Wiley StatsRef:
Statistics Reference Online

Sure Independence Screening

By *Jianqing Fan*[1] *and Jinchi Lv*[2]

Keywords: *big data, dimensionality reduction, iterative sure independence screening, nonparametric, robustness, scalability, sparsity, sure independence screening, sure screening property, two-scale statistical learning, ultrahigh dimensionality*

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, ultrahigh-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 of 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.

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[1]. Among many types of big data, ultrahigh-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 10 years or so. Scalability is a major challenge of ultrahigh-dimensional data analysis. Meanwhile, it is well known that additional intrinsic challenges of ultrahigh-dimensional data analysis include high collinearity, spurious correlation, and noise accumulation[2−5]. For example, in the presence of a large number of noise features, high-dimensional classification using all the features can behave like random guess[3]. To improve the scalability and reduce noise accumulation, one possible approach is reducing the dimensionality of the feature space from a very large

[1] Princeton University, Princeton, NJ, USA
[2] University of Southern California, Los Angeles, CA, USA

DOI: 10.1002/9781118445112.stat08043

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)

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 owing 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[3].

To appreciate the point, let us consider a simple simulation example that provides a prototype for the common goals desired by practitioners analyzing high-dimensional data. We generate 100 data sets from the Gaussian linear model given in Equation (1) with sample size $n = 120$, dimensionality $p = 1000$, design matrix $\mathbf{X} \sim N(\mathbf{0}, I_n \bigotimes \boldsymbol{\Sigma})$ for $\boldsymbol{\Sigma} = (0.5^{|j-k|})_{1\leq j,k\leq p}$, and error vector $\varepsilon \sim N(\mathbf{0}, \sigma^2 I_n)$ 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-X knockoffs method introduced in Ref. 6 coupled with the innovated scalable efficient estimation (ISEE) estimator in Ref. 7 to control the **False Discovery Rate** (FDR)[8] for feature selection, where the target FDR level is set as $q = 0.2$. For sparse model fitting, we employ **Lasso**[9], **SCAD**[10], and sure independence screening (SIS)[2] followed by the Lasso and smoothly clipped absolute deviation (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.

2 Sure Independence Screening

We now begin the journey of feature screening in ultrahigh-dimensional feature space. A common practice for feature screening is using independence learning that 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 ultrahigh-dimensional data analysis, the SIS was formally introduced and rigorously justified in Ref. 2 to address both scalability and noise accumulation issues. Let us consider the linear regression model

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

where $\mathbf{y} = (y_1, \ldots, y_n)^T$ is an n-dimensional response vector, $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)$ is an $n \times p$ design matrix consisting of p covariates $\mathbf{x}_j$s, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is a p-dimensional regression coefficient vector, and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T$ is an n-dimensional error vector. The focus of Ref. 2 is the ultrahigh-dimensional setting

DOI: 10.1002/9781118445112.stat08043

with $\log p = O(n^{\alpha})$ for some $0 < \alpha < 1$. To ensure model identifiability, the true regression coefficient vector $\boldsymbol{\beta}_0 = (\beta_{0,1}, \ldots, \beta_{0,p})^T$ is assumed to be sparse. The covariates $\mathbf{x}_j$s with indices in the support $\mathcal{M}_* = \mathrm{supp}(\boldsymbol{\beta}_0) = \{1 \le j \le p : \beta_{0,j} \ne 0\}$ 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 $\widehat{\mathrm{corr}}(\mathbf{x}_j, \mathbf{y})$ 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,

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

where $\widehat{\mathrm{corr}}$ 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 performed with any favorite regularization method of user's choice including Lasso, SCAD, and Dantzig selector[4,5,9−17]. 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[18,19].

The feature screening (Equation (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,

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

as $n \to \infty$. The property in Equation (3) was termed as the sure screening property in Ref. 2 which is crucial to the second step of refined variable selection. This is a weaker notation than the model selection consistency and is a more realistic task in high-dimensional inference, particularly for a screening method. Surprisingly, SIS was shown in Ref. 2 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 $\boldsymbol{\Sigma}$ and the p-dimensional random covariate vector multiplied by $\boldsymbol{\Sigma}^{-1/2}$ is assumed to have a spherical distribution. The sure screening property of SIS depends on the so-called concentration property for random design matrix $\mathbf{X}$ introduced in Ref. 2; see Ref. 20 for a similar concentration phenomenon of the large random design matrix.

The concentration property was originally verified for the scenario of Gaussian distributions, and later established in Ref. 21 for a wide class of elliptical distributions as conjectured previously. With such a property, the sure screening property (Equation (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 article[2]. In practice, the parameter d can be chosen by some data-driven methods such as the cross-validation (CV) and generalized information criterion (GIC)[22]. It can also be selected by a simple permutation method[23,24] 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 $\{1, \ldots, n\}$ 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 the false positive rate when applied to the randomly decoupled data. When $q = 1$, r^* is merely the largest spurious correlation. Now, select the model

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

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.

DOI: 10.1002/9781118445112.stat08043

3 Iterative and Conditional Sure Independence Screening

As 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 (false positive). Second, some important covariates that are jointly correlated but marginally uncorrelated with the response can be missed after the screening step (false negative). To address these issues, Ref. 2 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 Ref. 25 to the pseudo-likelihood framework beyond the linear model with more general loss functions. Reference 25 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 Ref. 26. 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 $\mathbf{y}$ the 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 $\phi \in (0, \infty)$, $(\theta_1, \ldots, \theta_n)^T = \mathbf{X}\boldsymbol{\beta}$, $b(\cdot)$ 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 Equation (5) give rise to different GLMs including the linear regression, logistic regression, and Poisson regression for continuous, binary, and count data of responses, respectively.

As the GLMs are widely used in applications, Ref. 27 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, assuming that the covariate x_j has been standardized. Then, one can construct the reduced model $\widehat{\mathcal{M}}$ as in Equation (2) with $\widehat{\beta}_j^M$ in place of $\widehat{\mathrm{corr}}(\mathbf{x}_j, \mathbf{y})$. 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 Ref. 27 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 (Equation (3)) when dimensionality p grows as high as nonpolynomially with sample size n. In addition, a specific bound was established on the size of

DOI: 10.1002/9781118445112.stat08043

the reduced model. The random decoupling method in Equation (4) can be employed here to choose the threshold values.

For the binary response, there is a huge literature on classification beyond logistic regression[15,28,29]. 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 the high-dimensional covariance matrix is challenging. One choice is to replace the covariance matrix $\boldsymbol{\Sigma}$ by its diagonal matrix diag$\{\boldsymbol{\Sigma}\}$, leading to the independence rule or naive Bayes method which pretends that the features were independent[30]. Fan and Fan[3] 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 Ref. 31 for the scenario of asymptotically perfect classification. To reduce the noise accumulation, Fan and Fan[3] further introduced the feature annealed independence rule (FAIR) based on feature selection using the two-sample t test[32], which was shown to possess an oracle property with an 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 ultrahigh-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 Ref. 33, where (Z_1, Z_2) is a pair of random variables and $\mathcal{H}$ stands for the vector space generated by a given set of functions such as the polynomials or spline bases.

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

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

where $\mathbf{m}_j(\boldsymbol{\theta}) = (m_j(\theta_1), \ldots, m_j(\theta_n))^T$ for $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)^T$, $m_j(\cdot)$s are some unknown smooth functions, and the rest of the notation is the same as in model (1). Fan et al.[23] extended the SIS method to a 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 $\hat{f}_j(\cdot)$ using a B-spline basis. Then, the empirical norms $\|\hat{f}_j\|_n$s are adopted as the marginal utilities to rank the importance of features, where $\|\hat{f}_j\|_n^2 = n^{-1}\sum_{i=1}^{n}\hat{f}_j(x_{ij})^2$ and $\mathbf{x}_j = (x_{1j}, \ldots, x_{nj})^T$. The reduced model $\widehat{\mathcal{M}}$ from feature screening can be constructed similarly to Equation (2) with the nonparametric marginal utilities. It was established in Ref. 23 that NIS can enjoy the sure screening property even in ultrahigh dimensions with a limited false selection rate. The SIS has also been generalized to other nonparametric and semiparametric settings[34−36].

Model misspecification can easily happen in real applications when we specify the wrong family of distributions or miss some important covariates[37−39]. 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

DOI: 10.1002/9781118445112.stat08043

correlation is known to be sensitive to the presence of outliers and not robust for heavy-tailed data. To address the robustness issue, Li et al.[40] extended the SIS method by replacing the Pearson correlation with the Kendall τ correlation coefficient, which is a robust measure of the correlation in a nonparametric sense[41,42]. To capture the nonlinear associations between the covariates and response, Li et al.[43] exploited the distance correlation in Ref. 44 to rank the marginal importance of features. There is growing literature on robust feature screening in ultrahigh dimensions[45−47].

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 $\{1, \ldots, p\}$ 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 Ref. 48 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 $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon$ can be linearly transformed into $\mathbf{d} = \mathbf{DX}^T\mathbf{y} = \mathbf{DG}\boldsymbol{\beta} + \mathbf{DX}^T\varepsilon$, and a graph-assisted m-variate χ^2-screening can be applied to the p-dimensional vector $\mathbf{d}$. Fan and Lv[7] 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 SIS 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 ultralarge Gaussian graphical models[7] and large interaction network screening and detection[49−53]. There are many other extensions of the general framework of SIS for scalable statistical learning and inference. See, for example, Ref. 54 for additional references on feature screening for ultrahigh-dimensional data.

Acknowledgments

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.

Related Articles

Big Data in Biosciences; **Dimension Reduction in Clustering**; **Lasso, the**; **Nonparametric Regression Model**; **Variable Selection via Regularization**.

References

[1] Fan, J., Han, F., and Liu, H. (2014) Challenges of big data analysis. *Natl. Sci. Rev.* **1**, 293–314.

[2] Fan, J. and Lv, J. (2008) Sure independence screening for ultrahigh dimensional feature space (with discussion). *J. R. Stat. Soc., Ser. B* **70**, 849–911.

DOI: 10.1002/9781118445112.stat08043

[3] Fan, J. and Fan, Y. (2008) High-dimensional classification using features annealed independence rules. *Ann. Stat.* **36**, 2605–2637.

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

[5] Fan, J., Lv, J., and Qi, L. (2011) Sparse high-dimensional models in economics (invited review article). *Annu. Rev. Econ.* **3**, 291–317.

[6] Candès, E.J., Fan, Y., Janson, L., and Lv, J. (2016) Panning for gold: model-X knockoffs for high-dimensional controlled variable selection. *J. Roy. Statist. Soc. Ser. B*, to appear.

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

[8] Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc., Ser. B* **57**, 289–300.

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

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

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

[12] Zhao, P. and Yu, B. (2006) On model selection consistency of lasso. *J. Mach. Learn. Res.* **7**, 2541–2563.

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

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

[15] Hastie, T., Tibshirani, R., and Friedman, J. (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd edn, Springer-Verlag, New York, Berlin Heidelberg.

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

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

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

[19] Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., and Rubin, D.B. (2013) *Bayesian Data Analysis*, 3rd edn, Chapman & Hall/CRC, London, Boca Raton, FL.

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

[21] Lv, J. (2013) Impacts of high dimensionality in finite samples. *Ann. Stat.* **41**, 2236–2262.

[22] Fan, Y. and Tang, C. (2013) Tuning parameter selection in high dimensional penalized likelihood. *J. R. Stat. Soc., Ser. B* **75**, 531–552.

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

[24] Zhao, S.D. and Li, Y. (2012) Principled sure independence screening for COX models with ultra-high-dimensional covariates. *J. Multivariate Anal.* **105**, 397–411.

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

[26] Barut, E., Fan, J., and Verhasselt, A. (2016) Conditional sure independence screening. *J. Am. Stat. Assoc.* **111**, 1266–1277.

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

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

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

[30] Bickel, P.J. and 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* **10**, 989–1010.

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

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

[33] Hall, P. and Miller, H. (2009) Using generalized correlation to effect variable selection in very high dimensional problems. *J. Comput. Graph. Stat.* **18**, 533–550.

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

DOI: 10.1002/9781118445112.stat08043

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

[36] Chu, W., Li, R., and Reimherr, M. (2016) Feature screening for time-varying coefficient models with ultrahigh dimensional longitudinal data. *Ann. Appl. Stat.* **10**, 596–617.

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

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

[39] Lv, J. and Liu, J.S. (2014) Model selection principles in misspecified models. *J. R. Stat. Soc., Ser. B* **76**, 141–167.

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

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

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

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

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

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

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

[47] Cui, H., Li, R., and Zhong, W. (2015) Model-free feature screening for ultrahigh dimensional discriminant analysis. *J. Am. Stat. Assoc.* **110**, 630–641.

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

[49] Hall, P. and Xue, J.-H. (2014) On selecting interacting features from high-dimensional data. *Comput. Stat. Data Anal.* **71**, 694–708.

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

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

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

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

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

DOI: 10.1002/9781118445112.stat08043