INDEPENDENCE SCREENING IN HIGH-DIMENSIONAL DATA

by

John Wauters

A Thesis submitted to the Faculty of the

GRADUATE INTERDISCIPLINARY PROGRAM IN STATISTICS

In Partial Fulfillment of the Requirements

For the Degree of

MASTER OF SCIENCE

In the Graduate College

THE UNIVERSITY OF ARIZONA

2016

STATEMENT BY AUTHOR

The thesis titled *Independence Screening in High-Dimensional Data* prepared by *John Wauters* has been submitted in partial fulfillment of requirements for a master's degree at the University of Arizona and is deposited in the University Library to be made available to borrowers under rules of the Library.

Brief quotations from this thesis are allowable without special permission, provided that an accurate acknowledgement of the source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the head of the major department or the Dean of the Graduate College when in his or her judgment the proposed use of the material is in the interests of scholarship. In all other instances, however, permission must be obtained from the author.

SIGNED: John C. Wauters

APPROVAL BY THESIS DIRECTOR

This thesis has been approved on the date shown below:

_____	*Defense date*
Yue Niu	12/12/2016
Professor of Mathematics	

Table of Contents:

List of Figures and Tables:

Figures

Tables

1 Abstract

High-dimensional data, data in which the number of dimensions exceeds the number of observations, is increasingly common in statistics. The term “ultra-high dimensional” is defined by Fan and Lv (2008) as describing the situation where log(p) is of order $O(n^a)$ for some a in the interval (0, ½). It arises in many contexts such as gene expression data, proteomic data, imaging data, tomography, and finance, as well as others. High-dimensional data present a challenge to traditional statistical techniques. In traditional statistical settings, models have a small number of features, chosen based on an assumption of what features may be relevant to the response of interest. In the high-dimensional setting, many of the techniques of traditional feature selection become computationally intractable, or does not yield unique solutions. Current research in modeling high-dimensional data is heavily focused on methods that screen the features before modeling; that is, methods that eliminate noise-features as a pre-modeling dimension reduction. Typically noise feature are identified by exploiting properties of independent random variables., thus the term “independence screening”

There are methods for modeling high-dimensional data without feature screening first (e.g. LASSO or SCAD), but simulation studies show screen-first methods perform better as dimensionality increases.

Many proposals for independence screening exist, but in my literature review certain themes recurred:

- The assumption of sparsity: that all the useful information in the data is actually contained in a small fraction of the features (the “active features”), the rest being essentially random noise (the “inactive” features).
- In many newer methods, initial dimension reduction by feature screening reduces the problem from the high-dimensional case to a classical case; feature selection then proceeds by a classical method.
- In the initial screening, removal of features independent of the response is highly desirable, as such features literally give no information about the response.
- For the initial screening, some statistic is applied pairwise to each feature in combination with the response; the specific statistic chosen so that in the case

that the two random variables are independent, a specific known value is expected for the statistic.

- Features are ranked by the absolute difference between the calculated statistic and the expected value of that statistic in the independent case, i.e. features that are most different from the independent case are most prefered.
- Proof is typically offered that, asymptotically, the method retains the true active features with probability approaching one.
- Where possible, an iterative version of the process is explored, as iterative versions do much better at identifying features that are active in their interactions, but not active individually.

2 Introduction

Ultra-high dimensional data is defined by Fan and Lv (2008) as data for which the log of the number of features is of the order of a power of the number of observations; that power being a positive number smaller than one-half. Informally high-dimensional data are data with more features than observations, often many times more features than observations. High-dimensional data are increasingly common in statistics. It can arise in many fields, medical imaging, hyperspectral imaging, genomics, tomography, finance and others. The literature is especially rich with examples from medicine, and especially oncology. In a typical example, a tissue sample is subjected to an automated process that provides gene expression levels for upwards of 10,000 genes, with the researcher analyzing this data for any number of classification purposes (sample is cancerous versus noncancerous, tumor is one type of cancer versus another, the patient will likely respond well or poorly to a particular treatment, etc.) Typically the number of samples available to the researcher is a small fraction of the number of gene expression measurements available per sample.

Classical statistics developed in a context of data with dimensionality lower than the number of observations. Many of the modeling techniques of classical statistics either do not work with high-dimensional data, or work poorly as dimensionality increases. With high-dimensional data matrices are often either singular or ill-conditioned, which causes many least squares methods to fail. Classical modeling methods that do still work with

high-dimensional data, such as forward selection methods, have increasing prediction errors as dimensionality increases.

One approach to the difficulties of high-dimensional data is “feature screening”; a dimension reduction by identifying “noise” features and removing them before model construction. Typically identifying noise features exploits properties of independent random variables; thus the term “independence screening.”

This thesis is an introduction to independence screening.

Feature screening should not be confused with feature selection. The objective of feature screening is first to retain the active features, and second eliminate inactive features. In pursuit of the first objective, some inactive features are retained to insure no active features are lost. The objective of feature selection is judicious selection of features for a parsimonious model that fits, but does not over-fit, the data. Feature selection often rejects active features in pursuit of a smaller, more parsimonious model.

There are modeling methods for high-dimensional data that do not employ screening. In the case of linear models, non-screening methods include the least absolute shrinkage and selection operator (LASSO), smoothly clipped absolute deviation (SCAD), and the Dantzig selector from Candes and Tao (2007). However, most of current research is in methods that employ screening, as numerical studies show non-screening methods have higher prediction error than screening methods as dimensionality increases.

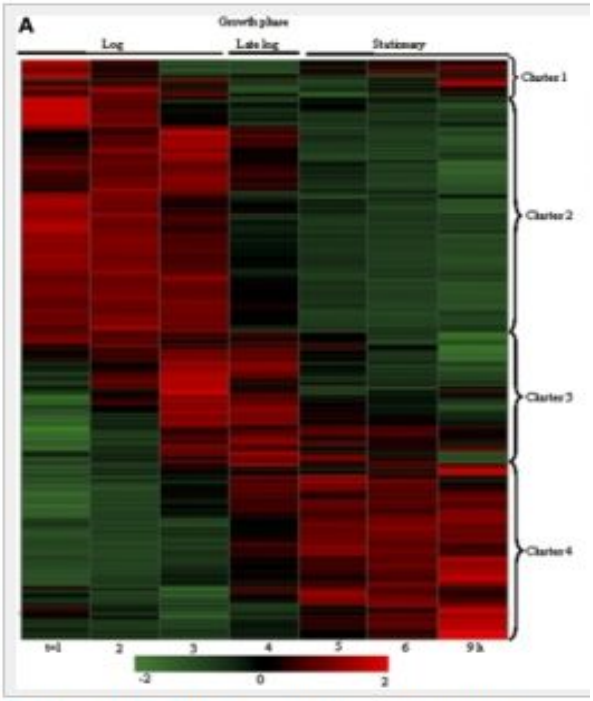
Independence screening depends on the sparsity assumption; the assumption that most of the useful information in the data is actually contained in a fraction of the features available (the ‘active’ features), and the other features are essentially noise (‘inactive’ features).

The methods described in this thesis determine some statistic for each feature, and rank the features by how different the statistic is from the what is expected in the case the feature is independent of the response. Then some number of features are screened in for use in model building. The methods differ chiefly in what statistic they use for ranking purposes.

The features are merely ranked; there is no formal testing of whether the feature is independent of the response, and no attempt is made to understand the distribution of the screening statistic used.

The number of features screened in is often n-1; not for any mathematically justified reason, but so that classical model building methods may be used. More data-driven methods of deciding how many features to screen in are an active area of research.

Figure 1



A heatmap plot from ten Broeke-Smits et al (2010) displaying high-dimensional genetic expression data.[1] Each column is an observation of gene expression for a bacteria sample at a different times, over three growth phases. Each row represents a different gene. The color in each rectangular space corresponds to a scaled and centered measured numerical value, with similar colors indicating similar values. Black represents values near the mean for that gene. Red indicates above-average expression for that gene. Green indicated below-average expression for that gene.

[1] Heatmap from ten Broeke-Smits *et al* (2010) https://openi.nlm.nih.gov/detailedresult.php?img=PMC2879529_gkq058f2&query=heatmap&lic=bync&req=4&npos=14
Used under non-commercial attribution creative commons licence.

The methods described in this thesis are known to have the "sure screening property"; that is asymptotically they screen in in the true active features with a probability approaching 1.

One of the limitations of independence screening methods is that they do poorly at identifying collections of features that are active in their interaction, but not active individually. Most of the methods in this thesis have an iterative variant that does better at identifying interacting features. Iterative versions also perform better generally in simulation studies. The typical schema for iterative versions is:

- Screen in some small number of features
- Model the data
- Using the residuals from the existing model as the new response, screen a small number of the remaining features
- Pool the newly screened and previously screened features and model again
- Repeat the cycle of screening a small number of additional features and constructing a new model with new residuals until either:
    - A predetermined number of features is reached, or
    - The set of screened features stabilizes

Iterative methods do not completely solve the problem of identifying interactions; they depend on one of the interacting pair to be active and screened in before the other member of the pair can be screened in. Research into better ways of identifying active interactions is ongoing.

The remainder of this thesis is structured as follows:

- Sections 3 through 6 are a brief literature review, describing screening methods for different classes of modeling problems.
- Section 3 addresses linear model methods.
- Section 4 addresses nonparametric methods.
- Section 5 addresses classification methods.
- Section 6 addresses survival model methods.
- In section 7 I present a more thorough review of three papers selected as significant in independence screening; which I refer to as the "Fan papers."
- In section 8 present my results of replicating select simulation studies and real data examples from the Fan papers.
- Section 9 presents some concluding remarks.

3 Linear Models and Generalized Linear Models

3.1 Discussion of Linear and Generalized Linear Models

Linear models are some of the oldest and most widely used statistical models. When attempting to model a continuous response by continuous features, a linear model is often the first attempt. Linear models follow the form:

$$E[Y] = \beta_0 + X\beta \quad (1)$$

where Y is a continuous random variable and X is a matrix of predictor variables.

These models quite flexible in that they can accommodate interactions by treating products of features as additional features, and also higher-order relationships with features by treating the powers of features as additional features.

The generalized linear model follows the form:

$$E[Y] = g^{-1}(\beta_0 + X\beta) \quad (2)$$

where again Y is a continuous random variable and X is a continuous random vector.

Depending on the choice of link function g(x), many distributions in the exponential family can be expressed as generalized linear models.

3.2 Fan and Lv (2008)

Fan and Lv introduce a method they call Sure Independence Screening (SIS). SIS ranks each feature by the absolute Pearson correlation between the feature and the response. They suggest choosing the n-1 highest ranking features as the screening. They present a theorem that the SIS method asymptotically retains the most important features with probability approaching one; a property they call the sure screening property. They demonstrate the SIS in combination with a number of feature selection methods on a variety of simulated data. They then apply SIS to real data, leukemia data originally published in Gollub *et al* (1999). Interestingly, Fan and Lv's simulation studies are all simulations of linear models, but their real data example applied SIS to a classification problem. Why they made that choice is not explained. They apply two

variations of SIS followed by classification, and compare the results to nearest shrunken centroids results.

Fan and Lv then introduce an iterative version of SIS (ISIS), and argue that the iterative version is needed because 1) Basic SIS can screen in unimportant features that happen to be highly correlated with important ones. 2) Basic SIS cannot identify groups of features that are marginally uncorrelated with the response, but are jointly correlated with the response, and 3) Basic SIS has no means of addressing collinearity between features. They argue that ISIS (iterative sure independence screening) addresses all of these issues. They then demonstrate ISIS on simulated data, comparing its performance to LASSO and SIS.

Their method of evaluation of ISIS differs slightly from their evaluation of SIS; SIS was evaluated in combination with a selection method applied after SIS. Also SIS is evaluated on the basis of model size, and also the difference between the estimated and actual beta vectors. In most of the simulation studies of ISIS, the evaluation is based on the success at screening in a set of features that includes the true model, but ISIS is not penalized for selecting additional inactive features. Also, in most of the simulation studies, ISIS is evaluated on its own, as opposed to in combination with a follow-up feature selection, as with the previous simulation studies of SIS. Why a different evaluation method was chosen is not clear.

For the above described simulation studies of ISIS, Fan and Lv report a very impressive 100% success rate at screening in the true active features, over hundreds of Monte Carlo runs, with differing true linear models, differing values of n and p, and differing correlation structures for the simulated features. That 100% success could be achieved over so many simulations and so many different models is surprising, and I was unable to replicate that result (see section 8.1 for more details).

These screening methods (SIS and ISIS) are easily grasped and easily implemented. This method is best used as a first attempt at linear models in high-dimensions.

3.3 Hall and Miller (2012)

As a variation on the theme of independence screening, Hall and Miller defined a new statistic called “generalized correlation,” which also has an expected value of zero

when the covariates are independent, and show a variation of SIS using the generalized correlation. They describe their technique as not prediction-based, and able to identify variables that are "influential, but not part of a predictive model." They explore their method's performance on real and simulated data, and show an argument for the validity of their method.

Their generalized correlation statistic between random variables Y and X_j is given by

$$\hat{\rho}_g(X_j, Y) = sup_{h \in H} \left(\sum_{i=1}^{n} \{h^2(X_{ij}) - \overline{h}_j^2\} \sum_{i=1}^{n} (Y_i - \overline{Y})^2\right)^{-1/2} \sum_{i=1}^{n} \{h(X_{ij}) - \overline{h}_j\}(Y_i - \overline{Y}) \quad (3)$$

where

$$\overline{h}_j = n^{-1} \sum_{i=1}^{n} h(X_{ij}) \quad (4)$$

and

$$\overline{Y} = n^{-1} \sum_{i=1}^{n} Y_i \quad (5)$$

where H is a class of function including all linear functions; in practice H is often the set of cubic splines.

The advantage of this method is flexibility; it is applicable with linear models, and also in a more general nonlinear case. This method should be used when the data is known or suspected to be significantly nonlinear.

3.4 Li, Peng, Zhang and Zhou (2012)

Li, Peng, Zhang and Zhou propose using still another variant using rank correlation in general transformation models of the form

$$H(Y_i) = x_i^T \beta + \varepsilon_i \text{ where H(•) is strictly monotonic} \quad (6)$$

This is called robust rank correlation screening (RRCS). They describe their method as being intended for '"large p, small n" paradigms when p can be as large as an exponential of the sample size n.' Their method is based on the Kendal Tau correlation rather than the Pearson correlation. Their method has four desirable properties when compared with previous screening methods:

1. The sure independence screening property can hold only under the existence of a second order moment of predictor variables, rather than exponential tails or alikeness, even when the number of predictor variables grows as fast as exponentially of the sample size.
2. It can be used to deal with semiparametric models such as transformation regression models and single-index models under monotonic constraint to the link function without involving nonparametric estimation even when there are nonparametric functions in the models.
3. The procedure can be largely used against outliers and influence points in the observations.
4. The use of indicator functions in rank correlation screening greatly simplifies the theoretical derivation due to the boundedness of the resulting statistics, compared with previous studies on variable screening.

They compare their method with previous methods (such as the SIS of Fan and Lv) using simulated data.

Their marginal rank correlation statistic for the jth feature is defined as

$$\omega_j = (1/(n(n-1))) \sum_{i \neq l}^{n} I(X_{ij} < X_{lj}) I(Y_i < Y_l) - \tfrac{1}{4} \quad (7)$$

which is one quarter of the Kendall τ correlation.

This method should be used when outliers are suspected, or when the features follow a heavy-tailed distribution.

4 Nonparametric Methods

Parametric statistics can be defined as statistics under the assumption that the random elements of the data follow distributions coming from a finite collection of known families, and individual distributions within those families can be uniquely identified by a fixed finite number of parameters. For example: within the family of Bernoulli distributions, a specific Bernoulli distribution can be identified by specifying the parameter p, the probability of a success; within the family of multivariate normal distributions of

some specified number of dimensions, a specific multivariate normal distribution can be identified by specifying a μ vector and a covariance matrix Σ.

Nonparametric statistics are used when the data do not fit a parametric distribution, or are too poorly understood to select an appropriate parametric model, or when an accurate parametric model would be excessively complicated.

4.1 Argument for Nonparametric

A nonparametric screening method is one that uses a nonparametric statistic to rank features, but otherwise follows our general schema of ranking features and preferentially screening in those most different from what is expected in the independent case.

A nonparametric modeling method is one that does not commit to a predetermined family of models; examples would include kernel regression, spline methods, or local polynomial regression, or rank-based methods.

Feature screening and feature selection are only loosely connected in the sense that a choice of one particular screening method does not commit one to a particular choice of selection method, or vice versa. More specifically, the choice to use a nonparametric screening method does not commit one to using a nonparametric modeling method; nor does the choice to use a parametric screening method commit one to using a parametric modeling method.

An argument for nonparametric modeling in the high-dimensional case can be made on the grounds that, having admitted that one does not know enough about the phenomenon we are trying to model to say which of very many potential features are important, how can one then assert that one does know enough about the phenomenon to say what form of model would best describe it.

With any feature screening method, it is highly desirable to have proof that it has the sure screening property. Such proofs typically have technical conditions, often including restrictions on the distribution of the underlying data that amount to “no tails heavier than the exponential distribution”. Some nonparametric statistics, such as rank correlation, or Kolmogorov-Smirnov statistic, are insensitive to heavy tailed distributions, and so do not require assuming the above regularity condition in order to have the sure

screening property. Also noteworthy is that some nonparametric statistics are insensitive to outliers, which can be helpful.

4.2 Liu *et al* (2014)

In the nonparametric case, using a varying coefficients model, Liu *et al* (2014) propose using kernel regression to model the data. They also propose using conditional correlation between the response and the individual features as feature screening. Their method screens in features most unlike the independent case, i.e. with correlations furthest from zero. They demonstrate this procedure has the sure screening property. They also develop an iterative version, and demonstrate their method on simulated data. For the kernel regression

$$\widehat{E}(Y|u) = \sum_{i=1}^{n} K_h(u_i - u)Y_i / [\sum_{i=1}^{n} K_h(u_i - u)] \; where \; K_h(t) = h^{-1}K(t/h) \tag{8}$$

for some kernel K(•), with h as a tuning parameter, and u is the variable controlling the varying coefficients. The screening statistic p^*_j for feature X_j is

$$\widehat{\rho}_j^{\,*} = \frac{1}{n}\sum_{i=1}^{n} \widehat{\rho}^{\,2}(X_j, \; Y|u_i) \tag{9}$$

where $\widehat{\rho}$ is the sample Pearson correlation. Their method proposes an unusual threshold, selecting a number of features d= integer part of $[n^{4/5}/\log(n^{4/5})]$.

This method should be used when a nonparametric model is desired.

4.3 He *et al* (2013)

He *et al* (2013) give a methodology based on quantile regression, intended for heterogenous data; that is, data from a mixture of dissimilar distributions. In the case that a feature is independent of the response, one expects the conditional quantile of the response to be no different than the unconditioned quantile. He *et al* use the squared difference between the conditioned and unconditioned quantiles to rate features, preferentially choosing the features with rating furthest from zero. They show this procedure has the sure selection property, with some mild conditions on the basis functions used to estimate the quantiles. Their framework has two distinctive features: (1) it allows the set of active variables to vary across quantiles, allowing it to accommodate

heterogeneity; (2) it is model-free and avoids the difficult task of specifying the form of a statistical model in a high-dimensional space. Their procedure uses spline approximations to model the marginal effects at a chosen quantile level. Under appropriate conditions, without requiring the existence of any moments, the new procedure has the sure screening property in ultra-high dimensions. This framework can work with censored data, such as from survival analysis. Numerical studies confirm the fine performance of the proposed method for various semiparametric models and its effectiveness to extract quantile-specific information from heteroscedastic data.

Their screening statistic for the j-th feature and the α th quantile is

$$\widehat{q}_{\alpha j} = \frac{1}{n}\sum_{i=1}^{n}\{B(X_{ij})^{T}\widehat{\Gamma}_{j} - F^{-1}_{Y,n}(\alpha)\}^{2} \quad (10)$$

where $B(x) = \{B_1(x), \ldots, B_{d_n}(x)\}^{T}$ B(x) is a set of basis functions approximating the quantile curve, X_{ij} is j-th feature of the i-th observation,

$$\widehat{\Gamma} = argmin_{\Gamma}\sum_{i=1}^{n}\rho_{\alpha}(Y_i - B(X_{ij})^{T}\Gamma) \quad (11)$$

where $\rho_{\alpha}(z) = z[\alpha - I(z < 0)]$ is the quantile loss function.

This method was invented specifically to model heteroskedastic data, but also works well with data with potential outliers, or when the features follow a heavy-tailed distribution.

4.4 Zhu *et al* (2012)

Zhu *et al* (2012) propose a model free screening procedure for ultrahigh-dimensional data. They argue that the high-dimensional data one often lacks information on the true model structure. In a model-free context, Zhu *et al* propose a new measure of marginal utility, and screen in features with marginal utilities furthest from zero. They show this method also has the sure screening property. Their novel method is computationally efficient, does not require specifying a particular model form, and performs well in empirical test with real and simulated data.

Their arguments proceed from an assumption that the conditional distribution of Y given X is dependent on X only through a linear combination of the components of X. In

other words, $F(y|x)=F_0(\bullet|B^T x_M)$ where $F_0(\bullet|B^T x_M)$ is an unknown distribution, and x_M is the vector containing only the active features from x.

Their screening statistic for feature k is their marginal utility, which when X_{ik} is the k-th dimension of the i-th observation, and $I(\bullet)$ is the indicator function, is defined as:

$$\widehat{\omega}_k = \frac{1}{n}\sum_{j=1}^{n}\{\frac{1}{n}\sum_{i=1}^{n} X_{ik} I(Y_i < Y_j)\}^2 \tag{12}$$

This method is model-free, and should be used when maximum flexibility is needed in choosing the final form of the model.

4.5 Li, Zhong, and Zhu (2012)

The distance correlation was introduced by Szekely *et al* (2007), to address a perceived shortcoming in the Pearson correlation; namely that independent random variables have zero correlation, but pairs of random variables with a correlation of zero are not necessarily independent. The distance correlations expected value is zero if and only if the two random variables are independent. Li, Zhong, and Zhu (2012) propose using the distance correlation for feature screening. Features with distance correlations furthest from zero are screened in. With some technical conditions, this method can be shown to have the sure screening property. They give an iterative version of the process. Their method does not require specifying a specific model, and numerical studies show better performance than the SIS of Fan and Lv (2008).

Their screening statistic for feature V with response U is

$$\widehat{dcorr}(U, V) = \widehat{dcov}(U, V)/\sqrt{\widehat{dcov}(U, U)\widehat{dcov}(V, V)} \tag{13}$$

where

$$\widehat{dcov}^2(U, V) = \widehat{S}_1 + \widehat{S}_2 - 2\widehat{S}_3 \tag{14}$$

where

$$\widehat{S}_1 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \|U_i - U_j\|\|V_i - V_j\| \tag{15}$$

$$\widehat{S}_2 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \|U_i - U_j\| \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \|V_i - V_j\| \tag{16}$$

$$\widehat{S}_3 = \frac{1}{n^3}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\|U_i - U_l\|\|V_j - V_l\| \quad (17)$$

and ||•|| is the Euclidean norm.

Like the previous method, this method is model-free, and should be used when maximum flexibility is needed in choosing the final form of the model. Also, this method uses a robust screening statistic, and should be used when outliers are suspected.

5 Classification

5.1 Discussion of Classification

Classification may be defined as the statistical problem of predicting a categorical response based upon (usually continuous) features. Classification with more than two possible classes is called “multiclass”, with only two classes it is called “binary.” A typical example would be to identify the species of an organism based on physical or genetic measurements.

Bickel and Levina (2004) show that traditional classification methods perform poorly in high-dimensions, due to the accumulation of noise. Even in favorable cases, such as when the covariance matrix is known, noise accumulation (the cumulative effect of multiple noise features) severely impairs traditional classification methods. As with the quantitative response models, the usual solution relies on statistics with known specific values in the case of independent random variables.

Tibshirani *et al* (2002) describe the method of nearest shrunken centroids. The shrinkage component of their method results in dimensions along which the classes are essentially identically distributed (as one would expect if that dimension is independent of the class) are shrunk to zero and effectively eliminated. Dimensions in which the distribution least resembles the independent case are preserved. To demonstrate their method, Tibshirani *et al* classify tumors into four categories with the nearest shrunken centroid method. They are able to shrink a set of 2,303 genes to a pool of 43 active genes, and achieve a test error rate of zero. The size of the shrinkage (which determines the number of features screened) is determined by 10-fold cross-validation.

5.2 Mai and Zou (2012) - Binary Classification

Mai and Zou (2012) propose screening for a binary classifier based on the Kolmogorov-Smirnov statistic, comparing the empirical cdf for each class for each feature. Features with Kolmogorov-Smirnov statistics furthest from zero are chosen. This method is model-free, and robust to heavy-tailed distributions and to possible outliers, but is limited to binary classification. This method has the sure screening property, under some technical conditions. With the notational simplification that the two classes are +1 and -1, the screening statistic for the j-th feature is

$$\widehat{\omega}_{nj} = sup_{x \in \Re} |\widehat{F}_{+j}(x) - \widehat{F}_{-j}(x)| \quad (18)$$

where $\widehat{F}_{+j}(x)$ *and* $\widehat{F}_{-j}(x)$ are respectively the empirical conditional distribution functions conditioned on Y=+1 and Y=-1 respectively.

This screening method should be used as a first attempt for binary classification.

5.3 Cui *et al* (2015) - Multiclass

For the multiclass case, Cui *et al* (2015) propose as a statistic a marginal utility defined as a weighted average of the Cramér-Von Mises distance between the unconditioned cdf and the cdfs conditioned on each class. They demonstrate that if the feature is independent of the class, then the expected Cramér-Von Mises distance is zero. This methodology once again follows the pattern of choosing as candidates features with statistics most unlike those expected in the case of independence. With technical conditions, one can prove this method also has the sure screening property. This method is also model-free, robust to potential outliers, robust to heavy-tailed distributions, and even accommodates diverging numbers of classes on the order of $O(n^k)$ for some non-negative k. Unlike previous methods, this method is applicable when the response is continuous but the features are categorical. The screening statistic for the j-th feature is

$$\widehat{MV}(X_j|Y) = \frac{1}{n} \sum_{k=1}^{K_n} \sum_{i=1}^{n} \widehat{p}_k [\widehat{F}_{jk}(X_{ij}) - \widehat{F}_j(X_{ij})] \quad (19)$$

Where X_{ij} is the j-th feature of the i-th observation,

$$\hat{p}_k = n^{-1} \sum_{i=1}^{n} I\{Y_i = y_k\} \tag{20}$$

$$\hat{F}_{jk} = n^{-1} \sum_{i=1}^{n} I\{X_{ij} \leq x,\ Y_i = y_k\} / \hat{p}_k \tag{21}$$

$$\hat{F}_j = n^{-1} \sum_{i=1}^{n} I\{X_{ij} \leq x\} \tag{22}$$

and $I\{\bullet\}$ is the indicator function.

This screening method should be used for multiclass classification problems.

5.4 Fan and Fan (2008) - Binary Classification

Fan and Fan propose a method called features annealed independence rules (FAIR). Their method first rates each feature by a two-sample t-score comparing the two classes. Fan and Fan then derive an expression for the optimum number of features to select based upon the upper bound of the misclassification rate. Once the number of features to be selected is known, features with the largest absolute t-scores are screened in. Classification then proceeds by nearest centroid in the reduced dimensional space.

Fan and Fan prove FAIR has the sure screening property. They show a numerical study comparing FAIR to nearest shrunken centroids and the independence classifier. Their numerical results show overall a strong positive association between features selected and misclassification rates. They show that FAIR and nearest shrunken centroids tend to select similar numbers of features, but FAIR tends to have lower classification error rates.

They also show several examples on real data. Datasets for leukemia, lung cancer, and prostate cancer are all analyzed: the leukemia data was classified between two forms of leukemia, the lung cancer data was classified between two forms of lung cancer, and the prostate data was classifying between cancerous and noncancerous samples. In each dataset, three groups of 100 Monte Carlo simulations are run, the three groups using 40%, 50% and 60% of the data as training data respectively. Their results show that: for leukemia data, FAIR and nearest shrunken centroids have similar rates of classification

errors; for the lung cancer data FAIR tended to perform more poorly than nearest shrunken centroids although generally with classification error rates below 5%; and for the prostate cancer data, error rates for FAIR and nearest shrunken centroids were comparable.

This method should be used as a general-purpose screening for binary classification problems.

6 Survival Models and Cox's Proportional Hazard

6.1 Survival Models Discussed

Survival analysis is a method for analyzing the time to an event (often termed a "failure"), with such events often being either mechanical failure of a device, or biological death of a patient or experimental subject, or time to recurrence of an illness. Modeling survival data is often complicated by frequent censoring of the data, meaning for many experimental units, the response is known to be in a particular range, but not known exactly. As a common example of censoring with survival data, a study of time-to-failure for a product may be halted before every experimental unit has failed. For those items that have not failed by the end of the study, their time-to-failure is longer than the length of the study, but it cannot be definitely known how much longer; the time to failure for those experimental units still operating at the end of the study is censored. In longitudinal health studies, patients may become unavailable to continue the study for any number of personal reasons, censoring the patient's time-to-event.

One common model for survival data is Cox's proportional hazard model, introduced in Cox (1972). The model is

$$h(t|x) = h_0(t)e^{x^T\beta} \tag{23}$$

and h(t) is interpreted as the instantaneous rate of failure at time t.

High-dimensional survival data is another area of active research. As with linear models and categorical data, the sparsity assumption is invoked to deal with the complexity of high-dimensional data. As before, some measure is found to initially screen

candidate features down to a classical-sized set, and classical feature selection is then performed.

6.2 Fan *et al* (2010)

Extending earlier work on SIS and ISIS, Fan *et al* develop an independence screening method for use with Cox's proportional hazard model in high-dimensions. They invoke the sparsity assumption. They define a marginal utility for a feature as the maximum of the partial likelihood of that feature, as shown in equation (24). In the case that a feature is independent of the response, the expected value of the marginal utility is zero, so following the pattern established previously, each feature is ranked by its marginal utility, and the features with marginal utilities furthest from zero are screened in.

Fan *et al* then show a feature selection method by penalized maximum likelihood, with their preference of penalty being the SCAD penalty. They show an iterative version of the screening that parallels ISIS.

They discuss variations on SIS and ISIS from Fan, Samworth, and Wu (2009). "variant 1" randomly divides the data in half, applies SIS/ISIS independently to each half, and then returns the intersection of the two lists of active features as its final estimate of the active features. The rationale being that in each half of the data, the true active features are very likely to be identified, and also some inactive features are likely to be screened in by chance, but the by-chance features in one half are less likely to be the same as the by-chance features in the other half. By intersecting the sets one expects to preserve the active features and eliminate many of the by-chance inactive ones. Variant 2 differs from variant 1 only in that the number of features screened in from each half of the data is increased to insure the intersection reaches a pre-specified size.

Fan *et al* show a number of simulation studies; 100 Monte Carlo repetition of six cases. Over the six cases they vary the size and dimensionality of the data, the size and coefficients of the true model, and the correlation structure among the features. In each case feature screening and selection is performed with both SIS and ISIS, in the original ("vanilla") form, and in each of the two variant discussed above, (total of six SIS/ISIS applications), and also LASSO for comparison. They evaluated each method in each case by; model size, probability that the final model included the true model, and difference between the true and estimated betas. To briefly summarize their results; ISIS performs

better than SIS, SIS performs much better than LASSO, but which variant of ISIS performs best (variant 1, variant 2, or vanilla) varies by case and by evaluation metric used.

Fan *et al* demonstrate their method on real data from a microarray genomic study of neuroblastoma patients due to Oberthuer *et al* (2006). Using vanilla ISIS, followed by SCAD penalized feature selection, they identify a model with 8 genes. They evaluate this model by comparing the log-likelihood of their model with the log-likelihoods of models obtained by excluding one of the 8 genes, and concluding that the mean decrease in log-likelihood was significant. They also produce 20 additional models by adding 2 random genes to their 8-gene model, and conclude the increase in log-likelihood from the additional genes is not significant.

The marginal utility, the screening statistic for the m-th feature, is defined as

$$u_m = max_{\beta_m}\left(\sum_{i=1}^{n} \delta_i x_{im}\beta_m - \sum_{i=1}^{n} \delta_i \{\sum_{j \in R(y_i)} exp(x_{jm}\beta_m)\}\right) \quad (24)$$

where R(t) = {i:$y_i \geq t$}, x_{im} is the m-th feature of the i-th observation, and δ_i is the censoring indicator for observation i. That is, $\delta_i = 0$ if the i-th observation is censored, and $\delta_i = 1$ otherwise.

This screening method should be used when the intent is to fit survival data to a Cox proportional hazard model.

7 The Fan Papers

The Fan papers, so called because they are all lead-authored by Professor Jianqing Fan of Princeton, were selected for more in-depth review because they are significant and often cited papers in the area of feature screening.

7.1 Sure Independence Screening for Ultra-High Dimensional Feature Space.

Fan and Lv (2008)

If two random variables are independent, they will have an expected correlation of zero. Based on this fact, Fan and Lv propose using pairwise correlation between the

response and each predictor to screen in features for linear models. Linear models are of the form

$$Y_i = X_i\beta + \varepsilon_i \quad (25)$$

where β is a px1 constant matrix, X_i is a 1xp random matrix, and ε_i are i.i.d. normal with zero mean and constant variance. Features with correlations furthest from zero are screened as candidates in some reasonable number, such as n-1 or the floor of n/log(n), and then the selection process is completed by some other method such as using Dantzig, SCAD, or LASSO on the candidates. Fan and Lv call this method "Sure Independence Screening" (SIS). They demonstrate that, with some technical conditions, with probability approaching one, the true active features will be selected by SIS, asymptotically as n increases without bound. They refer to this property as the "sure screening property" (SSP).

SIS is a "hard thresholding" method, meaning that features that are rated closer to zero than a specified threshold are eliminated (which comes to the same thing as fixing the coefficients of those features at zero), and the remaining features are unchanged. In contrast, a "soft thresholding" method would eliminate some features, and produce a shrinkage in coefficients of the remaining features.

Fan and Lv's statistic for feature screening is the sample Pearson correlation between the feature and the response.

Fan and Lv demonstrate SIS on a series of simulated data, and on a real dataset related to leukemia classification. The simulated data consists of 500 Monte Carlo runs of randomly generated features, a randomly generated beta, and responses generated according to the linear model with random noise. In each Monte Carlo run the data are modeled by the Dantzig selector, LASSO, SIS followed by SCAD, SIS followed by Dantzig selector, SIS followed by the Dantzig selector followed by SCAD, and SIS followed by Dantzig selector followed by adaptive LASSO. Each method was measured in terms of the median model size, and median estimation error; defined as the L2 norm of the difference between the true beta and the estimated beta. SIS followed by SCAD has the smallest median model size, and also the smallest estimation error of any of the methods.

Fan and Lv then do another series of simulated data, introducing correlation between the features. Also they increase the number of features in some of the second

series. The same six modeling methods are performed. SIS followed by SCAD continues to have the smallest median model size, and the smallest median estimation error.

Fan and Lv then apply SIS to leukemia data due to Golub *et al* (1999)[2]. These data consist of 7,129 genes and 72 samples in two classes: 47 in class ALL (acute lymphocytic leukemia), and 25 from class AML (acute myelogenous leukemia). Among those 72 samples, 38 (27 in class ALL and 11 in class AML) are assigned as training data, and the remaining 34 (20 in class ALL and 14 in class AML) are assigned as testing data.

They use two methods, SIS followed by smoothly clipped absolute deviation followed by linear discriminant (SIS-SCAD-LD), and SIS followed by smoothly clipped absolute deviation, followed by naive Bayes (SIS-SCAD-NB). To apply each method, they first apply SIS to select d = [2n/log(n)] = 20 genes with n=38.

This is the training set. Then SCAD is applied to get a family of models indexed by the regularization parameter λ. They choose the λ that produces a model the same size as the optimal number of features selected by FAIR in Fan and Fan (2008). This approach selects 16 genes. Using SIS-SCAD they obtained a linear model with 16 features. Finally, the SIS-SCAD-LD used the linear model to do the classification, and the SIS-SCAD-NB applied naive Bayes to the same group of 16 features. These methods were compared with nearest shrunken centroids. SIS-SCAD-LD has no training errors (out of a possible 38), one test error (out of a possible 34), and uses 16 genes (out of a possible 7,129). SIS-SCAD-NB has 4 training errors, one test error, and uses 16 genes. Nearest shrunken centroids has one training error, two test errors, and uses 21 genes.

Fan and Lv expand on SIS with an iterative variant, ISIS. Candidate features are selected in numbers on the order of n/log(n), and previously and newly screened features are subject to variable selection, possibly by SCAD, LASSO, or Dantzig selector. Then residuals are determined from the existing model, and a new round of screening proceeds against the residuals. The process continues until either the set of screened variables stabilizes, or until a predetermined size limit is reached, such as n-1 features. Then feature selection proceeds by a moderate-scale method such as LASSO, Dantzig, or SCAD.

One of the limitations of SIS is that it does a poor job of screening pairs of features that are collectively active in their interaction, but individually inactive. ISIS is able to

[2] Data available at http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi

select such pairs, provided one is active on its own. In general, numerical studies show ISIS to perform much better than SIS.

Fan and Lv demonstrate ISIS on a series of simulated data. They construct simulated datasets of either 100 or 1000 features, with true models with 3 to 5 active features, various covariance structures on the features, and values of n between 20 and 70. They model the various datasets by SIS, LASSO, and ISIS, evaluating each method on the fraction of the simulations in which the method correctly included all of the true active features. Note that the criterion is not including all and only the true active features; additional features are not penalized in this process. The performance of SIS and LASSO varied widely depending on the covariance structure, the value of n, the number of features, and the size of the true model. In the most difficult cases, LASSO and SIS correctly included all the true active features zero times out of 200 replications. However, the reported success rate across all the various models, replications, and covariance structures was 100% for ISIS.

They then revisit some of the earlier simulations, with the modification of changing the feature selection from SIS followed by SCAD to ISIS followed by SCAD. In terms of median model size and median estimation error of the true beta coefficient vector, ISIS-SCAD uniformly performed better than SIS-SCAD, which in the previous simulations had uniformly outperformed the other five methods evaluated (Dantzig selector, LASSO, SIS-Dantzig, SIS-Dantzig-SCAD, and SIS-DS-adaptive LASSO).

Fan and Lv then show, with some technical conditions:

- SIS has the sure screening property.
- SIS-Dantzig and SIS-SCAD give consistent estimates of the true beta coefficient vector.
- SIS-SCAD has “oracle properties”, that is, the estimated beta produced by SIS-SCAD will assign coefficients of zero to inactive features.
- estimation with the SIS-SCAD estimated beta will perform as well as if the true model were known.

7.2 High Dimensional Classification Using Features Annealed Independence Rules.

Fan and Fan (2008)

Fan and Fan propose features annealed independence rules (FAIR) for high-dimensional binary classification. Fan and Fan cite the modern trend of high throughput data, particularly from proteomics and microarray data, in which it is typical to see data with observations numbering on the order of tens, and dimensions numbering on the order of thousands. Classical methods function poorly when dimensionality is very large. Bickel and Levina (2004) show that Fisher discriminant analysis performs poorly, due to poorly conditioned matrices that are almost inevitable in high-dimensions. Even when the true covariance matrix is not ill-conditioned, in high-dimensions the sample covariance will often be singular, making Fisher's discriminant inapplicable. Bickel and Levina show that independence screening can overcome the above difficulties, but Fan and Fan demonstrate a more parsimonious method than Bickel and Levina.

According to Fan and Fan, the difficulty of high-dimensional classification stems from the large number of noise features, and the accumulation of estimation errors erode the classification rate. They argue that when most of the variability in the data can be explained with a small fraction of the features, failure to reduce the dimensionality and exclude the noise features will only increase misclassification. Parsimony is desirable not only for interpretability, but also to prevent noise accumulation.

Projection methods, such as principle component analysis, are frequently used for dimension reduction, both within and without the context of classification. Usually the directions found by theses methods place most of the weight in features with large classification power, and can perform poorly due to accumulation of noise, unless the projected vector is sparse; that is, principally built from only a small number of the original features.

Fan and Fan propose a specific form of features annealed independence rule (FAIR) that selects the m most statistically significant features, according to componentwise two-sample t-statistics between the two classes, and applying the independence classifier to the m features. They also address the question of determining the optimum value for m.

In the binary classification case, assuming both classes have identical covariance matrix Σ, the independence classification rule is

$$\delta(x) = (x-\mu)^T D^{-1}(\mu_1 - \mu_2) > 0 \qquad (26)$$

where μ_1 and μ_2 are the means of the two classes, μ is the grand mean, and $D=\text{diag}(\Sigma)$. The parameters can be estimated from the sample as

$$\hat{\mu}_k = \sum_{i=1}^{n_k} Y_{ki}/n_k;\ k = 1,2,\ \hat{\mu} = (\hat{\mu}_1 + \hat{\mu}_2)/2 \tag{27}$$

with

$$\hat{D} = diag\{(S_{1j}^2 + S_{2j}^2)/2,\ j = 1,...,p\} \tag{28}$$

and

$$S_{kj}^2 = \sum_{i=1}^{n_k} (Y_{kij} - \overline{Y}_{kj})/(n_k - 1) \tag{29}$$

and

$$\overline{Y}_{kj} = \sum_{i=1}^{n_k} Y_{ki}/n_k \tag{30}$$

With some technical conditions, Fan and Fan show that in high-dimensions the worst case misclassification rate converges in probability to 0.5 unless the signal levels are extremely high.

To identify the most useful features, Fan and Fan recommend the two-sample t-statistic. The two-sample t-statistic for feature j is defined as

$$T_j = (\overline{Y}_{1j} - \overline{Y}_{2j})/\sqrt{S_{1j}^2/n_1 + S_{2j}^2/n_2}\ ,\ j = 1 \ldots p \tag{31}$$

In this context Fan and Fan do not require the normality assumption, only that the noise vectors are i.i.d. Within each class, with mean **0** and some covariance matrix Σ_k, and also that the noise vectors are independent between classes.

With the above assumptions and some additional technical conditions, Fan and Fan prove that with probability approaching one, the two-sample t-statistic identifies all important features.

Fan and Fan apply the independence classifier to the selected features, resulting in FAIR.

Fan and Fan derive an expression for estimating the ideal number of features to select:

$$\hat{m}_0 = argmax_{1 \le m \le p} \left[\sum_{j=1}^{m} \hat{\alpha}_j^2 + m(n_1 - n_2)/(n_1 n_2) \right]^2 / [nm/(n_1 n_2) + \sum_{j=1}^{m} \hat{\alpha}_j^2] \quad (32)$$

where $\hat{\alpha}_j = \hat{\mu}_{1j} - \hat{\mu}_{2j}$

This leads to the FAIR classifier

$$\delta_{FAIR}(x) = \sum_{j=1}^{p} \hat{\alpha}_j (x_j - \mu_j) 1\{|\hat{\alpha}_j| > b\} / \hat{\sigma}_j^2 \quad (33)$$

where b is chosen so that

$$\sum_{j=1}^{p} 1\{|\hat{\alpha}_j| > b\} = \hat{m}_0 \quad (34)$$

Fan and Fan then present a simulation study. They generate 30 training observations and 200 test observation for each of two classes, with p = 4,500. The training data are used to train the FAIR classifier and the classification error for the test set is recorded. The process is repeated 100 times. As the dimensionality m grows, the misclassification rate steadily increases. When all 4,500 features are included, the mean misclassification rate is 0.2522. The misclassification rates for FAIR average 0.0154 with standard deviation 0.0085. The average optimal number of features across the 100 simulation is 29.71. By contrast nearest shrunken centroid chooses on average 28.43 features and misclassifies at a rate of 0.0216 with standard deviation 0.0179

Real data analyzed with FAIR include leukemia data due to Golub *et al* (1999)[3], prostate cancer data due to Singh *et al* (2002)[4], and lung cancer data due to Gordon *et al* (2002)[5].

The leukemia data is from high-density Affymetrix oligonucleotide arrays. There are 7,129 genes and 72 samples coming from two classes; 47 in class ALL (acute lymphocytic leukemia) and 25 from class AML (acute myelogenous leukemia). Among the 72 samples, 38 (27 in class ALL and 11 in class AML) are assigned to be training data, and 34 samples (20 in class ALL and 14 in class AML) are assigned to be test samples.

[3] Data available at http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi
[4] Data available at http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi
[5] Data available at http://datam.i2r.a-star.edu.sg/datasets/krbd/LungCancer/LungCancer-Harvard2.zip.

Before classification, Fan and Fan standardize each feature to mean zero and unit variance. Classification is done by both FAIR and nearest shrunken centroids for comparison. Both methods have a training error of one out of 38. FAIR has a better test error, one out of 34 versus three out of thirty-four. Fair selected nearly half as many genes as nearest shrunken centroids, 11 versus 21.

The data are then pooled, and over 300 Monte Carlo simulations randomly divided into training and test sets, with the proportion of training data being 40% for one hundred of the Monte Carlo simulations, 50% for another hundred simulations, and 60% for the remaining 100 simulations.

Of the 300 simulations, regardless of the fraction of the data used for training, in half of the cases FAIR has a lower error rate than nearest shrunken centroids.

The lung cancer data is used to classify type of lung cancer, the two possibilities being malignant pleural mesothelioma (MPM), and adenocarcinoma (ADCA). The data consists of 181 tissue samples, 31 MPM and 150 ADCA. The training set consists of 32 tissue samples, with 16 being ADCA and 16 being MPM. The testing set consists of the remaining 149 tissue samples; 15 from MPM and 134 from ADCA. Each sample is described by 12,533 genes.

As with the leukemia dataset, the lung cancer dataset is standardized to a mean of zero and a standard deviation of one. Again for comparison the data is classified by both FAIR and nearest shrunken centroids. With this data FAIR uses five more genes than nearest shrunken centroids (31 compared with 26), but still has better classification results. Both methods perfectly classify the training data, while FAIR has a smaller test error than nearest shrunken centroids; 7 errors out of 149 compared with 11 errors out of 149.

As was previously done with the leukemia data, the data are randomly divided into training and test sets, in three versions with the training data representing 40%, 50%, and 60% of the data in the three versions. As before each version has 100 Monte Carlo simulations. In each of the three versions, the median classification error was lower for nearest shrunken centroids than for FAIR.

The training data for prostate cancer consists of 102 patient samples, 52 of which are tumors, and 50 are normal. Approximately 12,600 gene expression levels are measured. Testing data is drawn from a separate experiment than the training data, and

consists of 25 tumor samples and 9 normal samples. The data are scaled to a variance of one and shifted to a mean of zero.

As before the data are classified by both FAIR and nearest shrunken centroids for comparison. With the prostate cancer dataset, the test error is the same (9 errors out of 34) for both methods, and the training error is larger for FAIR than nearest shrunken centroids (10 errors out of 102, compared with 8 out of 102), but FAIR uses a smaller set of features (2 genes compared with 6).

As before the data are pooled and randomly divided into training and testing sets, in three versions (100 simulations each) with 40%, 50% and 60% of the data being selected for the training set. In all three versions, the error rate for FAIR was no worse than nearest shrunken centroids.

7.3 High-Dimensional Variable Selection for Cox's Proportional Hazard Model.

Fan *et al* (2010) propose to extend their earlier work with SIS and ISIS to the context of survival data, modeled by Cox's proportional hazard model. Instead of Pearson correlation as a measure of independence, they now propose a marginal utility, defined as the maximum of the partial likelihood of a single covariate. They derive a formula for the partial likelihood that accounts for censoring of the data. As before, each covariate is ranked by the marginal utility statistic, and those features or covariates with utilities furthest from zero are preferentially chosen as potential features. As before, a shrinkage process is applied while determining the coefficients, which may have the effect of un-selecting a feature previously selected. Their chosen method of shrinkage-selection is smoothly clipped absolute deviation (SCAD). As with their previous work on sure independence screening (SIS), they recommend an iterative version of their method, to better identify features that are individually inactive, but active in their interaction with other features. The marginal utility, the screening statistic, is defined as

$$u_m = max_{\beta_m}\left(\sum_{i=1}^{n}\delta_i x_{im}\beta_m - \sum_{i=1}^{n}\delta_i\left\{\sum_{j\in R(y_i)} exp(x_{jm}\beta_m)\right\}\right) \qquad (35)$$

where R(t) = {i:$y_i \geq t$}, and δ_i is the censoring indicator for observation i. That is, δ_i = 1{i-th observation is not censored}

Following a variation suggested in Fan, Samworth, and Wu (2009), they propose dividing the data, screening each part of the data separately, and finding the intersections of the two lists of screened features. This method reduces false selections, as true features are likely to be in both screened sets, but false selections are likely to appear in only one. They call this "version 1", a label that can be applied to both SIS and ISIS. They note that this can quite quite drastically reduce the size of the pool of potential features, as a product of the reduction in false selections. To insure a larger pool of potential features will be screened, Fan, Samworth and Wu proposed "version 2", in which the data are still divided, and the intersection of the two lists of screened features is still used, but each half of the data the number of features screened is increased, allowing for a larger intersection of the two lists. To further distinguish these two variants, the original forms of SIS and ISIS are renamed "vanilla" SIS and ISIS respectively. Fan *et al* demonstrate these methods retain the sure screening property even in the context of survival data. They further demonstrate their methods on simulated and real data. The simulated data consists of six cases, each with different true beta coefficients, previously chosen by a specific random method, (inactive features indicated by a zero coefficient), and with different covariance structures among the features. In each case, 100 Monte Carlo simulations are performed. Each of the six cases was subjected to feature selection by six or seven methods, vanilla SIS and ISIS, variant 1 SIS and ISIS, variant 2 SIS and ISIS, and LASSO (but LASSO not being done with cases five or six). The various forms of SIS were followed by or incorporated SCAD as final feature selection. Tuning parameters for SCAD and LASSO were selected by Bayes Information Criterion. In all six cases the 100 Monte Carlo runs were evaluated by five criterion:

1) L1 norm of the difference between the true and estimated beta
2) the square of the L2 norm of the difference between the true and estimated beta
3) the proportion of runs in which the true model was included in the selected features
4) the proportion of runs in which the final model has the sure screening property, and
5) the median model size.

In cases one through four, LASSO was the worst performer by most criteria, but surprisingly not all. In all six cases, the iterative variants had median model sizes that

agreed with the true models. In all six cases, none of the *iterative* variants did poorer than 99% at correctly including the true model. In all six cases by all five criteria, the worst of the iterative variants was better than the best of the non-iterative variants. Which iterative variant performed best varied by case and criterion, but vanilla ISIS was best or tied for best most often.

Fan *et al* also demonstrated their methods on real data. They modeled survival data on neuroblastoma patients, originally due to Oberthuer *et al* (2006). Neuroblastoma is "an extracranial solid cancer. It is most common in childhood or infancy. In the United States several hundred new cases are reported each year. Neuroblastoma is a malignant pediatric tumor originating from the neural crest elements of the sympathetic nervous system." (Fan *et al,* pg 83) The study due to Oberthuer *et al* includes 251 patients of the German Neuroblastoma Trials NB90-NB2004, who were diagnosed between 1980 and 2004. The patients' ages range from zero to 296 months at diagnosis, with a median age of 15 months. Neuroblastoma specimens were analyzed from all 251 patients using a customized oligonucleotide microarray (a customized device for measuring abundances of specific short segments of DNA or RNA in a sample). The microarray comprised 10,163 probes. The goal was to study associations between gene expression and various clinical information, including survival time and three-year event-free survival.

Fan *et al* excluded five arrays as outliers, but do not elaborate on their criterion for outlier, even though the original Oberthuer paper does not mention outliers. Nor do Fan *et al* explain why the data were excluded, other than to label them as 'outliers'. Of the remaining 246 patients, 205 were censored as to overall survival time. Fan argues for using an iterative version of his methods, on the grounds that complex interaction effects are a reasonable supposition for complex oligonucleotide data. Each predictor was standardized to have a mean of zero and a variance of one. Screening and modeling processes by vanilla-ISIS, with each iteration selecting a number of genes d = n/log(n) = 43. ISIS is followed by SCAD penalized Cox regression. Their final model uses just 8 genes.

Analysis of a model of real data is hampered in that often we don't know the true model with real data. A X^2 test shows the model with these 8 selected genes to be significant. Based on the estimated coefficients and their standard errors, 6 of the 8 coefficients are significant at the $\alpha = 0.01$ level. Fitting the Cox model with only 7 of the features selected decrease the log-likelihood by (on average) 11.9431. Adding 2

additional features increases the log-likelihood by (on average) 0.9584. Based on these facts, Fan *et al* conclude the eight selected genes are “very important.”

8 Replication of Screening Methods

I selected a number of examples, drawn from the Fan papers, to replicate on simulated and real data. This section presents my results. All of my code is publicaly available at https://github.com/johnwauters/Thesis.

8.1 Methods from Fan and Lv (2008)

Following parts of section 3.3.1 in Fan and Lv (2008), the Dantzig selector, LASSO, SIS-SCAD, and SIS-Dantzig methods were all applied to sets of simulated data (500 replications). All the features of the simulated data are independent of each other, p=1000, n=200, true model size = 8. The four methods were evaluated and compared in terms of median model size, and median estimation error (L2 norm of difference between true coefficient beta vector and estimated coefficient beta). The tuning parameter for SCAD is tuned by BIC. The tuning parameter for LASSO is tuned by 10-fold cross-validation.

Table 1

	Dantzig Selector	LASSO	SIS-SCAD	SIS-Dantzig
Median model size	181.0	63	60.0	99.0
Median estimation error	1.360	0.883	1.303	1.314

Results of 500 Monte Carlo simulation of linear models, n=200, p=1000, true model size = 8, estimation error defined as the Euclidean norm of the difference between the true and estimated β vectors.

Following parts of section 3.3.2 in Fan and Lv, SIS-SCAD, and SIS-Dantzig were all applied to 200 replications of simulated data, p=20,000, true model size=14, n=800. The screening methods are again evaluated in terms of median selected model size and median estimation error (as defined above). The SCAD tuning parameter is again tuned by BIC.

Table 2

	SIS-SCAD	SIS-Dantzig	SIS-DZ-SCAD
Median model size	142.5	115.0	104
Median estimation error	2.783	2.786	2.762

Results of 200 Monte Carlo simulations of SIS-based screening and selection methods, with correlated features. Estimation error is defined as for the previous table. True model size = 14, p = 20,000, and n = 800.

Table 3

		n=20	n=50	n=70
p=100	SIS	0	.251	.645
p=100	LASSO	0	0.290	.890
p=100	ISIS	0.231[6]	1	1

Results of 200 Monte Carlo simulations, with correlated features. Reported value is the empirical probability of correctly including the true model in the result of the screening process.

[6] Of the results of my attempts to replicate the methods in this thesis, this one result (0.231 for ISIS when n=20) is only one I would say is clearly not close to the published results. I am at a loss to explain why I am getting this result in this case, even after working with a consulting statistician and expert programmer.

Following parts of 4.2.3 in Fan and Lv; SIS, ISIS, and LASSO all compared in terms of ability to extract the true model from 200 replications of simulated data, p=100, n = 20, 50, 70. The three methods are evaluated in terms of accuracy at including true model.

Following section 3.3.3 of Fan and Lv, applying SIS to real leukemia data due to Golub *et al* (1999).[7] The data consist of 72 samples with 7,129 gene expression levels measured by Affymetrix oligonucleotide arrays. The 72 observations consist of 47 of class ALL (acute lymphocytic leukemia) and 25 of class AML (acute myelogenous leukemia). The data are divided into training data (27 in class ALL and 11 in class AML), and testing data (20 in class ALL and 14 in class AML). The leukemia data are screened by SIS followed by SCAD, and then classified by both linear discriminant and by naive Bayes classification. The data is also classified by nearest shrunken centroid for comparison.

8.1.1 Additional Variant

In Fan and Lv's application of their method to the leukemia data, they divided the data into training and test sets only once. Also, Fan and Lv compared SIS-SCAD linear discriminant and SIS-SCAD naive Bayes to nearest shrunken centroids: the very different methods making it difficult to say if the difference in performance is due to the SIS, or the SCAD, or the method applied after the SIS-SCAD.

Table 4

Method	Training error	Test error	Number of genes
SIS-SCAD-LD	0	7	16
SIS-SCAD-NB	0	2	16
NSC	0	1	34

Classification of leukemia data by SIS-SCAD linear discriminant, SIS-SCAD naive Bayes, and nearest shrunken centroid.

Extending beyond Fan and Lv's analysis, a series of Monte Carlo simulations divide the available data into test and training sets. Of the 47 observations of class ALL,

[7] Data available at http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi

23 are selected for training data, and of the 25 observations of class AML, 12 are selected for training data. All observations not used as training data are used as test data. In this extension nearest shrunken centroids is compared with SIS followed by nearest shrunken centroids, a more direct comparison. After 300 Monte Carlo simulations, it seems clear that using SIS before using nearest shrunken centroids has no significant improvement in number of genes selected, or training error, or test error rates.

Table 5

	Improvement		
	Mean	Standard deviation	Empirical 95% confidence interval
Number of genes selected	267.8	799.0175	(-58.525, 1959.650)
Training error	0.1833	0.3875861	(0, 1)
Test error	-0.1933	1.166401	(-3, 2)

Here improvement is the difference between the values for SIS followed by nearest shrunken centroids, and ordinary nearest shrunken centroids. Results based on 300 Monte Carlo simulations.

8.2 Methods from Fan and Fan (2008)

Following section 5.2.3 of Fan and Fan, replication of classification of prostate cancer data. The prostate data is due to Singh *et al* (2002)[8]. These data consist of 136 patient samples, 59 are normal tissue and 77 are tumors. For each patient sample 12,600 gene expression levels are measured. The data are initially grouped into training and test sets; the initial training set contains 52 tumor observations and 50 normal observations, the test set contains 9 normal samples and 25 tumor samples. Features are pre-processed to be centered at zero and scaled to a standard deviation of one. The data are classified using Fan and Fan's features annealed independence rules (FAIR), and by nearest shrunken centroids (NSC) for comparison.

[8] Data available at http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi

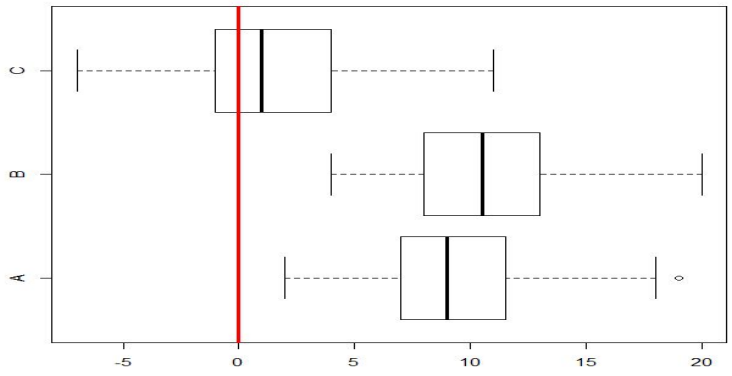
After that initial analysis, the training and test data are pooled. The data are then randomly divided into training and test sets over a series of Monte Carlo runs. A total of 300 Monte Carlo simulations are run, 100 each with the training data representing 40%, 50%, and 60% of the data respectively. For each Monte Carlo run, FAIR is compared to NSC in terms of test error and number of genes selected.

Table 6

	Training error	Test error	Number of selected genes
NSC	7/102	2/34	52
FAIR	8/102	4/34	5

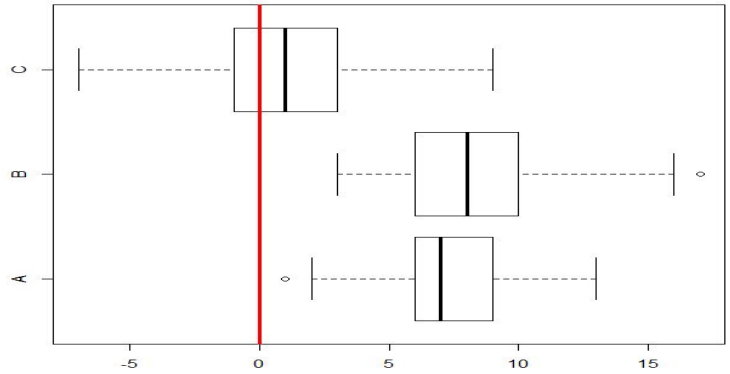
Initial classification of prostate cancer data by nearest shrunken centroids and by features annealed independence rules.

Figure 2



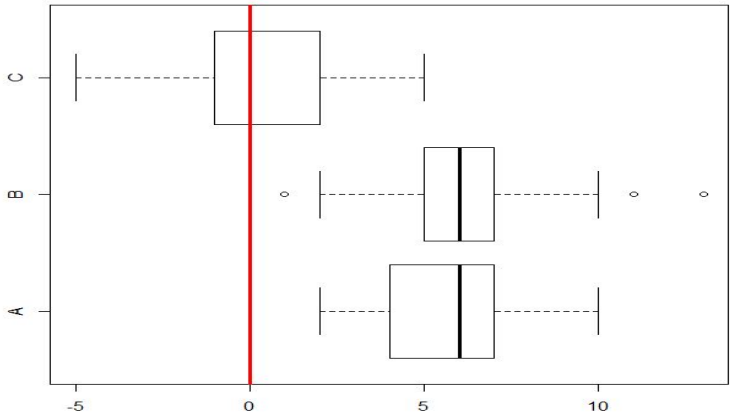
Boxplots of 100 test error rates, with 40% of prostate cancer data as training data. A = test error for FAIR, B = test error for nearest shrunken centroids, C = difference between FAIR and NSC. Red line marks zero.

Figure 3



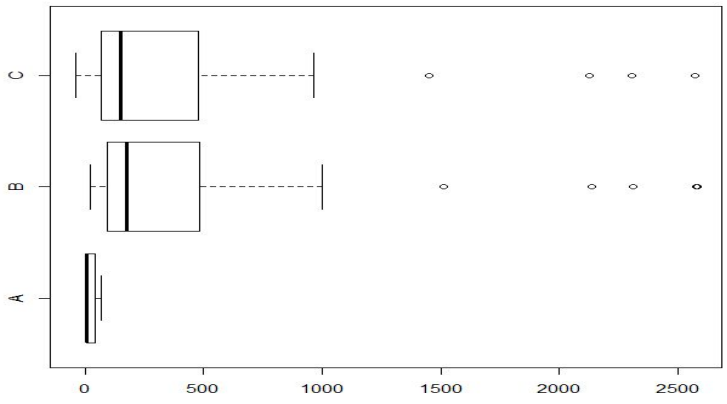
Boxplots of 100 test error rates, with 50% of data as training data. A = test error for FAIR, B = test error for nearest shrunken centroids, C = difference between FAIR and NSC. Red line marks zero.

Figure 4



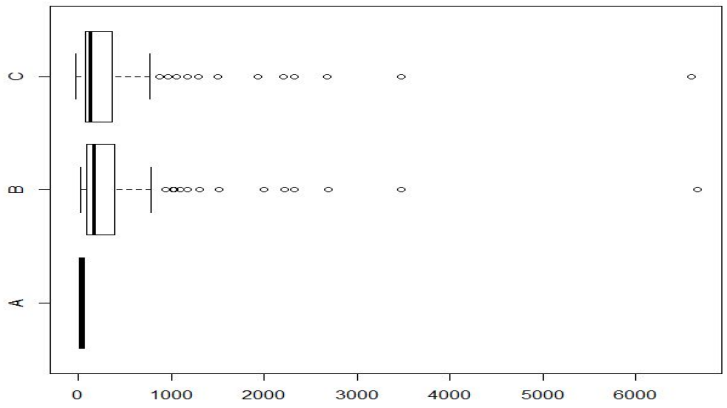
Boxplots of 100 test error rates, with 60% of data as training data. A = test error for FAIR, b = test error for nearest shrunken centroids, C = difference between FAIR and NSC. Red line marks zero.

Figure 5



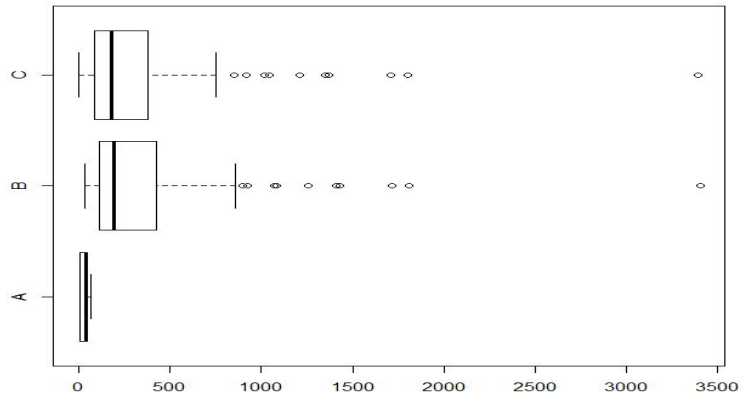
Boxplots of 100 gene selection rates, with 40% of data as training data. A = genes selected by FAIR, B = genes selected by nearest shrunken centroids, C = difference between NSC and FAIR.

Figure 6



Plots of 100 gene selection rates, with 50% of data as training data. A = genes selected by FAIR, B = genes selected by nearest shrunken centroids, C = difference between NSC and FAIR.

Figure 7



Plots of 100 gene selection rates, with 60% of data as training data. A = genes selected by FAIR, B = genes selected by nearest shrunken centroids, C = difference between NSC and FAIR.

8.3 Methods from Fan *et al* (2010)

Following Fan *et al* section 5, cases 5 and 6, simulated data are generated; n = 400, p = 1000. In case 5, all pairs of features have a correlation of 0.5; in case 6 feature 5 is uncorrelated with all the other features, feature 4 has correlation of $1/\sqrt{2}$ with all features except 4 and 5, and all other pairs of features have correlations of 0.5. Case 5 has a true model size of 6 and 23% censoring rate. Case 6 has a true model size of 5 and 36% censoring rate.

One-hundred Monte Carlo simulations are run for each case. In each simulation model screening is performed by vanilla ISIS and variant 1 ISIS (aggressive version based on splitting the data and intersecting the sets of screened in features). Each simulation is evaluated on:

- the difference between the true and estimated beta.
- the probability that the estimated model includes the true model.
- the size of the estimated model.

Fan *et al*'s methods are then applied to real data due to Rosenwald *et al* (2002), and later used in Bair and Tibshirani (2004).[9] The data is collected from 240 post-chemotherapy diffuse large-B-cell lymphoma patients. The data consists of DNA microarray gene expression levels for 7,399 genes, and survival times in years, with 129 of the 240 survival times being censored.

Following the pattern of Fan *et al*'s analysis of neuroblastoma data (section 6 of Fan *et al*), ISIS was applied to the data, and 10 features were selected for the final model. Following the pattern in Fan *et al*, the selected model was compared to each of the possible 9-feature submodels in terms of its log-likelihood. By removing one of the features, the log-likelihood decreases by an average of 4.950635 with a standard deviation of 3.637027808.

Continuing to follow the model of Fan *et al*, two randomly selected features are added to the model, and the change in log-likelihood is observed. Over 200 replications of

Table 7

	Vanilla ISIS	Variant 1 ISIS
$\|\beta - \hat{\beta}\|_1$	2.408319	1.719573
$\|\beta - \hat{\beta}\|_2^2$	0.4782417	0.3405675
P	1	1
MS	16	12

Table summarizes case 5. Reported values are medians over 100 Monte Carlo simulations. P is empirical probability that the estimated model includes the true model. MS is model size.

adding features, the change in log-likelihood averaged 1.042930108, with a standard deviation of 0.971471443. Of the 200 randomly selected pairs of features added to the model, none changed the log-likelihood more than 4.13, which is less than the average

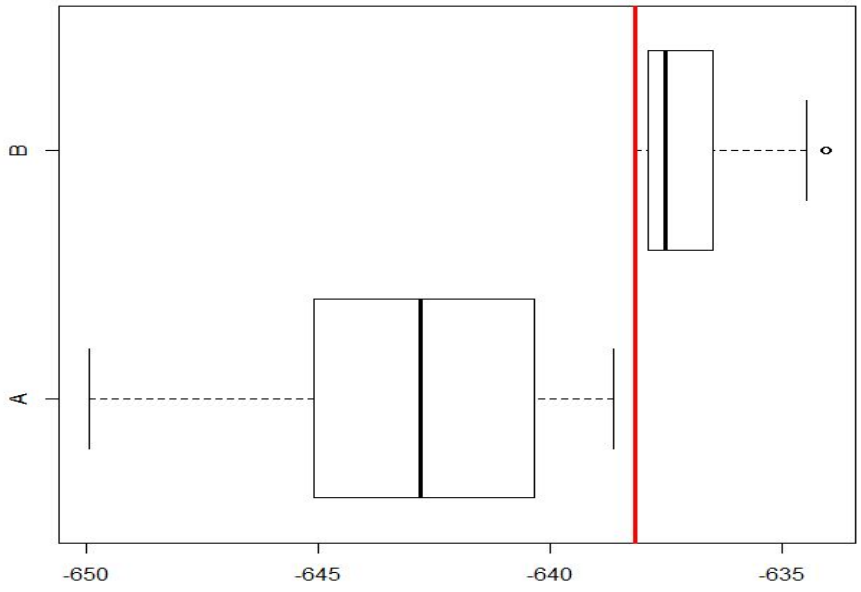
[9] Lymphoma data available at http://statweb.stanford.edu/~tibs/superpc/

Table 8

	Vanilla ISIS	Variant 1 ISIS
$\|\beta - \hat{\beta}\|_1$	1.914122	2.085366
$\|\beta - \hat{\beta}\|_2^2$	0.4126624	0.3831478
P	1	1
MS	16	16

Table summarizes case 6. Reported values are medians over 100 Monte Carlo simulations. P and MS are as in the previous table.

Figure 8



Plots of log-likelihoods of models of the lymphoma data. The red line represents the ten-feature model found by ISIS. A = ten models produced by removing one feature from the found model, B = two-hundred models produced by adding two randomly selected features to the found model.

change produced by omitting one of the selected features. Over a series of 20,000 Monte Carlo cycles of adding two randomly selected features to the model, more than 99% produced a smaller change in the log-likelihood than the mean change from removing one of the 10 selected features. If the above is interpreted as an approximate permutation test, it seems reasonable to say that the 10 identified features are much more important than an average feature.

9 Concluding Remarks

This thesis presented a selective overview of methods of feature screening for high-dimensional data. High-dimensional data is increasingly common as technology advances, and efficient methods of analyzing and modeling such data are needed. There are modeling methods for high-dimensional data that do not require feature screening first, but simulation studies show screening-first methods perform better than non-screening methods as dimensionality increases. A wide variety of screening methods exist, but there is a surprising consistency to them; very often these methods:

- Invoke the sparsity assumption.
- Calculate some statistic for each feature.
- Choose a statistic has a known specific expected value in the case that the feature and the response are independent.
- Screen in features most unlike the independent case, the number of features screened in often defaulting to n-1.
- Are implemented in an iterative version, in hopes of improved performance.

This thesis then presented a more in-depth review of three significant and often cited papers on feature screening in high-dimensions: Sure Independence Screening for Ultra-High Dimensional Feature Space, Fan and Lv (2008); High-Dimensional Variable Selection for Cox's Proportional Hazard Model, Fan *et al* (2010); and High Dimensional Classification Using Features Annealed Independence Rules, Fan and Fan (2008).

This thesis also presents results of replicating some of the methods described, on simulated and real data.

Current methods do poorly at identifying important interactions among the features. Iterative variants do better, but are usually dependent on a subset of the

interacting features being active and screened in individually. Finding practical methods of identifying important interactions in the most general case, without depending on most of the features being individually active, is an area of ongoing research.

Many of the independence screening methods screen in n-1 features, without solid argument for using that number of features. (FAIR being a notable exception.) Development of data-driven thresholds for feature screening is another area of ongoing research.

There does not appear to be much literature on combining feature screening with classification trees or regression trees, this may be an interesting line of research.

References

Bickel, P. J., & Levina, E. (2004). Some theory for Fisher's linear discriminant function,'naive
Bayes', and some alternatives when there are many more variables than observations. Bernoulli, 989-1010.

Bair, E., & Tibshirani, R. (2004). Semi-supervised methods to predict patient survival from gene expression data. *PLoS Biol*, *2*(4), e108.

Candes, E., & Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 2313-2351.

Cox, D. R. (1972) Regression and models with life tables. *J. Roy. Stat. Soc. Ser. B* **34** 187-220

Cui, H., Li, R., & Zhong, W. (2015). Model-free feature screening for ultrahigh dimensional discriminant analysis. *Journal of the American Statistical Association*, *110*(510), 630-641.

Fan, J., & Fan, Y. (2008). High dimensional classification using features annealed independence rules. *Annals of statistics*, *36*(6), 2605.

Fan, J., Feng, Y., & Wu, Y. (2010). High-dimensional variable selection for Cox's proportional hazards model. In *Borrowing Strength: Theory Powering Applications–A Festschrift for Lawrence D. Brown* (pp. 70-86). Institute of Mathematical Statistics.

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, *96*(456), 1348-1360.

Fan, J., & Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *70*(5), 849-911.

Fan, J., Samworth, R., & Wu, Y. (2009). Ultrahigh dimensional feature selection: beyond the linear model. *Journal of Machine Learning Research*, *10*(Sep), 2013-2038.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., ... & Bloomfield, C. D. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *science*, *286*(5439), 531-537.

Gordon, G. J., Jensen, R. V., Hsiao, L. L., Gullans, S. R., Blumenstock, J. E., Ramaswamy, S., ... & Bueno, R. (2002). Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. *Cancer research*, *62*(17), 4963-4967.

Hall, P., & Miller, H. (2012). Using generalized correlation to effect variable selection in very high dimensional problems. *Journal of Computational and Graphical Statistics*.

He, X., Wang, L., & Hong, H. G. (2013). Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data. *The Annals of Statistics*, *41*(1), 342-369.

Li, G., Peng, H., Zhang, J., & Zhu, L. (2012). Robust rank correlation based screening. *The Annals of Statistics*, 1846-1877.

Li, R., Zhong, W., & Zhu, L. (2012). Feature screening via distance correlation learning. *Journal of the American Statistical Association*, *107*(499), 1129-1139.

Liu, J., Li, R., & Wu, R. (2014). Feature selection for varying coefficient models with ultrahigh-dimensional covariates. *Journal of the American Statistical Association*, *109*(505), 266-274.

Liu, J., Zhong, W., & Li, R. (2015). A selective overview of feature screening for ultrahigh-dimensional data. *Science China Mathematics*, *58*(10), 1-22.

Mai, Q., & Zou, H. (2012). The Kolmogorov filter for variable screening in high-dimensional binary classification. *Biometrika*, ass062.

Oberthuer, A., Berthold, F., Warnat, P., Hero, B., Kahlert, Y., Spitz, R., ... & Schwab, M. (2006). Customized oligonucleotide microarray gene expression–based classification of neuroblastoma patients outperforms current clinical risk stratification. *Journal of clinical oncology*, *24*(31), 5070-5078.

Rosenwald, A., Wright, G., Chan, W. C., Connors, J. M., Campo, E., Fisher, R. I., ... & Hurt, E. M. (2002). The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *New England Journal of Medicine*, *346*(25), 1937-1947.

Singh, D., Febbo, P. G., Ross, K., Jackson, D. G., Manola, J., Ladd, C., ... & Lander, E. S. (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer cell*, *1*(2), 203-209.

Székely, G. J., Rizzo, M. L., & Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, *35*(6), 2769-2794.

ten Broeke-Smits, N. J., Pronk, T. E., Jongerius, I., Bruning, O., Wittink, F. R., Breit, T. M., ... & Boel, C. E. (2010). Operon structure of Staphylococcus aureus. *Nucleic acids research*, *38*(10), 3263-3274.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267-288.

Tibshirani, R., Hastie, T., Narasimhan, B., & Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences*, *99*(10), 6567-6572.

Wang, H., Li, R., & Tsai, C. L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika*, *94*(3), 553-568.

Zhu, L. P., Li, L., Li, R., & Zhu, L. X. (2012). Model-free feature screening for ultrahigh-dimensional data. *Journal of the American Statistical Association*.