Part 2: Feature Screening for Ultrahigh Dimensional Data

Feature screening

Also referred to as **Variable Screening**

Contents:

1. Model-based feature screening
   1.a. Marginal screening
   1.b. Joint screening
2. Model free feature screening

4.1 Model-based feature screening

Main ideas: (a) to create a marginal utility measure, and then (b) use the marginal utility measure screening predictors

Goal: reduce dimensionality of predictors from a large or huge scale (say, $\exp(O(n^{\xi}))$ for some $\xi > 0$) to a relatively large scale d (e.g, $o(n)$) by a **fast and efficient** screening method

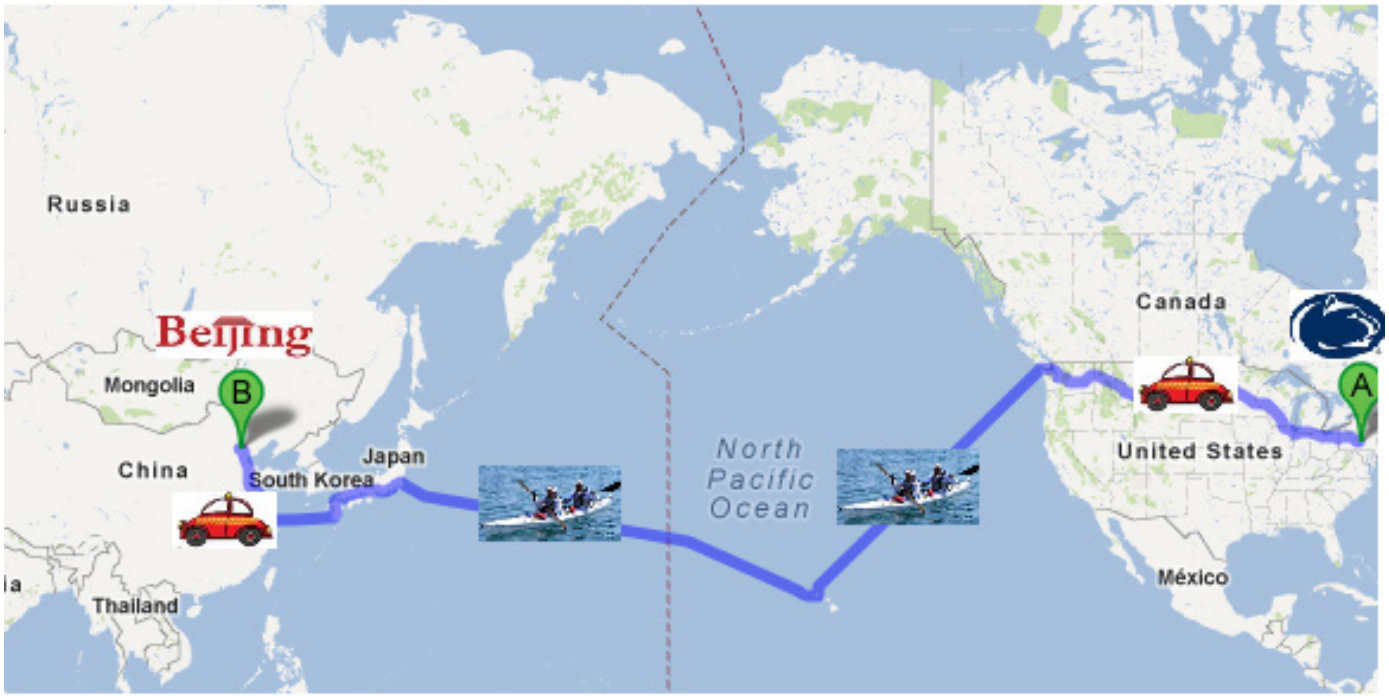
Fan and Lv (2008) proposed Sure Independence Screening (SIS) procedure for linear model

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

$\mathbf{X}$: $n \times p$ with n - the sample size n and p - the dimension of predictors.

Two-stage Approach: Intuition

Google map: driving direction from Penn State to Peking University



Two-stage Approach: Intuition

(1) Screening stage: a "Quick" way −− Fly from USA to Peking

(2) Post-Screening Selection Stage: one can drive from airport to Peking University.

Notation

Let $\{\mathbf{x}_i, Y_i\}$, $i = 1, \cdots, n$ be iid from a population $\{\mathbf{x}, Y\}$, where $\mathbf{x} = (X_1, \cdots, X_p)^T$ is the p-dimensional predictor variables, and Y is the response variable.

Denote $\mathbf{y} = (Y_1, \ldots, Y_n)^T$ and $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_n)^T = (\mathbf{X}_1, \ldots, \mathbf{X}_p)$, the $n \times p$ design matrix. Furthermore denote by X_{ij} the ith observation of the jth variable. Thus, $\mathbf{x}_i = (X_{i1}, \cdots, X_{ip})^T$. Let ε be a general random error and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$ with ε_i being an $n \times 1$ vector of random errors.

Let $\mathcal{M}_*$ stand for the true model with the size $s = |\mathcal{M}_*|$, and $\widehat{\mathcal{M}}$ is the selected model with the size $d = |\widehat{\mathcal{M}}|$. The definitions of $\mathcal{M}_*$ and $\widehat{\mathcal{M}}$ may be different for different models and contexts.

Feature Screening for Linear Models

Consider a linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1}$$

When the dimension p is greater than the sample size n, the least squares estimator of $\boldsymbol{\beta}$ is not well defined due to the singularity of $\mathbf{X}^T\mathbf{X}$. A useful technique in the classical linear regression to deal with singularity of the design matrix $\mathbf{X}$ is the ridge regression, defined by

$$\widehat{\boldsymbol{\beta}}_\lambda = (\mathbf{X}^T\mathbf{X} + \lambda I_p)^{-1}\mathbf{X}^T\mathbf{y},$$

where λ is a ridge parameter.
It is observed that if $\lambda \to 0$, then $\widehat{\boldsymbol{\beta}}_\lambda$ tends to the least squares estimator, if it is well-defined; and if $\lambda \to \infty$, then $\lambda\widehat{\boldsymbol{\beta}}_\lambda$ tends to $\mathbf{X}^T\mathbf{y}$. This implies that $\widehat{\boldsymbol{\beta}}_\lambda \propto \mathbf{X}^T\mathbf{y}$.

Correlation learning

In practice, all covariates and the response are **marginally standardized** so that their means and variances equals 0 and 1, respectively. Then $\frac{1}{n}\mathbf{X}^T\mathbf{y}$ becomes the vector consists of the sample version of Pearson correlations between the response and individual covariate.

This is the motivation of using Pearson correlation as a marginal utility for feature screening. Specifically, denote

$$\omega_j = \frac{1}{n}\mathbf{X}_j{}^T\mathbf{y}, \quad \text{for } j = 1, 2, \ldots, p. \tag{2}$$

Here it is assumed that both $\mathbf{X}_j$ and $\mathbf{y}$ are marginally standardized. Thus, ω_j indeed is the sample correlation between the jth predictor and the response variable.

Correlation learning

Fan and Lv (2008) suggested ranking all predictors according to $|\omega_j|$ and select the top predictors which are relatively strongly correlated with the response. To be specific, for any given $\gamma \in (0, 1)$, the $[\gamma n]$ top ranked predictors are selected to obtain the submodel

$$\widehat{\mathcal{M}}_\gamma = \{1 \leq j \leq p : |\omega_j| \text{ is among the first } [\gamma n] \text{ largest of all}\}, \qquad (3)$$

where $[\gamma n]$ denotes the integer part of γn.
It reduces the ultrahigh dimensionality down to a relatively moderate scale $[\gamma n]$, i.e. the size of $\widehat{\mathcal{M}}_\gamma$, and then the well-established penalized variable selection methods is applied to the submodel $\widehat{\mathcal{M}}_\gamma$.

Connection between multiple comparison and correlation learning

Suppose that a random sample with size n was collected from two groups:

Disease (Case): $\mathbf{x}_i = (X_{i1}, \cdots, X_{ip})^T$, $i = 1, \cdots, n_1$
Normal (Control): $\mathbf{x}_i = (X_{i1}, \cdots, X_{ip})^T$, $i = n_1 + 1, \cdots, n_1 + n_2 (= n)$

Of interest is to identify which X-variables are associated with the disease status when $p >> n$.

How to do it?

Two-sample t-test

Need to assume normality on X_{ij}'s. Assume that samples from two groups have the same marginal variance.

Denote $\bar{X}_{+j}$ and $\bar{X}_{-j}$ to be the sample mean for disease and normal groups, respectively. Two-sample t-test can be written as

$$t_j = \frac{\sqrt{n_1 n_2 / n}(\bar{X}_{+j} - \bar{X}_{-j})}{\sqrt{\{(n_1 - 1)s_{+j}^2 + (n_2 - 1)s_{-j}^2)\}/(n-2)}}$$

where s_{+j}^2 and s_{-j}^2 to be the sample variance for disease and normal groups, respectively.

Multiple comparison: Bonferroni correction, FDR

Bonferroni correction, FDR

Assume that $s_{+j}^2 = s_{-j}^2 = 1$ (by marginal standardized). Then

$$t_j = \sqrt{n_1 n_2 / n}(\bar{X}_{+j} - \bar{X}_{-j})$$

Ranking P-values of two-sample t-test is equivalent to ranking $|t_j|$, which is proportional to $|\bar{X}_{+j} - \bar{X}_{-j}|$. Define

$Y_i = (1/n_1)\sqrt{n_1 n_2 / n}$ if the ith subject comes from disease group;

$Y_i = -(1/n_2)\sqrt{n_1 n_2 / n}$ if the ith subject comes from normal group.

Let $\mathbf{y} = (Y_1, \cdots, Y_n)^T$ and $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_n)^T = (\mathbf{X}_1, \cdots, \mathbf{X}_j)$. Then

$$t_j = \mathbf{X}_j^T \mathbf{y}$$

Bonferroni correction and FDR method rank variables by $|\mathbf{X}_j^T \mathbf{y}|$. Multiple comparison method can be viewed as correlation learning for two-sample mean problem.

Multiple testing and feature screening

Methods in multiple tests include false discovery rate method (Benjamini and Hockberg, 1995) are to derive a critical value or proper thresholding for all t-values.

Feature screening based on ω_j is distinguished from the multiple test methods:

(a) The screening method aims to rank the importance of predictors rather than directly judge whether an individual variable is significant.

(b) $\Rightarrow$ Fine-tuning analysis based on the selected variables from the screening step is necessary.

Note that it is not involved any numerical optimization to evaluate ω_j or $\mathbf{X}^T\mathbf{y}$. Thus, feature screening based on the Pearson correlation can be carried out in a simple way at low computational burden.

Issues related to multiple comparison methods

Null model: all X_js are independent of the disease status.

Since $\bar{Y} = 0$, t_j is proportional to $r(X_j, Y)$, the sample correlation between X_j and Y, under assumption the sample variance of X_j being 1.

Under the null model, all $|t_j|$s should be close to zero since $\rho(X_j, Y) = 0$.

Geometric interpretation: scatter p points on n-dimensional space.

There exist at most n orthogonal vectors in n-dimensional vector space.

Note that $r(X_j, Y)$ is the cosine of angle between $\mathbf{X}_j$ and $\mathbf{y}$.

This implies that some values of $r(X_j, Y)$ are not close to zero. Note that ranking variables according to $|t_j|$ is equivalent to ranking variables based on $|r(X_j, Y)|$.

Illustration of spurious correlation

To explain more clearly the concept about spurious correlation, we simulate $n = 50$ data points iid $N(0,1)$ $\{X_j\}_{j=1}^p$ (with $p = 1000$) and iid response $Y \sim N(0,1)$.

In this **null** model: $Y = \varepsilon$, all covariates $\{X_j\}_{j=1}^p$ and the response Y are independent and follow $N(0,1)$.

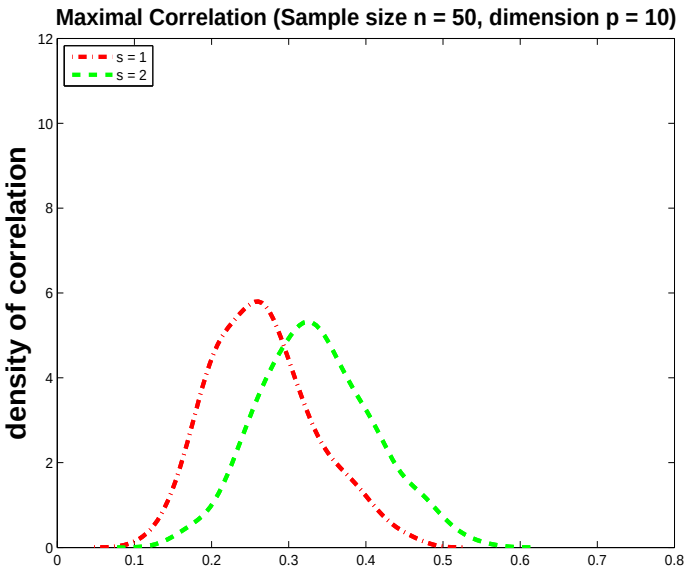
Estimate the maximum "linear" spurious correlation:

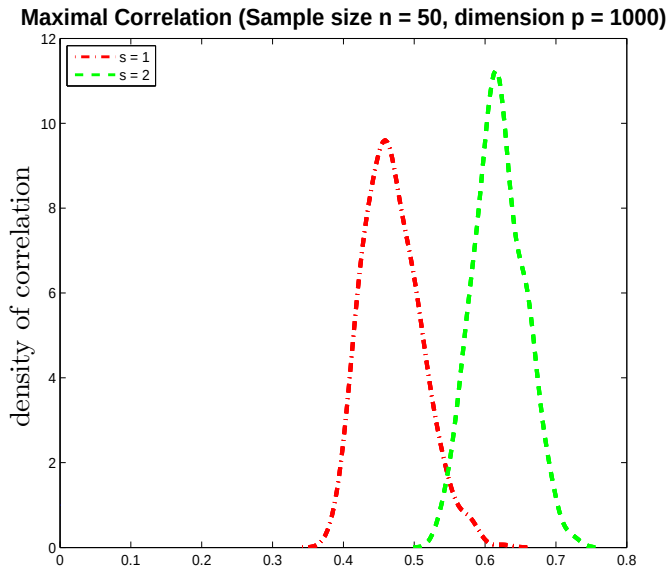
$$\zeta_n = \max_{1 \le j \le p} |\widehat{\text{corr}}(X_j, Y)|$$

In other words, one variable ($s = 1$) is selected to predict the realized noise vector best.

Idea can be extended to s variables: the correlation between the response and fitted values using the "best subset" of s-variables.

Spurious correlation





Density of ζ_n. Left: $s = 1, 2$, $(n, p) = (50, 10)$ based on 500 simulations.
Right: $s = 1, 2$, $(n, p) = (50, 1000)$ based on 500 simulations.

Strategy of high-dimensional data analysis

The phenomenon of spurious correlation explains why many variables identified by FDR are not really significant in ultrahigh dimensional setting.

Two-stage or several stage approach should be considered.

SIS based model selection techniques

Reduce model size from p to $d < n$ by using SIS.
PLS:

$$\frac{1}{2n}\|\mathbf{y} - \mathbf{X}_s\boldsymbol{\beta}_s\|^2 + \sum_{j=1}^{d} p_\lambda(|\beta_{sj}|)$$

where $\dim(\boldsymbol{\beta}_s) = d$.

Penalized least squares methods can be applied to further reduce model complexity.

In particular, LLA with initial value $\mathbf{0} \Rightarrow$ LASSO estimate in the first step, ...

Extensions of sure independent screening

The key idea of SIS is to apply a single componentwise regression. Three potential issues might arise with this approach.

1. Some unimportant predictors that are highly correlated with the important predictors can have higher priority for being selected by SIS than other important predictors that are relatively weakly related to the response.

2. An important predictor that is marginally uncorrelated but jointly correlated with the response cannot be picked by SIS and thus will not enter the estimated model.

3. The issue of collinearity between predictors adds difficulty to the problem of variable selection.

These three issues will be addressed in the extensions of SIS below, which allow us to use more fully the joint information of the covariates rather than just the marginal information in variable selection.

Iterative SIS (ISIS)

Step 1. We select a subset of k_1 variables $\mathcal{A}_1 = \{x_{i_1}, \cdots, x_{i_{k_1}}\}$ using an SIS-based model selection method such as SIS-SCAD or SIS-LASSO. Here $k_1 = [n/\log(n)]$.

Step 2. We regress y over the $x_{i_1}, \cdots, x_{i_{k_1}}$, and then treat the residuals of y over those variables as a new response, and apply the same method as in the previous step to remaining $p - k_1$ variables, which results in a subset of k_2 variables $\mathcal{A}_2 = \{x_{j_1}, \cdots, x_{j_{k_2}}\}$

Step 3. Repeat Step 2.

Remarks

In Step 2, fitting the residuals from the previous step on $\{X_1, ..., X_p\} \setminus \mathcal{A}_1$ can significantly weaken the priority of those unimportant variables that are highly correlated with the response through their associations with $X_{i_1}, \cdots, X_{i_{k1}}$, since the residuals are uncorrelated with those selected variables in $\mathcal{A}_1$. This helps to solve the first issue.

It also makes those important predictors that are missed in the previous step possible to survive, which addresses the second issue above. In fact, after variables in $\mathcal{A}_1$ enter 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.

Generalized correlation, rank correlation and transformation linear models

The SIS procedure performs well for the linear regression model with ultrahigh dimensional predictors.

It is well known that the Pearson correlation is used to measure linear dependence. In the presence of nonlinearity, one may try to make a transformation such as the Box-Cox transformation on the covariates.

This motivates people to consider the Pearson correlation between a transformed covariate and the response as the marginal utility.

Generalized correlation

To capture both linearity and nonlinearity, Hall and Miller (2009) defined the generalized correlation between the jth predictor X_j and Y to be,

$$\rho_g(X_j, Y) = \sup_{h \in \mathcal{H}} \frac{\text{cov}\{h(X_j), Y\}}{\sqrt{\text{var}\{h(X_j)\}\text{var}(Y)}}, \quad \text{for each } j = 1, \ldots, p, \tag{4}$$

where
$\mathcal{H}$ is a class of functions including all linear functions. For instance, it is a class of polynomial functions up to a given degree.

If $\mathcal{H}$ is restricted to be a class of all linear functions, $\rho_g(X_j, Y)$ is the absolute value of Pearson correlation between X_j and Y. Therefore, $\rho_g(X, Y)$ is considered as a generalization of the conventional Pearson correlation.

Generalized correlation

Suppose that $\{(X_{ij}, Y_i), i = 1, 2, \ldots, n\}$ is a random sample from the population (X_j, Y). The generalized correlation $\rho_g(X_j, Y)$ can be estimated by

$$\widehat{\rho}_g(X_j, Y) = \sup_{h \in \mathcal{H}} \frac{\sum_{i=1}^n \{h(X_{ij}) - \bar{h}_j\}(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^n \{h^2(X_{ij}) - \bar{h}_j^2\} \sum_{i=1}^n (Y_i - \overline{Y})^2}}, \tag{5}$$

where $\bar{h}_j = n^{-1} \sum_{i=1}^n h(X_{ij})$ and $\overline{Y} = n^{-1} \sum_{i=1}^n Y_i$.
In practice, $\mathcal{H}$ can be defined as a set of cubic splines (Hall and Miller, 2009).
How to choose an optimal transformation $h(\cdot)$ remains an open issue and the associated sure screening property is needed to justify.

Hall and Miller (2009) introduced a bootstrap method to determine a cutoff.

Let $r(j)$ be the ranking of the jth predictor X_j; that is, X_j has the $r(j)$ largest empirical generalized correlation of all. Let $r^*(j)$ be the ranking of the jth predictor X_j using the bootstrapped sample. Then, a nominal $(1-\alpha)$-level two-sided prediction interval of the ranking, $[\widehat{r}_-(j), \widehat{r}_+(j)]$, is computed based on the bootstrapped $r^*(j)$'s.

Hall and Miller (2009) recommended a criterion to regard the predictor X_j as influential if $\widehat{r}_+(j) < \frac{1}{2}p$ or some smaller fraction of p, such as $\frac{1}{4}p$.

Therefore, the proposed generalized correlation ranking reduces the ultrahigh p down to the size of the selected model $\widehat{\mathcal{M}}_k = \{j : \widehat{r}_+(j) < kp\}$, where $0 < k < 1/2$ is a constant multiplier to control the size of the selected model $\widehat{\mathcal{M}}_k$.

Rank correlation

Alternative to make transformations on predictors, one may make a transformation on the response and define a correlation between the transformed response and a covariate.
A general transformation regression model is defined to be

$$H(Y_i) = \mathbf{x}_i{}^T \boldsymbol{\beta} + \varepsilon_i, \tag{6}$$

Li, Peng, Zhang and Zhu (2012) proposed the rank correlation as a measure of the importance of each predictor by imposing an assumption on strict monotonicity on $H(\cdot)$ in model (6).

The marginal rank correlation:

$$\omega_j = \frac{1}{n(n-1)} \sum_{i \neq \ell}^{n} I(X_{ij} < X_{\ell j}) I(Y_i < Y_\ell) - \frac{1}{4}, \tag{7}$$

to measure the importance of the jth predictor X_j. According to the magnitudes of all ω_j's, the feature screening procedure based on the rank correlation selects a submodel

Defined the true model as $\mathcal{M}_* = \{1 \leq j \leq p : \beta_j \neq 0\}$ with the size $s = |\mathcal{M}_*|$.

Suppose that $\min_{j \in \mathcal{M}_*} E|X_j|$ is a positive constant free of p, and $H(\cdot)$ being an unspecified strictly increasing function. Under some technical conditions, the RRCS enjoys the sure screening property:

$$P\left(\mathcal{M}_* \subseteq \widehat{\mathcal{M}}_{\gamma_n}\right) \geq 1 - 2s\exp(-c_2 n^{1-2\kappa}) \to 1, \quad \text{as} \quad n \to \infty \tag{8}$$

hold for some constant c_2 provided that $\gamma_n = c_3 n^{-\kappa}$ for some constant c_3, $p = O(\exp(n^{\delta}))$ for some $\delta \in (0, 1)$ satisfying $\delta + 2\kappa < 1$ for any $\kappa \in (0, 0.5)$.

General strategy

Consider a simple linear regression model

$$Y = \beta_{j0} + X_j \beta_{j1} + \varepsilon_j^*, \tag{9}$$

where ε_j^* is a random error with $E(\varepsilon_j^* | X_j) = 0$.

Suppose that $\text{var}(X_j) = 1$ and $\text{var}(Y) = 1$. Then, $\beta_{j1} = \text{corr}(Y, X_j)$. The magnitude of β_{j1} can be used to rank the importance of the predictor.

Let $\mathbf{y}$ and $\mathbf{X}_j$ both be marginal standardized. Thus, $\widehat{\beta}_{j1} = n^{-1} \mathbf{X}_j{}^T \mathbf{y} = \omega_j$ in (2)

Note that the least squares estimate of β_{j0} is $\bar{Y} = 0$. Since $\|\mathbf{y}\|^2 = \|\mathbf{y} - \mathbf{X}_j \widehat{\beta}_{j1}\|^2 + n\|\widehat{\beta}_{j1}\|^2$ due to the fact $\mathbf{X}_j$ is marginally standardized (i.e, $\|\mathbf{X}_j\|^2 = n$). Thus, the greater $|\widehat{\beta}_{j1}|$ is, the smaller $\text{RSS}_j = \|\mathbf{y} - \mathbf{X}_j \widehat{\beta}_{j1}\|^2$.

Use $|\widehat{\beta}_{j1}|$ or RSS_j to rank importance of predictors.

Feature screening for generalized linear models

Given $\mathbf{x} = (X_1, \cdots, X_p)$, Y has a density function:

$$f_{Y|\mathbf{x}}(y|\mathbf{x}) = \exp\{y\theta(\mathbf{x}) - b(\theta(\mathbf{x})) + c(y)\}, \tag{10}$$

for some known functions $b(\cdot)$, $c(\cdot)$. Here Y can be continuous, binary or count.

Assume $E(Y|\mathbf{x}) = b'\{\theta(\mathbf{x})\} = g^{-1}(\beta_0 + \mathbf{x}^T\boldsymbol{\beta})$. Denote the negative log-likelihood of i-th observation $\{\mathbf{x}_i, Y_i\}$ to be $\ell(\beta_0 + \mathbf{x}_i{}^T\boldsymbol{\beta}, Y_i)$.

Note that the minimizer of the negative log-likelihood $\sum_{i=1}^n \ell(\beta_0 + \mathbf{x}_i{}^T\boldsymbol{\beta}, Y_i)$ is not well defined when $p > n$.

Feature screening for generalized linear models

Parallel to the least squares estimate for model (9), assume that each predictor is standardized to have mean zero and standard deviation one, and define the maximum marginal likelihood estimator (MMLE) $\widehat{\beta}_j^M$ for the jth predictor X_j as

$$\widehat{\beta}_j^M = (\widehat{\beta}_{j0}^M, \widehat{\beta}_{j1}^M) = \arg\min_{\beta_{j0},\beta_{j1}} \sum_{i=1}^{n} \ell(Y_i, \beta_{j0} + \beta_{j1} X_{ij}), \tag{11}$$

Similar to the marginal least squares estimate for model (9), it is reasonable to consider $|\widehat{\beta}_{j1}^M|$ as a marginal utility to rank the importance of X_j and select a submodel for a given prespecified threshold γ_n,

$$\widehat{\mathcal{M}}_{\gamma_n} = \{1 \le j \le p : |\widehat{\beta}_{j1}^M| \ge \gamma_n\}. \tag{12}$$

This was proposed for feature screening in generalized linear model by Fan and Song (2010)

Feature screening for generalized linear models

To establish the theoretical properties of MMLE, Fan and Song (2010) defined the population version of the marginal likelihood maximizer as

$$\boldsymbol{\beta}_j^M = (\beta_{j0}^M, \beta_{j1}^M) = \arg\min_{\beta_{j0},\beta_{j1}} E\ell(Y, \beta_{j0} + \beta_{j1}X_j).$$

They first showed that $\beta_{j1}^M = 0$ if and only if $\text{cov}(Y, X_j) = 0$ for $j = 1, \ldots, p$. Thus, when the important variables are correlated with the response, $\beta_{j1}^M \neq 0$.

Define the true model as $\mathcal{M}_* = \{1 \leq j \leq p : \beta_j \neq 0\}$ with the size $s = |\mathcal{M}_*|$. Fan and Song (2010) showed that under some conditions if $|\text{cov}(Y, X_j)| \geq c_1 n^{-\kappa}$ for $j \in \mathcal{M}_*$ and some $c_1 > 0$, then $\min_{j\in\mathcal{M}_*} |\beta_{j1}^M| \geq c_2 n^{-\kappa}$, for some $c_2, \kappa > 0$. Thus, the marginal signals β_{j1}^M's are stronger than the stochastic noise provided that X_j's are marginally correlated with Y.

Feature screening for generalized linear models

Under some technical assumptions, Fan and Song (2010) proved that if $n^{1-2\kappa}/(k_n^2 K_n^2) \to \infty$, where $k_n = b'(K_n B + B) + m_0 K_n^{\alpha}/s_0$, B is the upper bound of the true value of β_j^M, K_n is the supremum norm of $\mathbf{x}$, then for any $c_3 > 0$, there exists some $c_4 > 0$ such that

$$\begin{aligned} & P\left(\max_{1\le j\le p} |\widehat{\beta}_{j1}^M - \beta_{j1}^M| \ge c_3 n^{-\kappa}\right) \\ \le \quad & p\left\{\exp(-c_4 n^{1-2\kappa}/(k_n K_n)^2) + n m_1 \exp(-m_0 K_n^{\alpha})\right\}. \end{aligned}$$

In addition, assume that $|\mathrm{cov}(Y, X_j)| \ge c_1 n^{-\kappa}$ for $j \in \mathcal{M}_*$ and take $\gamma_n = c_5 n^{-\kappa}$ with $c_5 \le c_2/2$, the following inequality holds,

$$\begin{aligned} & P\left(\mathcal{M}_* \subseteq \widehat{\mathcal{M}}_{\gamma_n}\right) \qquad (13) \\ \ge \quad & 1 - s\left\{\exp(-c_4 n^{1-2\kappa}/(k_n K_n)^2) + n m_1 \exp(-m_0 K_n^{\alpha})\right\} \to 1, \end{aligned}$$

as $n \to \infty$.

Feature screening under a general parametric framework

Suppose that we are interested in exploring the relationship between $\mathbf{x}$ and Y. A general statistical framework is to minimize an objective function

$$Q(\beta_0, \boldsymbol{\beta}) = \sum_{i=1}^{n} L(Y_i, \beta_0 + \boldsymbol{\beta}^T \mathbf{x}_i), \tag{14}$$

where $L(\cdot, \cdot)$ is a loss function, and $\boldsymbol{\beta}$ is a regression coefficient and is assumed to be sparse. By taking different loss functions, many commonly-used statistical frameworks can be unified under (14).

Examples

Linear models. Let $L(Y_i, \beta_0 + \boldsymbol{\beta}^T \mathbf{x}_i) = (Y_i - \beta_0 - \boldsymbol{\beta}^T \mathbf{x}_i)^2$. $\Rightarrow$ the least squares method for a linear regression model.

Generalized linear models. Let $L(Y_i, \beta_0 + \boldsymbol{\beta}^T \mathbf{x}_i)$ be the negative logarithm of (quasi)-likelihood function of Y_i given $\mathbf{x}_i$, $\ell(Y_i, \beta_0 + \boldsymbol{\beta}^T \mathbf{x}_i)$, $\Rightarrow$ the likelihood method for generalized linear models.

Quantile linear regression. In the presence of heteroscedastic errors, people may consider quantile regression instead of the least squares estimate. For a given quantile level $\alpha \in (0, 1)$, define $\rho_\alpha(u) = u\{\alpha - I(u < 0)\}$, the quantile loss function, where $I(A)$ is the indicator function of a set A.

Taking $L(Y_i, \beta_0 + \boldsymbol{\beta}^T \mathbf{x}_i) = \rho_\alpha(Y_i - \beta_0 - \boldsymbol{\beta}^T \mathbf{x}_i) \Rightarrow$ the quantile regression. E.g., $\alpha = 1/2 \Rightarrow$ the median or the least absolute deviation regression.

Examples (continued)

Robust linear regression. In the presence of outliers or heavy-tailed errors, robust linear regression has been proved to be more suitable than the least squares method. The commonly-used loss function includes the L_1-loss: $\rho_1(u) = |u|$, the Huber loss function: $\rho_2(u) = (1/2)|u|^2 I(|u| \leq \delta) + \delta\{|u| - (1/2)\delta\} I(|u| > \delta)$ (Huber, 1964). Setting $L(Y_i, \beta_0 + \boldsymbol{\beta}^T \mathbf{x}_i) = \rho_k(Y_i - \beta_0 + \boldsymbol{\beta}^T \mathbf{x}_i)$, $k = 1$, or 2 leads a robust linear regression.

Classification. In machine learning, the response variable typically is set to be the input class label such as $Y_1 = \{-1, +1\}$ as in the classification method in the statistical literature. The hinge loss function is defined to be $h(u) = \{(1 - u) + |1 - u|\}/2$ (Vapnik, 1995). Taking $L(Y_i, \beta_0 + \boldsymbol{\beta}^T \mathbf{x}_i) = h(Y_i - \beta_0 - \boldsymbol{\beta}^T \mathbf{x}_i)$ in (14) leads a classification rule.

A general marginal utility

Similar to RSS_j defined in the beginning of this section, a natural marginal utility of the jth predictor is

$$L_j = \min_{\beta_{j0},\beta_{j1}} \frac{1}{n} \sum_{i=1}^{n} L(Y_i, \beta_{j0} + X_{ij}\beta_{j1}).$$

According to the definition, the smaller the value L_j, the more important the corresponding jth predictor. This was first proposed in Fan, Samworth and Wu (2009), in which numerical studies clearly demonstrates the potential of this marginal utility.

Iterative screening procedure

The marginal feature screening methodology may fail if the predictor is marginally uncorrelated but jointly related with the response, or jointly uncorrelated with the response but has higher marginal correlation than some important features. Fan, Samworth and Wu (2009) proposed an iterative feature screening procedure under the general statistical framework (14).

The proposed iterative procedure consists of the following steps.

S1. Compute the vector of marginal utilities $(L_1, \ldots, L_p)$ and select the set $\widehat{\mathcal{A}}_1 = \{1 \leq j \leq p : L_j \text{ is among the first } k_1 \text{ smallest of all}\}$. Then apply a penalized method such as the LASSO (Tibshirani, 1996) and SCAD (Fan and Li, 2001) to select a subset $\widehat{\mathcal{M}}$.

Iterative screening procedure (continued)

S2. Compute, for each $j \in \{1, \ldots, p\}/\widehat{\mathcal{M}}$,

$$L_j^{(2)} = \min_{\beta_0, \boldsymbol{\beta}_{\widehat{\mathcal{M}}}, \beta_j} \frac{1}{n} \sum_{i=1}^{n} L(Y_i, \beta_0 + \mathbf{x}^T{}_{i,\widehat{\mathcal{M}}} \boldsymbol{\beta}_{\widehat{\mathcal{M}}} + X_{ij}\beta_j),$$

where $\mathbf{x}_{i,\widehat{\mathcal{M}}}$ denotes the sub-vector of $\mathbf{x}_i{}^T$ consisting of those elements in $\widehat{\mathcal{M}}$. $L_j^{(2)}$ can be considered as the additional contribution of the jth predictor given the existence of predictors in $\widehat{\mathcal{M}}$. Then, select the set

$$\widehat{\mathcal{A}}_2 = \{j \in \{1, \ldots, p\}/\widehat{\mathcal{M}} : L_j^{(2)} \text{ is among the first } k_2 \text{ smallest of all}\}.$$

Iterative screening procedure (continued)

S3. Employ a penalized (pseudo-)likelihood method such as the LASSO and SCAD on the combined set $\widehat{\mathcal{M}} \bigcup \widehat{\mathcal{A}}_2$ to select a updated selected set $\widehat{\mathcal{M}}$. That is, use the penalized likelihood to obtain

$$\widehat{\boldsymbol{\beta}}_2 = \underset{\beta_0, \boldsymbol{\beta}_{\widehat{\mathcal{M}}}, \boldsymbol{\beta}_{\widehat{\mathcal{A}}_2}}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} L(Y_i, \beta_0 + \mathbf{x}^T{}_{i,\widehat{\mathcal{M}}} \boldsymbol{\beta}_{\widehat{\mathcal{M}}} + \mathbf{x}^T{}_{i,\widehat{\mathcal{A}}_2} \boldsymbol{\beta}_{\widehat{\mathcal{A}}_2}) + \sum_{j \in \widehat{\mathcal{M}} \bigcup \widehat{\mathcal{A}}_2} p_\lambda(|\beta_j|),$$

and the indices set of $\widehat{\boldsymbol{\beta}}_2$ that are none-zero yield a new updated set $\widehat{\mathcal{M}}$.

S4. Repeat S2 and S3 until $|\widehat{\mathcal{M}}| \leq d$, where d is the prescribed number and $d \leq n$. The indices set $\widehat{\mathcal{M}}$ is the final selected submodel.

This iterative procedure extends the Iterative SIS, without explicit definition of residuals, to a general statistical framework. It also allows the procedure to delete predictors from the previously selected set.

Nonparametric regression models

Nonparametric models become very useful in the absence of priori information about the model structure for the regression and may be used to enhance the model flexibility, especially for the ultrahigh dimensional data with much challenge to check model assumptions. Therefore, the feature screening techniques for the nonparametric models naturally drew great attention from researchers.

Additive model

$$Y = \sum_{j=1}^{p} m_j(X_j) + \varepsilon, \tag{15}$$

where $\{m_j(X_j)\}_{j=1}^{p}$ have mean 0 for identifiability.

An intuitive population level marginal screening utility is $E(f_j^2(X_j))$, where $f_j(X_j) = E(Y|X_j)$ is the projection of Y onto X_j.

With the sample $\{(\mathbf{x}_i, Y_i), i = 1, \ldots, n\}$, $f_j(x)$ can be estimated via a normalized B-spline basis $\mathbf{B}_j(x) = \{B_{j1}(x), \ldots, B_{jd_n}(x)\}^{\mathrm{T}}$:

$$\widehat{f}_{nj}(x) = \widehat{\boldsymbol{\beta}}_j{}^T \mathbf{B}_j(x), \quad 1 \le j \le p, \tag{16}$$

where

$$\widehat{\boldsymbol{\beta}}_j = \underset{\boldsymbol{\beta}_j \in \mathbb{R}^{d_n}}{\operatorname{argmin}} \sum_{i=1}^{n} (Y_i - \boldsymbol{\beta}_j{}^T \mathbf{B}_j(X_{ij})).$$

Additive model (continued)

The screened model index set is

$$\widehat{\mathcal{M}}_\nu = \{1 \le j \le p : \|\widehat{f}_{nj}\|_n^2 \ge \nu_n\}, \tag{17}$$

for some predefined threshold value ν_n, with $\|\widehat{f}_{nj}\|_n^2 = n^{-1}\sum_{i=1}^n \widehat{f}_{nj}(X_{ij})^2$.

Fan, Feng and Song (2011) define the true model $\mathcal{M}_* = \{j : Em_j(X_j)^2 > 0\}$, Under a set of conditions, they showed that

$$P\left(\mathcal{M}_* \subset \widehat{\mathcal{M}}_\nu\right) \to 1 \tag{18}$$

for $p = \exp\{n^{1-4\kappa}d_n^{-3} + nd_n^{-3}\}$.

Varying coefficient models (VCM)

Varying coefficient models are another class of popular nonparametric regression models in the literature and defined by

$$Y = \sum_{j=1}^{p} \beta_j(u) X_j + \varepsilon, \tag{19}$$

where $\beta_j(u)$'s are coefficient functions. An intercept can be included by setting $X_1 \equiv 1$.
Fan, Ma and Dai (2014) extended the screening procedure for the additive model to this varying coefficient model.
From a different perspective, Liu, Li and Wu (2014) proposed a conditional correlation screening procedure based on the kernel regression approach.

Fan, Ma and Dai (2014)'s method

Fan, Ma and Dai (2014) started from the marginal regression problem for each X_j, $j = 1, \ldots, p$, i.e., finding $a_j(u)$ and $b_j(u)$ to minimize $E\{(Y - a_j(u) - b_j(u)X_j)^2|u\}$.

Thus, the marginal contribution of X_j for predicting Y can be described as

$$u_j = E\{a_j(u) + b_j(u)X_j\}^2 - E\{a_0(u)^2\},$$

where $a_0(u) = E(Y|u)$.

With the sample $\{(u_i, \mathbf{x}_i, Y_i), i = 1, \ldots, n\}$, where $\mathbf{x}_i = (X_{i1}, \ldots, X_{ip})^T$, $a_j(u), b_j(u)$ and $a_0(u)$ can be estimated based on the normalized B-spline basis $\mathbf{B}(u) = \{B_1(u), \ldots, B_{d_n}(u)\}^T$.

Fan, Ma and Dai (2014)'s method

Therefore, the sample marginal utility for screening is

$$\widehat{\mu}_{nj} = \|\widehat{a}_j(u) + \widehat{b}_j(u)X_j\|_n^2 - \|\widehat{a}_0(u)\|_n^2, \tag{20}$$

where $\|\cdot\|_n^2$ is the sample average, defined in the same fashion as in the last section. The selected model is then $\widehat{\mathcal{M}}_\tau = \{1 \le j \le p : \widehat{u}_{nj} \ge \tau_n\}$ for the pre-specified data-driven threshold τ_n.

The authors proved the sure screening property of the proposed method under similar regularity conditions with Fan, Feng and Song (2011).

Liu, Li and Wu (2014)'s method

Liu, Li and Wu (2014) proposed another sure independent screening procedure for the ultrahigh dimensional varying coefficient model (19) based on the conditional correlation learning (CC-SIS).
Since VCM is indeed a linear model when conditioning on u, the conditional correlation between each predictor X_j and Y given u is defined similarly as the Pearson correlation:

$$\rho(X_j, Y|u) = \frac{\text{cov}(X_j, Y|u)}{\sqrt{\text{cov}(X_j, X_j|u)\text{cov}(Y, Y|u)}},$$

where $\text{cov}(X_j, Y|u) = E(X_j Y|u) - E(X_j|u)E(Y|u)$, and the population level marginal utility to evaluate the importance of X_j is $\rho^*_{j0} = E\{\rho^2(X_j, Y|u)\}$.

Liu, Li and Wu (2014)'s method

Liu, Li and Wu (2014) proposed kernel regression to estimate ρ_{j0}^*, or equivalently, the conditional means $E(Y|u)$, $E(Y^2|u)$, $E(X_j|u)$, $E(X_j^2|u)$ and $E(X_jY|u)$, with sample $\{(u_i, \mathbf{X}_i, Y_i), i = 1, \ldots, n\}$. For instance,

$$\widehat{E}(Y|u) = \sum_{i=1}^{n} \frac{K_h(u_i - u)Y_i}{\sum_{i=1}^{n} K_h(u_i - u)},$$

where $K_h(t) = h^{-1}K(t/h)$ is a rescaled kernel function. We can further obtain $\widehat{\rho}^2(X_j, Y|u_i)$, and $\widehat{\rho}_{j0}^*$:

$$\widehat{\rho}_j^* = \frac{1}{n}\sum_{i=1}^{n} \widehat{\rho}^2(X_j, Y|u_i). \tag{21}$$

Liu, Li and Wu (2014)'s method

The screened model is defined as

$$\widehat{\mathcal{M}} = \{j : 1 \leq j \leq p,\ \widehat{\rho}_j^* \text{ ranks among the first } d\},$$

with $d = [n^{4/5}/\log(n^{4/5})]$ follows the hard threshold by Fan and Lv (2008) but modified for the nonparametric regression.

Define $\mathcal{M}_* = \{1 \leq j \leq p :\ \beta_j(u) \neq 0 \text{ for some } u\}$ is the true model.

With these additional conditions, the authors also systematically studied the ranking consistency (Zhu, Li, Li and Zhu, 2011):

$$\liminf_{n\to\infty}\{\min_{j\in\mathcal{M}_*} \widehat{\rho}_j^* - \max_{j\in\mathcal{M}_*^c} \widehat{\rho}_j^*\} > 0 \quad \text{in probability.} \tag{22}$$

That is, with an overwhelming probability, the truly important predictors have larger $\widehat{\rho}_j^*$ than the unimportant ones.

Joint screening

1. **Forward selection**

2. **Sparse MLE**

Forward selection

Another way to cope with the situation with some individual covariates being jointly correlated but marginally uncorrelated with the response is the forward regression.

Wang (2009) carefully studied the property of forward regression with ultrahigh dimensional predictors. To achieve the sure screening property, the forward regression requires that there exist two positive constants $0 < \tau_1 < \tau_2 < \infty$ such that $\tau_1 < \lambda_{\min}(\text{cov}(\mathbf{x})) \leq \lambda_{\max}(\text{cov}(\mathbf{x})) < \tau_2$. Wang (2009) further proposed using the extended BIC (Chen and Chen, 2008) to determine the size of the active predictor set.

Weakness: computational cost would be high because of computing inverse of submatrix of $\mathbf{X}^T\mathbf{X}$.

Sure joint screening

Xu and Chen (2014, JASA) proposed sparse MLE as follows:
Let $\ell_n(\boldsymbol{\beta}) = \sum_{i=1}^{m}\{(\mathbf{x}_i^T\boldsymbol{\beta})y_i - b(\mathbf{x}_i^T\boldsymbol{\beta})\}$ the negative logarithm of likelihood of a random sample from generalized linear model with canonical link.
Sparse MLE of $\boldsymbol{\beta}$ is defined by

$$\widehat{\boldsymbol{\beta}}_k = \arg\max_{\boldsymbol{\beta}} \ell_n(\boldsymbol{\beta}) \quad \text{subject to} \quad \|\boldsymbol{\beta}_0\| \leq k$$

The corresponding minimization problem requires combinatorial optimization.

The key is to develop an algorithm to solve the problem.

Iterative Hard-Thresholding (IHT) Algorithm

IHT algorithm starts with approximating $\ell_n(\cdot)$ at a genetic $\boldsymbol{\beta}$ by

$$h_n(\boldsymbol{\gamma}|\boldsymbol{\beta}) = \ell_n(\boldsymbol{\beta}) + (\boldsymbol{\gamma} - \boldsymbol{\beta})^T \mathbf{S}_n(\boldsymbol{\beta}) - \frac{u}{2}\|\boldsymbol{\gamma} - \boldsymbol{\beta}\|_2^2 \tag{23}$$

for some scale parameter $u > 0$, where $\mathbf{S}_n(\boldsymbol{\beta}) = \ell_n'(\boldsymbol{\beta})$ is the score function.

Authors explain: the first two terms in (23) match the Taylor's expansion of $\ell_n(\boldsymbol{\beta})$ at $\boldsymbol{\beta}$, and the $\frac{u}{2}\|\boldsymbol{\gamma} - \boldsymbol{\beta}\|_2^2$ is a regularization term.

Using (23), an approximate solution to the original minimization problem is obtained by the following iterative procedure

$$\boldsymbol{\beta}^{(t+1)} = \arg\max_{\boldsymbol{\gamma}} h_n(\boldsymbol{\gamma}|\boldsymbol{\beta}^{(t)}) \quad \text{subject to} \quad \|\boldsymbol{\gamma}\|_0 \leq k.$$

There is a closed form solution for this minimization problem.

Denote $\boldsymbol{\gamma}^{(t)} = \boldsymbol{\beta}^{(t)} + u^{-1}\mathbf{S}_n(\boldsymbol{\beta}^{(t)})$. Sort elements of $|\widehat{\boldsymbol{\gamma}}|$ so that

$$|\gamma_{(1)}^{(t)}| \geq |\gamma_{(2)}^{(t)}| \geq \cdots \geq |\gamma_{(d)}^{(t)}|.$$

Then

$$\beta_j^{(t+1)} = \gamma_j^{(t)} I\{|\gamma_j^{(t)}| \geq |\gamma_{(k)}^{(t)}|\}$$

1. Start with $\boldsymbol{\beta}^{(0)} = \mathbf{0}$, the first step for linear regression model is corresponding to the correlation learning.

2. With proper choice of u, the algorithm may have ascent property.

3. The authors also established the oracle property.

One may consider another approximation to the logarithm of likelihood function:

$$h_n(\gamma|\beta) = \ell_n(\beta) + (\gamma - \beta)^T \mathbf{S}_n(\beta) - \frac{u}{2}(\gamma - \beta)^T \mathbf{W}(\gamma - \beta) \qquad (24)$$

where $\mathbf{W}$ is a weight matrix. How to choose $\mathbf{W}$?

Intuitively,

$$\ell_n(\gamma) = \ell_n(\beta) + (\gamma - \beta)^T \mathbf{S}_n(\beta) + \frac{1}{2}(\gamma - \beta)^T H(\beta^*)(\gamma - \beta)$$

where $H(\beta^*) = \ell''(\gamma)/\partial\gamma\partial\gamma^T|_{\gamma=\beta^*}$, the Hessian matrix of log-likelihood, and β^* lies between β and γ.

Take $\mathbf{W} = \text{diag}\{-H(\beta)\}$

More details can be found at Yang, Yu, Li and Buu (2016, Statistica Sinica).

3. Model-Free Feature Screening Procedures

It may be difficult to specify a regression model for ultrahigh dimensional data

It is desirable to develop screening procedures that work well for a very general model setting, and have comparable performance to the model-based SIS.

One approach: allow the underlying model very flexible to include most commonly-used regression models.

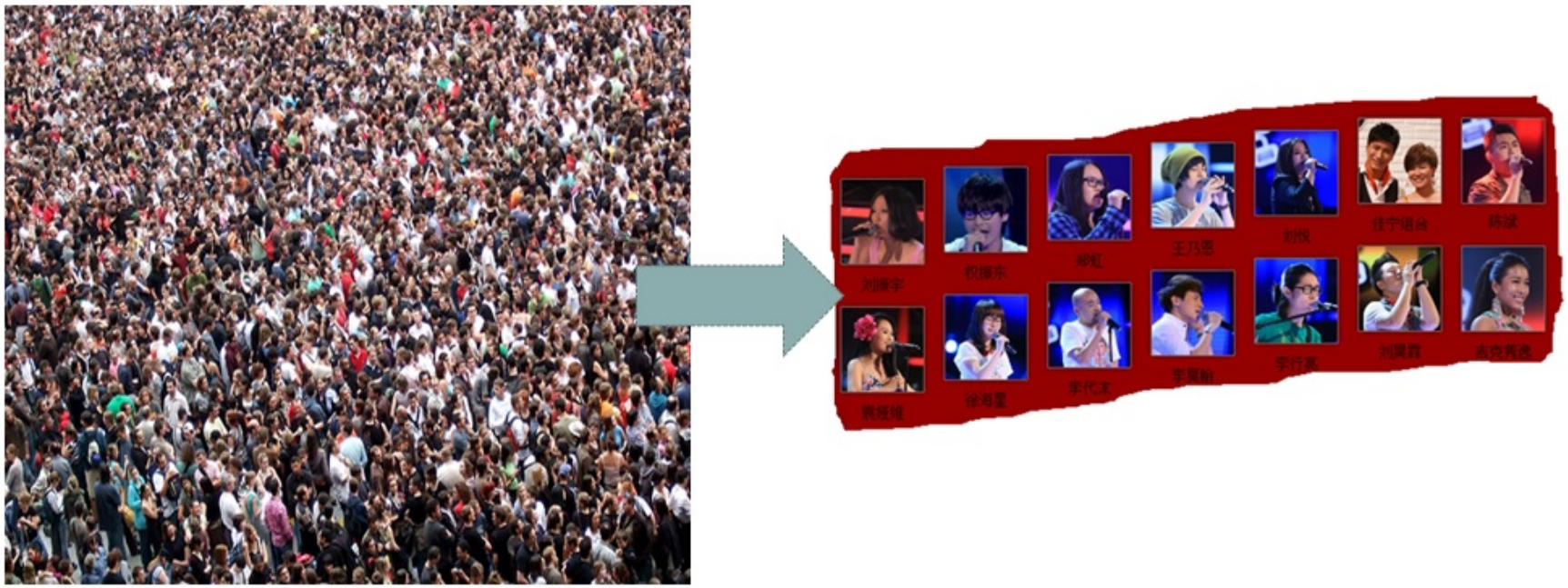
Two-stage Approach: Intuition

Huan Liu: *How do I choose several potentially great singers?*

Unlike the google map shown before, we don't have any ideas about all singers.

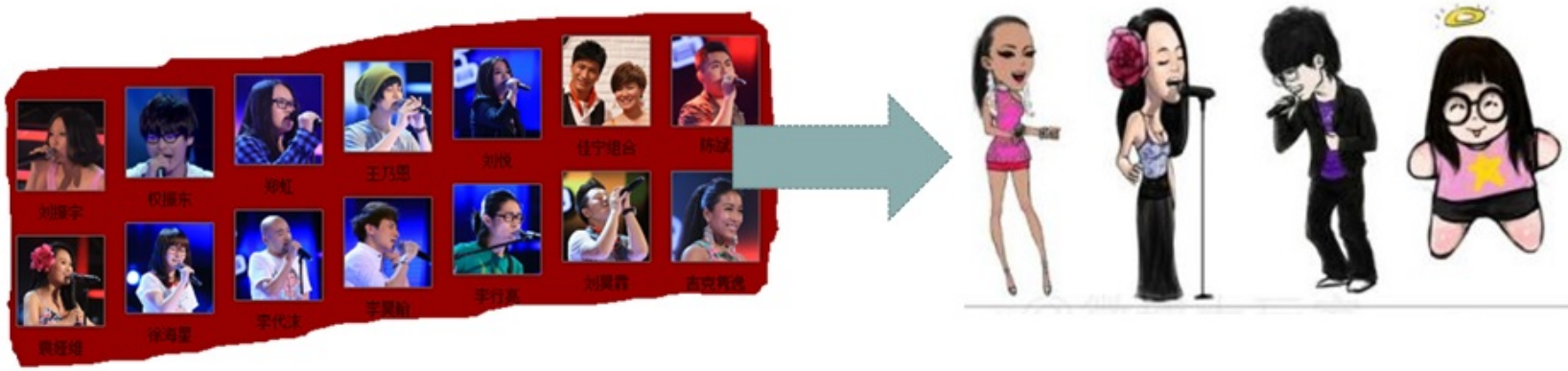
Two-stage Approach: Intuition

(1) Independence Screening Stage:
Score each person and pick the top 16 potential singers
"Quick and Dirty" way



Two-stage Approach: Intuition

(2) Post-Screening Selection Stage:
Choose the final 4 out of the 16 potential singers
Carefully and wisely



3.1 Sure Independence Ranking Screening (SIRS)

Material in this section is based on

Zhu, L, Li, L., Li, R. and Zhu, L.-X. (2011). Model-free feature screening for ultrahigh dimensional data. *Journal of American Statistical Association*. **106**, 1464 - 1475.

General Model Framework:

Y: the response variable with support Ψ_y,

Y can be both univariate and multivariate.

$\mathbf{x} = (X_1, \cdots, X_p)^T$: the feature/predictor/covariate vector.

$F(y \mid \mathbf{x})$: the conditional distribution function of Y given $\mathbf{x}$.

Define two index sets:

$$\begin{aligned} \mathcal{A} &= \{k : F(y \mid \mathbf{x}) \text{ functionally depends on } X_k \text{ for some } y \in \Psi_y\}, \\ \mathcal{I} &= \{k : F(y \mid \mathbf{x}) \text{ does not functionally depend on } X_k \text{ for any } y \in \Psi_y\}. \end{aligned}$$

If $k \in \mathcal{A}$, X_k is referred to as an active predictor, whereas if $k \in \mathcal{I}$, X_k is referred to as an inactive predictor. Let $\mathbf{x}_{\mathcal{A}}$, a $p_1 \times 1$ vector, consist of all X_k's with $k \in \mathcal{A}$. Similarly, let $\mathbf{x}_{\mathcal{I}}$, a $(p - p_1) \times 1$ vector, consist of all inactive predictors X_k with $k \in \mathcal{I}$.

Feature screening for multi-index models

The most general model for $F(y \mid \mathbf{x})$ is

$$F(y \mid \mathbf{x}) = F_0(y \mid \mathbf{x}_{\mathcal{A}}) \tag{A}$$

where $F_0(\cdot|\cdot)$ is an unknown function.
In practice, $F(y \mid \mathbf{x})$ depends on $\mathbf{x}$ only through $\mathbf{B}^T\mathbf{x}_{\mathcal{A}}$ for some $p_1 \times K$ constant **matrix** $\mathbf{B}$. Thus, we may consider

$$F(y \mid \mathbf{x}) = F_0(y \mid \mathbf{B}^T\mathbf{x}_{\mathcal{A}}), \tag{B}$$

where $F_0(\cdot|\cdot)$ is an unknown function.
Identifiability: not concerned here.
If we take $\mathbf{B} = I_{p_1}$, the identity matrix, then (A) $\equiv$ (B).

In (B), $K = 1$, 2 or 3, and may be much smaller than p_1. Model in (B) allows us to impose weaker conditions to establish the theoretic property of the proposed procedure.

Examples for continuous response

Models for a continuous response:

$$h(Y) = f_1(\boldsymbol{\alpha}_1^T \mathbf{x}_A) + \boldsymbol{\alpha}_2^T \mathbf{x}_A + f_2(\boldsymbol{\alpha}_3^T \mathbf{x}_A)\varepsilon, \tag{25}$$

where ε is independent of $\mathbf{x}$, $h(\cdot)$ is a monotone function, $f_2(\cdot)$ is a positive function, $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2$,and $\boldsymbol{\alpha}_3$ are unknown coefficients, and Here $h(\cdot), f_1(\cdot)$ and $f_2(\cdot)$ may be either known or unknown.

Model (25) is a special case of model (B).

Model (25) with $h(Y) = Y$ includes: linear model, partially linear model, single-index model, and partially linear single-index model.

Transformation regression model

Model (25) also includes the transformation regression model for a general transformation $h(Y)$.

In survival data analysis, the response Y is the time to event of interest, and a commonly used model for Y is the accelerated failure time model:

$$\log(Y) = \alpha_0 + \boldsymbol{\alpha}_1^T \mathbf{x}_A + \varepsilon,$$

where ε is independent of $\mathbf{x}$.

Examples for discrete response

Generalized partially linear single-index model for discrete responses such as binary outcomes and count responses:

$$g_1\{E(Y|\mathbf{x})\} = g_2(\boldsymbol{\alpha}_1^T \mathbf{x}_A) + \boldsymbol{\alpha}_2^T \mathbf{x}_A, \tag{26}$$

where $Y|\mathbf{x} \sim$ the exponential family, $g_1(\cdot)$ is a link function, $g_2(\cdot)$ is an unknown function,

In summary, many existing models with various types of response variables can be cast into the common model framework of (B).

The proposed feature screening approach developed under (B) offers a unified solution that works for a wide range of existing models.

Transformation regression model

Model (25) also includes the transformation regression model for a general transformation $h(Y)$.

In survival data analysis, the response Y is the time to event of interest, and a commonly used model for Y is the accelerated failure time model:

$$\log(Y) = \alpha_0 + \boldsymbol{\alpha}_1^T \mathbf{x}_A + \varepsilon,$$

where ε is independent of $\mathbf{x}$.

Examples for discrete response

Generalized partially linear single-index model for discrete responses such as binary outcomes and count responses:

$$g_1\{E(Y|\mathbf{x})\} = g_2(\boldsymbol{\alpha}_1^T \mathbf{x}_A) + \boldsymbol{\alpha}_2^T \mathbf{x}_A, \tag{27}$$

where $Y|\mathbf{x} \sim$ the exponential family, $g_1(\cdot)$ is a link function, $g_2(\cdot)$ is an unknown function,

In summary, many existing models with various types of response variables can be cast into the common model framework of (B).

The proposed feature screening approach developed under (B) offers a unified solution that works for a wide range of existing models.

Motivation

WLOG, assume that $E(X_k) = 0$ and $\text{var}(X_k) = 1$ for $k = 1, \ldots, p$.
Denote $\boldsymbol{\Omega}(y) = E\{\mathbf{x}F(y|\mathbf{x})\}$. Then

$$\boldsymbol{\Omega}(y) = \text{E}[XE\{I(Y < y)|\mathbf{x}\}] = \text{cov}\{\mathbf{x}, I(Y < y)\}.$$

Let $\Omega_k(y)$ be the k-th element of $\boldsymbol{\Omega}(y)$, and define

$$\omega_k = E\left\{\Omega_k^2(Y)\right\}, \quad k = 1, \ldots, p. \tag{28}$$

Gaussian Predictors: $\mathbf{x} \sim N_p(\mathbf{0}, \sigma^2 I_p)$ and $K = 1$.
Write $\mathbf{x} = (\mathbf{x}_A^T, \mathbf{x}_I^T)^T$, and $\mathbf{b} = (b_1, \ldots, b_p)^T = (\mathbf{B}^T, \mathbf{0}^T)^T$. Then

$$\boldsymbol{\Omega}(y) = E\{\mathbf{x}F_0(y|\mathbf{b}^T\mathbf{x})\} = c(y)\mathbf{b},$$

for a $c(y)$. Then $\omega_k = E\{c^2(Y)\}b_k^2$. If $E\{c^2(Y)\} > 0$, then

$$\omega_k = 0 \Leftrightarrow k \in I.$$

$\Rightarrow \omega_k$ may be used for feature screening in this setting.

Model free feature screening method

Let $\mathbf{1}(\cdot)$ be an indicator function. Our procedure is described as follows.

1. Define $r_j(y) = \mathrm{cov}\Big\{X_j, \mathbf{1}(Y < y)\Big\}$ for any fixed y. Calculate $\widehat{\omega}_j \widehat{=} \mathrm{E}_n\{r_j^2(Y)\} = n^{-1}\sum_{i=1}^n r_j^2(Y_i)$.
2. Sort $\widehat{\omega}_j$ in descending order: $\widehat{\omega}_{(1)} \geq \widehat{\omega}_{(2)} \geq \cdots \geq \widehat{\omega}_{(p)}$.
3. Let $N = [n/\log n]$, keep $\Big\{X_{(1)}, X_{(2)}, \cdots, X_{(N)}\Big\}$, screening out $\Big\{X_{(N+1)}, \cdots, X_{(p)}\Big\}$.

Technical Conditions:

(C1) The following inequality condition holds uniformly for p:

$$\frac{K^2\lambda_{\max}\left\{\mathrm{Cov}(\mathbf{x}_{\mathcal{A}},\mathbf{x}_{\mathcal{I}}{}^T)\mathrm{Cov}(\mathbf{x}_{\mathcal{I}},\mathbf{x}_{\mathcal{A}}{}^T)\right\}}{\lambda^2_{\min}\left\{\mathrm{Cov}(\mathbf{x}_{\mathcal{A}},\mathbf{x}_{\mathcal{A}}{}^T)\right\}} < \frac{\min_{k\in\mathcal{A}}\omega_k}{\lambda_{\max}\{\mathbf{\Omega}_{\mathcal{A}}\}}, \tag{29}$$

where $\mathbf{\Omega}_{\mathcal{A}} = \mathrm{E}\left\{\mathbf{\Omega}_{\mathcal{A}}(Y)\mathbf{\Omega}_{\mathcal{A}}(Y)^T\right\}$, $\mathbf{\Omega}_{\mathcal{A}}(y) = \{\Omega_1(y),\cdots,\Omega_{p_1}(y)\}^T$, and $\lambda_{\max}\{\mathbf{A}\}$ and $\lambda_{\min}\{\mathbf{A}\}$ denote the largest and smallest eigenvalues of a matrix $\mathbf{A}$, respectively. Note that $\lambda_{\min}(\mathbf{A})$ and $\lambda_{\max}(\mathbf{A})$ may depend on the dimension of $\mathbf{A}$. Throughout this article, when we say that "$a < b$ holds uniformly for p", it means that $\limsup_{p\to\infty}\{a(p) - b(p)\} < 0$.

(C2) The linearity condition:

$$\mathrm{E}\left\{\mathbf{x}|\mathbf{B}^T\mathbf{x}_A\right\} = \mathrm{Cov}\left(\mathbf{x},\mathbf{x}_A^T\right)\mathbf{B}\left\{\mathrm{Cov}(\mathbf{B}^T\mathbf{x}_A)\right\}^{-1}\mathbf{B}^T\mathbf{x}_A.$$

(C3) The moment condition: $\exists$ a constant $t_0 > 0$ such that

$$\max_{1\le k\le p}\mathrm{E}\left\{\exp(tX_k)\right\} < \infty, \text{ for } 0 < t \le t_0.$$

Condition (C1)

Condition (C1) dictates the correlations among the predictors, and is the key assumption to ensure that the proposed screening procedure works properly. We make the following remarks about this condition.
First, as the dimension K of $\mathbf{B}$ in (B) increases, the condition becomes more stringent. Therefore, a model with a small K is favored by our procedure. In many commonly used models, however, K is indeed small.

Second, for the left hand side of equation in Condition (C1), the numerator measures the correlation between the active predictors $\mathbf{x}_{\mathcal{A}}$ and the inactive ones $\mathbf{x}_{\mathcal{I}}$, while the denominator measures the correlation among the active predictors themselves. When $\mathbf{x}_{\mathcal{A}}$ and $\mathbf{x}_{\mathcal{I}}$ are uncorrelated, (C1) holds automatically.
For the proposed screening method to work well, this condition rules out the case in which there is strong collinearity between the active and inactive predictors, or among the active predictors themselves. This is very similar to Condition 4 of Fan and Lv (2008).

Third, the quantity $\min_{k \in \mathcal{A}} \omega_k$ on the right hand side of (29) reflects the signal strength of individual active predictors, which in turn controls the rate of probability error in selecting the active predictors. This aspect is similar to Condition 3 of Fan and Lv (2008), which requires the contribution of an active predictor to be sufficiently large.

Finally, we note that (29) is not scale invariant, since $\mathbf{\Sigma} = \text{Cov}(\mathbf{x}, \mathbf{x}^T)$ is not taken into account. This is similar to the linear SIS procedure of Fan and Lv (2008), which is based upon the covariance vector $\text{Cov}(\mathbf{x}, Y)$ alone without the term $\mathbf{\Sigma}$. Fan and Lv (2008) imposed the concentration property that implicitly requires the marginal variances of all predictors be of the same order. In our setup, we always marginally standardize all the predictors to have sample variance equal to one.

Condition (C2)

Condition (C2) holds if $\mathbf{x}$ follows a normal or an elliptically symmetric distribution. This condition was first proposed by Li (1991) and has been widely used in the dimension-reduction literature.

It is remarkable though that Condition (C2) is itself weaker than both the normality and the elliptical symmetry conditions because we only require it to hold for the true value of $\boldsymbol{\beta}$.

Furthermore, Hall and Li (1993) showed that the linearity condition holds asymptotically if the number of predictors p diverges while the dimension K remains fixed. For this reason, we view the linearity condition as a mild assumption in ultrahigh-dimensional regressions, where p is essentially very large and grows at a fast rate towards infinity.

Condition (C3)

Condition (C3) is concerned with the moments of the predictors, which assumes that all moments of the predictors are uniformly bounded. This condition holds for a variety of distributions, including the normal distribution and the distributions with bounded support. Compared with the usual conditions imposed in the feature screening literature, (C3) relaxed the normality assumption assumed by Fan and Lv (2008), in which both $\mathbf{x}$ and $Y \mid \mathbf{x}$ are assumed to be normally distributed.

Theorem 1. Under Conditions (C1)–(C3), the following inequality holds uniformly for p:

$$\max_{k\in\mathcal{I}} \omega_k < \min_{k\in\mathcal{A}} \omega_k. \tag{30}$$

Corollary 1. Under the linearity condition (C2) and for $k = 1, \cdots, p$, $\omega_k = 0$ if and only if $\text{cov}(\boldsymbol{\beta}^T \mathbf{x}_{\mathcal{A}}, X_k) = \mathbf{0}$.

Theorem 1 and Corollary 1 together offer more insights into the newly proposed utility measure ω_k. First, it is easy to see that, when X_k is independent of Y, $\omega_k = 0$. On the other hand, $k \in \mathcal{I}$ alone does not necessarily imply that $\omega_k = 0$. The quantity is zero only if X_k is uncorrelated with $\boldsymbol{\beta}^T \mathbf{x}_{\mathcal{A}}$. Theorem 1, however, ensures that ω_k of an inactive predictor is always smaller than ω_k of an active predictor, which is sufficient for the purpose of predictor ranking.

Consistency in Ranking

Theorem 2. In addition to the conditions in Theorem 1, we further assume that $p = o\{\exp(an)\}$ for any fixed $a > 0$. Then, for any $\varepsilon > 0$, there exists a sufficiently small constant $s_\varepsilon \in (0, \varepsilon/2)$ such that

$$P\left(\sup_{k=1,\ldots,p} |\widehat{\omega}_k - \omega_k| > \varepsilon\right) \leq 2p\exp\{n\log(1-\varepsilon s_\varepsilon/2)/3\}.$$

In addition, if we write $\delta = \min_{k\in\mathcal{A}} \omega_k - \max_{k\in\mathcal{I}} \omega_k$, then there exists a sufficiently small constant $s_{\delta/2} > 0$ such that

$$P\left(\max_{k\in\mathcal{I}} \widehat{\omega}_k < \min_{k\in\mathcal{A}} \widehat{\omega}_k\right) \geq 1 - 4p\exp\left\{n\log(1-\delta s_{\delta/2}/4)/3\right\}.$$

Note that $p = o(\exp(an))$. Thus, the right-hand side of the above equation approaches 1 with an exponential rate as $n \to \infty$. Theorem 2 justifies using $\widehat{\omega}_k$ to rank the predictors, and it establishes the consistency in ranking. That is, $\widehat{\omega}_k$ always ranks an active predictor above an inactive one in probability, and so guarantees a clear separation between the active and inactive predictors.

Provided an ideal cutoff is available, this property would lead to consistency in selection in the ultrahigh-dimensional setup. Next we propose a thresholding rule to obtain a cutoff value to separate the active and inactive predictors.

Soft-Thresholding cutoff:

We independently and randomly generate d auxiliary variables $\mathbf{z} \sim N_d(\mathbf{0}, \mathbf{I}_d)$ such that $\mathbf{z}$ is independent of both $\mathbf{x}$ and Y. The normality is not critical here, as we shall see later.

Regard the $(p+d)$ dimensional vector $(\mathbf{x}^T, \mathbf{z}^T)^T$ as the predictors and Y as the response. We calculate ω_k for $k = 1, \cdots, p+d$. Since $\mathbf{z}$ is truly inactive by construction, we have $\min_{k \in \mathcal{A}} \omega_k > \max_{\ell=1,\ldots,d} \omega_{p+\ell}$ by Theorem 1, and given a random sample $\{(\mathbf{x}_i, \mathbf{z}_i, Y_i), i = 1, \ldots, n\}$, it holds in probability that $\min_{k \in \mathcal{A}} \widehat{\omega}_k > \max_{\ell=1,\ldots,d} \widehat{\omega}_{p+\ell}$ by Theorem 2.

Define $C_d = \max_{\ell=1,\ldots,d} \widehat{\omega}_{p+\ell}$, which can be viewed as a benchmark that separates the active predictors from the inactive ones. This leads to the selection,

$$\widehat{\mathcal{A}}_1 = \{k : \widehat{\omega}_k > C_d\}. \tag{31}$$

We call (31) the soft thresholding selection.

Theorem 3

Let $r \in \mathbb{N}$, the set of natural numbers. We assume the exchangeability condition, that is, the inactive predictors $\{X_j, j \in \mathcal{I}\}$ and the auxiliary variables $\{Z_j, j = 1, \ldots, d\}$ are exchangeable in the sense that both the inactive and auxiliary variables are equally likely to be recruited by the soft cutoff procedure. Then

$$P\left(|\widehat{\mathcal{A}}_1 \cap \mathcal{I}| \geq r\right) \leq \left(1 - \frac{r}{p+d}\right)^d,$$

where $|\cdot|$ denotes the cardinality of a set.

Hard cutoff

Let $N = [n/\log n]$

$$\widehat{\mathcal{A}}_2 = \{k : \widehat{\omega}_k > \widehat{\omega}_{(N)}\}, \tag{32}$$

where $\widehat{\omega}_{(N)}$ denotes the N-th largest value among all $\widehat{\omega}_k$'s.

In practice, the data determine whether the soft or hard thresholding comes into play. Our experience has suggested that, when the number of active predictors p_1 is relatively small compared to the sample size n, the hard cutoff is more relevant; whereas when p_1 is large compared to n, the soft cutoff is more relevant. Consequently, we propose to combine the soft and hard cutoff, and construct the final active predictor index set as

$$\widehat{\mathcal{A}} = \widehat{\mathcal{A}}_1 \cup \widehat{\mathcal{A}}_2, \tag{33}$$

where the union of the two sets is taken.

Iterative SIRS

Parallel to ISIS, an iterative SIRS may be used to deal with "marginal independence, joint dependence".

An iterative procedure is given as follows.

Step 1. We first apply our proposed screening procedure for $\mathbf{y}$ and $\mathbf{X}$, where $\mathbf{X}$ denotes the $n \times p$ data matrix that stacks n sample observations $\mathbf{x}_1, \ldots, \mathbf{x}_n$ and $\mathbf{y} = (Y_1, \ldots, Y_n)^T$. Suppose $p_{(1)}$ predictors are selected, where $p_{(1)} < N = [n/\log n]$. We denote the set of indices of the selected predictors by $\widehat{\mathcal{A}}_{(1)}$, and the associated $n \times p_{(1)}$ data matrix by $\mathbf{X}_{\widehat{\mathcal{A}}_{(1)}}$.

Step 2. Let $\widehat{\mathcal{I}}_{(1)}$ denote the complement of $\widehat{\mathcal{A}}_{(1)}$, and $\mathbf{X}_{\widehat{\mathcal{I}}_{(1)}}$ denote the remaining $n \times (p - p_{(1)})$ data matrix. Next, we define the predictor residual matrix

$$\mathbf{X}_r = \left\{ \mathbf{I}_n - \mathbf{X}_{\widehat{\mathcal{A}}_1} \left(\mathbf{X}_{\widehat{\mathcal{A}}_1}{}^T \mathbf{X}_{\widehat{\mathcal{A}}_1} \right)^{-1} \mathbf{X}_{\widehat{\mathcal{A}}_1}{}^T \right\} \mathbf{X}_{\widehat{\mathcal{I}}_{(1)}}.$$

Apply again our proposed screening procedure for $\mathbf{y}$ and $\mathbf{X}_r$. Suppose $p_{(2)}$ predictors are selected, and the resulting index set is denoted by $\widehat{\mathcal{A}}_{(2)}$. Update the total selected predictor set by $\widehat{\mathcal{A}}_{(1)} \cup \widehat{\mathcal{A}}_{(2)}$.

Step 3. Repeat Step 2 $M - 1$ times until the total selected number of predictors $p_{(1)} + \ldots + p_{(M)}$ exceeds the pre-specified number $N = [n/\log n]$. The final selected predictor set is $\widehat{\mathcal{A}}_{(1)} \cup \ldots \cup \widehat{\mathcal{A}}_{(M)}$.

3.2 Feature screening Based on Distance Correlation

Material in this section is based on

Li, R., Zhong, W. and Zhu, L. (2012). Feature screening via distance correlation learning. *Journal of American Statistical Association*. **107**, 1129-1139.

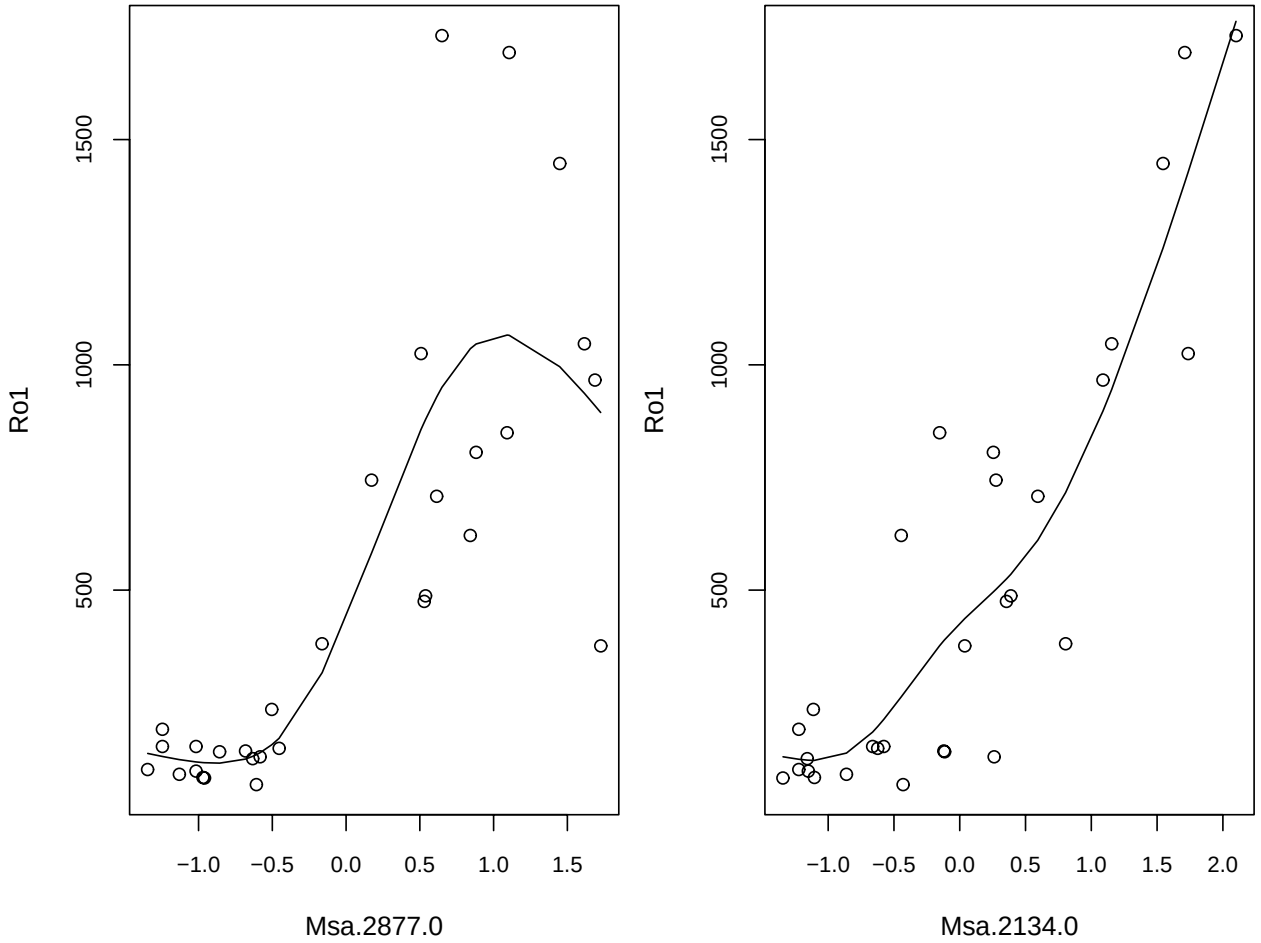
A Motivating Example

Cardiomyopathy Microarray Data, analyzed by Hall and Miller (2009), consists of $n = 30$ samples with $p = 6319$-dimensional $\mathbf{x}$.

- The Goal is to identify the most influential genes for overexpression of a G protein-coupled receptor (Ro1) in mice.
- The response Y – the Ro1 expression level
- The predictors X_k's – other gene expression levels.

$p = 6319 \gg n = 30$
A typical high or ultrahigh-dimensional data set.

- The scatter plots of Y versus the two genes with cubic spline fit curves.



- Plots show nonlinear relationship between X's and Y

Desired properties of DC-SIS

Desired properties for marginal screening procedures:

- Easy to implement and computational efficient.
- Good theoretical properties such as ranking consistency and sure screening property;
- **Model-free**: not require to specify a regression model btw x and y.

Goal: To proposed a screening procedure possess the above three properties. In addition, our procedure is directly applicable for

- continuous, categorial and count responses
- multivariate response and multivariate/groupwise predictors

Distance Correlation (Szekely, Rizzo and Bakirov, 2007)

Distance Correlation (dcorr) for two random vectors $\mathbf{X} \in \mathrm{R}^p$ and $\mathbf{y} \in \mathrm{R}^q$

$$\mathrm{dcorr}(\mathbf{X}, \mathbf{y}) \stackrel{\mathrm{def}}{=} \frac{\mathrm{dcov}(\mathbf{X}, \mathbf{y})}{\sqrt{\mathrm{dcov}(\mathbf{X}, \mathbf{X})\mathrm{dcov}(\mathbf{y}, \mathbf{y})}},$$

where dcov is the distance covariance:

$$\mathrm{dcov}(\mathbf{X}, \mathbf{y}) \stackrel{\mathrm{def}}{=} \int_{\mathrm{R}^{p+q}} |\phi_{\mathbf{X},\mathbf{y}}(\mathbf{t}, \mathbf{s}) - \phi_{\mathbf{X}}(\mathbf{t})\phi_{\mathbf{y}}(\mathbf{s})|^2 \, w(\mathbf{t}, \mathbf{s}) d\mathbf{t} d\mathbf{s},$$

with $\phi(\cdot)$ being the characteristic function and

$$w(\mathbf{t}, \mathbf{s}) = \left\{ c_p c_q \|\mathbf{t}\|^{1+p} \|\mathbf{s}\|^{1+q} \right\}^{-1}$$

with $c_d = \pi^{(1+d)/2} / \Gamma\{(1+d)/2\}$.

Some properties of Distance Correlation

- When $\mathbf{X}$ and $\mathbf{y}$ are univariate normal, $\mathrm{dcorr}(\mathbf{X}, \mathbf{y})$ is strictly increasing in $|\rho(\mathbf{X}, \mathbf{y})|$, where $\rho(\mathbf{X}, \mathbf{y})$ is the Pearson correlation.
- $\mathrm{dcorr}(\mathbf{X}, \mathbf{y})$ defined for $\mathbf{X}$ and $\mathbf{y}$ in arbitrary dimensions;
- $\mathrm{dcorr}(\mathbf{X}, \mathbf{y}) = 0 \Leftrightarrow \mathbf{X}$ and $\mathbf{y}$ are independent.

Sample distance correlation

As shown in Szekely, Rizzo and Bakirov (2007)

$$\mathrm{dcov}^2(\mathbf{u}, \mathbf{v}) = S_1 + S_2 - 2S_3,$$

where S_j, $j = 1$, 2 and 3, are defined below:

$$\begin{aligned} S_1 &= E\left\{\|\mathbf{u} - \widetilde{\mathbf{u}}\|\|\mathbf{v} - \widetilde{\mathbf{v}}\|\right\}, \\ S_2 &= E\left\{\|\mathbf{u} - \widetilde{\mathbf{u}}\|\right\} E\left\{\|\mathbf{v} - \widetilde{\mathbf{v}}\|\right\}, \\ S_3 &= E\left\{E\left(\|\mathbf{u} - \widetilde{\mathbf{u}}\|\,|\,\mathbf{u}\right) E\left(\|\mathbf{v} - \widetilde{\mathbf{v}}\|\,|\,\mathbf{v}\right)\right\}, \end{aligned}$$

where $(\widetilde{\mathbf{u}}, \widetilde{\mathbf{v}})$ is an independent copy of $(\mathbf{u}, \mathbf{v})$. Thus, the sample correlation is the moment estimator simply replacing S_j by $\widehat{S}_j$. E.g.

$$\widehat{S}_3 = \frac{1}{n^3} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{l=1}^{n} \|\mathbf{u}_i - \mathbf{u}_l\|\|\mathbf{v}_j - \mathbf{v}_l\|.$$

Statistical Setting

- $\mathbf{X} = (X_1, \ldots, X_p)$ are predictors and $\mathbf{y}$ is the response. Without assuming any model, we define the active subset

$$\begin{aligned} \mathcal{D} &= \{k : F(y|\mathbf{X}) \text{ functionally depends on } X_k, 1 \le k \le p\}, \\ &= \{k : X_k \text{ is important for } y\}; \end{aligned}$$

  Then inactive subset $\mathcal{I} = \mathcal{D}^c = \{1, 2, \ldots, p\} \setminus \mathcal{D}$.

- Sparsity assumption: $|\mathcal{D}| << p$ and hopefully $|\mathcal{D}| < n$

Goals of Screening:

- to remove as many variables from $\mathcal{I}$ as possible,
- to keep all important variables from $\mathcal{D}$,
- to reduce the dimension p to relative large one (hopefully $< n$).

Distance Correlation Sure Independence Screening

- DC-SIS uses sample distance correlation $\widehat{\omega}_k = \widehat{\text{dcorr}}^2(X_k, \mathbf{y})$ as a marginal utility to measure the importance of X_k, then define the submodel

$$\widehat{\mathcal{D}} = \{k : \widehat{\omega}_k \text{ is among the top } d \text{ largest of all}\}.$$

- DC-SIS reduces **very large size** p down to **size** d **at the sample level.**

Theoretical Properties of DC-SIS

Conditions for Ranking Consistency:

(A1) both $\mathbf{x}$ and $\mathbf{y}$ satisfy the sub-exponential tail probability uniformly in p. That is, there exists a positive constant s_0 such that

$$\max_{1\le k\le p} E\left\{\exp(s\|X_k\|^2)\right\} < \infty, \text{ and } E\{\exp(s\|\mathbf{y}\|^2)\} < \infty,$$

for all $0 < s \le 2s_0$.

(A2) $\liminf_{p\to\infty}\{\min_{k\in\mathcal{D}}\omega_k - \max_{k\in\mathcal{I}}\omega_k\} > 0$.

Condition (A1) follows immediately when $\mathbf{x}$ and $\mathbf{y}$ are bounded uniformly, or when they have multivariate normal distribution.

Condition (A2) requires in spirit that the active and inactive predictors should be weakly correlated.

Ranking Consistency

Theorem 1 (Ranking Consistency Property) If conditions (A1) and (A2) hold true for $p = o\{\exp(an)\}$ where $a > 0$ is a fixed constant, then

$$\liminf_{p\to\infty}\{\min_{k\in\mathcal{D}}\widehat{\omega}_k - \max_{k\in\mathcal{I}}\widehat{\omega}_k\} > 0, \text{ in probability.}$$

Conditions (A1) and (A2) are used to facilitate the technical proofs, although they may not be the weakest ones.

Conditions for Sure Screening Property

(B1) both $\mathbf{x}$ and $\mathbf{y}$ are uniformly bounded. That is, there exist positive constants a and b such that $\max_{1\le k\le p} \|X_k\| \le a$ and $\|\mathbf{y}\|_q \le b$; and

(B2) the minimum distance correlation of active predictors satisfies

$$\min_{k\in\mathcal{D}} \omega_k \ge 2cn^{-\kappa}, \text{ for some constants } c > 0 \text{ and } 0 \le \kappa < 1/2.$$

Although (B1) seems more stringent than (A1), it is mild in our context because we can consider using monotone transformations of $\mathbf{x}$ and $\mathbf{y}$ when they are not bounded. E.g, we can use $T(z) = \exp(z)/\{\exp(z)+1\}$ for every predictor and each response. Such marginal transformations of $\mathbf{x}$ and $\mathbf{y}$ do not change the definitions of active and inactive predictors. This is because $T(X_k) \perp\!\!\!\perp \mathbf{y}$ is equivalent to $X_k \perp\!\!\!\perp \mathbf{y}$ when the transformation function $T(\cdot)$ is monotonic.

Sure Screening Property

Under Condition (B1), it holds that

$$\Pr\left(\max_{1\leq k\leq p}|\widehat{\omega}_k-\omega_k|>cn^{-\kappa}\right)\leq O\left\{p\exp\left(-n^{1-2\kappa}/\xi_{a,b,c}\right)\right\}, \tag{34}$$

where $\xi_{a,b,c}$ is a positive constant depending on c and the boundaries a and b.
Under Conditions (B1) and (B2), we have that

$$\Pr\left(\mathcal{D}\subseteq\widehat{\mathcal{D}}^{\star}\right)\geq 1-O\left\{s_n\exp\left(-n^{1-2\kappa}/\xi_{a,b,c}\right)\right\}, \tag{35}$$

where s_n is the cardinality of $\mathcal{D}$.

Simulation Setting

- $\mathbf{x} = (X_1, X_2, \cdots, X_p)^T \sim MVN(\mathbf{0}, \Sigma)$, where $\Sigma = (\sigma_{ij})_{p\times p}$, $\sigma_{ij} = 0.8^{|i-j|}$,
- $\varepsilon \sim N(0, 1)$,
- $n = 200$ and $p = 1000$, so $p \gg n$,
- repeat each experiment 500 times,
- Three criteria to evaluate the performance:
  1. $\mathcal{S}$ – the minimum model size to include all active predictors;
  2. $\mathcal{P}_s$ – the empirical probability that every single active predictor is selected into $\widehat{\mathcal{D}}$;
  3. $\mathcal{P}_a$ – the empirical probability that all active predictors are selected into $\widehat{\mathcal{D}}$.

Example 1

$$
\begin{aligned}
\textbf{(1.a)}: Y &= c_1\beta_1 X_1 + c_2\beta_2 X_2 + c_3\beta_3 \mathbf{1}(X_{12} < 0) + c_4\beta_4 X_{22} + \varepsilon, \\
\textbf{(1.b)}: Y &= c_1\beta_1 X_1 X_2 + c_3\beta_2 \mathbf{1}(X_{12} < 0) + c_4\beta_3 X_{22} + \varepsilon, \\
\textbf{(1.c)}: Y &= c_1\beta_1 X_1 X_2 + c_3\beta_2 \mathbf{1}(X_{12} < 0) X_{22} + \varepsilon, \\
\textbf{(1.d)}: Y &= c_1\beta_1 X_1 + c_2\beta_2 X_2 + c_3\beta_3 \mathbf{1}(X_{12} < 0) + \exp(c_4|X_{22}|)\varepsilon,
\end{aligned}
$$

where $\mathbf{1}(\cdot)$ is an indicator function. Following Fan and Lv (2008), $\beta_j = (-1)^U(a + |Z|)$ for $j = 1, 2, 3$ and 4, where $a = 4\log n/\sqrt{n}$, $U \sim$ Bernoulli(0.4) and $Z \sim \mathcal{N}(0, 1)$. $(c_1, c_2, c_3, c_4) = (2, 0.5, 3, 2)$.

The minimum model size $\mathcal{S}$ to ensure the inclusion of all active predictors.

$\mathcal{S}$	SIS			SIRS			DC-SIS		
Model	50%	75%	95%	50%	75%	95%	50%	75%	95%
$p = 2000$ and $\sigma_{ij} = 0.5^{\lvert i-j \rvert}$									
(1.a)	5.0	7.0	21.2	5.0	7.0	45.1	4.0	6.0	18.0
(1.b)	1180.5	1634.5	1938.0	1386.0	1725.2	1942.4	24.5	73.0	345.1
(1.c)	1438.0	1745.0	1945.1	1320.0	1697.0	1946.0	22.0	59.0	324.1
(1.d)	1166.0	1637.0	1936.5	797.0	1432.2	1846.1	9.0	41.0	336.2
$p = 2000$ and $\sigma_{ij} = 0.8^{\lvert i-j \rvert}$									
(1.a)	16.0	97.0	729.4	18.0	112.8	957.1	11.0	31.2	507.2
(1.b)	852.0	1541.2	1919.0	1174.0	1699.2	1968.0	11.0	17.0	98.0
(1.c)	1249.5	1670.0	1951.1	1201.5	1685.2	1955.0	15.0	38.0	198.3
(1.d)	1107.5	1626.2	1930.0	728.0	1368.2	1900.1	17.0	73.2	653.1

REMARK: Models **(1.a)**-**(1.d)** are nonlinear in terms of xs because the mean function of Y is nonlinearly related to X_{12}.

In addition, models (**1.b**) and (**1.c**) contain interaction terms X_1X_2, and model (**1.d**) is heteroscedastic.

The empirical probabilities of each active predictor (denoted by $\mathcal{P}_s$) and all active predictors (denoted by $\mathcal{P}_a$) are chosen for a given model size.

		SIS					SIRS					DC-SIS				
		$\mathcal{P}_s$				$\mathcal{P}_a$	$\mathcal{P}_s$				$\mathcal{P}_a$	$\mathcal{P}_s$				$\mathcal{P}_a$
model	size	X_1	X_2	X_{12}	X_{22}	ALL	X_1	X_2	X_{12}	X_{22}	ALL	X_1	X_2	X_{12}	X_{22}	ALL
$p = 2000$ and $\sigma_{ij} = 0.5^{\lvert i-j \rvert}$																
(1.a)	d_1	1.0	1.0	.96	1.0	.96	1.0	1.0	.95	1.0	.94	1.0	1.0	.97	1.0	.96
	d_2	1.0	1.0	.98	1.0	.97	1.0	1.0	.96	1.0	.96	1.0	1.0	.98	1.0	.98
	d_3	1.0	1.0	.98	1.0	.98	1.0	1.0	.97	1.0	.97	1.0	1.0	.99	1.0	.98
(1.b)	d_1	.08	.07	.97	1.0	.03	.02	.03	.98	1.0	.0	.72	.70	.99	1.0	.58
	d_2	.12	.13	.98	1.0	.06	.05	.05	.99	1.0	.01	.85	.84	1.0	1.0	.76
	d_3	.15	.17	.99	1.0	.07	.06	.06	.99	1.0	.01	.89	.88	1.0	1.0	.82
(1.c)	d_1	.12	.13	.01	.99	.0	.04	.03	.51	1.0	.01	.93	.93	.77	1.0	.65
	d_2	.17	.18	.03	.99	.0	.07	.05	.67	1.0	.01	.97	.96	.84	1.0	.79
	d_3	.21	.21	.05	.99	.0	.09	.08	.75	1.0	.02	.98	.97	.89	1.0	.84
(1.d)	d_1	.42	.22	.14	.42	.02	1.0	.98	.87	.05	.04	1.0	.91	.81	.99	.73
	d_2	.48	.29	.22	.50	.03	1.0	.99	.91	.10	.09	1.0	.94	.87	1.0	.82
	d_3	.56	.32	.26	.54	.04	1.0	.99	.93	.12	.11	1.0	.96	.92	1.0	.88

$d_1 = [n/\log n]$, $d_2 = 2d_1$ and $d_3 = 3d_1$

The empirical probabilities of each active predictor (denoted by $\mathcal{P}_s$) and all active predictors (denoted by $\mathcal{P}_a$) are chosen for a given model size.

		SIS					SIRS					DC-SIS				
		$\mathcal{P}_s$				$\mathcal{P}_a$	$\mathcal{P}_s$				$\mathcal{P}_a$	$\mathcal{P}_s$				$\mathcal{P}_a$
model	size	X_1	X_2	X_{12}	X_{22}	ALL	X_1	X_2	X_{12}	X_{22}	ALL	X_1	X_2	X_{12}	X_{22}	ALL
$p = 2000$ and $\sigma_{ij} = 0.8^{\lvert i-j \rvert}$																
	d_1	1.0	1.0	.63	1.0	.63	1.0	1.0	.62	1.0	.62	1.0	1.0	.78	1.0	.77
(1.a)	d_2	1.0	1.0	.71	1.0	.72	1.0	1.0	.70	1.0	.69	1.0	1.0	.84	1.0	.84
	d_3	1.0	1.0	.77	1.0	.78	1.0	1.0	.75	1.0	.75	1.0	1.0	.86	1.0	.86
	d_1	.12	.13	.81	1.0	.06	.04	.04	.88	1.0	.02	.97	.98	.92	1.0	.88
(1.b)	d_2	.19	.19	.86	1.0	.12	.07	.07	.91	1.0	.03	.99	.99	.95	1.0	.94
	d_3	.22	.23	.88	1.0	.15	.09	.11	.93	1.0	.06	1.0	.99	.96	1.0	.96
	d_1	.17	.16	.03	.99	.0	.04	.04	.53	1.0	.02	1.0	1.0	.75	1.0	.75
(1.c)	d_2	.22	.22	.06	1.0	.01	.08	.08	.71	1.0	.03	1.0	1.0	.85	1.0	.86
	d_3	.27	.27	.10	1.0	.03	.10	.10	.81	1.0	.05	1.0	1.0	.90	1.0	.90
	d_1	.44	.38	.11	.45	.03	1.0	1.0	.73	.05	.04	.99	.98	.68	1.0	.67
(1.d)	d_2	.51	.46	.18	.53	.05	1.0	1.0	.81	.09	.08	1.0	.98	.76	1.0	.75
	d_3	.55	.49	.22	.57	.06	1.0	1.0	.84	.14	.11	1.0	.99	.80	1.0	.80

$d_1 = [n/\log n]$, $d_2 = 2d_1$ and $d_3 = 3d_1$

Example 2: Group-wise variable screening

$$\begin{aligned} Y &= c_1\beta_1 X_1 + c_2\beta_2 X_2 + c_3\beta_3\{\mathbf{1}(X_{12} < q_1) + 1.5\mathbf{1}(q_1 \le X_{12} < q_2) \\ &+ 2\mathbf{1}(q_2 \le X_{12} < q_3) + \mathbf{1}(X_{12} \ge q_3)\} + c_4\beta_4 X_{22} + \varepsilon, \end{aligned}$$

where q_1, q_2 and q_3 are the 25%, 50% and 75% quantiles of X_{12}, respectively. The coefficients c_i's and β_i's are the same as those in Example 1.

$$\widetilde{\mathbf{x}}_{12} = \{\mathbf{1}(X_{12} < q_1),\ \mathbf{1}(q_1 \le X_{12} < q_2),\ \mathbf{1}(q_2 \le X_{12} < q_3),\ \mathbf{1}(X_{12} \ge q_3)\}^T.$$

These four correlated variables naturally become a group. The active predictor vector in this example becomes $\mathbf{x} = (X_1, \ldots, X_{11}, \widetilde{\mathbf{x}}_{12}, X_{13}, \ldots, X_p)^T$. We remark here that the marginal utility of the grouped variable $\widetilde{\mathbf{x}}_{12}$ is defined by

$$\widehat{\omega}_{12} = \widehat{\mathrm{dcorr}}^2(\widetilde{\mathbf{x}}_{12}, Y).$$

Simulation Result for Example 2

The 5%, 25%, 50%, 75% and 95% quantiles of the minimum model size $\mathcal{S}$ out of 500 replications in Example 2 with $p = 2000$.

	5%	25%	50%	75%	95%
$\sigma_{ij} = 0.5^{\|i-j\|}$	4.0	4.0	4.0	5.0	12.0
$\sigma_{ij} = 0.8^{\|i-j\|}$	4.0	5.0	7.0	9.0	15.2

Example 3: Multiple Responses

Given $\mathbf{x}$, we generate $\mathbf{y} = (Y_1, Y_2)^T \sim N(\mathbf{0}, \mathbf{\Sigma}_{\mathbf{y}|\mathbf{x}})$.
Let $\mathbf{\Sigma}_{\mathbf{y}|\mathbf{x}} = (\sigma_{ij}(\mathbf{x}))_{2\times 2}$, where $\sigma_{11}(\mathbf{x}) = \sigma_{22}(\mathbf{x}) = 1$ and $\sigma_{12}(\mathbf{x}) = \sigma_{21}(\mathbf{x}) = \rho(\mathbf{x})$.
We consider two scenarios for the correlation function $\rho(\mathbf{x})$:

(3.a): $\rho(\mathbf{x}) = \sin(\mathbf{b}_1{}^T\mathbf{x})$, where $\mathbf{b}_1 = (0.8, 0.6, 0, \ldots, 0)^T$.

(3.b): $\rho(\mathbf{x}) = \left\{\exp(\mathbf{b}_2{}^T\mathbf{x}) - 1\right\} / \left\{\exp(\mathbf{b}_2{}^T\mathbf{x}) + 1\right\}$, where $\mathbf{b}_2 = (2 - U_1, 2 - U_2, 2 - U_3, 2 - U_4, 0, \ldots, 0)^T$ with $U_i \sim \mathrm{Uniform}[0, 1]$.

Simulation Results for Example 3

The minimum model size $\mathcal{S}$ with $p = 2000$.

	Model	5%	25%	50%	75%	95%
$\sigma_{ij} = 0.5^{\|i-j\|}$	**(3.a)**	4.0	9.0	18.0	39.3	112.3
	(3.b)	6.0	19.0	43.0	92.0	253.1
$\sigma_{ij} = 0.8^{\|i-j\|}$	**(3.a)**	2.0	3.0	6.0	12.0	40.0
	(3.b)	4.0	4.0	4.0	6.0	10.0

The proportions $\mathcal{P}_s$ and $\mathcal{P}_a$ with $p = 2000$.

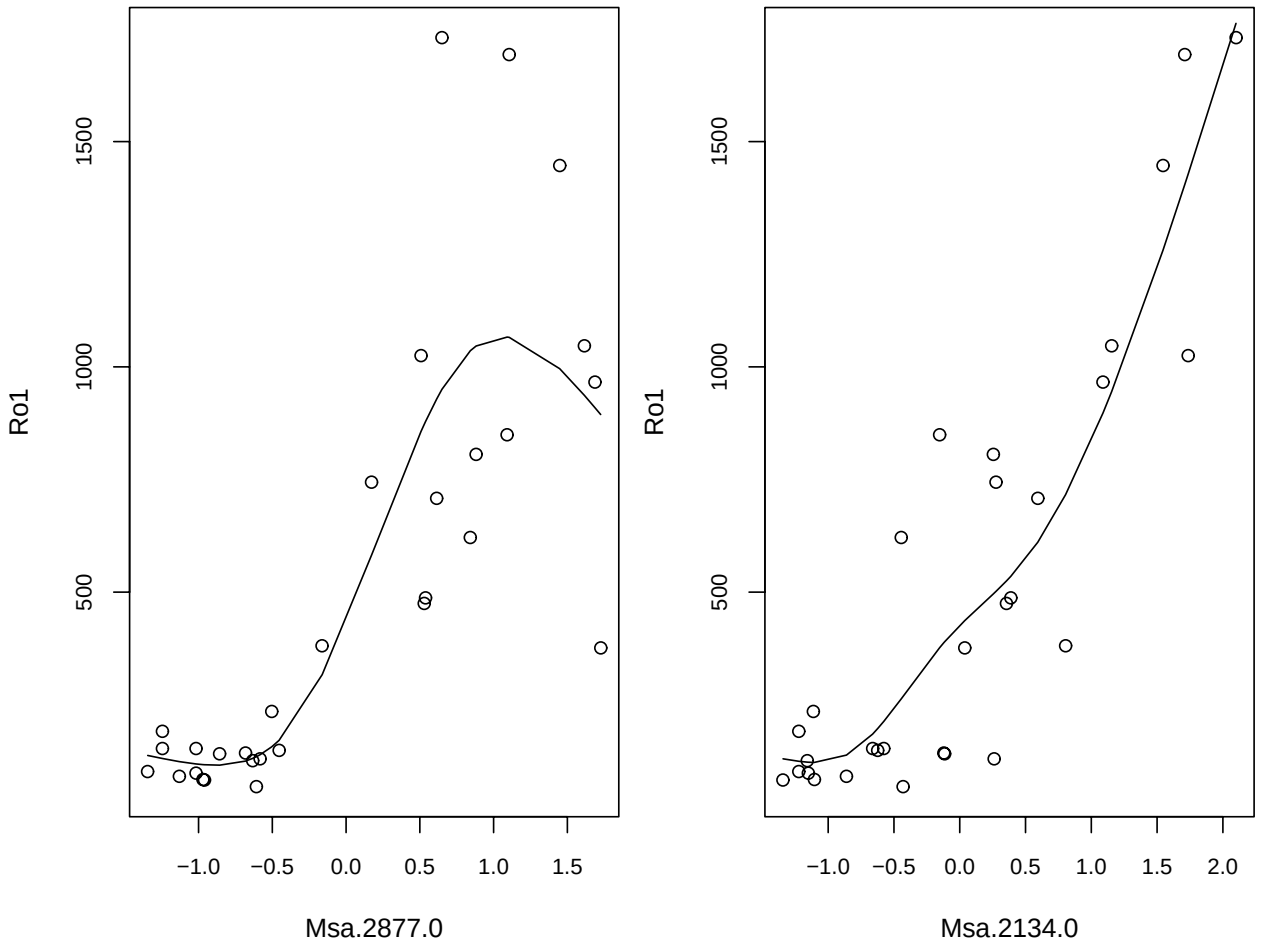
		(3.a)			**(3.b)**				
		$\mathcal{P}_s$		$\mathcal{P}_a$	$\mathcal{P}_s$				$\mathcal{P}_a$
	size	X_1	X_2	ALL	X_1	X_2	X_3	X_4	ALL
$\sigma_{ij} = 0.5^{\|i-j\|}$	d_1	0.95	0.76	0.74	0.71	0.98	0.98	0.72	0.47
	d_2	0.98	0.90	0.90	0.85	0.99	0.99	0.85	0.71
	d_3	1.00	0.95	0.95	0.91	0.99	1.00	0.90	0.81
$\sigma_{ij} = 0.8^{\|i-j\|}$	d_1	0.98	0.95	0.94	1.00	1.00	1.00	1.00	1.00
	d_2	1.00	0.98	0.99	1.00	1.00	1.00	1.00	1.00
	d_3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Real Data Example - Cardiomyopathy Microarray Data

This microarray dataset was once analyzed by Hall and Miller (2009).

- The Goal is to identify the most influential genes for overexpression of a G protein-coupled receptor (Ro1) in mice.
- The response Y – the Ro1 expression level
- The predictors X_k's – other gene expression levels.
- Dimension $p = 6319 \gg$ Sample size $n = 30$.

- The scatter plots of Y versus the top two genes identified by DC-SIS with cubic spline fit curves.



- DC-SIS can capture linear and nonlinear relationships.

Comparsion with Hall and Miller (2009)

- To compare the performance of DC-SIS with other screening methods, we fit an additive model using top two predictors,

$$Y = f_{k1}(X_{k1}) + f_{k2}(X_{k2}) + \varepsilon_k, \text{ for } k = 1, 2, 3.$$

Methods	Adjusted R^2	MSE
SIS	0.842	0.137
HM(2009)	0.845	0.129
DC-SIS	0.968	0.016

Summary

- DC-SIS is an extension of SIS using the Pearson's correlation;
- DC-SIS is easy to implement.
- DC-SIS possesses ranking consistency property and sure screening property
- DC-SIS is model-free manner;
- DC-SIS allows for groupwise predictors and multivariate responses.
- DC-SIS is applicable for continuous, categorical, count predictors/responses;
- Numerical studies showed that DC-SIS had better performance and wider applications than some existing screening methods.