

High-dimensional empirical likelihood inference

BY JINYUAN CHANG

*School of Statistics, Southwestern University of Finance and Economics,
Chengdu, Sichuan 611130, China*
changjinyuan@swufe.edu.cn

SONG XI CHEN

Guanghua School of Management, Peking University, Beijing 100871, China
csx@gsm.pku.edu.cn

CHENG YONG TANG

*Department of Statistical Science, Temple University, 1810 Liacouras Walk,
Philadelphia, Pennsylvania 19122, U.S.A.*
yongtang@temple.edu

AND TONG TONG WU

*Department of Biostatistics and Computational Biology, University of Rochester,
265 Crittenden Boulevard, Rochester, New York 14642, U.S.A.*
Tongtong_Wu@urmc.rochester.edu

SUMMARY

High-dimensional statistical inference with general estimating equations is challenging and remains little explored. We study two problems in the area: confidence set estimation for multiple components of the model parameters, and model specifications tests. First, we propose to construct a new set of estimating equations such that the impact from estimating the high-dimensional nuisance parameters becomes asymptotically negligible. The new construction enables us to estimate a valid confidence region by empirical likelihood ratio. Second, we propose a test statistic as the maximum of the marginal empirical likelihood ratios to quantify data evidence against the model specification. Our theory establishes the validity of the proposed empirical likelihood approaches, accommodating over-identification and exponentially growing data dimensionality. Numerical studies demonstrate promising performance and potential practical benefits of the new methods.

Some key words: Empirical likelihood; General estimating equation; High-dimensional statistical inference; Nuisance parameter; Over-identification.

1. INTRODUCTION

General estimating equations are broadly applicable for solving statistical inference problems, and they commonly involve over-identification: a general situation where the number of

restrictions is larger than that of the model parameters. Such a feature is advantageous. As a well-known example, the generalized method of moments (Hansen, 1982) allows incorporation of a flexible number of moment conditions in model building and subsequent statistical inferences; see also Hansen & Singleton (1982). Empirical likelihood (Owen, 2001), coupled with general estimating equations, has been demonstrated to be powerful for statistical inference since the seminal work of Qin & Lawless (1994). Without requiring specification of a full parametric probability distribution, empirical likelihood conveniently supports statistical inference with many desirable features including Wilks' type theorems, data adaptive yet shape-constraint-free confidence regions, and flexibility in combining multiple sources of data information.

Recently, there has been a surge in research on high-dimensional statistical problems. A class of approaches are facilitated by sparse model parameters whose many components are zeros. Penalized likelihood approaches have been demonstrated to be effective for estimating sparse model parameters; see the overview by Fan & Lv (2010), the monographs Bühlmann & van de Geer (2011) and Hastie et al. (2015), and references therein. Nevertheless, most existing penalized likelihood methods are constructed from conventional tools such as the least squares criterion or loglikelihood functions. Hence, they do not accommodate problems with general estimating equations, leaving this influential device less utilized.

High dimensionality is challenging for empirical likelihood; see Hjort et al. (2009) and Chen et al. (2009). Facilitated by empirical likelihood, Leng & Tang (2012) and Chang et al. (2015) consider penalized empirical likelihood with general estimating equations, and show that sparse estimators and statistical inferential procedures with good properties are achievable. However, those results only hold when the numbers of estimating equations and model parameters diverge at some slow polynomial rate of the sample size. Recently, Chang et al. (2018) introduced a new penalized empirical likelihood method that can accommodate exponentially growing numbers of estimating equations and model parameters. Their method effectively selects a subset of the estimating equations for estimating the nonzero components of the sparse model parameters. The study in Chang et al. (2018) only focuses on estimations, and does not cover broader concerns such as testing hypotheses or constructing confidence regions.

In this paper we consider two inference problems with general estimating equations using the empirical likelihood. To the best of our knowledge, this is the first attempt in the literature to accommodate over-identification in high-dimensional settings. In our presentation, we call a case low dimensional when it deals with either fixed or slowly diverging numbers of model parameters and estimating equations. The first problem is how to construct a confidence region for low-dimensional multiple components of the high-dimensional model parameters. Here, the estimation error associated with the other components of the model parameters, so-called nuisance parameters, is cumbersome. To overcome this difficulty, we propose to construct an empirical likelihood with a new set of low-dimensional estimating equations for those specified components. By projecting the original estimating equations with a linear transformation matrix whose rows are asymptotically orthogonal to the column space of the gradient matrix with respect to the nuisance parameters, the impact due to the estimation of the nuisance parameters becomes asymptotically negligible. Under the new construction, a valid confidence region can be constructed using the empirical likelihood ratio. The second problem is how to test whether or not a set of over-identified moment conditions are correctly specified. Our approach here is to calculate the marginal empirical likelihood ratios from a set of estimating functions evaluated at some consistent estimates. If a moment condition is misspecified, the corresponding marginal empirical likelihood ratio will diverge. Therefore, we propose a novel

high-dimensional over-identification test by assessing the maximum of the marginal empirical likelihood ratios.

Our investigation contributes to several areas. First, it expands the scope and deepens the understanding of high-dimensional empirical likelihood methods. We show that, by appropriate mapping, empirical likelihood still inherits the desirable merits for statistical inference with general estimating equations as in Qin & Lawless (1994). The key is to handle the high-dimensional nuisance parameters, a problem of foundational importance in the empirical likelihood literature; see, among others, Lazar & Mykland (1999), Chen & Cui (2006, 2007), Hjort et al. (2009), and the recent investigation of Bravo et al. (2020). Our treatment using a linear transformation is new, and it provides a crucial device of its own interests when investigating empirical likelihood; see § 3.1 for details and discussion. Second, our empirical likelihood-based over-identification test offers a new specification assessment tool to check the moment conditions. In conventional cases, the validity of the moment conditions can be assessed by the Sargan–Hansen J -test (Sargan, 1958; Hansen, 1982) and the empirical likelihood ratio test (Qin & Lawless, 1994). Unfortunately, those testing procedures cannot be applied to high-dimensional problems, because the test statistics are not even well defined when the number of estimating equations is larger than the sample size. To fill this gap, our method provides a suitable and viable solution in high-dimensional settings. Moreover, our approach is the first that can simultaneously handle multiple components of the model parameters and over-identification inference problems. To the best of our knowledge, existing high-dimensional methods of confidence set estimation focus on univariate analyses with no over-identification; see Zhang & Zhang (2013), van de Geer et al. (2014), Lee et al. (2016), Tibshirani et al. (2016) and Ning & Liu (2017). In the context of estimating equations, a recent study in Neykov et al. (2018) estimates the univariate confidence interval in high-dimensional just-identified settings, i.e., the same number of model parameters and estimating equations. Our approach can be applied more broadly. Our real data analysis with a longitudinal dataset from the Trial of Activity for Adolescent Girls demonstrates that the empirical likelihood methods with over-identification can provide an opportunity for potentially more accurate statistical inference in practice.

2. PRELIMINARIES

2.1. Notation

Let X_1, \dots, X_n be d -dimensional independently and identically distributed observations, and $\theta = (\theta_1, \dots, \theta_p)^T \in \Theta$ be a p -dimensional model parameter. With an r -dimensional estimating function $g(X; \theta) = \{g_1(X; \theta), \dots, g_r(X; \theta)\}^T$, a data model involving θ is specified by

$$E\{g(X_i; \theta_0)\} = 0, \quad (1)$$

where $\theta_0 = (\theta_{0,1}, \dots, \theta_{0,p})^T \in \Theta$ is the unknown truth. Here, one can view $\{g(X_i; \theta)\}_{i=1}^n$ as a triangular array, where r, d, p, X_i, θ and $g(\cdot; \cdot)$ may all depend on the sample size n .

For simplicity, in the following, when no confusion arises, we use $h_i(\theta)$ as equivalent to $h(X_i; \theta)$ for a generic q -dimensional function $h(\cdot; \cdot) = \{h_1(\cdot; \cdot), \dots, h_q(\cdot; \cdot)\}^T$ and denote by $h_{i,k}(\theta)$ the k th component of $h_i(\theta)$. Let $\bar{h}(\theta) = n^{-1} \sum_{i=1}^n h_i(\theta)$ and $\bar{h}_k(\theta) = n^{-1} \sum_{i=1}^n h_{i,k}(\theta)$. For a given index set $\mathcal{L} \subset \{1, \dots, q\}$, we denote by $h_{\mathcal{L}}(\cdot; \cdot)$ the subvector of $h(\cdot; \cdot)$ collecting the components indexed by \mathcal{L} . Analogously, let $h_{i,\mathcal{L}}(\theta) = h_{\mathcal{L}}(X_i; \theta)$ and $\bar{h}_{\mathcal{L}}(\theta) = n^{-1} \sum_{i=1}^n h_{i,\mathcal{L}}(\theta)$. For a matrix $B = (b_{i,j})_{s_1 \times s_2}$, let $B^{\otimes 2} = BB^T$, $|B|_{\infty} = \max_{1 \leq i \leq s_1, 1 \leq j \leq s_2} |b_{i,j}|$, and $\|B\|_s$ denote the matrix L_s -operator norm of B . When $s_2 = 1$, $|B|_s$ denotes the vector L_s -norm of the s_1 -dimensional vector B .

2.2. Current development of high-dimensional empirical likelihood

Since model estimation is the foundation of subsequent inference problems, let us start with an overview of the penalized empirical likelihood estimation approach in [Chang et al. \(2018\)](#):

$$\hat{\theta}_{\text{PEL}} = \arg \min_{\theta \in \Theta} \max_{\lambda \in \hat{\Lambda}_n(\theta)} \left[\sum_{i=1}^n \log\{1 + \lambda^\top g_i(\theta)\} + n \sum_{k=1}^p P_{1,\pi}(|\theta_k|) - n \sum_{j=1}^r P_{2,\nu}(|\lambda_j|) \right], \quad (2)$$

where $\lambda = (\lambda_1, \dots, \lambda_r)^\top$, $\hat{\Lambda}_n(\theta) = \{\lambda \in \mathbb{R}^r : \lambda^\top g_i(\theta) \in \mathcal{U} \text{ for any } i = 1, \dots, n\}$ with an open interval \mathcal{U} containing zero, and $P_{1,\pi}(\cdot)$ and $P_{2,\nu}(\cdot)$ are two penalty functions with tuning parameters π and ν , respectively. For any penalty function $P_\tau(\cdot)$ with tuning parameter τ , let $\rho(t; \tau) = \tau^{-1} P_\tau(t)$ for any $t \in [0, \infty)$ and $\tau \in (0, \infty)$. We restrict ourselves to the set of suitably behaved penalty functions.

DEFINITION 1. Let $\mathcal{P} = \{P_\tau(\cdot) : \rho(t; \tau) \text{ be increasing in } t \in [0, \infty) \text{ and have continuous derivative } \rho'(t; \tau) \text{ for any } t \in (0, \infty) \text{ with } \rho'(0^+; \tau) \in (0, \infty), \text{ where } \rho'(0^+; \tau) \text{ is independent of } \tau\}$.

Such a class is broad and general, including the L_1 penalty, the smoothly clipped absolute deviation penalty ([Fan & Li, 2001](#)), the minimax concave penalty ([Zhang, 2010](#)), amongst others. Let $\mathcal{S} = \{1 \leq k \leq p : \theta_{0,k} \neq 0\}$ with $|\mathcal{S}| = s \ll n$, i.e., the truth θ_0 is sparse. From [Chang et al. \(2018\)](#), $\hat{\theta}_{\text{PEL}}$ is consistent under some regularity conditions, stated in the following proposition.

PROPOSITION 1. Let $P_{1,\pi}(\cdot), P_{2,\nu}(\cdot) \in \mathcal{P}$ and $P_{2,\nu}(\cdot)$ be convex with bounded second-order derivative around 0. If Conditions [A1–A6](#) and the restrictions [\(A3\)](#) in the Appendix hold, there is a local minimizer $\hat{\theta}_{\text{PEL}} \in \Theta$ in [\(2\)](#) such that: (i) $|\hat{\theta}_{\text{PEL},\mathcal{S}} - \theta_{0,\mathcal{S}}|_\infty = O_p(\alpha_n)$ for some $\alpha_n \rightarrow 0$ as $n \rightarrow \infty$, and (ii) $\text{pr}(\hat{\theta}_{\text{PEL},\mathcal{S}^c} = 0) \rightarrow 1$ as $n \rightarrow \infty$.

From [\(A3\)](#) in the Appendix, Proposition [1](#) holds even if r and p grow exponentially with n . On the one hand, using a convex penalty $P_{2,\nu}(\cdot)$ makes the loss function concave with respect to λ , which leads to a unique maximizer $\lambda(\theta)$ in the inner optimization of [\(2\)](#) for each given θ . On the other hand, due to the convexity of $P_{2,\nu}(\cdot)$, there exists an asymptotic bias of order slower than $n^{-1/2}$ in $\hat{\theta}_{\text{PEL},\mathcal{S}}$. We observe that regularizing λ in [\(2\)](#) leads to a sparse solution $\hat{\lambda}$ corresponding to $\hat{\theta}_{\text{PEL}}$, which effectively selects components in $g(\cdot; \cdot)$. Write $\hat{\lambda} = (\hat{\lambda}_1, \dots, \hat{\lambda}_r)^\top$ and $\mathcal{R}_n = \text{supp}(\hat{\lambda})$. Similarly, denote $\rho_2(t; \nu) = \nu^{-1} P_{2,\nu}(t)$ for any $t > 0$ and $\hat{\eta} = (\hat{\eta}_1, \dots, \hat{\eta}_r)^\top$ with $\hat{\eta}_j = \nu \rho_2'(|\hat{\lambda}_j|; \nu) \text{sgn}(\hat{\lambda}_j)$ for $\hat{\lambda}_j \neq 0$ and $\hat{\eta}_j \in [-\nu \rho_2'(0^+), \nu \rho_2'(0^+)]$ for $\hat{\lambda}_j = 0$. Define $\hat{V}_{\mathcal{R}_n}(\hat{\theta}_{\text{PEL}}) = n^{-1} \sum_{i=1}^n g_{i,\mathcal{R}_n}(\hat{\theta}_{\text{PEL}})^{\otimes 2}$ and $\hat{J}_{\mathcal{R}_n} = \{[\nabla_{\theta_S} \bar{g}_{\mathcal{R}_n}(\hat{\theta}_{\text{PEL}})]^\top \hat{V}_{\mathcal{R}_n}^{-1/2}(\hat{\theta}_{\text{PEL}})\}^{\otimes 2}$. To achieve the best performance, the bias-corrected estimator is defined by

$$\hat{\theta}_{\text{PELbc}} = \hat{\theta}_{\text{PEL}} - \hat{\psi}_*, \quad (3)$$

where the p -dimensional vector $\hat{\psi}_*$ satisfies $\hat{\psi}_{*,\mathcal{S}^c} = 0$ and $\hat{\psi}_{*,\mathcal{S}} = \hat{\psi}$ with s -dimensional vector $\hat{\psi} = \hat{J}_{\mathcal{R}_n}^{-1} \{[\nabla_{\theta_S} \bar{g}_{\mathcal{R}_n}(\hat{\theta}_{\text{PEL}})]^\top \hat{V}_{\mathcal{R}_n}^{-1}(\hat{\theta}_{\text{PEL}}) \hat{\eta}_{\mathcal{R}_n}\}$. Let $V_{\mathcal{R}_n}(\theta_0) = E_{\mathcal{R}_n} \{g_{i,\mathcal{R}_n}(\theta_0)^{\otimes 2}\}$ and $J_{\mathcal{R}_n} = \{[E_{\mathcal{R}_n} \{[\nabla_{\theta_S} g_{i,\mathcal{R}_n}(\theta_0)]\}^\top V_{\mathcal{R}_n}^{-1/2}(\theta_0)]\}^{\otimes 2}$. Due to the randomness of the index set \mathcal{R}_n , we only take the expectation with respect to X_i and treat \mathcal{R}_n as given when we define $E_{\mathcal{R}_n} \{g_{i,\mathcal{R}_n}(\theta_0)^{\otimes 2}\}$ and $E_{\mathcal{R}_n} \{[\nabla_{\theta_S} g_{i,\mathcal{R}_n}(\theta_0)]\}$. Properties of $\hat{\theta}_{\text{PELbc}}$ are summarized in the following proposition.

PROPOSITION 2. Let $P_{1,\pi}(\cdot), P_{2,\nu}(\cdot) \in \mathcal{P}$ and $P_{2,\nu}(\cdot)$ be convex with bounded second-order derivative around 0. If Conditions [A1–A8](#) and the restrictions [\(A4\)](#) in the Appendix hold, $\hat{\theta}_{\text{PELbc}}$

in (3) satisfies: (i) $\hat{\theta}_{\text{PELbc},\mathcal{S}} - \theta_{0,\mathcal{S}} = -J_{\mathcal{R}_n}^{-1}[E_{\mathcal{R}_n}\{\nabla_{\theta_{\mathcal{S}}}\mathcal{G}_{i,\mathcal{R}_n}(\theta_0)\}]^T V_{\mathcal{R}_n}^{-1}(\theta_0)\bar{g}_{\mathcal{R}_n}(\theta_0) + \Delta_n$ with $|\Delta_n|_2 = O_p(\phi_n)$ for some $\phi_n = o(n^{-1/2})$; and (ii) $\text{pr}(\hat{\theta}_{\text{PELbc},\mathcal{S}^c} = 0) \rightarrow 1$ as $n \rightarrow \infty$.

To compute the bias-corrected estimator $\hat{\theta}_{\text{PELbc}}$, the support \mathcal{S} of θ_0 is needed. In practice, we may use the support of $\hat{\theta}_{\text{PEL}}$. Since $|\hat{\theta}_{\text{PEL}} - \theta_0|_\infty = O_p(\alpha_n)$ for some $\alpha_n \rightarrow 0$ as $n \rightarrow \infty$, if the signal strength of the nonzero components satisfies the condition $\alpha_n = o(\min_{k \in \mathcal{S}} |\theta_{0,k}|)$, such a support estimation is valid in the sense that $\text{pr}\{\text{supp}(\hat{\theta}_{\text{PEL}}) = \mathcal{S}\} \rightarrow 1$ as $n \rightarrow \infty$.

2.3. Two inference problems of interest

We study the following two problems:

- (a) Inference for multiple components of model parameters. Without loss of generality, we write $\theta = (\theta_{\mathcal{M}}^T, \theta_{\mathcal{M}^c}^T)^T$, where $\theta_{\mathcal{M}} \in \mathbb{R}^m$ contains the low-dimensional components of interest, and $\theta_{\mathcal{M}^c} \in \mathbb{R}^{p-m}$ contains the nuisance parameters. The construction of confidence regions for $\theta_{\mathcal{M}}$ will be shown in § 3.1.
- (b) Over-identification test. When $r > p$, a specification test is proposed in § 3.2 to check the validity of model (1) by testing the hypothesis $H_0 : E\{g_i(\theta_0)\} = 0$ for some $\theta_0 \in \Theta$ versus $H_1 : E\{g_i(\theta)\} \neq 0$ for any $\theta \in \Theta$.

In Problem (a) when $m = 1$, our method reduces to the special case of constructing a confidence interval for an individual component of θ . More generally when $m > 1$, we are estimating the confidence region for multiple components as specified by $\theta_{\mathcal{M}}$. Although $\hat{\theta}_{\text{PEL}}$ and $\hat{\theta}_{\text{PELbc}}$ in (2) and (3) provide consistent estimates for θ_0 , we cannot use their limiting distributions to solve Problem (a), mainly due to two reasons: (i) \mathcal{S} is generally unknown, and (ii) the limiting distributions of $\hat{\theta}_{\text{PEL},\mathcal{S}^c}$ and $\hat{\theta}_{\text{PELbc},\mathcal{S}^c}$ are also unknown. Problem (b) is known as the over-identification test. The Sargan–Hansen J -test (Sargan, 1958; Hansen, 1982) and the empirical likelihood ratio test (Qin & Lawless, 1994) can be used for such a purpose when r and p are fixed. When both r and p are less than n or diverge with n at some polynomial rate, by appropriate normalization, the Sargan–Hansen test and the empirical likelihood ratio test may still apply (Chang et al., 2015). However, when p and/or r is greater than n , neither applies because they both rely explicitly or implicitly on inverting a large sample covariance matrix that is not of full rank.

3. METHODOLOGY

3.1. Inference for low-dimensional components of model parameters

When r and p are fixed, the profile empirical likelihood approach of Qin & Lawless (1994) can be applied to solve Problem (a). Specifically, we consider the empirical likelihood function

$$L(\theta) = \sup \left\{ \prod_{i=1}^n \pi_i : \pi_i > 0, \sum_{i=1}^n \pi_i = 1, \sum_{i=1}^n \pi_i g_i(\theta) = 0 \right\} \quad (4)$$

for any $\theta \in \Theta$, and define the empirical likelihood estimator for θ_0 as $\check{\theta}_n = \arg \max_{\theta \in \Theta} L(\theta)$. The profile empirical likelihood ratio is defined as $\tilde{\ell}(\theta_{\mathcal{M}}) = \ell(\theta_{\mathcal{M}}, \bar{\theta}_{\mathcal{M}^c}) - \ell(\check{\theta}_n)$, where $\ell(\theta) = -2 \log\{n^n L(\theta)\}$, and $\bar{\theta}_{\mathcal{M}^c}$ minimizes $\ell(\theta_{\mathcal{M}}, \theta_{\mathcal{M}^c})$ with respect to $\theta_{\mathcal{M}^c}$ for a given $\theta_{\mathcal{M}}$. It is well known that $\tilde{\ell}(\theta_{0,\mathcal{M}}) \rightarrow \chi_m^2$ in distribution as $n \rightarrow \infty$. Then, $\{\theta_{\mathcal{M}} \in \mathbb{R}^m : \tilde{\ell}(\theta_{\mathcal{M}}) \leq \chi_{m,1-\alpha}^2\}$

provides a $100(1 - \alpha)\%$ confidence region for $\theta_{\mathcal{M}}$, where $\chi_{m,1-\alpha}^2$ denotes the $(1 - \alpha)$ -quantile of the chi-square distribution with m degrees of freedom.

The confidence regions constructed by empirical likelihood ratio have several advantages (Hall & La Scala, 1990). First, empirical-likelihood-based confidence regions are data driven, being free from shape constraints; see the plots of the estimated confidence regions from our simulation in the Supplementary Material. Second, though being nonparametric, empirical-likelihood-based confidence regions are Bartlett correctable, so the order of the coverage error can be reduced from n^{-1} to n^{-2} , with a simple correction for the mean of the empirical likelihood ratio statistics (Chen & Cui, 2006, 2007). Third, empirical likelihood methods require no further estimation such as the scale and the skewness, which is an appealing feature when solving high-dimensional problems. Fourth, the empirical likelihood method can be adapted to construct confidence regions for general smooth functions of the model parameter (Qin & Lawless, 1995).

Clearly, when both r and p are allowed to diverge with n , the profile empirical likelihood approach encounters substantial difficulty. First, calculating $\tilde{\ell}(\theta_{\mathcal{M}})$ is challenging due to the fact that it is generally a high-dimensional, nonconvex optimization problem. Second, the existing asymptotic analysis on the profile empirical likelihood ratio $\tilde{\ell}(\theta_{\mathcal{M}})$ cannot be generalized to the high-dimensional case. To illustrate this, let us first pretend that the truth of the nuisance parameters $\theta_{\mathcal{M}^c}$, denoted by θ_{0,\mathcal{M}^c} , is known. Then the empirical likelihood for $\theta_{\mathcal{M}} \in \mathbb{R}^m$ follows the conventional framework. When r is fixed, the empirical likelihood ratio $\ell(\theta_0) = -2 \log\{n^n L(\theta_0)\} \rightarrow \chi_r^2$ in distribution as $n \rightarrow \infty$, so $\{\theta_{\mathcal{M}} \in \mathbb{R}^m : \ell(\theta_{\mathcal{M}}, \theta_{0,\mathcal{M}^c}) \leq \chi_{r,1-\alpha}^2\}$ is a valid confidence region for $\theta_{\mathcal{M}}$. If θ_{0,\mathcal{M}^c} is replaced by a \sqrt{n} -consistent estimate $\tilde{\theta}_{\mathcal{M}^c}$, still keeping r fixed, $\ell(\theta_{0,\mathcal{M}}, \tilde{\theta}_{\mathcal{M}^c})$ generally converges to some weighted sum of chi-square distributions (Hjort et al., 2009). However, if the estimate $\tilde{\theta}_{\mathcal{M}^c}$ converges to θ_{0,\mathcal{M}^c} slower than $n^{-1/2}$, $\ell(\theta_{0,\mathcal{M}}, \tilde{\theta}_{\mathcal{M}^c})$ generally diverges with probability approaching one (Chang et al., 2013, 2016). When θ is high dimensional, the convergence rate of such estimators is generally slower than $n^{-1/2}$. Hence, a naive plugging in of $\tilde{\theta}_{\mathcal{M}^c}$ into (4) will not work. A key reason leading to the failure of empirical likelihood with high-dimensional problems is the errors from estimating the nuisance parameters.

Therefore, it is crucial to investigate the impact on empirical likelihood from the estimation of nuisance parameters. In conventional settings with a fixed number of model parameters, the first- and second-order properties of empirical likelihood ratio statistics are documented in Qin & Lawless (1994), Lazar & Mykland (1999) and Chen & Cui (2006, 2007). Hjort et al. (2009) consider nuisance parameters that can be functional valued and estimated by some nonparametric methods. The work of Bravo et al. (2020) demonstrates that by using estimated influence functions, the chi-squared-distributed empirical likelihood ratio statistics can be justified. Nevertheless, little exists in the literature on how to handle high-dimensional nuisance parameters when penalized estimation approaches are used. Though sparse and consistent parameter estimates are achievable with penalized empirical likelihood (Chang et al., 2018), the zero components in the estimates are essentially degenerated and, even worse, their influence functions do not exist, rendering the existing methods inapplicable. Hence, our challenge here is fundamentally different from, for example, functional-valued nuisance parameters, where smoothness around the truth leads to uniformly consistent and regular estimations.

To cope with nuisance parameters, we observe that for a consistent estimator $\theta_{\mathcal{M}^c}^*$ of θ_{0,\mathcal{M}^c} ,

$$Q_n = \bar{g}(\theta_{0,\mathcal{M}}, \theta_{\mathcal{M}^c}^*) - \bar{g}(\theta_0) = \{\nabla_{\theta_{\mathcal{M}^c}} \bar{g}(\theta_{0,\mathcal{M}}, \theta_{\mathcal{M}^c}^*)\}(\theta_{\mathcal{M}^c}^* - \theta_{0,\mathcal{M}^c}) + R_1, \quad (5)$$

where R_1 is asymptotically negligible. A strategy here is to find a linear transformation matrix $A_n = (a_1^n, \dots, a_m^n)^T \in \mathbb{R}^{m \times r}$ satisfying $|A_n Q_n|_2 = o_p(n^{-1/2})$, where each a_k^n is an r -dimensional

vector. Then, by utilizing $f^{A_n}(\cdot; \cdot) = A_n g(\cdot; \cdot)$ as the new m -dimensional estimating functions, the empirical likelihood constructed with $f^{A_n}(\cdot; \cdot)$ instead of $g(\cdot; \cdot)$ can be used for statistical inference for $\theta_{\mathcal{M}}$. Specifically, let $\ell_{A_n}^*(\theta_{\mathcal{M}}) = -2 \log\{n^n L_{A_n}^*(\theta_{\mathcal{M}}; \theta_{\mathcal{M}^c}^*)\}$, with

$$L_{A_n}^*(\theta_{\mathcal{M}}; \theta_{\mathcal{M}^c}^*) = \sup \left\{ \prod_{i=1}^n \pi_i : \pi_i > 0, \sum_{i=1}^n \pi_i = 1, \sum_{i=1}^n \pi_i f_i^{A_n}(\theta_{\mathcal{M}}, \theta_{\mathcal{M}^c}^*) = 0 \right\}. \quad (6)$$

From (5), an ideal choice of A_n should be such that $A_n \nabla_{\theta_{\mathcal{M}^c}} \bar{g}(\theta_{0, \mathcal{M}}, \theta_{\mathcal{M}^c}^*)$ being small in the sense that each row vector $(a_k^n)^T$ of A_n satisfies that $|(a_k^n)^T \{\nabla_{\theta_{\mathcal{M}^c}} \bar{g}(\theta_{0, \mathcal{M}}, \theta_{\mathcal{M}^c}^*)\}|_{\infty}$ diminishes to 0 as $n \rightarrow \infty$. Or equivalently, rows of A_n should be chosen as asymptotically orthogonal to the column space of $\nabla_{\theta_{\mathcal{M}^c}} \bar{g}(\theta_{0, \mathcal{M}}, \theta_{\mathcal{M}^c}^*)$, the $r \times (p - m)$ sample gradient matrix with respect to the nuisance parameters. As an additional key consideration, we note that the gradient with respect to $\theta_{\mathcal{M}}$ evaluated at $(\theta_{0, \mathcal{M}}, \theta_{\mathcal{M}^c}^*)$ should not vanish; otherwise, a flat estimating function at $\theta_{0, \mathcal{M}}$ is not informative. Thus, $A_n \nabla_{\theta_{\mathcal{M}}} \bar{g}(\theta_{0, \mathcal{M}}, \theta_{\mathcal{M}^c}^*)$ is required to be nonsingular. In practice, the true $\theta_{0, \mathcal{M}}$ is unknown so an estimate, denoted by $\theta_{\mathcal{M}}^*$, is needed when constructing A_n .

By putting the ideas together, we propose to find A_n row by row with the optimizations

$$a_k^n = \arg \min_{u \in \mathbb{R}^r} |u|_1 \quad \text{such that} \quad |\{\nabla_{\theta} \bar{g}(\theta^*)\}^T u - \xi_k|_{\infty} \leq \tau, \quad (7)$$

where $\theta^* = \{(\theta_{\mathcal{M}}^*)^T, (\theta_{\mathcal{M}^c}^*)^T\}^T$ is an initial estimate for θ_0 , τ is a tuning parameter and $\{\xi_k\}_{k=1}^m$ are the canonical basis of the linear space $\mathcal{M}_{\xi} = \{b = (b_1, \dots, b_p)^T : b_j = 0 \text{ for any } j = m + 1, \dots, p\}$, i.e., ξ_k is chosen such that its k th component is 1 and all other components are 0. Thus, a $100(1 - \alpha)\%$ -level confidence region for $\theta_{\mathcal{M}}$ is given as follows:

- (i) When m is fixed, $\mathcal{C}_{1-\alpha} = \{\theta_{\mathcal{M}} \in \mathbb{R}^m : \ell_{A_n}^*(\theta_{\mathcal{M}}) \leq \chi_{m, 1-\alpha}^2\}$, where $\chi_{m, 1-\alpha}^2$ is the $(1 - \alpha)$ -quantile of the chi-square distribution with m degrees of freedom.
- (ii) When m is diverging, $\mathcal{C}_{1-\alpha} = \{\theta_{\mathcal{M}} \in \mathbb{R}^m : \ell_{A_n}^*(\theta_{\mathcal{M}}) \leq m + z_{1-\alpha}(2m)^{1/2}\}$, where $z_{1-\alpha}$ is the $(1 - \alpha)$ -quantile of standard normal distribution $N(0, 1)$. The rationale is that $(\chi_m^2 - m)/(2m)^{1/2} \rightarrow N(0, 1)$ in distribution as $m \rightarrow \infty$.

To construct $f^{A_n}(\cdot; \cdot)$ in (6), one needs no more than the original estimating functions $g(\cdot; \cdot)$ and an initial estimate θ^* . This strategy is new and effective, as it can be generally adapted to handle nuisance parameters with empirical likelihood or as a development of its own interests. Theorem 1 in § 4.1 establishes the validity of the above procedure. Briefly speaking, for a given consistent initial estimate θ^* , the estimated confidence region is asymptotically valid as $n \rightarrow \infty$, allowing both r and p to diverge at some exponential rate of n . Requiring a consistent initial estimate θ^* is not restrictive, and can be broadly satisfied by sparse penalized estimates in cases such as linear models and generalized linear models. For more general problems associated with estimating equations, we advocate applying $\hat{\theta}_{\text{PEL}}$ given by (2). As an advantage of our method, it does not require the bias correction step when using the transformed estimating function $f^{A_n}(\cdot; \cdot)$. As a comparison, in Zhang & Zhang (2013) and van de Geer et al. (2014), bias correction is necessary to construct normal-distribution-based confidence intervals in high-dimensional linear models.

Let $\Gamma = E\{\nabla_{\theta} g_i(\theta_0)\}$. The existence of a_k such that $\Gamma^T a_k = \xi_k$ is an elementary requirement, since $\Gamma \in \mathbb{R}^{r \times p}$ with $r \geq p$, a_k may not be unique. A major challenge of the theoretical analysis is the identifiability of A_n from (7) for high-dimensional problems. We impose the following regularity condition.

Condition 1. For each $k = 1, \dots, m$ there is a nonrandom a_k satisfying $\Gamma^\top a_k = \xi_k$, $|a_k|_1 \leq C_1$ for some uniform constant $C_1 > 0$, and $\max_{1 \leq k \leq m} |a_k^n - a_k|_1 = O_p(\omega_n)$ for some $\omega_n \rightarrow 0$.

Let $\hat{\Gamma}_n = \nabla_{\theta} \bar{g}(\theta^*)$. It follows from the existence of a_k that $\xi_k = \hat{\Gamma}_n^\top a_k + (\Gamma - \hat{\Gamma}_n)^\top a_k = \hat{\Gamma}_n^\top a_k + e_k$. Under some mild conditions, $|\hat{\Gamma}_n - \Gamma|_\infty \rightarrow 0$ in probability. This, together with the assumption that $|a_k|_1 \leq C_1$, implies that e_k is stochastically small uniformly over all its components such that $|e_k|_\infty = o_p(1)$. This can be seen as an attempt to recover a nonrandom a_k with no noise asymptotically (Candès & Tao, 2007; Bickel et al., 2009). Thus, a_k^n from (7) satisfies $|a_k^n - a_k|_1 \rightarrow 0$ in probability if $\hat{\Gamma}_n$ satisfies the routine conditions for sparse signal recovering. It can be shown that C_1 in Condition 1 can be replaced by some diverging sequence γ_n and our main results remain valid. Theorem 1 in § 4.1 indicates that it only requires ω_n to satisfy $m\omega_n^2(m + \log r) = o(1)$ for the validity of our procedure. For the just-identified case where $r = p$, our assumption on the existence of $\Gamma^\top a_k = \xi_k$ is weaker than that of Neykov et al. (2018), which assumes Γ to be invertible to make $a_k = \Gamma^{-1} \xi_k$ unique.

Our statistical inferential procedure can be extended to broader cases of interest. For a general function $S(\theta_{\mathcal{M}}) \in \mathbb{R}^q$ of a specified $\theta_{\mathcal{M}}$, the formulation of Qin & Lawless (1995) can be applied to construct the confidence region for $S(\theta_{\mathcal{M}})$:

$$C_{1-\alpha} = \left\{ v \in \mathbb{R}^q : \min_{\theta_{\mathcal{M}}: S(\theta_{\mathcal{M}})=v} \ell_{A_n}^*(\theta_{\mathcal{M}}) \leq \chi_{q,1-\alpha}^2 \right\}.$$

For a univariate and monotone transformation $S(\cdot)$, the confidence region with empirical likelihood has the invariant property (Hall & La Scala, 1990). In the special case $S(\theta_{\mathcal{M}}) = L\theta_{\mathcal{M}}$ with $L \in \mathbb{R}^{q \times m}$, i.e., q linear combinations of $\theta_{\mathcal{M}}$, the validity of the confidence region defined in such a way can be established following the same idea as our analysis.

Empirical likelihood with over-identified general estimating equations may provide a unique opportunity for enhancing the accuracy of the statistical inference. For the inference of m -dimensional components $\theta_{\mathcal{M}}$, one may opt to find \tilde{m} ($\tilde{m} \geq m$) linear combinations of the original estimating function. The rationale is that $\Gamma^\top u = \xi_k$, as in Condition 1, may yield multiple linearly independent sparse solutions. Practically, an option is to implement (7) sequentially: upon finding a solution $a_k^{n,1}$, one runs (7) to find another solution $a_k^{n,2}$ subject to an additional linear constraint such that $(a_k^{n,1})^\top a_k^{n,2} = 0$. Using over-identification ($\tilde{m} > m$) for $\theta_{\mathcal{M}}$ is beneficial for improving the accuracy of statistical inference (Qin & Lawless, 1994). As shown in our simulation in § 5, such improvement is substantial.

3.2. Over-identification test

Over-identification provides an opportunity to develop a statistical test for checking the validity of model specification. In low-dimensional cases, the Sargan–Hansen J -test (Sargan, 1958; Hansen, 1982) and the empirical likelihood ratio test (Qin & Lawless, 1994) can be used. Qin & Lawless (1994) shows that $\ell(\check{\theta}_n) = -2 \log\{n^n L(\check{\theta}_n)\} \rightarrow \chi_{r-p}^2$ in distribution under H_0 , where $\check{\theta}_n$ is the maximizer of $L(\theta)$ in (4). It can be shown that the Sargan–Hansen J statistic is first-order equivalent to the empirical likelihood ratio statistic $\ell(\check{\theta}_n)$, therefore they share the same limiting distribution. When the paradigm shifts to high-dimensional settings, the asymptotic quadratic form no longer holds and the limiting χ_{r-p}^2 distribution becomes invalid.

Our over-identification test is developed from the marginal empirical likelihood ratios. Given $\hat{\theta}_n$, a consistent estimate of θ_0 under H_0 , we define the marginal empirical likelihood ratio for the

j th estimating function $g_j(\cdot; \cdot)$ in $g(\cdot; \cdot)$ as

$$\ell_j(\hat{\theta}_n) = 2 \max_{\lambda \in \hat{\Lambda}_{n,j}} \sum_{i=1}^n \log\{1 + \lambda g_{i,j}(\hat{\theta}_n)\},$$

where $\hat{\Lambda}_{n,j} = \{\lambda \in \mathbb{R} : \lambda g_{i,j}(\hat{\theta}_n) \in \mathcal{U} \text{ for any } i = 1, \dots, n\}$ with an open interval \mathcal{U} containing zero. Based on $\{\ell_j(\hat{\theta}_n)\}_{j=1}^r$, we propose the following test statistic:

$$T_n = \max_{j \in \mathcal{J}} \ell_j(\hat{\theta}_n), \tag{8}$$

where \mathcal{J} is a prescribed index set with $|\mathcal{J}| = q$. Since the calculation of $\ell_j(\hat{\theta}_n)$ only involves univariate optimizations, calculating T_n is highly scalable and can be done efficiently. The intuition of (8) is that when H_0 is true, each $\ell_j(\hat{\theta}_n)$ should take a relatively small value. In contrast, when H_0 is violated, one expects that at least some $\ell_j(\hat{\theta}_n)$ will be large.

The selection of the index set \mathcal{J} in (8) is the key to developing a powerful procedure for a high-dimensional over-identification test. In low-dimensional cases, a natural choice of \mathcal{J} is to include all r estimating functions. However, additional consideration is necessary when dealing with high-dimensional problems. To illustrate the idea, we add the subscript \mathcal{J} in T_n here to emphasize the dependency. In our method, the α -level critical value for $T_{n,\mathcal{J}}$ is selected as the $(1 - \alpha)$ -quantile of the distribution of $|\hat{G}_{\mathcal{J}}|_{\infty}^2$, where $\hat{G}_{\mathcal{J}}$ follows some q -dimensional multivariate normal distribution. Let $j_{\#} = \arg \max_{1 \leq j \leq r} \ell_j(\hat{\theta}_n)$. For any two different index sets \mathcal{J}_1 and \mathcal{J}_2 satisfying $j_{\#} \in \mathcal{J}_1 \cap \mathcal{J}_2$ and $\mathcal{J}_1 \subset \mathcal{J}_2$, it is easy to see that $T_{n,\mathcal{J}_1} = T_{n,\mathcal{J}_2}$. Due to the fact that $\hat{G}_{\mathcal{J}_1}$ is a subvector of $\hat{G}_{\mathcal{J}_2}$, the $(1 - \alpha)$ -quantile of the distribution of $|\hat{G}_{\mathcal{J}_1}|_{\infty}^2$ will be no larger than that of $|\hat{G}_{\mathcal{J}_2}|_{\infty}^2$. Hence, when too many components are included in constructing the test statistic, the associated critical value inevitably becomes too large, which will lead to power loss. To obtain a powerful test, we only need to select a small index set \mathcal{J} , with $j_{\#}$ being included to best maintain the signal for detecting the violation of H_0 ; see also Chang et al. (2017, § 2.3) for more discussion on such a phenomenon for an L_{∞} -type test statistic. Further, results in Chang et al. (2013, 2016) show that $\ell_j(\hat{\theta}_n)$ diverges fast if $|\bar{g}_j(\hat{\theta}_n)|$ does not converge to zero fast enough, the signal from violating H_0 that the over-identification test intends to detect. Thus, one should ideally include in the index set \mathcal{J} those j with large $|\bar{g}_j(\hat{\theta}_n)|$. The selection of \mathcal{J} will be elaborated more at the end of this section.

Obviously, the test statistic T_n in (8) depends on the estimate $\hat{\theta}_n$. Recall that $\mathcal{S} = \text{supp}(\theta_0) = \{1 \leq k \leq p : \theta_{0,k} \neq 0\}$ with $|\mathcal{S}| = s$. Our theoretical analysis requires $\hat{\theta}_n$ to satisfy the following two properties under H_0 :

Property 1. $\hat{\theta}_{n,\mathcal{S}} - \theta_{0,\mathcal{S}} = n^{-1} \sum_{i=1}^n m(X_i; \theta_0) + \Delta_n$ with $|\Delta_n|_2 = o_p(n^{-1/2})$;

Property 2. $\text{pr}(\hat{\theta}_{n,\mathcal{S}^c} = 0) \rightarrow 1$ as $n \rightarrow \infty$,

where $m(\cdot; \cdot)$ is the s -dimensional influence function of $\hat{\theta}_{n,\mathcal{S}}$. To require $|\Delta_n|_2 = o_p(n^{-1/2})$ in Property 1 is not stringent, and can be satisfied by penalized likelihood estimates up to a bias correction (Fan & Li, 2001). Property 2 is the oracle property. As seen below, Property 2 is not essential, but more involved characterization is required without it. In special cases including the linear and generalized linear models, we recommend applying bias correction or refitting the selected model to obtain less biased estimates, for example using the method in Belloni & Chernozhukov (2013). For more general models with estimating equations, $\hat{\theta}_n$ can be chosen

as the bias-corrected estimate $\hat{\theta}_{\text{PELbc}}$ given by (3) in § 2.2, which meets the requirements by Proposition 2 in § 2.2.

Denote $\hat{V}_{\mathcal{J}}(\hat{\theta}_n) = n^{-1} \sum_{i=1}^n g_{i,\mathcal{J}}(\hat{\theta}_n)^{\otimes 2}$ and $V_{\mathcal{J}}(\theta_0) = E_{\mathcal{J}}\{g_{i,\mathcal{J}}(\theta_0)^{\otimes 2}\}$. Here, when we define $V_{\mathcal{J}}(\theta_0)$ we only take the expectation with respect to X_i and treat the index set \mathcal{J} as given, which itself might be random. For any $j \in \mathcal{J}$, let $\hat{\sigma}_j^2(\hat{\theta}_n) = n^{-1} \sum_{i=1}^n g_{i,j}^2(\hat{\theta}_n)$. Based on the well-known self-Studentized property of the empirical likelihood ratio, it can be shown under H_0 that $\sup_{j \in \mathcal{J}} |\ell_j(\hat{\theta}_n) - n\{\bar{g}_j(\hat{\theta}_n)\}^2 \hat{\sigma}_j^{-2}(\hat{\theta}_n)| = o_p(1)$. By expanding $\bar{g}_{\mathcal{J}}(\hat{\theta}_n)$ around θ_0 , we have $n^{1/2}[\text{diag}\{\hat{V}_{\mathcal{J}}(\hat{\theta}_n)\}]^{-1/2} \bar{g}_{\mathcal{J}}(\hat{\theta}_n) = n^{-1/2} \sum_{i=1}^n w_i(\theta_0) + \tilde{\Delta}_n$, where $w_i(\theta_0) = [\text{diag}\{V_{\mathcal{J}}(\theta_0)\}]^{-1/2} \{g_{i,\mathcal{J}}(\theta_0) + [E_{\mathcal{J}}\{\nabla_{\theta_S} g_{i,\mathcal{J}}(\theta_0)\}] m_i(\theta_0)\}$ and $|\tilde{\Delta}_n|_{\infty} = o_p(n^{-1/2})$. Following the idea of Gaussian approximation (Chernozhukov et al., 2017), we can approximate the distribution of $T_n = \max_{j \in \mathcal{J}} \ell_j(\hat{\theta}_n)$ by that of $|\hat{G}|_{\infty}^2$, where $\hat{G} \sim N(0, \hat{W})$ for some \hat{W} .

Since $\hat{\theta}_n$ is estimated from the data $\mathcal{X}_n = \{X_1, \dots, X_n\}$, its influence function $m_i(\cdot)$ and the estimating function $g_i(\cdot)$ are dependent. As discussed below Proposition 2 in § 2.2, the unknown index set \mathcal{S} can be consistently estimated. To simplify the notation and without loss of generality, we assume \mathcal{S} is known. Otherwise, we can replace it in practice by $\hat{\mathcal{S}} = \text{supp}(\hat{\theta}_{\text{PEL}})$ for $\hat{\theta}_{\text{PEL}}$ in (2). To elaborate with details on \hat{W} , we present the framework by selecting $\hat{\theta}_n$ as $\hat{\theta}_{\text{PELbc}}$ given in (3). Recall that $\mathcal{R}_n = \text{supp}(\hat{\lambda})$ and $\hat{\lambda}$ corresponds to $\hat{\theta}_{\text{PEL}}$ in the inner optimization of (2). Singling out \mathcal{R}_n here is necessary to concretely present a synthetic framework.

To avoid loss of generality, we do not impose any relationship between the two index sets \mathcal{J} and \mathcal{R}_n in our theoretical analysis. Let $\mathcal{I} = \mathcal{R}_n \cup \mathcal{J}$. Both the estimating functions in $g(\cdot; \cdot)$ indexed by \mathcal{I} , and the covariance matrix of $\hat{\theta}_{n,\mathcal{S}}$ contribute to the joint distribution of $\{\ell_j(\hat{\theta}_n)\}_{j \in \mathcal{J}}$; see the Supplementary Material. For any $\mathcal{L} \subset \{1, \dots, r\}$, we define $V_{\mathcal{L}}(\theta_0) = E_{\mathcal{L}}\{g_{i,\mathcal{L}}(\theta_0)^{\otimes 2}\}$ and $J_{\mathcal{L}} = \{[E_{\mathcal{L}}\{\nabla_{\theta_S} g_{i,\mathcal{L}}(\theta_0)\}]^T V_{\mathcal{L}}^{-1/2}(\theta_0)\}^{\otimes 2}$. Again, both expectations are taken with respect to X_i and the index set \mathcal{L} is treated as given. To ensure the validity of \hat{W} given in (10), we rewrite $g(\cdot; \cdot)$ as

$$g(\cdot; \cdot) = \{g_{\mathcal{R}_n \cap \mathcal{J}}(\cdot; \cdot)^T, g_{\mathcal{R}_n \cap \mathcal{J}^c}(\cdot; \cdot)^T, g_{\mathcal{R}_n^c \cap \mathcal{J}}(\cdot; \cdot)^T, g_{\mathcal{I}^c}(\cdot; \cdot)^T\}^T.$$

Define $B = [E_{\mathcal{J}}\{\nabla_{\theta_S} g_{i,\mathcal{J}}(\theta_0)\}] J_{\mathcal{R}_n}^{-1} [E_{\mathcal{R}_n}\{\nabla_{\theta_S} g_{i,\mathcal{R}_n}(\theta_0)\}]^T V_{\mathcal{R}_n}^{-1}(\theta_0)$ with blocks

$$B = \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix}, \quad (9)$$

where B_{11} and B_{22} are $|\mathcal{R}_n \cap \mathcal{J}| \times |\mathcal{R}_n \cap \mathcal{J}|$ and $|\mathcal{R}_n^c \cap \mathcal{J}| \times |\mathcal{R}_n \cap \mathcal{J}^c|$ matrices. Let

$$\hat{Q} = \begin{pmatrix} I_{|\mathcal{R}_n \cap \mathcal{J}|} - \hat{B}_{11} & -\hat{B}_{12} & 0 \\ -\hat{B}_{21} & -\hat{B}_{22} & I_{|\mathcal{R}_n^c \cap \mathcal{J}|} \end{pmatrix},$$

where I_k is the identity matrix with order k , and \hat{B}_{ij} ($i, j = 1, 2$) are the corresponding estimates of B_{ij} in the matrix $\hat{B} = \{\nabla_{\theta_S} \bar{g}_{\mathcal{J}}(\hat{\theta}_n)\} \hat{J}_{*,\mathcal{R}_n}^{-1} \{\nabla_{\theta_S} \bar{g}_{\mathcal{R}_n}(\hat{\theta}_n)\}^T \hat{V}_{\mathcal{R}_n}^{-1}(\hat{\theta}_n)$ with $\hat{J}_{*,\mathcal{R}_n} = [\{\nabla_{\theta_S} \bar{g}_{\mathcal{R}_n}(\hat{\theta}_n)\}^T \hat{V}_{\mathcal{R}_n}^{-1/2}(\hat{\theta}_n)]^{\otimes 2}$. Last, we define

$$\hat{W} = \{[\text{diag}\{\hat{V}_{\mathcal{J}}(\hat{\theta}_n)\}]^{-1/2} \hat{Q} \hat{V}_{\mathcal{I}}^{1/2}(\hat{\theta}_n)\}^{\otimes 2}, \quad (10)$$

with $\hat{V}_{\mathcal{J}}(\hat{\theta}_n) = n^{-1} \sum_{i=1}^n g_{i,\mathcal{J}}(\hat{\theta}_n)^{\otimes 2}$ and $\hat{V}_{\mathcal{I}}(\hat{\theta}_n) = n^{-1} \sum_{i=1}^n g_{i,\mathcal{I}}(\hat{\theta}_n)^{\otimes 2}$.

For a given $\alpha \in (0, 1)$, the critical value is given by

$$\hat{c}v_\alpha = \inf\{t \in \mathbb{R} : \text{pr}(|\hat{G}_\infty|^2 > t \mid \mathcal{X}_n) \leq \alpha\}, \tag{11}$$

where $\hat{G} \sim N(0, \hat{W})$ with \hat{W} defined in (10). We reject H_0 if $T_n > \hat{c}v_\alpha$. Furthermore, $\hat{c}v_\alpha$ can be conveniently obtained by simulation with \hat{W} obtained from data. That is, one can generate independent $\hat{G}_1, \dots, \hat{G}_M$ from $N(0, \hat{W})$ for a large M and approximate $\hat{c}v_\alpha$ in (11) by $\hat{c}v_{\alpha, M} = \inf\{x \in \mathbb{R} : \hat{F}_M(x) \geq 1 - \alpha\}$, where $\hat{F}_M(x) = M^{-1} \sum_{b=1}^M I(|\hat{G}_b|^2 \leq x)$. The validity of the test is established in § 4.2. Theorem 2 justifies that the size of the test is asymptotically α under H_0 , and Theorem 3 elucidates the power of the test when H_0 is violated.

This section is concluded with a final remark that \mathcal{R}_n from (2) is an ideal candidate for \mathcal{J} if $\hat{\theta}_n$ is selected to be $\hat{\theta}_{\text{PELbc}}$. As we discussed before, the index set \mathcal{J} should include those j with large $|\bar{g}_j(\hat{\theta}_n)|$. Chang et al. (2018, Proposition 3) shows that components $g_j(\cdot; \cdot)$ with large values in $|\bar{g}_j(\hat{\theta}_n)|$ are included in \mathcal{R}_n . Furthermore, under H_1 , if $E\{g_{i,j}(\hat{\theta}_n)\} \neq 0$ for some j , its sample counterpart $|\bar{g}_j(\hat{\theta}_n)|$ tends to take some large value, and hence the corresponding index would fall into \mathcal{R}_n . In practice, we recommend using \mathcal{R}_n for the over-identification test, which is the one implemented in our numerical studies. Our simulation in § 5.2 shows that the over-identification test performs very well. By choosing \mathcal{J} in (8) as \mathcal{R}_n , the test is powerful compared with the one using all the estimating functions, especially when r is large.

4. THEORETICAL ANALYSIS

4.1. Inference for low-dimensional components

To establish theoretical guarantees for the validity of the confidence sets $\mathcal{C}_{1-\alpha}$ given in § 3.1, we assume the following regularity conditions.

Condition 2. For any X and $j = 1, \dots, p$, $g_j(X; \theta)$ is twice continuously differentiable with respect to θ . Both $\sup_{\theta \in \Theta} \max_{1 \leq j \leq r} \max_{1 \leq l \leq p} n^{-1} \sum_{i=1}^n |\partial g_{i,j}(\theta) / \partial \theta_l|^2 = O_p(1)$ and $\sup_{\theta \in \Theta} \max_{1 \leq j \leq r} \max_{1 \leq l_1, l_2 \leq p} n^{-1} \sum_{i=1}^n |\partial^2 g_{i,j}(\theta) / \partial \theta_{l_1} \partial \theta_{l_2}| = O_p(1)$ hold.

Condition 3. For some uniform constants $C_2 > 0$ and $\gamma > 4$, $\max_{1 \leq j \leq r} E\{\sup_{\theta \in \Theta} |g_{i,j}(\theta)|^\gamma\} < C_2$.

Condition 4. Assume $\max_{1 \leq j \leq r} n^{-1} \sum_{i=1}^n |g_{i,j}(\theta_0)|^2 = O_p(1)$.

Condition 5. Eigenvalues of $E\{g_i(\theta_0)^{\otimes 2}\}$ are uniformly bounded away from zero and infinity.

Condition 2 is standard on the first- and second-order derivatives of $g(\cdot; \cdot)$, ensuring its smoothness. If there exist two uniform envelope functions $B_{n,1}(\cdot)$ and $B_{n,2}(\cdot)$ with $E\{B_{n,1}^2(X_i)\} < \infty$ and $E\{B_{n,2}(X_i)\} < \infty$ such that $|\partial g_j(X; \theta) / \partial \theta_l| \leq B_{n,1}(X)$ and $|\partial^2 g_j(X; \theta) / \partial \theta_{l_1} \partial \theta_{l_2}| \leq B_{n,2}(X)$ ($j = 1, \dots, r; l, l_1, l_2 = 1, \dots, p$) for any $\theta \in \Theta$, then Condition 2 holds automatically. More generally, if there exist envelope functions $B_{n,jl}(\cdot)$ such that $|\partial g_j(X; \theta) / \partial \theta_l|^2 \leq B_{n,jl}(X)$ ($j = 1, \dots, r; l = 1, \dots, p$) for any $\theta \in \Theta$, and $|E\{B_{n,jl}^k(X_i)\}| \leq H_1 k! H_2^{k-2}$ for any $k \geq 2$, where H_1 and H_2 are two uniform positive constants independent of j and l , then Petrov (1995, Theorem 2.8) implies that $\sup_{1 \leq j \leq r} \sup_{1 \leq l \leq p} n^{-1} \sum_{i=1}^n B_{n,jl}(X_i) = O_p(1)$, provided $\log(rp) = o(n)$, and Condition 2 holds as well. Conditions 3 and 4 put constraints on the moments of estimating functions. In fact, the order $O_p(1)$ required in Conditions 2 and 4 can be replaced by $O_p(\varpi_n)$ with some diverging

sequence ϖ_n , and our main results remain valid. We use $O_p(1)$ here for ease of presentation. Condition 5 ensures the nonsingularity of the covariance matrix of $g_i(\theta_0)$. Under those conditions, we then have the following theorem.

THEOREM 1. *Under Conditions 1–5, if $|\theta_{\mathcal{M}}^* - \theta_{0,\mathcal{M}}|_1 = O_p(\xi_{1,n})$ and $|\theta_{\mathcal{M}^c}^* - \theta_{0,\mathcal{M}^c}|_1 = O_p(\xi_{2,n})$ for some $\xi_{1,n} \rightarrow 0$ and $\xi_{2,n} \rightarrow 0$, the following results hold:*

- (i) *If m is fixed, then $\ell_{A_n}^*(\theta_{0,\mathcal{M}}) \rightarrow \chi_m^2$ in distribution as $n \rightarrow \infty$, provided that $n\xi_{2,n}^2(\tau^2 + \xi_{1,n}^2 + \xi_{2,n}^2) = o(1)$ and $\omega_n^2 \log r = o(1)$.*
- (ii) *If m diverges with n , then $(2m)^{-1/2}\{\ell_{A_n}^*(\theta_{0,\mathcal{M}}) - m\} \rightarrow N(0, 1)$ in distribution as $n \rightarrow \infty$, provided that $m\xi_{2,n} = o(1)$, $m\omega_n^2(m + \log r) = o(1)$, $m^3n^{2/\gamma-1} = o(1)$ and $mn\xi_{2,n}^2(\tau^2 + \xi_{1,n}^2 + \xi_{2,n}^2) = o(1)$.*

To ensure the validity of the inferential procedure in § 3.1, a consistent initial estimate θ^* is required in Theorem 1. Theorem 1 also suggests that a faster convergence rate of θ^* would allow higher dimensionality of r . Define $s^* = |\{1 \leq k \leq m : \theta_{0,k} \neq 0\}|$ and select θ^* as $\hat{\theta}_{\text{PEL}}$ given by (2). It follows immediately from Proposition 1 that $\xi_{1,n} = s^*\alpha_n$ and $\xi_{2,n} = (s - s^*)\alpha_n$. Theorem 1 holds provided that $m(s - s^*)\alpha_n = o(1)$, $m\omega_n^2(m + \log r) = o(1)$, $m^3n^{2/\gamma-1} = o(1)$, $mn\tau^2(s - s^*)^2\alpha_n^2 = o(1)$ and $mns^2(s - s^*)^2\alpha_n^4 = o(1)$. For the bias-corrected penalized empirical likelihood estimator $\hat{\theta}_{\text{PELbc}}$ of Chang et al. (2018) in (3), \sqrt{n} -consistency is achievable for estimating each component of $\theta_{0,S}$. In such a case, $\xi_{1,n} = s^*n^{-1/2}$ and $\xi_{2,n} = (s - s^*)n^{-1/2}$, and Theorem 1 holds when $m\omega_n^2(m + \log r) = o(1)$, $m^3n^{2/\gamma-1} = o(1)$, $m\tau^2(s - s^*)^2 = o(1)$ and $m(s^2 + m)(s - s^*)^2n^{-1} = o(1)$. In addition, if m is fixed, Theorem 1 holds if $\omega_n^2 \log r = o(1)$, $s^2\tau^2 = o(1)$ and $s^4n^{-1} = o(1)$. Therefore, with a polynomial decay rate ω_n when approximating a_k in Condition 1, our method accommodates exponentially diverging r as $n \rightarrow \infty$.

Here are some example scenarios regarding the above conditions. In a just-identified case, let us consider linear regression $Y_i = W_i^T\theta_0 + \varepsilon_i$, where ε_i is independent of predictor variables $W_i = (W_{i,1}, \dots, W_{i,p})^T \in \mathbb{R}^p$, and the estimating function is $g(X_i; \theta) = W_i(Y_i - W_i^T\theta)$ with $X_i = (Y_i, W_i^T)^T$. Condition 2 is equivalent to $\max_{1 \leq j_1, j_2 \leq p} n^{-1} \sum_{i=1}^n W_{i,j_1}^2 W_{i,j_2}^2 = O_p(1)$. Then, Conditions 2–4 hold if W_i and ε_i are sub-Gaussian and $\log p = o(n)$. In an over-identified case, we consider a linear regression model $Y_i = W_i^T\theta_0 + \varepsilon_i$ with instrumental variables $Z_i \in \mathbb{R}^r$ for $r > p$ and the estimating function $g(X_i; \theta) = Z_i(Y_i - W_i^T\theta)$. Analogously to the just-identified case, Conditions 2–5 are met if W_i, Z_i and ε_i are sub-Gaussian, $\log p = o(n)$, and the eigenvalues of $E(\varepsilon_i^2 Z_i^{\otimes 2})$ are uniformly bounded away from zero and infinity. If ε_i is independent of Z_i , the last requirement is equivalent to the boundedness of the eigenvalues of $\text{cov}(Z_i)$.

4.2. Over-identification test

Let $q = |\mathcal{J}|$ and $h_n = |\mathcal{R}_n|$. To investigate the properties of the over-identification test in § 3.2, we impose the following condition.

Condition 6. With γ specified in Condition 3, there is a uniform constant $C_3 > 0$ such that $E\{|\partial g_{i,j}(\theta_0)/\partial \theta_l|^\gamma\} < C_3$ for any $j = 1, \dots, r$ and $l = 1, \dots, p$. In addition, assume

$$\sup_{\theta \in \Theta} \max_{j \in \mathcal{J}} \frac{1}{n} \sum_{i=1}^n |g_{i,j}(\theta)|^\gamma = O_p(1). \quad (12)$$

The following theorem is for the Type I error of the proposed over-identification test.

THEOREM 2. Assume B defined in (9) to satisfy that $\|B\|_\infty$ is bounded away from infinity, and all the eigenvalues of $([E_{\mathcal{R}_n}\{\nabla_{\theta_S} g_{i,\mathcal{R}_n}(\theta_0)\}]^T)^{\otimes 2}$ and $([E_{\mathcal{J}}\{\nabla_{\theta_S} g_{i,\mathcal{J}}(\theta_0)\}]^T)^{\otimes 2}$ are uniformly bounded away from zero and infinity. Select $\hat{\theta}_n = \hat{\theta}_{\text{PELbc}}$ defined as (3), with Proposition 2 in §2.2 being satisfied. Let $h_n \geq s$ and Conditions 2, 5 and 6 hold. If $s^3 h_n^2 \log^4 q = o(n)$, $sh_n^2(\log h_n) \log^4 q = o(n)$, $s^6 \log^2 q = o(n)$, $s^2 h_n \log^5 q = o(n)$, $n\phi_n^2 \log q = o(1)$ and $q^2(\log n)^{3\gamma+3} = o(n^{\gamma-2})$, then $\sup_{0 < \alpha < 1} |\text{pr}_{H_0}(T_n > \hat{c}v_\alpha) - \alpha| \rightarrow 0$ as $n \rightarrow \infty$, where $\hat{c}v_\alpha$ is estimated in (11).

Theorem 2 shows the Type I error of the proposed over-identification test is approximately α . If we select $\mathcal{J} = \mathcal{R}_n$, which means $q = h_n$, then Theorem 2 is valid if $s^3 h_n^2 \log^4 h_n = o(n)$, $sh_n^2 \log^5 h_n = o(n)$, $s^6 \log^2 h_n = o(n)$ and $n\phi_n^2 \log h_n = o(1)$. As discussed in Chang et al. (2018), Proposition 2 holds even if r and p grow at some exponential rate n with $h_n \ll n$. Therefore, the proposed over-identification test can be employed in the case that r and p diverge exponentially.

To show the test is consistent, we assume that under the alternative hypothesis H_1 , there exists some $\varsigma_n > 0$ that may decay to zero as $n \rightarrow \infty$ such that

$$\inf_{\theta \in \Theta} |E\{g_i(\theta)\}|_\infty \geq \varsigma_n. \tag{13}$$

Let $\theta_* = \arg \inf_{\theta \in \Theta} |E\{g_i(\theta)\}|_\infty$ and $j_* = \arg \max_{1 \leq j \leq r} |E\{g_{i,j}(\theta_*)\}|$. We impose the following condition.

Condition 7. For ς_n specified in (13), $|\bar{g}_{j_*}(\hat{\theta}_n) - E\{g_{i,j_*}(\hat{\theta}_n)\}| = o_p(\varsigma_n)$.

The following theorem states the power of the proposed over-identification test under the alternative hypothesis.

THEOREM 3. Let (13) and Condition 7 hold under H_1 . Select \mathcal{J} satisfying $\mathcal{J} \supseteq \mathcal{R}_n$. If (12) holds and $\varsigma_n^{-2} n^{2/\gamma-1} \log q = o(1)$, then $\text{pr}_{H_1}(T_n > \hat{c}v_\alpha) \rightarrow 1$ as $n \rightarrow \infty$.

5. NUMERICAL STUDIES

5.1. Confidence set estimations

The methods in §3.1 are implemented to construct confidence sets in two simulation examples: a linear regression model and an analysis of longitudinal data using over-identified estimating equations. We use the function `slim` in the R package `flare` (R Development Core Team, 2020) to solve the optimization (7) with the tuning parameter $\tau = 0.5(n^{-1} \log p)^{1/2}$, which meets the conditions in our theory. The estimate (2) is used as the initial estimate θ^* in (6) and (7). The smoothly clipped absolute deviation penalty with local quadratic approximation (Fan & Li, 2001) is employed for both penalty functions $P_{1,\pi}(\cdot)$ and $P_{2,\nu}(\cdot)$ in (2) for all the numerical experiments. The extended Bayesian information criterion (Chen & Chen, 2008) is applied to determine the tuning parameters π and ν by a two-dimensional grid search. All simulation experiments are repeated 1000 times.

Example 1 (Linear regression model). We consider a linear regression model $Y_i = Z_i^T \theta_0 + \varepsilon_i$, where $\theta_0 = (1.5, 1.2, 1, 0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3, 0, \dots, 0)^T$ with the first 10 components being nonzero, and $Z_i \sim N(0, \Sigma_z)$ with $\Sigma_z = (\sigma_{z,kl})_{p \times p}$ and $\sigma_{z,kl} = I(k = l) + 0.5I(k \neq l)$. This is a just-identified model. Our method is compared with a few existing alternatives: the desparsified method of van de Geer et al. (2014) implemented by the R package `hdi`, the estimating equation approach of Neykov et al. (2018) implemented by the R package `clime`, and

Table 1. Empirical frequencies (%) of the estimated confidence intervals covering the truth in the linear regression example

(n, p, r)	Method	Level	θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	θ_7	θ_8	θ_9	θ_{10}
(50,100,100)	EL	90%	90.3	88.5	89.6	88.9	90.1	90.0	88.8	90.6	87.5	89.4
		95%	94.5	93.8	94.1	93.8	94.5	95.0	94.1	94.9	94.4	94.2
		99%	98.7	98.8	98.4	98.0	98.8	98.8	97.9	98.6	98.9	98.2
	EE	90%	92.1	92.2	90.6	91.9	92.4	92.9	94.6	95.8	94.2	96.0
		95%	97.8	95.6	95.9	95.9	96.5	96.7	96.8	97.8	97.6	97.9
		99%	99.7	98.5	99.1	99.3	99.2	99.6	99.3	99.9	99.6	99.5
	Desparsity	90%	88.9	88.8	88.1	88.0	88.2	88.2	85.5	87.6	87.4	85.5
		95%	93.6	94.0	94.1	94.3	94.7	93.5	93.0	93.8	94.5	92.3
		99%	98.7	99.4	98.9	99.1	99.3	98.6	98.8	98.5	98.5	98.5
	Debias	90%	94.3	93.0	94.0	93.9	93.2	94.3	92.6	92.7	93.5	91.5
		95%	96.3	96.5	96.5	97.3	96.5	97.4	96.5	96.0	96.7	95.8
		99%	98.5	99.5	99.5	99.6	99.3	99.6	99.3	98.7	99.8	99.0
(100,500,500)	EL	90%	88.3	88.7	89.3	89.0	88.9	89.7	88.2	88.1	89.2	89.0
		95%	93.2	94.1	94.1	93.8	93.9	93.5	94.3	93.4	94.5	94.2
		99%	98.2	98.4	98.3	98.8	98.4	98.7	98.9	98.0	98.9	98.4
	EE	90%	93.4	92.8	92.7	91.1	91.8	91.6	94.0	94.6	95.7	95.3
		95%	97.3	96.0	96.3	95.5	95.4	95.3	96.4	97.3	98.0	97.8
		99%	99.3	98.9	98.8	99.0	99.1	99.0	99.5	99.6	99.8	99.5
	Desparsity	90%	89.0	87.9	89.3	88.2	89.7	86.9	89.2	87.8	87.6	89.5
		95%	94.9	94.1	94.7	94.4	94.0	93.2	95.2	94.1	95.1	94.3
		99%	98.5	99.4	99.2	99.0	98.9	99.1	98.6	99.1	98.9	99.1
	Debias	90%	93.5	94.5	92.9	92.3	90.9	92.0	92.9	94.0	91.4	91.8
		95%	97.4	96.9	96.0	95.7	95.7	95.5	96.4	96.7	95.5	96.0
		99%	99.6	99.1	99.1	99.1	99.0	99.1	99.7	99.3	99.2	99.1

EL, the proposed method; EE, Neykov et al. (2018)’s method; Desparsity, van de Geer et al. (2014)’s method; Debias, Javanmard & Montanari (2014)’s method.

Table 2. Average lengths of the 95% confidence intervals in the linear regression example with $(n, p, q) = (50, 100, 100)$

Method	θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	θ_7	θ_8	θ_9	θ_{10}
EL	0.815	0.769	0.771	0.781	0.758	0.749	0.770	0.760	0.789	0.782
EE	0.905	0.884	0.887	0.862	0.889	0.867	0.857	0.867	0.854	0.869
Desparsity	0.837	0.841	0.852	0.856	0.867	0.846	0.841	0.851	0.847	0.864
Debias	0.840	0.841	0.845	0.849	0.859	0.840	0.842	0.851	0.849	0.867

the debiased approach of Javanmard & Montanari (2014) using the code downloaded from the author’s website. We consider two settings with $(n, p, r) = (50, 100, 100)$ and $(100, 500, 500)$, respectively.

Table 1 reports the empirical frequencies of the estimated univariate confidence intervals that cover the truth. At each level, the empirical coverage probabilities are close to the nominal level. We observe that our method has similar coverage accuracy to the desparsified method and better coverage accuracy than the estimating equation approach and the debiased approach. The average lengths of the 95% confidence intervals are reported in Table 2. We can see that the proposed empirical-likelihood-based method outperforms the other methods. The two- and three-dimensional confidence regions constructed from our method are plotted in the Supplementary Material.

Table 3. Empirical frequencies (%) of the empirical-likelihood-based confidence intervals covering the truth in the repeated measurements example

(n, p, r)	Level	θ_1	θ_2	θ_3	θ_4	θ_5
(50,100,200)	90%	87.3	88.3	89.6	89.9	88.9
	95%	93.4	93.6	94.9	94.5	93.8
	99%	97.5	98.0	98.8	98.4	98.2
(100,200,400)	90%	89.3	89.1	92.5	92.5	88.9
	95%	93.8	94.5	96.4	96.2	94.8
	99%	98.0	98.9	99.2	98.9	98.6

Table 4. Comparison of the lengths of the 95% empirical-likelihood-based confidence intervals in the repeated measurements example with $(n, p, r) = (50, 100, 200)$

Method	θ_1	θ_2	θ_3	θ_4	θ_5
One estimating equation	0.325	0.329	0.323	0.322	0.323
Two estimating equations	0.289	0.293	0.285	0.288	0.285

Example 2 (Regression model with repeated measurements). For m_i ($i = 1, \dots, n$) repeated measurements, we consider the model $Y_{ij} = Z_{ij}^T \theta_0 + \epsilon_{ij}$ ($i = 1, \dots, n; j = 1, \dots, m_i$), where $\theta_0 = (3, 1.5, 0, 0, 2, 0, \dots, 0)^T$, $Z_{ij} \sim N(0, \Sigma_z)$ with $\Sigma_z = (\sigma_{z,kl})_{p \times p}$ and $\sigma_{z,kl} = 0.3^{|k-l|}$, and $(\epsilon_{i1}, \dots, \epsilon_{im_i})^T \sim N(0, \Sigma_\epsilon)$ with $\Sigma_\epsilon = (\sigma_{\epsilon,kl})_{p \times p}$ and $\sigma_{\epsilon,kl} = I(k=l) + 0.5I(k \neq l)$. Denote by $Y_i = (Y_{i1}, \dots, Y_{im_i})^T$ and $Z_i = (Z_{i1}^T, \dots, Z_{im_i}^T)^T$, respectively, the response variables and the corresponding predictor variables, and write $X_i = (Y_i^T, Z_i^T)^T$. To incorporate the within-subject dependence, we apply the estimating function $g(X_i; \theta) = [\{Z_i^T K_i^{-1/2} M_1 K_i^{-1/2} (Y_i - Z_i^T \theta)\}^T, \dots, \{Z_i^T K_i^{-1/2} M_\kappa K_i^{-1/2} (Y_i - Z_i^T \theta)\}^T]^T$, as in Qu et al. (2000), where $K_i \in \mathbb{R}^{m_i \times m_i}$ is a diagonal matrix of the conditional variance for subject i , and M_j ($j = 1, \dots, \kappa$) are working correlation matrices. We set $m_i = 3$ and $\kappa = 2$ in this simulation, with M_1 being the identity matrix of order 3 and M_2 being compound symmetry with diagonal elements of 1 and off-diagonal elements of 0.5. Since $\kappa = 2$, there are twice estimating equations as parameters, so this is an over-identified case with $r = 2p$. For the first five components of the parameters, the empirical frequencies that the estimated confidence intervals cover the truth are reported in Table 3. Similar to Example 1, we see satisfactory performance of the proposed method in the over-identified case. The two- and three-dimensional empirical-likelihood-based confidence regions are in the Supplementary Material.

The sequential procedure described at the end of § 3.1 is also implemented by finding two estimating equations for each individual component in this over-identified example. We found similar coverage accuracy for such a method. To see its advantage, Table 4 compares the lengths of the 95% confidence intervals using one estimating equation versus two estimating equations. It can be seen that the confidence intervals using two estimating equations are about 11% shorter than those using only one estimating equation, which shows a potential advantage of our method, since more information can be retained through the over-identification of estimating functions.

5.2. Over-identification test

To assess the over-identification test in § 3.2, we consider the mean of a normal distributed random vector $X = (X_1, \dots, X_p)^T$, where only the first component X_1 has a nonzero mean of 5 and the remaining components all have mean zero. The first p estimating functions are simply

Table 5. Empirical percentages of rejecting H_0 in the model specification test example. Case 1 corresponds to a correct model specification and Case 2 corresponds to a model misspecification; Method 1 uses the set of estimating functions selected by \mathcal{R}_n , and Method 2 uses all the estimating functions

	σ_{11}	(n, p)	Method 1	Method 2
Case 1	5^2	(50, 1)	0.056	0.056
		(50, 10)	0.061	0.061
		(50, 50)	0.061	0.002
		(50, 100)	0.058	0.002
		(100, 100)	0.047	0.002
Case 2	$5^2 \times 0.7$	(50, 1)	0.492	0.492
		(50, 10)	0.521	0.521
		(50, 50)	0.580	0.082
		(50, 100)	0.601	0.054
		(100, 100)	0.738	0.286
	$5^2 \times 0.5$	(50, 1)	0.915	0.915
		(50, 10)	0.911	0.911
		(50, 50)	0.883	0.143
		(50, 100)	0.890	0.257
		(100, 100)	0.994	0.381
	$5^2 \times 0.3$	(50, 1)	1.000	1.000
		(50, 10)	1.000	1.000
		(50, 50)	0.998	0.167
		(50, 100)	1.000	0.743
		(100, 100)	1.000	0.294

from the components of $g(X; \theta) = X - \theta$ with $\theta_0 = (5, 0, \dots, 0)^T$. In addition, we impose an extra moment restriction $g_{p+1}(X; \theta) = X_1^2 - \theta_1^2 - 25$, where θ_1 is the first component of θ . In this setting, the number of estimating equations is $r = p + 1$. We consider the following two cases:

Case 1. The covariance matrix $\Sigma = (\sigma_{ij})_{p \times p}$ is compound symmetry with diagonal $\sigma_{11} = 5^2$ and $\sigma_{ii} = 1$ for all $i \neq 1$. All off-diagonal elements $\sigma_{ij} = 0.3$ for $i \neq j$.

Case 2. The covariance matrix $\Sigma = (\sigma_{ij})_{p \times p}$ is compound symmetry with diagonal $\sigma_{11} = 5^2 \times a$ with $a < 1$ and $\sigma_{ii} = 1$ for all $i \neq 1$. All off-diagonal elements $\sigma_{ij} = 0.3$ for $i \neq j$.

Clearly, the moment conditions are correctly specified in Case 1, but not in Case 2. We conduct the experiments for a few settings of (n, p, r) in this example. We apply (11) to obtain the critical value of the test. Further, we compare the performances of the test by using two different choices of \mathcal{J} in (8). The first one, referred to as Method 1, uses the set \mathcal{R}_n of estimating functions selected by (2). The other one, referred to as Method 2, simply uses \mathcal{J} containing all estimating functions. We report in Table 5 the empirical percentages rejecting H_0 at $\alpha = 0.05$ level. In Case 1, we expect the rate to be close to 0.05, which is indeed the case for our advocated Method 1 of choosing \mathcal{R}_n as \mathcal{J} . Method 2 works well when the dimension is low, but it gets much worse with p and r getting close to n . In Case 2, the closer the rate is to 1, the better the power is for the testing procedure. We tried three settings with $a = 0.7, 0.5$ and 0.3 , respectively, where a smaller value in a can be viewed as more severe violation of H_0 . One can clearly see that the advocated method works quite well in terms of providing a more powerful test with the right choice of estimating functions. The power improves consistently for more severe violation of the null hypothesis. As

Table 6. The estimated regression coefficients and 95% confidence intervals for the selected variables associated with moderate to vigorous physical activity over time using penalized empirical likelihood, as compared to linear regression. The column C.I. Ratio lists the ratio of the 95% confidence intervals constructed from over-identified estimating functions and the linear models

Variable	Repeated	Linear Reg.	C.I. Ratio
TAAG (time)	-0.280 (-0.310, -0.210)	-0.297 (-0.356, -0.237)	0.840
Body mass index	-0.056 (-0.136, -0.016)	-0.098 (-0.163, -0.041)	0.984
Self-management strategies	0.072 (0.052, 0.172)	0.126 (0.065, 0.186)	0.992
Social support from friends	0.118 (0.048, 0.148)	0.079 (0.023, 0.135)	0.890
Smoker	-0.102 (-0.132, -0.022)	-0.044 (-0.100, 0.011)	0.991
Father's education	0.059 (0.029, 0.139)	0.087 (0.023, 0.151)	0.859
Mother's education	0.067 (0.037, 0.147)	0.073 (0.010, 0.137)	0.862
Number of parks within 1 mile	0.088 (0.058, 0.178)	0.126 (0.061, 0.182)	0.992

for the Method 2, it works well when p and r are small, but it becomes powerless in moderate high-dimensional cases, which is consistent with our discussions in § 3.2.

5.3. Multi-level longitudinal study of physical activity among girls

We analyse a dataset from a longitudinal study of physical activities among girls from adolescence into young adulthood. The main goal of this study is to identify individual, social and environmental factors associated with moderate to vigorous physical activity among those girls over time using a multilevel approach. An initial cohort of 730 girls were randomly recruited in Maryland in 2006, and 428 girls had complete assessments at all three study periods in 2006 ($n = 730$), 2009 ($n = 589$) and 2015 ($n = 460$) at ages 14, 17 and 23. The response variable, moderate to vigorous physical activity minutes, was assessed from accelerometers, and over 800 variables were collected, including: (i) demographic and psychosocial information, individual- and social-level variables; (ii) height, weight and triceps skinfold to assess body composition; and (iii) geographical information systems and self-report for neighbourhood-level variables.

In this example we consider an over-identified model specification with $r > p$; see the longitudinal data example in § 5.1. The same estimating equations and basis matrices M_1 and M_2 of size 3×3 as in § 5.1 are used. Eight predictor variables out of 34 screened variables were selected in the model for the logarithm of response; see Table 6. The second column of Table 6 provides the regression coefficients together with the 95% componentwise confidence intervals estimated by the approach in § 3.1 using the over-identified estimating equations. We see that none of the 95% confidence intervals contain 0, showing that all the selected variables are statistically significant in the model. We applied the over-identification test in § 3.2, and found no significant statistical evidence against the model specification with over-identification.

In the selected model, the first variable, TAAG, is an ordinal variable indicating the time of study when data were collected. As expected, physical activities decreased significantly over time among young females. The variable self-management strategies, an aggregated variable of eight questionnaire items, and social support from friends, a sum of three questionnaire items, are positively correlated with the response. Parents' education and number of parks within 1 mile of home have a positive impact on physical activities. On the other hand, BMI and being a smoker are negatively correlated with physical activities. Our findings are consistent with the previous results (Young et al., 2014, 2018; Grant et al., 2015).

As for comparisons, we apply an alternative approach using a linear regression model. The third column of Table 6 reports the componentwise point estimates and confidence intervals for

the eight selected variables. All the confidence intervals in this approach are wider than those from the over-identified estimating equations; the ratios of the interval lengths are reported in the fourth column of Table 6. In particular, the variable smoker is significant when applying the over-identified approach, but becomes insignificant if ignoring the within-subject dependence.

The two-dimensional confidence regions for TAAG, i.e., time, versus other covariates are given in the Supplementary Material. The constructed confidence regions are not symmetric at the estimate, reflecting the merit that the empirical-likelihood-based confidence region is data oriented and free of shape constraint.

6. DISCUSSION

It would be interesting to extend high-dimensional statistical inference to a setting with some unknown functional-valued parameters. A possible strategy is to apply the sieve method (Ai & Chen, 2003) to approximate the functional-valued parameters with some linear combinations of diverging numbers of given basis functions. Then the estimation of unknown functional-valued parameters is transferred to the estimation of the coefficients in the sieve approximation; and the frameworks of Chang et al. (2018) and this paper apply. Nevertheless, accommodating functional-valued parameters will introduce foundational changes in the settings. New developments in both theory and methods are beyond the scope of this study. We plan to carefully investigate the problem in a future project.

ACKNOWLEDGEMENT

We are grateful to the editor, associate editor and referees for their constructive comments. Chang and Chen were supported by the Fundamental Research Funds for the Central Universities and the National Natural Science Foundation of China. Chang was also supported by the Fok Ying-Tong Education Foundation, and the Center of Statistical Research and the Joint Lab of Data Science and Business Intelligence at Southwestern University of Finance and Economics. Tang and Wu were supported by the U.S. National Science Foundation. Wu was also supported by the U.S. National Institutes of Health.

SUPPLEMENTARY MATERIAL

Supplementary material available at Biometrika online includes all the technical proofs, figures of the empirical-likelihood-based confidence regions in § 5, and an extra simulation example.

APPENDIX

Write $\theta_0 = (\theta_{0,1}, \dots, \theta_{0,p})^T$. Recall that $\mathcal{S} = \{1 \leq k \leq p : \theta_{0,k} \neq 0\}$. To investigate the properties of the penalized empirical likelihood estimator $\hat{\theta}_{\text{PEL}}$ defined in (2), we need the following technical conditions.

Condition A1. Assume that $\inf_{\theta \in \{\theta = (\theta_{\mathcal{S}}^T, \theta_{\mathcal{S}^c}^T)^T \in \Theta : |\theta_{\mathcal{S}} - \theta_0, \mathcal{S}|_{\infty} > \varepsilon, \theta_{\mathcal{S}^c} = 0\}} |E\{g_i(\theta)\}|_{\infty} \geq \Delta(\varepsilon)$ for any $\varepsilon > 0$, where $\Delta(\cdot)$ is a function satisfying $\liminf_{\varepsilon \rightarrow 0^+} \varepsilon^{-\beta} \Delta(\varepsilon) \geq K_1$ for some uniform constants $K_1 > 0$ and $\beta \in (0, 1]$.

For some $C_* \in (0, 1)$, define $\mathcal{M}_{\theta} = \{1 \leq j \leq r : |\bar{g}_j(\theta)| \geq C_* \nu \rho_2'(0^+)\}$ for any $\theta \in \Theta$. Let $b_n = \max\{a_n, \nu^2\}$ with $a_n = \sum_{k=1}^p P_{1,\pi}(|\theta_{0,k}|)$, and

$$l_n = \max_{\theta \in \{\theta = (\theta_S^T, \theta_{S^c}^T)^T \in \Theta : |\theta_S - \theta_{0,S}|_\infty \leq c_n, \theta_{S^c} = 0\}} |\mathcal{M}_\theta| \tag{A1}$$

for some $c_n \rightarrow 0$ satisfying $b_n^{1/(2\beta)} c_n^{-1} \rightarrow 0$ with β as specified in Condition A1. We assume $l_n \geq s$ with $s = |S|$.

Condition A2. For any X and $j = 1, \dots, r$, $g_j(X; \theta)$ is continuously differentiable with respect to θ , and $\max_{1 \leq j \leq r} \max_{k \notin S} E\{\sup_{\theta \in \Theta} |\partial g_{i,j}(\theta) / \partial \theta_k|\} \leq K_2$ for some uniform constant $K_2 > 0$.

Condition A3. For some uniform constants $K_3 > 0$ and $\gamma > 4$, $\max_{1 \leq j \leq r} E\{\sup_{\theta \in \Theta} |g_{i,j}(\theta)|^\gamma\} \leq K_3$.

Condition A4. Let $V_{\mathcal{F}}(\theta_0) = E\{g_{i,\mathcal{F}}(\theta)^{\otimes 2}\}$ for any $\mathcal{F} \subset \{1, \dots, r\}$. There exist uniform constants $0 < K_4 < K_5$ such that $K_4 < \lambda_{\min}\{V_{\mathcal{F}}(\theta_0)\} \leq \lambda_{\max}\{V_{\mathcal{F}}(\theta_0)\} < K_5$ for any \mathcal{F} with $|\mathcal{F}| \leq l_n$, where l_n is defined in (A1).

Condition A5. The following hold: $\sup_{\theta \in \Theta} \max_{1 \leq j \leq r} \max_{1 \leq k \leq p} n^{-1} \sum_{i=1}^n |\partial g_{i,j}(\theta) / \partial \theta_k| = O_p(1)$ and $\sup_{\theta \in \Theta} \max_{1 \leq j \leq r} n^{-1} \sum_{i=1}^n |g_{i,j}(\theta)|^4 = O_p(1)$.

Condition A6. For c_n specified in (A1), $\max_{k \in S} \sup_{0 < t < |\theta_{0,k}| + c_n} P'_{1,\pi}(t) = O(\chi_n)$ for some $\chi_n \rightarrow 0$.

Conditions A1–A6 are simplified versions of Conditions 1–6 in Chang et al. (2018), and we refer to Chang et al. (2018) for detailed discussion of their validity. With the additional assumption $b_n = o(\min_{k \in S} |\theta_{0,k}|^{2\beta})$, that the signal strength of the nonzero components of θ_0 does not diminish to zero too fast, Condition A6 can be replaced by

$$\max_{k \in S} \sup_{c|\theta_{0,k}| < t < c^{-1}|\theta_{0,k}|} P'_{1,\pi}(t) = O(\chi_n). \tag{A2}$$

for some constant $c \in (0, 1)$. For those asymptotically unbiased penalties like the smoothly clipped absolute deviation penalty (Fan & Li, 2001) and the minimax concave penalty (Zhang, 2010), $\chi_n = 0$ in (A2) for sufficiently large n if $b_n = o(\min_{k \in S} |\theta_{0,k}|^{2\beta})$. Define $\kappa_n = \max\{l_n^{1/2} n^{-1/2}, s^{1/2} \chi_n^{1/2} b_n^{1/(4\beta)}\}$. Based on Conditions A1–A6, Proposition 1 holds provided that the following restrictions are satisfied:

$$\begin{aligned} \log r = o(n^{1/3}), \quad s^2 l_n b_n^{1/\beta} = o(1), \quad l_n^2 n^{-1} \log r = o(1), \quad \max\{b_n, l_n \kappa_n^2\} = o(n^{-2/\gamma}), \\ l_n^{1/2} \kappa_n = o(v) \quad \text{and} \quad l_n^{1/2} \max\{l_n v, s^{1/2} \chi_n^{1/2} b_n^{1/(4\beta)}\} = o(\pi). \end{aligned} \tag{A3}$$

The convergence rate α_n specified in Proposition 1 equals $b_n^{1/(2\beta)}$. If $b_n = o(\min_{k \in S} |\theta_{0,k}|^{2\beta})$ and $P_{1,\pi}(\cdot)$ is selected as the asymptotically unbiased penalty, χ_n specified in (A2) can be selected as 0. Since $a_n = O(s\pi)$, the restrictions (A3) in this scenario can be simplified to $\log r = o(n^{1/3})$, $l_n = o(\min\{n^{1/2}(\log r)^{-1/2}, n^{1/2-1/\gamma}\})$, and the tuning parameters v and π satisfy $l_n n^{-1/2} = o(v)$, $v = o(\min\{s^{-\beta} l_n^{-\beta/2}, n^{-1/\gamma}\})$, $l_n^{3/2} v = o(\pi)$ and $\pi = o(\min\{s^{-2\beta-1} l_n^{-\beta}, s^{-1} n^{-2/\gamma}\})$. If $\log r = o(n^{1/3})$ and $l_n = o(\min\{n^{(\gamma-4)/(5\gamma)} s^{-2/5}, n^{1/(2\beta+5)} s^{-(4\beta+2)/(2\beta+5)}\})$, there exist suitable selections of v and π satisfying these conditions. Furthermore, we need two additional conditions, listed below, to construct Proposition 2.

Condition A7. For any X and $j = 1, \dots, p$, $g_j(X; \theta)$ is twice continuously differentiable with respect to θ , and $\sup_{\theta \in \Theta} \max_{1 \leq j \leq r} \max_{k_1, k_2 \in S} n^{-1} \sum_{i=1}^n |\partial^2 g_{i,j}(\theta) / \partial \theta_{k_1} \partial \theta_{k_2}|^2 = O_p(1)$.

Condition A8. Let $Q_{\mathcal{F}} = (E\{\nabla_{\theta_S} g_{i,\mathcal{F}}(\theta_0)\})^{\otimes 2}$ for any $\mathcal{F} \subset \{1, \dots, r\}$. There exist uniform constants $0 < K_6 < K_7$ such that $K_6 < \lambda_{\min}(Q_{\mathcal{F}}) \leq \lambda_{\max}(Q_{\mathcal{F}}) < K_7$ for any \mathcal{F} with $s \leq |\mathcal{F}| \leq l_n$, where l_n is defined as in (A1).

Based on Conditions A1–A8, Proposition 2 holds provided that the following restrictions are satisfied:

$$\begin{aligned} \log r &= o(n^{1/3}), \quad b_n = o(n^{-2/\gamma}), \quad ns\chi_n^2 = o(1), \quad nl_n\kappa_n^4 \max\{s, n^{2/\gamma}\} = o(1), \\ l_n^2(\log r) \max\{s^3 b_n^{1/\beta}, l_n n^{-1} \log r\} &= o(1), \quad nl_n s^2 \max\{l_n^2 \nu^4, s^2 \chi_n^2 b_n^{1/\beta}\} = o(1), \\ l_n^{1/2} \kappa_n &= o(\nu) \quad \text{and} \quad l_n^{1/2} \max\{l_n \nu, s^{1/2} \chi_n^{1/2} b_n^{1/(4\beta)}\} = o(\pi). \end{aligned} \quad (\text{A4})$$

Notice that $a_n = O(s\pi)$. Under the reasonable case $\chi_n = 0$ and $l_n \sim s$, restrictions (A4) hold provided that $\log r = o(n^{1/3})$ and $s = o[\min\{n^{1/9}, n^{1/(10\beta+7)}(\log r)^{-2\beta/(10\beta+7)}, n^{(\gamma-4)/(7\gamma)}\}]$, and the tuning parameters ν and π satisfy the conditions $\pi = o[\min\{n^{-2/\gamma} s^{-1}, s^{-5\beta-1}(\log r)^{-\beta}\}]$, $\nu = o[\min\{n^{-1/\gamma}, s^{-5\beta/2}(\log r)^{-\beta/2}, n^{-1/4} s^{-5/4}\}]$, $s^{3/2} \nu = o(\pi)$ and $sn^{-1/2} = o(\nu)$. Due to $s = o[\min\{n^{1/9}, n^{1/(10\beta+7)}(\log r)^{-2\beta/(10\beta+7)}, n^{(\gamma-4)/(7\gamma)}\}]$, there exist suitable selections of ν and π satisfying these conditions.

REFERENCES

- AI, C. & CHEN, X. (2003). Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica*, **71**, 1795–843.
- BELLONI, A. & CHERNOZHUKOV, V. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli*, **19**, 521–47.
- BICKEL, P., RITOV, Y. & TSYBAKOV, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.*, **37**, 1705–32.
- BRAVO, F., ESCANCIANO, J. C. & VAN KEILEGOM, I. (2020). Two-step semiparametric empirical likelihood inference. *Ann. Statist.*, **48**, 1–26.
- BÜHLMANN, P. & VAN DE GEER, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. New York: Springer.
- CANDÈS, E. & TAO, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n . *Ann. Statist.*, **35**, 2313–51.
- CHANG, J., CHEN, S. X. & CHEN, X. (2015). High-dimensional generalized empirical likelihood for moment restrictions with dependent data. *J. Economet.*, **185**, 283–304.
- CHANG, J., TANG, C. Y. & WU, Y. (2013). Marginal empirical likelihood and sure independence feature screening. *Ann. Statist.*, **41**, 2132–48.
- CHANG, J., TANG, C. Y. & WU, Y. (2016). Local independence feature screening for nonparametric and semiparametric models by marginal empirical likelihood. *Ann. Statist.*, **44**, 515–39.
- CHANG, J., TANG, C. Y. & WU, T. T. (2018). A new scope of penalized empirical likelihood with high-dimensional estimating equations. *Ann. Statist.*, **46**, 3185–216.
- CHANG, J., ZHENG, C., ZHOU, W.-X. & ZHOU, W. (2017). Simulation-based hypothesis testing of high-dimensional means under covariance heterogeneity. *Biometrics*, **73**, 1300–10.
- CHEN, J. & CHEN, Z. (2008). Extended Bayesian information criterion for model selection with large model space. *Biometrika*, **95**, 759–71.
- CHEN, S. X. & CUI, H. J. (2006). On Bartlett correction of empirical likelihood in the presence of nuisance parameters. *Biometrika*, **93**, 215–20.
- CHEN, S. X. & CUI, H. J. (2007). On the second-order properties of empirical likelihood with moment restrictions. *J. Economet.*, **141**, 492–516.
- CHEN, S. X., PENG, L. & QIN, Y. (2009). Effects of data dimension on empirical likelihood. *Biometrika*, **96**, 711–22.
- CHERNOZHUKOV, V., CHETVERIKOV, D. & KATO, K. (2017). Central limit theorems and bootstrap in high dimensions. *Ann. Probab.*, **45**, 2309–52.
- FAN, J. & LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Statist. Assoc.*, **96**, 1348–60.
- FAN, J. & LV, J. (2010). A selective overview of variable selection in high-dimensional feature space. *Statist. Sinica*, **20**, 101–48.
- GRANT, E. M., YOUNG, D. R. & WU, T. T. (2015). Predictors for physical activity in adolescent girls using statistical shrinkage techniques for hierarchical longitudinal mixed effects models. *PLOS ONE*, **10**, e0125431.
- HALL, P. & LA SCALA, B. (1990). Methodology and algorithms of empirical likelihood. *Int. Statist. Rev.*, **58**, 109–27.
- HANSEN, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica*, **50**, 1029–54.
- HANSEN, L. P. & SINGLETON, K. J. (1982). Generalized instrumental variables estimation of nonlinear rational expectations models. *Econometrica*, **50**, 1269–86.

- HASTIE, T., TIBSHIRANI, R. & WAINWRIGHT, M. (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations*. Abingdon: Taylor & Francis.
- HJORT, N. L., MCKEAGUE, I. & VAN KEILEGOM, I. (2009). Extending the scope of empirical likelihood. *Ann. Statist.*, **37**, 1079–111.
- JAVANMARD, A. & MONTANARI, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *J. Mach. Learn. Res.*, **15**, 2869–909.
- LAZAR, N. A. & MYKLAND, P. A. (1999). Empirical likelihood in the presence of nuisance parameters. *Biometrika*, **86**, 203–11.
- LEE, J. D., SUN, D. L., SUN, Y. & TAYLOR, J. E. (2016). Exact post-selection inference, with application to the lasso. *Ann. Statist.*, **44**, 907–27.
- LENG, C. & TANG, C. Y. (2012). Penalized empirical likelihood and growing dimensional general estimating equations. *Biometrika*, **99**, 703–16.
- NEYKOV, M., NING, Y., LIU, J. S. & LIU, H. (2018). A unified theory of confidence regions and testing for high-dimensional estimating equations. *Statist. Sci.*, **33**, 427–43.
- NING, Y. & LIU, H. (2017). A general theory of hypothesis tests and confidence regions for sparse high-dimensional models. *Ann. Statist.*, **45**, 158–95.
- OWEN, A. B. (2001). *Empirical Likelihood*. New York: Chapman and Hall/CRC.
- PETROV, V. V. (1995). *Limit Theorems of Probability Theory: Sequences of Independent Random Variables*. Oxford: Oxford University Press.
- QIN, J. & LAWLESS, J. (1994). Empirical likelihood and general estimating equations. *Ann. Statist.*, **22**, 300–25.
- QIN, J. & LAWLESS, J. (1995). Estimating equations, empirical likelihood and constraints on parameter. *Can. J. Statist.*, **23**, 145–59.
- QU, A., LINDSAY, B. G. & LI, B. (2000). Improving estimating equations using quadratic inference functions. *Biometrika*, **87**, 823–36.
- R DEVELOPMENT CORE TEAM (2020). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0, <http://www.R-project.org>.
- SARGAN, J. D. (1958). The estimation of economic relationships using instrumental variables. *Econometrica*, **26**, 393–415.
- TIBSHIRANI, R. J., TAYLOR, J., LOCKHART, R. & TIBSHIRANI, R. (2016). Exact post-selection inference for sequential regression procedures. *J. Am. Statist. Assoc.*, **111**, 600–20.
- VAN DE GEER, S., BÜLMANN, P., RITOV, Y. & DEZEURE, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.*, **42**, 1166–202.
- YOUNG, D., SAKSVIG, B. I., WU, T. T., ZOOK, K., LI, X., CHAMPALOUX, S., GRIESER, M., LEE, S. & TREUTH, M. S. (2014). Multilevel correlates of physical activity for early, mid, and late adolescent girls. *J. Phys. Act. Health*, **11**, 950–60.
- YOUNG, D. R., COHEN, D., KOEBNICK, C., MOHAN, Y., SAKSVIG, B. I., SIDELL, M. & WU, T. T. (2018). Longitudinal associations of physical activity among females from adolescence to young adulthood. *J. Adolesc. Health*, **63**, 466–73.
- ZHANG, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.*, **38**, 894–942.
- ZHANG, C.-H. & ZHANG, S. S. (2013). Confidence intervals for low-dimensional parameters in high-dimensional linear models. *J. R. Statist. Soc. B*, **76**, 217–42.

[Received on 11 April 2019. Editorial decision on 20 January 2020]