ICCM 2007 · Vol. II · 735–747

Variable Screening in High-dimensional Feature Space

Jianqing Fan*

Abstract

Variable selection in high-dimensional space characterizes many contemporary problems in scientific discovery and decision making. Fan and Lv [8] introduced the concept of sure screening to reduce the dimensionality. This article first reviews the part of their ideas and results and then extends them to the likelihood based models. The techniques are then applied to disease classifications in computational biology and portfolio selection in finance.

2000 Mathematics Subject Classification: 68Q32, 62J99.
Keywords and Phrases: Feature screening, Variable selection, Portfolio allocation, Classification.

1 Introduction

Exponential increases in computing power and falling costs have had a profound impact on the development of mathematical science. Modern technologies allow scientists to collect data in unprecedent size and complexity. Examples include microarrays, proteomics, brain images, videos, functional data, longitudinal data, high-frequency financial data, warehouse sales, among others. Quantitative methods have been widely employed in different scientific and engineering disciplines, which empower knowledge discovery and policy making. The field of statistics and applied mathematics have experienced extraordinary growth in the last two decades. Many data-analytic techniques have been developed and many new phenomena have been unveiled. They have become indispensable tools in contemporary scientific research, technological invention, knowledge discovery, and policy making.

The frontiers of scientific research have also posed many challenges to the development of mathematical science. Many new techniques are needed in order to confront the complex problems from various scientific disciplines. Many new theories have to be developed to understand the properties of procedures in use,

*Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08540. E-mail: jqfan@princeton.edu. Financial support from the NSF grants DMS-0704337, DMS-0714554 and NIH grant R01-GM072611 is gratefully acknowledged.

to introduce new and more powerful techniques that push all theory, methods and applications forward.

Many high-dimensional statistical learning problems can be abstracted as follows. We have one or more output variables Y and their associated features or covariates $X_1, \cdots, X_p$. We would like to model the relation between Y and $X_1, \cdots, X_p$ based on a sample of size n. Unlike traditional statistics, the dimensionality p is large and is thought mathematically as a function n, diverging to infinite. The dimensionality can even be much larger than sample size. For example, in disease classification using microarray gene expression data (Tibshirani *et al.*[19], Fan and Ren [10]), the number of arrays is usually in the order of tens while the number of gene expression profiles is in the order of tens of thousands; in the study of protein-protein interactions, the number of features can be in the order of millions, but sample size can be in the order of thousands. In portfolio allocation in finance, the number of investable stocks can be in the order of thousands, though the number of trading days that are most informative is in the order of hundreds. Although this problem does not have a response variable, we will reduce it to the familiar statistical multiple regression problem.

In the high dimensional statistical endeavor, two important issues emerge: Computational speed and statistical accuracy. For example, in selecting stocks among 2000 investible ones in the emerging market, there are 2^{2000} possible subsets. Selection decisions depend on the returns of these 2000 stocks and their associated risks as well as their correlation matrix, involving millions of parameters. Since each parameters are estimated with some errors, accumulation of these millions of estimation errors can have adverse effects on the performance of selected portfolios.

In high-dimensional scientific discovery, it is very reasonable to assume that the relevant features lie in low-dimensional manifolds. One way to avoid the aforementioned "curse-of-dimensionality" is to introduce the sure screening strategy, as initiated in Fan and Lv [8]. The basic idea is to apply a fast but crude method to screen variables that have weak correlation with the response variable, and then to show that the screened variables are indeed irrelevant, with probability tending to one. After that, applying relatively low-dimensional techniques to select further the features and to estimate relevant parameters. This reduces simultaneously the dimensionality and ensures the estimation accuracy.

2 Sure independent screening

This sections give a review of the techniques and results from Fan and Lv [8]. This enables us to develop further the Sure Independent Screening (SIS) and iterative SIS (ISIS) for more general statistical models.

Let $\mathbf{y} = (Y_1, \cdots, Y_n)^T$ be an n-vector of responses and $\mathbf{x}_1, \cdots, \mathbf{x}_n$ be their associated covariate vectors so that the data collected from the i^{th} individual is $(\mathbf{x}_i^T, Y_i)^T$, which are assumed as an i.i.d. realization from the population $(X_1, \cdots, X_p, Y)^T$. In the theoretical study of high-dimensional problem, it is helpful to think that p grows with sample size n so that we will write p as p_n

whenever the stress of dependence on n is needed in the theoretical study. It is also helpful to differentiate two cases: $p_n = O(n^\xi)$ and $\log p_n = O(n^\xi)$ for some $\xi > 0$. Whenever such distinctions are needed, the former case is referred to as "high-dimensional" and the latter as "ultra high-dimensional".

Consider the high or untra high dimensional regression model

$$Y = \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon = \mathbf{x}^T \boldsymbol{\beta} + \varepsilon, \tag{2.1}$$

where ε is a random noise and $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)^T$ is a p-vector of parameters and $\mathbf{x} = (X_1, \cdots, X_p)^T$. Putting the above model in the matrix form for the sample, we have

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2.2}$$

where $\boldsymbol{\varepsilon} = (\varepsilon_1, \cdots, \varepsilon_n)^T$ and $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_n)^T$, an $n \times p$ design matrix.

2.1 SIS and thresholded ridge regression

Let $\mathcal{M}_* = \{1 \leqslant i \leqslant p : \beta_i \neq 0\}$ be the true sparse model with nonsparsity number $s = |\mathcal{M}_*|$. The other $p - s$ variables can also be correlated with the response variable via their links to the predictors contained in the model. Let $\boldsymbol{\omega} = (\omega_1, \cdots, \omega_p)^T$ be a p-vector obtained by the componentwise regression, that is,

$$\boldsymbol{\omega} = \mathbf{X}^T \mathbf{y}, \tag{2.3}$$

where, with slight abuse of notation, the $n \times p$ data matrix $\mathbf{X}$ is first standardized columnwise.

For any given $\gamma \in (0, 1)$, define a submodel

$$\mathcal{M}_\gamma = \{1 \leqslant i \leqslant p : |\omega_i| \text{ is among the first } [\gamma n] \text{ largest of all}\}. \tag{2.4}$$

In other words, we select $[\gamma n]$ variables that have the largest magnitude of correlation with the response variable, without even consulting the contributions of other covariates to the regression. It turns out that such a simple procedure does have the sure screening property in the following sense:

$$P(\mathcal{M}_* \subset \mathcal{M}_\gamma) \to 1 \quad \text{as } n \to \infty \tag{2.5}$$

for some given γ. More precisely, we have

Theorem 1. *Suppose that*

(a) $p > n$ *and* $\log p = O(n^\xi)$ *for some* $\xi > 0$

(b) $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ *for some* $\sigma > 0$ *and* $\boldsymbol{\Sigma}^{-1/2}\mathbf{x}$ *has a spherically symmetric distribution satisfying the concentration property, where* $\boldsymbol{\Sigma}$ *is the covariance matrix of* $\mathbf{x}$*;*

(c) $\operatorname{var}(Y) < \infty$ *and for some* $\kappa \geqslant 0$ *and* $c > 0$*,*

$$\min_{i \in \mathcal{M}_*} |\beta_i| \geqslant \frac{c}{n^\kappa} \quad \text{and} \quad \min_{i \in \mathcal{M}_*} |\operatorname{cov}(\beta_i^{-1} Y, X_i)| \geqslant c;$$

(d) *there exist some* $\tau \geqslant 0$ *and* $c^* > 0$ *such that*

$$\lambda_{max}(\boldsymbol{\Sigma}) \leqslant c^* n^{\tau}.$$

If $2\kappa + \tau < 1$, *then there exists some* $\theta < 1 - 2\kappa - \tau$ *such that when* $\gamma \sim cn^{-\theta}$ *with* $c > 0$, *we have for some* $C > 0$,

$$P(\mathcal{M}_* \subset \mathcal{M}_\gamma) = 1 - O(\exp(-Cn^{1-2\kappa}/\log n)).$$

The theorem implies that the non-sparsity number $s \leqslant [\gamma n]$. It demonstrates that SIS can reduce dimension dimensionality p from exponentially growing down to a relatively large scale $d = [\gamma n] = O(n^{1-\theta}) < n$ for some $\theta > 0$ and the reduced model $\mathcal{M}_\gamma$ still contains all the variables in the true model with an overwhelming probability. In particular, we can choose the submodel size $d = n - 1$ or $n/\log n$, to be conservative. Although SIS is proposed to reduce dimensionality from p to d that is below sample size n, nothing stops us from applying it with final model size $d \geqslant n$. It is obvious that larger d means larger probability to include the true model $\mathcal{M}_*$ in the final model $\mathcal{M}_\gamma$.

When there are more predictors than observations, the noise in estimation can be very large, causing by over fitting and noise accumulation. To reduce the noise, regularization is frequently used. Let $\boldsymbol{\omega}^\lambda = (\omega_1^\lambda, \cdots, \omega_p^\lambda)^T$ be a p-vector obtained by the ridge regression:

$$\boldsymbol{\omega}^\lambda = \left(\mathbf{X}^T\mathbf{X} + \lambda I_p\right)^{-1}\mathbf{X}^T\mathbf{y}, \tag{2.6}$$

where $\lambda > 0$ is a regularization parameter. Then, one can also screen variables based on the magnitude of ω_i^λ in the same manner as (2.4). Note that

$$\lambda\boldsymbol{\omega}^\lambda \to \boldsymbol{\omega} \quad \text{as } \lambda \to \infty \tag{2.7}$$

and the ordering in (2.4) does not depend on the scaling factor λ. We conclude that the componentwise regression $\boldsymbol{\omega}$ corresponds to the ridge regression with $\lambda = \infty$, the most regularized estimator in the family. Fan and Lv [8] made further connections between the SIS and iterative thresholded ridge regression screener and established its sampling property. That provides rationale why SIS works.

2.2 Model selection in high-dimensional space

After screening variables, we can now apply more refined techniques to further select the variables and estimate regression coefficients. Due to the sure screening property, the variables for further analysis should be in the set $\mathcal{M}_\gamma$ of size $n-1$ or $n/\log n$. For simplicity of notation, we still use $X_1, \cdots, X_d$ to denote the selected variables.

We now briefly review several well-developed techniques for high-dimensional model selection. These include LASSO and SCAD in Fan and Li [7]) and Fan and Peng [9], adaptive LASSO in Zou [21], elastic net Zou and Hastie [20], and the Dantzig selector in Candes and Tao [2]. They can be combined with our sure screening method to select the variables further and to estimate the non-vanishing parameters.

2.2.1 Non-convex penalized least-squares

The penalized least squares can be written as

$$\ell(\boldsymbol{\beta}) = \frac{1}{2n}\sum_{i=1}^{n}\left(Y_i - \mathbf{x}_i^T\boldsymbol{\beta}\right)^2 + \sum_{j=1}^{d} p_{\lambda_j}(|\beta_j|), \tag{2.8}$$

where $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_d)^T \in \mathbf{R}^d$ and $p_{\lambda_j}(\cdot)$ is a penalty function indexed by a regularization parameter λ_j. By minimizing (2.8), we hope to achieve simultaneously selecting the model and estimating the parameters (the variables with estimated coefficients zero are not selected).

Antoniadis and Fan [1] and Fan and Li [7] showed that the the singularity of $p_\lambda(\cdot)$ at the origin is needed in order to select simultaneously the variables and estimate parameters. An example of this is LASSO of Tibshirani [18] with $p_\lambda(\beta) = \lambda|\beta|$. Recent invention of the creative LARS algorithm (Efron, *et al.*[4]) makes it possible to compute the LASSO $\widehat{\boldsymbol{\beta}}_\lambda$ for all λ with $O(np)$ operations.

It is observed and shown in Fan and Li [7] and Zou [21] that LASSO has biases in estimating non-vanishing parameters. Antoniadis and Fan [1] and Fan and Li [7] advocated to use the non-convex penalized least-squares to reduce the bias. In particular, they proposed using the smoothly clipped absolute deviation (SCAD) penalty function $p_\lambda(\cdot)$, which is a non-decreasing quadratic spline on $[0, \infty)$, linear on $(0, \lambda)$ and constant on $[a\lambda, \infty)$ for some $a > 2$:

$$p'_\lambda(|\beta|) = \lambda\left\{I(|\beta| \leqslant \lambda) + \frac{(a\lambda - |\beta|)_+}{(a-1)\lambda} I(|\beta| > \lambda)\right\}. \tag{2.9}$$

The MCP of Zhang [22]

$$p'_\lambda(|\beta|) = (a\lambda - |\beta|)_+/a, \tag{2.10}$$

removes the linear part of the SCAD and minimizes the maximum of the concavity. The elastic net in Zou and Hastie [20] uses $p_\lambda(|\beta|) = \lambda_1|\beta| + \lambda_2\beta^2$.

An algorithm for optimizing penalized likelihood, called local quadratic approximation, was proposed in Fan and Li [7] and studied in Hunter and Li [13]. Recently, Zou and Li [24] proposed the local linear approximation to take the advantage of the innovative LARS algorithm by Efron *et al.*developed in 2004. To see this, assume that $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{1,0}, \cdots, \boldsymbol{\beta}_{d,0})^T$ is the current value of the estimate. If some components are zero, delete the variables from the model. For those components with non-vanishing coefficients, by approximating

$$p_{\lambda_j}(|\beta_j|) \approx p_{\lambda_j}(|\beta_{j,0}|) + \frac{\partial}{\partial\beta_j} p_{\lambda_j}(|\beta_{j,0}|)\mathrm{sgn}(\beta_{j,0})((|\beta_j| - |\beta_{j,0}|), \tag{2.11}$$

the updated estimate is to minimize

$$\frac{1}{2n}\sum_{i=1}^{n}\left(Y_i - \mathbf{x}_i^T\boldsymbol{\beta}\right)^2 + \sum_{j=1}^{d} w_j|\beta_j|, \tag{2.12}$$

where $w_j = \frac{\partial}{\partial \beta_j} p_{\lambda_j}(|\beta_{j,0}|)\text{sgn}(\beta_{j,0}) \geqslant 0$ (The condition that $p_\lambda(\cdot)$ is nondecreasing over $[0, \infty)$ was imposed by Fan and Li [7]). The estimator that minimizes (2.12) is called the adaptive LASSO by Zou [21], although the construction of the adaptive LASSO of Zou [21] is different from ours here. In this sense, the penalty function p_λ dictates the scheme on how the weights w_j vary according to the current value $\beta_{j,0}$ in each iteration. The weights are usually larger when $|\beta_{j,0}|$ is closer to zero; see for example (2.11) and (2.12).

The oracle property was established in Fan and Li [7] when p is finite and in Fan and Peng [9] when $p_n = o(n^{1/5})$. Zou [21] shows that the adaptive LASSO possesses the oracle property, too. See also further theoretical studies by Zhang and Huang [23] and Zhang [22].

2.2.2 Dantzig selector

The Dantzig selector $\widehat{\boldsymbol{\beta}}_{\text{DS}}$, proposed in Candes and Tao [2], is the solution to the following ℓ_1-regularization problem:

$$\min_{\boldsymbol{\zeta} \in \mathbf{R}^d} \|\boldsymbol{\zeta}\|_1 \quad \text{s.t.} \quad \left\| \mathbf{X}^T \mathbf{r} \right\|_\infty \leqslant \lambda_d, \tag{2.13}$$

where $\lambda_d > 0$ is a tuning parameter, $\mathbf{r} = \mathbf{y} - \mathbf{X}\boldsymbol{\zeta}$ is an n-vector of the residuals. They pointed out that the above convex optimization problem can easily be recast as a linear program:

$$\min_{\mathbf{u}} \sum_{i=1}^{d} u_i \quad \text{s.t.} -\mathbf{u} \leqslant \boldsymbol{\zeta} \leqslant \mathbf{u} \text{ and } -\lambda_d \mathbf{1} \leqslant \mathbf{X}^T (\mathbf{y} - \mathbf{X}\boldsymbol{\zeta}) \leqslant \lambda_d \mathbf{1},$$

where $\mathbf{u} = (u_1, \cdots, u_d)^T$ and $\boldsymbol{\zeta} \in \mathbf{R}^d$, and $\mathbf{1}$ is a d-vector of ones.

In the seminal paper by Candes and Tao [2], they showed that with the choice of $\lambda_d = \sqrt{2(\log d)/n}\sigma$, if the design matrix satisfies the uniform uncertainty condition (see also Zhang, 2007), then

$$\left\| \widehat{\boldsymbol{\beta}}_{DS} - \boldsymbol{\beta} \right\|^2 = O_p\left((\log d) s\sigma^2 / n \right),$$

recalling that s is the number of nonvanishing components of $\boldsymbol{\beta}$. In other words, it mimicks the oracle performance within a factor of $\log d$.

When the dimensionality is ultra high, i.e., $p_n = \exp(O(n^\xi))$ for some $\xi > 0$, then a direct application of the Dantzig selector to the orignal problem results in a loss of a factor $\log p_n = O(n^\xi)$ which could be too large to be acceptable. On the other hand, with the dimensionality first reduced by SIS the loss is now merely of a factor $\log d$, which is less than $\log n$.

2.3 Iterative SIS

SIS utilizes only the marginal information about the covariates, and can fail when the technical conditions in Theorem 2.1 fail. Fan and Lv [8] noted the three potential issues associated with SIS:

(a) an important predictor that is marginally uncorrelated but jointly correlated with the response can not be picked by SIS [the second condition in (c) in Theorem 2.1 rules this out];
(b) some unimportant predictors that are highly correlated with the important predictors can have higher priority to be selected by SIS than other important predictors that are relatively weakly related to the response (the first condition in (c) in Theorem 2.1 and other conditions rules this out);
(c) the issue of collinearity between predictors adds difficulty to the problem of variable selection [The assumption (d) in Theorem 2.1 prevents this from happening].

They addressed the issue by the following iterative SIS (ISIS), which uses more fully the joint information of the covariates and maintains computational expedience.

The ISIS works as follows. In the first step, we apply the SIS to screen the variables down to $n/\log n$ (say) and then employ LASSO or SCAD to further select a subset of k_1 variabes $\mathcal{A}_1 = \{X_{i_1}, \cdots, X_{i_{k_1}}\}$. Then we have an n-vector of the residuals from regressing the response Y over $X_{i_1}, \cdots, X_{i_{k_1}}$. In the next step, we treat those residuals as the new responses and apply the same method as in the previous step to the remaining $p - k_1$ variables, which results in a subset of k_2 variables $\mathcal{A}_2 = \{X_{j_1}, \cdots, X_{j_{k_2}}\}$. We can keep on doing this until we get ℓ disjoint subsets $\mathcal{A}_1, \cdots, \mathcal{A}_\ell$ whose union $\mathcal{A} = \cup_{i=1}^{\ell} \mathcal{A}_i$ has a size d, which is less than n. In practical implementation, we can choose, for example, the largest l such that $|\mathcal{A}| < n$. This achieves the variable screening. Finally, one can apply a penalized least-squares in Section 2.2 to further select variables and to estimate non-vanishing coefficients.

Fan and Lv [8] showed that the ISIS improves significantly the performance of SIS in the difficult cases mentioned at the beginning of this subsection. This can be understood as follows. Fitting the residuals from the previous step on $\{X_1, \cdots, X_p\} \setminus \mathcal{A}_1$ significantly weakens the priority of those unimportant variables that are highly correlated with the response through their associations with $X_{i_1}, \cdots, X_{i_{k_1}}$. This is due to the fact that the residuals and the variables $X_{i_1}, \cdots, X_{i_{k_1}}$ are uncorrelated. This helps solving (b). It also makes those important predictors that are missed in the previous step possible to survive, which addresses (a). In fact, after variables in $\mathcal{A}_1$ entering into the model, those that are marginally weakly correlated with Y purely due to the presence of variables in $\mathcal{A}_1$ should now be correlated with the residuals.

3 Variable screening for generalized likelihood models

In classification, the class label can be 0 or 1. Fan and Lv [8] pointed that SIS is equivalent to using a version of two-sample t test to select features and Fan and Fan [5] showed the sure screening property indeed holds. However, the issues outlined in Section 2.3 also exist for classification problem.

For classification problem, it is more elegant to cast the problem as a logistic regression problem. Then the conditional log-likelihood can be written as

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} [Y_i \log p_i(\boldsymbol{\beta}) + (1 - Y_i) \log\{1 - p_i(\boldsymbol{\beta})\}], \tag{3.1}$$

where $p_i(\beta)$ is usually modeled through a link function such as the logit-link

$$\log \frac{p_i(\boldsymbol{\beta})}{1 - p_i(\boldsymbol{\beta})} = \mathbf{x}_i^T \boldsymbol{\beta}.$$

When the feature vector is large, maximizing (3.1) is nearly impossible: the solution might not exist in addition to computation cost.

In statistical learning, various loss functions have been introduced for classifications. See Hastie, Tibshirani and Friedman [12]. The likelihood function (3.1) is replaced by the negative pseudo-likelihood function (translating into loss):

$$Q(\boldsymbol{\beta}) = \sum_{i=1}^{n} L(Y_i, \mathbf{x}_i^T \boldsymbol{\beta}). \tag{3.2}$$

For example, if the class-label is written as $\{-1, 1\}$, the hinge loss $L(Y_i, \mathbf{x}_i^T \boldsymbol{\beta}) = (1 - Y_i \mathbf{x}_i^T \boldsymbol{\beta})_+$, corresponding to the support-vector machine classifier; and $L(Y_i, \mathbf{x}_i^T \boldsymbol{\beta}) = \exp(-Y_i \mathbf{x}_i^T \boldsymbol{\beta})$ corresponding to the AdaBoost.

On the other hand, the framework (3.2) accommodates other likelihood or quasi-likelihood based models. If the conditional distribution of Y_i given the covariates $\mathbf{x}_i$ is Poisson distribution with rate $\lambda_i(\boldsymbol{\beta})$, then

$$L(Y_i, x_i^T \boldsymbol{\beta}) = -Y_i \log \lambda_i(\boldsymbol{\beta}) + \lambda_i(\boldsymbol{\beta}) + \log(Y_i!),$$

and $\lambda_i(\boldsymbol{\beta})$ is related to the covariates through modeling

$$\log \lambda_i(\boldsymbol{\beta}) = \mathbf{x}_i^T \boldsymbol{\beta}.$$

Suppose that we wish to select features based on the pseudo-likelihood model as in (3.2). One can extend directly SIS by minimizing

$$Q(\beta_j) = \sum_{i=1}^{n} L(Y_i, X_{ij} \beta_j)$$

for each component, resulting an estimate ω_j, where X_{ij} is the j^{th} component of $\mathbf{x}_i$. We can then apply the screening procedure as in (2.4). With the selected variables in $\mathcal{M}$, one can then apply the non-concave penalized likelihood to further select the variables and to estimate non-vanishing coefficients as in Fan and Li [7]. In other words, we minimize

$$\sum_{i=1}^{n} L(Y_i, \mathbf{x}_{i,\mathcal{M}}^T \boldsymbol{\beta}_{\mathcal{M}}) + \sum_{j \in \mathcal{M}} p_\lambda(|\beta_j|), \tag{3.3}$$

where $\mathbf{x}_{i,\mathcal{M}}$ is a sub-vector of $\mathbf{x}_i$ with elements in $\mathcal{M}$.

The question then arises how to generalize ISIS to the pseudo-likelihood to enhance the performance of SIS in variable screening. The challenge is to define an appropriate 'residual'. They are several versions of residuals in the statistical literature such as those in the generalized linear models. It is unclear whether it works for the variable screening purpose.

A possible extension of ISIS to the pseudo-likelihood based problem is as follows. After obtaining the first k_1 variables with index $\mathcal{A}_1 = \{i_1, \cdots, i_{k_1}\}$ from fitting (3.3) with estimated coefficients $\widehat{\boldsymbol{\beta}}_{\mathcal{A}_1}$, instead of defining the residuals, we minimize marginally

$$Q_1(\beta_j) = \sum_{i=1}^{n} L(Y_i, \mathbf{x}_{i,\mathcal{A}_1}^T \widehat{\boldsymbol{\beta}}_{\mathcal{A}_1} + X_{ij}\beta_j) \tag{3.4}$$

for $j \in \mathcal{A}_1^c$, where $\mathbf{x}_{i,\mathcal{A}_1}$ is a sub-vector of $\mathbf{x}_i$ with elements in $\mathcal{A}_1$. This is again a univariate optimization problem and can easily be computed.

Note that for the least-squares case,

$$L(Y_i, \mathbf{x}_{i,\mathcal{A}_1}^T \widehat{\boldsymbol{\beta}}_{\mathcal{A}_1} + X_{ij}\beta_j) = (r_i - X_{ij}\beta_j)^2,$$

where $r_i = Y_i - \mathbf{x}_{i,\mathcal{A}_1}^T \boldsymbol{\beta}_{\mathcal{A}_1}$ is the residual from the submodel $\mathcal{A}_1$. Hence, the above method is an extension of the least-squares procedure without explicit definition of the residuals. It is in the same spirit as the AdaBoost algorithm of Freund and Schapire [11].

Let ω_j be the maximizer of (3.4). We can then apply (2.4) to screen variables among $p - k_1$ variables in $\mathcal{A}_1^c$, resulting in a subset $\mathcal{M}_2$. Now, apply the penalized likelihood

$$\sum_{i=1}^{n} L(Y_i, \mathbf{x}_{i,\mathcal{A}_1}^T \widehat{\boldsymbol{\beta}}_{\mathcal{A}_1} + \mathbf{x}_{i,\mathcal{M}_2}^T \boldsymbol{\beta}_{\mathcal{M}_2}) + \sum_{j \in \mathcal{M}_2} p_\lambda(|\beta_j|) \tag{3.5}$$

to estimate $\boldsymbol{\beta}_{\mathcal{M}_2}$, resulting in the estimate $\widehat{\boldsymbol{\beta}}_{\mathcal{M}_2}$. Let $\mathcal{A}_2$ be the subset of $\mathcal{M}_2$ with non-vanishing elements of $\widehat{\boldsymbol{\beta}}_{\mathcal{M}_2}$. This recruits variables $\mathcal{A}_2^* = \mathcal{A}_1 \cup \mathcal{A}_2$.

Further recruitment can be accomplished by the following regression problem that is similar to (3.4):

$$Q_2(\beta_j) = \sum_{i=1}^{n} L(Y_i, \mathbf{x}_{i,\mathcal{A}_2^*}^T \widehat{\boldsymbol{\beta}}_{\mathcal{A}_2^*} + X_{ij}\beta_j) \tag{3.6}$$

for $j \in \mathcal{A}_2^{*c}$ to select the variables in $\mathcal{A}_2^{*c}$, resulting in $\mathcal{M}_3$ and then further screen variables via the penalized likelihood

$$\sum_{i=1}^{n} L(Y_i, \mathbf{x}_{i,\mathcal{A}_2^*}^T \widehat{\boldsymbol{\beta}}_{\mathcal{A}_2^*} + \mathbf{x}_{i,\mathcal{M}_3}^T \boldsymbol{\beta}_{\mathcal{M}_3}) + \sum_{j \in \mathcal{M}_3} p_\lambda(|\beta_j|). \tag{3.7}$$

Iterating the procedure in this matter results in the selected variables $\mathcal{A}$ in the same manner as ISIS.

With the selected variables in $\mathcal{A}$, with model size smaller than n, we can apply the penalized likelihood

$$\sum_{i=1}^{n} L(Y_i, \mathbf{x}_{i,\mathcal{A}}^T \boldsymbol{\beta}_{\mathcal{A}}) + \sum_{j\in\mathcal{A}} p_\lambda(|\beta_j|) \tag{3.8}$$

to select the final model and to estimate the coefficients. This is clearly a feasible extension of ISIS of Fan and Lv [8].

4 Portfolio selection and multiple regression

This section shows how the portfolio selection problem can be reduced to a multiple regression problem. To simplify the notation, we only consider the population version with understanding that the sample version can easily be extended.

Suppose that we have p assets with returns $X_1, \cdots, X_p$ at period t, in which the dependence of the return on t is dropped. Let $\mathbf{X}$ be the vector of the returns. A portfolio with weights $\mathbf{w}$, satisfying $\mathbf{w}^T\mathbf{1} = 1$ (total allocation is 100%), have the risk

$$\text{var}(\mathbf{w}^T\mathbf{X}) = \mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w}. \tag{4.1}$$

The global minimum portfolio is to find $\mathbf{w}$ that minimizes (2.1).

Suppose that the riskless interest rate is r_0 and instead of investing all money in the risky assets $\mathbf{X}$, we invest $(1-\mathbf{w}^T\mathbf{1})$ on the riskless asset, and $\mathbf{w}$ on the risky assets $\mathbf{X}$ (so that the total investment is 100%). The expected return of such a portfolio, consisting of $(p+1)$ assets, is $(1-\mathbf{w}^T\mathbf{1})r_0 + \mathbf{w}^T\boldsymbol{\mu}$, with the risk given by (4.1), where $\boldsymbol{\mu} = E\mathbf{X}$. The portfolio allocation problem of Markowitz [15] is to maximize the expected utility function

$$(1-\mathbf{w}^T\mathbf{1})r_0 + \mathbf{w}^T\boldsymbol{\mu} - \frac{\lambda}{2}\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w}, \tag{4.2}$$

for a given $\lambda > 0$. The problem is equivalent to minimizing (4.1) with the constraint

$$(1-\mathbf{w}^T\mathbf{1})r_0 + \mathbf{w}^T\boldsymbol{\mu} = c, \tag{4.3}$$

for some constant $c > 0$, namely to construct a portfolio of assets that minimizes the risk with the expected return c,

In practice, instead of targeting the expected return c in (4.3), many mutual funds or exchange-traded funds(ETF) set the target percentages on each sector or each region, such as 25% in energy, 15% in banking, and so on. This translates into the constraints

$$\mathbf{B}\mathbf{w} = \mathbf{c} \tag{4.4}$$

for a given matrix $\mathbf{B}$ and vector $\mathbf{c}$. Clearly, Markowitz's problem (4.2) is a specific problem (4.4).

It is not difficult to find the optimal allocation vectors for all aforementioned optimization problems. For example, the global minimum portfolio, minimizing (4.1) subject to the constraint $\mathbf{w}^T\mathbf{1} = 1$, is given by

$$\mathbf{w}^* = \boldsymbol{\Sigma}^{-1}\mathbf{1}/\mathbf{1}^T\boldsymbol{\Sigma}^{-1}\mathbf{1} \tag{4.5}$$

with minimum risk $(\mathbf{1}^T\mathbf{\Sigma}^{-1}\mathbf{1})^{-1}$. However, in practical implementation, one needs to estimate $\mathbf{\Sigma}$ based on a sample of size n. For $p = 2000$, as mentioned in the introduction, there are over two millions of estimated parameters, yielding estimated vector $\hat{\mathbf{w}}^* = \widehat{\mathbf{\Sigma}}^{-1}\mathbf{1}/\mathbf{1}^T\widehat{\mathbf{\Sigma}}^{-1}\mathbf{1}$. However, the risk of the portfolio $\hat{\mathbf{w}}^*\mathbf{\Sigma}\hat{\mathbf{w}}^*$ might be very different the optimal risk $(\mathbf{1}^T\mathbf{\Sigma}^{-1}\mathbf{1})^{-1}$, due to the accumulation of estimation errors. Indeed, Fan, Fan and Lv [6] showed that the convergence rate is extremely slow. This means that the number of selected assets can not be too large. This is a desirable property in practice, as the costs for research, monitoring, and transactions of 2000 stocks are expensive.

For both statistical and practical reasons, one wishes to find sparse allocation that minimizes the portfolio variance (4.1).

4.1 Global minimum portfolio

To make connections with the least-squares problem, let $\mu_j = EX_j$ and $X_j^* = X_j - X_1 - (\mu_j - \mu_1)$. Set $Y = X_1 - \mu_1$. Then, using $w_1 = 1 - w_2 - \cdots - w_p$, we have

$$\text{var}(\mathbf{w}^T\mathbf{X}) = E\big(Y - w_2X_2^* - \cdots - w_pX_p^*\big)^2. \tag{4.6}$$

Thus, the global minimum portfolio becomes the least-squares problems with respect to the parameters $w_2, \cdots, w_p$. The sparse global minimum portfolio can be obtained by the penalized least-squares

$$E\big(Y - w_2X_2^* - \cdots - w_pX_p^*\big)^2 + \lambda\sum_{j=2}^{p} p_\lambda(|w_j|). \tag{4.7}$$

The penalty function can be LASSO $p_\lambda(|w_j|) = \lambda|w_j|$ (Tibshirani [18]), SCAD (2.9) (Fan and Li [7]), elastic net $p_\lambda(|w_j|) = \lambda_1|w_j| + \lambda_2|w_j|^2$ (Zou and Hastie [20]), and MCP (2.10) (Zhang [22]). In particular, let $w_j(\lambda)$ be the solution to the penalized L_1 problem. It is clear that $\lambda = \infty$ results in the solution $w_j(\lambda) = 0$ for all j, and $\lambda = 0$ corresponds to chooses all $p - 1$ assets. Using the LARS algorithm of Efron et al [4], we are able to obtain all solutions $w_j(\lambda)$, including $w_1(\lambda) = 1 - w_2(\lambda) - \cdots - w_p(\lambda)$ associated with X_1, the risk $\sigma^2(\lambda)$ of the allocated portfolio, and the number of selected assets $N(\lambda)$. Similar paths can be obtained for other penalty functions.

In practice, one can choose X_1 to be the asset with the minimum risk. It can also be a tradable portfolio such as the index or exchange-traded fund (ETF) as long as it is liquidly traded.

For optimal portfolio with constraints (4.3), we can proceed the same way as above except that the constraints need to be observed. New algorithm needs to be developed to take care of the constraints.

4.2 Factor models

The Capital Assess Pricing Model (CAPM, Sharpe [17]; Lintner, [14]) imposes the following structure on the excess return (over the riskless interest rate):

$$X_i = \beta_i f + \varepsilon_i,$$

where f is the excess return of the market portfolio and β_i is the market "β", and ε_i is the idiosyncratic noise, independent of f. Then,

$$\text{var}(\mathbf{w}^T\mathbf{X}) = (\mathbf{w}^T\boldsymbol{\beta})^2\text{var}(f) + \|\mathbf{w}\|_\sigma^2, \tag{4.8}$$

where $\|\mathbf{w}\|_\sigma^2 = \sum_{j=1}^p \sigma_j^2 w_j^2$ with $\sigma_j^2 = \text{var}(\varepsilon_j)$. To find the sparse solution of the global minimum portifolio, one can apply the penalized least-squares approach, which minimizes

$$(\mathbf{w}^T\boldsymbol{\beta})^2\text{var}(f) + \|\mathbf{w}\|_\sigma^2 + \sum_{j=1}^p p_\lambda(|w_j|), \tag{4.9}$$

subject to the constraint (4.3). If $p_\lambda(|w_j|) = \lambda|w_j|$, the above problem is very similar to the elastic net in Zou and Hastie [20].

A further extension of the CAPM model is the multi-factor model (Ross, 1976; Chamberlian and Rothchild, 1982), which admits

$$\mathbf{X} = \mathbf{B}\mathbf{f} + \boldsymbol{\varepsilon},$$

where $\mathbf{f}$ is a vector of k factors that influence the returns, $\mathbf{B}$ is a $p \times k$ factor loading matrix, and $\boldsymbol{\varepsilon}$ is the vector of idiosyncratic noises. Let $\boldsymbol{\Omega} = \text{var}(\mathbf{f})$ and $\mathbf{B}^* = \mathbf{B}\boldsymbol{\Omega}^{1/2}$. Then,

$$\text{var}(\mathbf{w}^T\mathbf{X}) = (\mathbf{w}^T\mathbf{B}^*\mathbf{B}^{*T}\mathbf{w})^2 + \|\mathbf{w}\|_\sigma^2. \tag{4.10}$$

Write $B^* = (\mathbf{b}_1^*, \cdots, \mathbf{b}_k^*)$. Then, (4.10) can be expressed as

$$\text{var}(\mathbf{w}^T\mathbf{X}) = \sum_{j=1}^k (\mathbf{b}_j^T\mathbf{w})^2 + \|\mathbf{w}\|_\sigma^2. \tag{4.11}$$

The sparse solution can also be found via penalized least-squares problem.

References

[1] Antoniadis, A. & Fan, J., Regularized wavelet approximations (with discussion), *Jour. Ameri. Statist. Assoc.*, 96 (2001), 939-967.

[2] Candes, E. & Tao, T., The Dantzig selector: statistical estimation when p is much larger than n (with discussion), *Ann. Statist.*, to appear.

[3] Chamberlain, G. & Rothschild, M., Arbitrage, factor structure, and mean-variance analysis on large asset markets, *Econometrica*, 51 (1983), 1281–1304.

[4] Efron, B., Hastie, T., Johnstone, I. & Tibshirani, R., Least angle regression (with discussions), *Ann. Statist.*, 32 (2004), 409-499.

[5] Fan, J. & Fan, Y., High dimensional classification using shrunken independence rule, *Ann. Statist*, to appear.

[6] Fan, J., Fan, Y. & Lv, J., Large dimensional covariance matrix estimation via a factor model, *Journal of Econometrics*, to appear.

[7] Fan, J. & Li, R., Variable selection via nonconcave penalized likelihood and its oracle properties, *Jour. Ameri. Statist. Assoc.*, 96 (2001), 1348-1360.

[8] Fan, J. & Lv, J., Sure independence screening for ultra-high dimensional feature space, *Jour. Roy. Statist. Assoc. B*, to appear.
[9] Fan, J. & Peng, H., On non-concave penalized likelihood with diverging number of parameters, *Ann. Statist.*, 32 (2004), 928-961.
[10] Fan, J. & Ren, Y., Statistical analysis of DNA microarray data, *Clinical Cancer Research* 12 (2006), 4469–4473.
[11] Freund, Y. & Schapire, R.E., A decision-theoretic generalization of on-line learning and an application to boosting. *Jour. Comput. Sys. Sci.*, 55 (1997), 119–139.
[12] Hastie, T.J., Tibshirani, R. & Friedman, J., *The elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, 2001.
[13] Hunter, D. & Li, R., Variable selection using MM algorithms. *Ann. Statist.*. 33 (2005), 1617-1642.
[14] Lintner, J., The valuation of risky assets and the selectin of risky investments in stock portfolios and capital budgets, *Rev. Econ. Statist.*, 47 (1965), 13-37.
[15] Markowitz, H. M., Portfolio selection. *Journal of Finance*, 7 (1952), 77–91.
[16] Ross, S.A., The arbitrage theory of capital asset pricing, *Journal of Economic Theory*, 3 (1976), 341-360.
[17] Sharpe, W., Capital asset prices: A theory of market equilibrium under conditons of risks, *Jour. Fin.*, 19 (1964), 425-442.
[18] Tibshirani, R., Regression shrinkage and selection via lasso, *Jour. Roy. Statist. Soc. B.*, 58 (1996), 267-288.
[19] Tibshirani, R., Hastie, T., Narasimhan, B. & Chu, G., Class prediction by nearest shrunken centroids, with applications to DNA microarrays. *Statist. Sci.*, 18 (2003), 104-117.
[20] Zou, H. & Hastie, T., Regularization and variable selection via the elastic net, *Jour. Roy. Statist. Soc. B*, 67 (2005), 301-320.
[21] Zou, H., The adaptive Lasso & its oracle properties, *J. Amer. Statist. Assoc.*, 101 (2006), 1418–1429.
[22] Zhang, C.-H., Penalized linear unbiased selection, *Manuscript*.
[23] Zhang, C.-H. & Huang, J., The sparsity and bias of the LASSO selection in high-dimensional linear regression. *Ann. Statist.*, to appear.
[24] Zou, H. & Li, R., One-step sparse estimates in nonconcave penalized likelihood models (with discussion), *Ann. Statist.*, to appear.