The Annals of Statistics
2016, Vol. 44, No. 2, 515–539
DOI: 10.1214/15-AOS1374

LOCAL INDEPENDENCE FEATURE SCREENING FOR NONPARAMETRIC AND SEMIPARAMETRIC MODELS BY MARGINAL EMPIRICAL LIKELIHOOD

BY JINYUAN CHANG[*,†,1], CHENG YONG TANG[‡,2] AND YICHAO WU[§,3]

Southwestern University of Finance and Economics[*], *University of Melbourne*[†], *Temple University*[‡] *and North Carolina State University*[§]

We consider an independence feature screening technique for identifying explanatory variables that locally contribute to the response variable in high-dimensional regression analysis. Without requiring a specific parametric form of the underlying data model, our approach accommodates a wide spectrum of nonparametric and semiparametric model families. To detect the local contributions of explanatory variables, our approach constructs empirical likelihood locally in conjunction with marginal nonparametric regressions. Since our approach actually requires no estimation, it is advantageous in scenarios such as the single-index models where even specification and identification of a marginal model is an issue. By automatically incorporating the level of variation of the nonparametric regression and directly assessing the strength of data evidence supporting local contribution from each explanatory variable, our approach provides a unique perspective for solving feature screening problems. Theoretical analysis shows that our approach can handle data dimensionality growing exponentially with the sample size. With extensive theoretical illustrations and numerical examples, we show that the local independence screening approach performs promisingly.

1. Introduction. High-dimensional data are becoming increasingly available, and they have triggered surging investigation and development of new theory and methods; see Hastie, Tibshirani and Friedman (2009), Fan and Lv (2010) and Bühlmann and van de Geer (2011) for overview and discussions. Independence feature screening is a class of rapidly developing approaches that is particularly useful in preliminary analysis for preprocessing data to reduce the scale of high-dimensional statistical problems; see, among others, Fan and Lv (2008) and Fan and Song (2010) for the independence screening methods for linear and gener-

Received February 2015; revised August 2015.

[1] Supported by the Fundamental Research Funds for the Central Universities, NSFC (Grant No. 11501462) and a grant from the Australian Research Council. The bulk of this work was carried out during Jinyuan Chang's Ph.D. study at Peking University.

[2] Supported in part by NSF Grants IIS-1546087 and SES-1533956, and the Fox School of Business and Management, Temple University.

[3] Supported in part by NSF Grant DMS-10-55210 and NIH/NCI Grants R01-CA-149569 and P01-CA-142538.

MSC2010 subject classifications. Primary 62G99; secondary 62H99.

Key words and phrases. Empirical likelihood, high-dimensional data analysis, nonparametric and semiparametric models, sure independence screening.

alized linear models, Mai and Zou (2013) for variable screening in classification problems, Zhu et al. (2011), Li et al. (2012) and Li, Zhong and Zhu (2012) for feature screening methods using more general types of correlations.

Broadly speaking, independence feature screening methods rely on ranking estimations measuring the marginal contributions of explanatory variables. For example, Fan and Lv (2008) and Fan and Song (2010) consider ranking magnitudes of marginal estimators under some parametric models. In linear and generalized linear models, marginal estimator-based ranking can be viewed as equivalent to marginal correlation based ranking [Fan and Lv (2008), Chang, Tang and Wu (2013a)]. Various generalized versions of the correlation and conditional correlation are also considered as ranking criteria [Zhu et al. (2011), Li et al. (2012), Fan, Ma and Dai (2014), Liu, Li and Wu (2014)]. Recently, Fan, Feng and Song (2011) consider a ranking measure based on aggregating local contributions from an explanatory variable in a framework of nonparametric additive models using marginal penalized splines approach. In classification problems, Mai and Zou (2013) propose to use the so-called Kolmogrov filter to construct a ranking criterion based on aggregating sample distributional discrepancies between the two groups of interest at all observed values of a predictor.

We consider in this paper an independence feature screening method for a general class of regression problems covering the nonparametric additive models, semiparametric single-index models and multiple-index models, and varying coefficient models as special cases. There are two building blocks for constructing the screening criterion in our approach. The first one is the nonparametric regression applied marginally on one explanatory variable at a time. For overview of nonparametric regression methods, see Fan and Gijbels (1996) and Härdle (1990). The second building block is empirical likelihood [Owen (1988)] constructed locally for the marginal nonparametric regression. Instead of acquiring some marginal estimators, our approach is capable of objectively and conveniently assessing the strength of data evidence for testing the local contributions of a given explanatory variable. Moreover, as has been noted in the literature, an independence feature screening procedure may miss explanatory variables that are marginally unrelated but jointly related to the response [Fan and Lv (2008)]. To address this issue, many iterative versions of feature screening methods have been proposed. We borrow the idea of Zhu et al. (2011) and propose an iterative version of our local independence feature screening procedure.

Our study carries innovative contributions from a few aspects. First and foremost, the perspective of our approach is unique compared with other existing ones. Our approach directly targets at quantifying the strength of data evidence against the null hypothesis that explanatory variables are not locally contributing to the response variable. Hence, it actually requires no estimation. Moreover, as shown in our theoretical analysis, the fundamental statistic in our approach is self-Studentized, automatically incorporating variance of the marginal statistical approach. All existing approaches for nonparametric and semiparametric models require estimating marginal contributions and incorporate no effect from the level of variations of the marginal estimators. As a consequence, ranking the nonstandardized magnitudes of the marginal estimators may not best reflect the marginal contributions from predictors. Additionally, there may be difficulties when identifying the marginal effect becomes an issue, for example, in single- and multiple-index models. We show in our numerical examples that our approach outperforms others especially when the signal of the marginal contribution is weak, and when the variation of the response variable is more complex, all thanks to the unique perspective of the proposed marginal empirical likelihood approach. Second, existing approaches are typically investigated within specific families of models while our approach targets at detecting generic local contributions to the response variable from an explanatory variable. Thus, our method is suitable for capturing more general nonlinear effects in explanatory variables for solving a broad range of high-dimensional problems. Our theoretical analysis establishes the validity for feature screening in a general and broad setup, allowing data dimensionality to grow exponentially with the sample size, and our numerical and real data examples demonstrate that our method performs very promisingly.

Our investigation also contributes in solving challenging empirical likelihood problems for high-dimensional nonparametric and semiparametric statistical problems. In existing literature, much effort has been devoted into extending the empirical likelihood of Owen (1988, 1991) for parametric models to nonparametric and semiparametric models; see, among others, Chen (1996), Chen and Qin (2000) and the review in Chen and Van Keilegom (2009). For high-dimensional data, it remains open for solving empirical likelihood problems in nonparametric and semiparametric scenarios where merits such as robustness and other nonparametric features are highly desirable [Chang, Chen and Chen (2015), Chen, Peng and Qin (2009), Hjort, McKeague and Van Keilegom (2009), Leng and Tang (2012), Tang and Leng (2010)]. Recently, Chang, Tang and Wu (2013a) investigate marginal empirical likelihood for general high-dimensional parametric models specified by the estimating equations. Nevertheless, studying high-dimensional empirical likelihood beyond parametric models remains open because formulating and characterizing the empirical likelihood locally itself is known to be an important and difficult problem [Chen and Van Keilegom (2009)]. Our study on the local feature screening procedure solves the problem of constructing and characterizing empirical likelihood locally, which ideally fits a broad class of nonparametric and semiparametric models. Additionally, for summarizing the contribution from one specific predictor, we propose and justify an approach for aggregating the data evidence for local contributions, which in turn delivers the validity of our feature screening procedure. Remarkably, our approach can handle exponential data dimensionality even in the nonparametric and semiparametric settings where the convergence rate of nonparametric kernel regression is known to be slower [Fan and Gijbels (1996), Härdle (1990)].

The rest of the paper is organized as follows. Section 2 introduces the methodology of independence feature screening for nonparametric models and presents the corresponding theoretical properties. In Section 3, we apply this unified screening approach to deal with problems in nonparametric additive models, single-index models and multiple-index models, and varying coefficient models. As a methodological extension, we outline the iterative version of our unified screening approach in Section 4. Our simulation studies in Section 5 demonstrate the effectiveness of this method. We conclude with a discussion in Section 6, and relegate a real data analysis and the proofs to the supplementary file of this paper [Chang, Tang and Wu (2016)].

2. Main results.

2.1. *Methods.* Suppose that we have a random sample $\{(\mathbf{X}_i, Y_i)\}_{i=1}^n$ from the data model

$$Y = m(\mathbf{X}) + \varepsilon, \tag{2.1}$$

where $\mathbf{X} = (X_1, \dots, X_p)^{\mathrm{T}}$ and ε is the random error with $\mathbb{E}(\varepsilon|\mathbf{X}) = 0$. In our study, no specification of $m(\mathbf{X})$ is required. The data dimensionality p of the explanatory variable vector $\mathbf{X}$ can grow exponentially with the sample size n, but the true model is very sparse in the sense that there is only a small fraction of the explanatory variables contributing to the response variable Y. Let $\mathcal{M}_* = \{1 \le j \le p : \mathbb{E}(Y|\mathbf{X})$ varies with the value of $X_j\}$, and we call the variables indexed by $\mathcal{M}_*$ contributing explanatory variables. Without loss of generality, we assume that $\mathbb{E}(Y) = 0$ implying that $\mathbb{E}\{m(\mathbf{X})\} = 0$.

Since $\mathbf{X}$ is high-dimensional and without any prior information on which of them are contributing in explaining Y, a natural idea is to investigate the marginal contribution from each explanatory variable in explaining Y to justify whether it is relevant. For such a purpose, we consider marginal nonparametric regression problems:

$$\min_{f_j \in \mathscr{L}_2} \mathbb{E}\big[\{Y - f_j(X_j)\}^2\big] \qquad (j = 1, \dots, p), \tag{2.2}$$

where $\mathscr{L}_2$ denotes the class of square integrable functions. Note that $\mathbb{E}(Y|X_j)$ is the minimizer of (2.2). Naturally, we use $f_j(x) = \mathbb{E}(Y|X_j = x)$ to evaluate the marginal contribution of X_j locally at $X_j = x$. If an explanatory variable X_j is not contributing to Y marginally, then $f_j(x) = 0$ for all $x \in \mathcal{X}$. Here, $\mathcal{X}$ is the support of X_j. This motivates us to investigate a feature screening procedure by assessing whether $\mathbb{E}(Y|X_j) \equiv 0$ or not for each $j = 1, \dots, p$. However, how to develop a nice way for summarizing the impact due to X_j is not straightforward due to the fact that $f_j(x)$ is a function of x so that one needs to assess $f_j(x)$ for all values over the support of X_j.

We consider the Nadaraya–Watson (NW) estimator for $f_j(x)$:

$$\hat{f}_j(x) = \frac{n^{-1} \sum_{i=1}^n \mathcal{K}_h(X_{ij} - x) Y_i}{n^{-1} \sum_{i=1}^n \mathcal{K}_h(X_{ij} - x)}, \tag{2.3}$$

where $\mathcal{K}_h(u) = h^{-1}\mathcal{K}(\frac{u}{h})$ for some kernel function $\mathcal{K}$ and h is the bandwidth. For capturing the marginal variable effect, the choice of the NW estimator does not compromise the general applicability of the marginal empirical likelihood with other nonparametric approaches, for example, the local linear estimator [Fan and Gijbels (1996)], etc. Intuitively, $\hat{f}_j(x)$ should be small for all $x \in \mathcal{X}$ if X_j does not marginally contribute to Y.

Empirical likelihood [Owen (1988, 2001)] is an influential nonparametric likelihood approach. Without requiring to assume full parametric distributions, empirical likelihood shares some desirable merits of the conventional likelihood such as χ^2-distributed likelihood ratios and Bartlett correctability; see Chen and Van Keilegom (2009) for a review. For assessing $f_j(x) = 0$ at a given x without distributional assumptions, we construct the following empirical likelihood:

$$\begin{aligned}&\mathrm{EL}_j(x,0)\\ &\quad = \sup\left\{\prod_{i=1}^n w_i : w_i \geq 0, \sum_{i=1}^n w_i = 1, \sum_{i=1}^n w_i \mathcal{K}_h(X_{ij} - x)Y_i = 0\right\}.\end{aligned} \tag{2.4}$$

By applying the Lagrange multiplier method for solving (2.4), we obtain the empirical likelihood ratio:

$$\begin{aligned}\ell_j(x,0) &= -2\log\{\mathrm{EL}_j(x,0)\} - 2n\log n\\ &= 2\sum_{i=1}^n \log\{1 + \lambda\mathcal{K}_h(X_{ij} - x)Y_i\},\end{aligned} \tag{2.5}$$

where λ is the univariate Lagrange multiplier solving $0 = \sum_{i=1}^n \frac{\mathcal{K}_h(X_{ij}-x)Y_i}{1+\lambda\mathcal{K}_h(X_{ij}-x)Y_i}$. Since the denominator in (2.3) converges to the density of X_j evaluated at x, a large value of $\ell_j(x,0)$ is taken as evidence against $f_j(x) = 0$ provided that the density of X_j is bounded away from 0 at x. Hence, $\ell_j(x,0)$ is indeed a statistic for testing whether or not the numerator in (2.3) has zero mean locally at x. If 0 is not in the convex hull of $\{\mathcal{K}_h(X_{ij} - x)Y_i\}_{i=1}^n$, we define $\ell_j(x,0) = \infty$ as a strong evidence of the local contribution from X_j.

For assessing $\mathbb{E}(Y|X_j) \equiv 0$, we propose to use

$$\ell_j(0) = \sup_{x \in \mathcal{X}_n} \ell_j(x,0) \tag{2.6}$$

for each $j = 1, \dots, p$, where $\mathcal{X}_n$ is a partition of the support $\mathcal{X}$ into several intervals. For feature screening purpose, we propose selecting the set of explanatory variables by

$$\widehat{\mathcal{M}}_{\gamma_n} = \{1 \leq j \leq p : \ell_j(0) \geq \gamma_n\}. \tag{2.7}$$

We call our method a local independence feature screening approach by observing that in (2.4) it is the local correlation evaluated at x between X_j via $\mathcal{K}_h(X_j - x)$ and Y is assessed. Such a notion is seen as extending the initial proposal of independence feature screening using correlations as in Fan and Lv (2008). As the empirical likelihood ratio $\ell_j(x,0)$ is self-Studentized, the statistic $\ell_j(0)$ is robust to the heterogeneous cases. Compared with L_2-type screening statistics, the L_∞-type screening statistic $\ell_j(0)$ is seen to be more suitable when $\mathbb{E}(Y|X_j = x)$ is away from zero locally at some x instead of globally. We will give the specification of γ_n later in Theorem 1 under which the proposed approach has the sure screening property, that is, capable of identifying the set of contributing explanatory variables. However, choosing γ_n is generally hard in practice. Thus, following the convention of existing screening methods, we suggest running procedure to preprocess data by selecting a prespecified number of variables.

Practically implementing the proposed method is convenient. In practice, a natural choice is to evaluate the statistic (2.6) by $\ell_j(0) = \max_{1\le i\le n} \ell_j(X_{ij}, 0)$, where $\{X_{ij}\}_{i=1}^n$ are the n observations of the jth explanatory variable. Since evaluating $\ell_j(x,0)$ only involves univariate optimization when solving (2.5) using the Lagrange multiplier method, the screening statistics can be carried out easily by the existing algorithms. As for the bandwidth h in the marginal NW estimator (2.3), we note that conventional bandwidth selection methods such as cross-validation and the reference rule [Fan and Gijbels (1996)] can be applied. Our theory in the next section demonstrates the validity of the variable screening procedure for a range of bandwidth applied in (2.3) including ones selected by methods like cross-validation and the reference rule. Our numerical examples also show that the approach implemented with bandwidth selected by the cross-validation performs satisfactorily.

2.2. *Theoretical properties*. Throughout this paper, we use $\|\cdot\|_2$ and $\|\cdot\|_\infty$ to denote the L_2-norm and sup-norm, respectively, and $C^r(\mathcal{I})$ denotes the class of all continuous functions defined over $\mathcal{I}$ that are r times differentiable. We assume the following conditions.

(A.1) The marginal projections $\{f_j\}_{j=1}^p$ belong to $C^r(\mathcal{X})$. If $r=0$, f_j's satisfy the Lipschitz condition with order $\alpha \in (0,1]$, that is $|f_j(s) - f_j(t)| \le K_1|s-t|^\alpha$ for any $s,t \in \mathcal{X}$, where K_1 is a positive constant uniformly for any $j=1,\ldots,p$. In addition, there exists a constant K_2 such that $|f_j^{(r)}(x)| \le K_2$ for any $x \in \mathcal{X}$ and $j=1,\ldots,p$.

(A.2) The marginal density function g_j of X_j satisfies $0 < K_3 \le g_j(x) \le K_4 < \infty$ on $\mathcal{X}$ for $j=1,\ldots,p$. In addition, we assume that each g_j belongs to $C^r(\mathcal{X})$ for the r given in (A.1) and $|g_j^{(r)}(x)| \le K_5$ for any $x \in \mathcal{X}$ and $j=1,\ldots,p$.

(A.3) There exist nonnegative constants $c_1 > 0$ and $\kappa \in [0, \frac{\max\{r,\alpha\}}{2\max\{r,\alpha\}+2})$ such that $\min_{j\in\mathcal{M}_*} \|f_j\|_\infty \ge c_1 n^{-\kappa}$, where r and α are specified in (A.1).

(A.4) Let $\|\mathcal{X}_n\|$ be the largest length of the intervals in the partition $\mathcal{X}_n$, and there exists some positive constant ξ such that $\|\mathcal{X}_n\| = n^{-\xi}$.

(A.5) There exist positive constants K_6, K_7 and γ such that $\mathbb{P}(|Y| \ge u) \le K_6 \exp(-K_7 u^\gamma)$ for any $u > 0$.

(A.6) For r specified in (A.1), if $r \geq 1$, the kernel function $\mathcal{K}(\cdot)$ is of order r, that is, $\int \mathcal{K}(u)\,du = 1$, $\int u^k \mathcal{K}(u)\,du = 0$ for $k = 1, \dots, r-1$ and $\int u^r \mathcal{K}(u)\,du > 0$. If $r = 0$, the kernel function satisfies $\mathcal{K}(u) \geq 0$ for any u and $\int \mathcal{K}(u)\,du = 1$.

Here, (A.1) is a general condition describing the continuity of each $f_j(x) = \mathbb{E}(Y|X_j = x)$. If the first derivation of each f_j exists, then $r \geq 1$ and $\alpha = 1$. Assumption (A.2) is standard for kernel regression implying that the density of X_j does not vanish on its support; see, for example, Härdle (1990), and it implies bounded support of the explanatory variables. For ease in presentation, we take the same support $\mathcal{X} = [a, b]$ for all X_j which can be easily satisfied because some location-scale transformation can always be applied in practice if otherwise. The condition in (A.3) is for identifying $\mathcal{M}_*$, which requires that the minimal signal strength measured by $\|f_j\|_\infty$ cannot be too weak. The restriction of the minimal signal strength depends on the continuity of f_j via r. The smoother f_j's are, the weaker the condition on the signal strength is required, and the minimal signal strength cannot vanish at a rate faster than $n^{-1/2}$. Assumption (A.4) regularizes the partition of the support $\mathcal{X}$ to be of size at least $O(n^{\xi})$. Assumption (A.5) on the tail distribution of the response variable is a conventional technical requirement for Cramér-type large deviations. For example, $\gamma = 2$ if the response variable Y is a normal or sub-Gaussian distribution, and $\gamma = \infty$ if Y has a compact support. Assumption (A.6) specifies the requirement for the kernel function so that the bias due to kernel smoothing is not dominating; see Müller (1987) for more detail about higher order kernel functions.

For the parameters r, α and γ specified in above assumptions, let

$$\varrho_1 = \max\{r, \alpha\}, \qquad \varrho_2 = \max\{\gamma, 2\} + 2 \quad \text{and} \quad \delta = \max\left\{\frac{2}{\gamma} - 1, 0\right\}. \tag{2.8}$$

The parameter ϱ_1 characterizes the continuity of the marginal projections and densities. Parameters ϱ_2 and δ are related to the tail probabilistic behavior of response variable Y. Meanwhile, we assume that the bandwidth h used in (2.4) satisfies $h \asymp n^{-w}$ for some positive w whose specification is discussed later.

PROPOSITION 1. *Under assumptions* (A.1)–(A.6), *pick* $w \in [\frac{\kappa}{\varrho_1}, 1)$ *and* $\xi > \kappa + 2w$, *then there exists a uniform constant* C_1 *such that for any* $j \in \mathcal{M}_*$ *and* $L \to \infty$,

$$\mathbb{P}\left\{\ell_j(0) < \frac{c_1^2 K_3^2 n^{1-2\kappa-2w}}{2L^2}\right\} \leq \exp(-C_1 L^{\gamma}) \\ + \exp(-C_1 n^{\min\{1-2\kappa-w, (1-\kappa-w)/(1+\delta)\}}).$$

Proposition 1 gives a uniform result for all explanatory variables contributing in the true model. The maximum distance between the adjacent two points that are used to construct our procedure should be $o(n^{-\kappa-2w})$. Specifically, with large probability and uniformly for all $j \in \mathcal{M}_*$, the diverging rate of $\ell_j(0)$ is not slower than $n^{1-2\kappa-2w}L^{-2}$. If $j \notin \mathcal{M}_*$, that is, the explanatory variable X_j does not have the marginal contribution to Y (i.e., $f_j = 0$), following the argument in Owen (1988) and Chang, Tang and Wu (2013a), it can be shown that the corresponding $\ell_j(0)$ is $O_p(1)$. Hence, $n^{1/2-\kappa-w}L^{-1}$ is required to diverge as $n \to \infty$ for sure independence screening. Furthermore, we note that the requirement for the bandwidth used in Proposition 1 is mild, which can be naturally satisfied by the conventional optimal bandwidth $h = O(n^{-1/5})$ selected by the cross-validation method.

Let $L = n^{1/2-\kappa-w-\tau}$ for some $\tau \in (0, \frac{1}{2} - \kappa - w)$. A more clear uniform result related to the probabilistic behavior of the statistics $\ell_j(0)$ for $j \in \mathcal{M}_*$ is described in the following corollary.

COROLLARY 1. *Under assumptions* (A.1)–(A.6), *pick* $w \in [\frac{\kappa}{\varrho_1}, \frac{1}{2} - \kappa)$, $\tau \in (0, \frac{1}{2} - \kappa - w)$ *and* $\xi > \kappa + 2w$, *then*

$$\max_{j\in\mathcal{M}_*} \mathbb{P}\{\ell_j(0) < \tfrac{1}{2}c_1^2K_3^2n^{2\tau}\} \le \exp\{-C_1 n^{(1/2-\kappa-w-\tau)\gamma}\} + \exp(-C_1 n^{\min\{1-2\kappa-w,(1-\kappa-w)/(1+\delta)\}}),$$

where C_1 *is given in Proposition* 1.

Choosing threshold level $\gamma_n = \frac{1}{2}c_1^2K_3^2n^{2\tau}$ in (2.7) and noting that

$$\begin{aligned}\mathbb{P}(\mathcal{M}_* \subsetneq \widehat{\mathcal{M}}_{\gamma_n}) &= \mathbb{P}\{\text{There exists } j \in \mathcal{M}_* \text{ such that } \ell_j(0) < \tfrac{1}{2}c_1^2K_3^2n^{2\tau}\}\\ &\le s \max_{j\in\mathcal{M}_*} \mathbb{P}\{\ell_j(0) < \tfrac{1}{2}c_1^2K_3^2n^{2\tau}\},\end{aligned}$$

we establish the sure screening property of our approach in the following theorem based on Corollary 1.

THEOREM 1. *Under assumptions* (A.1)–(A.6), *pick* $w \in [\frac{\kappa}{\varrho_1}, \frac{1}{2} - \kappa)$, $\gamma_n = \frac{1}{2}c_1^2K_3^2n^{2\tau}$ *for some* $\tau \in (0, \frac{1}{2} - \kappa - w)$, *and* $\xi > \kappa + 2w$, *then*

$$\mathbb{P}(\mathcal{M}_* \subset \widehat{\mathcal{M}}_{\gamma_n}) \ge 1 - s\exp\{-C_1 n^{(1/2-\kappa-w-\tau)\gamma}\} - s\exp(-C_1 n^{\min\{1-2\kappa-w,(1-\kappa-w)/(1+\delta)\}}),$$

where C_1 *is given in Proposition* 1.

Theorem 1 implies that our local independence feature screening method can handle nonpolynomial dimensionality: $\log p = o(n^{\epsilon})$ for $\epsilon = \min\{1-2\kappa-w, (\frac{1}{2}-\kappa-w-\tau)\gamma\}$. By noting that $w \ge \frac{\kappa}{\varrho_1}$, the highest dimensionality is achieved with the optimal $\epsilon = \min\{1 - 2\kappa - \frac{\kappa}{\varrho_1}, (\frac{1}{2} - \kappa - \frac{\kappa}{\varrho_1})\gamma\}$ when τ is close enough to zero.

It actually depends on κ, ϱ_1 and γ, that is, the signal strength, smoothness of f_j's and the tail probabilistic behavior of Y. If Y follows a normal or sub-Gaussian distribution such that $\gamma = 2$, the corresponding highest dimensionality satisfies $\log p = o(n^{1-2\kappa-2\kappa/\varrho_1})$. Furthermore, if the projections f_j's have derivatives of all orders such that $\varrho_1 = r = \infty$, then the highest dimensionality satisfies $\log p = o(n^{1-2\kappa})$.

In what follows, we consider the size of the selected set $\widehat{\mathcal{M}}_{\gamma_n}$ under an ideal case that

$$\max_{j \notin \mathcal{M}_*} \|f_j\|_\infty = o(n^{-\kappa}). \tag{2.9}$$

The key is to investigate the probabilistic behavior of $\mathbb{P}\{\ell_j(0) \geq \frac{1}{2}c_1^2 K_3^2 n^{2\tau}\}$ for each $j \notin \mathcal{M}_*$ which is given in the next proposition.

PROPOSITION 2. *Under assumptions* (A.1)–(A.2) *and* (A.4)–(A.6), *suppose* $\max_{j \notin \mathcal{M}_*} \|f_j\|_\infty = O(n^{-\eta})$ *for some* $\eta > \frac{(2\varrho_1+1)\kappa}{2\varrho_1}$. *Pick* $w \in [\frac{\kappa}{\varrho_1}, \min\{\frac{1}{2} - \kappa, 2(\eta - \kappa)\})$, $\tau \in (\max\{\frac{1}{2} - \eta - \frac{w}{2}, 0\}, \frac{1}{2} - \kappa - w)$ *and* $\xi > \kappa + 2w$. *If* $\inf_{u \in [a,b]} \mathbb{E}(Y^2 | X_j = u) \geq \rho$ *for some positive* ρ *holds for any* $j \notin \mathcal{M}_*$, *then there exists a uniform positive constant* C_2 *such that for any* $j \notin \mathcal{M}_*$,

$$\mathbb{P}\{\ell_j(0) \geq \tfrac{1}{2}c_1^2 K_3^2 n^{2\tau}\} \leq \exp(-C_2 n^{\min\{\eta\gamma, (1-w)\gamma/\varrho_2, 2\tau, \gamma(1-w)/6\}}).$$

From Proposition 2, we can find that the quantities on the right-hand side are decreasing as w is increasing. Thus, the optimal $w = \frac{\kappa}{\varrho_1}$ is the same as the one for the best dimensionality p discussed previously. Hence, the optimal bandwidth in our screening procedure is $w = \frac{\kappa}{\varrho_1}$, which is quite sensible because intuitively the smoother each f_j is, the larger the bandwidth is allowed. The corresponding upper bound for $\mathbb{P}\{\ell_j(0) \geq \frac{1}{2}c_1^2 K_3^3 n^{2\tau}\}$ is given in the following corollary.

COROLLARY 2. *Under assumptions* (A.1)–(A.2) *and* (A.4)–(A.6), *suppose* $\max_{j \notin \mathcal{M}_*} \|f_j\|_\infty = O(n^{-\eta})$ *for some* $\eta > \frac{(2\varrho_1+1)\kappa}{2\varrho_1}$. *Pick* $w = \frac{\kappa}{\varrho_1}$, $\tau \in (\max\{\frac{1}{2} - \eta - \frac{\kappa}{2\varrho_1}, 0\}, \frac{1}{2} - \frac{(\varrho_1+1)\kappa}{\varrho_1})$ *and* $\xi > \frac{(\varrho_1+2)\kappa}{\varrho_1}$. *If* $\inf_{u \in [a,b]} \mathbb{E}(Y^2 | X_j = u) \geq \rho$ *for some positive* ρ *holds for any* $j \notin \mathcal{M}_*$, *then there exists a uniform positive constant* C_3 *such that for any* $j \notin \mathcal{M}_*$,

$$\mathbb{P}\{\ell_j(0) \geq \tfrac{1}{2}c_1^2 K_3^2 n^{2\tau}\} \leq p \exp(-C_3 n^{\min\{\eta\gamma, (\varrho_1-\kappa)\gamma/(\varrho_1\varrho_2), 2\tau, (\varrho_1-\kappa)\gamma/(6\varrho_1)\}}).$$

By noting that

$$\begin{aligned}
|\widehat{\mathcal{M}}_{\gamma_n}| &= \sum_{j \in \mathcal{M}_*} I\left\{\ell_j(0) \geq \frac{1}{2}c_1^2 K_3^2 n^{2\tau}\right\} + \sum_{j \notin \mathcal{M}_*} I\left\{\ell_j(0) \geq \frac{1}{2}c_1^2 K_3^2 n^{2\tau}\right\} \\
&\leq s + \sum_{j \notin \mathcal{M}_*} I\left\{\ell_j(0) \geq \frac{1}{2}c_1^2 K_3^2 n^{2\tau}\right\},
\end{aligned}$$

we have $\mathbb{P}(|\widehat{\mathcal{M}}_{\gamma_n}| > s) \le \sum_{j \notin \mathcal{M}_*} \mathbb{P}\{\ell_j(0) \ge \frac{1}{2}c_1^2 K_3^2 n^{2\tau}\}$. Hence, from Corollary 2, we obtain the following theorem for the size of $\widehat{\mathcal{M}}_{\gamma_n}$.

THEOREM 2. *Under assumptions* (A.1)–(A.2) *and* (A.4)–(A.6), *suppose* $\max_{j \notin \mathcal{M}_*} \|f_j\|_\infty = O(n^{-\eta})$ *for some* $\eta > \frac{(2\varrho_1+1)\kappa}{2\varrho_1}$. *Pick* $w = \frac{\kappa}{\varrho_1}$, $\xi > \frac{(\varrho_1+2)\kappa}{\varrho_1}$ *and* $\gamma_n = \frac{1}{2}c_1^2 K_3^2 n^{2\tau}$ *for some* $\tau \in (\max\{\frac{1}{2} - \eta - \frac{\kappa}{2\varrho_1}, 0\}, \frac{1}{2} - \frac{(\varrho_1+1)\kappa}{\varrho_1})$. *If* $\inf_{u \in [a,b]} \mathbb{E}(Y^2 | X_j = u) \ge \rho$ *for some positive* ρ *holds for any* $j \notin \mathcal{M}_*$, *then*

$$\mathbb{P}(|\widehat{\mathcal{M}}_{\gamma_n}| > s) \le p \exp\bigl(-C_3 n^{\min\{\eta\gamma, (\varrho_1-\kappa)\gamma/(\varrho_1\varrho_2), 2\tau, (\varrho_1-\kappa)\gamma/(6\varrho_1)\}}\bigr),$$

where C_3 *is given in Corollary* 2.

This theorem shows that our screening procedure well controls the set size of the recruited variables. With large probability, the number of the recruited variables is not larger than the true size s. From Theorems 1 and 2 with $w = \frac{\kappa}{\varrho_1}$, we have that $\mathbb{P}(\widehat{\mathcal{M}}_{\gamma_n} = \mathcal{M}_*) \to 1$ as $n \to \infty$ provided that $\log p = o(n^{\min\{\eta\gamma, (\varrho_1-\kappa)\gamma/(\varrho_1\varrho_2), 2\tau, (\varrho_1-\kappa)\gamma/(6\varrho_1), 1-2\kappa-\kappa/\varrho_1, (1/2-\kappa-\kappa/\varrho_1-\tau)\gamma\}})$. This selection consistency property demonstrates that our approach performs very well by distinguishing the true contributing variables from false ones under condition (2.9). Aiming to obtain the optimal diverging rate for p, we can select

$$\tau = \begin{cases} \dfrac{\gamma}{\gamma+2}\biggl(\dfrac{1}{2} - \kappa - \dfrac{\kappa}{\varrho_1}\biggr), & \text{if } \eta > \dfrac{1}{\gamma+2} + \dfrac{\gamma\kappa}{\gamma+2} + \dfrac{(\gamma-2)\kappa}{2(\gamma+2)\varrho_1}; \\ \dfrac{1}{2} - \eta - \dfrac{\kappa}{2\varrho_1} + \varsigma, & \text{if } \eta \le \dfrac{1}{\gamma+2} + \dfrac{\gamma\kappa}{\gamma+2} + \dfrac{(\gamma-2)\kappa}{2(\gamma+2)\varrho_1}, \end{cases}$$

where ς can be chosen to be positive and converging to 0 as $n \to \infty$. Hence, $\mathbb{P}(\widehat{\mathcal{M}}_{\gamma_n} = \mathcal{M}_*) \to 1$ as $n \to \infty$ provided that

$$\log p = \begin{cases} o\bigl(n^{\min\{(\varrho_1-\kappa)\gamma/(\varrho_1\varrho_2), \gamma(1-2\kappa-2\kappa/\varrho_1)/(\gamma+2), (\varrho_1-\kappa)\gamma/(6\varrho_1)\}}\bigr), \\ \qquad \text{if } \eta > \dfrac{1+\gamma\kappa}{\gamma+2} + \dfrac{(\gamma-2)\kappa}{2(\gamma+2)\varrho_1}; \\ o\bigl(n^{\min\{(\varrho_1-\kappa)\gamma/(\varrho_1\varrho_2), (\varrho_1-\kappa)\gamma/(6\varrho_1), (\eta-\kappa-\kappa/(2\varrho_1))\gamma\}}\bigr), \\ \qquad \text{if } \eta \le \dfrac{1+\gamma\kappa}{\gamma+2} + \dfrac{(\gamma-2)\kappa}{2(\gamma+2)\varrho_1}. \end{cases}$$

More specifically, if the response variable Y has a compact support which means $\gamma = \infty$, the above selection consistency holds if $\log p = o(n^{1-2\kappa-2\kappa/\varrho_1})$. Additionally, the smoothness of the projections f_j's also affects the allowable dimensionality. When all $f_j \in C^\infty(\mathcal{X})$ implying that $\varrho_1 = r = \infty$, the allowable dimensionality turns out to be $\log p = o(n^{1-2\kappa})$. If Y follows a normal or sub-Gaussian distribution that $\gamma = 2$ and $\eta = \infty$ which can be guaranteed by partial orthogonal condition [Huang, Horowitz and Wei (2010)], the selection consistency holds if $\log p = o(n^{\min\{1/2-\kappa-\kappa/\varrho_1, 1/3-\kappa/(3\varrho_1)\}})$. It is worthwhile to note that though we show that our approach can identify the set of contributing variables with probability tending to 1, practical performances can vary because first the results are valid asymptotically, and second choosing the threshold level γ_n to achieve the perfect variable selection is difficult.

3. Applications to some special models. Our local independence feature screening approach does not require a specific form of the underlying model. Now we elaborate how the proposed approach can be applied in three families of popular nonparametric and semiparametric models: the nonparametric additive models, the single-index models and multiple-index models, and varying coefficient models; and we also compare our results with existing ones.

3.1. *Nonparametric additive models.* The nonparametric additive model introduced by Stone (1985) has the form $Y = \sum_{j=1}^{p} s_j(X_j) + \varepsilon$, where $s_1(\cdot), \dots, s_p(\cdot)$ are unknown functions with zero mean and $\mathbb{E}(\varepsilon|\mathbf{X}) = 0$. It is a special case of model (2.1) with $m(\mathbf{X}) = \sum_{j=1}^{p} s_j(X_j)$. For this model, the true model can be defined as $\mathcal{M}_* = \{1 \le j \le p : \mathbb{E}\{s_j^2(X_j)\} > 0\}$. Recall that $f_j(x) = \mathbb{E}(Y|X_j = x)$ following the earlier discussion in Section 2. If we identify the true model by

$$\min_{j\in\mathcal{M}_*} \|f_j\|_\infty \ge c_1 n^{-\kappa} \tag{3.1}$$

for some nonnegative κ, our local independence feature screening procedure proposed in Section 2 can be directly applied here to surely identify the contributing explanatory variables.

Let us carefully compare our approach with the one in Fan, Feng and Song (2011) that specifically targets at the feature screening problem for nonparametric additive models. The screening procedure of Fan, Feng and Song (2011) includes four steps. First, each f_j is expanded by using the B-spline basis functions $\{\Psi_{j,k}\}_{k=1}^{\infty}$ and use the truncated version $\tilde{f}_{nj} = \sum_{k=1}^{d_n} \beta_{j,k}\Psi_{j,k}$ to approximate the projection f_j. Second, to estimate the coefficients $\beta_{j,k}$'s and obtain the corresponding estimation of each projection f_j. Third, to estimate each $\mathbb{E}\{f_j^2(X_j)\}$ by $n^{-1}\sum_{i=1}^{n} \hat{f}_{nj}^2(X_{ij})$ where $\hat{f}_{nj}$ is the estimation of f_j obtained in the second step. Fourth, to screen features via the corresponding magnitudes of these estimates. To ensure the sure screening property, they assume that each f_j belongs to the class of functions whose rth derivative satisfies the Lipschitz continuity of order $\theta \in (0, 1]$ and $d = r + \theta > 0.5$, where r is a nonnegative integer, and identify the true model by the condition

$$\min_{j\in\mathcal{M}_*} \mathbb{E}\{f_j^2(X_j)\} \ge C d_n n^{-2\tilde{\kappa}} \tag{3.2}$$

for some positive constant C. Here, $0 < \tilde{\kappa} < \frac{d}{2d+1}$ and d_n is the truncation parameter used in approximating each f_j which satisfies $d_n \ge C n^{2\tilde{\kappa}/(2d+1)}$ for some positive constant C. By Theorem 1 of Fan, Feng and Song (2011), the sure screening property holds for their procedure if $\log p = o(n^{1-4\tilde{\kappa}} d_n^{-3})$. When $d_n = O(n^{2\tilde{\kappa}/(2d+1)})$, (3.2) implies $\min_{j\in\mathcal{M}_*} \mathbb{E}\{f_j^2(X_j)\} \ge Cn^{-4d\tilde{\kappa}/(2d+1)}$ for some positive constant C. Actually, if the density of each X_j is uniformly bounded away from zero, these conditions are sufficient for the identification condition of our approach given in (3.1) to hold with $\kappa = \frac{2d\tilde{\kappa}}{2d+1}$. Fan, Feng and Song (2011) also assume that the error ε satisfies $\mathbb{E}\{\exp(B|\varepsilon|)|\mathbf{X}\} \le C$ for some positive constants B and C which implies there exist two positive constants b_1 and b_2 such that $\mathbb{P}(|\varepsilon| \ge u) \le b_1 \exp(-b_2 u)$ for any $u > 0$. See Lemma 2.2 in Petrov (1995). On the other hand, they also assume that $\|m\|_\infty \le \widetilde{B}$ for some positive constant $\widetilde{B}$. These two conditions together imply that $\gamma = 1$ in our assumption (A.5). In this case, the sure screening property of our approach given in Theorem 1 holds if $\log p = o(n^{1/2-\kappa-\kappa/\varrho_1}) = o(n^{1/2-2d\tilde{\kappa}(1+1/\varrho_1)/(2d+1)})$. When $d + \frac{1}{2} > (10 + 6d - \frac{2d}{\varrho_1})\tilde{\kappa}$, their procedure can handle faster diverging p than that of ours; otherwise, our method is stronger than theirs. This shows that our procedure can handle faster diverging p when the signal strength level is weak, that is, κ is large.

In the specific case when the number of basis functions $d_n = O(n^{1/(2d+1)})$ which leads to the optimal rate for the B-spline approach [Stone (1985)], their approach can handle the dimensionality $\log p = o(n^{1-4\tilde{\kappa}-3/(2d+1)})$. For such a selection of d_n, the allowable dimensionality of our approach under which the sure screening property holds is $\log p = o(n^{1/2+(1+1/\varrho_1)\{1/(4d+2)-\tilde{\kappa}\}})$. To make the approach of Fan, Feng and Song (2011) work with a high-dimensional setting for such a selection of d_n, the smooth parameter d should be larger than 1 implying the existence of the first derivation of each f_j, which is not required in our approach. From this point of view, our approach can handle the situation where each nonparametric marginal projection f_j does not have the first derivative but being just continuous. When $d > 1$, the parameter r in (A.1) and (A.2) satisfies $1 \le r < d \le r + 1$. If the minimum signal does not diminish to 0 [Huang, Horowitz and Wei (2010), Lin and Zhang (2006)], then $\tilde{\kappa} = 0$. In this case, our approach and their approach can handle the nonpolynomial dimensionality $\log p = o(n^{1/2+(1+1/r)/(4d+2)})$ and $\log p = o(n^{1-3/(2d+1)})$, respectively. If $2d \le 6 + r^{-1}$, our approach allows faster diverging p than that of their approach; otherwise, their result is stronger than ours. This can be viewed as a price paid for our approach by allowing weaker requirement on the continuity of each f_j and without requiring $m(\mathbf{X})$ to be bounded.

We note that the above diverging rates comparison is established in a case in favor of the approach of Fan, Feng and Song (2011) by using their identification condition with the smallest d_n. If d_n diverges faster than $n^{2\tilde{\kappa}/(2d+1)}$, the parameter κ appeared in our identification condition will be smaller than $\frac{2d\tilde{\kappa}}{2d+1}$ and the allowable dimensionality of our approach will be improved. Additionally, our approach has a very good control of the size of the recruited variables. From Theorem 2, the set of the recruited variables is not larger than the true contributing covariates with large probability, which together with Theorem 1, imply the selection consistency of our approach in nonparametric additive models.

3.2. *Single-index and multiple-index models.* The single-index model that is recognized as a particularly useful variation of the linear regression model has the form $Y = s(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}) + \varepsilon$, where $s(\cdot)$ is the conditional mean function that is not explicitly specified; see Brillinger (1983) for more details. This kind of models is a special case of model (2.1) with $m(\mathbf{X}) = s(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})$. Since single-index model requires identifiability condition [Brillinger (1983)], parameters in marginal single-index models is not identifiable. Therefore, marginal estimator-based ranking procedure cannot be applied to handle the feature screening problem in the single-index model.

In this case, our local independence feature screening approach conveniently applies, because our marginal empirical likelihood based approach does not require estimating parameters in the marginal models. Since our marginal empirical likelihood approach is capable of assessing $\mathbb{E}(Y|X_j) \equiv 0$, identifying the parameter in a marginal single-index model is not necessary for our local independence feature screening approach. Specifically, if we identify the true model by $\min_{j\in\mathcal{M}_*} \|f_j\|_\infty \geq c_1 n^{-\kappa}$, then the local independence feature screening procedure and its properties discussed in Section 2 also directly apply here. The effective application of our approach in single-index models demonstrates that the marginal empirical likelihood based approach is advantageous in independence feature screening. Such a merit is due to the new insight of the marginal empirical likelihood approach for screening variables by assessing the evidence against the null hypothesis that the explanatory variable is not contributing marginally.

More generally, we may consider the screening for multiple-index model $Y = s(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{X}, \dots, \boldsymbol{\beta}_K^{\mathrm{T}}\mathbf{X}) + \varepsilon$, where K is a known integer, $\boldsymbol{\beta}_k$ $(k = 1, \dots, K)$ are sparse vectors of unknown parameters, and s is an unknown function. The marginal condition for the jth component of $\mathbf{X}$ given by (2.2) still applies in the multiple-index models. Nevertheless, in multiple-index models, identification is also an issue, which is actually more complicated and challenging than that in the single-index models when considering the marginal contributions. Since our concern in this model can also be transformed to assessing $\mathbb{E}(Y|X_j) \equiv 0$ or not, the local independence feature screening procedure given in the last section can also be applied.

3.3. *Varying coefficient models.* Varying coefficient model is useful for studying the variable-dependent effects in the regressions. Many methods have been proposed for estimation of this model. See, for example, Fan and Zhang (2000) for the local polynomial smoothing method, Huang, Wu and Zhou (2002) and Qu and Li (2006) for basis expansion and spline method. The varying coefficient model has the following form:

$$Y = \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}(Z) + \varepsilon, \tag{3.3}$$

where $\mathbf{X} = (X_1, \dots, X_p)^{\mathrm{T}}$ is a $p \times 1$ vector of explanatory variables, Z is a scalar variable that takes values on a compact interval $\mathcal{Z}$, and ε is the error satisfies $\mathbb{E}(\varepsilon|\mathbf{X}, Z) = 0$ almost surely. Here, $\boldsymbol{\beta}(z) = (\beta_1(z), \dots, \beta_p(z))^{\mathrm{T}}$ is unknown but smooth in z. One way to screen explanatory variables could be ignoring the impact due to Z, and applying some marginal approaches. However, it is not difficult to see that there are situations, for example, when $\int \beta(z)\,dz = 0$, that is, the parameter effect is zero in an average, univariate screening procedure ignoring Z will not be able to identify the component in $\mathbf{X}$ even if it is contributing.

To overcome such a difficulty, we may adjust the marginal nonparametric regression for varying coefficient models to incorporate the effect of Z. The marginal version of (3.3) is

$$Y = a_j(Z) + X_j b_j(Z) + \tilde{\varepsilon}_j, \tag{3.4}$$

where $\mathbb{E}(\tilde{\varepsilon}_j|X_j, Z) = 0$ almost surely for $j = 1, \dots, p$. Here, $b_j(Z)$ can be interpreted as the marginal contribution of X_j in explaining Y via Z. It can be shown that $b_j(Z) = \frac{\mathrm{cov}(X_j, Y|Z)}{\mathrm{var}(X_j|Z)}$. Thus, the marginal effect $b_j(z) = 0$ for any z is equivalent to $\mathrm{cov}(X_j, Y|Z = z) = 0$ for any z. In the simple case when $\mathbb{E}(Y|Z) = 0$ or some constant free of Z, it is equivalent to $\mathrm{cov}(X_j, Y|Z) = \mathbb{E}(X_j Y|Z) \equiv 0$, which essentially share the same form $\mathbb{E}(Y|X_j) \equiv 0$ as in the local independence featuring screening. In the general situation, $\mathbb{E}(Y|Z) \neq 0$ so that one needs to assess $0 \equiv \mathrm{cov}(X_j, Y|Z) = \mathbb{E}[X_j\{Y - \mathbb{E}(Y|Z)\}|Z]$ with the nuisance function $\mathbb{E}(Y|Z)$ estimated. For a kernel function $\widetilde{\mathcal{K}}(\cdot)$, $\mathbb{E}(Y|Z = z)$ can be estimated by

$$\widehat{\mathbb{E}}(Y|Z = z) = \frac{n^{-1}\sum_{i=1}^n \widetilde{\mathcal{K}}_{\tilde{h}}(Z_i - z)Y_i}{n^{-1}\sum_{i=1}^n \widetilde{\mathcal{K}}_{\tilde{h}}(Z_i - z)}, \tag{3.5}$$

where $\widetilde{\mathcal{K}}_{\tilde{h}}(u) = \tilde{h}^{-1}\widetilde{\mathcal{K}}(u\tilde{h}^{-1})$ with bandwidth $\tilde{h}$. Let $\widetilde{Y}_i = Y_i - \widehat{\mathbb{E}}(Y|Z = Z_i)$, then the marginal empirical likelihood is constructed as

$$\mathrm{EL}_j(z, 0) = \sup\left\{\prod_{i=1}^n w_i : w_i \geq 0, \sum_{i=1}^n w_i = 1, \sum_{i=1}^n w_i \mathcal{K}_h(Z_i - z) X_{ij} \widetilde{Y}_i = 0\right\}. \tag{3.6}$$

We then propose using

$$\ell_j(0) = \sup_{z \in \mathcal{Z}_n} \ell_j(z, 0), \tag{3.7}$$

where $\ell_j(z, 0) = -2\log\{\mathrm{EL}_j(z, 0)\} - 2n\log n$, and $\mathcal{Z}_n$ is a partition of $\mathcal{Z}$. In practice, a natural choice is to evaluate the statistic (3.7) by $\ell_j(0) = \max_{1 \leq i \leq n} \ell_j(Z_i, 0)$ where $\{Z_i\}_{i=1}^n$ are the n observations of variable Z. This new $\ell_j(0)$ can be used for local independence feature screening for varying coefficient models analogously to that in Section 2.

The analogous assumptions corresponding to (A.1)–(A.5) in Section 2 are given as follows.

(A.1)$'$ For each $j = 1, \dots, p$, let $\tilde{f}_j(z) = \operatorname{cov}(X_j, Y|Z = z)$. Assume $\{\tilde{f}_j\}_{j=1}^p$ belong to $C^r(\mathcal{Z})$. If $r = 0$, we assume that $\tilde{f}_j$'s satisfy the Lipschitz condition with order $\alpha \in (0, 1]$, that is, $|\tilde{f}_j(s) - \tilde{f}_j(t)| \le \widetilde{K}_1 |s - t|^\alpha$ for any $s, t \in \mathcal{Z}$, where $\widetilde{K}_1$ is a positive constant uniformly for any $j = 1, \dots, p$. In addition, there exists a constant $\widetilde{K}_2$ such that $|\tilde{f}_j^{(r)}(z)| \le \widetilde{K}_2$ for any $z \in \mathcal{Z}$ and $j = 1, \dots, p$.

(A.2)$'$ The density function g of Z satisfies $0 < \widetilde{K}_3 \le g(z) \le \widetilde{K}_4 < \infty$ on $\mathcal{Z}$. In addition, we assume that g belongs to $C^r(\mathcal{Z})$ for the r given in (A.1)$'$ and $|g^{(r)}(z)| \le \widetilde{K}_5$ for any $z \in \mathcal{Z}$.

(A.3)$'$ There exist nonnegative constants $\tilde{c}_1 > 0$ and $\kappa \in [0, \frac{\max\{r,\alpha\}}{2\max\{r,\alpha\}+2})$ such that $\min_{j \in \mathcal{M}_*} \|\tilde{f}_j\|_\infty \ge \tilde{c}_1 n^{-\kappa}$, where r and α are specified in (A.1)$'$.

(A.4)$'$ There exists some positive constant ξ such that $\|\mathcal{Z}_n\| = n^{-\xi}$.

(A.5)$'$ There exist positive constants $\widetilde{K}_6, \widetilde{K}_7, \gamma_1$ and γ_2 such that $\mathbb{P}(|Y| \ge u) \le \widetilde{K}_6 \exp(-\widetilde{K}_7 u^{\gamma_1})$ and $\mathbb{P}(|X_j| \ge u) \le \widetilde{K}_6 \exp(-\widetilde{K}_7 u^{\gamma_2})$ for any $u > 0$ and $j = 1, \dots, p$.

Assumption (A.5)$'$ and Lemma 2 in Chang, Tang and Wu (2013b) yield that $\mathbb{P}(|X_j Y| \ge u) \le 2\widetilde{K}_6 \exp(-\widetilde{K}_7 u^\gamma)$ for any $u > 0$ and $j = 1, \dots, p$, where $\gamma = \frac{\gamma_1\gamma_2}{\gamma_1+\gamma_2}$. To investigate the theoretical properties of (3.7) with nonparametric estimation (3.5), we need the following two extra conditions:

(A.7) $\mathbb{E}(Y|Z = z)$ belongs to $C^r(\mathcal{Z})$, where r is given in (A.1)$'$. If $r = 0$, we assume $\mathbb{E}(Y|Z = z)$ satisfies the Lipschitz condition with order α where α is specified in (A.1)$'$.

(A.8) For r specified in (A.1)$'$, if $r \ge 1$, the kernel function $\widetilde{\mathcal{K}}(\cdot)$ is of order r. If $r = 0$, the kernel function satisfies $\widetilde{\mathcal{K}}(u) \ge 0$ for any u and $\int \widetilde{\mathcal{K}}(u)\,du = 1$.

For $\gamma = \frac{\gamma_1\gamma_2}{\gamma_1+\gamma_2}$, we define ϱ_1, ϱ_2 and δ as (2.8). In addition, let $\delta_1 = \max\{\frac{2}{\gamma_1} - 1, 0\}$. Meanwhile, we assume that the bandwidths h and $\tilde{h}$ in constructing marginal empirical likelihood (3.6) and the NW estimator (3.5) for $\mathbb{E}(Y|Z = z)$ satisfy $h \asymp n^{-w}$ and $\tilde{h} \asymp n^{-\phi}$, where w and ϕ will be specified later. The property of the marginal empirical likelihood for varying coefficient models are given in the following theorem.

THEOREM 3. *Under assumptions* (A.1)$'$–(A.5)$'$, (A.6) *and* (A.7)–(A.8), *pick* $w \in [\frac{\kappa}{\varrho_1}, \frac{1}{2} - \kappa)$, $\phi \in (\frac{\kappa}{\varrho_1}, 1 - 2\kappa)$, $\xi > \kappa + 2w$, *and* $\gamma_n = \frac{1}{2}\tilde{c}_1^2 \widetilde{K}_3^2 n^{2\tau}$ *for some* $\tau \in (0, \frac{1}{2} - \kappa - w)$, *then there exists a uniform positive constant* C_4 *such that*

$$\begin{aligned}\mathbb{P}(\mathcal{M}_* \subset \widehat{\mathcal{M}}_{\gamma_n}) \ge 1 &- s\exp\{-C_4 n^{(1/2-\kappa-w-\tau)\gamma}\} - s\exp\{-C_4 n^{(\phi\varrho_1-\kappa)\gamma_2}\} \\ &- s\exp\bigl(-C_4 n^{\min\{1-2\kappa-w, \frac{1-\kappa-w}{1+\delta}\}}\bigr) \\ &- s\exp\bigl(-C_4 n^{\min\{\frac{\{3-2(\phi+\kappa+w+\tau)\}\gamma_2}{(2+2\delta_1)\gamma_2+2}, \frac{\{2-\phi-2(\kappa+w+\tau)\}\gamma_2}{\gamma_2+2}\}}\bigr) \\ &- s\exp\bigl(-C_4 n^{\min\{\frac{(1-2\kappa-\phi)\gamma_2}{\gamma_2+2}, \frac{(1-\kappa-\phi)\gamma_2}{(1+\delta_1)\gamma_2+1}\}}\bigr).\end{aligned}$$

Theorem 3 provides a general result for the sure screening property of the marginal empirical likelihood for varying coefficient models. The first two terms and the fourth term on the right-hand side of above inequality are the same as those in Theorem 1. The extra terms are due to the kernel estimation of $\mathbb{E}(Y|Z)$. The analogues of Proposition 1 and Theorem 2 are also valid for varying coefficient models using the above marginal empirical likelihood.

Fan, Ma and Dai (2014), Liu, Li and Wu (2014) and Song, Yi and Zou (2014) also consider the feature screening for ultra-high dimensional varying coefficient models. Fan, Ma and Dai (2014) and Liu, Li and Wu (2014) considered the same model as (3.3) while Song, Yi and Zou (2014) allows both Y and $\mathbf{X}$ to depend on Z. Fan, Ma and Dai (2014) estimate $a_j(\cdot)$ and $b_j(\cdot)$ in (3.4) simultaneously via the B-spline basis functions expansion approach. Fan, Ma and Dai (2014) propose an estimator $\hat{u}_j$ for $\mathbb{E}\{\frac{\operatorname{cov}^2(X_j,Y|Z)}{\operatorname{var}(X_j|Z)}\}$ for each $j = 1,\dots,p$ measuring the marginal contribution of X_j. Then they propose to select the covariates via ranking $|\hat{u}_j|$'s. Liu, Li and Wu (2014) study conditional Pearson correlation as a measure of marginal contribution for varying coefficient models. For each $j = 1,\dots,p$ and z, they estimate conditional Pearson correlation $\rho(X_j, Y|Z = z)$ via kernel smoothing method and construct an estimation $\hat{\rho}_j^*$ for $\mathbb{E}\{\rho^2(X_j, Y|Z)\}$. Then they use the magnitude of $|\hat{\rho}_j^*|$ to determine whether X_j is an important explanatory variable or not. The main idea of Song, Yi and Zou (2014) is similar to that of Fan, Ma and Dai (2014). In the sequel, we carefully compare our procedure with those proposed in Fan, Ma and Dai (2014) and Liu, Li and Wu (2014).

Fan, Ma and Dai (2014) identify $\mathcal{M}_*$ via $\min_{j\in\mathcal{M}_*}\mathbb{E}\{\operatorname{cov}^2(X_j,Y|Z)\} \ge Cd_n n^{-2\tilde{\kappa}}$ for some positive constants C and $\tilde{\kappa} < \frac{2d+1}{8d+10}$. Here, d_n is the number of approximation terms to $a_j(\cdot)$ and $b_j(\cdot)$ in the estimation step. Details of d_n can be found in Section 3.1 for discussion of additive models. To guarantee the validity of the approach in Fan, Ma and Dai (2014), $d_n \ge Cn^{2\tilde{\kappa}/(2d+1)}$ is required for some positive constant C and $d = r + \theta$, where r is an integer and $\theta \in (0, 1]$ are employed to describe the smoothness of each a_j and b_j, that is, the rth derivatives of all a_j and b_j are Lipschitz continuous of order θ. In Liu, Li and Wu (2014), the identification condition is given by $\min_{j\in\mathcal{M}_*}\mathbb{E}\{\rho^2(X_j,Y|Z)\} \ge Cn^{-2\bar{\kappa}}$ for some positive constants C and $\bar{\kappa} < \hbar$. Here, $\hbar < \frac{1}{3}$ is a parameter employed to describe the decay rate of the bandwidth for the kernel smoothing step in their procedure, that is, the bandwidth is chosen as $O(n^{-\hbar})$. Based on the moments condition and the assumptions $\inf_{z\in\mathcal{Z}}\min_{1\le j\le p}\operatorname{var}(X_j|Z = z) > 0$ and $\inf_{z\in\mathcal{Z}}\operatorname{var}(Y|Z = z) > 0$, the identification condition in Liu, Li and Wu (2014) is essentially equivalent to $\min_{j\in\mathcal{M}_*}\mathbb{E}\{\operatorname{cov}^2(X_j,Y|Z)\} \ge Cn^{-2\bar{\kappa}}$ for some positive constant C.

From three aspects, we compare our identification condition and theoretical results with those of Fan, Ma and Dai (2014) and Liu, Li and Wu (2014). First, if the density of Z is uniformly bounded away from zero and infinity on its support $\mathcal{Z}$, their L_2-type requirement is a sufficient condition for ours proposed in assumption (A.3)$'$. But their identification conditions rule out the case where

$\mathrm{cov}(X_j, Y|Z = z)$ only contribute largely at several local small intervals on $\mathcal{Z}$. Second, our method works for weaker signal strength than theirs. Fan, Ma and Dai (2014) can handle the case that $[\mathbb{E}\{\mathrm{cov}^2(X_j, Y|Z)\}]^{1/2} \geq Cn^{-d\tilde{\kappa}/(2d+1)}$ for some $\tilde{\kappa} < \frac{2d+1}{8d+10}$. Therefore, the weakest signal strength can be handled by their method cannot decay at a rate faster than $O(n^{-d/(8d+10)})$. On the other hand, the weakest signal strength our method can handle is at the rate of $O(n^{-\varrho_1/(2\varrho_1+2)})$. By noting that $r < d \leq r+1$, we have $\frac{\varrho_1}{2\varrho_1+2} > \frac{d}{8d+10}$ which implies our method can accommodate weaker signal strength than Fan, Ma and Dai (2014). If condition (C4) of Liu, Li and Wu (2014) holds, the parameter r in our notation system is not smaller than 2. Thus, $\frac{\varrho_1}{2\varrho_1+2} \geq \frac{1}{3}$ which implies our method can also accommodate weaker signal strength than that of Liu, Li and Wu (2014). Third, we compare the nonpolynomial dimensionality allowed by different methods. As our theoretical assumptions are close to that proposed in Liu, Li and Wu (2014), we compare our result with theirs. When $\gamma_1 = \gamma_2 = r = 2$, which are assumed in Liu, Li and Wu (2014), from Corollary 3, our method can handle $\log p = o(n^{\min\{4\phi-2\kappa, 1/2-\kappa-\phi/2\}})$ when τ is chosen to be close enough to zero and $w = \frac{\kappa}{\varrho_1} = \frac{\kappa}{2}$, where the result for the method of Liu, Li and Wu (2014) is $\log p = o(n^{\hbar-\kappa})$ for some $\hbar < \frac{1}{3}$. Notice that $\kappa < \hbar < \frac{1}{3}$ in their setting. If we choose $\phi = \frac{1}{5}$, then $\min(4\phi - 2\kappa, \frac{1}{2} - \kappa - \frac{\phi}{2}) = \frac{2}{5} - \kappa > \hbar - \kappa$, which means our procedure can accommodate faster diverging p than theirs.

4. Iterative feature screening. It is possible that some predictors are marginally unrelated but jointly related to the response as illustrated by Example 4.2.2 of Fan and Lv (2008). As observed in the literature, marginal utility-based feature screening methods may miss this type of predictors badly [Fan and Lv (2008), Fan, Samworth and Wu (2009)]. To overcome this difficulty, several versions of iterative feature screening have been proposed. Fan and Lv (2008) proposed to regress the response on the recruited predictors and use the regression residual as the new “response” to recruit further from the remaining predictors. While recruiting additional predictors, Fan, Samworth and Wu (2009) consider the joint model with both the recruited predictors and each additional predictor and use the conditional contribution of each additional predictor given those predictors that are already selected to recruit further from the remaining predictors. Both versions need to fit a joint model on the recruited predictors with or without an additional feature where a parametric model specification is required. However our proposed empirical likelihood-based screening is for a general nonparametric model (2.1) and, in this sense, model-free. Consequently, the aforementioned two versions of iterative feature screening cannot be extended to our case. Next, we will borrow the idea of Zhu et al. (2011) and propose an iterative version for our proposed empirical likelihood-based screening.

Next, we detail our iterative screening procedure.

Step 1. We first apply our proposed empirical likelihood based screening to $\{(\mathbf{X}_i, Y_i), i = 1, \dots, n\}$ and denote the selected set of predictors by $\widehat{\mathcal{A}}_1$.

Step 2. Apply COSSO [Lin and Zhang (2006)] to data $\{(\mathbf{X}_{i\widehat{\mathcal{A}}_1}, Y_i), i = 1, \dots, n\}$ and use $\widehat{\mathcal{M}}_1$ to denote the subset of $\widehat{\mathcal{A}}_1$ that are retained by the COSSO.

Step 3. For any $j \notin \widehat{\mathcal{M}}_1$, we regress X_j on the predictors with indices in $\widehat{\mathcal{M}}_1$ based on the data $\{(\mathbf{X}_{i\widehat{\mathcal{M}}_1}, X_{ij}), i = 1, \dots, n\}$ and denote the residual by $\hat{\varepsilon}_{ij}$, $i = 1, \dots, n$. For any $j \notin \widehat{\mathcal{M}}_1$, treat $\hat{\varepsilon}_{ij}$, $i = 1, \dots, n$ as the pseudo predictor and apply our proposed empirical likelihood-based screening to recruit a subset $\widehat{\mathcal{A}}_2$ of predictors.

Step 4. Apply COSSO to data $\{(\mathbf{X}_{i\widehat{\mathcal{M}}_1 \cup \widehat{\mathcal{A}}_2}, Y_i), i = 1, \dots, n\}$ and use $\widehat{\mathcal{M}}_2$ to denote the subset of $\widehat{\mathcal{M}}_1 \cup \widehat{\mathcal{A}}_2$ that are retained by the COSSO.

Step 5. Repeat Steps 3 and 4 until $\widehat{\mathcal{M}}_k = \widehat{\mathcal{M}}_{k-1}$ or the number of predictors in $\widehat{\mathcal{M}}_k$ reaches some prespecified number.

5. Numerical examples.

We next demonstrate the performance of the local independence feature screening methods using five examples with comparisons to appropriate existing alternative approaches. In what follows, we denote our method by EL when reporting the results. When implementing the local independence feature screening approach, we select the bandwidth h by the cross-validation method [Fan and Gijbels (1996)], and the Epanechnikov kernel $K(u) = \frac{3}{4}(1 - u^2)I(|u| \le 1)$ is applied for the marginal regressions.

EXAMPLE 1. This example is taken from Example 3 of Fan, Feng and Song (2011). Data are generated from model $Y = 5g_1(X_1) + 3g_2(X_2) + 4g_3(X_3) + 6g_4(X_4) + \sigma\varepsilon$ with $g_1(x) = x$, $g_2(x) = (2x - 1)^2$, $g_3(x) = \frac{\sin(2\pi x)}{2-\sin(2\pi x)}$, and $g_4(x) = 0.1\sin(2\pi x) + 0.2\cos(2\pi x) + 0.3\sin^2(2\pi x) + 0.4\cos^3(2\pi x) + 0.5\sin^3(2\pi x)$. Here, predictors X_j's are i.i.d. random variables of Unif(0, 1) distribution, and $\varepsilon \sim N(0, 1)$ is independent of X_j's. We set $p = 1000$. Data sets of size $n = 400$ are used. We apply the proposed screening method to reduce the number of predictors from 1000 to 20. We consider different signal noise ratios by varying σ^2 at four different levels while Fan, Feng and Song (2011) chose a specific value of $\sigma^2 = 1.74$. Table 1 reports the frequency of the important predictors being selected among 100 repetitions for different values of σ^2. A comparison is made with Fan, Feng and Song (2011). It is observed that the proposed empirical likelihood-based local independence feature screening performs similarly as the method proposed by Fan, Feng and Song (2011). Contributing features X_1, X_2 and X_4 are selected by both methods for all repetitions. Yet for feature X_2 with a slightly weaker contribution, our method does slightly better.

EXAMPLE 2. This is a nonlinear predictor effect example with heterogeneous conditional variance. Data are generated from model $Y = -3h_1(X_1) +$

TABLE 1
Simulation results for Example 1

σ^2	**Method**	X_1	X_2	X_3	X_4
1	EL	100	97	100	100
	Fan, Feng and Song (2011)	100	90	100	100
	Zhu et al. (2011)	100	1	100	100
	Li, Zhong and Zhu (2012)	100	27	100	100
	Chang, Tang and Wu (2013a)	100	1	100	100
1.74	EL	100	96	100	100
	Fan, Feng and Song (2011)	100	88	100	100
	Zhu et al. (2011)	100	3	100	100
	Li, Zhong and Zhu (2012)	100	24	100	100
	Chang, Tang and Wu (2013a)	100	1	100	100
2	EL	100	95	100	100
	Fan, Feng and Song (2011)	100	87	100	100
	Zhu et al. (2011)	100	2	100	100
	Li, Zhong and Zhu (2012)	100	24	100	100
	Chang, Tang and Wu (2013a)	100	1	100	100
3	EL	100	93	100	100
	Fan, Feng and Song (2011)	100	85	100	100
	Zhu et al. (2011)	100	1	100	100
	Li, Zhong and Zhu (2012)	100	20	100	100
	Chang, Tang and Wu (2013a)	100	1	100	100

$2.5h_2(X_2) - 2h_3(X_3) + 1.5h_4(X_4) + \sigma\varepsilon$ with $h_1(x) = (2x - 1)^2$, $h_2(x) = \frac{\cos(2\pi x)}{2+\sin(2\pi x)}$, $h_3(x) = \frac{\cos(2\pi x)}{2-\cos(2\pi x)}$, and $h_4(x) = \cos\{(2x - 1)\pi\}$. Here, X_j's are independent and uniform over [0, 1], ε is independent of X_j's and has normal distribution with mean zero and its heterogeneous conditional variance is generated by $\operatorname{var}(\varepsilon) = \frac{4}{x_1^2+x_2^2+x_3^2+x_4^2}$. The noise level is govern by σ with different values in the simulations. We set $p = 1000$ and $n = 100$. We apply the proposed screening method to reduce the number of predictors from 1000 to 20. For comparison purposes, we also apply the methods on data generated from a model with homogenous conditional variance while other settings are the same. The results are reported in Table 2 for the two cases. There are a few observations. First, those methods incorporating less local impact perform poorly in this nonlinear effect example, demonstrating the importance and substantial effect of the local feature of the marginal contributions to the response variable. Second, in this example with heterogeneous conditional variance, our method outperforms others, and is better than the one of Fan, Feng and Song (2011), especially when the noise level is relatively higher implying the signal is relatively weaker. This is consistent with our theory and the finding from Example 1, and it shows that our method is advantageous for detecting nonlinear effects. It also demonstrates that when the signal is weak, and when the situation is more difficult due to high level and more complex

TABLE 2
Simulation results for Example 2 with σ^2 controlling the overall noise level

		Homogeneous variance				Heterogeneous variance			
σ^2	Method	X_1	X_2	X_3	X_4	X_1	X_2	X_3	X_4
0.5	EL	99	97	97	100	83	90	81	94
	Fan, Feng and Song (2011)	94	99	90	100	76	86	78	92
	Zhu et al. (2011)	3	4	4	4	4	6	2	2
	Li, Zhong and Zhu (2012)	36	59	38	68	23	43	22	47
	Chang, Tang and Wu (2013a)	3	4	1	2	4	6	0	0
1.0	EL	94	98	89	99	66	74	67	85
	Fan, Feng and Song (2011)	88	95	86	97	53	67	58	84
	Zhu et al. (2011)	3	4	2	3	4	4	0	4
	Li, Zhong and Zhu (2012)	27	50	32	60	16	34	16	37
	Chang, Tang and Wu (2013a)	3	5	1	2	3	5	0	1
1.5	EL	88	95	87	97	57	61	48	73
	Fan, Feng and Song (2011)	81	92	81	96	43	55	41	76
	Zhu et al. (2011)	3	6	2	2	4	4	1	2
	Li, Zhong and Zhu (2012)	22	44	26	51	14	24	13	26
	Chang, Tang and Wu (2013a)	3	5	0	2	3	5	0	1
2.0	EL	84	86	82	94	46	52	38	65
	Fan, Feng and Song (2011)	77	87	79	93	34	50	35	61
	Zhu et al. (2011)	4	5	0	2	4	5	1	2
	Li, Zhong and Zhu (2012)	21	39	21	47	14	20	11	20
	Chang, Tang and Wu (2013a)	3	5	0	1	5	4	0	2

variations, our method delivers more promising results thanks to the feature of the marginal empirical likelihood approach.

EXAMPLE 3. Data are generated from model $Y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon$ with independent error $\varepsilon \sim N(0, 1)$. Jointly Gaussian covariates satisfy $\mathbb{E}(X_j) = 0$ and $\operatorname{var}(X_j) = 1$ for $j = 1, \dots, p$ with $\operatorname{cov}(X_j, X_4) = \frac{1}{\sqrt{2}}$ for $j \neq 4$ and $\operatorname{cov}(X_j, X_{j'}) = \frac{1}{2}$ if j and j' are distinct elements of $\{1, \dots, p\} \setminus \{4\}$. True regression coefficients are given by $\beta_1 = \beta_2 = \beta_3 = 2$, $\beta_4 = -3\sqrt{2}$, and $\beta_j = 0$ for $j > 4$ such that X_4 is marginally independent of the response Y. Yet X_4 is the most important predictor variable in the joint model. This example is to illustrate that the iterative version of the proposed screening procedure works effectively. We borrow the idea of Zhu et al. (2011) to define the iterative version of our screening procedure as laid out in Section 4. Fan, Feng and Song (2011) only considered the case with $(n = 400, p = 1000)$ while we consider two cases: $(n = 300, p = 1000)$ and $(n = 400, p = 2000)$. Simulation results over 100 repetitions are reported in Table 3, where we report the frequency of important predictors being selected and

TABLE 3
Simulation results for Example 3

(n, p)	Iterative screening method	X_1	X_2	X_3	X_4	Average for $x_j, j \geq 5$
(300, 1000)	EL	100	100	100	100	0.0351
	Fan, Feng and Song (2011)	100	100	100	100	0.1175
	Zhu et al. (2011)	100	100	100	100	N/A
	Li, Zhong and Zhu (2012)	100	100	100	100	N/A
	Chang, Tang and Wu (2013a)	100	100	100	99	0.0281
(400, 2000)	EL	100	100	100	100	0.0141
	Fan, Feng and Song (2011)	100	100	100	100	0.0612
	Zhu et al. (2011)	100	100	100	100	N/A
	Li, Zhong and Zhu (2012)	100	100	100	100	N/A
	Chang, Tang and Wu (2013a)	100	100	100	100	0.0220

the average frequency of unimportant predictors being selected. It shows that the iterative screening based on empirical likelihood performs similarly as the nonparametric screening proposed in Fan, Feng and Song (2011). In terms of average frequency of unimportant predictors being selected, our new method has slight advantage.

EXAMPLE 4. In this example, we consider a single-index type model. Data are generate from $Y = m(\mathbf{X}) + \sigma\varepsilon$, where $m(\mathbf{X})$ is generated from $\exp\{-\frac{1}{2}(X_1^2/0.8^2 + X_2^2/0.9^2 + X_3^2/1.0 + X_4^2/1.1^2)\}$ by appropriately scaling it to have zero mean and unit variance, predictors are independently generated from standard normal distribution and $\varepsilon \sim N(0, 1)$ is independent of X_j's. We set $p = 1000$ and $n = 100$, and vary the noise level as 0.5 and 1.0, respectively. We apply the proposed screening method to reduce the number of predictors from 1000 to 20 and compare it with the method of Fan, Feng and Song (2011) and additional two methods in Zhu et al. (2011) and Li, Zhong and Zhu (2012). In this example, the signals are strongest locally at 0 while decay exponentially fast at other locations, and X_1 is the strongest and X_4 is the weakest in their signal strength according to the coefficients. Note that the iterative screening of Fan, Feng and Song (2011) is residual-based while that of Zhu et al. (2011) is projection-based. Thus to be fair, we only compare in terms of the noniterative version. The frequencies of important predictors being selected over 100 repetitions are reported in Table 4 for different methods. We see that our method performs much better than that of Fan, Feng and Song (2011) thanks to the merit of our method in detecting local contributions. Additionally, we see that correlation based methods completely fail in this case while the distance correlation based method of Li, Zhong and Zhu (2012) can still detect signal, while our method performs the best.

TABLE 4

Simulation results for Example 4

σ	**Method**	X_1	X_2	X_3	X_4
0.5	EL	96	92	81	63
	Fan, Feng and Song (2011)	92	77	62	32
	Zhu et al. (2011)	1	1	5	3
	Li, Zhong and Zhu (2012)	58	29	26	9
	Chang, Tang and Wu (2013a)	0	1	2	5
1.0	EL	81	69	60	36
	Fan, Feng and Song (2011)	73	62	45	19
	Zhu et al. (2011)	1	0	7	4
	Li, Zhong and Zhu (2012)	37	12	17	7
	Chang, Tang and Wu (2013a)	1	1	2	4

EXAMPLE 5. We consider the varying coefficient model in this example. We generate data from model $Y = X_1\beta_1(Z) + X_2\beta_2(Z) + X_3\beta(Z) + X_4\beta(Z) + \varepsilon$, where predictors are multivariate normal with $\mathbb{E}(X_j) = 0$, $\text{var}(X_j) = 1$, and zero correlation, $\varepsilon \sim N(0, 0.1)$ is independent of X_j's, and Z is independently generated from the standard uniform distribution over [0, 1]. The varying coefficients are given by $\beta_1(z) = \sin(2\pi z + \frac{\pi}{4})$, $\beta_2(z) = \sin(2\pi z)$, $\beta_3(z) = \cos(2\pi z)$ and $\beta_4(z) = \sin(2\pi z + \frac{3\pi}{4})$. We fix the dimensionality $p = 1000$ and vary the sample size from 100 to 200. We try to reduce the dimensionality from 1000 to 20 and compare our method to Fan, Ma and Dai (2014), Liu, Li and Wu (2014) and Song, Yi and Zou (2014) in terms of the noniterative version due to the same reason as in the above example. Table 5 summarizes the results over 100 repetitions in terms of how often important predictors are selected. It shows that our methods perform competitively. Note that the methods of Fan, Ma and Dai (2014), Liu, Li and Wu (2014), Song, Yi and Zou (2014) are developed specially for the varying coefficient model.

6. Discussion. We have proposed and investigated a local independent feature screening method using the marginal empirical likelihood in conjunction with marginal kernel smoothing methods to detect contributing explanatory variables in a general model setting. We show that our method is broadly applicable in a wide class of nonparametric and semiparametric models for high-dimensional data analysis. Theory and numerical examples show that our approach works promisingly. When the minimal signal is weak or the collinearity level among the explanatory variables is high, independence feature screening methods will face substantial difficulty. How to solve the variable selection problem under such a scenario remains open, and we hope to work along this direction with the marginal empirical likelihood approach.

TABLE 5
Simulation results for Example 5

n	Method	X_1	X_2	X_3	X_4
100	EL	97	93	96	96
	Song, Yi and Zou (2014)	84	85	82	89
	Liu, Li and Wu (2014)	92	98	88	98
	Fan, Ma and Dai (2014)	93	95	97	99
	Fan, Feng and Song (2011)	13	8	9	11
	Zhu et al. (2011)	3	6	4	5
	Li, Zhong and Zhu (2012)	4	6	8	9
	Chang, Tang and Wu (2013a)	4	2	4	3
200	EL	100	100	100	100
	Song, Yi and Zou (2014)	100	100	100	100
	Liu, Li and Wu (2014)	100	100	100	100
	Fan, Ma and Dai (2014)	100	100	100	100
	Fan, Feng and Song (2011)	16	13	11	15
	Zhu et al. (2011)	5	9	3	6
	Li, Zhong and Zhu (2012)	8	15	7	7
	Chang, Tang and Wu (2013a)	2	7	1	3

Our method is based on the empirical likelihood, and thus necessarily inherits its intensive computation. Fortunately, the marginal screening methods are highly scalable by exploring the response variable's dependence on each individual predictor at a time. Consequently, they are naturally suited for parallel computing. With parallel computing, the computational intensiveness issue of our new method can be alleviated significantly, making it a practically appealing candidate method.

Acknowledgements. We thank the Co-Editor, the Associate Editor and three referees for very constructive comments and suggestions which have improved the presentation of the paper.

SUPPLEMENTARY MATERIAL

Supplement to "Local independence feature screening for nonparametric and semiparametric models by marginal empirical likelihood" (DOI: 10.1214/15-AOS1374SUPP; .pdf). This supplement contains a real data analysis and all technical proofs.

REFERENCES

BRILLINGER, D. R. (1983). A generalized linear model with "Gaussian" regressor variables. In *A Festschrift for Erich L. Lehmann* (P. J. Bickel, K. A. Doksum and J. L. Hodges, eds.) 97–114. Wadsworth, Belmont, CA. MR0689741

BÜHLMANN, P. and VAN DE GEER, S. (2011). *Statistics for High-Dimensional Data*: *Methods*, *Theory and Applications*. Springer, Heidelberg. MR2807761

CHANG, J., CHEN, S. X. and CHEN, X. (2015). High dimensional generalized empirical likelihood for moment restrictions with dependent data. *J. Econometrics* **185** 283–304. MR3300347

CHANG, J., TANG, C. Y. and WU, Y. (2013a). Marginal empirical likelihood and sure independence feature screening. *Ann. Statist.* **41** 2123–2148. MR3127860

CHANG, J., TANG, C. Y. and WU, Y. (2013b). Supplement to "Marginal empirical likelihood and sure independence feature screening." DOI:10.1214/13-AOS1139SUPP.

CHANG, J., TANG, C. Y. and WU, Y. (2016). Supplement to "Local independence feature screening for nonparametric and semiparametric models by marginal empirical likelihood." DOI:10.1214/15-AOS1374SUPP.

CHEN, S. X. (1996). Empirical likelihood confidence intervals for nonparametric density estimation. *Biometrika* **83** 329–341. MR1439787

CHEN, S. X., PENG, L. and QIN, Y.-L. (2009). Effects of data dimension on empirical likelihood. *Biometrika* **96** 711–722. MR2538767

CHEN, S. X. and QIN, Y. S. (2000). Empirical likelihood confidence intervals for local linear smoothers. *Biometrika* **87** 946–953. MR1813987

CHEN, S. X. and VAN KEILEGOM, I. (2009). A review on empirical likelihood methods for regression. *TEST* **18** 415–447. MR2566404

FAN, J., FENG, Y. and SONG, R. (2011). Nonparametric independence screening in sparse ultra-high-dimensional additive models. *J. Amer. Statist. Assoc.* **106** 544–557. MR2847969

FAN, J. and GIJBELS, I. (1996). *Local Polynomial Modelling and Its Applications*. *Monographs on Statistics and Applied Probability* **66**. Chapman & Hall, London. MR1383587

FAN, J. and LV, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **70** 849–911. MR2530322

FAN, J. and LV, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statist. Sinica* **20** 101–148. MR2640659

FAN, J., MA, Y. and DAI, W. (2014). Nonparametric independence screening in sparse ultra-high-dimensional varying coefficient models. *J. Amer. Statist. Assoc.* **109** 1270–1284. MR3265696

FAN, J., SAMWORTH, R. and WU, Y. (2009). Ultrahigh dimensional feature selection: Beyond the linear model. *J. Mach. Learn. Res.* **10** 2013–2038. MR2550099

FAN, J. and SONG, R. (2010). Sure independence screening in generalized linear models with NP-dimensionality. *Ann. Statist.* **38** 3567–3604. MR2766861

FAN, J. and ZHANG, J.-T. (2000). Two-step estimation of functional linear models with applications to longitudinal data. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **62** 303–322. MR1749541

HÄRDLE, W. (1990). *Applied Nonparametric Regression*. *Econometric Society Monographs* **19**. Cambridge Univ. Press, Cambridge. MR1161622

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2009). *The Elements of Statistical Learning*: *Data Mining*, *Inference*, *and Prediction*, 2nd ed. Springer, New York. MR2722294

HJORT, N. L., MCKEAGUE, I. W. and VAN KEILEGOM, I. (2009). Extending the scope of empirical likelihood. *Ann. Statist.* **37** 1079–1111. MR2509068

HUANG, J., HOROWITZ, J. L. and WEI, F. (2010). Variable selection in nonparametric additive models. *Ann. Statist.* **38** 2282–2313. MR2676890

HUANG, J. Z., WU, C. O. and ZHOU, L. (2002). Varying-coefficient models and basis function approximations for the analysis of repeated measurements. *Biometrika* **89** 111–128. MR1888349

LENG, C. and TANG, C. Y. (2012). Penalized empirical likelihood and growing dimensional general estimating equations. *Biometrika* **99** 703–716. MR2966779

LI, R., ZHONG, W. and ZHU, L. (2012). Feature screening via distance correlation learning. *J. Amer. Statist. Assoc.* **107** 1129–1139. MR3010900

LI, G., PENG, H., ZHANG, J. and ZHU, L. (2012). Robust rank correlation based screening. *Ann. Statist.* **40** 1846–1877. MR3015046

LIN, Y. and ZHANG, H. H. (2006). Component selection and smoothing in multivariate nonparametric regression. *Ann. Statist.* **34** 2272–2297. MR2291500

LIU, J., LI, R. and WU, R. (2014). Feature selection for varying coefficient models with ultrahigh-dimensional covariates. *J. Amer. Statist. Assoc.* **109** 266–274. MR3180562

MAI, Q. and ZOU, H. (2013). The Kolmogorov filter for variable screening in high-dimensional binary classification. *Biometrika* **100** 229–234. MR3034336

MÜLLER, H.-G. (1987). Weighted local regression and kernel methods for nonparametric curve fitting. *J. Amer. Statist. Assoc.* **82** 231–238. MR0883351

OWEN, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* **75** 237–249. MR0946049

OWEN, A. (1991). Empirical likelihood for linear models. *Ann. Statist.* **19** 1725–1747. MR1135146

OWEN, A. B. (2001). *Empirical Likelihood*. Chapman & Hall/CRC, New York.

PETROV, V. V. (1995). *Limit Theorems of Probability Theory*: *Sequences of Independent Random Variables*. *Oxford Studies in Probability* **4**. Oxford Univ. Press, New York. MR1353441

QU, A. and LI, R. (2006). Quadratic inference functions for varying-coefficient models with longitudinal data. *Biometrics* **62** 379–391. MR2227487

SONG, R., YI, F. and ZOU, H. (2014). On varying-coefficient independence screening for high-dimensional varying-coefficient models. *Statist. Sinica* **24** 1735–1752. MR3308660

STONE, C. J. (1985). Additive regression and other nonparametric models. *Ann. Statist.* **13** 689–705. MR0790566

TANG, C. Y. and LENG, C. (2010). Penalized high-dimensional empirical likelihood. *Biometrika* **97** 905–919. MR2746160

ZHU, L.-P., LI, L., LI, R. and ZHU, L.-X. (2011). Model-free feature screening for ultrahigh-dimensional data. *J. Amer. Statist. Assoc.* **106** 1464–1475. MR2896849

J. CHANG
SCHOOL OF STATISTICS
SOUTHWESTERN UNIVERSITY OF FINANCE
AND ECONOMICS
CHENGDU, SICHUAN 611130
CHINA
AND
SCHOOL OF MATHEMATICS AND STATISTICS
UNIVERSITY OF MELBOURNE
PARKVILLE, VICTORIA 3010
AUSTRALIA
E-MAIL: jinyuan.chang@unimelb.edu.au

C. Y. TANG
DEPARTMENT OF STATISTICS
TEMPLE UNIVERSITY
1810 NORTH 13TH STREET
PHILADELPHIA, PENNSYLVANIA 19122-6083
USA
E-MAIL: yongtang@temple.edu

Y. WU
DEPARTMENT OF STATISTICS
NORTH CAROLINA STATE UNIVERSITY
2311 STINSON DRIVE
RALEIGH, NORTH CAROLINA 27695-8203
USA
E-MAIL: wu@stat.ncsu.edu