Sure Independence Screening *

Jianqing Fan and Jinchi Lv
Princeton University and University of Southern California

August 16, 2017

Abstract

Big data is ubiquitous in various fields of sciences, engineering, medicine, social sciences, and humanities. It is often accompanied by a large number of variables and features. While adding much greater flexibility to modeling with enriched feature space, ultra-high dimensional data analysis poses fundamental challenges to scalable learning and inference with good statistical efficiency. Sure independence screening is a simple and effective method to this endeavor. This framework of two-scale statistical learning, consisting of large-scale screening followed by moderate-scale variable selection introduced in Fan and Lv (2008), has been extensively investigated and extended to various model settings ranging from parametric to semiparametric and nonparametric for regression, classification, and survival analysis. This article provides an overview on the developments of sure independence screening over the past decade. These developments demonstrate the wide applicability of the sure independence screening based learning and inference for big data analysis with desired scalability and theoretical guarantees.

Key words: Big data; Scalability; Sure independence screening; Iterative sure independence screening; Two-scale statistical learning; Ultra-high dimensionality; Dimensionality reduction; Sure screening property; Sparsity; Efficiency

*Jianqing Fan is Frederick L. Moore '18 Professor of Finance, Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544, USA (E-mail: jqfan@princeton.edu). Jinchi Lv is McAlister Associate Professor in Business Administration, Data Sciences and Operations Department, Marshall School of Business, University of Southern California, Los Angeles, CA 90089, USA (E-mail: *jinchilv@marshall.usc.edu*). This work was supported by National Science Foundation Grants DMS-1662139 and DMS-1712591 and a grant from the Simons Foundation. We sincerely thank the Editor of Wiley StatsRef: Statistics Reference Online for the kind invitation to write this review article.

1 Introduction

Big data has emerged in recent years as a prominent feature of many applications from different disciplines of sciences, engineering, medicine, social sciences, and humanities, enabling more capacity for refined discoveries, recommendations, and policies [18]. Among many types of big data, ultra-high dimensional data in which the number of features p can be much larger than the number of observations n is central to a spectrum of tasks of statistical learning and inference in the past ten years or so. Scalability is a major challenge of ultra-high dimensional data analysis. Meanwhile it is well known that additional intrinsic challenges of ultra-high dimensional data analysis include high collinearity, spurious correlation, and noise accumulation [20, 15, 21, 22]. For example, in the presence of a large number of noise features high-dimensional classification using all the features can behave like random guess [15]. To improve scalability and reduce noise accumulation, one possible approach is reducing the dimensionality of the feature space from a very large scale to a moderate one in a computationally fast way and implementing refined learning and inference in the much reduced feature space.

The ideas of feature screening have been widely employed in practice partly for computational reasons. In addition to the gain in computational efficiency, one can in fact also expect improved statistical efficiency in estimation and inference due to alleviated noise accumulation by dimensionality reduction. For the aforementioned classification problem, one can reduce the number of features by applying two-sample t-test to each variable and then implement a classification procedure using the selected variables. This approach is a specific case of the sure independence screening and has high classification power [15].

To appreciate the point, let us consider a simple simulation example which provides a prototype for the common goals desired by practitioners analyzing high-dimensional data. We generate 100 data sets from Gaussian linear model given in (1) with sample size $n = 120$, dimensionality $p = 1000$, design matrix $\mathbf{X} \sim N(\mathbf{0}, I_n \otimes \mathbf{\Sigma})$ for $\mathbf{\Sigma} = (0.5^{|j-k|})_{1\le j,k\le p}$, and error vector $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 I_n)$ for $\sigma = 1$. The true regression coefficient vector $\boldsymbol{\beta}_0$ has the first $s = 10$ components being nonzero and each nonzero component is selected randomly from $\{\pm 1\}$. For each data set, we apply the model-free knockoffs method introduced in [7] coupled with the ISEE estimator in [27] to control the false discovery rate (FDR) [2] for feature selection, where the target FDR level is set as $q = 0.2$. For sparse model fitting, we employ Lasso [49], SCAD [19], and sure independence screening (SIS) [20] followed by the Lasso and SCAD as variable selectors, referred to as SIS-Lasso and SIS-SCAD, respectively. With the set of identified covariates $\widehat{S}$ by the model-free knockoffs procedure, we can also construct an estimate for the error standard deviation $\widehat{\sigma}$. Table 1 summarizes the simulation

results for the FDR, power, and estimated error standard deviation $\widehat{\sigma}$ over 100 replications. From Table 1, we see that feature screening using SIS can also boost the accuracy of large-scale estimation and inference.

Table 1: The means (standard errors) of different measures by all the methods for the simulation example in Section 1

Method	FDR	Power	$\widehat{\sigma}$
Lasso	0.158(0.015)	0.789(0.030)	1.248(0.039)
SCAD	0.150(0.018)	0.711(0.038)	1.224(0.045)
SIS-Lasso	0.167(0.017)	0.841(0.029)	1.173(0.041)
SIS-SCAD	0.147(0.017)	0.903(0.025)	1.033(0.032)

2 Sure independence screening

We now begin the journey of feature screening in ultra-high dimensional feature space. A common practice for feature screening is using independence learning which treats the features as independent and thus applies marginal regression techniques. Yet the theoretical properties of such computationally expedient procedures were not well understood for a long while. Motivated by the aforementioned fundamental challenges of ultra-high dimensional data analysis, the sure independence screening (SIS) was formally introduced and rigorously justified in [20] to address both issues of scalability and noise accumulation. Let us consider the linear regression model

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

where $\mathbf{y} = (y_1, \cdots, y_n)^T$ is an n-dimensional response vector, $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_p)$ is an $n \times p$ design matrix consisting of p covariates $\mathbf{x}_j$'s, $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)^T$ is a p-dimensional regression coefficient vector, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \cdots, \varepsilon_n)^T$ is an n-dimensional error vector. The focus of [20] is the ultra-high dimensional setting with $\log p = O(n^{\alpha})$ for some $0 < \alpha < 1$. To ensure model identifiability, the true regression coefficient vector $\boldsymbol{\beta}_0 = (\beta_{0,1}, \cdots, \beta_{0,p})^T$ is assumed to be sparse. The covariates $\mathbf{x}_j$'s with indices in the support $\mathcal{M}_* = \text{supp}(\boldsymbol{\beta}_0) = \{1 \le j \le p : \beta_{0,j} \neq 0\}$ are called important variables, while the remaining covariates are referred to as noise variables.

The SIS is a two-scale learning framework in which large-scale screening is first applied to reduce the dimensionality from p to a moderate one d, say, below sample size n, and moderate-scale learning and inference are then conducted on the much reduced feature space.

In particular, the SIS ranks all the p features using the marginal utilities based on the marginal correlations $\widehat{\text{corr}}(\mathbf{x}_j, \mathbf{y})$ of $\mathbf{x}_j$'s with the response $\mathbf{y}$ and retains the top d covariates with the largest absolute correlations collected in the set $\widehat{\mathcal{M}}$; that is,

$$\widehat{\mathcal{M}} = \{1 \le j \le p : |\widehat{\text{corr}}(\mathbf{x}_j, \mathbf{y})| \text{ is among the top } d \text{ largest ones}\}, \tag{2}$$

where $\widehat{\text{corr}}$ denotes the sample correlation. This achieves the goal of variable screening. The variable selection step of SIS using features in the reduced set $\widehat{\mathcal{M}}$ from the screening step can be done with any favorite regularization method of user's choice including Lasso, SCAD, and Dantzig selector [49, 19, 14, 52, 6, 4, 34, 21, 5, 22, 35]. The SIS ideas can also be incorporated into large-scale Bayesian estimation and inference, where the marginal utilities can be replaced by the Bayesian counterpart [44, 29].

The feature screening (2) can be implemented expeditiously. An important question is whether it contains all the important covariates in the set $\mathcal{M}_*$ with asymptotic probability one; that is,

$$\mathbb{P}\left\{\mathcal{M}_* \subset \widehat{\mathcal{M}}\right\} \to 1 \tag{3}$$

as $n \to \infty$. The property in (3) was termed as the sure screening property in [20] which is crucial to the second step of refined variable selection. Surprisingly, SIS was shown in [20] to enjoy the sure screening property under fairly general conditions, with a relatively small size of $\widehat{\mathcal{M}}$. Specifically, the p covariates $\mathbf{x}_j$'s are allowed to be correlated with covariance matrix $\mathbf{\Sigma}$ and the p-dimensional random covariate vector multiplied by $\mathbf{\Sigma}^{-1/2}$ is assumed to have a spherical distribution. The sure screening property of SIS depends upon the so-called concentration property for random design matrix $\mathbf{X}$ introduced in [20]; see [30] for similar concentration phenomenon of large random design matrix.

The concentration property was originally verified for the scenario of Gaussian distributions, and later established in [45] for a wide class of elliptical distributions as conjectured previously. With such a property, the sure screening property (3) can hold for $d = o(n)$, leading to the suggestion of choosing $d = n - 1$ or $[n/(\log n)]$ for SIS in the original paper [20]. In practice, the parameter d can be chosen by some data-driven methods such as the cross-validation and generalized information criterion [28]. It can also be selected by a simple permutation method [17, 53] that controls the false positive rate at a prescribed level q. Let $\{(\mathbf{X}_i, Y_i)\}_{i=1}^n$ be the original sample for the covariates and response. One can apply a random permutation π of $\{1, \cdots, n\}$ to obtain the randomly permuted decoupled data $\{(\mathbf{X}_{\pi(i)}, Y_i)\}_{i=1}^n$. This does not change the marginal distributions of $\{\mathbf{X}_i\}_{i=1}^n$ and $\{Y_i\}_{i=1}^n$, but makes the associations between covariates and response in $\{(\mathbf{X}_{\pi(i)}, Y_i)\}_{i=1}^n$ vanish. For the randomly permuted data, denote by r^* the top qth percentile of the absolute marginal

sample correlation, which has proportion q of false positive rate when applied to the randomly decoupled data. When $q = 1$, r^* is merely the largest spurious correlation. Now select the model

$$\widehat{\mathcal{M}} = \{1 \le j \le p : |\widehat{\text{corr}}(\mathbf{x}_j, \mathbf{y})| \ge r^*\}. \tag{4}$$

One can also randomly permute the data multiple times and use the median of r^* to improve the stability. This simple permutation idea is applicable to other screening methods discussed in this article.

3 Iterative and conditional sure independence screening

Since the marginal utilities are employed to rank the importance of features, SIS can suffer from some potential issues associated with independence learning. First, some noise covariates strongly correlated with the important ones can have higher marginal utilities than other important ones. Second, some important covariates that are jointly correlated but marginally uncorrelated with the response can be missed after the screening step. To address these issues, [20] further introduced an extension of the SIS method, called the iterative SIS. The main idea is to iteratively update the estimated set of important variables using SIS conditional on the estimated set of variables from the previous step. Intuitively, such an iterative procedure can help recruit important covariates that have very weak or no marginal associations with the response in the presence of other important ones identified from earlier steps. The method of iterative SIS was extended in [23] to the pseudo-likelihood framework beyond the linear model with more general loss functions. [23] also introduced a sample splitting strategy to reduce the false positive rate, where some exchangeability conditions were invoked.

When there is some additional knowledge about the importance of a certain set of covariates, it is helpful to utilize this prior information and rank the importance of features by replacing simple marginal correlations with the marginal correlations conditional on such a set of variables. This approach of conditional SIS was introduced and justified in [1]. It also intends to provide understandings on the iterative SIS.

4 Sure independence screening for generalized linear models and classification

When the response is discrete, it is more suitable to consider the fitting of models beyond the linear one. The generalized linear model (GLM) provides a natural extension of the

linear model for both continuous and discrete responses. The GLM with a canonical link assumes that the conditional distribution of response $\mathbf{y}$ given design matrix $\mathbf{X}$ belongs to the canonical exponential family, having the following density function with respect to some fixed measure

$$f_n(\mathbf{y}; \mathbf{X}, \boldsymbol{\beta}) \equiv \prod_{i=1}^{n} f_0(y_i; \theta_i) = \prod_{i=1}^{n} \left\{ c(y_i) \exp \left[\frac{y_i \theta_i - b(\theta_i)}{\phi} \right] \right\}, \tag{5}$$

where $\{f_0(y;\theta) : \theta \in \mathbb{R}\}$ is a family of distributions in the regular exponential family with dispersion parameter $\phi \in (0, \infty)$, $(\theta_1, \cdots, \theta_n)^T = \mathbf{X}\boldsymbol{\beta}$, $b(\cdot)$ and $c(\cdot)$ are some known functions, and the remaining notation is the same as in model (1). Different choices of function $b(\theta)$ in (5) give rise to different GLMs including the linear regression, logistic regression, and Poisson regression for continuous, binary, and count data of responses, respectively.

Since the GLMs are widely used in applications, [24] extended the SIS idea to this more general class of models. Specifically, two measures of feature importance were considered. The first one is the magnitude of the maximum marginal likelihood estimator (MMLE) $\widehat{\beta}_j^M$ which is defined as the maximizer of the quasi-likelihood function $\ell(\beta_j) = \log f_n(\mathbf{y}; \mathbf{x}_j, \beta_j)$ from marginal regression. Then one can construct the reduced model $\widehat{\mathcal{M}}$ as in (2) with $\widehat{\beta}_j^M$ in place of $\widehat{\text{corr}}(\mathbf{x}_j, \mathbf{y})$. The second one is the marginal likelihood ratio test statistic $\widehat{L}_j$ for testing the significance of each covariate $\mathbf{x}_j$ separately. It was shown in [24] that with both marginal utilities $|\widehat{\beta}_j^M|$ and $\widehat{L}_j$, the SIS for the GLM can continue to enjoy the sure screening property (3) when dimensionality p grows nonpolynomially with sample size n. In addition, a specific bound was established on the size of the reduced model. The random decoupling method in (4) can be employed here to choose the threshold values.

For the binary response, there is a huge literature on classification beyond logistic regression [16, 34, 8]. The idea of independence learning used in SIS has also been exploited widely for feature screening and selection in high-dimensional classification. For the classical two-class Gaussian classification problem with common covariance matrix $\boldsymbol{\Sigma}$, the optimal Fisher's linear discriminant function depends on the inverse of the unknown covariance matrix $\boldsymbol{\Sigma}$. It is well known that estimating high-dimensional covariance matrix is challenging. One choice is to replace the covariance matrix $\boldsymbol{\Sigma}$ by its diagonal matrix $\text{diag}\{\boldsymbol{\Sigma}\}$, leading to the independence rule or naive Bayes method which pretends that the features were independent [3]. [15] formally characterized the phenomenon of noise accumulation in high-dimensional classification which reveals that the independence rule using all the features can perform as bad as random guess when there are a large number of noise features having no discriminative power; see also [32] for the scenario of asymptotically perfect classification. To reduce the noise accumulation, [15] further introduced the features annealed independence rule (FAIR)

based on feature selection using the two-sample t test [50], which was shown to possess an oracle property with explicit classification error rate. The main ideas of FAIR share the same spirit as SIS in that marginal utilities are exploited to rank the importance of features and the two-scale learning framework is formally introduced and justified for ultra-high dimensional regression and classification.

5 Nonparametric and robust sure independence screening

When there exist nonlinear relationships between the covariates and the response, one can use measures of nonlinear correlations in place of the Pearson correlation for linear association. One of such measures is the generalized correlation $\sup_{h\in\mathcal{H}} \mathrm{corr}(h(Z_1), Z_2)$ introduced in [31], where (Z_1, Z_2) is a pair of random variables and $\mathcal{H}$ standards for the vector space generated by a given set of functions such as the polynomials.

Nonparametric models provide flexible alternatives to parametric ones. In particular, the additive model has been widely used for high-dimensional data analysis to alleviate the well-known curse of dimensionality associated with fully nonparametric models. This model assumes that

$$\mathbf{y} = \sum_{j=1}^{p} \mathbf{m}_j(\mathbf{x}_j) + \boldsymbol{\varepsilon}, \tag{6}$$

where $\mathbf{m}_j(\boldsymbol{\theta}) = (m_j(\theta_1), \cdots, m_j(\theta_n))^T$ for $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_n)^T$, $m_j(\cdot)$'s are some unknown smooth functions, and the rest of notation is the same as in model (1). [17] extended the SIS method to high-dimensional additive model (6) and introduced the nonparametric independence screening (NIS). For each covariate $\mathbf{x}_j$, marginal nonparametric regression is employed to provide an estimated function $\widehat{f}_j(\cdot)$ using a B-spline basis. Then the empirical norms $\|\widehat{f}_j\|_n$'s are adopted as the marginal utilities to rank the importance of features, where $\|\widehat{f}_j\|_n^2 = n^{-1}\sum_{i=1}^n \widehat{f}_j(x_{ij})^2$ and $\mathbf{x}_j = (x_{1j}, \cdots, x_{nj})^T$. The reduced model $\widehat{\mathcal{M}}$ from feature screening can be constructed similarly to (2) with the nonparametric marginal utilities. It was established in [17] that NIS can enjoy the sure screening property even in ultra-high dimensions with limited false selection rate. The SIS has also been generalized to other nonparametric and semiparametric settings [10, 9, 11].

Model misspecification can easily happen in real applications when we specify the wrong family of distributions or miss some important covariates [51, 13, 46]. Thus it is important to design statistical learning and inference procedures that are robust to a certain level of model misspecification. In particular, the Pearson correlation is known to be sensitive to the presence of outliers and not robust for heavy-tailed data. To address the robustness issue, [41] extended the SIS method by replacing the Pearson correlation with the Kendall

τ correlation coefficient, which is a robust measure of correlation in a nonparametric sense [38, 39]. To capture the nonlinear associations between the covariates and response, [42] exploited the distance correlation in [48] to rank the marginal importance of features. There is a growing literature on robust feature screening in ultra-high dimensions [54, 47, 12].

6 Multivariate sure independence screening and the beyond

The computational expediency of the SIS comes from the use of marginal screening. To address the potential drawbacks of independence learning, it would be helpful to exploit the joint information among the covariates in the two-scale learning framework. However, naively considering k-dimensional submodels of $\{1, \cdots, p\}$ involves the screening in a space of size $\binom{p}{k} = O(p^k)$ whose computational complexity grows rapidly even for a small k. A computationally tractable multivariate screening method, called the covariate assisted screening and estimation (CASE), was introduced in [37] under the Gaussian linear model (1). The key assumption is that the Gram matrix $\mathbf{G} = \mathbf{X}^T\mathbf{X}$ is nonsparse but sparsifiable in the sense that there exists some $p \times p$ linear filtering matrix $\mathbf{D}$ such that the matrix $\mathbf{DG}$ is sparse. Then the Gaussian linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ can be linearly transformed into $\mathbf{d} = \mathbf{DX}^T\mathbf{y} = \mathbf{DG}\boldsymbol{\beta} + \mathbf{DX}^T\boldsymbol{\varepsilon}$, and a graph-assisted m-variate χ^2-screening can be applied to the p-dimensional vector $\mathbf{d}$. [27] also suggested a way to exploit the joint information among the covariates while using marginal screening ideas, where the features are linearly transformed by the innovated transformation. These new features can be used for ranking the importance of original features. Certainly the area of multivariate sure independence screening awaits further developments.

The ideas of feature screening with SIS have also been applied and adapted to a wide range of large-scale statistical learning problems such as ultra-large Gaussian graphical models [27] and large interaction network screening and detection [33, 36, 26, 25, 40]. There are many other extensions of the general framework of sure independence screening for scalable statistical learning and inference. See, for example, [43] for additional references on feature screening for ultra-high dimensional data.

References

[1] Barut, E., J. Fan, and A. Verhasselt (2016). Conditional sure independence screening. *Journal of the American Statistical Association 111*, 1266–1277.

[2] Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. B 57*, 289–300.

[3] Bickel, P. J. and E. Levina (2004). Some theory for fisher's linear discriminant function, "naive bayes", and some alternatives when there are many more variables than observations. *Bernoulli 10*, 989–1010.

[4] Bickel, P. J., Y. Ritov, and A. Tsybakov (2009). Simultaneous analysis of lasso and dantzig selector. *Ann. Statist. 37*, 1705–1732.

[5] Bühlmann, P. and S. van de Geer (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer.

[6] Candes, E. and T. Tao (2007). The dantzig selector: Statistical estimation when p is much larger than n (with discussion). *Ann. Statist. 35*, 2313–2404.

[7] Candès, E. J., Y. Fan, L. Janson, and J. Lv (2016). Panning for gold: Model-free knockoffs for high-dimensional controlled variable selection. *Manuscript*.

[8] Cannings, T. I. and R. J. Samworth (2017). Random-projection ensemble classification (with discussion). *J. Roy. Statist. Soc. B 79*, 959–1035.

[9] Chang, J., C. Y. Tang, and Y. Wu (2016). Local independence feature screening for nonparametric and semiparametric models by marginal empirical likelihood. *Ann. Statist. 44*, 515–539.

[10] Cheng, M.-Y., T. Honda, J. Li, and H. Peng (2014). Nonparametric independence screening and structure identification for ultra-high dimensional longitudinal data. *Ann. Statist. 42*, 1819–1849.

[11] Chu, W., R. Li, and M. Reimherr (2016). Feature screening for time-varying coefficient models with ultrahigh dimensional longitudinal data. *Annals of Applied Statistics 10*, 596–617.

[12] Cui, H., R. Li, and W. Zhong (2015). Model-free feature screening for ultrahigh dimensional discriminant analysis. *Journal of American Statistical Association 110*, 630–641.

[13] Cule, M., R. Samworth, and M. Stewart (2010). Maximum likelihood estimation of a multi-dimensional log-concave density (with discussion). *J. Roy. Statist. Soc. B 72*, 545–600.

[14] Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani (2004). Least angle regression (with discussion). *Ann. Statist. 32*, 407–499.

[15] Fan, J. and Y. Fan (2008). High-dimensional classification using features annealed independence rules. *The Annals of Statistics 36*, 2605–2637.

[16] Fan, J., Y. Fan, and Y. Wu (2010). High dimensional classification. *High-dimensional Statistical Inference (T. T. Cai and X. Shen, eds., World Scientific, New Jersey)*, 3–37.

[17] Fan, J., Y. Feng, and R. Song (2011). Nonparametric independence screening in sparse ultra-high-dimensional additive models. *J. Amer. Statist. Assoc. 106*, 544–557.

[18] Fan, J., F. Han, and H. Liu (2014). Challenges of big data analysis. *National Science Review 1*, 293–314.

[19] Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc. 96*, 1348–1360.

[20] Fan, J. and J. Lv (2008). Sure independence screening for ultrahigh dimensional feature space (with discussion). *Journal of the Royal Statistical Society Series B 70*, 849–911.

[21] Fan, J. and J. Lv (2010). A selective overview of variable selection in high dimensional feature space (invited review article). *Statistica Sinica 20*, 101–148.

[22] Fan, J., J. Lv, and L. Qi (2011). Sparse high-dimensional models in economics (invited review article). *Annual Review of Economics 3*, 291–317.

[23] Fan, J., R. Samworth, and Y. Wu (2009). Ultrahigh dimensional variable selection: beyond the linear model. *J. Mach. Learn. Res. 10*, 1829–1853.

[24] Fan, J. and R. Song (2010). Sure independence screening in generalized linear models with np-dimensionality. *Ann. Statist. 38*, 3567–3604.

[25] Fan, Y., Y. Kong, D. Li, and J. Lv (2016). Interaction pursuit with feature screening and selection. *Manuscript*.

[26] Fan, Y., Y. Kong, D. Li, and Z. Zheng (2015). Innovated interaction screening for high-dimensional nonlinear classification. *Ann. Statist. 43*, 1243–1272.

[27] Fan, Y. and J. Lv (2016). Innovated scalable efficient estimation in ultra-large gaussian graphical models. *Ann. Statist. 44*, 2098–2126.

[28] Fan, Y. and C. Tang (2013). Tuning parameter selection in high dimensional penalized likelihood. *Journal of the Royal Statistical Society Series B 75*, 531–552.

[29] Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin (2013). *Bayesian Data Analysis (3rd edition)*. Chapman & Hall/CRC.

[30] Hall, P., J. S. Marron, and A. Neeman (2005). Geometric representation of high dimension, low sample size data. *J. Roy. Statist. Soc. Ser. B 67*, 427–444.

[31] Hall, P. and H. Miller (2009). Using generalized correlation to effect variable selection in very high dimensional problems. *Journal of Computational and Graphical Statistics 18*, 533–550.

[32] Hall, P., Y. Pittelkow, and M. Ghosh (2008). Theoretic measures of relative performance of classifiers for high dimensional data with small sample sizes. *J. Roy. Statist. Soc. B 70*, 158–173.

[33] Hall, P. and J.-H. Xue (2014). On selecting interacting features from high-dimensional data. *Computational Statistics & Data Analysis 71*, 694–708.

[34] Hastie, T., R. Tibshirani, and J. Friedman (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd edition)*. Springer.

[35] James, G., D. Witten, T. Hastie, and R. Tibshirani (2013). *An Introduction to Statistical Learning: with Applications in R*. Springer.

[36] Jiang, B. and J. S. Liu (2014). Variable selection for general index models via sliced inverse regression. *Ann. Statist. 42*, 1751–1786.

[37] Ke, Z. T., J. Jin, and J. Fan (2014). Covariate assisted screening and estimation. *Ann. Statist. 42*, 2202–2242.

[38] Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika 30*, 81–93.

[39] Kendall, M. G. (1962). *Rank Correlation Methods (3rd ed.)*. Griffin & Co, London.

[40] Kong, Y., D. Li, Y. Fan, and J. Lv (2017). Interaction pursuit in high-dimensional multi-response regression via distance correlation. *Ann. Statist. 45*, 897–922.

[41] Li, G., H. Peng, J. Zhang, and L. Zhu (2012). Robust rank correlation based screening. *Ann. Statist. 40*, 1846–1877.

[42] Li, R., W. Zhong, and L. Zhu (2012). Feature screening via distance correlation learning. *J. Amer. Statist. Assoc. 107*, 1129–1139.

[43] Liu, J., W. Zhong, and R. Li (2015). A selective overview of feature screening for ultrahigh dimensional data. *Science China: Mathematics 58*, 2033–2054.

[44] Liu, J. S. (2001). *Monte Carlo Strategies in Scientific Computing*. Springer-Verlag, New York.

[45] Lv, J. (2013). Impacts of high dimensionality in finite samples. *The Annals of Statistics 41*, 2236–2262.

[46] Lv, J. and J. S. Liu (2014). Model selection principles in misspecified models. *Journal of the Royal Statistical Society Series B 76*, 141–167.

[47] Mai, Q. and H. Zou (2013). The kolmogorov filter for variable screening in high-dimensional binary classification. *Biometrika 100*, 229–234.

[48] Székely, G. J., M. L. Rizzo, and N. K. Bakirov (2007). Measuring and testing dependence by correlation of distances. *Ann. Statist. 35*, 2769–2794.

[49] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B 58*, 267–288.

[50] Tibshirani, R., T. Hastie, B. Narasimhan, and G. Chu (2003). Class prediction by nearest shrunken centroids, with applications to dna microarrays. *Statist. Sci. 18*, 104–117.

[51] White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica 50*, 1–25.

[52] Zhao, P. and B. Yu (2006). On model selection consistency of lasso. *Journal of Machine Learning Research 7*, 2541–2563.

[53] Zhao, S. D. and Y. Li (2012). Principled sure independence screening for cox models with ultra-high-dimensional covariates. *Journal of multivariate analysis 105*, 397–411.

[54] Zhu, L., L. Li, R. Li, and L. Zhu (2011). Model-free feature screening for ultrahigh-dimensional data. *J. Amer. Statist. Assoc. 106*, 1464–1475.