
	gcorr	2.25	1.67
$n = 500$	lasso	3.98	2.45
	corr	3.54	3.28
	gLARS	3.43	1.95
	gcorr	3.23	2.94
$n = 1000$	lasso	4.32	3.27
	corr	4.10	3.87
	gLARS	3.91	2.36
	gcorr	3.93	3.66

generally show comparable performance. The experiment was repeated 100 times and the average results are presented in Table 2 for various n.

The main feature of the results is that whereas the lasso and group LARS are better at detecting weaker main effects compared to conventional and generalized correlation, respectively, they fail to select the second of each correlated pair of variables. Of course, this is a consequence of using model fitting as a surrogate for variable selection; adding a highly correlated random variable does not greatly improve predictive accuracy, but it nevertheless produces influential variables which, from most practical viewpoints, should be detected by a highly performing variable selector. Thus the results highlight the risk of using a prediction-based method as a means of detecting influential variables. Also, some loss of detection power is observed when moving to nonlinear methods, particularly for lower sample sizes. However, given that this enables the user to detect genuine nonlinear patterns should they exist, the loss appears tolerable.

Figure 4 shows typical, randomly chosen results for our bootstrapped ranking approach for various n in this simulation. For this purpose we used 100 bootstrap resamples and took $\alpha = 0.1$. Of note is the increased ability with which weaker trends are identified, and the increased stability in the ranking of genuine variables as n increases. The theoretical basis for these ideas is covered in Section 4.

As discussed in Section 3.2, there are practical considerations when choosing the level at which variables are classified as significant. Figure 5 gives the number of variables admitted when $n = 500$. It shows four variables appearing very strongly, seen in the flat section of the curve before 3% of p, and then the number of variables admitted grows
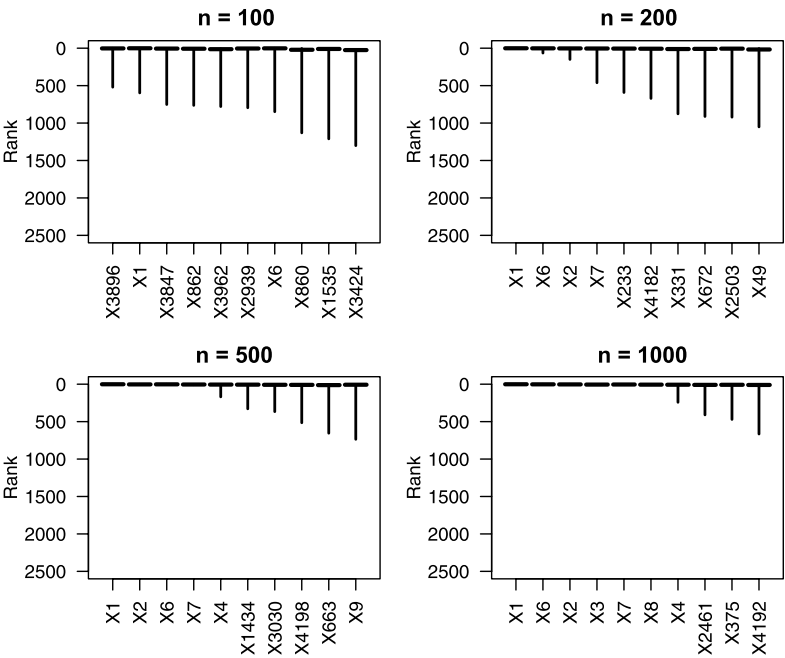

Figure 4.   Top ten variables by $\hat{r}_+$ for Example 4 with various n.
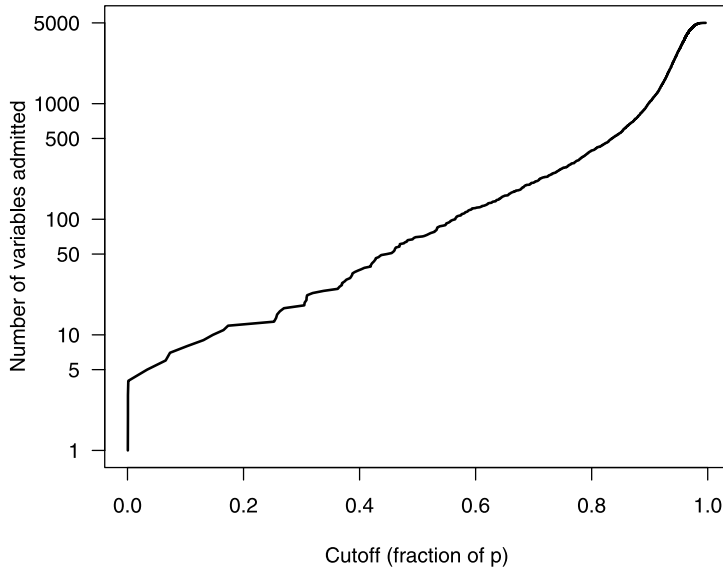
Figure 5.   Number of variables admitted at various cutoffs for Example 4 with $n = 500$.

exponentially. The proposed $p/2$ level admits only a moderate number of variables (70), but at fractions larger than this the number of variables tends to be unwieldy. Although any choice of cutoff between 3% and 50% might be considered reasonable, and would largely be driven by a user's tolerance of false positives, any presentation should highlight the relative strength of the top four variables. We emphasize that the final choice for the cutoff should ideally be based on the dataset itself.

Example 5 (A nonlinear situation)**:**   For this simulation study we took the first component to have a nonlinear impact on Y_i and to have contamination of errors-in-variables type: $X_{i1} = W_i + \delta_i$ and $Y_i = W_i^2 - 1 + \epsilon_i$. Here each W_i was taken to be uniform on $[-2, 2]$, and the two error terms, δ_i and ϵ_i, were both normal with zero mean and standard deviation $\frac{3}{4}$. Also, $X_{i2}, X_{i3}, \ldots, X_{i,5,000}$ were taken to be independent $N(0, 1)$ random variables. The simulations were run with $n = 200$, prediction bands for the ranking used $\alpha = 0.02$, and 500 bootstrap simulations were performed.

In this case, if ranking is based on conventional correlation, then X_{i1} does not appear influential, due to its nonlinear relationship with Y_i. This is true of other linear-based approaches; for instance, the lasso fails to detect X_{i1}. Thus the generalized correlation of (3.2) was used, where $\mathcal{H}$ was a basis of natural cubic splines constructed in the same way as in Example 1. As Figure 6 demonstrates, under the second criterion, X_{i1} emerges strongly as the top variable, with only three false positives if we use a cutoff at $\frac{1}{2}p$. The natural cubic-spline fit captures the relationship between X_{i1} and Y_i, although the plot in Figure 6 suggests there is some bias at the limits of X_{i1}.

Example 6 (A highly nonlinear situation)**:**   Here we report the results of simulating a model with highly nonlinear structure. Let $W_{i1}, \ldots, W_{i6}$ and $X_{i5}, \ldots, X_{i,5,000}$ be inde-
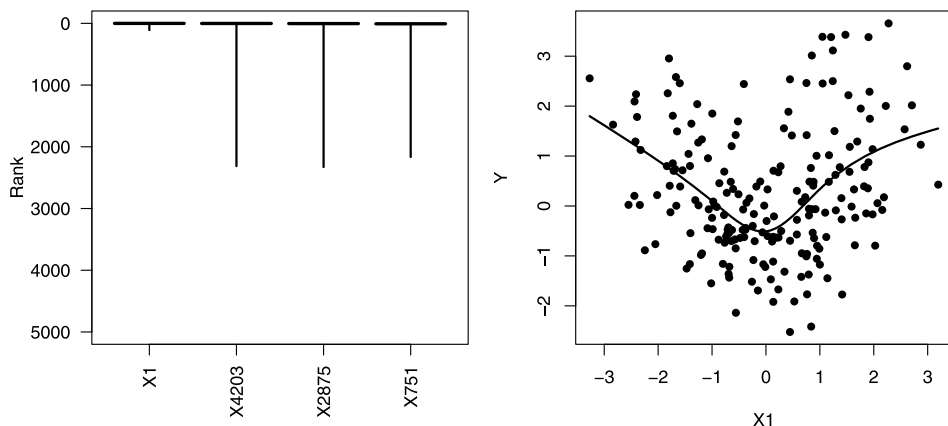
Figure 6.   Top variables by $\hat{r}_+$ for Example 5 and the cubic-spline fit for X_1.

pendent standard normal random variables, and put

$$Y_i = 2\sin\left\{\frac{\pi}{2}(W_{i1} + 0.5W_{i2})\right\} + \sum_{j=3}^{5} W_{ij}^2 + 0.4e^{W_{i6}} + Z_{i0},$$

$X_{i1} = 2W_{i1}^2 + Z_{i1}$, $X_{i2} = 2W_{i2} + Z_{i2}$, $X_{i3} = W_{i3}W_{i4} + Z_{i3}$, and $X_{i4} = W_{i6} + Z_{i4}$, with each of the Z_{ij} being normal random variables with mean zero and standard deviation 0.1. This simulation was run using natural cubic splines for $\mathcal{H}$, as in Example 5, with 500 observations, 500 bootstrap simulations, and used prediction level of $\alpha = 0.02$.

The variables with the lowest 99% percentile ranking are plotted in Figure 7. A comparison of the lengths of prediction intervals shows immediately that just two variables, X_{i3} and X_{i4}, appear influential. Two marginal false positives also have a markedly smaller degree of influence. What is interesting is that X_{i1} and X_{i2} do not appear influential; this is due to the heavy codependence of X_{i1} and X_{i2} in producing Y_i. This highlights a drawback of measuring the correlation of individual variables; sometimes the combination of several variables may be influential, whereas individually they are not. Note that if a variable $X_{i,5,001} = W_{i1} + 0.5W_{i2} + Z_{i5}$, with Z_{i5} normal with mean zero and standard deviation 0.1, were constructed, then this would present as influential in the simulation. Interestingly, the lasso (weakly) detects X_{i2} but fails to detect X_{i3}, its inconsistency resulting from the highly nonlinear behavior of the system.

5. THEORETICAL PROPERTIES

We shall state and prove a result describing the sensitivity of the rankings given by the method described in Section 2. Let $h = h_j$ denote the function for which the supremum in the definition of ψ_j, in (3.2), is achieved. We take $\mathcal{H}$ to be a class of polynomials—see assumption (5.1)(b) below—and in that case the supremum is achieved at a particular element of $\mathcal{H}$. Because our methodology is invariant under changes to the scales of Y_i
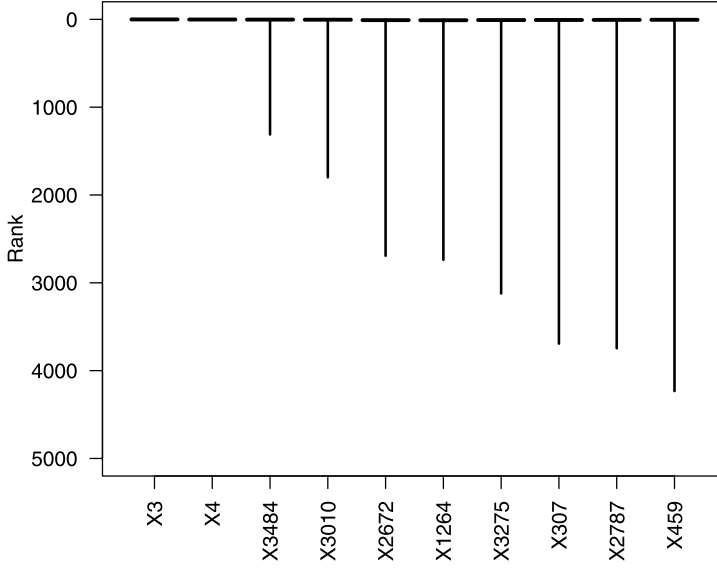
Figure 7.   Top ten variables by $\hat{r}_+$ for Example 6.

and to the components of X_i, then in formulating our assumptions below we may assume without loss of generality that $\text{var}\{h_j(X_{ij})\} = \text{var}(Y_i) = 1$ for each i and j; see (5.1)(c) below. In all other respects, except where constrained by (5.1)(e), we allow the distribution of (X_i, Y_i) to vary with n. We think of p, too, as a function of n, diverging to infinity as n increases, but diverging at no faster than a polynomial rate; see (5.1)(d). Our main other assumption is a moment condition, (5.1)(e):

(a)   the pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$ are independent and identically distributed;

(b) $\mathcal{H}$ is the class of polynomial functions of degree up to but not exceeding the positive integer $d \geq 1$;

(c) $\text{var}\{h_j(X_{ij})\} = \text{var}(Y_i) = 1$, for each i and j;   (5.1)

(d)   for a constant $\gamma > 0$ and all sufficiently large n, $p \leq \text{const.} \, n^\gamma$; and

(e)   for a constant $C > 4d(\gamma + 1)$, $\sup_n \max_{j \leq p} E|X_{1j}|^C < \infty$ and $\sup_n E|Y_1|^C < \infty$.

Given constants $0 < c_1 < c_2 < \infty$, write $\mathcal{I}_1(c_1)$ and $\mathcal{I}_2(c_2)$ for the sets of indices j for which $|\text{cov}(X_i, Y_i)| \leq c_1(n^{-1}\log n)^{1/2}$ and $|\text{cov}(X_i, Y_i)| \geq c_2(n^{-1}\log n)^{1/2}$, respectively.

Theorem 2.   *Assume* (5.1). *If, in the definitions of* $\mathcal{I}_1(c_1)$ *and* $\mathcal{I}_2(c_2)$, *the constants* c_1 *and* c_2 *are chosen sufficiently small and sufficiently large, respectively, then, in the correlation-based ranking at* (3.3), *with probability converging to* 1 *as* $n \to \infty$, *all the indices in* $\mathcal{I}_2(c_2)$ *are listed before any of the indices in* $\mathcal{I}_1(c_1)$.

Theorem 2 argues that the sensitivity point for component ranking based on correlation, or covariance, is on the scale of $(n^{-1}\log n)^{1/2}$. In particular, components for which the covariances are at least as large as sufficiently large constant multiples of $(n^{-1}\log n)^{1/2}$, are very likely to be ranked ahead of covariances which are of smaller order than this. To appreciate the clarity of the implications of this result, assume for simplicity that $\mathcal{H}$ is the set of linear functions, suppose that exactly q (a fixed number) components of X are correlated with Y, and have correlation coefficients whose absolute values are bounded above a positive constant; and that all the other components have correlations with Y which are uniformly of smaller order than $(n^{-1}\log n)^{1/2}$. For example, this would be the case if all the latter components of X were uncorrelated with Y. Then, with probability converging to 1 as p increases, all the q correlated components are listed together in the first q places of the ranking at (3.3), and all the other components are listed together in the last $p - q$ places.

APPENDIX: PROOF OF THEOREM 2

Using moderate-deviation formulas for probabilities associated with sums of independent random variables (see, e.g., Rubin and Sethuraman 1965 and Amosova 1972), it can be shown that if $b > 0$ is given, and if $\sup_n \max_{j \le p} E|X_{1j}|^C < \infty$ for some $C > 4d(b+1)$, then

$$P\{|\hat{\psi}_j - \psi_j| > c_0(n^{-1}\log n)^{1/2} \text{ for some } 1 \le j \le p\} = O(\delta),$$

where c_0 is a constant and $\delta = pn^{-b}(\log n)^{-1/2}$. Hence, with probability equal to $1 - O(\delta)$, $|\hat{\psi}_j| \le 2c_0(n^{-1}\log n)^{1/2}$ for all j such that $|\psi_j| \le c_0(n^{-1}\log n)^{1/2}$, and $|\hat{\psi}_j| > 2c_0(n^{-1}\log n)^{1/2}$ for all j for which $|\psi_j| > 3c_0(n^{-1}\log n)^{1/2}$. It follows that if, in the definitions of the sets $\mathcal{I}_1(c_1)$ and $\mathcal{I}_2(c_2)$ of indices, $c_1 \le c_0$ and $c_2 > 3c_0$, then, in the ranking at (3.3), with probability equal to $1 - O(\delta)$, all the indices in $\mathcal{I}_1$ are placed ahead of all the indices in $\mathcal{I}_2$. Provided $p \le \text{const}\, n^\gamma$ (as specified in (5.1)(c)), and $b \ge \gamma$, we have $\delta \to 0$ as $p \to \infty$.

SUPPLEMENTAL MATERIALS

Data and Computer Code: The data and computer code used in this article are available in a single archive. A read me file (read-me.rtf) is included and describes the contents of the archive in details. Note that the data used in Example 3 from the 2008 KDD cup could not be included in this archive for confidentiality reasons. (hall-miller-supplements.zip, zip archive)

REFERENCES

Amosova, N. N. (1972), "Limit Theorems for the Probabilities of Moderate Deviations," *Vestnik Leningradskogo Universiteta, No.* 13, *Matematika, Mehanika, Astronomija Vyp.*, 3 (1972), 5–14, 148.

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009), "Simultaneous Analysis of the Lasso and Dantzig Selector," *The Annals of Statistics*, to appear.

Breiman, L. (1995), "Better Subset Regression Using the Nonnegative Garrote," *Technometrics*, 37, 373–384.

——— (2001), "Random Forests," *Machine Learning*, 45, 5–32.

Candes, E., and Tao, T. (2007), "The Dantzig Selector: Statistical Estimation When p Is Much Larger Than n," *The Annals of Statistics*, 35, 2313–2351.

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006), *Measurement Error in Nonlinear Models. A Modern Perspective* (3rd ed.), Boca Raton, FL: Chapman & Hall.

Chen, S. S., and Donoho, D. L. (1998), "Atomic Decomposition by Basis Pursuit," *SIAM Journal on Scientific Computing*, 20, 33–61.

Csörgő, S., and Hall, P. (1982), "Estimable Versions of Griffiths' Measure of Association," *The Australian Journal of Statistics*, 24, 296–308.

Donoho, D. L. (2006a), "For Most Large Underdetermined Systems of Linear Equations the Minimal ℓ_1-Norm Solution Is Also the Sparsest Solution," *Communications on Pure and Applied Mathematics*, 59, 797–829.

——— (2006b), "For Most Large Underdetermined Systems of Equations, the Minimal ℓ_1-Norm Near-Solution Approximates the Sparsest Near-Solution," *Communications on Pure and Applied Mathematics*, 59, 907–934.

Donoho, D. L., and Elad, M. (2003), "Optimally Sparse Representation in General (Nonorthogonal) Dictionaries via ℓ_1 Minimization," *Proceedings of the National Academy of Sciences of the United States of America*, 100, 2197–2202.

Donoho, D. L., and Huo, X. (2001), "Uncertainty Principles and Ideal Atomic Decomposition," *Institute of Electrical and Electronics Engineers. Transactions on Information Theory*, 47, 2845–2862.

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G., and Picard, D. (1995), "Wavelet Shrinkage: Asymptopia? With Discussion and a Reply by the Authors," *Journal of the Royal Statistical Society*, Ser. B, 57, 301–369.

Efron, B., Hastie, T. J., Johnstone, I., and Tibshirani, R. J. (2004), "Least Angle Regression" (with discussion), *The Annals of Statistics*, 32, 407–499.

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

Fan, J., and Li, R. (2001), "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties," *Journal of the American Statistical Association*, 96, 1348–1360.

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

Gao, H.-Y. (1998), "Wavelet Shrinkage Denoising Using the Non-Negative Garrote," *Journal of Computational and Graphical Statistics*, 7, 469–488.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gassenbeek, 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.

Griffiths, R. C. (1972), "Linear Dependence in Bivariate Distributions," *The Australian Journal of Statistics*, 14, 182–187.

Grindea, S., and Postelnicu, V. (1977), "Some Measures of Association," in *Proceedings of the Fifth Conference on Probability Theory (Braşov, 1974)*, eds. B. Bereanu, M. Iosifescu, and G. Popescu, Bucharest: Editura Academia Republicii Socialiste România, pp. 197–203.

Hall, P., Titterington, D. M., and Xue, J. (2007b), "Tilting Methods for Assessing the Influence of Components in a Classifier," unpublished manuscript.

McCullagh, P., and Nelder, J. A. (1989), *Generalized Linear Models*, London: Chapman & Hall.

Meier, L., van de Geer, S., and Bühlmann, P. (2008), "The Group Lasso for Logistic Regression," *Journal of the Royal Statistical Society*, Ser. B, 70, 53–71.

Rubin, H., and Sethuraman, J. (1965), "Probabilities of Moderate Deviations," *Sankhyā*, Ser. A, 27, 325–346.

Schechtman, E., and Yitzhaki, S. (1987), "A Measure of Association Based on Gini's Mean Difference," *Communications in Statistics. Theory and Methods*, 16, 207–231.

Segal, M. R., Dahlquist, K. D., and Conklin, B. R. (2003), "Regression Approach for Microarray Data Analysis," *Journal of Computational Biology*, 10, 961–980.

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society*, Ser. B, 58, 267–288.

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

Tropp, J. A. (2005), "Recovery of Short, Complex Linear Combinations via ℓ_1 Minimization," *Institute of Electrical and Electronics Engineers. Transactions on Information Theory*, 51, 1568–1570.

Yuan, M., and Lin, Y. (2006), "Model Selection and Estimation in Repression With Grouped Variables," *Journal of the Royal Statistical Society*, Ser. B, 68, 49–67.

Zou, H. (2006), "The Adaptive Lasso and Its Oracle Properties," *Journal of the American Statistical Association*, 101, 1418–1429.