Ann Inst Stat Math (2017) 69:63–96
DOI 10.1007/s10463-015-0534-9

Conditional sure independence screening by conditional marginal empirical likelihood

Qinqin Hu[1] · Lu Lin[2]

Received: 11 April 2014 / Revised: 26 May 2015 / Published online: 23 August 2015

Abstract In many applications, researchers often know a certain set of predictors is related to the response from some previous investigations and experiences. Based on the conditional information, we propose a conditional screening feature procedure via ranking conditional marginal empirical likelihood ratios. Due to the use of centralized variable, the proposed screening approach works well when there exist either or both hidden important variables and unimportant variables that are highly marginal correlated with the response. Moreover, the new method is demonstrated effective in scenarios with less restrictive distributional assumptions by inheriting the advantage of empirical likelihood approach and is computationally simple because it only needs to evaluate the conditional marginal empirical likelihood ratio at one point, without parameter estimation and iterative algorithm. The theoretical results reveal that the proposed procedure has sure screening properties. The merits of the procedure are illustrated by extensive numerical examples.

Keywords Empirical likelihood · Sure screening · Variable selecting · High dimensional data analysis

The research was supported by NNSF projects (11171188, 11071145, 11221061, and 11231005) and the 111 project (B12023) of China, NSF and SRRF projects (ZR2010AZ001 and BS2011SF006) of Shandong Province of China.

✉ Qinqin Hu
qinqinhulv@gmail.com

Lu Lin
linlu@sdu.edu.cn

[1] School of Mathematics and Statistics, Shandong University, Weihai, Weihai 264209, People's Republic of China

[2] Shandong University Qilu Securities Institute for Financial Studies, Shandong University, Jinan 250100, People's Republic of China

1 Introduction

Ultrahigh-dimensional data are frequently collected in various frontiers of areas, such as finance, biomedical imaging, and genomics. Many challenges to statistical inference in ultrahigh-dimensional scenarios can be imposed when the number of covariates p may be much larger than the sample size n. In such situations, the sparsity principle is frequently adopted and useful in the analysis of ultrahigh-dimensional data. The sparsity assumption requires that only a limit number of predictors contribute to the response. As a result, variable selection has attracted increasing interests.

Over the last 10 years, there are a great deal of developments in statistical theory and computing on variable selection techniques for ultrahigh-dimensional feature space, see Hastie et al. (2009), Fan et al. (2011b), and Bühlmann and van de Geer (2011) for overviews. To reduce dimension, Tibshirani (1996), Fan and Li (2001), Candes and Tao (2007), Bickel et al. (2009), Fan and Lv (2011), and Zhang and Zhang (2012) proposed techniques to select variables and estimate parameters simultaneously by solving a high-dimensional optimization problem. Efron et al. (2004) and Fan and Lv (2011) proposed various efficient algorithms. However, there are still huge computational challenges when the number of variables grows exponentially with sample size. Fan and Lv (2008), Fan et al. (2009), Hall et al. (2009), Hall and Miller (2009), Fan and Song (2010), Fan et al. (2011a), Li et al. (2012), and Chang et al. (2013a) suggested to screen variables by ranking marginal utility such as marginal correlation with the response. However, due to the correlation among the predictors, the sample marginal screening can screen out hidden important variables who have a big impact on response but are weakly marginal correlated with the response. It also can recruit those variables who have strong marginal utility but are conditionally independent with the response given other variables.

In many applications, based on some previous investigations and experiences, researchers often know a set of certain predictors $X_{\mathcal{C}}$ is related to the response Y in advance. As shown in Barut et al. (2012), conditional information can help reducing the correlation among the variables. They proposed a conditional sure independence screening (CSIS) by the known active predictors which makes it possible to recover the hidden importance variables and reduce the number of false negatives. But the CSIS in Barut et al. (2012) has a strongly restrictive for distributional model assumptions and needs to estimate $\beta_{\mathcal{C}}$ repeatedly when individually measuring the strength of the conditional contribution of the rest variables given $X_{\mathcal{C}}$.

Lots of literature show that the empirical likelihood approach (Owen 1988, 2001) has a nice performance when there is less restrictive distributional assumption for statistical inferences, the details can be found in Qin and Lawless (1994), Newey and Smith (2004), and Chen and Van Keilegom (2009) and so on. Recently, the empirical likelihood approach has also been extended to deal with high-dimensional data; see Hjort et al. (2009), Chen et al. (2009), Tang and Leng (2010), Leng and Tang (2012), Chang et al. (2013a), Chang et al. (2015a), and Chang et al. (2015b). The properties of marginal empirical likelihood approach, where the available features are assessed one at a time individually, are systematically studied in Chang et al. (2013a). The marginal empirical likelihood approach only involves univariate optimizations, which means such a method provides a convenient device for both theoretical analysis and practical implementation. Chang et al. (2013a) found the probabilistic behavior of the marginal empirical likelihood ratios as functions of the parameters of interest that can be evaluated at arbitrary value. The theoretical analyses reveal that the marginal empirical likelihood ratio should not be large when evaluated at the truth and the marginal empirical likelihood ratio statistic diverges with large probability when there is deviation of fixed parameter value from the truth. Therefore, Chang et al. (2013a) proposed a screening procedure based on the marginal empirical likelihood approach (EL-SIS) by ranking the marginal empirical likelihood ratio at zero $l_j(0)$. But the screening procedure based on the marginal empirical likelihood approach (EL-SIS) is severely affected by the correlation among the predictors like other marginal screening procedures.

In this paper, we propose a unified conditional sure screening feature procedure by conditional marginal empirical likelihood ratio, which can be equally applied in both linear models and generalized linear models. It is known that high correlation among variables is a fatal difficulty for marginal feature screenings. In our paper, by centralizing the covariates, the proposed screening procedure is able to handle the issue that there exist either or both hidden important variables and unimportant variables that are highly marginal correlated with the response. Although the iterative version of marginal screening procedures can alleviate the mentioned fatal issue, the iterative algorithms are of computational redundance. Due to the conditional information, our proposed procedure can remedy such problem without iterative algorithm. Hence, our proposal is of computational simplicity. On the other hand, comparing to the conditional sure independence screening (CSIS), our proposed procedure preforms much better when heterogeneity exists in the conditional variance. Owing to inherited the advantage of empirical likelihood, the conditional marginal empirical likelihood ratio statistic is a self-studentized quantity (Owen 2001) while CSIS relies on the ranking of features based on magnitudes of conditional maximum marginal likelihood estimators. Therefore, the proposed procedure is able to incorporate the level of uncertainties associated with the estimators to conduct feature screening. In simulation studies, we will show the newly proposed procedure performs much better than CSIS in heteroscedastic examples. In addition, the proposed screening procedure only needs to evaluate the conditional marginal empirical likelihood at one point, without estimating β_C^M repeatedly and all candidates β_j^M. It must be stressed that the proposed screening procedure gives better results than both EL-SIS and CSIS when the heteroscedastic models have hidden important variables or unimportant variables that are highly marginal correlated with the response. As a result, the newly proposed procedure not only inherits the advantages of EL-SIS and CSIS, but also has flexibility in practice. Our theoretical results reveal that the proposed screening procedure is able to identify the rest features that contribute to the response when the number of rest predictors grows exponentially with the sample size.

The rest of the paper is organized as follows. We introduce the conditional marginal empirical likelihood (CMEL) and the corresponding screening procedure in Sect. 2. Section 3 gives the theoretical properties. Section 4 shows some simulation studies. We conclude with some discussions in Sect. 5. For the ease of presentation, the detailed proofs are collected in the Appendix.

2 Conditional marginal empirical likelihood

In this section, we first introduce our conditional marginal moment condition and the related conditional marginal empirical likelihood for linear models and generalized linear models, respectively, and then get a generalized conditional marginal empirical likelihood with a unified conditional marginal moment function for both models. Based on the properties of the related unified conditional marginal empirical likelihood ratio (CMELR) when evaluated at the truth value or not, we propose a convenient screening feature procedure by evaluating CMELR at zero. Finally, we give the sample version of the CMELR such that the screening feature procedure can be easily applied in practice.

2.1 Conditional marginal empirical likelihood for linear models

We first consider the following linear regression model:

$$Y = X^{\mathrm{T}}\beta + \varepsilon, \tag{1}$$

where $X = (X_1, \ldots, X_p)^{\mathrm{T}}$ is a p-dimensional vector of predictors, Y is the response variable, ε is the random error with conditional zero mean given X, and $\beta = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ is the vector of unknown parameters. We use β^* to denote the true value of the parameters. Hereinafter we assume that the predictors are standardized such that $\mathbb{E}(X_j) = 0$ and $\mathbb{E}(X_j^2) = 1$ for $j = 1, \ldots, p$. Define two index sets as

$$\mathcal{A} = \{k : \beta_k^* \neq 0\}, \quad \bar{\mathcal{A}} = \{k : \beta_k^* = 0\}.$$

Actually $\mathcal{A}$ is the active index set that corresponds to the active predictors, and $\bar{\mathcal{A}}$ is the complement set of $\mathcal{A}$. The model sparsity is assumed in the sense of that the cardinality $s_{\mathcal{A}} = |\mathcal{A}|$ of $\mathcal{A}$ is small, which is satisfied in many practical applications such as in finance, biology, and clinical studies. Let $X_{\mathcal{A}}$ be the corresponding $s_{\mathcal{A}}$-dimensional vector of the active predictors.

As was mentioned in Introduction, in many practical application, researchers have already known certain predictors are important for the response Y by some previous investigations and experiences, which means that a set of active predictors has been determined in advance. Without loss of generality, suppose that these known active predictors are the first $s_{\mathcal{C}}$ components $X_1, \ldots, X_{s_{\mathcal{C}}}$ of X. Denote $X_{\mathcal{C}} = (X_1, \ldots, X_{s_{\mathcal{C}}})^{\mathrm{T}}$, $X_{\mathcal{D}} = (X_{s_{\mathcal{C}}+1}, \ldots, X_p)^{\mathrm{T}}$, and partition the parameters β as $\beta = (\beta_{\mathcal{C}}^{\mathrm{T}}, \beta_{\mathcal{D}}^{\mathrm{T}})^{\mathrm{T}}$, correspondingly. Our aim is then to identify the set $\mathcal{D} \cap \mathcal{A} = \{j \in \mathcal{D} : \beta_j^* \neq 0\}$.

There exist various screening methods for the ranking of features based on magnitudes of some marginal estimators, see Fan and Lv (2008), Fan and Song (2010), Fan et al. (2011a), Barut et al. (2012), Zhu et al. (2011), and Lin et al. (2013), among others. Comparing to other methods, Chang et al. (2013a) first employ the idea of marginal hypothesis testing to handle the feature screening problem while the other methods all deal such a problem by simple marginal estimations. This motivates us to use empirical likelihood ratio evaluated at zero as a criterion for feature screening, since such statistic can be used to against the null hypothesis that the marginal effect is negligible. This new methodology can be used for cases of heteroscedasticity and misspecification, as clearly stated in Chang et al. (2013a, 2015b), because empirical likelihood approach requires a less restrictive distribution assumption. Moreover, the marginal empirical likelihood approach only involves univariate optimizations, it provides a convenient device for both theoretical analysis and practical implementation. Therefore, we construct marginal empirical likelihood under the situation with a known set of active predictors in advance.

To apply marginal empirical likelihood with conditional information, for any given vector or matrix $\beta_{\mathcal{C}}$, let us consider the following moment condition:

$$\mathbb{E}\{[X_j - \mathbb{E}(X_j|X_{\mathcal{C}}^{\mathrm{T}}\beta_{\mathcal{C}})][Y - \mathbb{E}_L(Y|(X_{\mathcal{C}}, X_j))]\} = 0, \quad j \in \mathcal{D}, \tag{2}$$

where $\mathbb{E}_L(Y|(X_{\mathcal{C}}, X_j))$ is the best linear regression fit of Y by using $X_{\mathcal{C}}$ and X_j. It implies

$$\mathbb{E}\{[X_j - \mathbb{E}(X_j|X_{\mathcal{C}}^{\mathrm{T}}\beta_{\mathcal{C}})][Y - X_{\mathcal{C}}^{\mathrm{T}}\beta_{\mathcal{C}}^M - X_j\beta_j^M]\} = 0, \quad j \in \mathcal{D}.$$

Since the centralized variable $X_j - \mathbb{E}(X_j|X_{\mathcal{C}}^{\mathrm{T}}\beta_{\mathcal{C}})$ is completely uncorrelated with $X_{\mathcal{C}}^{\mathrm{T}}\beta_{\mathcal{C}}^M$ for any $\beta_{\mathcal{C}}$ by the property of conditional expectation, the above moment condition is equivalent to

$$\mathbb{E}\{[X_j - \mathbb{E}(X_j|X_{\mathcal{C}}^{\mathrm{T}}\beta_{\mathcal{C}})][Y - X_j\beta_j^M]\} = 0, \quad j \in \mathcal{D}. \tag{3}$$

We use (3) as our conditional marginal moment condition (CMMC).

Note that β_j^M in CMMC (3) measures the strength of the conditional contribution of X_j given $X_{\mathcal{C}}$, called as the conditional marginal signal. We can see that $\beta_j^M = 0$ is equivalent to that the response Y and the centralized variable $X_j - \mathbb{E}(X_j|X_{\mathcal{C}}^{\mathrm{T}}\beta_{\mathcal{C}})$ are marginally uncorrelated. Moreover, conditional marginal signal β_j^M has significant advantages in screening feature than marginal signal β_j^{MUC}, where β_j^{MUC} is the covariance between X_j and Y without conditional information. We can use the following two similar examples mentioned in the introduction of Barut et al. (2012) to explain the point of view.

Example 1 The case when there exist the hidden important variable in models. Consider model (1) with the true $\beta^* = (1, 2, 3, 4, 5, -13.5, 0, \ldots, 0)^{\mathrm{T}}$, and all variables follow the standard normal distribution with equal correlation 0.9, and ε follows the standard normal distribution. By this setting, $\beta_6^{\mathrm{MUC}} = 0$, that means X_6 is a hidden important variable. Hence X_6 cannot be selected by ranking marginal correlation with the response. Now, we assume that $X_{\mathcal{C}} = \{X_1, X_2, X_3, X_4, X_5\}$ is known in advance. For simplicity, we consider the following *linearity condition* among the predictors:

$$\text{(LC)} \quad \mathbb{E}(X|X^{\mathrm{T}}\beta) = \mathrm{cov}(X, X^{\mathrm{T}})\beta\{\mathrm{cov}(X^{\mathrm{T}}\beta)\}^{-1}\beta^{\mathrm{T}}X \quad \text{for any } \beta. \tag{4}$$

Condition LC is a regular condition and it always holds when X follows normal, or more generally, elliptical distribution. According to CMMC (3) and Condition LC, choosing $\beta_C = I_5$, a 5 × 5 identity matrix, we can evaluate that $\beta_6^M = -13.4958$. It is clear that the conditional marginal signal β_6^M is very closed to the true β_6^*.

Example 2 The case when there are unimportant variables that are highly marginal correlated with the response. Consider model (1) with the true $\beta^* = (5, 1, 0, \dots, 0)^{\mathrm{T}}$, equi-correlation 0.9 among all covariates except X_2, which is independent of the rest of the covariates. Hence, marginal correlation for all unimportant variables ($\beta_j^{\mathrm{MUC}} = 4.5, j = 3, \dots, p$) are higher than that for important variable X_2 ($\beta_2^{\mathrm{MUC}} = 1$). Marginal screening can fail to recruit X_2. Similarly, assume that $X_{\mathcal{C}} = \{X_1\}$, using CMMC (3) and Condition LC, choosing $\beta_{\mathcal{C}} = 1$, we have the conditional marginal signals $\beta_2^M = 1 = \beta_2^*$ and $\beta_j^M = 0 = \beta_j^*$ for any $j \neq 1, 2$.

The above two examples illustrate a screening procedure based on CMMC(3) can reduce the impact of a strong correlation among the predictors due to centralized variables. Although sometimes there is remarkable difference between β_j^M and the true value of the parameter β_j^* in models (1), Theorem 1 given below shows that $\beta_j^M = \beta_j^*$ under some regularity conditions. Moreover, Remarks below illustrate that the condition used here is weaker than ones used for the case when the set $\mathcal{C}$ is unknown.

Theorem 1 *If the centralized variables, $X_j - \mathbb{E}(X_j|X_{\mathcal{C}}^{\mathrm{T}}\beta_{\mathcal{C}})$ and $X_k - \mathbb{E}(X_k|X_{\mathcal{C}}^{\mathrm{T}}\beta_{\mathcal{C}})$, are uncorrelated, where $j \neq k$, $j \in \mathcal{D}$ and $k \in \mathcal{D} \cap \mathcal{A}$, i.e.,*

$$\mathbb{E}\{[X_j - \mathbb{E}(X_j|X_{\mathcal{C}}^{\mathrm{T}}\beta_{\mathcal{C}})][X_k - \mathbb{E}(X_k|X_{\mathcal{C}}^{\mathrm{T}}\beta_{\mathcal{C}})]\} = 0, \quad j \neq k, j \in \mathcal{D}, k \in \mathcal{D} \cap \mathcal{A}, \tag{5}$$

then

$$\beta_j^M = \beta_j^* \quad \textit{for any } j \in \mathcal{D}.$$

With regard to Theorem 1, we have the following remarks:

Remark 1 Notice that $\mathbb{E}\{[X_j - \mathbb{E}(X_j|X_{\mathcal{C}}^{\mathrm{T}}\beta_{\mathcal{C}})]\mathbb{E}(X_k|X_{\mathcal{C}}^{\mathrm{T}}\beta_{\mathcal{C}})\} = 0$ for any vector $\beta_{\mathcal{C}}$ or matrix $\beta_{\mathcal{C}}$. Condition (5) is equivalent to $\mathbb{E}\{[X_j - \mathbb{E}(X_j|X_{\mathcal{C}}^{\mathrm{T}}\beta_{\mathcal{C}})]X_k\} = 0$ for any $j \neq k$, $j \in \mathcal{D}$ and $k \in \mathcal{D} \cap \mathcal{A}$, implying the centralized variable $X_j - \mathbb{E}(X_j|X_{\mathcal{C}}^{\mathrm{T}}\beta_{\mathcal{C}})$ is completely uncorrelated with X_k for any $j \neq k$, $j \in \mathcal{D}$ and $k \in \mathcal{D} \cap \mathcal{A}$. Even so, this condition does not mean X_j and X_k ($j \neq k$, $j \in \mathcal{D}$, and $k \in \mathcal{D} \cap \mathcal{A}$) are uncorrelated. For example, the variables are generated as $X_j \sim N(0, 1)$ ($j = 1, \dots, p$) with $\mathbb{E}(X_j X_t) = \rho$ ($\rho > 0$), and $\mathbb{E}(X_j X_p) = \rho^{1/2}$ for all $t = 1, \dots, p-1$ and $j \neq t$. The active set is $\mathcal{A} = \{1, 2, 3, 4, 5\}$. Choose the conditional set $X_{\mathcal{C}} = \{X_p\}$ under linearity condition (4); it can be easily verified that the condition (5) holds. Moreover, this condition has no requirement on the relationship among conditional set $X_{\mathcal{C}}$.

Remark 2 If without the information about the index set $\mathcal{C}$, we need to use the marginal moment condition

$$\mathbb{E}\{X_j[Y - X_j\beta_j^{\mathrm{MUC}}]\} = 0, \quad j = 1, \dots, p$$

to construct marginal empirical likelihood. In this case, $\beta_j^{\mathrm{MUC}} = \mathbb{E}(X_j Y)$ is the correlation coefficient between X_j and Y. However, β_j^{MUC} is not equal to β_j^* unless X_j and X_k are uncorrelated for all $j \neq k$ $(j, k = 1, \ldots, p)$ (Chang et al. 2013a). Notice that condition (5) holds clearly in this scenario.

Remark 3 With respect to the number of required conditional equations, if without the conditional set, it is $(p-1)s_{\mathcal{A}}$. However, the number of equations in condition (5) is $(s_{\mathcal{D}}-1)(s_{\mathcal{A}}-s_{\mathcal{CA}})$, where $s_{\mathcal{D}}$ and $s_{\mathcal{CA}}$ are the cardinalities of $\mathcal{D}$ and $\mathcal{C} \cap \mathcal{A}$, respectively. It is clear that even the conditional set only include inactive variables, the number of required conditional equations reduces by $s_{\mathcal{C}} \times s_{\mathcal{A}}$. Moreover, increasing an active variable in the conditional set can reduce $(s_{\mathcal{D}}-1)$ required conditions.

According to the above remarks, Theorem 1 gives a weaker condition to make sure $\beta_j^M = \beta_j^*$. On the other hand, it must be pointed out that conditional set $X_{\mathcal{C}}$ does not necessarily have to contain active variables, which is clearly shown in our proof of Theorem 1 given in the Appendix. Moreover, according to the example in remark 1, we know that condition (5) is weaker than the condition that X_j and X_k are uncorrelated for all $j \neq k$ $(j, k = 1, \ldots, p)$, even the conditional set $X_{\mathcal{C}}$ only includes inactive predictors. In addition, it is clear that condition (5) has no requirement on the relationship among conditional set.

Therefore, our conditional marginal moment condition (CMMC) in (3) can be conveniently used to deal with the case when there exist strong correlations between X_j and X_k $(j, k = 1, \ldots, p,\ j \neq k)$. The condition set $X_{\mathcal{C}}$ can contain some active predictors known in advance and some inactive predictors which are strongly correlated with other predictors. Or, the conditional set $X_{\mathcal{C}}$ can only contain those inactive predictors. Furthermore, we will see in the simulation studies that our method has excellent performance even if some of the conditional variables are inactive, and compared to other original screening procedures, it performs well even all conditional variables are randomly selected inactive variables. These show that our method is convenient and flexible.

Next we will construct a conditional marginal empirical likelihood for linear models as follows. Let (X_i, Y_i) be collected independent data,

$$g_{ij}^{(cl)}(\beta) = [X_{ij} - \mathbb{E}(X_j | X_{i\mathcal{C}}^{\mathrm{T}} \beta_{\mathcal{C}})][Y_i - X_{ij}\beta] \quad (j \in \mathcal{D}),$$

where X_{ij} and $X_{i\mathcal{C}}$ are the ith observations of X_j and $X_{\mathcal{C}}$, respectively. By the CMMC (3), we define the following conditional marginal empirical likelihood (CMEL) as

$$EL_j(\beta) = \sup\left\{\prod_{i=1}^n w_i : w_i \geq 0, \sum_{i=1}^n w_i = 1, \sum_{i=1}^n w_i g_{ij}^{(cl)}(\beta) = 0\right\},$$

for $j \in \mathcal{D}$, and further define the conditional marginal empirical likelihood ratio (CMELR) as

$$l_j(\beta) = 2\sum_{i=1}^n \log\{1 + \lambda g_{ij}^{(cl)}(\beta)\}, \tag{6}$$

where λ is the lagrange multiplier satisfying

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

2.2 Conditional marginal empirical likelihood for generalized linear models

In this part, we will show that the above conditional marginal empirical likelihood approach can be applied in generalized linear models. Assume that the conditional probability density of the random variable Y belongs to an exponential family:

$$f(y|x, \theta) = \exp(y\theta(x) - b(\theta(x)) + c(x; y)),$$

where $b(\cdot)$ and $c(\cdot)$ are specific known functions in the canonical parameter $\theta(x)$ (McCullagh and Nelder 1989). In this case, the mean function $\mu = \mathbb{E}(Y|X) = b'(\theta(X))$. Suppose that the second derivative of $b(\cdot)$ is continuous and positive/negative, the canonical parameter is modeled by a linear function $\theta(X) = X^{\mathrm{T}}\beta^*$. Particularly, for linear models, $\mu = b'(\theta(X)) = \theta(X) = X^{\mathrm{T}}\beta^*$. Let $\mathbb{E}(X_j) = 0$ and $\mathbb{E}(X_j^2 = 1)$ $(j = 1, \ldots, p)$, without loss of generality. Consider sparse models, i.e., the active set $\mathcal{A} = \{k : \beta_k^* \neq 0\}$ is small. Furthermore, assume that a set $X_{\mathcal{C}}$ of variables is known to be related to the response Y. Denote by $X_{\mathcal{D}}$ the set of the rest of variables.

Under generalized linear models, based on the definition of the conditional linear expectation in Barut et al. (2012), the moment condition (2) is equivalent to

$$\mathbb{E}\{[X_j - \mathbb{E}(X_j|X_{\mathcal{C}}^{\mathrm{T}}\beta_{\mathcal{C}})][Y - b'(X_{\mathcal{C}}^{\mathrm{T}}\beta_{\mathcal{C}}^M + X_j\beta_j^M)]\} = 0. \tag{7}$$

In the above notation, we assume that the intercept is incorporated in the vector $X_{\mathcal{C}}$.

The following Lemma 1 reveals that the conditional marginal signal β_j^M is in fact a measurement of the correlation between the centralized variable $X_j - \mathbb{E}(X_j|X_{\mathcal{C}}^{\mathrm{T}}\beta_{\mathcal{C}})$ and the response as in linear models.

Lemma 1 *For any $j \in \mathcal{D}$, the conditional marginal signal $\beta_j^M = 0$ if and only if*

$$\mathbb{E}\{[X_j - \mathbb{E}(X_j|X_{\mathcal{C}}^{\mathrm{T}}\beta_{\mathcal{C}})]Y\} = 0.$$

Note that the conditional marginal signal β_j^M is not equal to the true parameter β_j^* usually. In order to guarantee that the conditional marginal signals provide useful probes for the true parameters, we need to ensure that the conditional marginal signal $|\beta_j^M|$ exceeds a related threshold when the corresponding truth $|\beta_j^*|$ exceeds a certain threshold. This will be shown in the following theorem and condition:

Condition 1 For $j \in \mathcal{D} \cap \mathcal{A}$, there exists $c_1 > 0$ and $\kappa \in [0, \frac{1}{2})$ such that

$$|\mathbb{E}\{[X_j - \mathbb{E}(X_j|X_{\mathcal{C}}^{\mathrm{T}}\beta_{\mathcal{C}})]Y\}| \geq c_1 n^{-\kappa}.$$

For screening feature, the above is an identification condition for $j \in \mathcal{D} \cap \mathcal{A}$ and can be viewed as a requirement for the minimal signal strengthen. It is the same as the first part of condition 1 in Barut et al. (2012).

Theorem 2 *If Condition 1 holds, then there exits* $c_2 > 0$ *such that*

$$\min_{j \in \mathcal{D} \cap \mathcal{A}} |\beta_j^M| \geq c_2 n^{-\kappa},$$

provided that $b''(\cdot)$ *is bounded.*

Theorem 2 implies that there exists a threshold $\gamma > 0$ such that the set $\{j \in \mathcal{D} : |\beta_j^M| > \gamma\}$ must contain the target set $\mathcal{D} \cap \mathcal{A}$. Hence, we can select non-zero conditional marginal signal β_j^M instead of directly selecting $\beta_j^* \neq 0$.

To apply the conditional marginal empirical likelihood approach for generalized linear models, similar to $g_{ij}^{(cl)}(\beta)$ given in the previous subsection, we now define new estimating functions as

$$g_{ij}^{(cg)}(\beta, \tilde{\beta}_{\mathcal{C}}) = [X_{ij} - \mathbb{E}(X_j | X_{i\mathcal{C}}^{\mathrm{T}} \beta_{\mathcal{C}})][Y_i - b'(X_{i\mathcal{C}}^{\mathrm{T}} \tilde{\beta}_{\mathcal{C}} + X_{ij}\beta)], \quad j \in \mathcal{D}.$$

By the same argument used above, for generalized linear model, the conditional marginal empirical likelihood ratio is defined as

$$l_j(\beta, \tilde{\beta}_{\mathcal{C}}) = 2 \sum_{i=1}^{n} \log\{1 + \lambda g_{ij}^{(cg)}(\beta, \tilde{\beta}_{\mathcal{C}})\}, \quad j \in \mathcal{D}, \tag{8}$$

where λ is the lagrange multiplier satisfying

$$0 = \sum_{i=1}^{n} \frac{g_{ij}^{(cg)}(\beta, \tilde{\beta}_{\mathcal{C}})}{1 + \lambda g_{ij}^{(cg)}(\beta, \tilde{\beta}_{\mathcal{C}})}.$$

2.3 CSIS by CMELR

The marginal empirical likelihood ratios (6) and (8) can be viewed as functions of the parameters of interest. Moreover, the theoretical analyses in Chang et al. (2013a) show that the marginal empirical likelihood ratio should not be large when evaluated at the truth value, and the marginal empirical likelihood ratio statistics has high probability to take large value when evaluated at the false values. Hence, in our cases, the conditional marginal empirical likelihood ratios $l_j(0)$ and $l_j(0, \widehat{\beta}_{\mathcal{C}}^M)$ in (6), and (8) should not be large if $\beta_j^M = 0$ and they diverge with large probability if $\beta_j^M \neq 0$, where $\widehat{\beta}_{\mathcal{C}}^M$ can be the maximum marginal likelihood estimator in generalized linear model. That means we can use $l_j(0)$ and $l_j(0, \widehat{\beta}_{\mathcal{C}}^M)$ as devices for feature screening for linear models and generalized linear models, respectively.

However, for generalized linear models, we need to estimate marginal signals $\beta_{\mathcal{C}}^M$ if directly using the conditional marginal empirical likelihood ratio in (8). In order to reduce redundant computation and construct a unified feature screening method for both linear models and generalized linear models, we use a unified conditional marginal moment condition (UMMC) as

$$\mathbb{E}\{[X_j - \mathbb{E}(X_j|X_{\mathcal{C}}^{\mathrm{T}}\beta_{\mathcal{C}})]Y\} - \alpha_j = 0, \tag{9}$$

where α_j is denoted as the correlation coefficient between the centralized variable $X_j - \mathbb{E}(X_j|X_{\mathcal{C}}^{\mathrm{T}}\beta_{\mathcal{C}})$ and the response Y.

As the discussion in the previous subsection, to guarantee that α_j can be used as a tool for recruiting the corresponding index j, we need to ensure that $|\alpha_j|$ exceeds a related threshold when the corresponding truth $|\beta_j^*|$ exceeds a certain threshold. Note that it holds directly in generalized linear models due to Condition 1. The following lemma shows that it also holds for linear models without Condition 1.

Lemma 2 *Suppose that condition in Theorem 1 holds, for any $j \in \mathcal{D}$, the true $\beta_j^* \neq 0$ if and only if $\mathbb{E}\{[X_j - \mathbb{E}(X_j|X_{\mathcal{C}}^{\mathrm{T}}\beta_{\mathcal{C}})]Y\} \neq 0$, i.e., $\alpha_j \neq 0$.*

Hence we can conclude that a unified conditional marginal empirical likelihood with the same estimating functions $g_{ij}^{(c)}(\alpha) = [X_{ij} - \mathbb{E}(X_j|X_{i\mathcal{C}}^{\mathrm{T}}\beta_{\mathcal{C}})]Y_i - \alpha$ can be equally applied for both linear models and generalized linear models, and the corresponding conditional marginal empirical likelihood ratio (CMELR) can be defined as

$$l_j(\alpha) = 2\sum_{i=1}^{n} \log\{1 + \lambda g_{ij}^{(c)}(\alpha)\}, \quad j \in \mathcal{D}, \tag{10}$$

where λ is the lagrange multiplier satisfying

$$0 = \sum_{i=1}^{n} \frac{g_{ij}^{(c)}(\alpha)}{1 + \lambda g_{ij}^{(c)}(\alpha)}.$$

Moreover,we can use $l_j(0)$ defined in (10) as a convenient device for feature screening in both linear models and generalized linear models. More specifically, the feature screening is to keep the variables X_j with indices in

$$\mathcal{D} \cap \mathcal{A}_{\gamma_n} = \{j \in \mathcal{D} : l_j(0) \geq \gamma_n\}, \tag{11}$$

for a given thresholding parameter γ_n.

In this way, we just need to evaluate the conditional marginal empirical likelihood ratio at one point and avoid estimating the marginal signal β_j^M or $\beta_{\mathcal{C}}^M$. Moreover, the feature screening method based on empirical likelihood approach requires less strict distribution assumptions. In next section, we will give theoretical results to show the sure screening properties of the proposed screening procedure.

Notice that in practice, we cannot directly use the likelihood ratio $l_j(0)$ in (11) for screening feature because it contains unknown $\mathbb{E}(X_j|X_{\mathcal{C}}^{\mathrm{T}}\beta_{\mathcal{C}})$ in estimating function $g_{ij}^{(c)}(0)$. To get an estimator $\widehat{l_j}(0)$ of $l_j(0)$, we first need to estimate $\mathbb{E}(X_j|X_{\mathcal{C}}^{\mathrm{T}}\beta_{\mathcal{C}})$.

Assume $\frac{1}{n}\sum_{i=1}^{n} X_{ij} = 0$ and $\frac{1}{n}\sum_{i=1}^{n} X_{ij}^2 = 1$ for any $j \in \{1, \ldots, p\}$, where X_{ij} is the i-th observation of X_j. For simplicity, we here need Condition LC (4). Without it, we can use nonparametric method to construct estimator. We consider the following two cases:

Case 1 $\mathbb{E}(X_{\mathcal{C}} X_{\mathcal{C}}^{\mathrm{T}})$ is a positive definite matrix. Since our theoretical results hold for any vector or matrix $\beta_{\mathcal{C}}$, we then use the following most simple choice:

$$\beta_{\mathcal{C}} = I_{s_{\mathcal{C}}},$$

a $s_{\mathcal{C}} \times s_{\mathcal{C}}$ identity matrix. Then, we estimate $\mathrm{cov}(X_j, X_{\mathcal{C}}^{\mathrm{T}})$ and $\mathbb{E}(X_{\mathcal{C}} X_{\mathcal{C}}^{\mathrm{T}})$, respectively, by

$$\widehat{\mathrm{cov}}(X_j, X_{\mathcal{C}}^{\mathrm{T}}) = \frac{1}{n}\sum_{k=1}^{n} X_{kj} X_{k\mathcal{C}}^{\mathrm{T}}, \quad \widehat{\mathbb{E}}(X_{\mathcal{C}} X_{\mathcal{C}}^{\mathrm{T}}) = \frac{1}{n}\sum_{k=1}^{n} X_{k\mathcal{C}} X_{k\mathcal{C}}^{\mathrm{T}},$$

where $X_{k\mathcal{C}}$ is the k-th observation of $X_{\mathcal{C}}$. Hence we can obtain the estimator as

$$\widehat{\mathbb{E}}(X_j | X_{i\mathcal{C}}^{\mathrm{T}}) = \frac{1}{n}\sum_{k=1}^{n} X_{kj} X_{k\mathcal{C}}^{\mathrm{T}} \left\{ \frac{1}{n}\sum_{k=1}^{n} X_{k\mathcal{C}} X_{k\mathcal{C}}^{\mathrm{T}} \right\}^{-1} X_{i\mathcal{C}}.$$

Particularly, if $\mathbb{E}(X_{\mathcal{C}} X_{\mathcal{C}}^{\mathrm{T}}) = I_{s_{\mathcal{C}}}$, then the above estimator can be rewritten as

$$\widehat{\mathbb{E}}(X_j | X_{i\mathcal{C}}^{\mathrm{T}}) = \frac{1}{n}\sum_{k=1}^{n} X_{kj} X_{k\mathcal{C}}^{\mathrm{T}} X_{i\mathcal{C}}.$$

Case 2 $\mathbb{E}(X_{\mathcal{C}} X_{\mathcal{C}}^{\mathrm{T}})$ is singular or approximately singular. Let $B = \frac{1}{n}\sum_{k=1}^{n} X_{k\mathcal{C}} X_{k\mathcal{C}}^{\mathrm{T}}$ and $\lambda_1 \geq \ldots \geq \lambda_t > 0$ denote the nonzero eigenvalues of B, where $t \leq s_{\mathcal{C}}$. We can easily find a matrix Γ such that $\Gamma^{\mathrm{T}} B \Gamma = \Lambda$, where Λ denotes a diagonal matrix with the diagonal elements $(\lambda_1, \ldots, \lambda_t)$. Note the matrix Γ can consist of the first t columns of Γ_1, where Γ_1 satisfies $\Gamma_1^{\mathrm{T}} B \Gamma_1 = \Lambda_1$, Λ_1 denotes a diagonal matrix whose diagonal elements consist of all eigenvalues of B. Under this situation, we choose

$$\beta_{\mathcal{C}} = \Gamma,$$

and then get the following estimator:

$$\widehat{\mathbb{E}}(X_j | X_{i\mathcal{C}}^{\mathrm{T}} \Gamma) = \frac{1}{n}\sum_{k=1}^{n} X_{kj} X_{k\mathcal{C}}^{\mathrm{T}} \Gamma \Lambda^{-1} \Gamma^{\mathrm{T}} X_{i\mathcal{C}}.$$

For both the cases above, denote $\widehat{g_{ij}^{(c)}}(0) = [X_{ij} - \widehat{\mathbb{E}}(X_j | X_{i\mathcal{C}}^{\mathrm{T}} \beta_{\mathcal{C}})] Y_i$, the estimator of $g_{ij}^{(c)}(0)$. We then obtain the estimated conditional marginal empirical likelihood ratio at zero as

$$\widehat{l_j}(0) = 2\sum_{i=1}^{n} \log\{1 + \hat{\lambda}\widehat{g_{ij}^{(c)}}(0)\},$$

where $\hat{\lambda}$ is the lagrange multiplier satisfying

$$0 = \sum_{i=1}^{n} \frac{\widehat{g_{ij}^{(c)}}(0)}{1 + \hat{\lambda}\widehat{g_{ij}^{(c)}}(0)}.$$

Finally, we select the index set of active variables as

$$\widehat{\mathcal{D} \cap \mathcal{A}}_{\gamma_n} = \{j \in \mathcal{D} : \widehat{l_j}(0) \geq \gamma_n\},$$

where γ_n is a predefined threshold value. This method will be referred to as conditional sure independence screening based on conditional marginal empirical likelihood ratio or CMELR-CSIS for short.

It is important to note that the threshold level γ_n in practice is generally difficult to identify explicitly, because it involves unknown constants. Thus, we choose hard thresholding rule (Fan and Lv 2008) in practice such that $\widehat{\mathcal{D} \cap \mathcal{A}}_{\gamma_n}$ recruits fixed number $d = \lfloor n/\log n \rfloor$ or $d = n$ of candidate variables, where $\lfloor m \rfloor$ denotes the largest integer that is less than or equal to m.

3 Sure screening properties

A useful screening approach is expected to retain all important variables while removing the others, which means the procedure possesses sure screening properties. In this section, we derive the sure screening properties of the proposed screening procedure with respect to the population aspect and sample aspect, simultaneously. Moveover, we give a bound on the size of the selected set of variables. The following lemmas and theorems state the details.

To get the sure screening properties, we assume that the response Y has bounded variance and the following regular condition holds.

Condition 2 There are positive constants K_1, K_2, γ_1 and γ_2 such that

$$\mathbb{P}\{|X_j - \mathbb{E}(X_j|X_{\mathcal{C}}^{\mathrm{T}}\beta_{\mathcal{C}})| > u\} \leq K_1 \exp\{-K_2 u^{\gamma_1}\},$$

for any $j \in \mathcal{D}$ and any $u > 0$, and

$$\mathbb{P}\{|Y| > u\} \leq K_1 \exp\{-K_2 u^{\gamma_2}\},$$

for any $u > 0$.

Condition 2 is required to ensure the large deviation results that are used to get the exponential convergence rate, because we impose no distributional assumptions. This is a regular condition appearing in lots of literature. For instance, the first part of

Condition 2 is similar to the first part of condition D in Fan and Song (2010), Condition (C3) in Zhu et al. (2011), the first part of Condition 2(ii) in Barut et al. (2012), and the first part of A.2 in Chang et al. (2013a); the second part of Condition 2 is same as the second part of A.2 in Chang et al. (2013a). According to the argument in Chang et al. (2013a), the second part of Condition 2 is actually weaker than the second part of Condition D in Fan and Song (2010) and the second part of Condition 2(ii) in Barut et al. (2012).

The following lemma shows that the goal set $\mathcal{D} \cap \mathcal{A}$ can be clearly distinguished by the conditional marginal empirical likelihood ratio (CMELR) valued at zero.

Lemma 3 *Under Conditions* 1 *and* 2*, there exists a positive constant* C_1 *such that, for any* $\tau \in (0, \frac{1}{2} - \kappa)$*,*

$$\max_{j \in \mathcal{D}\cap\mathcal{A}} \mathbb{P}\{l_j(0) < c_1^2 n^{2\tau}\} \\ \leq \begin{cases} \exp\left\{-C_1 n^{(1-2\kappa)\wedge\frac{(1-2\kappa-2\tau)\gamma}{2}}\right\}, & \text{if } (1-2\kappa)(1+2\delta) < 1, \\ \exp\left\{-C_1 n^{\frac{1-\kappa}{1+\delta}\wedge\frac{(1-2\kappa-2\tau)\gamma}{2}}\right\}, & \text{if } (1-2\kappa)(1+2\delta) \geq 1, \end{cases}$$

where C_1 *depends only on* K_1, K_2, γ_1 *and* γ_2 *given in Condition 2,* $\gamma = \frac{\gamma_1\gamma_2}{\gamma_1+\gamma_2}$ *and* $\delta = \max\{\frac{2}{\gamma} - 1, 0\}$.

According to Lemma 3, we can get the sure screening property of our method on population aspect directly. When $l_j(0)$ is replaced by its estimator $\widehat{l_j}(0)$, the next lemma shows that the estimator has the same properties as shown in Lemma 3.

Lemma 4 *Under condition* 1 *and* 2*, if* $\max_i |X_{ik}Y_i| = O_p(n^{\omega})$ *where* $\omega < 1/2 - \kappa$*,* $k \in \mathcal{C}$*, there exists a positive constant* C_2 *such that, for any* $\tau \in (0, \frac{1}{2} - \kappa)$*,*

$$\max_{j \in \mathcal{D}\cap\mathcal{A}} \mathbb{P}\{\widehat{l_j}(0) < c_1^2 n^{2\tau}\} \\ \leq \begin{cases} \exp\left\{-C_2 n^{(1-2\kappa)\wedge\frac{(1-2\kappa-2\tau)\gamma}{2}}\right\}, & \text{if } (1-2\kappa)(1+2\delta) < 1, \\ \exp\left\{-C_2 n^{\frac{1-\kappa}{1+\delta}\wedge\frac{(1-2\kappa-2\tau)\gamma}{2}}\right\}, & \text{if } (1-2\kappa)(1+2\delta) \geq 1, \end{cases}$$

where C_2 *depends only on* K_1, K_2, γ_1 *and* γ_2 *defined in Condition 2,* $\gamma = \frac{\gamma_1\gamma_2}{\gamma_1+\gamma_2}$ *and* $\delta = \max\{\frac{2}{\gamma} - 1, 0\}$.

Due to Lemma 4, we can easily establish the sure screening property of the proposed procedure in sample version. That is shown in the following theorem:

Theorem 3 *Under the conditions in Lemma* 4*, there exists a positive constant* C_2 *such that, for any* $\tau \in (0, \frac{1}{2} - \kappa)$ *and* $\gamma_n = c_1^2 n^{2\tau}$*,*

$$\begin{aligned}&\mathbb{P}\{\mathcal{D}\cap\mathcal{A}\subset\widehat{\mathcal{D}\cap\mathcal{A}}_{\gamma_n}\}\\&\quad\geq\begin{cases}1-s_{\mathcal{D}\mathcal{A}}\exp\left\{-C_2 n^{(1-2\kappa)\wedge\frac{(1-2\kappa-2\tau)\gamma}{2}}\right\}, & \text{if } 1-2\kappa)(1+2\delta)<1,\\ 1-s_{\mathcal{D}\mathcal{A}}\exp\left\{-C_2 n^{\frac{1-\kappa}{1+\delta}\wedge\frac{(1-2\kappa-2\tau)\gamma}{2}}\right\}, & \text{if } 1-2\kappa)(1+2\delta)\geq 1,\end{cases}\end{aligned}$$

where C_2 depends only on K_1, K_2, γ_1 and γ_2 given in Condition 2, $\gamma=\frac{\gamma_1\gamma_2}{\gamma_1+\gamma_2}$, $\delta=\max\{\frac{2}{\gamma}-1,0\}$, and $s_{\mathcal{D}\mathcal{A}}=|\mathcal{D}\cap\mathcal{A}|$, the size of the set of non-sparse elements.

Based on Theorem 3, we know that our proposed screening procedure can handle the dimensionality of order

$$\log q=\begin{cases}o\left(n^{(1-2\kappa)\wedge\frac{(1-2\kappa-2\tau)\gamma}{2}}\right), & \text{if } (1-2\kappa)(1+2\delta)<1,\\ o\left(n^{\frac{1-\kappa}{1+\delta}\wedge\frac{(1-2\kappa-2\tau)\gamma}{2}}\right), & \text{if } (1-2\kappa)(1+2\delta)\geq 1,\end{cases}$$

where $q=s_{\mathcal{D}}$ is the size of $X_{\mathcal{D}}$. It is very similar to that in Chang et al. (2013a) with the number p of all predictors being replaced by the number q of unknown predictors in the set $X_{\mathcal{D}}$. However, as the result in Chang et al. (2013a), our result is also weaker than that in Fan and Lv (2008) as a price paid for allowing more general error distribution and is a stronger result than those in Fan and Song (2010) and Barut et al. (2012) in a certain setting, see the details in Chang et al. (2013a).

We have already stated the sure screening properties of our proposed procedure (CMELR-CSIS) in population and sample level. However, a good screening procedure does not only possess sure screening, but also retains a small set of variables after thresholding. For population level, according to the argument in Chang et al. (2013a), we can directly obtain that with large probability, the size of $\mathcal{D}\cap\mathcal{A}_{\gamma_n}$ in (11) is not larger than the number of the true contributing explanatory variables. Now we investigate how large the set $\widehat{\mathcal{D}\cap\mathcal{A}}_{\gamma_n}$ is. We can notice that

$$\begin{aligned}|\widehat{\mathcal{D}\cap\mathcal{A}}_{\gamma_n}|&=\sum_{j\in\mathcal{D}\cap\mathcal{A}} I\{\widehat{l_j}(0)\geq c_1^2 n^{2\tau}\}+\sum_{j\notin\mathcal{D}\cap\mathcal{A}} I\{\widehat{l_j}(0)\geq c_1^2 n^{2\tau}\}\\&\leq s_{\mathcal{D}\mathcal{A}}+\sum_{j\notin\mathcal{D}\cap\mathcal{A}} I\{\widehat{l_j}(0)\geq c_1^2 n^{2\tau}\},\end{aligned}$$

where $s_{\mathcal{D}\mathcal{A}}=|\mathcal{D}\cap\mathcal{A}|$, the size of the set of non-sparse elements. Then

$$\mathbb{P}\{|\widehat{\mathcal{D}\cap\mathcal{A}}_{\gamma_n}|>s_{\mathcal{D}\mathcal{A}}\}\leq\sum_{j\notin\mathcal{D}\cap\mathcal{A}}\mathbb{P}\{\widehat{l_j}(0)\geq c_1^2 n^{2\tau}\}.$$

Hence, we need to know the magnitudes of $\widehat{l_j}(0)$ for $j\notin\mathcal{D}\cap\mathcal{A}$.

Lemma 5 *Under condition* 1 *and* 2*, if* $\max_{j\notin\mathcal{D}\cap\mathcal{A}}|\mathbb{E}\{[X_j-\mathbb{E}(X_j|X_{\mathcal{C}}^{\mathrm{T}}\beta_{\mathcal{C}})]Y\}|=O(n^{-\eta})$ *where* $\eta>\kappa$ *and* $\min_{j\notin\mathcal{D}\cap\mathcal{A}}\mathbb{E}\{[X_j-\mathbb{E}(X_j|X_{\mathcal{C}}^{\mathrm{T}}\beta_{\mathcal{C}})]^2Y^2\}\geq c_3$ *for some* $c_3>0$*, there exists a positive constant* C_3 *such that, for any* $j\notin\mathcal{D}\cap\mathcal{A}$ *and any* $\tau\in((\frac{1}{2}-\eta)\vee\omega,\frac{1}{2}-\kappa)$,

$$\begin{aligned}&\mathbb{P}\{\widehat{l_j}(0) \geq c_1^2 n^{2\tau}\}\\ &\leq \begin{cases}\exp(-C_3 n^{2\tau}) + \exp(-C_3 n^{\frac{\gamma}{6}}), & \text{if } \gamma < 2 \ \textit{and} \ (\eta \wedge \frac{1-2\omega}{2}) > \frac{1}{4};\\ \exp(-C_3 n^{\gamma(\eta\wedge\frac{1-2\omega}{2})}) + \exp(-C_3 n^{\frac{\gamma}{6}}), & \text{if } \gamma < 2 \ \textit{and} \ (\eta \wedge \frac{1-2\omega}{2}) \leq \frac{1}{4};\\ \exp(-C_3 n^{\gamma(\eta\wedge\frac{1-2\omega}{2})}) + \exp(-C_3 n^{\frac{\gamma}{6}}), & \text{if } \gamma \geq 2 \ \textit{and} \ (\eta \wedge \frac{1-2\omega}{2}) \leq \frac{1}{\gamma+2};\\ \exp(-C_3 n^{\frac{\gamma}{\gamma+2}}) + \exp(-C_3 n^{2\tau}), & \text{if } \gamma \geq 4 \ \textit{and} \ (\eta \wedge \frac{1-2\omega}{2}) > \frac{1}{\gamma+2};\\ \exp(-C_3 n^{\frac{\gamma}{6}}) + \exp(-C_3 n^{2\tau}), & \text{if } 2 \leq \gamma < 4 \ \textit{and} \ (\eta \wedge \frac{1-2\omega}{2}) > \frac{1}{\gamma+2},\end{cases}\end{aligned}$$

where ω satisfies $\max_i |X_{ik}Y_i| = O_p(n^{\omega})$ and $\omega \in [0, \frac{1}{2})$, $k \in \mathcal{C}$, C_3 depends only on K_1, K_2, γ_1, and γ_2 given in Condition 2, $\gamma = \frac{\gamma_1\gamma_2}{\gamma_1+\gamma_2}$.

Then following by the argument between Theorem 3 and Lemma 5, we can obtain the following theorem:

Theorem 4 *Under the conditions in Lemma 5, there exists a positive constant C_3 such that, for any $\tau \in ((\frac{1}{2} - \eta) \vee \omega, \frac{1}{2} - \kappa)$ and $\gamma_n = c_1^2 n^{2\tau}$,*

$$\begin{aligned}&\mathbb{P}\{|\widehat{\mathcal{D} \cap \mathcal{A}_{\gamma_n}}| > s_{\mathcal{D}\mathcal{A}}\}\\ &\leq \begin{cases}q\exp(-C_3 n^{2\tau}) + q\exp(-C_3 n^{\frac{\gamma}{6}}), & \text{if } \gamma < 2 \ \textit{and} \ (\eta \wedge \frac{1-2\omega}{2}) > \frac{1}{4};\\ q\exp(-C_3 n^{\gamma(\eta\wedge\frac{1-2\omega}{2})}) + q\exp(-C_3 n^{\frac{\gamma}{6}}), & \text{if } \gamma < 2 \ \textit{and} \ (\eta \wedge \frac{1-2\omega}{2}) \leq \frac{1}{4};\\ q\exp(-C_3 n^{\gamma(\eta\wedge\frac{1-2\omega}{2})}) + q\exp(-C_3 n^{\frac{\gamma}{6}}), & \text{if } \gamma \geq 2 \ \textit{and} \ (\eta \wedge \frac{1-2\omega}{2}) \leq \frac{1}{\gamma+2};\\ q\exp(-C_3 n^{\frac{\gamma}{\gamma+2}}) + q\exp(-C_3 n^{2\tau}), & \text{if } \gamma \geq 4 \ \textit{and} \ (\eta \wedge \frac{1-2\omega}{2}) > \frac{1}{\gamma+2};\\ q\exp(-C_3 n^{\frac{\gamma}{6}}) + q\exp(-C_3 n^{2\tau}), & \text{if } 2 \leq \gamma < 4 \ \textit{and} \ (\eta \wedge \frac{1-2\omega}{2}) > \frac{1}{\gamma+2},\end{cases}\end{aligned}$$

where q is the size of $X_{\mathcal{D}}$, ω satisfies $\max_i |X_{ik}Y_i| = O_p(n^{\omega})$ and $\omega \in [0, \frac{1}{2})$, $k \in \mathcal{C}$, C_3 depends only on K_1, K_2, γ_1 and γ_2 given in Condition 2, $\gamma = \frac{\gamma_1\gamma_2}{\gamma_1+\gamma_2}$.

Theorem 4 shows that our proposed procedure in sample level also has a very good control of the size of the selected set of variables. As shown later in our simulation results, our proposed procedure (CMELR-CSIS) performs very well.

4 Simulation studies

In this part, we use several simulation examples to demonstrate the performances of our proposed screening procedure (CMELR-CSIS). We compare it with some competitors, such as empirical likelihood-based screening procedure (EL-SIS) proposed in Chang et al. (2013a), conditional sure independence screening (CSIS) proposed by Barut et al. (2012) and sure independence screening (SIS) in Fan and Lv (2008) for linear models or the GLM-SIS in Fan and Song (2010) for generalized linear models. Meanwhile, we compare our method with the corresponding iterative version of the above competitors (denote by EL-ISIS, ISIS, and GLM-ISIS, respectively).

To evaluate the performance of the CMELR-CSIS, when comparing our proposed procedure to the original version of the competitors, we focus on the accuracy of ranking the predictors without thresholding. Our reports include the median minimum model size (MMMS) of the selected models as well as its RSD (in parentheses) over 200 repetitions, where the RSD is defined as the associated interquartile range of minimum model size divided by 1.34 across 200 simulations. The minimum model size (MMS) of the selected models are required for each method to have a sure screening, that is, to contain the true model. The MMMS is used as a measure of the effectiveness of a screening method and avoids choosing the threshold parameter. On the other hand, we compare the conditional screening procedures to the iterative algorithms on the complexity of the procedures and the accuracy of feature screening when applying a proposed thresholding rule, where the proposed threshoding is used to control the model size of the selected models in conditional screenings and the first step of iterative algorithms. Since feature screening procedure generally serves as a preliminary massive reduction step, and is often followed by a conventional variable selection for further refinement, feature screening is more concerned with recruiting all the truly important predictors. Furthermore, the conditional screening approaches are non-iterative algorithms which have much less computational cost. Therefore, we record the proportion that all active variables are correctly recruited in 200 repetitions and the mean computing time (minute) for one repetition (in parentheses) when the hard thresholding is $d = \lfloor n/2 \rfloor$, where $\lfloor m \rfloor$ denotes the largest integer that is less than or equal to m, n is the sample size.

In the simulation studies, we vary the sample size from 200 to 400 for different scenarios and the number of predictors range from 2000 to 10,000. Example 1 and 2 show that for linear models, the conditional screening procedures have excellent performance whether there exist hidden important variables or the unimportant variables are highly marginal correlated with the response. The CMELR-CSIS gives better results in heteroscedastic models which are shown in Example 3. Example 4 shows that the CMELR-CSIS performs better than EL-SIS and CSIS simultaneously when the heteroscedastic models exist hidden important variables or unimportant variables that are highly marginal correlated with the response. The results of the mentioned screening procedure for generalized models are reported in Example 5. The robustness of the CMELR-CSIS to the conditional set is demonstrated in Example 6. The last example shows that the more active variables the conditional set includes, the better performance the CMELR-CSIS has. When there is no information on conditional set, an effective method for constructing CMELR-CSIS is provided in Example 7.

Example 1 The goal of this example is to demonstrate that conditional screening procedures (CMELR-CSIS and CSIS) can make it possible to recover the hidden important explanatory variables. Similar to the first example for linear models mentioned in the introduction of Barut et al. (2012), we consider that variables are generated as $X_j \sim N(0, 1)$ and $\mathrm{cov}(X_i, X_j) = 0.9$ for all $i, j = 1, \ldots, p$ and $i \neq j$, the response is generated as

$$Y = X_1 + 2X_2 + 3X_3 + 4X_4 + 5X_5 - 13.5X_6 + \varepsilon,$$

Table 1 Simulation results for Example 1

(n, p)	SIS	EL-SIS	CSIS	CMELR-CSIS
(400, 5000)	1256.5 (1192.9)	1294.5 (1233.6)	1 (0)	1 (0)
(400, 10,000)	2538 (2626.5)	2558.5 (2804.9)	1 (0)	1 (0)
(200, 5000)	2280 (1356.3)	2274.5 (1366.8)	1 (0)	1 (0)
	ISIS	EL-ISIS	CSIS	CMELR-CSIS
(400, 5000)	1 (0.82)	1 (303.5)	1 (0.019)	1 (0.015)

with $\varepsilon \sim N(0, 1)$ being independent of explanatory variables. Here X_6 is a hidden important variable because it is marginally uncorrelated with the response Y. For the conditional screening, the conditional set is chosen as $X_{\mathcal{C}} = \{X_1, X_2, X_3, X_4, X_5\}$. Simulation results over 200 repetitions with the number of variables $p = 5000$, 10,000 and the size of random samples $n = 200, 400$ are reported in Table 1. It shows that the SIS and the EL-SIS perform poorly when there exist such a hidden important explanatory variable in model; however, as the iterative procedures of the SIS and EL-SIS (namely ISIS and EL-ISIS, respectively), our CMELR-CSIS and CSIS (proposed in Barut et al. 2012) have excellent performances, and moreover, have less much computational cost.

Example 2 The goal of this example is to illustrate that the conditional screening procedures (CMELR-CSIS and CSIS) have much better performance than the unconditional screening (SIS and EL-SIS) when there are inactive variables that are highly marginal correlated with the response. To see this, we consider the second setting in Barut et al. (2012), where the variables are generated as $X_j \sim N(0, 1)$ $(j = 1, \ldots, p)$ with equal correlation 0.9 among all covariates except X_2, which is independent of the rest of the covariates. The response is generated as

$$Y = 5X_1 + X_2 + \varepsilon,$$

where $\varepsilon \sim N(0, 1)$. In this setting, the marginal correlation between all the unimportant variables and the response is 4.5, which is higher than 1, the marginal correlation between the important variable X_2 and the response. We consider $p = 5000$, 10,000 and vary the random sample size from $n = 200$ to 400. Results over 200 repetitions are listed in Table 2. It reveals that the SIS and the EL-SIS almost break down in this scenario but the iterative algorithms (ISIS and EL-ISIS) and the conditional screening procedures (CMELR-CSIS and CSIS) still work well. In this case, the iterative algorithms also consume more time than conditional screening. It is worth noting that in this situation the CMELR-CSIS is more robust to the sample size than CSIS based on the simulation results.

Example 3 The previous two examples show that our proposed procedure CMELR-CSIS performs as well as the CSIS. Since the empirical likelihood approach requires a less restrictive distributional assumption, as pointed by Chang et al. (2013a), we

Table 2 Simulation results for Example 2

(n, p)	SIS	EL-SIS	CSIS	CMELR-CSIS
(400, 5000)	3765.5 (3350)	4998 (11.9)	1 (1)	1 (0)
(400, 10,000)	7544 (7294)	9998 (13.4)	295 (65.3)	1 (0)
(200, 5000)	4317.5 (3576.5)	4999 (44.8)	202 (305.2)	1 (0)
	ISIS	EL-ISIS	CSIS	CMELR-CSIS
(400, 5000)	1 (0.67)	1 (44.38)	1 (0.013)	1 (0.017)

Table 3 Simulation results for Example 3

n	c	SIS	EL-SIS	CSIS	CMELR-CSIS
300	1.5	1252 (517.16)	329 (471.26)	1160 (571.27)	208.5 (417.16)
	2.5	527.5 (522.01)	19.5 (66.04)	420 (428.73)	11 (32.84)
	3.5	265 (518.65)	3 (9.33)	198 (472.76)	2 (4.10)
200	5	791 (584.32)	85 (175.74)	652.5 (622.76)	50 (117.91)
	7	311.5 (480.22)	12 (34.33)	233.5 (373.88)	6 (24.25)
	9	122.5 (301.49)	3 (6.71)	81.5 (217.16)	2 (2.23)

use a heteroscedastic example to demonstrate the advantage of the newly procedure. Consider the case when variables are generated as $X_j \sim N(0, 1)$ with $\text{cov}(X_i, X_j) = 0$ for $i \neq j$ and 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 $\varepsilon \sim N(0, 1)$, where $c > 0$ controls the signal level. We consider the first predictor known in advance for conditional screening. Results over 200 repetitions with $p = 2000$ and $n = 200, 300$ are reported in Table 3 for three different values of c. Based on the results in Table 3, we find that all the screening feature methods are affected by the heteroscedasticity, especially when the signal level is low. However, the CMELR-CSIS and the EL-SIS need smaller model size to have all the relevant variables in each setting.

Example 4 This example is to show that comparing to EL-SIS and CSIS, our screening feature approach has an obvious advantage when the heteroscedastic models have hidden important variables or unimportant variables that are highly marginal correlated with the response. First, consider the following heteroscedastic model,

$$Y = c(X_1 + 2X_2 - 2.7X_3) + \varepsilon/(X_1^2 + X_2^2 + X_3^2),$$

where variables are generated as $X_j \sim N(0, 1)$ and $\text{cov}(X_i, X_j) = 0.9$ for all $i, j = 1, \ldots, p$ and $i \neq j$, and $\varepsilon \sim N(0, 1)$. It entails X_3 is the hidden important variable. The conditional set consists of the first two predictors for conditional screening. The second model is chosen as

Table 4 Simulation results for Example 4

Heteroscedastic model exist hidden important variable					
n	c	SIS	EL-SIS	CSIS	CMELR-CSIS
200	2.5	1518.5 (566.4)	1133 (825.8)	185.5 (397.8)	21.5 (70.5)
	3	1333 (644.4)	892.5 (784.3)	118 (341.1)	3(12 (51.12))
	3.5	1097.5 (624.6)	506 (664.2)	77 (226.1)	3 (30.22)
		ISIS	EL-ISIS	CSIS	CMELR-CSIS
200	3.5	0.66 (0.045)	0.79 (62.66)	0.925 (0.0049)	0.975 (0.0044)
Heteroscedastic model exist unimportant variables that are highly marginal correlated with the response					
n	c	SIS	EL-SIS	CSIS	CMELR-CSIS
200	0.5	1345.5 (870.5)	676 (938.8)	1699 (252.2)	208 (622.8)
	0.8	1300 (1046.6)	612 (1035.1)	1389 (39.7)	16 (124.3)
	1	1203 (1079.5)	594.5 (861.6)	1338.5 (470.1)	4 (33.6)
		ISIS	EL-ISIS	CSIS	CMELR-CSIS
200	1	0.015 (0.073)	0.58 (62.14)	0.405 (0.0046)	0.975 (0.0045)

$$Y = c(5X_1 + X_2) + \varepsilon/(X_1^2 + X_2^2),$$

where variables and model error ε are generated just like in Example 2. The first predictor is known in advance for conditional screening method. Table 4 records the results over 200 repetitions with $p = 2000$ and $n = 200$. Since there exist high correlated between predictors in the above two models, we also consider the results given by the iterative algorithms. Following the results in Table 4, it is clear that our screening approach CMELR-CSIS gets better results than EL-SIS and CSIS simultaneously. Notice that EL-ISIS performs better than another iterative algorithm ISIS in this example and CSIS in second model, but the proposed screening procedure CMELR-CSIS has better behavior on computational cost and quality of screening result. These ensure that the proposed screening procedure CMELR-CSIS is flexible.

Example 5 In this example, we consider the performances of the mentioned methods in the cases with a binary response via logistic regressions. The conditional distribution of the response Y given $X = x$ is binomial distribution with probability of success $\mathbb{P}(x) = \exp(x\beta^*)(1 + \exp(x\beta^*))^{-1}$. We generate covariates just like Example 1 and Example 2. In this case, we get the results over 200 repetitions with $n = 200, 400$ and $p = 2000$. The details reported in Table 5 show CMELR-CSIS puts up a good show as in the linear models. Notice that due to the model complexity and the small coefficient of hidden important variable, under the covariate conditions of Example 2, the conditional and iterative screenings perform poorer than linear models and another generalized linear model, excepting the newly proposed approach CMELR-CSIS.

Table 5 Simulation results for Example 5

Example 1 for logistic model				
n	GLM-SIS	EL-SIS	CSIS	CMELR-CSIS
400	963 (744.02)	962 (822.38)	1 (0)	1 (0)
200	1253.5 (698.88)	1333.5 (684.32)	1 (0)	1 (0)
	GLM-ISIS	EL-ISIS	CSIS	CMELR-CSIS
400	1 (0.99)	1 (97.33)	1 (0.31)	1 (0.0064)
Example 2 for logistic model				
n	GLM-SIS	EL-SIS	CSIS	CMELR-CSIS
400	1999 (0)	1999 (0)	19 (41.79)	1 (0.75)
200	1999 (0)	1999 (0)	300 (362.68)	10.5 (75)
	GLM-ISIS	EL-ISIS	CSIS	CMELR-CSIS
400	0.49 (0.21)	0.5 (13.46)	0.67 (0.44)	0.985 (0.0064)

Example 6 We evaluate the performance of the CEMLR-CSIS under three different types of conditional sets to check its robustness to the choice of the conditional set in this example. The first type conditional set consists of only active predictors, the second type includes both active and inactive predictors, and the last one randomly chooses inactive predictors. For a comprehensive comparison, we consider different correlation structures within a large number of correlated covariates. Similar to Barut et al. (2012) and Fan and Song (2010), the variables are generated as

$$X_j = \frac{\varepsilon_j + a_j \varepsilon}{\sqrt{1 + a_j^2}},$$

where ε and $\{\varepsilon_j\}_{j=1}^{[\frac{2p}{3}]}$ are i.i.d standard normal random variables, $\{\varepsilon_j\}_{j=[\frac{2p}{3}]+1}^{p}$ are i.i.d standard laplace random variables. The constants $a_1 = \cdots = a_{[\frac{2p}{3}]}$ are chosen such that the correlation $\rho = \text{corr}(X_i, X_j) = 0, 0.4, 0.8$ among the first $[\frac{2p}{3}]$ variables and the rest constants equal zero. We fix true regression coefficients $\beta^* = \{1, 2, 1, 2, 0, \ldots, 0, 1, 2\}^{\mathrm{T}}$ which have six non-zero parameters.

We consider the following conditional sets: $\mathcal{C}_1 = \{1, 2, 3, p\}, \mathcal{C}_2 = \{1, 2, 5, [\frac{2p}{3}] + 1\}$ and $\mathcal{C}_3$ consists of 3 randomly selected variables for the first $[\frac{2p}{3}]$ predictors which are correlated and one randomly selected inactive predictors from the rest. We can notice that X_1, X_2, X_3, and X_p are active variables, X_5 and $X_{[\frac{2p}{3}]+1}$ are inactive variables. In this example, we consider the number of predictors $p = 10{,}000/5000$ for linear model and $p = 2000$ for logistic model, respectively. Then we get simulation results over 200 repetitions with $n = 400$ for linear models in Table 6 and for logistic models in Table 7.

Table 6 Simulation results for Example 6 for linear model

p	ρ		0.00	0.40	0.80
10,000	SIS		6 (1.49)	6668 (1.49)	6671 (11.38)
	EL-SIS		6 (2.43)	6668 (2.24)	6673 (25.93)
	CSIS	$\mathcal{C}_1$	2 (0)	2 (0)	2 (3.17)
		$\mathcal{C}_2$	4 (0)	18.5 (83.58)	551.5 (743.09)
		$\mathcal{C}_3$	6 (2.43)	3904 (2451.31)	6231 (572.20)
	CMELR-CSIS	$\mathcal{C}_1$	2 (0)	2 (0)	2 (0)
		$\mathcal{C}_2$	4 (0)	6 (15.30)	10.5 (43.28)
		$\mathcal{C}_3$	6 (3.17)	1811.5 (2972.2)	703 (1416.42)
5000	SIS		6 (0.15)	3335 (1.49)	3336 (9.89)
	EL-SIS		6 (1.49)	3335 (1.68)	3338.5 (9.70)
	CSIS	$\mathcal{C}_1$	2 (0)	2 (0)	2.5 (3.17)
		$\mathcal{C}_2$	4 (0)	15 (52.05)	362.5 (480.60)
		$\mathcal{C}_3$	6 (0.75)	2206 (1211.57)	3135 (263.25)
	CMELR-CSIS	$\mathcal{C}_1$	2 (0)	2 (0)	2 (0)
		$\mathcal{C}_2$	4 (0)	5 (7.46)	7 (13.25)
		$\mathcal{C}_3$	6 (1.49)	1118 (1426.12)	260 (497.95)
5000	ISIS		1 (0.070)	1 (0.22)	1 (0.17)
	EL-ISIS		1 (198.4)	1 (245.6)	1 (194.8)
	CSIS	$\mathcal{C}_1$	1 (0.014)	1 (0.013)	1 (0.066)
		$\mathcal{C}_2$	1 (0.017)	0.85 (0.016)	0.405 (0.022)
		$\mathcal{C}_3$	1 (0.016)	0.12 (0.016)	0.02 (0.016)
	CMELR-CSIS	$\mathcal{C}_1$	1 (0.015)	1 (0.014)	1 (0.042)
		$\mathcal{C}_2$	1 (0.016)	0.985 (0.015)	0.98 (0.020)
		$\mathcal{C}_3$	1 (0.016)	0.38 (0.015)	0.365 (0.015)

The simulation results in the two tables show that, compared with the original version of the competitors, CMELR-CSIS has excellent performance even when the conditional set $X_{\mathcal{C}}$ only includes inactive variables. In the case of high correlation among variables, unconditional screening procedures are close to collapse, but CMELR-CSIS still works well when the conditional set contains active variables. For the worst cases, CMELR-CSIS reduces the minimum model size approximatively by two-thirds, which means CMELR-CSIS performs well even all conditional variables are randomly selected inactive variables. It might be because in the case of high correlation among variables, CMELR-CSIS can reduce the impact of a strong correlation among the predictors due to the use of centralized variables, even when the conditional set includes only inactive variables. Under the assumption $\mathbb{E}(X_j) = 0$ and $\mathbb{E}(X_j^2) = 1$, the concern of the unconditional screening feature procedures is $\mathrm{cov}(X_j, X_k)$, while according to Theorem 1, the concern of CMELR-CSIS becomes $\mathrm{cov}(X_j - \mathbb{E}(X_j|X_{\mathcal{C}}^T\beta_{\mathcal{C}}), X_k - \mathbb{E}(X_k|X_{\mathcal{C}}^T\beta_{\mathcal{C}}))$, for any $j, k \in \{1, \ldots, p\}$ and $j \neq k$. In this example, when $\mathrm{cov}(X_j, X_k) = 0.4, 0.8$, it is clear that $\mathrm{cov}(X_j - \mathbb{E}(X_j|X_{\mathcal{C}}^T\beta_{\mathcal{C}}), X_k - \mathbb{E}(X_k|X_{\mathcal{C}}^T\beta_{\mathcal{C}})) = 0.1333,\ 0.0615$, respectively.

Table 7 Simulation results for Example 6 for logistic model

ρ	0.00	0.40	0.80
GLM-SIS	7 (3.92)	1335 (2.99)	1336.5 (6.53)
EL-SIS	7.5 (4.48)	1335 (3.73)	1337 (9.33)
CSIS			
$\mathcal{C}_1$	2 (0)	5 (8.2)	44 (112.5)
$\mathcal{C}_2$	4 (0.75)	43.5 (95.15)	309.5 (382.65)
$\mathcal{C}_3$	7 (3.73)	772.5 (350.18)	1083.5 (289.74)
CMELR-CSIS			
$\mathcal{C}_1$	2 (0)	2 (1.49)	3 (5.41)
$\mathcal{C}_2$	4 (0.75)	17 (42.53)	120 (214.18)
$\mathcal{C}_3$	7 (4.48)	490 (391.04)	382 (385.26)
GLM-ISIS	0.97 (0.21)	0.79 (0.37)	0.30 (0.26)
EL-ISIS	0.97 (370.8)	0.80 (407.7)	0.31 (406.9)
CSIS			
$\mathcal{C}_1$	1 (0.35)	0.98 (0.41)	0.88 (0.36)
$\mathcal{C}_2$	1 (0.36)	0.80 (0.42)	0.46 (0.38)
$\mathcal{C}_3$	0.98 (0.34)	0.13 (0.36)	0.01 (0.33)
CMELR-CSIS			
$\mathcal{C}_1$	1 (0.0074)	0.97 (0.0067)	0.96 (0.0069)
$\mathcal{C}_2$	1 (0.0073)	0.87 (0.0069)	0.84 (0.0068)
$\mathcal{C}_3$	0.98 (0.0072)	0.30 (0.0067)	0.39 (0.0065)

The covariance among the centralized variables is clearly weaker than the covariance among the variables. Comparing to the iterative procedures, if some of the conditional variables are active, the CMELR-CSIS has similar nice screening results in linear models, and it performs better when generalized linear model has a large number of high correlated covariates. Moreover, the CMELR-CSIS has less much computational cost.

Example 7 This simulation example consists of two parts. The first part is to confirm that the more active variables the conditional set has, the better performance the proposed approach (CMELR-CSIS) gives. The second part is to give a solution, a two-stage method, to construct our proposed conditional screening (CMELR-CSIS) when there is no information on conditional set.

First, we consider the models as in Example 6 with $n = 400$ and $p = 5000$. We use $C0-C4$ to denote that there are 0–4 true active predictors, respectively, in conditional set. The simulation results in Table 8 demonstrate that when the conditional set includes active variables, the proposed approach (CMELR-CSIS) performs better, even it only has one active variable.

As shown in the previous sections, usually one can use prior knowledge and experience to choose a suitable conditional set. However, even if there is no information about the set of active variables in advance, the following two-stage procedure can be used

Table 8 Simulation results for the first part of Example 7

Linear models			
ρ	0.20	0.40	0.80
SIS	2524 (1049.07)	3335 (1.49)	3337 (8.58)
EL-SIS	2591 (1094.59)	3335 (1.68)	3336.5 (10.45)
CMELR-CSIS			
C0	272 (790.11)	845 (1185.82)	234.5 (392.54)
C1	45.5 (165.86)	216.5 (560.82)	46.5 (182.46)
C2	4 (1.49)	4 (8.77)	10 (21.08)
C3	3 (0)	3 (0)	5 (10.63)
C4	2 (0)	2 (0)	2 (0)
Generalized linear models			
ρ	0.20	0.40	0.80
GLM-SIS	1056.5 (398.32)	1336 (3.17)	1338 (13.43)
EL-SIS	1060 (376.31)	1336 (3.17)	1339 (16.79)
CMELR-CSIS			
C0	174.5 (256.53)	494.5 (462.5)	400.5 (449.44)
C1	94 (185.82)	174.5 (336.19))	202 (255.60)
C2	13 (26.31)	34.5 (108.58)	107 (238.43)
C3	3 (0.75)	3 (2.99)	5 (14.37)
C4	2 (0)	2 (0)	3 (8.58)

for searching conditional set and then constructing CMELR-CSIS. In the first stage, an unconditional screening is employed to determine some active variable. By the use of these selected variable as condition, our method can be used in the second stage to construct CMELR-CSIS. Since the two screenings have less computational cost, the resulting two-stage method has less computational cost as well. Table 9 reports the results of unconditional screening procedures (SIS, GLM-SIS, EL-SIS, ISIS, GLM-ISIS, and EL-ISIS) and CMELR-CSIS in which the conditional set consists of the first $d = 4$ largest ranked variables in SIS/GLM-SIS or EL-SIS. We find the two-stage method works well and this method can be viewed as a good way to choose a nice conditional set when we do not have any information.

5 Discussion

In this paper, we propose a new screening procedure, the CMELR-CSIS, for ultrahigh-dimensional models. The CMELR-CSIS is a unified feature screening method; it can be equally applied in both linear models and generalized linear models. We use centralized variables to reduce the impact of the correlation among the predictors. On the other hand, the proposed approach can get nice results with less restrictive distribution assumptions, which inherits the merits of empirical likelihood approach. Moreover, the new screening procedure has a high computational efficiently, because it only

Table 9 Simulation results for the second part of Example 7

Linear models			
ρ	0.20	0.40	0.80
SIS	2761 (855.78)	3335 (1.49)	3336.5 (8.40)
EL-SIS	2780 (716.98)	3335 (2.99)	3337.5 (12.13)
SIS + CMELR-CSIS	2 (0)	3 (0.75)	16.5 (56.72)
EL-SIS + CMELR-CSIS	3 (0.75)	4 (0.75)	32 (54.10)
ISIS	1 (0.21)	1 (0.22)	1 (0.17)
EL-ISIS	1 (204.6)	1 (240.5)	1 (196.7)
SIS + CMELR-CSIS	1 (0.11)	1 (0.12)	0.905 (0.11)
EL-SIS + CMELR-CSIS	1 (0.36)	1 (0.42)	0.86 (0.54)
Generalized linear models			
ρ	0.20	0.40	0.80
GLM-SIS	1106.5 (405.78)	1336 (7.65)	1336.5 (6.53)
EL-SIS	1120 (341.04)	1336 (8.96)	1337 (9.33)
GLM-SIS + CMELR-CSIS	2 (1.49)	9.5 (26.68)	593 (786.94)
EL-SIS + CMELR-CSIS	3 (2.24)	7.5 (18.28)	245 (396.45)
GLM-ISIS	1 (0.24)	0.77 (0.35)	0.30 (0.26)
EL-ISIS	1 (401.5)	0.76 (406.4)	0.31 (405.8)
SIS + CMELR-CSIS	1 (0.25)	0.96 (0.24)	0.30 (0.24)
EL-SIS + CMELR-CSIS	1 (0.15)	0.96 (0.18)	0.45 (0.23)

needs to evaluate the conditional marginal empirical likelihood ratio at zero, without iterative algorithm or estimating marginal signals. Our theoretical results show that the CMELR-CSIS has sure screening properties, and simulation studies demonstrate that it gets satisfactory results under all simulation cases. Extending CMELR-CSIS method to semiparametric models or general models is beyond the scope of the current paper and is an interesting topic for future research.

6 Proofs

Proof of Theorem 1 First, we can notice that Eq. (5) is equivalent to

$$\mathbb{E}\{X_j X_k\} = \mathbb{E}\{X_k \mathbb{E}(X_k | X_{\mathcal{C}}^T \beta_{\mathcal{C}})\},$$

for any $j \neq k$, $j \in \mathcal{D}$ and $k \in \mathcal{D} \cap \mathcal{A}$, because

$$\mathbb{E}\{[X_j - \mathbb{E}(X_j | X_{\mathcal{C}}^T \beta_{\mathcal{C}})]\mathbb{E}(X_k | X_{\mathcal{C}}^T \beta_{\mathcal{C}})\} = 0.$$

Since β_j^M satisfies the conditional marginal moment condition (CMMC)

$$\mathbb{E}\{[X_j - \mathbb{E}(X_j | X_{\mathcal{C}}^T \beta_{\mathcal{C}})][Y - X_j \beta_j^M]\} = 0,$$

for any $j \in \mathcal{D}$, that implies

$$\begin{aligned}
&\mathbb{E}\{[X_j - \mathbb{E}(X_j|X_{\mathcal{C}}^T\beta_{\mathcal{C}})]X_j\}\beta_j^M \\
&\quad = \mathbb{E}\{[X_j - \mathbb{E}(X_j|X_{\mathcal{C}}^T\beta_{\mathcal{C}})]Y\} \\
&\quad = \mathbb{E}\{[X_j - \mathbb{E}(X_j|X_{\mathcal{C}}^T\beta_{\mathcal{C}})][X_{\mathcal{C}\cap\mathcal{A}}^{\mathrm{T}}\beta_{\mathcal{C}\cap\mathcal{A}}^* + X_{\mathcal{D}\cap\mathcal{A}}^{\mathrm{T}}\beta_{\mathcal{D}\cap\mathcal{A}}^*]\} \\
&\quad = \mathbb{E}\{[X_j - \mathbb{E}(X_j|X_{\mathcal{C}}^T\beta_{\mathcal{C}})]X_{\mathcal{D}\cap\mathcal{A}}^{\mathrm{T}}\beta_{\mathcal{D}\cap\mathcal{A}}^*\},
\end{aligned}$$

the last equation holds because

$$\begin{aligned}
&\mathbb{E}\{[X_j - \mathbb{E}(X_j|X_{\mathcal{C}}^{\mathrm{T}}\beta_{\mathcal{C}})]X_{\mathcal{C}\cap\mathcal{A}}^{\mathrm{T}}\beta_{\mathcal{C}\cap\mathcal{A}}^*\} \\
&\quad = \mathbb{E}\{[X_j X_{\mathcal{C}\cap\mathcal{A}}^{\mathrm{T}}\beta_{\mathcal{C}\cap\mathcal{A}}^* - \mathbb{E}(X_j X_{\mathcal{C}\cap\mathcal{A}}^{\mathrm{T}}\beta_{\mathcal{C}\cap\mathcal{A}}^*|X_{\mathcal{C}}^{\mathrm{T}}\beta_{\mathcal{C}})]\} = 0.
\end{aligned}$$

Note that

$$\begin{aligned}
0 &< \mathbb{E}\{[X_j - \mathbb{E}(X_j|X_{\mathcal{C}}^{\mathrm{T}}\beta_{\mathcal{C}})]^2|X_{\mathcal{C}}^{\mathrm{T}}\beta_{\mathcal{C}}\} \\
&= \mathbb{E}\{[X_j - \mathbb{E}(X_j|X_{\mathcal{C}}^{\mathrm{T}}\beta_{\mathcal{C}})]X_j|X_{\mathcal{C}}^{\mathrm{T}}\beta_{\mathcal{C}}\}.
\end{aligned}$$

The above equation holds because $\mathbb{E}\{[X_j - \mathbb{E}(X_j|X_{\mathcal{C}}^{\mathrm{T}}\beta_{\mathcal{C}})]\mathbb{E}(X_j|X_{\mathcal{C}}^{\mathrm{T}}\beta_{\mathcal{C}})|X_{\mathcal{C}}^{\mathrm{T}}\beta_{\mathcal{C}}\} = 0$. Therefore, we can get

$$\mathbb{E}\{[X_j - \mathbb{E}(X_j|X_{\mathcal{C}}^{\mathrm{T}}\beta_{\mathcal{C}})]X_j\} > 0$$

for any $j \in \mathcal{D}$.

If $j \in \bar{\mathcal{A}} \cap \mathcal{D}$, then we can get $\mathbb{E}\{[X_j - \mathbb{E}(X_j|X_{\mathcal{C}}^T\beta_{\mathcal{C}})]X_{\mathcal{D}\cap\mathcal{A}}^{\mathrm{T}}\beta_{\mathcal{D}\cap\mathcal{A}}^*\} = 0$, which means $\beta_j^M = 0 = \beta_j^*$.

If $j \in \mathcal{A} \cap \mathcal{D}$, i.e., $\beta_j^* \neq 0$, then

$$\begin{aligned}
\mathbb{E}\{[X_j - \mathbb{E}(X_j|X_{\mathcal{C}}^T\beta_{\mathcal{C}})]X_j\}\beta_j^M &= \mathbb{E}\{[X_j - \mathbb{E}(X_j|X_{\mathcal{C}}^T\beta_{\mathcal{C}})]X_{\mathcal{D}\cap\mathcal{A}\backslash j}^{\mathrm{T}}\beta_{\mathcal{D}\cap\mathcal{A}\backslash j}^*\} \\
&\quad + \mathbb{E}\{[X_j - \mathbb{E}(X_j|X_{\mathcal{C}}^T\beta_{\mathcal{C}})]X_j\}\beta_j^* \\
&= \mathbb{E}\{[X_j - \mathbb{E}(X_j|X_{\mathcal{C}}^T\beta_{\mathcal{C}})]X_j\}\beta_j^*.
\end{aligned}$$

Hence $\beta_j^* = \beta_j^M$. Therefore, we can prove our result by the above discussion. □

Proof of Lemma 1 By the definition of β_j^M, we have

$$\mathbb{E}\{[X_j - \mathbb{E}(X_j|X_{\mathcal{C}}^{\mathrm{T}}\beta_{\mathcal{C}})][Y - b'(X_{\mathcal{C}}^{\mathrm{T}}\beta_{\mathcal{C}}^M + X_j\beta_j^M)]\} = 0.$$

If $\beta_j^M = 0$, then the above equation is equivalent to

$$\mathbb{E}\{[X_j - \mathbb{E}(X_j|X_{\mathcal{C}}^{\mathrm{T}}\beta_{\mathcal{C}})][Y - b'(X_{\mathcal{C}}^{\mathrm{T}}\beta_{\mathcal{C}}^M)]\} = 0.$$

According to the property of conditional expectation, $\mathbb{E}\{[X_j - \mathbb{E}(X_j|X_{\mathcal{C}}^{\mathrm{T}}\beta_{\mathcal{C}})] b'(X_{\mathcal{C}}^{\mathrm{T}}\beta_{\mathcal{C}}^{M})\} = 0$, so the above two equations are equivalent to

$$\mathbb{E}\{[X_j - \mathbb{E}(X_j|X_{\mathcal{C}}^{\mathrm{T}}\beta_{\mathcal{C}})]Y\} = 0.$$

Next we show that $\mathbb{E}\{[X_j - \mathbb{E}(X_j|X_{\mathcal{C}}^{\mathrm{T}}\beta_{\mathcal{C}})]Y\} = 0$; then β_j^M must be zero.

First we note that $\mathbb{E}\{[X_j - \mathbb{E}(X_j|X_{\mathcal{C}}^{\mathrm{T}}\beta_{\mathcal{C}})]Y\} = 0$ means that

$$\mathbb{E}\{[X_j - \mathbb{E}(X_j|X_{\mathcal{C}}^{\mathrm{T}}\beta_{\mathcal{C}})]b'(X_{\mathcal{C}}^{\mathrm{T}}\beta_{\mathcal{C}}^{M} + X_j\beta_j^M)\} = 0.$$

Since function $b(\theta)$ is strictly convex or strictly concave in θ, it implies $b = \inf_\theta |b''(\theta)| > 0$. Then if $\beta_j^M \neq 0$, denote $\tilde{X}_j = X_j - \mathbb{E}(X_j|X_{\mathcal{C}}^{\mathrm{T}}\beta_{\mathcal{C}})$ and $w = X_{\mathcal{C}}^{\mathrm{T}}\beta_{\mathcal{C}}^{M} + X_j\beta_j^M$

$$\begin{aligned}
&|\mathbb{E}\{[X_j - \mathbb{E}(X_j|X_{\mathcal{C}}^{\mathrm{T}}\beta_{\mathcal{C}})]b'(X_{\mathcal{C}}^{\mathrm{T}}\beta_{\mathcal{C}}^{M} + X_j\beta_j^M)\}| \\
&\quad = |\mathbb{E}\{\tilde{X}_j b'(w)\}| \\
&\quad = |\mathbb{E}\{\mathbb{E}\{\tilde{X}_j|w\}b'(w)\}| \\
&\quad = |\mathbb{E}(\tilde{X}_j w)\{\mathrm{cov}(w)\}^{-1}\mathbb{E}[wb'(w)]|;
\end{aligned}$$

the last equation holds by linearity condition. It is clear that there exists $0 < \tilde{w} < w$,

$$\begin{aligned}
|\mathbb{E}[wb'(w)]| &= |\mathbb{E}[w^2 b''(\tilde{w})]| \\
&\geq |\mathbb{E}[b''(\tilde{w})w^2 I(w^2 \leq 1)]| \\
&\geq \inf_{0\leq\theta\leq 1} |b''(\theta)|\mathbb{E}[w^2 I(w^2 \leq 1)] \\
&> 0.
\end{aligned}$$

In addition,

$$\begin{aligned}
\mathbb{E}[\tilde{X}_j w] &= \mathbb{E}[(X_j - \mathbb{E}(X_j|X_{\mathcal{C}}^{\mathrm{T}}\beta_{\mathcal{C}}))(X_{\mathcal{C}}^{\mathrm{T}}\beta_{\mathcal{C}}^{M} + X_j\beta_j^M)] \\
&= \mathbb{E}[(X_j - \mathbb{E}(X_j|X_{\mathcal{C}}^{\mathrm{T}}\beta_{\mathcal{C}}))X_j\beta_j^M].
\end{aligned}$$

Therefore,

$$|\mathbb{E}\{[X_j - \mathbb{E}(X_j|X_{\mathcal{C}}^{\mathrm{T}}\beta_{\mathcal{C}})]b'(X_{\mathcal{C}}^{\mathrm{T}}\beta_{\mathcal{C}}^{M} + X_j\beta_j^M)\}| = r|\beta_j^M| > 0,$$

where $r = |\mathbb{E}[(X_j - \mathbb{E}(X_j|X_{\mathcal{C}}^{\mathrm{T}}\beta_{\mathcal{C}}))X_j]\{\mathrm{cov}(w)\}^{-1}\mathbb{E}[wb'(w)]| > 0$.

It leads to a contradiction, which means β_j^M must be zero. Therefore, we get our result. □

Proof of Theorem 2 Since β_j^M satisfied the equation

$$\mathbb{E}\{[X_j - \mathbb{E}(X_j|X_{\mathcal{C}}^{\mathrm{T}}\beta_{\mathcal{C}})][Y - b'(X_{\mathcal{C}}^{\mathrm{T}}\beta_{\mathcal{C}}^{M} + X_j\beta_j^M)]\} = 0,$$

which means

$$\mathbb{E}\{[X_j - \mathbb{E}(X_j|X_{\mathcal{C}}^{\mathrm{T}}\beta_{\mathcal{C}})]Y\} = \mathbb{E}\{[X_j - \mathbb{E}(X_j|X_{\mathcal{C}}^{\mathrm{T}}\beta_{\mathcal{C}})]b'(X_{\mathcal{C}}^{\mathrm{T}}\beta_{\mathcal{C}}^M + X_j\beta_j^M)\}.$$

Then we can get

$$\begin{aligned}&|\mathbb{E}\{[X_j - \mathbb{E}(X_j|X_{\mathcal{C}}^{\mathrm{T}}\beta_{\mathcal{C}})]b'(X_{\mathcal{C}}^{\mathrm{T}}\beta_{\mathcal{C}}^M + X_j\beta_j^M)\}|\\ &\quad= |\mathbb{E}\{[X_j - \mathbb{E}(X_j|X_{\mathcal{C}}^{\mathrm{T}}\beta_{\mathcal{C}})][b'(X_{\mathcal{C}}^{\mathrm{T}}\beta_{\mathcal{C}}^M + X_j\beta_j^M) - b'(X_{\mathcal{C}}^{\mathrm{T}}\beta_{\mathcal{C}}^M)]\}|\\ &\quad\le B\mathbb{E}\{[X_j - \mathbb{E}(X_j|X_{\mathcal{C}}^{\mathrm{T}}\beta_{\mathcal{C}})]X_j\}|\beta_j^M|,\end{aligned}$$

where $B = \sup_\theta b''(\theta)$. The first equation holds because $\mathbb{E}\{[X_j - \mathbb{E}(X_j|X_{\mathcal{C}}^{\mathrm{T}}\beta_{\mathcal{C}})]$ $b'(X_{\mathcal{C}}^{\mathrm{T}}\beta_{\mathcal{C}}^M)\} = 0$.

Since $\mathbb{E}\{[X_j - \mathbb{E}(X_j|X_{\mathcal{C}}^{\mathrm{T}}\beta_{\mathcal{C}})]X_j\} > 0$, we have

$$\begin{aligned}|\beta_j^M| &\ge \{B\mathbb{E}\{[X_j - \mathbb{E}(X_j|X_{\mathcal{C}}^{\mathrm{T}}\beta_{\mathcal{C}})]X_j\}\}^{-1}|\mathbb{E}\{[X_j - \mathbb{E}(X_j|X_{\mathcal{C}}^{\mathrm{T}}\beta_{\mathcal{C}})]b'(X_{\mathcal{C}}^{\mathrm{T}}\beta_{\mathcal{C}}^M + X_j\beta_j^M)\}|\\ &\ge \{B\mathbb{E}\{[X_j - \mathbb{E}(X_j|X_{\mathcal{C}}^{\mathrm{T}}\beta_{\mathcal{C}})]X_j\}\}^{-1}|\mathbb{E}\{[X_j - \mathbb{E}(X_j|X_{\mathcal{C}}^{\mathrm{T}}\beta_{\mathcal{C}})]Y\}|.\end{aligned}$$

Therefore, if condition *C1* holds, we can get $|\beta_j^M| \ge c_2 n^{-\kappa}$ for any $j \in \mathcal{D} \cap \mathcal{A}$. It means that $\min_{j\in\mathcal{D}\cap\mathcal{A}}|\beta_j^M| \ge c_2 n^{-\kappa}$. □

Proof of Lemma 2 First, together with CMMC (3) in linear models and the definition of α_j in (9), we can get

$$\alpha_j = \mathbb{E}\{[X_j - \mathbb{E}(X_j|X_{\mathcal{C}}^{\mathrm{T}}\beta_{\mathcal{C}})]X_j\}\beta_j^M.$$

It implies that $\alpha \ne 0$ is equivalent to $\beta_j^M \ne 0$ because $\mathbb{E}\{[X_j - \mathbb{E}(X_j|X_{\mathcal{C}}^{\mathrm{T}}\beta_{\mathcal{C}})]X_j\} > 0$. On the other hand, based on Theorem 1, we can get the conditional marginal signal $\beta_j^M = \beta_j^*$ for any $j \in \mathcal{D}$.

Therefore, we can get our result directly. □

Proof of Lemma 3 Note that $l_j(0) = 2\sum_{i=1}^n \log\{1 + \lambda g_{ij}^{(c)}(0)\}$, where λ satisfies $0 = \sum_{i=1}^n \frac{g_{ij}^{(c)}(0)}{1+\lambda g_{ij}^{(c)}(0)}$ and $g_{ij}^{(c)}(0) = [X_{ij} - \mathbb{E}(X_j|X_{i\mathcal{C}}^{\mathrm{T}}\beta_{\mathcal{C}})]Y_i$.

According to Condition 1 and the definition of $g_{ij}^{(c)}(0)$, we directly obtain

$$\min_{j\in\mathcal{D}\cap\mathcal{A}}|\mathbb{E}\{g_j^{(c)}(0)\}| \ge c_1 n^{-\kappa}. \tag{12}$$

On the other hand, following Lemma 2 in Chang et al. (2013b) and Condition 2, we can get the following inequality;

$$\mathbb{P}\{|[X_{ij} - \mathbb{E}(X_j | X_{i\mathcal{C}}^{\mathrm{T}} \beta_{\mathcal{C}})] Y_i| > u\} \leq 2K_1 \exp\{-K_2 u^{\gamma}\},$$

It means that

$$\mathbb{P}\{|g_{ij}^{(c)}(0)| > u\} \leq 2K_1 \exp\{-K_2 u^{\gamma}\}. \tag{13}$$

Together with (12) and(13) and following the same argument as the proof of Theorem 1, Proposition 2 and Corollary 1 in Chang et al. (2013b), we can prove our result. □

Proof of Lemma 4 Following Owen (2001), we can get

$$\widehat{l_j}(0) = 2 \max_{\hat{\lambda} \in \Lambda_n} \sum_{i=1}^{n} \log(1 + \hat{\lambda} \widehat{g_{ij}^{c}}(0)),$$

where $\Lambda_n = \{\hat{\lambda} : 1 + \hat{\lambda} \widehat{g_{ij}^{c}}(0) \geq n^{-1}$, for all $i = 1, \ldots, n\}$. To simplify the notation, we use $\widehat{g_{ij}}$ for $\widehat{g_{ij}^{c}}(0)$. Pick $\hat{\lambda} = (n^{\epsilon} \max_{l} |\widehat{g_{lj}}|)^{-1}$ for some $\epsilon > 0$; then $\hat{\lambda} \in \Lambda_n$ for sufficiently large n. Pick $t > 0$; we have

$$\mathbb{P}\{\widehat{l_j}(0) < 2t\} \leq \mathbb{P}\left\{ \sum_{i=1}^{n} \log[1 + \frac{\widehat{g_{ij}}}{n^{\epsilon} \max_{l} |\widehat{g_{lj}}|}] < t \right\}.$$

Note that

$$\log\left[1 + \frac{\widehat{g_{ij}}}{n^{\epsilon} \max_{l} |\widehat{g_{lj}}|}\right] = \frac{\widehat{g_{ij}}}{n^{\epsilon} \max_{l} |\widehat{g_{lj}}|} - \frac{1}{2(1 + c_i)^2} \frac{\widehat{g_{ij}}^2}{n^{2\epsilon} \max_{l} |\widehat{g_{lj}}|^2},$$

where $|c_i| \leq n^{-\epsilon}$; then we have

$$\sum_{i=1}^{n} \log\left[1 + \frac{\widehat{g_{ij}}}{n^{\epsilon} \max_{l} |\widehat{g_{lj}}|}\right] = \sum_{i=1}^{n} \frac{\widehat{g_{ij}}}{n^{\epsilon} \max_{l} |\widehat{g_{lj}}|} + R_n,$$

where $|R_n| \leq n^{1-2\epsilon}$. Notice that $\max_{l} |\widehat{g_{lj}}| = \max_{l} |g_{lj} + \widehat{g_{lj}} - g_{lj}| \leq \max_{l} |g_{lj}| + \max_{l} |\widehat{g_{lj}} - g_{lj}|$; then we can get

$$\begin{aligned}
&\mathbb{P}\{\widehat{l_j}(0) < 2t\} \\
&\leq \mathbb{P}\left\{ \sum_{i=1}^{n} \frac{\widehat{g_{ij}}}{n^{\epsilon} \max_{l} |\widehat{g_{lj}}|} < t + n^{1-2\epsilon} \right\} \\
&= \mathbb{P}\left\{ \sum_{i=1}^{n} \widehat{g_{ij}} < (t n^{\epsilon} + n^{1-\epsilon}) \max_{l} |\widehat{g_{lj}}| \right\}
\end{aligned}$$

$$\begin{aligned}
&= \mathbb{P}\left\{\sum_{i=1}^{n} g_{ij} + \sum_{i=1}^{n}(\widehat{g_{ij}} - g_{ij}) < (tn^{\epsilon} + n^{1-\epsilon})\max_{l}|\widehat{g_{lj}}|\right\}\\
&= \mathbb{P}\left\{\sum_{i=1}^{n} g_{ij} < (tn^{\epsilon} + n^{1-\epsilon})\max_{l}|\widehat{g_{lj}}| - \sum_{i=1}^{n}(\widehat{g_{ij}} - g_{ij})\right\}\\
&\le \mathbb{P}\left\{\sum_{i=1}^{n} g_{ij} < (tn^{\epsilon} + n^{1-\epsilon})\max_{l}|\widehat{g_{lj}}| + \sum_{i=1}^{n}|\widehat{g_{ij}} - g_{ij}|\right\}\\
&\le \mathbb{P}\left\{\sum_{i=1}^{n} g_{ij} < (tn^{\epsilon} + n^{1-\epsilon})\max_{l}|\widehat{g_{lj}}| + n\max_{i}|\widehat{g_{ij}} - g_{ij}|\right\}\\
&\le \mathbb{P}\left\{\sum_{i=1}^{n} g_{ij} < (tn^{\epsilon} + n^{1-\epsilon})\max_{l}|g_{lj}| + (tn^{\epsilon} + n^{1-\epsilon} + n)\max_{i}|\widehat{g_{ij}} - g_{ij}|\right\}\\
&\le \mathbb{P}\left\{\frac{1}{\sqrt{n}\sigma}\sum_{i=1}^{n}(g_{ij} - \mu_j) < \frac{(tn^{\epsilon-\frac{1}{2}} + n^{\frac{1}{2}-\epsilon})\max_{l}|g_{lj}| - \sqrt{n}\mu_j}{\sigma}\right.\\
&\qquad \left. + \frac{((tn^{\epsilon-\frac{1}{2}} + n^{\frac{1}{2}-\epsilon}) + \sqrt{n})\max_{i}|\widehat{g_{ij}} - g_{ij}|}{\sigma}\right\}\\
&\le \mathbb{P}\left\{\frac{1}{\sqrt{n}\sigma}\sum_{i=1}^{n}(g_{ij} - \mu_j) < \frac{(tn^{\epsilon-\frac{1}{2}} + n^{\frac{1}{2}-\epsilon})M - \sqrt{n}\mu_j}{\sigma}\right.\\
&\qquad \left. + \frac{((tn^{\epsilon-\frac{1}{2}} + n^{\frac{1}{2}-\epsilon}) + \sqrt{n})\max_{i}|\widehat{g_{ij}} - g_{ij}|}{\sigma}\} + \mathbb{P}\{\max_{l}|g_{lj}| > M\right\},
\end{aligned}$$

where $g_{ij} = g_{ij}^{(c)}(0)$, $\mu_j = \mathbb{E}g_{ij}$ and $\sigma^2 = \mathbb{E}\{(g_{ij} - \mu_j)^2\}$.

Under the linearity condition, denote $a = \text{cov}(X, X_{\mathcal{C}}^{T})\beta_{\mathcal{C}}\{\text{cov}(X_{\mathcal{C}}^{T}\beta_{\mathcal{C}})\}^{-1}$, $\mathbb{E}(X_j|X_{\mathcal{C}}^{T}\beta_{\mathcal{C}}) = a\beta_{\mathcal{C}}^{T}X_{\mathcal{C}}$. By $\widehat{g_{ij}} = [X_{ij} - \widehat{\mathbb{E}}(X_j|X_{i\mathcal{C}}^{\text{T}}\beta_{\mathcal{C}})]Y_i = [X_{ij} - \widehat{E_i}]Y_i$, where $\widehat{E_i} = \widehat{\mathbb{E}}(X_j|X_{i\mathcal{C}}^{\text{T}}\beta_{\mathcal{C}})$ is the estimator of $E_i = \mathbb{E}(X_j|X_{i\mathcal{C}}^{\text{T}}\beta_{\mathcal{C}}) = a\beta_{\mathcal{C}}^{T}X_{i\mathcal{C}}$, we have $|E_i - \widehat{E_i}| = |(\hat{a} - a)\beta_{\mathcal{C}}^{T}X_{i\mathcal{C}}|$ and $|\hat{a} - a| = O_p(n^{-1/2})$, where $\hat{a}$ is the estimator in Sect. 2, and $\max_{i}|X_{ik}Y_i| = O_p(n^{\omega})$, where $\omega \in (0, 1/2)$, $k \in \mathcal{C}$. Then we can get $\max_{i}|\widehat{g_{ij}} - g_{ij}| = O_p(n^{\omega-1/2})$.

For $L \to \infty$, pick ϵ such that $n^{\epsilon} = \frac{L}{\mu_j}$. Choose $\eta \in (0, \frac{2}{3})$ and let $M = \eta L$ and $2t = \frac{n\mu_j^2}{L^2}$, then $\frac{tn^{\epsilon}M}{n\mu_j} = \frac{\eta}{2}$ and $\frac{tn^{1-\epsilon}M}{n\mu_j} = \eta$.

$$\frac{((tn^{\epsilon-\frac{1}{2}} + n^{\frac{1}{2}-\epsilon}) + \sqrt{n})\max_{i}|\widehat{g_{ij}} - g_{ij}|}{\sigma} = O_p(\sqrt{n}\max_{i}|\widehat{g_{ij}} - g_{ij}|) = O_p(n^{\omega}),$$

and

$$\frac{(tn^{\epsilon-\frac{1}{2}}+n^{\frac{1}{2}-\epsilon})M-\sqrt{n}\mu_j}{\sigma}=O_p(n^{\frac{1}{2}}|\mu_j|).$$

Moreover, under condition 1, we can get $n^{\frac{1}{2}}|\mu_j|\geq c_1 n^{\frac{1}{2}-\kappa}$ for any $j\in\mathcal{D}\cap\mathcal{A}$; hence $n^{\omega}=o_p(n^{\frac{1}{2}}|\mu_j|)$ following our assumption $\kappa\leq\frac{1}{2}-\omega$. It implies that we can neglect

$$\frac{((tn^{\epsilon-\frac{1}{2}}+n^{\frac{1}{2}-\epsilon})+\sqrt{n})\max_i|\widehat{g_{ij}}-g_{ij}|}{\sigma}$$

or replace it by $O_p(n^{\frac{1}{2}}|\mu_j|)$. Hence, we have

$$\begin{aligned}\mathbb{P}\{\widehat{l_j}(0)<2t\}\leq\mathbb{P}&\left\{\frac{1}{\sqrt{n}\sigma}\sum_{i=1}^{n}(g_{ij}-\mu_j)<\frac{(tn^{\epsilon-\frac{1}{2}}+n^{\frac{1}{2}-\epsilon})M-\sqrt{n}\mu_j}{\sigma}\right\}\\&+\mathbb{P}\{\max_l|g_{lj}|>M\},\end{aligned}$$

which means

$$\begin{aligned}\mathbb{P}\{\widehat{l_j}(0)<\frac{n\mu_j^2}{L^2}\}\leq\mathbb{P}&\left\{\frac{1}{\sqrt{n}\sigma}\sum_{i=1}^{n}(g_{ij}-\mu_j)<\frac{(\frac{3}{2}\eta-1)\sqrt{n}\mu_j}{\sigma}\right\}\\&+K_1\exp\{-K_2M^{\gamma}+\log n\}.\end{aligned}\tag{14}$$

Since $|\mu_j|$ can be bounded by a uniform constant. And from Condition 1, $n\mu_j^2\geq c_1^2n^{1-2\kappa}$ for any $j\in\mathcal{D}\cap\mathcal{A}$,

$$\mathbb{P}\left\{\widehat{l_j}(0)<\frac{c_1^2n^{1-2\kappa}}{L^2}\right\}\leq\mathbb{P}\left\{\widehat{l_j}(0)<\frac{n\mu_j^2}{L^2}\right\}.$$

Then following (14) and Lemma 1 in Chang et al. (2013b), we can get

$$\begin{aligned}&\mathbb{P}\left\{\widehat{l_j}(0)<\frac{c_1^2n^{1-2\kappa}}{L^2}\right\}\\&\leq\begin{cases}\exp(-C_2n^{1-2\kappa})+\exp(-C_2L^{\gamma}), & \text{if}\quad(1-2\kappa)(1+2\delta)<1;\\ \exp(-C_2n^{\frac{1-\kappa}{1+\delta}})+\exp(-C_2L^{\gamma}), & \text{if}\quad(1-2\kappa)(1+2\delta)\geq1.\end{cases}\end{aligned}$$

Finally, choosing $L=n^{\frac{1}{2}-\kappa-\tau}$ for some $\tau\in(0,\frac{1}{2}-\kappa)$, we can get our result. □

Proof of Theorem 3 Notice that

$$\begin{aligned}\mathbb{P}\{\mathcal{D}\cap\mathcal{A}\subsetneq\widehat{\mathcal{D}\cap\mathcal{A}}_{\gamma_n}\} &= \mathbb{P}\{\text{There exists } j\in\mathcal{D}\cap\mathcal{A} \text{ such that } \widehat{l_j}(0)<c_1^2n^{2\tau}\}\\ &\le s_{re}\max_{j\in\mathcal{D}\cap\mathcal{A}}\mathbb{P}\{\widehat{l_j}(0)<c_1^2n^{2\tau}\};\end{aligned}$$

then we can get our result directly by Lemma 4. □

Proof of Lemma 5 Keep the notation in proof of Lemma 4, and first note that $l_j(0) = 2\sum_{i=1}^n\log\{1+\lambda\widehat{g_{ij}}\}$, where λ satisfies $0=\sum_{i=1}^n\frac{g_{ij}}{1+\lambda g_{ij}}$ and $g_{ij} = [X_{ij}-\mathbb{E}(X_j|X_{i\mathcal{C}}^{\mathrm{T}}\beta_{\mathcal{C}})]Y_i$. By Taylor expansion, we have

$$\begin{aligned}l_j(0) = {}& n\left(\frac{1}{n}\sum_{i=1}^n g_{ij}^2\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n g_{ij}\right)^2 - n\left(\frac{1}{n}\sum_{i=1}^n g_{ij}^2\right)^{-1}\left\{\frac{1}{n}\sum_{i=1}^n\frac{\lambda^2g_{ij}^3}{(1+c_{i2}\lambda g_{ij})^3}\right\}^2\\ &+\frac{2}{3}\sum_{i=1}^n\frac{\lambda^3g_{ij}^3}{1+c_{i1}\lambda g_{ij})^3}\\ =:{}& I_1+I_2+I_3.\end{aligned}$$

Define

$$\mathcal{M}=\left\{|\lambda|<\frac{4|n^{-1}\sum_{i=1}^n g_{ij}|}{3n^{-1}\sum_{i=1}^n g_{ij}^2} \quad\text{and}\quad \left|\frac{1}{n}\sum_{i=1}^n g_{ij}\right|\max_l|g_{lj}|<\frac{1}{4n}\sum_{i=1}^n g_{ij}^2\right\}.$$

Hence,

$$\begin{aligned}&\mathbb{P}\{l_j(0)\ge c_1^2n^{2\tau}\}\\ &\le\mathbb{P}\left\{I_1\ge\frac{c_1^2n^{2\tau}}{2}\right\}+\mathbb{P}\left\{I_3\ge\frac{c_1^2n^{2\tau}}{2},\mathcal{M}\text{ holds}\right\}+\mathbb{P}\left\{\mathcal{M}^c\right\}\\ &\le\mathbb{P}\left\{\left(\frac{1}{n}\sum_{i=1}^n g_{ij}\right)^2\ge\frac{c_1^2\mathbb{E}(g_{ij}-\mu_j)^2}{4n^{1-2\tau}}\right\}+\mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^n g_{ij}^2<\frac{\mathbb{E}(g_{ij}-\mu_j)^2}{2}\right\}\\ &\quad+\mathbb{P}\left\{C\left(\sum_{i=1}^n|g_{ij}|^3\right)\left|\frac{1}{n}\sum_{i=1}^n g_{ij}\right|^3\left(\frac{1}{n}\sum_{i=1}^n g_{ij}^2\right)^{-3}\ge\frac{c_1^2n^{2\tau}}{2}\right\}\\ &\quad+\mathbb{P}\left\{|\lambda|>\frac{4|n^{-1}\sum_{i=1}^n g_{ij}|}{3n^{-1}\sum_{i=1}^n g_{ij}^2}\right\}\\ &=:\mathcal{R}_1+\mathcal{R}_2+\mathcal{R}_3+\mathcal{R}_4.\end{aligned}\tag{15}$$

According to the proof of Proposition 4 in Chang et al. (2013b), we can obtain that

$$
\begin{aligned}
&\mathbb{P}\{l_j(0) \geq c_1^2 n^{2\tau}\}\\
&\quad \leq \mathcal{R}_1 + \mathcal{R}_2 + \mathcal{R}_3 + \mathcal{R}_4\\
&\quad \leq \begin{cases}
\exp(-C_3 n^{2\tau}) + \exp(-C_3 n^{\frac{\gamma}{6}}), & \text{if } \gamma < 2 \quad \text{and} \quad \eta > \frac{1}{4};\\
\exp(-C_3 n^{\gamma\eta}) + \exp(-C_3 n^{\frac{\gamma}{6}}), & \text{if } \gamma < 2 \quad \text{and} \quad \eta \leq \frac{1}{4};\\
\exp(-C_3 n^{\gamma\eta}) + \exp(-C_3 n^{\frac{\gamma}{6}}), & \text{if } \gamma \geq 2 \quad \text{and} \quad \eta \leq \frac{1}{\gamma+2};\\
\exp(-C_3 n^{\frac{\gamma}{\gamma+2}}) + \exp(-C_3 n^{2\tau}), & \text{if } \gamma \geq 4 \quad \text{and} \quad \eta > \frac{1}{\gamma+2};\\
\exp(-C_3 n^{\frac{\gamma}{6}}) + \exp(-C_3 n^{2\tau}), & \text{if } 2 \leq \gamma < 4 \quad \text{and} \quad \eta > \frac{1}{\gamma+2}.
\end{cases}
\end{aligned}
$$

On the other hand , we have $\widehat{l_j}(0) = 2\sum_{i=1}^n \log\{1 + \hat{\lambda}\widehat{g_{ij}}\}$, where $\hat{\lambda}$ satisfies $0 = \sum_{i=1}^n \frac{\widehat{g_{ij}}}{1+\hat{\lambda}\widehat{g_{ij}}}$ and $\widehat{g_{ij}} = [X_{ij} - \widehat{\mathbb{E}}(X_j|X_{i\mathcal{C}}^{\mathrm{T}}\beta_{\mathcal{C}})]Y_i$. Due to the same technique, we can get the sample version in (15) as following:

$$
\begin{aligned}
&\mathbb{P}\{\widehat{l_j}(0) \geq c_1^2 n^{2\tau}\}\\
&\leq \mathbb{P}\left\{\hat{I}_1 \geq \frac{c_1^2 n^{2\tau}}{2}\right\} + \mathbb{P}\left\{\hat{I}_3 \geq \frac{c_1^2 n^{2\tau}}{2}, \widehat{\mathcal{M}} \text{ holds}\right\} + \mathbb{P}\{\widehat{\mathcal{M}}^c\}\\
&\leq \mathbb{P}\left\{\left(\frac{1}{n}\sum_{i=1}^n \widehat{g_{ij}}\right)^2 \geq \frac{c_1^2 \mathbb{E}(g_{ij} - \mu_j)^2}{4n^{1-2\tau}}\right\} + \mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^n \widehat{g_{ij}}^2 < \frac{\mathbb{E}(g_{ij} - \mu_j)^2}{2}\right\}\\
&\quad + \mathbb{P}\left\{C\left(\sum_{i=1}^n |\widehat{g_{ij}}|^3\right)\left|\frac{1}{n}\sum_{i=1}^n \widehat{g_{ij}}\right|^3 \left(\frac{1}{n}\sum_{i=1}^n \widehat{g_{ij}}^2\right)^{-3} \geq \frac{c_1^2 n^{2\tau}}{2}\right\}\\
&\quad + \mathbb{P}\left\{|\hat{\lambda}| > \frac{4|n^{-1}\sum_{i=1}^n \widehat{g_{ij}}|}{3n^{-1}\sum_{i=1}^n \widehat{g_{ij}}^2}\right\}\\
&=: \widehat{\mathcal{R}}_1 + \widehat{\mathcal{R}}_2 + \widehat{\mathcal{R}}_3 + \widehat{\mathcal{R}}_4,
\end{aligned}
$$

where $\hat{I}_1$, $\hat{I}_3$, and $\widehat{\mathcal{M}}$ are the sample version of I_1, I_3 and $\mathcal{M}$, respectively.

In fact, $\mathbb{P}\left\{\widehat{l_j}(0) \geq c_1^2 n^{2\tau}\right\}$ also is bounded by $\mathcal{R}_1 + \mathcal{R}_2 + \mathcal{R}_3 + \mathcal{R}_4$. Firstly, note that

$$
\mathbb{P}\left\{\left|\frac{1}{n}\sum_{i=1}^n \widehat{g_{ij}}\right| \geq \tilde{C}n^{-\varepsilon}\right\} \leq \mathbb{P}\left\{\left|\frac{1}{n}\sum_{i=1}^n g_{ij}\right| \geq \tilde{C}n^{-\varepsilon}\right\} \tag{16}
$$

$$
\mathbb{P}\left\{\frac{1}{n}\sum_{i=}^n \widehat{g_{ij}}^2 < C\right\} \leq \mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^n g_{ij}^2 \leq C'\right\}, \tag{17}
$$

where ε, C, C' and $\tilde{C}$ are some positive constants. Because

$$\left|\frac{1}{n}\sum_{i=1}^{n}\widehat{g_{ij}}\right| \le \left|\frac{1}{n}\sum_{i=1}^{n}g_{ij}\right| + \max_i|g_{ij} - \widehat{g_{ij}}|$$

and $\max_i|\widehat{g_{ij}} - g_{ij}| = O_p(n^{\omega-\frac{1}{2}})$, following the same argument in proof of Lemma 4, we can neglect $\max_i|\widehat{g_{ij}} - g_{ij}|$ when $\varepsilon \le \frac{1}{2} - \omega$, and get (16). And

$$\frac{1}{n}\sum_{i=1}^{n}\widehat{g_{ij}}^2 = \frac{1}{n}\sum_{i=1}^{n}g_{ij}^2 + 2\frac{1}{n}\sum_{i=1}^{n}g_{ij}(\widehat{g_{ij}} - g_{ij}) + \frac{1}{n}\sum_{i=1}^{n}(\widehat{g_{ij}} - g_{ij})^2 \ge \frac{1}{2n}\sum_{i=1}^{n}g_{ij}^2;$$

then

$$\mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^{n}\widehat{g_{ij}}^2 < C\right\} \le \mathbb{P}\left\{\frac{1}{2n}\sum_{i=1}^{n}g_{ij}^2 < C\right\},$$

which implies (17) holds.

Therefore, choosing $\varepsilon = \frac{1}{2} - \tau$, we can easily get $\widehat{\mathcal{R}}_1 + \widehat{\mathcal{R}}_2 \le \mathcal{R}_1 + \mathcal{R}_2$. Due to the same technique, we have $\widehat{\mathcal{R}}_3 \le \mathcal{R}_3$. For the last part, since $\max_i|\widehat{g_{ij}}| \le \max_i|g_{ij}| + \max_i|\widehat{g_{ij}} - g_{ij}|$ and $\max_i|\widehat{g_{ij}} - g_{ij}| = O_p(n^{\omega-\frac{1}{2}})$, we have $\mathbb{P}\{\max_i|\widehat{g_{ij}}| > Cn^{\varepsilon}\} \le \mathbb{P}\{\max_i|g_{ij}| > Cn^{\varepsilon}\}$. And choose $\varepsilon \in (0, \eta \wedge (\frac{1}{2} - \omega)]$ in (16). Then we can get $\widehat{\mathcal{R}}_4 \le \mathcal{R}_4$ by (16) and (17) and the same argument in the proof of Lemma 6 in Chang et al. (2013b).

Finally, following by the proof of Proposition 4 in Chang et al. (2013b), we can get our result. □

Acknowledgements We thank the Editor, the Associate Editor and two referees for their constructive comments and suggestions, which have helped greatly improve the article. We are very grateful to Drs. Jinyuan Chang and Yichao Wu for sharing with us programs for implementing their methods.

References

Barut, E., Fan, J., Verhasselt, A. (2012). Conditional sure independence screening. http://arxiv.org/abs/1206.1024.

Bickel, P. J., Ritov, Y., Tsybakov, A. B. (2009). Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, *37*, 1705–1732.

Bühlmann, P., Van de Geer, S. (2011). *Statistics for high-dimensional data*: *Methods, theory and applications*. New York: Springer.

Candes, E., Tao, T. (2007). The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, *35*, 2313–2351.

Chang, J., Tang, C. Y., Wu, Y. (2013a). Marginal empirical likelihood and sure independence feature screening. *The Annals of Statistics*, *41*, 2123–2148.

Chang, J., Tang, C. Y., Wu, Y. (2013b). Supplement to "Marginal empirical likelihood and sure independence feature screening.". doi:10.1214/13-AOS1139SUPP.

Chang, J., Chen, S. X., Chen, X. (2015a). High dimensional generalized empirical likelihood for moment restrictions with dependent data. *Journal of Econometrics*, *185*, 283–304.

Chang, J., Tang, C. Y., Wu, Y. (2015b). Local independence feature screening for nonparametric and semiparametric models by marginal empirical likelihood. http://arxiv.org/abs/1502.07061.

Chen, S. X., Van Keilegom, I. (2009). A review on empirical likelihood methods for regression (with dicussions). *TEST*, *18*, 415–447.

Chen, S. X., Peng, L., Qin, Y. L. (2009). Effects of data dimension on empirical likelihood. *Biometrika*, *96*, 711–722.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R. (2004). Least angle regression (with discussions). *The Annals of Statistics*, *32*, 407–499.

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

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

Fan, J., Lv, J. (2011). Nonconcave penalized likelihood with np-dimensionality. *IEEE Transactions on Information Theory*, *57*, 5467–5484.

Fan, J., Song, R. (2010). Sure independence screening in generalized linear models with np-dimensionality. *The Annals of Statistics*, *38*, 3567–3604.

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

Fan, J., Feng, Y., Song, R. (2011a). Nonparametric independence screening in sparse ultra-high-dimensional additive models. *Journal of the American Statistical Association*, *106*, 544–557.

Fan, J., Lv, J., Qi, L. (2011b). Sparse high-dimensional models in economics. *Annual Review of Economics*, *3*, 291–317.

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

Hall, P., Titterington, D. M., Xue, J. H. (2009). Tilting methods for assessing the influence of components in a classifier. *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, *71*, 783–803.

Hastie, T., Tibshirani, R., Friedman, J. (2009). *The elements of statistical learning*: *data mining, inference and prediction*. New York: Springer.

Hjort, N. L., McKeague, I. W., Van Keilegom, I. (2009). Extending the scope of empirical likelihood. *The Annals of Statistics*, *37*, 1079–1111.

Leng, C., Tang, C. Y. (2012). Penalized empirical likelihood and growing dimensional general estimating equations. *Biometrika*, *99*, 703–716.

Li, G., Peng, H., Zhang, J., Zhu, L. (2012). Robust rank correlation based screening. *The Annals of Statistics*, *40*, 1846–1877.

Lin, L., Sun, J., Zhu, L. X. (2013). Nonparametric feature screening. *Computational Statistics and Data Analysis*, *67*, 162–174.

McCullagh, P., Nelder, J. A. (1989). *Generalized Linear Models*. New York: Chapman and Hall/CRC.

Newey, W., Smith, R. J. (2004). Higher order properties of gmm and generalised empirical likelihood estimators. *Econometrica*, *72*, 219–255.

Owen, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, *75*, 237–249.

Owen, A. B. (2001). *Empirical Likelihood*. New York: Chapman and Hall/CRC.

Qin, J., Lawless, J. (1994). Empirical likelihood and general estimating equations. *The Annals of Statistics*, *22*, 300–325.

Tang, C. Y., Leng, C. (2010). Penalized high-dimensional empirical likelihood. *Biometrika*, *97*, 905–920.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. *Series B* (*Statistical Methodology*), *58*, 267–288.

Zhang, C. H., Zhang, T. (2012). A general theory of concave regularization for high-dimensional sparse estimation problems. *Statistical Science*, *27*, 576–593.

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