VARIABLE SELECTION IN VERY-HIGH DIMENSIONAL REGRESSION AND CLASSIFICATION

Peter Hall    Hugh Miller
University of Melbourne & UC Davis
A variety of linear model-based methods have been proposed for variable selection. It is argued that a response variable, $Y_i$, might be expressible as a linear form in a long $p$-vector, $X_i = (X_{i1}, \ldots, X_{ip})$, of explanatory variables, plus error:

$$Y_i = \alpha + \beta_1 X_{i1} + \ldots + \beta_p X_{ip} + \text{error}.$$ 

Variable selection can be effected by choosing many of the coefficients $\beta_j$ to be zero.

Methods include the lasso (Tibshirani, 1996; Chen and Donoho, 1998), the nonnegative garotte (Breiman, 1995; Gao, 1998), soft thresholding (e.g. Donoho et al., 1995), nonconcave penalised likelihood (Fan and Li, 2001), $L_1$ minimisation (Tropp, 2005; Donoho, 2006); the Dantzig selector (Candes and Tao, 2007); features annealed independence rules (Fan and Fan, 2007); and sure independence screening (Fan and Lv, 2008).
Typical applications include those where $X_{ij}$ equals the expression level of the $j$th 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.

In cases such as this, and probably in all applications of the linear model, the response is unlikely to be an actual linear function of $X_i$. 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.
It is known that, even when the linear model is correct, many methods based on the linear model have similar theoretical properties (Bickel, Ritov and Tsybakov, 2007). Of perhaps greater concern is their performance when the model is at best an approximation to the influence of the explanatory variable on the response.

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 counter example, that non-existence of conventional correlation does not equate to absence of a relationship.

More generally, using an ill-fitting model to solve a variable-selection problem can result in reduced performance.
Even if a linear predictive model is a reasonable approximation to the truth, there is reason to be concerned that accessing influential components via the model can result in strongly influential variables being completely overlooked. For example, if \( Y_i = X_{i1} + \text{error} \), and if \( X_{i2} \) is highly correlated with \( X_{i1} \) but does not actually appear in the model, then a model-based analysis can result in \( X_{i2} \) being overlooked.

This might be justifiable from a statistical viewpoint, but generally not from a practical one. For example, if the expression level of the gene represented by \( X_{i2} \) is strongly correlated with \( Y \) then a biologist would generally want to be informed of that relationship by the statistician doing the analysis.

In a slightly different setting, 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. Therefore, non-identifiability in a model-fitting sense can lead to difficulties with variable selection.
Example 1: Cardiomyopathy Microarray Data

These data were used by Segal et al. (2003) to evaluate regression-based approaches to microarray analysis. The aim was to determine which genes were influential for over-expression 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. In particular, in this example $Y_i$ was measured in the continuum; it wasn’t just a zero-one variable.

Our analysis will be based on ranking, over $j$, the 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. In this example we shall take $\mathcal{H}$ to be a set of cubic splines.
This approach led 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 et al. (2003), but the second was not.

The figure indicates why this is so. It shows the scatterplots and corresponding cubic-spline fits. While Msa.2877.0 displays an essentially linear relationship, which is identified by many existing techniques, Msa.1166.0 exhibits clear nonlinear behaviour, 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 modelling approaches to variable selection, and was discussed earlier.
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.

In this example the response variables \( Y_i \) are just zero or one. Methods based on linear correlation, of which those that we shall proposed are a generalisation, are analogous to minimising the deviance of a normal model with identity link under the generalised linear model framework. This suggests that binary data could be treated by minimising the deviance formula for Bernoulli data with a logistic link for each \( X_i \), and using this to rank the values of

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

where each \( Y_i \) equals zero or one and \( \mathcal{H} \) is a class of functions, for example the class of polynomials of a given degree. In the analysis reported here we took \( \mathcal{H} \) to be the set of all linear functions.
On the present occasion 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; Fan and Fan, 2007; Hall et al., 2007; Fan and Lv, 2008). However, the set found in the present analysis seems to represent 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 (1), a classifier was chosen that minimised 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, as the table shows:

<table>
<thead>
<tr>
<th>Method</th>
<th>Training errors</th>
<th>Test errors</th>
<th>No. of genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Golub et al. (1999)</td>
<td>3</td>
<td>4</td>
<td>50</td>
</tr>
<tr>
<td>Tibshirani et al. (2002)</td>
<td>1</td>
<td>2</td>
<td>21</td>
</tr>
<tr>
<td>Fan and Fan (2007)</td>
<td>1</td>
<td>1</td>
<td>11</td>
</tr>
<tr>
<td>Hall et al. (2007)</td>
<td>1–3</td>
<td>2</td>
<td>1–50</td>
</tr>
<tr>
<td>Fan and Lv (2008)</td>
<td>0–4</td>
<td>1</td>
<td>16</td>
</tr>
<tr>
<td>Correlation approach</td>
<td>0</td>
<td>1</td>
<td>5</td>
</tr>
</tbody>
</table>

Table 1: Comparison of misclassification errors on training and test samples

Note 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, then it can be advantageous to focus on that problem from the start.
Let $\mathcal{H}$ denote a vector space of functions, which for simplicity we take to include all linear functions. 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 generalised measure of correlation between $Y_i$ and the $j$th component $X_{ij}$ of $X_i$ is given and estimated by

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

respectively, where $\bar{h}_j = n^{-1} \sum_i h(X_{ij})$.

Since the factors $\text{var}(Y_1)$ and $\sum_i (Y_i - \bar{Y})^2$, in the denominators in (2), do not depend on $j$, they may be replaced by constants without affecting our ranking-based methodology. Therefore we shall work instead with

$$
\psi_j = sup_{h \in \mathcal{H}} \frac{\text{cov}\{h(X_{1j}), Y_1\}}{\sqrt{\text{var}\{h(X_{1j})\}}} \quad \text{and} \quad \hat{\psi}_j = sup_{h \in \mathcal{H}} \frac{\sum_i \{h(X_{ij}) - \bar{h}_j\} (Y_i - \bar{Y})}{\sqrt{n \sum_i \{h(X_{ij})^2 - \bar{h}_j^2\}}},
$$
By restricting $\mathcal{H}$ to just its linear elements we obtain 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 $h$.

The maximiser that yields $\hat{\psi}_j$ is just the solution to the least-squares problem in $\mathcal{H}$, and so computation is very fast.

In particular, the least-squares problem has an explicit analytic solution. This avoids a potentially cumbersome optimisation problem and allows “basis expansions” of $X_{ij}$.

Global modelling 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.
We order the estimators $\hat{\psi}_j$ as $\hat{\psi}_{j_1} \geq \ldots \geq \hat{\psi}_{j_p}$, say, and take
\[
\hat{j}_1 \succeq \ldots \succeq \hat{j}_p
\]
(3)

to represent an empirical ranking of the component indices of $X$ in order of their impact, expressed through a generalised coefficient of correlation.

In (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 $j$th 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 $j$th component is defined to be the value of $r$ for which $\hat{j}_r = j$. 
Using Bootstrap Methods to Assess the Authority of Correlation Ranking

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 $D = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$.

Compute the corresponding version of the ranking at (3), denoted by $\hat{j}_1^* \succeq \ldots \succeq \hat{j}_p^*$, and calculate too the corresponding bootstrap version, $\hat{r}^*(j)$ say, of $\hat{r}(j)$. Given a value $\alpha$, 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, i.e. an interval $[\hat{r}_-(j), \hat{r}_+(j)]$ where

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

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 $\hat{r}(j)$.

We have empirical methods for choosing $\alpha$, and they will be implemented shortly in two numerical examples.
Recall that in this example the aim was to use microarray evidence, in the form of gene expression level data on 6,319 genes, to distinguish between two types of acute leukemia.

We found that two genes, Msa.2877.0 and Msa.1166.0, had greatest correlation with the response variable. (We measured correlation via cubic splines.) The first of these genes is identified by conventional linear-regression approaches, but the second is not.

The extent to which the data support this conclusion can be deduced from the marked jump in the length of the prediction intervals, represented by vertical lines in the figure, between the second and third most highly ranked genes.
Recall that in this example the set $\mathcal{H}$ was constrained to include only linear functions of $X_{ij}$. Therefore the approach was analogous to ranking the absolute values of conventional correlation coefficients.

Our bootstrap plot of influential genes shows that the first two or three genes stand out, 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.
It can be proved theoretically that the sensitivity point for component ranking based on generalised correlation, or generalised covariance, is on the scale of \((n^{-1} \log n)^{1/2}\). In particular, components for which the generalised 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 polynomials of given degree; 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 sufficiently large positive constant multiple of \((n^{-1} \log n)^{1/2}\); let \(p \leq C n^\gamma\), for constants \(C, \gamma > 0\); and assume 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 \(q\) relatively highly correlated components are listed together in the first \(q\) places of the component ranking, and all the other components are listed together in the last \(p - q\) places.