Published in final edited form as:
Ann Stat. 2013 August 1; 41(4): . doi:10.1214/13-AOS1139.

MARGINAL EMPIRICAL LIKELIHOOD AND SURE INDEPENDENCE FEATURE SCREENING

Jinyuan Chang[1], **Cheng Yong Tang**, and **Yichao Wu**[2]
Peking University, University of Colorado Denver and North Carolina State University

Abstract

We study a marginal empirical likelihood approach in scenarios when the number of variables grows exponentially with the sample size. The marginal empirical likelihood ratios as functions of the parameters of interest are systematically examined, and we find that the marginal empirical likelihood ratio evaluated at zero can be used to differentiate whether an explanatory variable is contributing to a response variable or not. Based on this finding, we propose a unified feature screening procedure for linear models and the generalized linear models. Different from most existing feature screening approaches that rely on the magnitudes of some marginal estimators to identify true signals, the proposed screening approach is capable of further incorporating the level of uncertainties of such estimators. Such a merit inherits the self-studentization property of the empirical likelihood approach, and extends the insights of existing feature screening methods. Moreover, we show that our screening approach is less restrictive to distributional assumptions, and can be conveniently adapted to be applied in a broad range of scenarios such as models specified using general moment conditions. Our theoretical results and extensive numerical examples by simulations and data analysis demonstrate the merits of the marginal empirical likelihood approach.

Key words and phrases

Empirical likelihood; high-dimensional data analysis; sure independence screening; large deviation

1. Introduction

High-dimensional data are more frequently encountered in current practical problems of finance, biomedical sciences, geological studies and many more areas. Statistical methods for high-dimensional data analysis have received increasing interests to deal with large volume of data containing considerably many features; see Bühlmann and van de Geer (2011), Hastie, Tibshirani and Friedman (2009) and Fan and Lv (2010) for overviews. A

J. Chang, Guanghua School of Management, Peking University, Beijing, 100871, China, changjinyuan1986@pku.edu.cn
C. Y. Tang, Business School, University of Colorado Denver, Campus Box 165, PO Box 173364, Denver, Colorado 80217-3364, USA, CHENGYONG.TANG@ucdenver.edu
Y. Wu, Department of Statistics, North Carolina State University, 2311 Stinson Drive, Raleigh, North Carolina 27695-8203, USA, wu@stat.ncsu.edu

[1] Supported in part by a National Natural Science Foundation of China Key Program Grant (No. 11131002) and Center for Statistical Science at Peking University. This work was completed during Jinyuan Chang's Ph.D. study at Peking University.

[2] Supported in part by NSF Grants DMS-09-05561 and DMS-10-55210 and NIH/NCI Grant R01-CA149569.

SUPPLEMENTARY MATERIAL

Supplement to "Marginal empirical likelihood and sure independence feature screening" (DOI: 10.1214/13-AOS1139SUPP; .pdf). This supplement contains all technical proofs.

fundamental objective of statistical analysis with high-dimensional data is to identify relevant features, so that effective models can be subsequently constructed and applied to solve practical problems.

Recently, independence feature screening methods have been considered, see, for example, Fan and Lv (2008), Fan and Song (2010) and Fan, Feng and Song (2011) for linear models, generalized linear models and nonparametric additive models, respectively. Fan and Lv (2008) and Fan and Song (2010) performed screening by ranking the absolute values of marginal estimates of model coefficients, and Fan, Feng and Song (2011) carried out screening by ranking integrated squared marginal nonparametric curve estimates. Fan and Song (2010) also discussed independence screening by examining the magnitudes of the likelihood ratios. More recently, Wang (2012) considered a sure independence screening by a factor profiling approach; Xue and Zou (2011) studied sure independence screening and sparse signal recovery; see also Zhu et al. (2011) and Li, Zhong and Zhu (2012) for recent development using model-free approaches for feature screening, Li et al. (2012) for a robust rank correcation based approach, and Zhao and Li (2012) for an estimating equation based feature screening approach.

The empirical likelihood approach [Owen (1988, 2001)] is demonstrated effective in scenarios with less restrictive distributional assumptions for statistical inferences; see Qin and Lawless (1994), Newey and Smith (2004) and reference therein. We refer to Chen and Van Keilegom (2009) as a review and discussion of recent development in the empirical likelihood approach. The scope of the empirical likelihood approach recently has also been extended to deal with high-dimensional data; see Hjort, McKeague and Van Keilegom (2009), Chen, Peng and Qin (2009), Tang and Leng (2010), Leng and Tang (2012), and Chang, Chen and Chen (2013). Though demonstrated effective in statistical inferences, the empirical likelihood approach encounters substantial difficulty when data dimensionality is high. More specifically, the data dimensionality p cannot exceed the sample size n in the conventional empirical likelihood construction. In addition, p can be at most $o(n^{1/2})$ or even slower under which asymptotic properties are established [Chang, Chen and Chen (2013), Chen, Peng and Qin (2009), Hjort, McKeague and Van Keilegom (2009), Leng and Tang (2012), Tang and Leng (2010)]. Therefore, to practically more effectively apply the empirical likelihood approach, a pre-screening procedure is necessary to reduce the candidates of target features.

In this study, we systematically examine the properties of a marginal empirical likelihood approach where the available features are assessed one at a time individually. The marginal empirical likelihood approach only involves univariate optimizations, so that it provides a convenient device for both theoretical analysis and practical implementation. Our analysis reveals the probabilistic behavior of the marginal empirical likelihood ratios as functions of the parameters of interest that can be evaluated at arbitrary values, which itself is a problem of individual interest because existing studies of the empirical likelihood approach generally focus on its properties when evaluated at the truth, or at values in a small neighborhood of the truth. Based on our finding, we propose to conduct feature screening by using the marginal empirical likelihood ratio evaluated at zero. We find that a unified screening procedure can be applied in both linear models and generalized linear models. We also demonstrate how the marginal empirical likelihood approach can be conveniently adapted to solve a broad range of problems for models specified by general moment conditions. Hence, the marginal empirical likelihood approach provides a general and adaptive procedure for solving a broad class of practical problems for feature screening. Our theoretical analyses show that the proposed screening procedure based on the marginal empirical likelihood approach is selection consistent—that is, being able to identify the features that contribute to the response variable when the number of explanatory variables p grows exponentially with sample size n.

Our study contributes to the sure independence feature screening for high-dimensional data analysis from the following two substantial aspects. First of all, a fundamental difference of our approach to all existing approaches is that the marginal empirical likelihood ratio statistic is a self-studentized quantity [Owen (2001)] while other existing screening methods generally rely on the ranking of features based on magnitudes of some marginal estimators. Therefore, our approach is able to incorporate additionally the level of uncertainties associated with the estimators to conduct feature screening. This clearly extends the scope of existing feature screening approaches by considering more aspects of marginal statistical approaches. We show in our simulation studies that when heterogeneity exists in the conditional variance, our approach performs much better than a least-squares based approach. Second, our screening procedure inherits the non-parametric merits of the empirical likelihood approach. Specifically, our approach requires no strict distributional assumptions such as normally distributed errors in the linear models, or exponential family distributed response in the generalized linear models. This generalizes the scope and applicability of our approach. As a result, we show that the marginal empirical likelihood approach provides a unified framework for feature screening in linear regression models and generalized linear models, and can be conveniently applied for solving a broad class of general problems.

The rest of this paper is organized as follows. We elaborate the method of the marginal empirical likelihood approach in Section 2. Properties of the proposed approach are given in Section 3. Section 4 extends the marginal empirical likelihood approach to a broad framework including models specified by general moment conditions, and presents an iterative sure screening procedure using profile empirical likelihood. Numerical examples are given in Section 5. We conclude with some discussions in Section 6. All technical details are contained in the supplementary material of this paper [Chang, Tang and Wu (2013)].

2. Methodology

2.1. Marginal empirical likelihood for linear models

Let us motivate the marginal empirical likelihood approach by first considering the multiple linear regression model

$$Y = \mathbf{X}^{\mathrm{T}}\beta + \varepsilon, \quad (2.1)$$

where $\mathbf{X} = (X_1, \ldots, X_p)^{\mathrm{T}}$ is the vector of explanatory variables, ε is the random error with zero mean, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ is the vector of unknown parameters. Hereinafter, we also use $\boldsymbol{\beta}$ to denote the truth of the parameter whenever no confusion arises. Without loss of generality, we assume hereinafter that the explanatory variables are standardized such that $\mathbb{E}(X_j) = 0$ and [redacted] ($j = 1, \ldots, p$). For effective and interpretable practical applications, one may reasonably expect that among the large number of explanatory variables, only a small fraction of them contribute to the response variable. We therefore denote by $\mathcal{M}_* = \{1 \leq j \leq p : \beta_j \neq 0\}$ the collection of the effective explanatory variables in the true sparse model whose size is characterized by its cardinality $s = |\mathcal{M}_*|$. Here we assume that s is much smaller than p, reflecting the case in many practical applications like in finance, biology and clinical studies.

In the recent literature of high-dimensional data analysis, various marginal approaches have been applied for locating the true model $\mathcal{M}_*$; see, for example, Fan and Lv (2008), Fan and

Song (2010) and Fan, Feng and Song (2011). Among those approaches, a popular way is to assess the marginal contribution from a given explanatory variable X_j. Commonly applied criteria for measuring the marginal contribution are the magnitudes of some marginal estimators [Fan, Feng and Song (2011), Fan and Lv (2008), Fan and Song (2010)]. Subsequently, the candidate models are chosen from the top ranked explanatory variables.

To apply a marginal empirical likelihood approach for the linear regression model (2.1), let us consider the marginal moment condition of the least squares estimator:

$$\mathbb{E}\{X_j(Y-X_j\beta_j^M)\}=0 \quad (j=1,\ldots,p), \quad (2.2)$$

where β_j^M is interpreted as the marginal contribution of covariate X_j to Y. From (2.2), we can see that $\beta_j^M=\mathbb{E}(X_jY)$ is the covariance between X_j and Y so that $\beta_j^M=0$ is equivalent to that Y and X_j are marginally uncorrelated. Here we note the remarkable difference between β_j^M and β_j where the latter is the truth of the parameter in (2.1). In general, $\beta_j^M \neq \beta_j$ unless $\mathbb{E}$ $(X_iX_j) = 0$ for all $i \neq j$. In addition to that from β_j in the model (2.1), β_j^M also contains aggregated contribution from other components that may be correlated with X_j. Thus, the correlation level among covariates has significant impact on the performance of a screening procedure based on (2.2); more discussions on this are given in a later section containing the main results.

A marginal empirical likelihood for linear models can be constructed as follows. Note that $\mathbb{E}(X_j^2)=1$, therefore (2.2) is equivalent to

$$\mathbb{E}(X_jY-\beta_j^M)=0. \quad (2.3)$$

Let $\{(\mathbf{X}_i,Y_i)\}_{i=1}^n$ be collected independent data, $g_{ij}(\beta) = X_{ij}Y_i-\beta$ $(j = 1,\ldots, p)$ and X_{ij} means the j th component of the ith observation $\mathbf{X}_i$. Based on (2.3), we define the following marginal empirical likelihood:

$$\mathrm{EL}_j(\beta)=\sup\left\{\prod_{i=1}^n w_i : w_i \geq 0, \sum_{i=1}^n w_i=1, \sum_{i=1}^n w_i g_{ij}(\beta)=0\right\} \quad (2.4)$$

for $j = 1,\ldots, p$. For any given β in the convex hull of $\{X_{ij}Y_i\}_{i=1}^n$, the marginal empirical likelihood ratio is defined as

$$\ell_j(\beta)=-2\log\{\mathrm{EL}_j(\beta)\}-2n\log n=2\sum_{i=1}^n \log\{1+\lambda g_{ij}(\beta)\}, \quad (2.5)$$

where λ is the Lagrange multiplier satisfying

$$0=\sum_{i=1}^n \frac{g_{ij}(\beta)}{1+\lambda g_{ij}(\beta)}. \quad (2.6)$$

2.2. Extended coverage to generalized linear models

A merit of the marginal empirical likelihood approach is that the formulation by (2.4) and (2.5) only requires the moment condition (2.3), rather than specific distributional assumption of ε in model (2.1). This entitles our approach robustness against the violation of distributional model assumptions, and thus it can be extended and adapted to a broader framework. Now we elaborate how the above marginal empirical likelihood approach can be equally applied when the response variable Y is in the exponential family with the density function taking the canonical form [McCullagh and Nelder (1989)]:

$$f(y)=\exp\{y\theta-b(\theta)+c(y)\} \quad (2.7)$$

for some suitable known functions $b(\cdot)$, $c(\cdot)$ and canonical parameter θ. Further extensions of the marginal empirical likelihood approach are discussed in a later section. We refer to Kolaczyk (1994) and Chen and Cui (2003) for conventional applications of the empirical likelihood to generalized linear models. Following the convention of generalized linear models, we denote the mean function by $\mu = \mathbb{E}(Y|\mathbf{X}) = b'(\theta)$ where θ is modeled by a linear function $\beta_0 + \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}$ with $\boldsymbol{\beta} = (\beta_1,\ldots, \beta_p)^{\mathrm{T}}$, and use $V(\mu)$ to denote the variance of Y expressed as a function of μ.

For any $j = 1,\ldots, p$, the moment condition based on the marginal likelihood approach in Fan and Song (2010) for β_j is

$$\mathbb{E}\left\{\frac{Y-\mu_j}{V(\mu_j)}\frac{\partial\mu_j}{\partial\beta_j}\right\}=0, \quad (2.8)$$

where $\mu_j=b^{'}(\beta_0+\beta_j^M X_j)$ is the implied mean function that is modeled marginally only using X_j. Here the β_j^M is again interpreted as the marginal contribution of X_j to the response variable Y ; see also Fan and Song (2010). By the property of the exponential family distribution, $\frac{\partial\mu}{\partial\beta_j}=X_j b^{''}(\theta)$ and $V(\mu) = b''(\theta)$. Then (2.8) becomes

$$\mathbb{E}\{X_j(Y-\mu_j)\}=0. \quad (2.9)$$

For linear models, $b(\theta)=\frac{\theta^2}{2}$, then $b'(\theta) = \theta$, so that (2.9) becomes $\mathbb{E}\{X_j(Y-X_j\beta_j^M)\}=0$ by noting that $\beta_0 = 0$ in the linear model case, which is exactly (2.2). Hence, (2.9) is a natural extension of (2.2) in the generalized linear models.

One way to apply the marginal approach can be generalizing the definition in (2.4) to be $g_{ij}(\beta) = X_{ij}\{Y_i - b'(\beta_0 + \beta X_{ij})\}$ $(j = 1,\ldots, p)$. However, such a modification is actually not necessary. To see this, we note that when the marginal contribution $\beta_j^M=0$, then the marginal moment condition (2.9) becomes $\mathbb{E}[X_j\{Y - b'(\beta_0)\}] = 0$. Hence, it implies that the covariance between X_j and Y is 0, which exactly shares the same implication of (2.3) as in the linear models. From this perspective, (2.9) and (2.3) are essentially equivalent. Additionally, the response variable in practice can always be centered to have zero mean. This fact eliminates the concern on the intercept β_0 in the generalized linear models when considering a marginal empirical likelihood approach. As a result, we conclude that a unified marginal empirical likelihood construction (2.4) with the same $g_{ij}(\beta) = X_{ij}Y_i - \beta$ can be equally applied for both linear models and generalized linear models with centered response variable Y. The implication of this unified construction is also intuitively very clear by interpreting β as the covariance between a covariate and the response variable.

Furthermore, we note that the distributional assumption (2.7) is actually not required in our marginal empirical likelihood approach. Therefore our approach is not restricted to the exponential family (2.7). Since we only require the marginal moment condition (2.9), our approach can be applied with the quasi-likelihood approach and it also works with misspecified variance functions [McCullagh and Nelder (1989)].

The marginal empirical likelihood ratio (2.5) with $g_{ij}(\beta) = X_{ij}Y_i - \beta$ evaluated at $\beta = 0$—that is, $\ell_j(0)$—has a very clear practical interpretation by noting that it can be used to test the null hypothesis $H_0{:}\beta_j^M=0$. By noting additionally the intuitively clear fact that $\ell_j(0)$ should not be large if $\beta_j^M=0$, we can see that $\ell_j(0)$ can be used as a device for feature screening. More specifically, we have the following procedure:

Step 1: Evaluating $\ell_j(0)$ for all $j = 1,\ldots, p$, where $\ell_j(\cdot)$ is defined in (2.5) with $g_{ij}(\beta) = X_{ij}Y_i - \beta$. If 0 is not in the convex hull of $\{X_{ij}Y_i\}_{i=1}^n$, we define $\ell_j(0)=\infty$ as a strong evidence of significance in predicting Y using X_j.

Step 2: Given a threshold level γ_n, select a set of variables by

$$\widehat{\mathscr{M}}_{\gamma_n}=\{1 \le j \le p{:}\ell_j(0) \ge \gamma_n\}.$$

We specify in the next section the requirement for γ_n so that the screening procedure is consistent. On the other hand, however, explicitly identifying γ_n in practice is generally difficult because it involves unknown constants. Thus, a screening procedure can be practically implemented in a way such that $\widehat{\mathscr{M}}_{\gamma_n}$ recruits candidate features until certain size such as $n^{1/2}$ is achieved.

We remark that the evaluation of $\ell_j(\beta)$ in (2.5) in practice is actually very easy by noting that all optimizations involved are univariate, which is very convenient for practical applications. On the other hand, our procedure only needs to evaluate the marginal empirical likelihood ratio (2.5) at $\beta = 0$ and avoids the estimation of β_j^M when conducting the feature screening.

3. Main results

Now we present main results for the marginal empirical likelihood ratio in (2.5) with the unified specification $g_{ij}(\beta) = X_{ij}Y_i - \beta$ that are generally applicable for both linear models and generalized linear models. In our discussion hereinafter, let $\rho_j = \mathbb{E}(X_jY)$. If $\rho_j = 0$, it is well known that $\ell_j(0)$ is asymptotically chi-square distributed with 1 degree of freedom [Owen (1988, 2001)]. If $\rho_j \ne 0$, however, the properties of $\ell_j(0)$ is generally less clear, which is also a question of independent interest. Specifically, if $\beta = \rho_j + \tau\sigma n^{-1/2}$ where $\sigma^2 = \mathrm{var}(X_jY)$, it can be shown following the same argument of Owen (1988) that $\ell_j(\beta) \xrightarrow{d} \chi_1^2(\tau^2)$ as $n\to\infty$ under some regularity conditions where τ^2 is a noncentrality parameter. But if $\beta - \rho_j$ converges to zero at a rate slower than $n^{-1/2}$, the exact diverging rate of $\ell_j(\beta)$ is less clear in existing literature.

We first present a general result that shows that the empirical likelihood ratio $\ell_j(\beta)$ is no longer $O_p(1)$ when $\beta - \rho_j$ converges to 0 but $n^{1/2}(\beta - \rho_j)$ diverges.

Proposition 1—Suppose that $U_1,\ldots, U_n$ are independent and identically distributed random variables with $\mathbb{E}(|U_i|^{\nu}) < \infty$ for some $\nu \ge 3$. Replacing $g_{ij}(\beta)$ in (2.5) and (2.6) by $U_i - \mu$ for all $i = 1,\ldots, n$, we obtain $\ell(\mu)$. If $|\mu - \mu_0| = O(n^{-w})$ for some $w \in (\frac{1}{\nu}, \frac{1}{2})$, then

$$\frac{\ell(\mu)}{n(\mu - \mu_0)^2\sigma^{-2}} \xrightarrow{p} 1 \quad as\, n \to \infty,$$

where $\mu_0 = \mathbb{E}(U_i)$ and $\sigma^2 = \mathbb{E}\{(U_i - \mu_0)^2\}$.

We note that Chen, Gao and Tang (2008) contains a related result showing that the empirical likelihood ratio is diverging when evaluated at values far enough from the truth. Our Proposition 1 contains the specific diverging rate of the empirical likelihood ratio. Proposition 1 implies that if $\beta - \rho_j$ converges to zero at a rate slower than $n^{-1/2}$, $\ell_j(\beta) = O_p\{n(\beta - \rho_j)^2\}$. On the other hand, if $\beta - \rho_j$ does not weaken to zero, our Theorem 1 presented later shows that $\ell_j(\beta)$ has high probability to take large value. On the other hand, as clearly shown in our proof of Proposition 1 given in Chang, Tang and Wu (2013), the statistics $\ell_j(0)$ is self-studentized, and hence it incorporates the level of uncertainties from using the finite sample moment conditions. Such a feature is desirable because in practice levels of uncertainties corresponding to different covariates can be different when contributing to the response variable of interest. This may confound the ranking for feature screening based on marginal estimators themselves without considering their standard errors, not mentioning incorporating the level of uncertainties is difficult especially when handling high-dimensional statistical problems.

An effective marginal screening procedure requires two conditions: (i) if $j \in \mathcal{M}_*$, then ρ_j takes nonnegligible value; and (ii) if $j \notin \mathcal{M}_*$, then ρ_j takes negligible value. Actually, the first requirement is closely related to recruiting the true signals that contribute to the response, and the second one affects the size of selected variable set that may contain false signals. Fan and Lv (2008) shows that under the identification condition $\min_{j \in \mathcal{M}_*} |\rho_j| \ge f_n > 0$ for some function f_n, the first requirement is fulfilled. A common assumption for f_n is $f_n = O(n^{-\kappa})$ for some $\kappa \in (0, \frac{1}{2})$.

Our next theoretical analysis imposes the following two assumptions:

A.1: The random variable Y has bounded variance and there exists a positive constant c_1 such that

$$\min_{j \in \mathcal{M}_*} |\mathbb{E}(X_j Y)| = \min_{j \in \mathcal{M}_*} |\mathrm{cov}(Y, X_j)| \ \ge c_1 n^{-\kappa}$$

for some $\kappa \in [0, \frac{1}{2})$.

A.2: There are positive constants K_1, K_2, γ_1 and γ_2 such that

$$\begin{array}{ll} \mathbb{P}\{|X_j| \ \ge u\} \le K_1 \exp(-K_2 u^{\gamma_1}) & \text{for each } j = 1, \ldots, p \text{ and any } u > 0, \\ \mathbb{P}\{|Y| \ \ge u\} \le K_1 \exp(-K_2 u^{\gamma_2}) & \text{for any } u > 0. \end{array}$$

Assumption A.1 can be viewed as a requirement for the minimal signal strength, and we call it the identification condition for $j \in \mathcal{M}_*$. For linear models, the assumption A.1 is same as condition 3 in Fan and Lv (2008) that is commonly assumed in sure independence feature screening. For generalized linear models, Fan and Song (2010) imposes the identification condition as $\min_{j\in\mathcal{M}_*}|\mathrm{cov}(b'(\mathbf{X}^{\mathrm{T}}\beta), X_j)| \geq c_1 n^{-\kappa}$. By noticing that $\mathrm{cov}(b'(\mathbf{X}^{\mathrm{T}}\beta), X_j) = \mathbb{E}(X_j Y)$, their identification condition for $j \in \mathcal{M}_*$ is also same as A.1. Since we impose no distributional assumptions, A.2 is assumed to ensure the large deviation results that are used to get the exponential convergence rate. The first part of A.2 is same as the first part of condition D in Fan and Song (2010). For linear regression model, the second part of condition D in Fan and Song (2010) is equivalent to that $\mathbf{X}^{\mathrm{T}}\beta$ satisfies the Cramér condition such that there exists a positive constant H such that $\mathbb{E}\{\exp(t\mathbf{X}^{\mathrm{T}}\beta)\} < \infty$ for any $|t| < H$. If the error ε is independent of covariates and satisfies the Cramér condition, then we can obtain that the variable Y also satisfies the Cramér condition. From Lemma 2.2 in Petrov (1995), a random variable W satisfies Cramér condition is equivalent to that there are positive constants b_1 and b_2 such that $\mathbb{P}\{|W| \geq u\} \leq b_1 \exp(-b_2 u)$ for any $u > 0$. Therefore, our assumption here is actually weaker than that in Fan and Song (2010). On the other hand, A.2 is also a general technical assumption in the literature of large derivations. For example, $\gamma_1 = 2$ if X_j's follow normal distribution or sub-Gaussian distribution, and $\gamma_1 = \infty$ if X_j's have compact support.

We now establish the following general result for the distribution of empirical likelihood ratio which is the foundation for our future theoretical results.

Theorem 1—Suppose that $U_1,\ldots, U_n$ are independent and identically distributed random variables. Assume that there exist three positive constants $\tilde{K}_1$, $\tilde{K}_2$ and γ such that $\mathbb{P}\{|U_i| > u\} \leq \tilde{K}_1 \exp(-\tilde{K}_2 u^{\gamma})$ for all $u>0$. Define $\mu_0 = \mathbb{E}(U_i)$, $\delta = \max\{\frac{2}{\gamma}-1, 0\}$, $H = 2^{1+\delta}$ and $\overline{\Delta} = \frac{n^{1/2}\sigma}{2K}$, where $\sigma^2 = \mathbb{E}\{(U_i - \mu_0)^2\}$ and $K > \sigma$ is a sufficiently large positive constant depending only on $\tilde{K}_1$, $\tilde{K}_2$, γ and μ_0, then for $L \to \infty$, there exists a positive constant C only depending on $\tilde{K}_1$, $\tilde{K}_2$ and γ such that

$$\mathbb{P}\left\{\ell(\mu) < \frac{n(\mu-\mu_0)^2}{L^2}\right\} \leq \begin{cases} \exp\left\{-\frac{n(\mu-\mu_0)^2}{4H\sigma^2}\right\} + \exp(-CL^{\gamma}), & if\, n^{1/2}|\mu-\mu_0| \leq \sigma(H^{1+\delta}\overline{\Delta})^{1/(1+2\delta)}; \\ \exp\left\{-\frac{1}{4}\left(\frac{n|\mu-\mu_0|}{2K}\right)^{1/(1+\delta)}\right\} + \exp(-CL^{\gamma}), & if\, n^{1/2}|\mu-\mu_0| > \sigma(H^{1+\delta}\overline{\Delta})^{1/(1+2\delta)}; \end{cases}$$

where $\ell(\mu)$ is defined in Proposition 1.

The proof of Theorem 1 is given in Chang, Tang and Wu (2013), where the main idea is applying large deviation theory [Petrov (1995), Saulis and Statulevičius (1991)].

Theorem 1 reveals the magnitude of the empirical likelihood ratio statistic evaluated at arbitrary values. When $\mu - \mu_0$ does not diminish to 0, Theorem 1 implies that the empirical likelihood ratio statistic diverges with large probability where the diverging rate synthetically depends on the sample size n, some diverging L and the deviation of μ from the truth. Here L is a general technical device whose diverging rate is arbitrary. As a direct result of Theorem 1, we have the following proposition for $\ell_j(0)$.

Proposition 2—Under assumptions A.1 and A.2, there exists a positive constant C_1 depending only on K_1, K_2, γ_1 and γ_2 appeared in assumption A.2 such that for any $j \in \mathcal{M}_*$ and $L \to \infty$,

$$\mathbb{P}\left\{\ell_j(0) < \frac{c_1^2 n^{1-2\kappa}}{L^2}\right\} \leq \begin{cases} \exp(-C_1 n^{1-2\kappa}) + \exp(-C_1 L^{\gamma}), & if (1-2\kappa)\,(1-2\delta) < 1; \\ \exp(-C_1 n^{(1-\kappa)/(1+\delta)}) + \exp(-C_1 L^{\gamma}), & if (1-2\kappa)\,(1+2\delta) \geq 1; \end{cases}$$

where $\gamma=\frac{\gamma_1\gamma_2}{\gamma_1+\gamma_2}$ and $\delta=\max\{\frac{2}{\gamma}-1,0\}$.

Proposition 2 is a uniform result for all features contributing in the true model. Specifically, with large probability and uniformly for all $j\in\mathcal{M}_*$, the diverging rate of $\ell_j(0)$ is not slower than $n^{1-2\kappa}L^{-2}$. From Proposition 1, if $|\mathbb{E}(X_jY)|=O(n^{-w})$ for some $w\in(0,\frac{1}{2})$ and some $j\in\mathcal{M}_*$, then $\ell_j(0)\xrightarrow{p}\infty$. This can be viewed as a requirement such that the signal strength cannot diminish to 0 at a too fast rate. Therefore, $n^{1/2-\kappa}L^{-1}\to\infty$ as $n\to\infty$ is required for sure independence screening. By choosing $L=n^{1/2-\kappa-\tau}$ for some $\tau\in(0,\frac{1}{2}-\kappa)$, we obtain the following corollary more specifically summarizing that the set $\mathcal{M}_*$ can be distinguished by examining the marginal empirical likelihood ratio $\ell_j(0)$ $(j=1,\ldots,p)$.

Corollary 1—Under assumptions A.1 and A.2, there exists a positive constant C_1 depending only on K_1, K_2, γ_1 and γ_2 appeared in assumption A.2 such that, for any $\tau\in(0,\frac{1}{2}-\kappa)$,

$$\max_{j\in\mathscr{M}_*}\mathbb{P}\{\ell_j(0)<c_1^2n^{2\tau}\}\le\begin{cases}\exp\{-C_1n^{(1-2\kappa)\wedge((1-2\kappa-2\tau)\gamma/2)}\}, & if(1-2\kappa)(1+2\delta)<1;\\ \exp\{-C_1n^{((1-\kappa)/(1+\delta))\wedge((1-2\kappa-2\tau)\gamma/2)}\}, & if(1-2\kappa)(1+2\delta)\ge 1;\end{cases}$$

where $\gamma=\frac{\gamma_1\gamma_2}{\gamma_1+\gamma_2}$ and $\delta=\max\{\frac{2}{\gamma}-1,0\}$.

Summarizing above results, we formally establish the screening properties of the marginal empirical likelihood approach.

Theorem 2—Under assumptions A.1 and A.2, there exists a positive constant C_1 depending only on K_1, K_2, γ_1 and γ_2 appeared in assumption A.2 such that, for any $\tau\in(0,\frac{1}{2}-\kappa)$ and $\gamma_n=c_1^2n^{2\tau}$,

$$\mathbb{P}\{\mathscr{M}_*\subset\widehat{\mathscr{M}}_{\gamma_n}\}\ge\begin{cases}1-s\exp\{-C_1n^{(1-2\kappa)\wedge((1-2\kappa-2\tau)\gamma/2)}\}, & if(1-2\kappa)(1+2\delta)<1;\\ 1-s\exp\{-C_1n^{((1-\kappa)/(1+\delta))\wedge((1-2\kappa-2\tau)\gamma/2)}\}, & if(1-2\kappa)(1+2\delta)\ge 1;\end{cases}$$

where $\gamma=\frac{\gamma_1\gamma_2}{\gamma_1+\gamma_2}$ and $\delta=\max\{\frac{2}{\gamma}-1,0\}$.

Theorem 2 implies the sure screening property for our procedure with nonpolynomial dimensionality:

$$\log p=\begin{cases}o(n^{(1-2\kappa)\wedge((1-2\kappa-2\tau)\gamma/2)}), & \text{if }(1-2\kappa)(1+2\delta)<1;\\ o(n^{((1-\kappa)/(1+\delta))\wedge((1-2\kappa-2\tau)\gamma/2)}), & \text{if }(1-2\kappa)(1+2\delta)\ge 1.\end{cases}$$

When the covariates and error are normal, $\gamma_1=2$ and $\gamma_2=2$. Then $\gamma=1$, $\delta=1$ and $\log p=o(n^{1/2-\kappa})$ which is weaker than that in Fan and Lv (2008) where $\log p=o(n^{1-2\kappa})$ is allowed. This can be viewed as a price paid for allowing nonnormal covariate and more general error distribution. Furthermore, we compare our result and that in Fan and Song (2010). The Lemma 1 in Fan and Song (2010) means that $\gamma_2=1$. The corresponding parameters under their this setting are $\gamma=\frac{\gamma_1}{\gamma_1+1}$ and $\delta=\frac{\gamma_1+2}{\gamma_1}$, respectively. Then, we can handle the nonpolynomial dimensionality

$$\log p = o\left(n^{(1-2\kappa)\gamma_1/(2\gamma_1+2)}\right)$$

in this setting, which is actually a stronger result than that in Fan and Song (2010) where log $p = o(n^{(1-2\kappa)\gamma_1/A})$ and $A = \max\{\gamma_1 + 4, 3\gamma_1 + 2\}$.

Now we investigate how large the set $\widehat{\mathscr{M}}_{\gamma_n}$ is. This question is closely related to the asymptotic property of $\ell_j(0)$ for $j \notin \mathscr{M}_*$. Essentially, we need to know the magnitudes of $\ell_j(0)$ for $j \notin \mathscr{M}_*$. We first consider the simple case $\rho_j = 0$ for any $j \notin \mathscr{M}_*$ and have the following result.

Proposition 3—Under assumptions A.1 and A.2, if $\rho_j = 0$, there is a positive constant C_2 depending only on K_1, K_2, γ_1 and γ_2 appeared in assumption A.2 such that, for any $\tau \in (0, \frac{1}{2}-\kappa)$,

$$\mathbb{P}\{\ell_j(0) \ge c_1^2 n^{2\tau}\} \le \begin{cases} \exp(-C_2 n^{2\tau}), & if\ \gamma<4\ and\ \tau \le \frac{\gamma}{12}; \\ \exp(-C_2 n^{\gamma/6}), & if\ \gamma<4\ and\ \tau>\frac{\gamma}{12}; \\ \exp(-C_2 n^{2\tau}), & if\ \gamma \ge 4\ and\ \tau \le \frac{\gamma}{2\gamma+4}; \\ \exp(-C_2 n^{\gamma/(\gamma+2)}), & if\ \gamma \ge 4\ and\ \tau>\frac{\gamma}{2\gamma+4}; \end{cases}$$

where $\gamma=\frac{\gamma_1\gamma_2}{\gamma_1+\gamma_2}$.

The assumption $\rho_j = 0$ for any $j \notin \mathscr{M}_*$ can be guaranteed by the partial orthogonality condition, that is, $\{X_j : j \notin \mathscr{M}_*\}$ is independent of $\{X_j : j \in \mathscr{M}_*\}$. The orthogonality condition is essentially the assumption made in Huang, Horowitz and Ma (2008) who showed the model selection consistency in the case with the ordinary linear model and bridge regression. This proposition gives the property of $\ell_j(0)$ for any $j \notin \mathscr{M}_*$ which can be used to establish the theoretical result for the size of $\widehat{\mathscr{M}}_{\gamma_n}$ where $\gamma_n=c_1^2 n^{2\tau}$. Note that

$$\begin{aligned} |\widehat{\mathscr{M}}_{\gamma_n}| &= \sum_{j\in\mathscr{M}_*} I\{\ell_j(0) \ge c_1^2 n^{2\tau}\} + \sum_{j\notin\mathscr{M}_*} I\{\ell_j(0) \ge c_1^2 n^{2\tau}\} \\ &\le s + \sum_{j\notin\mathscr{M}_*} I\{\ell_j(0) \ge c_1^2 n^{2\tau}\}, \end{aligned}$$

then

$$\mathbb{P}\{|\widehat{\mathscr{M}}_{\gamma_n}|>s\} \le \sum_{j\notin\mathscr{M}_*} \mathbb{P}\{\ell_j(0) \ge c_1^2 n^{2\tau}\}.$$

By Proposition 3, we obtain the following theorem.

Theorem 3—Under assumptions A.1 and A.2, if $\rho_j = 0$ for any $j \notin \mathscr{M}_*$, then there exists a positive constant C_2 depending only on K_1, K_2, γ_1 and γ_2 appeared in assumption A.2 such that, for any $\tau \in (0, \frac{1}{2}-\kappa)$ and $\gamma_n=c_1^2 n^{2\tau}$,

$$\mathbb{P}\{|\widehat{\mathscr{M}}_{\gamma_n}|>s\} \le \begin{cases} p\exp(-C_2 n^{2\tau}), & if\ \gamma<4\ and\ \tau \le \frac{\gamma}{12}; \\ p\exp(-C_2 n^{\gamma/6}), & if\ \gamma<4\ and\ \tau>\frac{\gamma}{12}; \\ p\exp(-C_2 n^{2\tau}), & if\ \gamma \ge 4\ and\ \tau \le \frac{\gamma}{2\gamma+4}; \\ p\exp(-C_2 n^{\gamma/(\gamma+2)}), & if\ \gamma \ge 4\ and\ \tau>\frac{\gamma}{2\gamma+4}; \end{cases}$$

where $\gamma=\frac{\gamma_1\gamma_2}{\gamma_1+\gamma_2}$.

From Theorem 3, we have $\mathbb{P}\{|\widehat{\mathcal{M}}_{\gamma_n}| > s\} \le p\exp\{-C_2 n^{(2\tau)\wedge(\gamma/6)\wedge(\gamma/(\gamma+2))}\}$ which means that the event $\{|\widehat{\mathcal{M}}_{\gamma_n}| \le s\}$ occurs with probability approaching to 1 if $\log p = o(n^{(2\tau)\wedge(\gamma/6)\wedge(\gamma/(\gamma+2))})$. On the other hand, following Theorem 2, we have $\mathbb{P}\{\mathcal{M}_* \subset \widehat{\mathcal{M}}_{\gamma_n}\} \to 1$ provided $\log p = o(n^{((1-2\kappa-2\tau)\gamma/2)\wedge(1-2\kappa)})$. Combining these two results together, we can obtain that

$$\mathbb{P}\{\widehat{\mathscr{M}}_{\gamma_n}=\mathscr{M}_*\} \to 1 \quad \text{if}\log p=o(n^{(\gamma/6)\wedge((1-2\kappa)\gamma/(\gamma+2))})$$

and

$$\tau=\frac{(1-2\kappa)\gamma}{2\gamma+4}.$$

This property shows the selection consistency of our procedure. In a more general case without partial orthogonality condition, we could consider the size of the set $\widehat{\mathcal{M}}_{\gamma_n}$ under the setting

$$\max_{j\notin\mathscr{M}_*}|\rho_j|=o\left(n^{-\kappa}\right),$$

which is an assumption imposed in Fan and Song (2010).

Proposition 4—Under assumptions A.1 and A.2, if $\max_{j\notin \mathcal{M}_*}|\rho_j| = O(n^{-\eta})$ where $\eta > \kappa$ and $\min_{j\notin\mathscr{M}_*}\mathbb{E}(X_j^2Y^2) \ge c_2$ for some $c_2 > 0$, there exists a positive constant C_3 depending only on K_1, K_2, γ_1 and γ_2 appeared in assumption A.2 and c_2 such that, for any $j \notin \mathcal{M}_*$ and $\tau \in (\frac{1}{2}-\eta, \frac{1}{2}-\kappa)$,

$$\mathbb{P}\{\ell_j(0) \ge c_1^2 n^{2\tau}\} \le \begin{cases} \exp(-C_3 n^{2\tau})+\exp(-C_3 n^{\gamma/6}), & if\ \gamma<2\ and\ \eta>\frac{1}{4}; \\ \exp(-C_3 n^{\gamma\,\eta})+\exp(-C_3 n^{\gamma/6}), & if\ \gamma<2\ and\ \eta \le \frac{1}{4}; \\ \exp(-C_3 n^{\gamma\,\eta})+\exp(-C_3 n^{\gamma/6}), & if\ \gamma \ge 2\ and\ \eta \le \frac{1}{\gamma+2}; \\ \exp(-C_3 n^{\gamma/(\gamma+2)})+\exp(-C_3 n^{2\tau}), & if\ \gamma \ge 4\ and\ \eta>\frac{1}{\gamma+2}; \\ \exp(-C_3 n^{\gamma/6})+\exp(-C_3 n^{2\tau}), & if 2 \le \gamma<4\ and\ \eta>\frac{1}{\gamma+2}; \end{cases}$$

where $\gamma=\frac{\gamma_1\gamma_2}{\gamma_1+\gamma_2}$.

If $\rho_j = 0$ for any $j \notin \mathcal{M}_*$, then $\eta = \infty$. Hence, this proposition reduces to Proposition 3. Following the same argument between Proposition 3 and Theorem 3, we can obtain the following theorem related to the size of $\widehat{\mathcal{M}}_{\gamma_n}$.

Theorem 4—Under assumptions A.1 and A.2, if $\max_{j \notin \mathcal{M}_*} |\rho_j| = O(n^{-\eta})$ where $\eta > \kappa$ and $\min_{j \notin \mathscr{M}_*} \mathbb{E}(X_j^2 Y^2) \ge c_2$ for some $c_2 > 0$, then there exists a positive constant C_3 only depending on K_1, K_2, γ_1 and γ_2 appeared in assumption A.2 and c_2 such that, for any $\tau \in (\frac{1}{2}-\eta, \frac{1}{2}-\kappa)$ and $\gamma_n = c_1^2 n^{2\tau}$,

$$\mathbb{P}\{|\widehat{\mathscr{M}}_{\gamma_n}| > s\} \le \begin{cases} p\exp(-C_3 n^{2\tau}) + p\exp(-C_3 n^{\gamma/6}), & if\ \gamma < 2\ and\ \eta > \frac{1}{4}; \\ p\exp(-C_3 n^{\gamma\,\eta}) + p\exp(-C_3 n^{\gamma/6}), & if\ \gamma < 2\ and\ \eta \le \frac{1}{4}; \\ p\exp(-C_3 n^{\gamma\,\eta}) + p\exp(-C_3 n^{\gamma/6}), & if\ \gamma \ge 2\ and\ \eta \le \frac{1}{\gamma+2}; \\ p\exp(-C_3 n^{\gamma/(\gamma+2)}) + p\exp(-C_3 n^{2\tau}), & if\ \gamma \ge 4\ and\ \eta > \frac{1}{\gamma+2}; \\ p\exp(-C_3 \eta^{\gamma/6}) + p\exp(-C_3 n^{2\tau}), & if 2 \le \gamma < 4\ and\ \eta > \frac{1}{\gamma+2}; \end{cases}$$

where $\gamma = \frac{\gamma_1 \gamma_2}{\gamma_1 + \gamma_2}$.

In summary, our results show that the marginal empirical likelihood approach has a very good control of the size of the recruited variables. With large probability, the set of the recruited variables is not larger than the true contributing explanatory variables. As shown later in our simulation results, the marginal empirical likelihood approach perform very well in terms of the set of false selected variables by the marginal empirical likelihood approach.

4. Extensions

4.1. A broad framework

The marginal empirical likelihood can be applied in a general framework besides the linear models and generalized linear models. Based on general estimating equations approach [Hansen (1982), Qin and Lawless (1994)], we can also apply the screening procedure based on the marginal empirical likelihood. We will demonstrate that the marginal empirical likelihood approach provides an effective device to combine information that can be used to enhance the performance of a screening procedure.

Let $\mathbf{Z}_i \in \mathbb{R}^d$ ($i = 1, \ldots, n$) be generic observations, $\beta = (\beta_1, \ldots, \beta_p)^{\mathrm{T}} \in \mathbb{R}^p$ be parameter of interest and $\mathbf{g}(\mathbf{Z}; \beta) = (g_1(\mathbf{Z}; \beta), \ldots, g_r(\mathbf{Z}; \beta))^{\mathrm{T}}$ be the r-dimensional estimating function such that $\mathbb{E}\{\mathbf{g}(\mathbf{Z}; \beta)\} = \mathbf{0}$. Let $\mathcal{M}_* = \{1 \le j \le p : \beta_j \ne 0\}$ be the true model with size $|\mathcal{M}_*| = s$. We are interested in how to construct a sure feature screening procedure to recover $\mathcal{M}_*$ in the general estimating equation setting. To motivate the marginal empirical likelihood approach, let us consider the estimating function evaluated at

$$\beta^{(j)} = (\underbrace{0, \ldots, 0}_{j-1}, \beta, \underbrace{0, \ldots, 0}_{p-j})^{\mathrm{T}} \quad (j = 1, \ldots, p).$$

In practice, many components in $\mathbf{g}(\mathbf{Z}; \beta^{(j)})$ do not involve the unknown parameter; see, for example, the estimating function constructed from the least-squares method and our example given later. Therefore, we denote by

$$\mathbf{g}^{(j)}(\boldsymbol{Z};\beta)=(g_1^{(j)}(\boldsymbol{Z};\beta),\ldots,g_{r_j}^{(j)}(\boldsymbol{Z};\beta))^{\mathrm{T}}$$

an r_j ($r_j \geq 1$)-dimensional estimating function collecting the components in $\mathbf{g}(\mathbf{Z};\boldsymbol{\beta}^{(j)})$ that depend on the unknown parameter. Usually $r_j > 1$ is small and not all components of $\mathbf{Z}$ are involved in $\mathbf{g}^{(j)}(\mathbf{Z};\beta)$. A remarkable advantage of this broad framework is that it provides a device for feature screening using more flexibly constructed conditions so that additional data information can be more effectively incorporated.

Correspondingly, we define the marginal empirical likelihood for β as

$$\mathrm{EL}_j(\beta)=\sup\left\{\prod_{i=1}^n w_i : w_i \geq 0, \sum_{i=1}^n w_i=1, \sum_{i=1}^n w_i \mathbf{g}^{(j)}(\boldsymbol{Z}_i;\beta)=\mathbf{0}\right\}. \quad (4.1)$$

Then screening can be done based on the ranking of $\mathrm{EL}_j(0)$ or equivalently using the corresponding marginal empirical likelihood ratio evaluated at 0—that is, $\ell_j(0)$. The steps of the procedure are the same as those described earlier. A concrete example of this scenario is given as follows.

Example (Quadratic inference function (QIF) approach [Qu, Lindsay and Li (2000)])—Longitudinal data arise commonly in biomedical research with repeated measurements from the same subject or within the same cluster. Let Y_{it} and $\mathbf{X}_{it}$($i = 1, \ldots, n$, $t = 1, \ldots, m_i$) be the response and covariates of the ith subject measured at time t. Let $\mathbb{E}(Y_{it})=\mu(\mathbf{X}_{it}^{\mathrm{T}}\beta)=\mu_{it}$ where $\boldsymbol{\beta} \in \mathbb{R}^p$ is the parameter of interest. Incorporating the dependence among the repeated measurements is essential for efficient inference. Liang and Zeger (1986) proposed to estimate $\boldsymbol{\beta}$ by solving $0=\sum_{i=1}^n \dot{\mu}_i^{\mathrm{T}}\mathbf{W}_i^{-1}(\boldsymbol{Y}_i-\mu_i)$. Here for the ith subject, $\mathbf{Y}_i = (\mathrm{Y}_{i1}, \ldots, \mathrm{Y}_{im_i})^{\mathrm{T}}$, $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{im_i})^{\mathrm{T}}$, $\dot{\mu}_i=\frac{\partial\mu_i}{\partial\beta}$ and $\mathbf{W}_i=\mathbf{v}_i^{1/2}\mathbf{R}\mathbf{v}_i^{1/2}$, where $\mathbf{v}_i$ is a diagonal matrix of the conditional variances of subject i and $\mathbf{R}$ is a working correlation matrix that may depend on some unknown parameter. This approach uses estimating function $\mathbf{g}(\boldsymbol{Z}_i;\beta)=\dot{\mu}_i^{\mathrm{T}}\mathbf{W}_i^{-1}(\boldsymbol{Y}_i-\mu_i)$, where $\boldsymbol{Z}_i=(\boldsymbol{Z}_{i1}^{\mathrm{T}},\ldots,\boldsymbol{Z}_{im_i}^{\mathrm{T}})^{\mathrm{T}}, \boldsymbol{Z}_{it}=(Y_{it},\mathbf{X}_{it}^{\mathrm{T}})^{\mathrm{T}}$ and $r = p$. More recently, Qu, Lindsay and Li (2000) proposed to model $\mathbf{R}^{-1}$ by $\sum_{i=1}^m a_i\mathbf{M}_i$, where $\mathbf{M}_1, \ldots, \mathbf{M}_m$ are known matrices and $a_1, \ldots, a_m$ are unknown constants. Then $\boldsymbol{\beta}$ can be estimated by the quadratic inference functions approach [Qu, Lindsay and Li (2000)] that uses

$$\mathbf{g}(\boldsymbol{Z}_i;\beta)=\begin{pmatrix}\dot{\mu}_i^{\mathrm{T}}\mathbf{v}_i^{-1/2}\mathbf{M}_1\mathbf{v}_i^{-1/2}(\boldsymbol{Y}_i-\mu_i)\\ \vdots \\ \dot{\mu}_i^{\mathrm{T}}\mathbf{v}_i^{-1/2}\mathbf{M}_m\mathbf{v}_i^{-1/2}(\boldsymbol{Y}_i-\mu_i)\end{pmatrix} \quad (i=1,\ldots,n). \quad (4.2)$$

This falls into our framework with $r > p$ when $m > 1$, and with $r = p$ if $m = 1$. When applying the marginal approach, we note that $\mathbf{g}^{(j)}(\mathbf{Z};\beta)$ is an m-dimensional estimating function. The marginal screening by empirical likelihood can be conveniently applied to this scenario, and we note that the existing independence screening methods cannot be directly applied when $m > 1$.

In a concurrent and independent work, Zhao and Li (2012) considered feature screening using estimating functions when $r = p$. By using our notations, their approach are based on $\mathbf{g}^{(j)}(\mathbf{Z}; 0)$—the marginal estimating function evaluated at 0. Their screening procedure are based on ranking the absolute value of $\mathbf{g}^{(j)}(\mathbf{Z}; 0)$ for $j = 1, \ldots, p$. Our approach is different as seen from the above marginal empirical likelihood construction. In addition, analogous to that in linear models and generalized linear models, the marginal empirical likelihood constructed from using the marginal estimating function is also capable of incorporating the level of uncertainties associated with finite sample estimating functions.

We now characterize the properties of the screening procedure in the framework of models specified by estimating equations. For any vector $\mathbf{a} =(a_1, \ldots, a_q)^{\mathrm{T}} \in \mathbb{R}^q$, we use $\|\mathbf{a}\|_\infty = \max_{i=1, \ldots, q} |a_q|$ and $\|\mathbf{a}\|_2=(\sum_{i=1}^q a_i^2)^{1/2}$ to denote its L_∞ and L_2 norms, respectively. Aiming to establish the theoretical results, we need the following two assumptions.

A.3: There exists a positive constant c_3 such that

$$\min_{j\in\mathscr{M}_*} \|\mathbb{E}\{\mathbf{g}^{(j)}(\boldsymbol{Z};0)\}\|_\infty \geq c_3 n^{-\kappa}$$

for some $\kappa \in [0, \frac{1}{2})$.

A.4: There are positive constants K_3, K_4 and γ_3 such that

$$\mathbb{P}\{\|\mathbf{g}^{(j)}(\boldsymbol{Z};0)\|_2 \geq u\} \leq K_3\exp(-K_4 u^{\gamma_3})$$

for each $j = 1, \ldots, p$ and any $u > 0$.

Assumption A.3 is a general identification condition for the set $\mathscr{M}_*$ when considering the broad framework of models specified by general estimating equations. It means that the weakest signals reflected by $\| \mathbb{E}\{\mathbf{g}^{(j)}(\mathbf{Z}; 0)\}\|_\infty$ ($j \in \mathscr{M}_*$) cannot vanish at a rate faster than $n^{-1/2}$. Assumption A.3 is not stringent, and it reduces to A.1 in special cases of linear models and generalized linear models. A similar assumption is also made in Zhao and Li (2012). Assumption A.4, which is a counterpart of A.2 in general cases, is required for establishing exponential inequality when analyzing large deviations. Zhao and Li (2012) assumed boundness of all components in $\mathbf{g}^{(j)}(\mathbf{Z}; 0)$, which implies A.4.

Theorem 5—Under assumptions A.3–A.4, there exists a positive constant C_4 depending only on K_3, K_4 and γ_3 appeared in assumption A.4 such that, for any $\tau \in (0, \frac{1}{2}-\kappa)$ and $\gamma_n=c_3^2 n^{2\tau}$,

$$\mathbb{P}\{\mathscr{M}_* \subset \widehat{\mathscr{M}}_{\gamma_n}\} \leq \begin{cases} 1-s\exp\{-C_4 n^{(1-2\kappa)\wedge((1-2\kappa-2\tau)\gamma_3/2)}\}, & if(1-2\kappa)(1+2\delta)<1; \\ 1-s\exp\{-C_4 n^{((1-\kappa)/(1+\delta))\wedge((1-2\kappa-2\tau)\gamma_3/2)}\}, & if(1-2\kappa)(1+2\delta) \geq 1; \end{cases}$$

where $\delta= \max\{\frac{2}{\gamma_3}-1, 0\}$.

This theorem is a natural extension of Theorem 2 in the broad framework for models specified by general estimating equations. In special cases, we have considered for linear models and generalized linear models, $\mathbf{g}^{(j)}(\mathbf{Z}; 0) = X_j Y (j = 1, \ldots, p)$, and γ_3 in assumption A.4 is equal to $\frac{\gamma_1\gamma_2}{\gamma_1+\gamma_2}$ where γ_1 and γ_2 are specified in A.2.

Let

$$\mathbf{u}_j = \mathbb{E}\{\mathbf{g}^{(j)}(\boldsymbol{Z};0)\} \quad \text{for each } j=1,\ldots,p.$$

We now consider the size of $\widehat{\mathcal{M}}_{\gamma_n}$ in the setting

$$\max_{j \notin \mathscr{M}_*} \|\mathbf{u}_j\|_\infty = o(n^{-\kappa}).$$

This specification also reduces to those considered in special cases of linear models and generalized linear models. The counterpart of Theorem 4 for establishing the selection consistency is given as follows.

Theorem 6—Under assumptions A.3 and A.4, if $\max_{j \notin \mathcal{M}_*} \|\mathbf{u}_j\|_\infty = O(n^{-\eta})$ where $\eta > \kappa$ and $\min_{j \notin \mathcal{M}_*} \lambda_{\min}(\mathbb{E}\{\mathbf{g}^{(j)}(\mathbf{Z}; 0)\mathbf{g}^{(j)}(\mathbf{Z}; 0)^{\mathrm{T}}\}) \geq c_4$ for some $c_4 > 0$, where $\lambda_{\min}(A)$ means the smallest eigenvalue of A, then there exists a positive constant C_5 depending only on K_3, K_4 and γ_3 appeared in assumption A.4 and c_4 such that, for any $\tau \in (\frac{1}{2}-\eta, \frac{1}{2}-\kappa)$and, $\gamma_n = c_3, n^{2\tau}$

$$\mathbb{P}\{|\widehat{\mathscr{M}}_{\gamma_n}| > s\} \leq \begin{cases} p\exp(-C_5 n^{2\tau}) + p\exp(-C_5 n^{\gamma_3/6}), & if\, \gamma_3 < 2 \; and\, \eta > \frac{1}{4}; \\ p\exp(-C_5 n^{\gamma_3\eta}) + p\exp(-C_5 n^{\gamma_3/6}), & if\, \gamma_3 < 2 \; and\, \eta \leq \frac{1}{4}; \\ p\exp(-C_5 n^{\gamma_3\eta}) + p\exp(-C_5 n^{\gamma_3/6}), & if\, \gamma_3 \geq 2 \; and\, \eta \leq \frac{1}{\gamma_3+2}; \\ p\exp(-C_5 n^{\gamma_3/(\gamma_3+2)}) + p\exp(-C_5 n^{2\tau}), & if\, \gamma_3 \geq 4 \; and\, \eta > \frac{1}{\gamma_3+2}; \\ p\exp(-C_5 n^{\gamma_3/6}) + p\exp(-C_5 n^{2\tau}), & if\, 2 \leq \gamma_3 < 4 \; and\, \eta > \frac{1}{\gamma_3+2}. \end{cases}$$

Combining the Theorems 5 and 6, we can see that the screening procedure using the marginal empirical likelihood ratio is valid in a broad framework for identifying the set of the effective features.

4.2. Iterative screening procedure

As we can see from the main results, the proposed marginal empirical likelihood screening procedure works ideally for the case with explanatory variables that are independent of each other. To deal with challenging situations with correlated explanatory variables, we propose to use the following iterative sure independence screening procedure.

Step 1: Rank explanatory variables according to $\ell_j(0)$ by (2.5) and select top ranked explanatory variables with largest values of $\ell_j(0)$'s until some desirable number of features are included. Denote the set of select explanatory variables by $\widehat{\mathcal{M}}_1$.

Step 1′: Apply penalized empirical likelihood [Leng and Tang (2012), Tang and Leng (2010)] to explanatory variables in $\widehat{\mathcal{M}}_1$ and denote the final model by $\widehat{\mathcal{A}}_1$.

Step 2: Let $\widehat{\mathcal{A}}_k \subset \{1, \ldots, p\}$ be the selected model at the kth step. At the kth iteration, for each $j \notin \widehat{\mathcal{A}}_k$, denote by

$$\mathrm{EL}_{\{j\}\cup\widehat{\mathscr{A}}_k}(\mu) = \sup\left\{\prod_{i=1}^{n} w_i : w_i \geq 0, \sum_{i=1}^{n} w_i = 1, \sum_{i=1}^{n} w_i \mathbf{X}_{i,\{j\}\cup\widehat{\mathscr{A}}_k} Y_i = \mu\right\}$$

the empirical likelihood for the combined covariates, and denote by

$$\widetilde{\mathrm{EL}}_j(\mu)=\sup_{\mu_j=\mu}\{\mathrm{EL}_{\{j\}\cup\widehat{\mathscr{A}}_k}(\mu)\}$$

the profile empirical likelihood evaluated at μ. Rank explanatory variable j in $\widehat{\mathscr{A}}_k^c$ according to $\widetilde{\mathrm{EL}}_j(0)$ and select the top ranked until some desirable number of features are included. Denote the selected set by $\widehat{\mathscr{M}}_{k+1}$.

Step 2′: Apply penalized empirical likelihood to explanatory variables in $\widehat{\mathscr{A}}_k \cup \widehat{\mathscr{M}}_{k+1}$ and denote the final model by $\widehat{\mathscr{A}}_{k+1}$.

Step 3: Repeat steps 2 and 2′ when either $\widehat{\mathscr{A}}_{k+1} = \widehat{\mathscr{A}}_k$ or the size of $\widehat{\mathscr{A}}_{k+1}$ reaches a pre-specified number.

The above iterative screening procedure incorporates the profile empirical likelihood. The rationale behind it is to capture the joint impact that may be invisible using the marginal screening procedure if correlations exist among those covariates. Our iterative screening procedure shares some similar features of the analogous ones in Fan and Lv (2008) and Fan and Song (2010). However, on the other hand, the iterative procedure using the profile empirical likelihood ratio shares the feature of the marginal empirical likelihood approach by incorporating the level of uncertainties. In addition, we note that the above iterative procedure is generally applicable in a broad framework.

5. Numerical examples

In this section, we use five simulation examples and a real data example to demonstrate the performance of the proposed empirical likelihood-based screening procedure (denoted by EL-SIS) and corresponding iterative procedure (denoted by EL-ISIS). Depending on the example setting, we compare it with the screening methods proposed in Fan and Lv (2008) (denoted by LS-SIS and LS-ISIS) and Fan and Song (2010) (denoted by GLM-SIS and GLM-ISIS) for linear regression models and generalized linear models, respectively. Whenever appropriate, we compare to the robust rank correlation based screening (RRC-SIS and RRC-ISIS) studied by Li et al. (2012). For all simulation examples, we begin with p = 1000 explanatory variables and screen to a much smaller number d of explanatory variables. The respective SCAD penalized variable selection is further applied to these selected explanatory variables to get the corresponding final model. Results over 200 repetitions are reported. For each case, we report the number of repetitions that each important explanatory variable is selected in the final model and also the average number of unimportant explanatory variables being selected.

Example 1—This example has a very standard setting with three important explanatory variables and is taken from Fan and Lv (2008). Covariates are generated as $X_j \sim N(0, 1)$ and $\mathrm{cov}(X_j, X_{j'}) = 1$ if $j = j'$ and 0.3 otherwise. The response is generated as $Y = 5X_1 + 5X_2 + 5X_3 + \varepsilon$ with error being independent of the explanatory variables. We consider three different error distribution $N(0, 1)$, $N(0, 2^2)$, and t_4 for ε. Random samples of size $n = 100$ are used and we set $d = \lfloor n/(2 \log n)\rfloor = 10$, where $\lfloor a \rfloor$ denotes the largest integer that is less than or equal to a. Results over 200 repetitions are reported in Table 1, where we report the number of repetitions that each of the important explanatory variables X_1, X_2 and X_3 is selected. For unimportant explanatory variables, Table 1 reports their average number of repetitions for each being selected. It shows that the proposed empirical likelihood-based screening methods perform very competitively when compared to the least squares-based screening or the robust rank correlation-based screening.

Example 2—The second example is also from Fan and Lv (2008) and has a hidden important explanatory variable, which is important but marginally uncorrelated with the response. This example is to illustrate that the proposed iterative empirical likelihood-based screening works effectively in such challenging cases. Covariates are generated as $X_j \sim N(0, 1)$ and $\mathrm{cov}(X_j, X_{j'}) = 1$ if $j = j'$ and 0.3 otherwise except $\mathrm{cov}(X_4, X_j) = \sqrt{0.3}$ for $j \neq 4$. The response is generated as $Y = 5X_1 + 5X_2 + 5X_3 - 15\sqrt{0.3}X_4 + \varepsilon$ with ε being independent of explanatory variables. We consider three different error distribution $N(0, 1)$, $N(0, 2^2)$, and t_4. Results over 200 repetitions with $n = 100$ and $d = \lfloor n/(2 \log n) \rfloor = 10$ are reported in Table 2. It shows that the empirical likelihood-based screening is challenged by the hidden important explanatory variable X_4 but the corresponding iterative screening can easily pick it up. Overall the performance of the empirical likelihood-based screening methods is very similar to that of the least squares-based screening methods and is better than the robust rank correlation-based screening. Note that iterative version of the robust rank correlation-based screening is residual-based. This explains the improvement of the robust rank correlation-based screening.

Example 3—The performances of the empirical likelihood-based screening and the least squares-based screening methods are very similar in the previous two examples. It is known that the empirical likelihood approach requires a less restrictive distributional assumption. We next use a heteroscedastic example to show the advantage of the empirical likelihood-based screening. Explanatory variables are generated as $X_j \sim N(0, 1)$ with $\mathrm{cov}(X_j, X_{j'}) = 0$ for $j \neq j'$. The response is generated as $Y = c(X_1 - X_2 + X_3) + \varepsilon/(X_1^2 + X_2^2 + X_3^2)$ with independent $\varepsilon \sim N(0, 1)$ and $c > 0$ controls the signal level. Results over 200 repetitions with $n = 70$ and $d = \lfloor n/(2 \log n) \rfloor = 8$ are reported in Table 3 for three different values of c. It shows that the performance of the least squares-based screening is severely affected by the heteroscedasticity especially when the signal level is low. On the other hand, the proposed empirical likelihood-based screening works much better and similarly as the robust rank correlation-based screening.

Example 4—We now consider an example with the extended scope. In this example, we generate data from the longitudinal data example as in Section 4.2 with $m = 4$ means 4 repeated measurements generated. In particular, the following model is generated:

$$Y_{il} = \mathbf{X}_{il}^{\mathrm{T}} \beta + \varepsilon_{il} \quad (i = 1, \ldots, n; l = 1, \ldots, m).$$

Here $\mathbf{X}_{il}$ is generated from multivariate normal $N(\mathbf{0}, \Sigma)$ with $\Sigma = (\sigma_{jk})_{j,k=1, \ldots, p}$ and $\sigma_{jk} = 0.5^{|j-k|}$. The error vector $\varepsilon_i = (\varepsilon_{i1}, \ldots, \varepsilon_{im})^{\mathrm{T}}$ is generated from multivariate normal distribution with unit variance. The correlation structure of ε is specified as AR(1) with parameter 0.8; see Diggle et al. (2002) for reference for the correlation structure. The first five components of the true β is set to be $c \cdot (2.0, -2.0, 0, 0, 2.0)^{\mathrm{T}}$ where c is used to control the signal strength, and all other components of β are zero. We use two sets of basis matrices in (4.2). We take $\mathbf{M}_1 = \mathbf{I}$ as the identity matrix. The second basis matrix $\mathbf{M}_2$ is a matrix with two main off-diagonals being 1 and 0 elsewhere corresponding to the AR(1) working correlation [Qu, Lindsay and Li (2000)]. We then apply the marginal empirical likelihood procedure as in Section 4 using the marginal estimating function of (4.2). Here we note that the marginal estimating function is 4-dimensional. By ignoring the correlation structure of the longitudinal data, the least squares-based screening and robust rank correlation-based screening procedures can be applied. Results over 200 repetitions with $n = 60$ and $d = 15$ are reported in Table 4. From Table 4, we clearly see that the marginal empirical likelihood approach works much better than the alternative ones, especially when signal is relatively weak. The improvement can be seen as the results of incorporating additional data structural information. Hence, we demonstrate an advantage of the marginal empirical likelihood approach of being adaptive and flexible.

In the review process, one referee pointed out that our comparison to the LS-SIS is not fair as it is based on the ordinary least squares. It is more reasonable to compare to a weighted least squares-based screening by adjusting to correlation among longitudinal observations. To address this issue, we implement this weighted least squares-based screening by using the R package "geepack," which can estimate both the correlation structure and regression parameter once a parametric form of the correlation structure is specified. Table 4 is updated accordingly with GEE-SIS denoting this weighted least squares-based screening method. It shows that our newly proposed EL-SIS still does better than the GEE-SIS even though a correct parametric correlation structure, AR(1), is specified.

Example 5—This is an extension of Example 2 to the case with a binary response using logistic regression. Covariates are generated as $X_j \sim N(0, 1)$ and $\text{cov}(X_j, X_{j'}) = 1$ if $j = j'$ and 0.3 otherwise except $\text{cov}(X_4, X_j) = \sqrt{0.3}$ for $j \neq 4$. The binary response is generated from Bernoulli distribution with success probability given by $\{1+\exp(-4X_1-4X_2-4X_3+12\sqrt{0.3}X_4)\}^{-1}$. Results over 200 repetitions with $n = 400$ and $d = 10$ are reported in Table 5. A similar performance pattern is observed. For this example, the result for the iterative version of the robust rank correlation-based screening is not presented since it is not clear how to define a residual-based iterative procedure.

A real data example—Glioblastoma is the most common primary malignant brain tumor of adults and one of the most lethal of all cancers [Horvath et al. (2006)]. The median survival of glioblastoma patients is 15 months from the time of diagnosis. We next apply our proposed methods to a microarray gene expression dataset of glioblastoma patients reported in Horvath et al. (2006). The dataset has been analyzed by Pan, Xie and Shen (2010) and Li and Li (2008) among many others. Drawn from two different studies, the data consist of two independent sets. We use the set with 50 samples. We use the log survival time, measured in years, as the response. The second sample with a outlier response is excluded and the other 49 samples are used in our analysis. Explanatory variables are gene expression profiles of 1523 genes measured on Affymetrix HG-U133A arrays.

We apply the least squares-based and empirical likelihood-based screening methods with $d = 6$. LS-SIS selects "GSN", "FOS", "COL11A1", "AVPR1A", "SELE", and "TBL1X" as important gene explanatory variables while EL-SIS selects "GSN", "JAK2", "COL11A1", "CDK6", "ADCYAP1R1", and "TBL1X". Note that they select some common genes ("GSN" and "COL11A1") and some different genes. LS-ISIS selects "GSN", "COL11A1", "THBS1", "SELE", "TBL1X", and "GCGR". EL-ISIS selects "DUSP7", "COL11A1", "BST1", "ADCYAP1R1", "TBL1X", and "GCGR". Similarly two genes ("TBL1X" and "GCGR") are recruited by the iterative screening methods based on both the least squares and empirical likelihood. The robust rank correlation-based screening performs similarly with 2–3 overlapping genes.

6. Discussion

Screening based on marginal model fitting has enjoyed great popularity in the recent literature. However, most, if not all, of the marginal screening methods studied thus far are based on some restrictive distributional assumptions. Yet these assumptions may not be realistic in applications. Thus motivated we propose a new screening method based on marginal empirical likelihood, which is known to be less restrictive. It has been demonstrated to be effective using both theoretical sure screening property and numerical evidences. Further extensions using empirical likelihood are being investigated.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank the Editor, the Associate Editor and two referees for very constructive comments and suggestions which have improved the presentation of the paper. We are very grateful to Drs. Gaorong Li, Yi Li, Heng Peng and Sihai Dave Zhao for sharing with us programs for implementing their methods and Drs. Hongzhe Li and Wei Pan for sharing the real data.

References

Bühlmann, P.; van de Geer, S. Theory and Applications. Springer; Heidelberg: 2011. Statistics for High-dimensional Data: Methods.

Chang J, Chen SX, Chen X. High dimensional generalized empirical likelihood for moment restrictions with dependent data. 2013 Available at arXiv:1308.5732.

Chang J, Tang CY, Wu Y. Supplement to “Marginal empirical likelihood and sure independence feature screening”. 201310.1214/13-AOS1139SUPP

Chen SX, Cui H. An extended empirical likelihood for generalized linear models. Statist Sinica. 2003; 13:69–81.

Chen SX, Gao J, Tang CY. A test for model specification of diffusion processes. Ann Statist. 2008; 36:167–198.

Chen SX, Peng L, Qin Y-L. Effects of data dimension on empirical likelihood. Biometrika. 2009; 96:711–722.

Chen SX, Van Keilegom I. A review on empirical likelihood methods for regression. TEST. 2009; 18:415–447.

Diggle, PJ.; Heagerty, PJ.; Liang, K-Y.; Zeger, SL. Oxford Statistical Science Series. 2. Oxford Univ. Press; Oxford: 2002. Analysis of Longitudinal Data; p. 25

Fan J, Feng Y, Song R. Nonparametric independence screening in sparse ultra-high-dimensional additive models. J Amer Statist Assoc. 2011; 106:544–557.

Fan J, Lv J. Sure independence screening for ultrahigh dimensional feature space. J R Stat Soc Ser B Stat Methodol. 2008; 70:849–911.

Fan J, Lv J. A selective overview of variable selection in high dimensional feature space. Statist Sinica. 2010; 20:101–148.

Fan J, Song R. Sure independence screening in generalized linear models with NP-dimensionality. Ann Statist. 2010; 38:3567–3604.

Hansen LP. Large sample properties of generalized method of moments estimators. Econometrica. 1982; 50:1029–1054.

Hastie, T.; Tibshirani, R.; Friedman, J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2. Springer; New York: 2009.

Hjort NL, McKeague IW, Van Keilegom I. Extending the scope of empirical likelihood. Ann Statist. 2009; 37:1079–1111.

Horvath S, Zhang B, Carlson M, Lu KV, Zhu S, Felciano RM, Laurance MF, Zhao W, Shu Q, Lee Y, Scheck AC, Liau LM, Wu H, Geschwind DH, Febbo PG, Kornblum HI, Cloughesy TF, Nelson SF, Mischel PS. Analysis of oncogenic signaling networks in glioblastoma identifies ASPM as a molecular target. Proc Natl Acad Sci USA. 2006; 103:17402–17407. [PubMed: 17090670]

Huang J, Horowitz JL, Ma S. Asymptotic properties of bridge estimators in sparse high-dimensional regression models. Ann Statist. 2008; 36:587–613.

Kolaczyk ED. Empirical likelihood for generalized linear models. Statist Sinica. 1994; 4:199–218.

Leng C, Tang CY. Penalized empirical likelihood and growing dimensional general estimating equations. Biometrika. 2012; 99:703–716.

Li C, Li H. Network-constrained regularization and variable selection for analysis of genomic data. Bioinformatics. 2008; 24:1175–1182. [PubMed: 18310618]

Li R, Zhong W, Zhu L. Feature screening via distance correlation learning. J Amer Statist Assoc. 2012; 107:1129–1139.

Li G, Peng H, Zhang J, Zhu L. Robust rank correlation based screening. Ann Statist. 2012; 40:1846–1877.

Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika. 1986; 73:13–22.

McCullagh, P.; Nelder, JA. Generalized Linear Models. Chapman & Hall/CRC; New York: 1989.

Newey WK, Smith RJ. Higher order properties of GMM and generalized empirical likelihood estimators. Econometrica. 2004; 72:219–255.

Owen AB. Empirical likelihood ratio confidence intervals for a single functional. Biometrika. 1988; 75:237–249.

Owen, AB. Empirical Likelihood. Chapman & Hall/CRC; New York: 2001.

Pan W, Xie B, Shen X. Incorporating predictor network in penalized regression with application to microarray data. Biometrics. 2010; 66:474–484. [PubMed: 19645699]

Petrov, VV. Oxford Studies in Probability. Vol. 4. Oxford Univ. Press; New York: 1995. Limit Theorems of Probability Theory: Sequences of Independent Random Variables.

Qin J, Lawless J. Empirical likelihood and general estimating equations. Ann Statist. 1994; 22:300–325.

Qu A, Lindsay BG, Li B. Improving generalised estimating equations using quadratic inference functions. Biometrika. 2000; 87:823–836.

Saulis, L.; Statulevičius, VA. Limit Theorems for Large Deviations Mathematics and Its Applications (Soviet Series). Kluwer Academic; Dordrecht: 1991. p. 73Translated and revised from the 1989 Russian original

Tang CY, Leng C. Penalized high-dimensional empirical likelihood. Biometrika. 2010; 97:905–919.

Wang H. Factor profiled sure independence screening. Biometrika. 2012; 99:15–28.

Xue L, Zou H. Sure independence screening and compressed random sensing. Biometrika. 2011; 98:371–380.

Zhao SD, Li Y. Sure screening for estimating equations in ultra-high dimensions. 2012 Unpublished manuscript.

Zhu L-P, Li L, Li R, Zhu L-X. Model-free feature screening for ultrahigh-dimensional data. J Amer Statist Assoc. 2011; 106:1464–1475.

Table 1

Simulation result for Example 1

ε	Method	X_1	X_2	X_3	Unimportant explanatory variables
$N(0, 1)$	LS-SIS	199	199	200	1.406219
	RRC-SIS	199	199	199	1.407222
	EL-SIS	194	183	185	1.442327
	LS-ISIS	200	200	200	0.965898
	RRC-ISIS	200	200	200	0.800401
	EL-ISIS	200	200	200	0.659980
$N(0, 2^2)$	LS-SIS	199	199	200	1.406219
	RRC-SIS	198	198	199	1.409228
	EL-SIS	192	182	183	1.447342
	LS-ISIS	200	200	200	1.404213
	RRC-ISIS	200	200	200	1.403210
	EL-ISIS	200	200	200	0.980943
t_4	LS-SIS	199	199	200	1.406219
	RRC-SIS	198	199	199	1.408225
	EL-SIS	193	186	187	1.438315
	LS-ISIS	200	200	200	1.383149
	RRC-ISIS	200	200	200	1.362086
	EL-ISIS	200	199	200	0.635908

Table 2

Simulation result for Example 2 with a hidden important explanatory variable X_4

ε	Method	X_1	X_2	X_3	X_4 (hidden)	Unimportant explanatory variables
$N(0, 1)$	LS-SIS	198	197	195	0	1.415663
	RRC-SIS	196	197	194	0	1.418675
	EL-SIS	198	198	197	0	1.412651
	LS-ISIS	200	200	199	196	1.125502
	RRC-ISIS	200	199	200	111	1.157631
	EL-ISIS	199	199	200	193	0.853414
$N(0, 2^2)$	LS-SIS	198	197	194	0	1.416667
	RRC-SIS	196	196	194	0	1.419679
	EL-SIS	198	196	194	0	1.417671
	LS-ISIS	199	200	199	196	1.210843
	RRC-ISIS	199	199	200	96	1.311245
	EL-ISIS	200	200	198	188	0.912651
t_4	LS-SIS	197	197	197	0	1.414659
	RRC-SIS	195	198	196	0	1.416667
	EL-SIS	197	198	196	0	1.414659
	LS-ISIS	199	200	200	196	1.209839
	RRC-ISIS	200	200	199	100	1.305221
	EL-ISIS	200	198	200	185	0.824297

Table 3

Simulation result for Example 3

c	Method	X_1	X_2	X_3	Unimportant explanatory variables
1	LS-SIS	149	147	156	1.151454
	RRC-SIS	191	185	190	1.037111
	EL-SIS	190	184	191	1.038114
1.5	LS-SIS	173	171	174	1.085256
	RRC-SIS	194	191	193	1.025075
	EL-SIS	196	192	194	1.021063
2	LS-SIS	182	182	180	1.059178
	RRC-SIS	194	194	195	1.020060
	EL-SIS	199	195	194	1.015045

Table 4

Simulation result for the longitudinal data estimation function example with *c* controlling the signal strength

c	Method	X_1	X_2	X_5	Unimportant explanatory variables
1	LS-SIS	90	73	153	2.692076
	GEE-SIS	111	111	168	2.617854
	RRC-SIS	84	66	136	2.722166
	EL-SIS	135	128	191	2.553661
1.5	LS-SIS	153	153	195	2.506520
	GEE-SIS	165	160	196	2.486459
	RRC-SIS	142	136	193	2.536610
	EL-SIS	176	187	199	2.445336
2	LS-SIS	183	183	200	2.441324
	GEE-SIS	183	184	200	2.440321
	RRC-SIS	179	176	200	2.452357
	EL-SIS	192	196	200	2.419258
2.5	LS-SIS	195	195	200	2.417252
	GEE-SIS	196	195	200	2.416249
	RRC-SIS	192	190	200	2.425276
	EL-SIS	198	197	200	2.412237
3	LS-SIS	199	198	200	2.410231
	GEE-SIS	198	197	200	2.412237
	RRC-SIS	199	198	200	2.410231
	EL-SIS	200	198	200	2.409228

Table 5

Simulation result for Example 5

Method	X_1	X_2	X_3	X_4 (hidden)	Unimportant explanatory variables
GLM-SIS	200	200	200	0	1.405623
RRC-SIS	200	200	200	0	1.405623
EL-SIS	200	200	200	0	1.405623
GLM-ISIS	200	200	200	200	0.324297
EL-ISIS	199	200	200	199	0.764056