A note on marginal correlation based screening

Somak Dutta and Vivekananda Roy
Department of Statistics, Iowa State University, USA

Abstract

Independence screening methods such as the two sample $t$-test and the marginal correlation based ranking are among the most widely used techniques for variable selection in ultrahigh dimensional data sets. In this sort note, simple examples are used to demonstrate potential problems with the independence screening methods in the presence of correlated predictors.

1 Introduction

Modern scientific research in diverse fields such as engineering, finance, genetics and neuroimaging involve data sets with hundreds of thousands of variables. For example, a typical problem in genetics is to study the association between a phenotype, say, resistance to a particular disease or yield of plants, and genotype involving millions of SNPs. Nevertheless, only a few of these variables are believed to be important. Thus, variable selection plays a crucial role in the modern scientific discoveries.

A variety of methods using different penalizations have been proposed for variable selection in the linear models, such as the Lasso (Tibshirani, 1996), SCAD (Fan and Li, 2001), elastic net (Zou and Hastie, 2005), adaptive Lasso (Zou, 2006) and others. These methods are very useful unless the number of predictors is much larger than the sample size (Fan, Samworth and Wu, 2009; Wang, 2009). In the ultra-high dimensional set up, generally variable screening is performed to reduce the number of variables before applying any of the aforementioned variable selection methods for choosing important variables and parameter estimation. One widely used variable screening procedure is based on marginal correlation or two-sample $t$ test (Fan and Lv, 2008; Li, Zhong and Zhu, 2012). According the review article of Saeys, Inza and Larranaga (2007), “the two sample $t$-test and ANOVA

Key words and phrases. correlation, feature selection; screening; sure independence; two-sample $t$ test
are among the most widely used techniques in microarray studies” for feature selection. In particular, one popular method is to decide which variables should remain in the model based on ranking of marginal Pearson correlations (Fan and Lv, 2008). As mentioned in Fan and Lv (2008), (see also Cho and Fryzlewicz, 2012; Fan et al., 2009) there can be several potential issues with this method, although they are shown to have sure screening property (that is, with probability tending to one, important variables survive the screening) under certain conditions.

In this short note, through examples, we illustrate the problems with marginal correlation based screening in the presence of correlated predictors. In particular, assuming either autoregressive (order 1) or equicorrelation covariance structure for normally distributed predictor variables, it is shown that, with high probability, important variables do not survive such screening for linear regression models. Since the examples considered here are fairly simple, we hope that the article can serve the purpose of providing warning against the use of independence screening such as the two-sample $t$-test, and the marginal correlation ranking, without further investigation.

2 Examples

Suppose the random vector $\mathbf{x} = (x_1, \ldots, x_p)^T$ has a multivariate normal distribution with zero mean vector and covariance matrix $\mathbf{R}$. Given $\mathbf{x}$ and a random variable $\epsilon \sim N(0, \sigma^2)$ independent of $\mathbf{x}$, assume that

$$y = \beta_0 + \beta^T \mathbf{x} = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon,$$

where $\beta \equiv (\beta_1, \ldots, \beta_p) \in \mathbb{R}^p$ and all but a few $\beta$’s are zero.

In the following examples we shall show that given a covariance matrix $\mathbf{R}$, the nonzero $\beta$’s can take values that would make some of the corresponding $x_i$’s marginally uncorrelated with the response. We consider here two popular correlation structures. The first correlation structure is autoregressive where correlation between $x_i$ and $x_j$ is $\rho^{|i-j|}$ for some $\rho$ between 0 and 1. The second correlation structure is compound symmetric where the correlation between $x_i$ and $x_j$ is $\rho \in (0, 1)$ whenever $i \neq j$.

The general setup of the simulations is as follows. We consider a particular multivariate Gaussian distribution for the random vector $\mathbf{x}$ and fix which of the $x_i$’s would have nonzero coefficients. Then for given value of the correlation parameter, we choose the values of the nonzero coefficients in such a way that some of those $x_i$’s become marginally uncorrelated with the response $y$. Then we simulate $n$ independent realizations from the resulting joint model and compute all the sample correlations between $y$ and $x_i$, denoted as $w_i, 1 \leq i \leq p$. As mentioned
in the introduction, a popular method of screening is to retain those features with largest absolute marginal correlations, that is, variables with \( \lceil n / \log(n) \rceil \) or \( \lceil n^{1-\theta} \rceil \) first largest \( |w_i| \) values for some \( 0 < \theta < 1 \) (Fan and Lv, 2008; Fan et al., 2009). In our simulation examples, we say a variable does not survive screening if its absolute marginal correlation is not among the largest \( n \) of all. Based on these \( w_i \)'s, we observe if a particular subset of important variables survive screening or not. We repeat this process 100 times and report the proportion of times the important variables fail to survive screening.

**Example 1**  In this example, we consider the autoregressive correlation design and show using a simple example that an important variable may fail to survive screening based on its marginal correlation with the response. To this end, suppose the \((i, j)\)th entry of \( R \) is \( r_{ij} = \rho^{|i-j|} \) where 0 < \( \rho < 1 \). Note that the largest eigenvalue of \( R \) is bounded. Next, suppose \( \beta \in \mathbb{R}^p \) be such that \( \beta_j = 0 \) if \( j \notin \{1, 3\} \) and

\[
\begin{pmatrix}
1 & \rho^2 \\
\rho & \rho
\end{pmatrix}
\begin{pmatrix}
\beta_1 \\
\beta_3
\end{pmatrix}
= \begin{pmatrix} 0 \\
a \end{pmatrix},
\]

where \( a \neq 0 \). Then, \( \text{Cov}(y, x_1) = 0 \) even though \( \beta_1 \neq 0 \) and \( \text{Cov}(y, x_2) = a \neq 0 \) even if \( \beta_2 = 0 \). For a concrete example, suppose \( \rho = 1/4 \) and \( a = 3 \), and \( \sigma = 1 \). Then the solutions are \( \beta_1 = -0.8 \) and \( \beta_3 = 12.8 \). For a given value of the sample size \( n \), we set \( p = 2n \) and generate data from the model (1) under the given setup. We repeat this process 100 times and obtain the proportion of times \( x_1 \) failed to survive the screening. In Table 1 we report these proportions for increasing values of \( n \). Clearly as \( n \to \infty \), variable \( x_1 \) does not survive screening based on marginal correlation with non-negligible probability.

In this example, the marginal correlation between \( y \) and \( x_1 \) is exactly zero.

In the next example, we consider the equicorrelation matrix and demonstrate that the important variables fail to survive screening even if the marginal correlations between \( y \) and each of the important variables are bounded away from zero.

**Example 2:**  Suppose \( R \) is the equicorrelation matrix with correlation parameter \( \rho \). That is, the \((i, j)\)th element of \( R \) is \( \rho \) if \( i \neq j \) and 1 if \( i = j \). Note that, the largest eigenvalue of \( R \) is \( 1 + (p-1)\rho \) which is \( O(n) \) if \( p = O(n) \). Without loss,
suppose $\beta_i \neq 0$, for $i \leq 5$ and $\beta_i = 0$ for $i > 5$. The covariance between $y$ and $x_i$ is

$$\text{Cov}(y, x_i) = \begin{cases} 
\beta_i + \rho \sum_{j \neq i} \beta_j & \text{if } \beta_i \neq 0 \\
\rho \sum_{j=1}^{5} \beta_j & \text{if } \beta_i = 0.
\end{cases}$$

Then if we choose $\beta_1, \ldots, \beta_5$ by solving the following system of linear equations

$$\begin{pmatrix}
1 & \rho & \rho & \rho & \rho \\
\rho & 1 & \rho & \rho & \rho \\
\rho & \rho & 1 & \rho & \rho \\
\rho & \rho & \rho & 1 & \rho \\
\rho & \rho & \rho & \rho & 1
\end{pmatrix}
\begin{pmatrix}
\beta_1 \\
\beta_2 \\
\beta_3 \\
\beta_4 \\
\beta_5
\end{pmatrix} =
\begin{pmatrix}
1 \\
1 \\
1 \\
4
\end{pmatrix},$$

then we have

$$\text{Cov}(y, x_i) = \begin{cases} 
1 & \text{if } i \leq 4 \\
4 & \text{if } i > 5.
\end{cases}$$

Thus, although $\beta_i \neq 0$ ($i \leq 4$), the marginal correlations between $y$ and each $x_i$ ($i \leq 4$) are all equal but uniformly smaller in magnitude than $\text{Cor}(y, x_j), j > 5$. Accordingly, based on a sample of size $n$ (where $n << p$), the sample marginal correlation coefficients between $y$ and each of $x_1, \ldots, x_4$ will be much smaller in magnitude than the same between $y$ and each of the unimportant variables. Accordingly, with probability tending to one, $x_1, \ldots, x_4$ will fail to survive screening.

The following simulation study confirms the conclusion. We set $p = 2n$, $\rho = 0.10$ and $\sigma^2 = 1$. We consider $n = 100, 500$ and 1000. In Table 2 (a) we report the proportion of those cases where a particular $x_i$ ($i \leq 4$) failed to survive screening.

This example is high-dimensional because $p = 2n$, but $p$ increases linearly with $n$. In Table 2 (b) we consider the above setup, except with $p = n^2$. We see that in this case, the probabilities of not surviving the screening increases to one much faster. In ultra-high-dimensional cases this problem is further exacerbated. However, in ultra-high dimensional problems, the largest eigenvalue of the covariance matrix $R$ increases at a larger rate than it does in this problem.

**References**

CHO, H. and FRYZLEWICZ, P. (2012). High dimensional variable selection via tilting. *Journal of the Royal Statistical Society, Series B*, 74 593–622.

FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle property. *Journal of the American Statistical Association*, 96 1348–1360.
Table 2: Proportion of times $x_1, \ldots, x_4$ failed to survive screening in Example 2. Left: $p = 2n$, Right: $p = n^2$.

(a) $p = 2n$  

| $n$ | $x_1$ | $x_2$ | $x_3$ | $x_4$ |
|-----|-------|-------|-------|-------|
| 100 | 0.68  | 0.61  | 0.59  | 0.54  |
| 500 | 0.83  | 0.86  | 0.90  | 0.92  |
| 1000| 0.96  | 0.96  | 0.96  | 0.97  |

(b) $p = n^2$  

| $n$ | $x_1$ | $x_2$ | $x_3$ | $x_4$ |
|-----|-------|-------|-------|-------|
| 25  | 0.95  | 0.98  | 0.96  | 0.98  |
| 50  | 0.98  | 0.96  | 0.97  | 0.98  |
| 100 | 1.00  | 1.00  | 1.00  | 0.97  |

FAN, J. and LV, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society*, Series B, 70 849–911.

FAN, J., SAMWORTH, R. and WU, Y. (2009). Ultrahigh dimensional feature selection: beyond the linear model. *Journal of Machine Learning Research*, 10 2013–2038.

LI, R., ZHONG, W. and ZHU, L. (2012). Feature screening via distance correlation learning. *Journal of the American Statistical Association*, 107 1129–1139.

SAEYS, Y., INZA, I. and LARRAÑAGA, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23 2507–2517.

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

WANG, H. (2009). Forward regression for ultra-high dimensional variable screening. *Journal of the American Statistical Association*, 104 1512–1524.

ZOU, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101 1418–1429.

ZOU, H. and HASTIE, T. (2005). Regularization and variable selection via the Elastic Net. *Journal of the Royal Statistical Society*, Series B, 67 301–320.