Statistical Science
2018, Vol. 33, No. 3, 427–443
https://doi.org/10.1214/18-STS661

A Unified Theory of Confidence Regions and Testing for High-Dimensional Estimating Equations

Matey Neykov, Yang Ning, Jun S. Liu and Han Liu

Abstract. We propose a new inferential framework for constructing confidence regions and testing hypotheses in statistical models specified by a system of high-dimensional estimating equations. We construct an influence function by projecting the fitted estimating equations to a sparse direction obtained by solving a large-scale linear program. Our main theoretical contribution is to establish a unified Z-estimation theory of confidence regions for high-dimensional problems. Different from existing methods, all of which require the specification of the likelihood or pseudo-likelihood, our framework is likelihood-free. As a result, our approach provides valid inference for a broad class of high-dimensional constrained estimating equation problems, which are not covered by existing methods. Such examples include, noisy compressed sensing, instrumental variable regression, undirected graphical models, discriminant analysis and vector autoregressive models. We present detailed theoretical results for all these examples. Finally, we conduct thorough numerical simulations, and a real dataset analysis to back up the developed theoretical results.

Key words and phrases: Post-regularization inference, estimating equations, confidence regions, hypothesis tests, Dantzig selector, instrumental variables, graphical models, discriminant analysis, vector autoregressive models.

1. INTRODUCTION

Let us observe a sample of n, q-dimensional random vectors $\{\mathbf{Z}_i\}_{i=1}^n$. Denote with $\mathbf{Z}$ the $n \times q$ data matrix obtained by stacking all the vectors $\mathbf{Z}_i$. Let the function $\mathbf{t}(\mathbf{Z}, \boldsymbol{\beta}) : \mathbb{R}^{n\times q} \times \mathbb{R}^d \mapsto \mathbb{R}^d$ specify estimating equations $\mathbf{t}(\mathbf{Z}, \boldsymbol{\beta}) = 0$ (Godambe, 1991) for a d-dimensional unknown parameter $\boldsymbol{\beta}$, and further let $E_{\mathbf{t}}(\boldsymbol{\beta}) = \lim_{n\to\infty} \mathbb{E}\mathbf{t}(\mathbf{Z}, \boldsymbol{\beta})$[1] denote the limiting expected value of the function $\mathbf{t}(\mathbf{Z}, \boldsymbol{\beta})$ as $n \to \infty$. As an example given n i.i.d. observations $\mathbf{Z}_i$ and a function $\mathbf{h}$, this reduces to the classical Z-estimation setup $\mathbf{t}(\mathbf{Z}, \boldsymbol{\beta}) = n^{-1}\sum_{i=1}^n \mathbf{h}(\mathbf{Z}_i, \boldsymbol{\beta})$ and $E_{\mathbf{t}}(\boldsymbol{\beta}) = \mathbb{E}\mathbf{h}(\mathbf{Z}, \boldsymbol{\beta})$. For the purpose of parameter estimation, it is usually assumed that the estimating equation is unbiased in the sense that the true value $\boldsymbol{\beta}^*$ is the unique solution to $E_{\mathbf{t}}(\boldsymbol{\beta}) = 0$. When the dimension d is fixed and much smaller than the sample size n, inference on $\boldsymbol{\beta}^*$ can be obtained by solving the estimating equations $\mathbf{t}(\mathbf{Z}, \boldsymbol{\beta}) = 0$, and the asymptotic properties follow from

Matey Neykov is Assistant Professor, Department of Statistics and Data Science, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, USA (e-mail: mneykov@stat.cmu.edu). Yang Ning is Assistant Professor, Department of Statistical Science, Cornell University, Ithaca, New York 14853, USA (e-mail: yn265@cornell.edu). Jun S. Liu is Associate Professor, Department of Statistics, Harvard University, Cambridge, Massachusetts 02138, USA (e-mail: jliu@stat.harvard.edu). Han Liu is Professor, Department of Electrical Engineering and Computer Science and Department of Statistics, Northwestern University, Evanston, Illinois 60208, USA (e-mail: hanliu@northwestern.edu).

[1] Here and throughout, such limits should be understood with d being fixed to its current value.

the classical Z-estimation theory (van der Vaart, 1998). However, when $d > n$, directly solving $\mathbf{t}(\mathbf{Z}, \boldsymbol{\beta}) = 0$ is an ill-posed problem. To avoid this problem, a popular approach is to impose the sparsity assumption on $\boldsymbol{\beta}^*$, which motivates constrained Z-estimators in the following generic form (Cai, Liang and Rakhlin, 2014):

$$\begin{aligned} \widehat{\boldsymbol{\beta}} &= \operatorname{argmin} \|\boldsymbol{\beta}\|_1 \\ &\quad \text{subject to } \|\mathbf{t}(\mathbf{Z}, \boldsymbol{\beta})\|_\infty \le \lambda, \end{aligned} \tag{1.1}$$

where λ is a regularization parameter.

Assume that we can partition $\boldsymbol{\beta}$ as $(\theta, \boldsymbol{\gamma})$, where θ is a univariate parameter of interest and $\boldsymbol{\gamma}$ is a $(d-1)$-dimensional nuisance parameter. Similarly, we denote $\widehat{\boldsymbol{\beta}} = (\widehat{\theta}, \widehat{\boldsymbol{\gamma}})$ and $\boldsymbol{\beta}^* = (\theta^*, \boldsymbol{\gamma}^*)$. The goal of this paper is to develop a general estimating equation based framework to obtain valid confidence regions for θ^* under the regime that d is much larger than n. The proposed framework has a large number of applications. For instance, given a convex and smooth loss function (or negative log-likelihood) $\ell : \mathbb{R}^q \times \mathbb{R}^d \mapsto \mathbb{R}$, with i.i.d. data $\mathbf{Z}_i$, the inference on $\boldsymbol{\beta}$ can be conducted based on solving equations specified by the score function $\mathbf{t}(\mathbf{Z}, \boldsymbol{\beta}) = n^{-1} \sum_{i=1}^n \frac{\partial \ell(\mathbf{Z}_i, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$. Hence, inference on many high-dimensional problems with specifications of the loss function or the likelihood can be addressed through our framework. More importantly, the estimating equation method has an advantage over likelihood methods in that it usually only requires the specification of a few moment conditions rather than the entire probability distribution (Godambe, 1991). To see the advantage of our framework, we consider the following examples, which are naturally handled by estimating equations.

1.1 Examples

Linear Regression via Dantzig Selector (Candes and Tao, 2007). Assume that a linear model (also referred to as noisy compressed sensing) is specified by the following moment condition $\mathbb{E}(Y|\boldsymbol{X}) = \boldsymbol{X}^T\boldsymbol{\beta}^*$. Let $\mathbf{X} \in \mathbb{R}^{n\times d}$ be the design matrix stacking the i.i.d. covariates $\{\boldsymbol{X}_i\}_{i=1}^n$ and $\boldsymbol{Y} \in \mathbb{R}^n$ be the response vector with independent entries Y_i. Given the moment condition, we can easily construct the estimating function as $\mathbf{t}((\boldsymbol{Y}, \mathbf{X}), \boldsymbol{\beta}) = n^{-1}\mathbf{X}^T(\mathbf{X}\boldsymbol{\beta} - \boldsymbol{Y})$ and $E_{\mathbf{t}}(\boldsymbol{\beta}) = \mathbb{E}\mathbf{t}((\boldsymbol{Y}, \mathbf{X}), \boldsymbol{\beta})$. In addition, $E_{\mathbf{t}}(\boldsymbol{\beta}) = 0$ has the true value $\boldsymbol{\beta}^*$ as its unique root, provided that the second moment matrix $\boldsymbol{\Sigma}_{\boldsymbol{X}} := n^{-1}\mathbb{E}\mathbf{X}^T\mathbf{X}$ is positive definite. In the high-dimensional setting, Candes and Tao (2007) estimated $\boldsymbol{\beta}$ by the following Dantzig selector:

$$\begin{aligned} \widehat{\boldsymbol{\beta}} &= \operatorname{argmin} \|\boldsymbol{\beta}\|_1 \\ &\quad \text{such that } \|n^{-1}\mathbf{X}^T(\mathbf{X}\boldsymbol{\beta} - \boldsymbol{Y})\|_\infty \le \lambda. \end{aligned}$$

Instrumental Variables Regression (IVR). Similar to the previous example, consider the linear model $Y = \boldsymbol{X}^T\boldsymbol{\beta}^* + \varepsilon$. In economics' applications, it is not always reasonable to believe that the error and the design variables are uncorrelated, that is, $\mathbb{E}[\boldsymbol{X}\varepsilon] = 0$, which is a key condition ensuring the unbiasedness of the estimating equation and consequently the consistency of the Dantzig selector estimate. In such cases, one may use a set of *instrumental variables* $\boldsymbol{W} \in \mathbb{R}^d$ which are correlated with $\boldsymbol{X}$ but satisfy $\mathbb{E}[\boldsymbol{W}\varepsilon] = 0$ and $\mathbb{E}[\varepsilon^2|\boldsymbol{W}] = \sigma^2$. Let $\mathbf{X}, \mathbf{W} \in \mathbb{R}^{n\times d}$ be the design matrix and instrumental variable matrix stacking the i.i.d. covariates $\{\boldsymbol{X}_i\}_{i=1}^n$ and instrumental variables $\{\boldsymbol{W}_i\}_{i=1}^n$, respectively, and $\boldsymbol{Y} \in \mathbb{R}^n$ be the response vector with independent entries Y_i. Using the instrumental variables, one can construct the estimating function $\mathbf{t}((\boldsymbol{Y}, \mathbf{X}, \mathbf{W}), \boldsymbol{\beta}) = n^{-1}\mathbf{W}^T(\mathbf{X}\boldsymbol{\beta} - \boldsymbol{Y})$ with $E_{\mathbf{t}}(\boldsymbol{\beta}) = \mathbb{E}\mathbf{t}((\boldsymbol{Y}, \mathbf{X}, \mathbf{W}), \boldsymbol{\beta})$. In addition, $E_{\mathbf{t}}(\boldsymbol{\beta})$ has $\boldsymbol{\beta}^*$ as its unique root, provided that the second moment matrix $\boldsymbol{\Sigma}_{\boldsymbol{WX}} := n^{-1}\mathbb{E}\mathbf{W}^T\mathbf{X}$ is of full rank. Inspired by Gautier and Tsybakov (2011), we consider the following estimator $\widehat{\boldsymbol{\beta}}$:

$$\begin{aligned} \widehat{\boldsymbol{\beta}} &= \operatorname{argmin} \|\boldsymbol{\beta}\|_1 \\ &\quad \text{such that } \|n^{-1}\mathbf{W}^T(\mathbf{X}\boldsymbol{\beta} - \boldsymbol{Y})\|_\infty \le \lambda. \end{aligned}$$

Graphical Models via CLIME/SKEPTIC (Cai, Liu and Luo, 2011, Liu, Han and Zhang, 2012). Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be i.i.d. copies of $\boldsymbol{X} \in \mathbb{R}^d$ with $\mathbb{E}(\boldsymbol{X}) = 0$ and $\operatorname{Cov}(\boldsymbol{X}) = \boldsymbol{\Sigma}_{\boldsymbol{X}}$. It is well known that in the case when $\boldsymbol{X}$ are Gaussian, the precision matrix $\boldsymbol{\Omega}^* = (\boldsymbol{\Sigma}_{\boldsymbol{X}})^{-1}$ induces a graph, encoding conditional independencies of the variables $\boldsymbol{X}$. More generally, this observation can be extended to transelliptical distributions (Liu, Han and Zhang, 2012).

Let $\boldsymbol{\Sigma}_n = n^{-1}\sum_{i=1} \boldsymbol{X}_i\boldsymbol{X}_i^T$ be the sample covariance of $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ [recall $\mathbb{E}(\boldsymbol{X}_i) = 0$]. Based on the second moment condition $\boldsymbol{\Sigma}_{\boldsymbol{X}}\boldsymbol{\Omega}^* = \mathbf{I}_d$, Cai, Liu and Luo (2011) proposed the CLIME estimator of $\boldsymbol{\Omega}^*$:

$$\begin{aligned} \widehat{\boldsymbol{\Omega}} &= \operatorname{argmin} \|\boldsymbol{\Omega}\|_1 \\ &\quad \text{subject to } \|\boldsymbol{\Sigma}_n\boldsymbol{\Omega} - \mathbf{I}_d\|_{\max} \le \lambda. \end{aligned} \tag{1.2}$$

In this case, we have $\mathbf{t}(\mathbf{X}, \boldsymbol{\Omega}) = \boldsymbol{\Sigma}_n\boldsymbol{\Omega} - \mathbf{I}_d$, and $E_{\mathbf{t}}(\boldsymbol{\Omega}) = \boldsymbol{\Sigma}_{\boldsymbol{X}}\boldsymbol{\Omega} - \mathbf{I}_d$. Under the more general setting of transelliptical graphical models, Liu, Han and Zhang (2012) substituted the sample covariance $\boldsymbol{\Sigma}_n$ with a nonparametric estimate based on Kendall's tau (see Remark 2). Doing so breaks down the i.i.d. decomposition of the estimating equation described above, but continues to belong to our formulation (1.1).

Discriminant Analysis (Cai and Liu, 2011). Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be d-dimensional random vectors, coming from two populations with different means $\boldsymbol{\mu}_1 = \mathbb{E}(\boldsymbol{X})$, $\boldsymbol{\mu}_2 = \mathbb{E}(\boldsymbol{Y})$, and a common covariance matrix $\boldsymbol{\Sigma} = \operatorname{Cov}(\boldsymbol{X}) = \operatorname{Cov}(\boldsymbol{Y})$. Given some training samples, we are interested in classifying a new observation $\boldsymbol{O}$ into population 1 or population 2. It is well known (e.g., see Mardia, Kent and Bibby, 1979, Theorem 11.2.1) that, under certain conditions, the Bayes' classification rule takes the form

$$\psi(\boldsymbol{O}) = I\big((\boldsymbol{O} - \boldsymbol{\mu})^T \boldsymbol{\Omega}\boldsymbol{\delta} > 0\big),$$

where $I(\cdot)$ is an indicator function, $\boldsymbol{\mu} = (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)/2$, $\boldsymbol{\delta} = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ and $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$. Specifically, the observation $\boldsymbol{O}$ is classified into population 1 if and only if $\psi(\boldsymbol{O}) = 1$.

To implement $\psi(\boldsymbol{O})$ in practice, one has to estimate the unknown parameters $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ and $\boldsymbol{\Omega}$. Assume we observe n_1 and n_2 training samples from population 1 and population 2 denoted by $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{n_1} \in \mathbb{R}^d$ and $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_{n_2} \in \mathbb{R}^d$. We assume that

$$\begin{aligned} &\boldsymbol{X}_i = \boldsymbol{\mu}_1 + \boldsymbol{U}_i, \quad i = 1, \ldots, n_1 \quad \text{and} \\ &\boldsymbol{Y}_i = \boldsymbol{\mu}_2 + \boldsymbol{U}_{i+n_1}, \quad i = 1, \ldots, n_2, \end{aligned} \tag{1.3}$$

where $\boldsymbol{U}_i$ are i.i.d. copies of $\boldsymbol{U} = (U_1, \ldots, U_d)^T$, which satisfies $\mathbb{E}(\boldsymbol{U}) = 0$ and $\operatorname{Cov}(\boldsymbol{U}) = \boldsymbol{\Sigma}$. Define the sample means as $\bar{\boldsymbol{X}} = \frac{1}{n_1}\sum_{i=1}^{n_1} \boldsymbol{X}_i$ and $\bar{\boldsymbol{Y}} = \frac{1}{n_2}\sum_{i=1}^{n_2} \boldsymbol{Y}_i$, and the sample covariances as $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{X}} = \frac{1}{n_1}\sum_{i=1}^{n_1}(\boldsymbol{X}_i - \bar{\boldsymbol{X}})(\boldsymbol{X}_i - \bar{\boldsymbol{X}})^T$ and $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{Y}} = \frac{1}{n_2}\sum_{i=1}^{n_2}(\boldsymbol{Y}_i - \bar{\boldsymbol{Y}})(\boldsymbol{Y}_i - \bar{\boldsymbol{Y}})^T$. Furthermore, let $\widehat{\boldsymbol{\Sigma}}_n = \frac{n_1}{n}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{X}} + \frac{n_2}{n}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{Y}}$ be the weighted average of $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{X}}$ and $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{Y}}$.

In the high-dimensional setting with $d \gg n$, we cannot directly estimate $\boldsymbol{\Omega}$ by $\widehat{\boldsymbol{\Sigma}}_n^{-1}$, since the sample covariance is not invertible. Noting that the classification rule solely depends on $\boldsymbol{\beta}^* = \boldsymbol{\Omega}\boldsymbol{\delta}$, Cai and Liu (2011) proposed a direct approach to estimate $\boldsymbol{\beta}^*$, rather than estimating $\boldsymbol{\Omega}$ and $\boldsymbol{\delta}$ separately. Their estimated classification rule is as follows:

$$\begin{aligned} &\widehat{\psi}(\boldsymbol{O}) = I\big((\boldsymbol{O} - (\bar{\boldsymbol{X}} + \bar{\boldsymbol{Y}})/2)^T \widehat{\boldsymbol{\beta}} > 0\big) \quad \text{where} \\ &\widehat{\boldsymbol{\beta}} = \operatorname{argmin} \|\boldsymbol{\beta}\|_1 \\ &\qquad \text{subject to } \big\|\widehat{\boldsymbol{\Sigma}}_n\boldsymbol{\beta} - (\bar{\boldsymbol{X}} - \bar{\boldsymbol{Y}})\big\|_\infty \le \lambda. \end{aligned} \tag{1.4}$$

Clearly, the latter formulation constitutes a high-dimensional estimating equation as in (1.1), with $\mathbf{t}((\{\boldsymbol{X}_i\}_{i=1}^{n_1}, \{\boldsymbol{Y}_i\}_{i=1}^{n_2}), \boldsymbol{\beta}) = \widehat{\boldsymbol{\Sigma}}_n\boldsymbol{\beta} - (\bar{\boldsymbol{X}} - \bar{\boldsymbol{Y}})$ and $E_{\mathbf{t}}(\boldsymbol{\beta}) = \boldsymbol{\Sigma}\boldsymbol{\beta} - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$.

Vector Autoregressive Models (Han, Lu and Liu, 2015). Let $\{\boldsymbol{X}_t\}_{t=-\infty}^{\infty}$ be a stationary sequence of mean 0 random vectors in $\mathbb{R}^d$ with covariance matrix $\boldsymbol{\Sigma}$. The sequence $\{\boldsymbol{X}_t\}_{t=-\infty}^{\infty}$ is said to follow a lag-1 autoregressive model if

$$\boldsymbol{X}_t = \mathbf{A}^T \boldsymbol{X}_{t-1} + \boldsymbol{W}_t, \quad t \in \mathbb{Z} := \{\ldots, -1, 0, 1, \ldots\},$$

where $\mathbf{A}$ is a $d \times d$ transition matrix, and the noise vectors $\boldsymbol{W}_t$ are i.i.d. with $\boldsymbol{W}_t \sim N(0, \boldsymbol{\Psi})$ and independent of the history $\{\boldsymbol{X}_s\}_{s<t}$. Under the additional assumption that $\det(\mathbf{I}_d - \mathbf{A}^T z) \neq 0$ for all $z \in \mathbb{C}$ with $|z| \le 1$, it can be shown that $\boldsymbol{\Psi}$ can be selected so that the process is stationary, i.e. for all t: $\boldsymbol{X}_t \sim N(0, \boldsymbol{\Sigma})$. Let $\boldsymbol{\Sigma}_i := \operatorname{Cov}(\boldsymbol{X}_0, \boldsymbol{X}_i)$, where $\boldsymbol{\Sigma}_0 := \boldsymbol{\Sigma}$. A simple calculation under the lag-1 autoregressive model leads to the following Yule–Walker equation: $\boldsymbol{\Sigma}_i := \boldsymbol{\Sigma}_0 \mathbf{A}^i$, for any $i \in \mathbb{N}$. A special case of the above equation with $i = 1$ yields that

$$\mathbf{A} = \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\Sigma}_1. \tag{1.5}$$

Assume that the data $(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_T)$ follow the lag-1 autoregressive model. By equation (1.5), Han, Lu and Liu (2015) proposed the following estimator of $\mathbf{A}$ in the high-dimensional setting

$$\begin{aligned} \widehat{\mathbf{A}} = &\operatorname*{argmin}_{\mathbf{M} \in \mathbb{R}^{d\times d}} \sum_{1 \le j,k \le d} |M_{jk}| \\ &\text{subject to } \|\boldsymbol{S}_0\mathbf{M} - \boldsymbol{S}_1\|_{\max} \le \lambda, \end{aligned} \tag{1.6}$$

where $\lambda > 0$ is a tuning parameter, $\boldsymbol{S}_0 = T^{-1}\sum_{t=1}^{T} \boldsymbol{X}_t \boldsymbol{X}_t^T$ and $\boldsymbol{S}_1 = (T-1)^{-1}\sum_{t=1}^{T-1} \boldsymbol{X}_t \boldsymbol{X}_{t+1}^T$ are estimators of $\boldsymbol{\Sigma}_0$ and $\boldsymbol{\Sigma}_1$, respectively, and T is the number of observations. In this case, we have that $\mathbf{t}(\{\boldsymbol{X}_t\}_{t=1}^{T}, \mathbf{M}) = \boldsymbol{S}_0\mathbf{M} - \boldsymbol{S}_1$, and $E_{\mathbf{t}}(\mathbf{M}) = \boldsymbol{\Sigma}_0\mathbf{M} - \boldsymbol{\Sigma}_1$.

1.2 Related Methods

Having explored a few examples falling into the estimating equation framework (1.1), we move on to outline some related works on high-dimensional inference. Recently, significant progress has been made toward understanding the post-regularization inference for the LASSO estimator in the linear and generalized linear models. For instance, Lockhart et al. (2014), Taylor et al. (2014), Lee et al. (2013), Tian and Taylor (2018) suggested conditional tests based on covariates which have been selected by the LASSO. We stress the fact that this type of tests are of fundamentally different nature compared to our work.

Another important class of methods is based on the bias correction of L_1 or nonconvex regularized estimators. In particular, Zhang and Zhang (2014) proposed the low dimensional projection estimator (LDPE) for the inference in linear models. The method is further

extended by Belloni, Chernozhukov and Wei (2013), van de Geer et al. (2014) to the generalized linear models. Recently, Ning and Liu (2014, 2017) proposed a decorrelated score test in a likelihood based framework. The difference between our method and this class of methods will be discussed in more detail in the next section. It is also worth mentioning two recent papers focusing on linear models; Zhu and Bradic (2016), Cai and Guo (2017). These papers set out to understand how to perform a more general testing of projections on a potentially dense loading vector in the linear model. In contrast, our work considers the inference on the component of $\boldsymbol{\beta}$, which is a special case of the aforementioned papers, but handles the more general setting of estimating equations.

A different score related approach is considered by Voorman, Shojaie and Witten (2014), which is testing a null hypothesis depending on the tuning parameter, and hence differs from our work. For the nonconvex penalty, under the oracle properties, the asymptotic normality property of the estimators is established by Fan and Lv (2011), which requires strong conditions, such as the minimal signal condition. In contrast, our work does not rely on oracle properties or variable selection consistency. P-values and confidence intervals based on sample splitting and subsampling are suggested by Meinshausen, Meier and Bühlmann (2009), Meinshausen and Bühlmann (2010), Shah and Samworth (2013), Wasserman and Roeder (2009). However, the sample splitting procedures may lead to certain efficiency loss. In a recent paper by Lu et al. (2015), the authors developed a new inferential method based on a variational inequality technique for the LASSO procedure which provably produces valid confidence regions. In contrast to our work, their method needs the dimension d to be fixed, and it may not be applicable to the inference problem based on the formulation (1.1).

In addition to the above works, three relevant papers on Z-estimation are Loh (2017), Belloni, Chernozhukov and Kato (2015), Belloni, Chernozhukov and Hansen (2014). The first work considered the M-estimators and influence function in robust regression. The latter considers Z-estimators, establishes validity of a bootstrap procedure to construct simultaneous confidence intervals for an increasing number of parameters, and studies in detail the LAD case. Their approach is based on the "orthogonal moment condition," which essentially achieves the debiasing feature needed to obtain confidence regions despite of the high dimensionality of the nuisance parameters.

1.3 Contributions

Our first contribution is to propose a new procedure for high-dimensional inference in the estimating equation framework. In order to construct confidence regions, our method projects the general estimating equation onto a certain sparse direction, which can be easily estimated by solving a large-scale linear program. Thus, the proposed inferential procedure is a general methodology and can be directly applied to many inference problems, including all aforementioned examples. We note that such a projection idea is first proposed by Zhang and Zhang (2014). Our method is different in that it directly targets the influence function of the estimating equation. Below we highlight the differences between our method and Zhang and Zhang (2014), Ning and Liu (2014, 2017).

In the linear model setting, Zhang and Zhang (2014) search for a projection direction which coincides with the least squares score equation for the parameter of interest θ, that is, they aim to estimate a vector $\mathbf{w}$ satisfying $\mathbf{w}^T(\boldsymbol{Y} - \theta\mathbf{X}_{*1}) = 0$. Specifically, they estimate $\mathbf{w}$ by approximately solving $\mathbf{w}^T\mathbf{X}_{*,-1} \approx 0$. In the present paper, we propose a different estimate of $\mathbf{w}$, which satisfies the same condition. More importantly, we extend this idea to general estimating equation settings and provide a very natural and compelling motivation based on influence function expansions. Ning and Liu (2014, 2017) define the decorrelated score function $n^{-1}\sum_{i=1}^{n}[\partial\ell(\boldsymbol{Z}_i, \boldsymbol{\beta})/\partial\theta - \mathbf{w}^T\partial\ell(\boldsymbol{Z}_i, \boldsymbol{\beta})/\partial\boldsymbol{\gamma}]$, where $\ell(\boldsymbol{Z}_i, \boldsymbol{\beta})$ is the log-likelihood for data $\boldsymbol{Z}_i$, and $\mathbf{w}^T\partial\ell(\boldsymbol{Z}_i, \boldsymbol{\beta})/\partial\boldsymbol{\gamma}$ is the sparse projection of the θ-score function $\partial\ell(\boldsymbol{Z}_i, \boldsymbol{\beta})/\partial\theta$ to the $(d-1)$-dimensional nuisance score space $\operatorname{span}\{\partial\ell(\boldsymbol{Z}_i, \boldsymbol{\beta})/\partial\boldsymbol{\gamma}\}$. While the score function can be treated as a special case of estimating equation, such a construction cannot be directly extended to general estimating equations. The reason is that it is unclear how to disentangle the estimating equation for the parameter of interest and the space of nuisance estimating equations and, therefore, the projection method in Ning and Liu (2014, 2017) is not applicable. To address this challenge, motivated from the classical influence function representation, we propose a different projection approach, which directly estimates the influence function of the equation.

Our second contribution is to establish a unified Z-estimation theory of confidence intervals. In particular, we construct a Z-estimator $\widetilde{\theta}$ that is consistent and asymptotically normal, and its asymptotic variance can be consistently estimated. Furthermore, the pointwise asymptotic normality results can be strengthened by

showing that $\widetilde{\theta}$ is uniformly asymptotically normal for $\boldsymbol{\beta}^*$ belonging to a certain parameter space (deferred to the Supplementary Material, Neykov et al., 2018). Moreover, owing to the flexibility of the estimating equations framework, we are able to push the theory through for non-i.i.d. data, relaxing the assumptions made in most existing work. In terms of relative efficiency, when the estimating equation corresponds to the score function, our estimator $\widetilde{\theta}$ is semiparametrically efficient. The theoretical properties of hypothesis tests have also been established, but for space limitations the proofs will be omitted and can be provided by the authors upon request.

Our third contribution is to apply the proposed framework to establish theoretical results for the previous motivating examples including the noisy compressed sensing with moment condition, instrumental variable regression, graphical models, transelliptical graphical models, linear discriminant analysis and vector autoregressive models. To the best of our knowledge, many of the aforementioned problems (e.g., instrumental variables regression, linear discriminant analysis and vector autoregressive models) have not been equipped with any inferential procedures.

Finally, we further emphasize the difference between our method and the class of methods based on the bias correction of regularized estimators. Compared to these methods in Zhang and Zhang (2014), Javanmard and Montanari (2014), van de Geer et al. (2014), Ning and Liu (2017), our framework differs in the following three aspects. First, all of the above propositions start from a likelihood, or more generally a loss function. In contrast, our framework directly handles the estimating equations and is likelihood-free, enabling us to perform inference in many examples (e.g., the motivating examples discussed in Section 1.1) where the likelihood or the loss function is unavailable or difficult to formulate. For instance, in the instrumental variable regression it is not clear how to devise a loss function, while the problem naturally falls into the realm of estimating equations. This leads to different methodological development from the previous work, which will be explained later in detail. Second, some of the existing work is only tailored for the linear and generalized linear models. In contrast, our framework covers a much broader class of statistical models specified by estimating equations, such as linear discriminant analysis and vector autoregressive models whose inferential properties have not been studied before. Third, the estimating equation framework gives us more flexibility to handle dependent data, whereas the existing work requires the data to be independent.

1.4 Organization of the Paper

The paper is organized as follows. In Section 2, we propose our generic inferential procedure for high-dimensional estimating equations, and layout the foundations of the general theoretical framework. In Section 4, we apply the general theory to study the motivating examples including the Dantzig selector, instrumental variables regression, graphical models, discriminant analysis and autoregressive models. Numerical studies and a real data analysis are presented in Section 5, and a discussion is provided in Section 6.

1.5 Notation

The following notation is used throughout the paper. For a vector $\mathbf{v} = (v_1, \ldots, v_d)^T \in \mathbb{R}^d$, let $\|\mathbf{v}\|_q = (\sum_{i=1}^d v_i^q)^{1/q}$, $1 \le q < \infty$, $\|\mathbf{v}\|_0 = |\operatorname{supp}(\mathbf{v})|$, where $\operatorname{supp}(\mathbf{v}) = \{j : v_j \neq 0\}$, and $|A|$ denotes the cardinality of a set A. Furthermore, let $\|\mathbf{v}\|_\infty = \max_i |v_i|$. For a matrix $\mathbf{M}$, denote with $\mathbf{M}_{*j}$ and $\mathbf{M}_{j*}$ the jth column and row of $\mathbf{M}$ correspondingly. Moreover, let $\|\mathbf{M}\|_{\max} = \max_{ij} |M_{ij}|$, $\|\mathbf{M}\|_p = \max_{\|\mathbf{v}\|_p=1} \|\mathbf{M}\mathbf{v}\|_p$ for $p \ge 1$. If $\mathbf{M}$ is positive semidefinite let $\lambda_{\max}(\mathbf{M})$ and $\lambda_{\min}(\mathbf{M})$ denote the largest and smallest eigenvalues correspondingly. For a set $S \subset \{1, \ldots, d\}$ let $\mathbf{v}_S = \{v_j : j \in S\}$ and S^c be the complement of S. We denote with ϕ, Φ, $\overline{\Phi}$ the p.d.f., c.d.f. and tail probability of a standard normal random variable correspondingly. Furthermore, we will use $\rightsquigarrow$ to denote weak convergence.

For a random variable X, we define its ψ_ℓ norm for any $\ell \ge 1$ as

$$\|X\|_{\psi_\ell} = \sup_{p \ge 1} p^{-1/\ell} \big(\mathbb{E}|X|^p\big)^{1/p}. \tag{1.7}$$

In the present paper, we mainly use the ψ_1 and ψ_2 norms. Random variables with bounded ψ_1 and ψ_2 norms are called *subexponential* and *sub-Gaussian* correspondingly (Vershynin, 2012). It can be shown that a random variable is subexponential if there exists a constant $K_1 > 0$ such that $\mathbb{P}(|\boldsymbol{X}| > t) \le \exp(1 - t/K_1)$ for all $t \ge 0$. Similarly, a random variable is sub-Gaussian, if there exists a $K_2 > 0$ such that $\mathbb{P}(|\boldsymbol{X}| > t) \le \exp(1 - t^2/K_2^2)$ for all $t \ge 0$. Finally, for two sequences of positive numbers $\{a_n\}$ and $\{b_n\}$ we will write $a_n \asymp b_n$ if there exist positive constants $c, C > 0$ such that $\limsup_n a_n/b_n \le C$ and $\liminf_n a_n/b_n \ge c$.

2. HIGH-DIMENSIONAL ESTIMATING EQUATIONS

In this section, we present the intuition behind the construction of our projection, and formulate the main

results of our theory. Recall that $\boldsymbol{\beta} = (\theta, \boldsymbol{\gamma}) \in \mathbb{R}^d$, where θ is a univariate parameter of interest and $\boldsymbol{\gamma}$ is a $(d-1)$-dimensional nuisance parameter. We are interested in constructing a confidence interval for θ. In fact, our results can be extended in a simple manner to cases with $\boldsymbol{\theta}$ being a finite and fixed-dimensional vector, but we do not pursue this development in the present manuscript. Throughout the paper, we assume without loss of generality that θ is the first component of $\boldsymbol{\beta}$.

In the conventional framework, where the dimension d is fixed and less than the sample size n, one can estimate the d-dimensional parameter $\boldsymbol{\beta}$ by the Z-estimator, which is the root (assumed to exist) of the following system of d equations (Godambe, 1991):

$$\mathbf{t}(\mathbf{Z}, \boldsymbol{\beta}) = 0. \tag{2.1}$$

Under certain regularity conditions, the Z-estimator is consistent, and one has the following influence function expansion of the parameter $\widehat{\theta}$, where $\widehat{\boldsymbol{\beta}} = (\widehat{\theta}, \widehat{\boldsymbol{\gamma}})$ is the solution to (2.1) (Newey and McFadden, 1994, van der Vaart, 1998):

$$\sqrt{n}(\widehat{\theta} - \theta^*) = -\sqrt{n}[E_{\mathbf{T}}(\boldsymbol{\beta}^*)]_{1*}^{-1}\mathbf{t}(\mathbf{Z}, \boldsymbol{\beta}^*) + o_p(1). \tag{2.2}$$

In the preceding display, we assume $E_{\mathbf{T}}(\boldsymbol{\beta}) := \lim_{n\to\infty} \mathbb{E}\mathbf{T}(\mathbf{Z}, \boldsymbol{\beta})$[2] is invertible, where $\mathbf{T}(\mathbf{Z}, \boldsymbol{\beta}) := \frac{\partial}{\partial \boldsymbol{\beta}}\mathbf{t}(\mathbf{Z}, \boldsymbol{\beta})$. It is noteworthy to observe that in contrast to the Hessian matrix of the log-likelihood (or more generally any smooth loss function), the Jacobian matrix $\mathbf{T}(\mathbf{Z}, \boldsymbol{\beta})$ need not be symmetric in general (refer to the IVR model for an example). Under further conditions, the right-hand side of (2.2) converges to a normal distribution, hence guaranteeing the asymptotic normality of the estimator $\widehat{\theta}$.

In the case when $d > n$, the estimating equation (2.1) is ill-posed as one has more parameters than samples, resulting in multiple solutions for $\boldsymbol{\beta}$. To deal with such situations, under the sparsity assumption on $\boldsymbol{\beta}^*$, we solve the constrained optimization program (1.1):

$$\widehat{\boldsymbol{\beta}} = \operatorname{argmin} \|\boldsymbol{\beta}\|_1 \quad \text{subject to } \|\mathbf{t}(\mathbf{Z}, \boldsymbol{\beta})\|_\infty \le \lambda,$$

which is the first stage of our algorithm. Due to the constraint in (1.1), the limiting distribution of the estimator $\widehat{\boldsymbol{\beta}}$, and $\widehat{\theta}$ in particular, becomes intractable as expansion (2.2) is no longer valid. Hence, instead of focusing on the left-hand side of (2.2), in order to construct a theoretically tractable estimator of θ we consider a direct approach by estimating the influence function on the right-hand side. Emulating expression (2.2), we propose the following projected estimating function along the direction $\widehat{\mathbf{v}}$:

$$\widehat{S}(\boldsymbol{\beta}) = \widehat{\mathbf{v}}^T\mathbf{t}(\mathbf{Z}, \boldsymbol{\beta}),$$

where $\widehat{\mathbf{v}}$ is defined as the solution to the optimization problem

$$\widehat{\mathbf{v}} = \operatorname{argmin} \|\mathbf{v}\|_1 \quad \text{such that } \|\mathbf{v}^T\mathbf{T}(\mathbf{Z}, \widehat{\boldsymbol{\beta}}) - \mathbf{e}_1\|_\infty \le \lambda'. \tag{2.3}$$

In (2.3), λ' is an additional tuning parameter, and $\mathbf{e}_1$ is a d-dimensional row vector $(1, 0, \ldots, 0)$, where the position of 1 corresponds to that of θ among $\boldsymbol{\beta}$. It is easily seen that $\widehat{\mathbf{v}}^T$ is a natural estimator of $\mathbf{v}^{*T} := [E_{\mathbf{T}}(\boldsymbol{\beta}^*)]_{1*}^{-1}$ in the high-dimensional setting, which is an essential term in the right-hand side of (2.2). Thus, $\widehat{S}(\boldsymbol{\beta})$ can be viewed as an estimate of the influence function for estimating θ in high dimensions. To better understand our method, consider the linear model example. In this case, we have

$$\widehat{S}(\boldsymbol{\beta}) = n^{-1}\widehat{\mathbf{v}}^T\mathbf{X}^T(\mathbf{X}\boldsymbol{\beta} - \boldsymbol{Y}),$$

where

$$\widehat{\mathbf{v}} = \operatorname{argmin} \|\mathbf{v}\|_1 \quad \text{such that } \|\mathbf{v}^T\boldsymbol{\Sigma}_n - \mathbf{e}_1\|_\infty \le \lambda'.$$

We can see that $\widehat{\mathbf{v}}$ corresponds to the first column of the CLIME estimator for the inverse covariance matrix of $\boldsymbol{X}_i$.

We emphasize that the construction of $\widehat{\mathbf{v}}$ does not depend on knowing which is the estimating equation for θ and which is the nuisance estimating equation space, and thus the projection is different from the decorrelated score method in Ning and Liu (2017). In fact, the lack of a valid loss (or likelihood) function corresponding to the general estimating equations is the main difficulty for applying the existing likelihood based inference methods.

Recall that $\widehat{\boldsymbol{\beta}} = (\widehat{\theta}, \widehat{\boldsymbol{\gamma}})$. By plugging in the estimator $\widehat{\boldsymbol{\gamma}}$, we obtain the projected estimating equation $\widehat{S}(\theta, \widehat{\boldsymbol{\gamma}})$ for the parameter of interest θ. Similar to the classical estimating equation approach, we propose to estimate θ by a Z-estimator $\widetilde{\theta}$, which is the root of $\widehat{S}(\theta, \widehat{\boldsymbol{\gamma}}) = 0$. In practice, we can solve $\widetilde{\theta}$ by the standard Newton–Raphson algorithm. When θ is multidimensional, the Newton–Raphson algorithm may require more computational cost. In the following section, we lay out the foundations of a unified theory guaranteeing that the estimator $\widetilde{\theta}$ is asymptotically normal.

We conclude this section by summarizing our two-step procedure in the Algorithm 1.

[2]Recall that such limits are taken with the current d fixed.

Algorithm 1 Test Statistic for High-Dimensional Estimating Equations

Input: Data $\{\mathbf{Z}_i\}_{i=1}^n$, Equation $\mathbf{t}$; Tuning parameters λ, λ',

1. Solve the optimization problem (1.1), to obtain an estimate $\widehat{\boldsymbol{\beta}}$:
$$\widehat{\boldsymbol{\beta}} = \operatorname{argmin} \|\boldsymbol{\beta}\|_1 \quad \text{subject to } \|\mathbf{t}(\mathbf{Z}, \boldsymbol{\beta})\|_\infty \le \lambda;$$
2. Calculate the projection direction $\widehat{\mathbf{v}}^T$ through the following optimization based on (2.3):
$$\widehat{\mathbf{v}} = \operatorname{argmin} \|\mathbf{v}\|_1 \quad \text{such that } \|\mathbf{v}^T \mathbf{T}(\mathbf{Z}, \widehat{\boldsymbol{\beta}}) - \mathbf{e}_1\|_\infty \le \lambda';$$
3. Output the sparse projected test function $\widehat{S}(\boldsymbol{\beta}) = \widehat{\mathbf{v}}^T \mathbf{t}(\mathbf{Z}, \boldsymbol{\beta})$. Solve
$$\widehat{S}(\theta, \widehat{\boldsymbol{\gamma}}) = 0$$
to obtain the corrected estimate $\widetilde{\theta}$. ($\widehat{\boldsymbol{\gamma}}$ is directly obtained from the first step estimate $\widehat{\boldsymbol{\beta}}$.)

REMARK 1. Before we move to lay out our framework, we remark that the tests we develop are for one parameter only. They can be easily generalized to the setting with fixed dimensional parameters. In cases when one is interested in performing multiple testing with an increasing number of parameters, then different strategies such as the multiplied bootstrap developed by Chernozhukov, Chetverikov and Kato (2014) can be applied.

3. A GENERAL THEORETICAL FRAMEWORK

In this section, we provide generic sufficient conditions which guarantee the existence and asymptotic normality of $\widetilde{\theta}$, which is the root of
$$\widehat{S}(\theta, \widehat{\boldsymbol{\gamma}}) = 0,$$
as defined in Algorithm 1. Here, $\widehat{\boldsymbol{\gamma}}$ is directly obtained from the $\widehat{\boldsymbol{\beta}}$ estimate of optimization (1.1). Due to space limitations, we only present results on the confidence intervals, and the results on uniformly valid confidence intervals are deferred to the Supplementary Material. The results and proofs on hypothesis testing can be obtained from the authors upon request.

We assume that $\mathbf{t}(\mathbf{Z}, \boldsymbol{\beta})$ is twice differentiable in $\boldsymbol{\beta}$. Recall that we further require $\boldsymbol{\beta}^*$ to be the unique solution to $E_{\mathbf{t}}(\boldsymbol{\beta}) = 0$, where $E_{\mathbf{t}}(\boldsymbol{\beta}) = \lim_{n\to\infty} \mathbb{E}\mathbf{t}(\mathbf{Z}, \boldsymbol{\beta})$ is the limiting value of $\mathbb{E}\mathbf{t}(\mathbf{Z}, \boldsymbol{\beta})$ as we hold d fixed to its present value. For any $\boldsymbol{\beta}$, we let $S(\boldsymbol{\beta}) := \mathbf{v}^{*T}\mathbf{t}(\mathbf{Z}, \boldsymbol{\beta})$, where $\mathbf{v}^{*T} := [E_{\mathbf{T}}(\boldsymbol{\beta}^*)]_{1*}^{-1}$. Let $\mathbb{P}_{\boldsymbol{\beta}}$ be the probability measure under the parameter $\boldsymbol{\beta}$. We use the shorthand notation $\mathbb{P}^* = \mathbb{P}_{\boldsymbol{\beta}^*}$, to indicate the measure under the true parameter $\boldsymbol{\beta}^*$. For any vector $\boldsymbol{\beta} = (\theta, \boldsymbol{\gamma})$, we use the following shorthand notation $\boldsymbol{\beta}_{\check{\theta}} = (\check{\theta}, \boldsymbol{\gamma})$ to indicate that θ is replaced by $\check{\theta}$. Before we proceed to define our abstract assumptions and present the results, we first motivate them and give an informal description below.

3.1 Motivation and Informal Description

Throughout this section, we build our theory based on the premisses that the estimators $\widehat{\boldsymbol{\beta}}$ and $\widehat{\mathbf{v}}$ can be shown to be L_1 consistent, that is, $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 = o_p(1)$ and $\|\widehat{\mathbf{v}} - \mathbf{v}^*\|_1 = o_p(1)$. This is expected to hold for estimators solving programs (1.1) and (2.3) owing to the fact that both programs aim to minimize the L_1 norm of the parameters. The L_1 consistency [see (3.5)] is central in what follows. Under this presumption, the key idea in our theory is the successful control of the deviations of the "plug-in" equation $\widehat{S}(\theta, \widehat{\boldsymbol{\gamma}}) = \widehat{S}(\widehat{\boldsymbol{\beta}}_\theta)$ about the equation $S(\theta, \boldsymbol{\gamma}^*) = S(\boldsymbol{\beta}^*_\theta)$ [recall $S(\boldsymbol{\beta}^*_\theta) := \mathbf{v}^{*T}\mathbf{t}(\mathbf{Z}, \boldsymbol{\beta}^*_\theta)$], that is, we aim to establish $\widehat{S}(\widehat{\boldsymbol{\beta}}_\theta) = S(\boldsymbol{\beta}^*_\theta) + o_p(1)$. By the mean value theorem,

$$\begin{aligned}\widehat{S}(\widehat{\boldsymbol{\beta}}_\theta) &= S(\boldsymbol{\beta}^*_\theta) + \widehat{\mathbf{v}}^T \mathbf{T}(\mathbf{Z}, \widetilde{\boldsymbol{\beta}}_\nu)(\widehat{\boldsymbol{\beta}}_\theta - \boldsymbol{\beta}^*_\theta) \\ &\quad + (\widehat{\mathbf{v}} - \mathbf{v}^*)^T \mathbf{t}(\mathbf{Z}, \boldsymbol{\beta}^*_\theta),\end{aligned} \tag{3.1}$$

where $\widetilde{\boldsymbol{\beta}}_\nu$ is a point on the line segment joining $\widehat{\boldsymbol{\beta}}_\theta$ with $\boldsymbol{\beta}^*_\theta$. Owing to the L_1 consistency of $\widehat{\mathbf{v}}$ and $\widehat{\boldsymbol{\beta}}$, (3.1) can indeed be rewritten in the form $\widehat{S}(\widehat{\boldsymbol{\beta}}_\theta) = S(\boldsymbol{\beta}^*_\theta) + o_p(1)$, provided that $\|\widehat{\mathbf{v}}^T \mathbf{T}(\mathbf{Z}, \widetilde{\boldsymbol{\beta}}_\nu)\|_\infty = O_p(1)$ and $\|\mathbf{t}(\mathbf{Z}, \boldsymbol{\beta}^*_\theta)\|_\infty = O_p(1)$. A sufficient and also sensible condition for these bounds, is to desire $\|\mathbf{t}(\mathbf{Z}, \boldsymbol{\beta}^*_\theta) - E_{\mathbf{t}}(\boldsymbol{\beta}^*_\theta)\|_\infty = o_p(1)$, and $\|\widehat{\mathbf{v}}^T \mathbf{T}(\mathbf{Z}, \widetilde{\boldsymbol{\beta}}_\nu) - \mathbf{v}^{*T} \times E_{\mathbf{T}}(\boldsymbol{\beta}^*_\theta)\|_\infty = o_p(1)$, where $E_{\mathbf{t}}(\boldsymbol{\beta}^*_\theta)$ and $E_{\mathbf{T}}(\boldsymbol{\beta}^*_\theta)$ are the limiting expected values of $\mathbf{t}(\mathbf{Z}, \boldsymbol{\beta}^*_\theta)$ and $\widehat{\mathbf{v}}^T \mathbf{T}(\mathbf{Z}, \widetilde{\boldsymbol{\beta}}_\nu)$, respectively. It is therefore rational to believe that the latter L_∞-norms converge to 0; see Assumption 1. Furthermore, to show $\sqrt{n}$ consistency of the equations one needs to require an additional scaling condition on the latter convergence rates; see (3.8).

3.2 Main Results

We now formalize our intuition above by requiring the following assumption.

ASSUMPTION 1 (Concentration). There exists a neighborhood $\mathcal{N}_{\theta^*}$ of θ^*, such that, for all $\theta \in \mathcal{N}_{\theta^*}$,

$$\lim_{n\to\infty} \mathbb{P}^*\big(\|\mathbf{t}(\mathbf{Z}, \boldsymbol{\beta}^*_\theta) - E_{\mathbf{t}}(\boldsymbol{\beta}^*_\theta)\|_\infty \le r_1(n,\theta)\big) = 1, \tag{3.2}$$

$$\lim_{n\to\infty} \mathbb{P}^*\big(|\mathbf{v}^{*T}\mathbf{t}(\mathbf{Z}, \boldsymbol{\beta}^*_\theta) - \mathbf{v}^{*T}E_{\mathbf{t}}(\boldsymbol{\beta}^*_\theta)| \le r_2(n,\theta)\big) = 1, \tag{3.3}$$

$$\lim_{n\to\infty} \mathbb{P}^*\Big(\sup_{\nu\in[0,1]} \|\widehat{\mathbf{v}}^T\mathbf{T}(\mathbf{Z}, \widetilde{\boldsymbol{\beta}}_\nu) - \mathbf{v}^{*T}E_{\mathbf{T}}(\boldsymbol{\beta}^*_\theta)\|_\infty \le r_3(n,\theta)\Big) = 1, \tag{3.4}$$

where $\widetilde{\boldsymbol{\beta}}_\nu = \nu\widehat{\boldsymbol{\beta}}_\theta + (1-\nu)\boldsymbol{\beta}^*_\theta$, $\sup_{\theta\in\mathcal{N}_{\theta^*}} \max(r_1(n,\theta), r_2(n,\theta), r_3(n,\theta)) = o(1)$, and the following condition holds:

$$\sup_{\theta\in\mathcal{N}_{\theta^*}} \|E_{\mathbf{t}}(\boldsymbol{\beta}^*_\theta)\|_\infty < \infty,$$

$$\sup_{\theta\in\mathcal{N}_{\theta^*}} \|\mathbf{v}^{*T}[E_{\mathbf{T}}(\boldsymbol{\beta}^*_\theta)]_{-1}\|_\infty < \infty,$$

where $[\mathbf{A}]_{-1}$ represents a submatrix of $\mathbf{A}$ with the first column removed.

Condition (3.2) means that the equation $\mathbf{t}(\mathbf{Z}, \boldsymbol{\beta}^*_\theta)$ concentrates on its limiting value $E_{\mathbf{t}}(\boldsymbol{\beta}^*_\theta)$ for any θ in a small neighborhood of θ^*. Similarly, condition (3.3) implies that the projection of the estimating equation on $\mathbf{v}^*$ also concentrates on its limiting value locally around θ^*, and is automatically implied by (3.2) when $\|\mathbf{v}^*\|_1 = O(1)$. Finally, condition (3.4) means that the projection of the Jacobian matrix $\mathbf{T}(\mathbf{Z}, \widetilde{\boldsymbol{\beta}}_\nu)$ on $\widehat{\mathbf{v}}$ concentrates on its limiting value $\mathbf{v}^{*T}E_{\mathbf{T}}(\boldsymbol{\beta}^*_\theta)$ in a neighborhood of θ^*. These conditions are mild, and can be validated for all examples we consider. The two extra boundedness assumptions ensure that the limiting expected values of the estimating function and its derivative projected on the sparse direction $\mathbf{v}^*$ do not blow up in a neighborhood of θ^*, that is, the estimating function behaves nicely around the true solution.

ASSUMPTION 2 (L_1 Consistency). Let the estimators $\widehat{\boldsymbol{\beta}}$ and $\widehat{\mathbf{v}}$ satisfy

$$\begin{aligned}\lim_{n\to\infty} \mathbb{P}^*\big(\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \le r_4(n)\big) = 1,\\ \lim_{n\to\infty} \mathbb{P}^*\big(\|\widehat{\mathbf{v}} - \mathbf{v}^*\|_1 \le r_5(n)\big) = 1,\end{aligned} \tag{3.5}$$

where $\max(r_4(n), r_5(n)) = o(1)$.

As mentioned previously, (3.5) is expected to hold due to the formulations of (1.1) and (2.3). In particular, (3.5) has been verified for all examples we consider. Assumptions 1 and 2 suffice to show the following consistency result.

THEOREM 1 (Consistency). *Let the (stochastic) map $\theta \mapsto \widehat{S}(\widehat{\boldsymbol{\beta}}_\theta)$ be either continuous or nondecreasing, and has a single root $\widetilde{\theta}$. Furthermore, suppose that, for any $\varepsilon > 0$,*

$$\mathbf{v}^{*T}[E_{\mathbf{t}}(\boldsymbol{\beta}^*_{\theta^*-\varepsilon})]\mathbf{v}^{*T}[E_{\mathbf{t}}(\boldsymbol{\beta}^*_{\theta^*+\varepsilon})] < 0. \tag{3.6}$$

Under Assumptions 1 and 2, we have that

$$\lim_{n\to\infty} \mathbb{P}^*\big(|\widetilde{\theta} - \theta^*| > \varepsilon\big) = 0.$$

Condition (3.6) implies that the scalars $\mathbf{v}^{*T}[E_{\mathbf{t}}(\boldsymbol{\beta}^*_{\theta^*-\varepsilon})]$ and $\mathbf{v}^{*T}[E_{\mathbf{t}}(\boldsymbol{\beta}^*_{\theta^*+\varepsilon})]$ have opposite signs for all $\varepsilon > 0$, which in turn guarantees that θ^* is a unique root of the map $\mathbf{v}^{*T}[E_{\mathbf{t}}(\boldsymbol{\beta}^*_\theta)]$. Hence under (3.6) the population equation $\widehat{S}(\widehat{\boldsymbol{\beta}}_\theta)$ is unbiased. The condition (3.6) holds for numerous examples and is also commonly used in the classical asymptotic theory; see Section 5 of van der Vaart (1998). In fact, the conclusion of Theorem 1 remains valid if one solves the equation approximately in the sense that $\widehat{S}(\widehat{\boldsymbol{\beta}}_{\widetilde{\theta}}) = o_p(1)$. To establish the asymptotic normality of $\widetilde{\theta}$, we require the following assumptions.

ASSUMPTION 3 (CLT). Assume that for $\sigma^2 = \mathbf{v}^{*T}\boldsymbol{\Sigma}\mathbf{v}^*$, it holds that

$$\sigma^{-1}n^{1/2}S(\boldsymbol{\beta}^*) \rightsquigarrow N(0,1),$$

where $\boldsymbol{\Sigma} = \lim_{n\to\infty} n\operatorname{Cov}\mathbf{t}(\mathbf{Z}, \boldsymbol{\beta}^*)$, and assume that $\sigma^2 \ge C > 0$ for some constant C.

Assumption 3 ensures that the right-hand side of expansion (3.1) converges to a normal distribution when scaled appropriately. This CLT condition is mild and in many cases will hold true. For example, the CLT will hold whenever the equation $\mathbf{t}(\mathbf{Z}, \boldsymbol{\beta}^*)$ is an average of i.i.d. terms (modulo verifying Lyapunov or Lindeberg conditions). This is the case since the function $S(\boldsymbol{\beta}^*) = \mathbf{v}^{*T}\mathbf{t}(\mathbf{Z}, \boldsymbol{\beta}^*)$ will naturally decompose to average of i.i.d. terms in such a situation. For some types of dependent data, Assumption 3 holds by applying the martingale central limit theorem (e.g., autoregressive models). Thus, one of the advantages of our framework is that we can handle dependent data, which are not covered by the existing methods. We show such an example in Section 4.4.

ASSUMPTION 4 (Bounded Jacobian Derivative). Suppose there exists a constant $\gamma > 0$ such that

$$\left|\mathbf{v}^T\frac{\partial}{\partial\theta}[\mathbf{T}(\mathbf{Z}, (\theta, \boldsymbol{\gamma}))]_{*1}\right| \le \psi(\mathbf{Z}), \tag{3.7}$$

for any $\mathbf{v}$ and $\boldsymbol{\beta}$ satisfying $\|\mathbf{v}-\mathbf{v}^*\|_1<\gamma$ and $\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_1<\gamma$, where $\psi:\mathbb{R}^{n\times q}\mapsto\mathbb{R}$ is an integrable function with $\mathbb{E}^*\psi(\mathbf{Z})<\infty$.

Inequality (3.7) is a technical condition ensuring that $\mathbf{v}^T\frac{\partial}{\partial\theta}[\mathbf{T}(\mathbf{Z},\boldsymbol{\beta})]_{*1}$ is bounded by an integrable function in a small neighborhood, and hence does not behave too erratically, so that the dominated convergence theorem can be applied. This is a standard condition, which is also assumed in Theorem 5.41 of van der Vaart (1998) to establish the asymptotic normality of Z-estimator in the classical low dimensional regime. It is easily seen that this condition is mild and holds for linear estimating equations.

ASSUMPTION 5 (Scaling). Assume the convergence rates in Assumptions 1 and 2 satisfy

$$n^{1/2}\big(r_4(n)r_3(n,\theta^*)+r_5(n)r_1(n,\theta^*)\big)=o(1). \tag{3.8}$$

Assumption 5 is a technical condition, which says that the multiplication of the estimation errors of $\widehat{\boldsymbol{\gamma}}$ (or $\widehat{\mathbf{v}}$) by the error of the concentration inequalities (Assumption 1) is negligible in the bias of the final estimate $\widetilde{\theta}$. This assumption is crucial for the $n^{1/2}$-consistency of $\widetilde{\theta}$, and can be verified in all of our examples. We are now in a position to state the main result of this section.

THEOREM 2 (Asymptotic Normality). *Assume the conditions from Theorem 1 and Assumptions 3, 4 and 5 hold. If $\widehat{\sigma}^2$ is a consistent estimator of σ^2, then for any $t\in\mathbb{R}$, we have*

$$\lim_{n\to\infty}\big|\mathbb{P}^*(\widehat{U}_n\le t)-\Phi(t)\big|=0$$

$$\text{where } \widehat{U}_n=\frac{n^{1/2}}{\widehat{\sigma}}\big(\widetilde{\theta}-\theta^*\big).$$

Some generic sufficient conditions for the consistency of $\widehat{\sigma}$ are shown in Proposition B.1 in Section B.1 of the Supplementary Material. In our examples, we will develop consistent estimates of the variance σ^2 case by case. Given a consistent estimator $\widehat{\sigma}^2$, Theorem 2 implies that we can construct a $(1-\alpha)\%$ confidence interval of θ^* in the following way:

$$\lim_{n\to\infty}\mathbb{P}^*\big(\theta^*\in\big[\widetilde{\theta}-\Phi^{-1}(1-\alpha/2)\widehat{\sigma}/\sqrt{n},\ \widetilde{\theta}+\Phi^{-1}(1-\alpha/2)\widehat{\sigma}/\sqrt{n}\big]\big)=1-\alpha. \tag{3.9}$$

We now note a property of our estimator $\widetilde{\theta}$ in cases when the estimating equation comes from a log-likelihood, that is, $\mathbf{t}(\mathbf{Z},\boldsymbol{\beta})=n^{-1}\sum_{i=1}^n\mathbf{h}(\mathbf{Z}_i,\boldsymbol{\beta})$ with $\mathbf{h}(\mathbf{Z}_i,\boldsymbol{\beta})$ being the gradient of the log-likelihood for $\mathbf{Z}_i$. Denote $\mathbf{H}(\mathbf{Z},\boldsymbol{\beta})=\frac{\partial}{\partial\boldsymbol{\beta}}\mathbf{h}(\mathbf{Z},\boldsymbol{\beta})$. According to the information identity $-\mathbb{E}\mathbf{H}(\mathbf{Z},\boldsymbol{\beta}^*)=\operatorname{Cov}\mathbf{h}(\mathbf{Z},\boldsymbol{\beta}^*)$, we have $\mathbf{v}^{*T}\boldsymbol{\Sigma}\mathbf{v}^*=(\boldsymbol{\Sigma}^{-1})_{11}$. In this case, the Z-estimator $\widetilde{\theta}$ is efficient (van der Vaart, 1998), because the variance $(\boldsymbol{\Sigma}^{-1})_{11}$ coincides with the inverse of the information bound for θ.

4. IMPLICATIONS OF THE GENERAL THEORETICAL FRAMEWORK

In this section, we apply the general theory of Section 3 to the motivating examples we listed in the Introduction.

4.1 Linear Model and Instrumental Variables Regression

In this section, we consider the linear model via Dantzig selector and the instrumental variables regression. As seen in the Introduction, the instrumental variables regression can be viewed as a generalization of the linear regression, by substituting $\boldsymbol{W}\equiv\boldsymbol{X}$. For simplicity, we only present the results for the linear regression and defer the development of the inference theory for instrumental variables regression to Appendix C of Supplementary Material.

Recall that $\boldsymbol{\beta}:=(\theta,\boldsymbol{\gamma})$, and let $\boldsymbol{\Sigma}_n=n^{-1}\mathbf{X}^T\mathbf{X}$ be the empirical estimator of the second moment matrix $\boldsymbol{\Sigma}_{\boldsymbol{X}}$. Our goal is to construct confidence intervals for the parameter θ. In the linear regression case, we can easily show that $\widehat{S}(\boldsymbol{\beta})$ reduces to

$$\widehat{S}(\boldsymbol{\beta})=n^{-1}\widehat{\mathbf{v}}^T\mathbf{X}^T(\mathbf{X}\boldsymbol{\beta}-\boldsymbol{Y}),$$

where

$$\widehat{\mathbf{v}}=\operatorname{argmin}\|\mathbf{v}\|_1 \quad \text{subject to } \big\|\mathbf{v}^T\boldsymbol{\Sigma}_n-\mathbf{e}_1\big\|_\infty\le\lambda', \tag{4.1}$$

is an estimator of $\mathbf{v}^*=\boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1}\mathbf{e}_1^T$. We impose the following assumption.

ASSUMPTION 6. Assume that the error $\varepsilon:=Y-\boldsymbol{X}^T\boldsymbol{\beta}^*$ and the predictor $\boldsymbol{X}$ are both coordinate-wise sub-Gaussian, that is,

$$\|\varepsilon\|_{\psi_2}:=K<\infty,\qquad \sup_{j\in\{1,\dots,d\}}\|X_j\|_{\psi_2}:=K_{\boldsymbol{X}}<\infty,$$

for some fixed constants $K,K_{\boldsymbol{X}}>0$. Furthermore, assume that the variance $\operatorname{Var}(\varepsilon)\ge C_\varepsilon>0$, the random variables ε and $\boldsymbol{X}$ are independent, and the second moment matrix $\boldsymbol{\Sigma}_{\boldsymbol{X}}$ satisfies $\lambda_{\min}(\boldsymbol{\Sigma}_{\boldsymbol{X}})\ge\delta>0$, where δ is some fixed constant.

While assumption that the smallest eigenvalue of $\mathbf{\Sigma}_X$ is bounded away from 0 could be somewhat restrictive given that the dimension of $\mathbf{\Sigma}_X$ is allowed to increase, it ensures that the second moment matrix of the covariates is nondegenerate. To construct confidence intervals for θ, we consider $\widehat{U}_n = \widehat{\Delta}^{-1} n^{1/2}(\widetilde{\theta} - \theta^*)$, where $\widetilde{\theta}$ is defined as the solution to $\widehat{S}(\theta, \widehat{\boldsymbol{\gamma}}) = 0$, and

$$\widehat{\Delta} := \widehat{\mathbf{v}}^T \mathbf{\Sigma}_n \widehat{\mathbf{v}} n^{-1} \sum_{i=1}^{n} (Y_i - \boldsymbol{X}_i^T \widehat{\boldsymbol{\beta}})^2, \tag{4.2}$$

is an estimator of the asymptotic variance $\Delta := \mathbf{v}^{*T} \mathbf{\Sigma}_X \mathbf{v}^* \operatorname{Var}(\varepsilon)$. In high-dimensional models, it is often reasonable to assume that the vector $\boldsymbol{\beta}^*$ is sparse. Additionally, if we are in a setting where X_1 is expected to be conditionally uncorrelated with many entries of the vector $\boldsymbol{X}_{-1}$, it is also reasonable to postulate that $\mathbf{v}^*$ is sparse. Let s and $s_{\mathbf{v}}$ denote the sparsity of $\boldsymbol{\beta}^*$ and $\mathbf{v}^*$ correspondingly, that is, $\|\boldsymbol{\beta}^*\|_0 = s$ and $\|\mathbf{v}^*\|_0 = s_{\mathbf{v}}$. The next corollary of the general Theorem 2 shows the asymptotic normality of $\widehat{U}_n$ in linear models. To simplify the presentation of our result, we will assume that $\|\mathbf{v}^*\|_1$ is bounded, although this is not needed in our proofs.

COROLLARY 1. *Assume that Condition 6 holds, and*

$$\max(s_{\mathbf{v}}, s) \log d / \sqrt{n} = o(1), \quad \sqrt{\log d / n} = o(1).$$

Then with $\lambda \asymp \sqrt{\log d / n}$ *and* $\lambda' \asymp \sqrt{\log d / n}$, $\widehat{U}_n$ *satisfies, for any* $t \in \mathbb{R}$,

$$\lim_{n \to \infty} \left| \mathbb{P}^*(\widehat{U}_n \le t) - \Phi(t) \right| = 0.$$

The proof of Corollary 1 can be found in Appendix F of the Supplementary Material. The conditions in Corollary 1 agree with the existing conditions in Zhang and Zhang (2014), van de Geer et al. (2014). In fact, under the additional assumption $s_{\mathbf{v}}^3 / n = o(1)$, we can show that $\widehat{U}_n$ is uniformly asymptotically normal; see Remark F.1 of the Supplementary Material. Finally, we comment that a similar asymptotic normality result under the instrumental variables regression is shown in Corollary C.1 of the Supplementary Material.

4.2 Graphical Models

We begin with introducing several assumptions which we need throughout the development. First, let $\mathbf{\Sigma}_X$ satisfy $\lambda_{\min}(\mathbf{\Sigma}_X) \ge \delta > 0$, where δ is some fixed constant. Similar to Section 4.1, we assume that $\boldsymbol{X}$ is coordinatewise sub-Gaussian, that is,

$$K_X := \max_{j \in \{1, \ldots, d\}} \|X_j\|_{\psi_2} < \infty, \tag{4.3}$$

for some fixed constant $K_X > 0$. Our goal is to construct confidence intervals for a component of $\mathbf{\Omega}^*$, where $\mathbf{\Omega}^* = (\mathbf{\Sigma}_X)^{-1}$. Without loss of generality, we focus on the parameter Ω^*_{1m} for some $m \in \{1, \ldots, d\}$. When $\boldsymbol{X}$ are coming from a Gaussian distribution, the confidence intervals for Ω^*_{1m} provide uncertainty assessment on whether X_1 is independent of X_m given the rest of the variables.

There are a number of recent works considering the inferential problems for Gaussian graphical models (Janková and van de Geer, 2015, Chen et al., 2016, Ren et al., 2015, Liu, 2013) and Gaussian copula graphical models (Gu et al., 2015, Barber and Kolar, 2015). Our framework differs from these existing procedures in the following two aspects. First, our method is based on the estimating equations rather than the likelihood and (node-wise) pseudo-likelihood. Second, we only require each component of $\boldsymbol{X}$ is sub-Gaussian, whereas the majority of the existing methods require the data to be sampled from Gaussian or Gaussian copula distributions.

Let $\boldsymbol{\beta}^* := \mathbf{\Omega}^*_{*m}$, be the mth column of $\mathbf{\Omega}^*$. Then the CLIME estimator of $\boldsymbol{\beta}^*$ given by (1.2) reduces to

$$\widehat{\boldsymbol{\beta}} = \operatorname{argmin} \|\boldsymbol{\beta}\|_1 \quad \text{subject to } \left\| \mathbf{\Sigma}_n \boldsymbol{\beta} - \mathbf{e}_m^T \right\|_\infty \le \lambda,$$

where $\mathbf{e}_m^T$ is a unit column vector with 1 in the mth position and 0 otherwise. Phrasing this problem in the terminology of Section 3, we can construct d estimating equations: $\mathbf{t}(\mathbf{X}, \boldsymbol{\beta}) = \mathbf{\Sigma}_n \boldsymbol{\beta} - \mathbf{e}_m^T$. Let us decompose the vector $\boldsymbol{\beta}$ as $\boldsymbol{\beta} := (\theta, \boldsymbol{\gamma})$. Then the projected estimating equation for θ is given by

$$\widehat{S}(\boldsymbol{\beta}) = \widehat{\mathbf{v}}^T (\mathbf{\Sigma}_n \boldsymbol{\beta} - \mathbf{e}_m^T),$$

where

$$\begin{aligned} \widehat{\mathbf{v}} = & \operatorname{argmin} \|\mathbf{v}\|_1 \\ & \text{such that } \left\| \mathbf{v}^T \mathbf{\Sigma}_n - \mathbf{e}_1 \right\|_\infty \le \lambda'. \end{aligned} \tag{4.4}$$

Here, $\widehat{\mathbf{v}}$ is an estimate of $\mathbf{v}^* := (\mathbf{\Sigma}_X)^{-1}_{*1} = \mathbf{\Omega}^*_{*1}$. Notice that, due to the symmetry of $\widehat{\boldsymbol{\beta}}$ and $\widehat{\mathbf{v}}$, if we take $\lambda = \lambda'$, it suffices to simply solve the CLIME optimization (1.2) once in order to evaluate $\widehat{S}(\widehat{\boldsymbol{\beta}})$, as $\widehat{\boldsymbol{\beta}} = \widehat{\mathbf{\Omega}}_{*m}$ and $\widehat{\mathbf{v}} = \widehat{\mathbf{\Omega}}_{*1}$. This pleasant consequence for CLIME shows that in this special case the number of tuning parameters in the generic procedure described in Section 2 can be reduced to 1, and hence the computation is simplified.

The solution $\widetilde{\theta}$ to the equation $\widehat{S}(\theta, \widehat{\boldsymbol{\gamma}}) = 0$ has the following closed form expression:

$$\widetilde{\theta} = \widehat{\theta} - \frac{\widehat{\mathbf{v}}^T (\mathbf{\Sigma}_n \widehat{\boldsymbol{\beta}} - \mathbf{e}_m^T)}{\widehat{\mathbf{v}}^T \mathbf{\Sigma}_{n,*1}}. \tag{4.5}$$

To establish the asymptotic normality of $\widetilde{\theta}$, we impose the following assumption.

ASSUMPTION 7. There exists a constant $\alpha_{\min} > 0$ such that

$$\Delta \geq \alpha_{\min}\|\boldsymbol{\beta}^*\|_2^2\|\mathbf{v}^*\|_2^2 \quad \text{where } \Delta = \operatorname{Var}(\mathbf{v}^{*T}\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{\beta}^*).$$

We note that Assumption 7 is natural. For example, when $\boldsymbol{X} \sim N(0, \boldsymbol{\Sigma}_{\boldsymbol{X}})$, Isserlis' theorem yields that for any two vectors $\boldsymbol{\xi}$ and $\boldsymbol{\theta}$,

$$\begin{aligned}\operatorname{Var}(\boldsymbol{\xi}^T\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{\theta}) &= (\boldsymbol{\xi}^T\boldsymbol{\Sigma}_{\boldsymbol{X}}\boldsymbol{\xi})(\boldsymbol{\theta}^T\boldsymbol{\Sigma}_{\boldsymbol{X}}\boldsymbol{\theta}) + (\boldsymbol{\xi}^T\boldsymbol{\Sigma}_{\boldsymbol{X}}\boldsymbol{\theta})^2 \\ &\geq \lambda_{\min}^2(\boldsymbol{\Sigma}_{\boldsymbol{X}})\|\boldsymbol{\xi}\|_2^2\|\boldsymbol{\theta}\|_2^2,\end{aligned}$$

which clearly implies Assumption 7, if $\lambda_{\min}^2(\boldsymbol{\Sigma}_{\boldsymbol{X}})$ is lower bounded by a constant.

Denote $\|\boldsymbol{\beta}^*\|_0 = s$ and $\|\mathbf{v}^*\|_0 = s_{\mathbf{v}}$. To simplify the presentation of our result, we will assume that $\|\mathbf{v}^*\|_1$ and $\|\boldsymbol{\beta}^*\|_1$ are bounded quantities, although this is not needed in our proofs. The following corollary yields the asymptotic normality of $\widehat{U}_n = \widehat{\Delta}^{-1/2}n^{-1/2}(\widetilde{\theta} - \theta^*)$, where $\widehat{\Delta} := n^{-1}\sum_{i=1}^n(\widehat{\mathbf{v}}^T(\boldsymbol{X}_i\boldsymbol{X}_i^T - \boldsymbol{\Sigma}_n)\widehat{\boldsymbol{\beta}})^2$ is an estimator of Δ.

COROLLARY 2. *Let Assumption 7 and (4.3) hold. Furthermore, assume that*

$$\begin{aligned}&\max(s_{\mathbf{v}}^2, s^2)\log d\log(nd)/n = o(1), \\ &\exists k > 2 : (s_{\mathbf{v}}s)^k/n^{k-1} = o(1),\end{aligned} \tag{4.6}$$

and $\operatorname{Var}((\mathbf{v}^{*T}\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{\beta}^*)^2) = o(n)$, $\mathbb{E}(\mathbf{v}^{*T}\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{\beta}^*)^2 = O(1)$. *Let the tuning parameters be* $\lambda \asymp \sqrt{\log d/n}$ *and* $\lambda' \asymp \sqrt{\log d/n}$. *Then for all* $t \in \mathbb{R}$,

$$\lim_{n\to\infty}\left|\mathbb{P}^*(\widehat{U}_n \leq t) - \Phi(t)\right| = 0.$$

The proof of Corollary 2 can be found in Appendix H of the Supplementary Material. In addition, we provide a stronger result on uniform confidence intervals for θ in Corollary H.1 of the Supplementary Material. Once again, the first part of condition (4.6) agrees with Ren et al. (2015), Liu (2013). In addition, when the data are known to be Gaussian one could use the alternative estimator $\widetilde{\Delta} := \widehat{v}_1\widehat{\beta}_m + \widehat{v}_m\widehat{\beta}_1$ of Δ, which can also be shown to be consistent under the assumption $\max(s_{\mathbf{v}}, s)\sqrt{\log d/n} = o(1)$. The second part of condition (4.6) is mild, since it is only slightly stronger than $n^{-1}s_{\mathbf{v}}s = o(1)$. Unlike Janková and van de Geer (2015), we do not assume irrepresentable conditions.

REMARK 2 (Transelliptical Graphical Models). Our estimating equation based methods for constructing confidence intervals can be extended to transelliptical graphical models (Liu, Han and Zhang, 2012). The key idea is to replace the same covariance matrix $\boldsymbol{\Sigma}_n$ in (1.2) and (4.4) by

$$\widehat{S}_{jk}^{\tau} = \begin{cases}\sin\left(\dfrac{\pi}{2}\widehat{\tau}_{jk}\right), & j \neq k; \\ 1, & j = k,\end{cases}$$

where

$$\begin{aligned}\widehat{\tau}_{jk} = &\frac{2}{n(n-1)} \\ &\times \sum_{1\leq i<i'\leq n}\operatorname{sign}\big((X_{ij} - X_{i'j})(X_{ik} - X_{i'k})\big).\end{aligned}$$

Similar to Corollary 2, the asymptotic normality of the estimator $\widetilde{\theta}$ is established. The details are shown in Appendix D of the Supplementary Material.

4.3 Sparse Linear Discriminant Analysis

In this section, we consider an application of the general theory to the sparse linear discriminant analysis problem. The consistency and rates of convergence of the classification rule $\widehat{\psi}(\boldsymbol{O})$ (1.4) have been established by Cai and Liu (2011) in the high-dimensional setting. In the following, we apply the theory of Section 3 to construct confidence intervals for θ, where θ is the first component of $\boldsymbol{\beta}$, that is, $\boldsymbol{\beta} = (\theta, \boldsymbol{\gamma})$. Note that if $\theta = 0$, then it implies that the first feature of $\boldsymbol{O}$ is not needed in the Bayes' rule $\psi(\boldsymbol{O})$. Hence, our procedure can be used to assess whether a certain feature is significant in the classification.

By the identity $\boldsymbol{\beta}^* = \boldsymbol{\Omega}\boldsymbol{\delta}$, we can construct the d-dimensional estimating equations $\mathbf{t}((\mathbf{X}, \mathbf{Y}), \boldsymbol{\beta}) = \widehat{\boldsymbol{\Sigma}}_n\boldsymbol{\beta} - (\bar{\boldsymbol{X}} - \bar{\boldsymbol{Y}})$. Then the projected estimating equation for θ is given by

$$\widehat{S}(\boldsymbol{\beta}) = \widehat{\mathbf{v}}^T(\widehat{\boldsymbol{\Sigma}}_n\boldsymbol{\beta} - (\bar{\boldsymbol{X}} - \bar{\boldsymbol{Y}})),$$

where

$$\widehat{\mathbf{v}} = \operatorname{argmin}\|\mathbf{v}\|_1 \quad \text{such that } \left\|\mathbf{v}^T\widehat{\boldsymbol{\Sigma}}_n - \mathbf{e}_1\right\|_\infty \leq \lambda',$$

is an estimator of $\mathbf{v}^* = (\boldsymbol{\Sigma}^{-1})_{*1}$. Solving the equation $\widehat{S}(\theta, \widehat{\boldsymbol{\gamma}}) = 0$ gives us the Z-estimator $\widetilde{\theta}$. To establish the asymptotic normality of $\widetilde{\theta}$, we impose the following assumption.

ASSUMPTION 8. Assume that $\boldsymbol{U}$ satisfies the following moment assumption:

$$\operatorname{Var}(\mathbf{v}^{*T}\boldsymbol{U}\boldsymbol{U}^T\boldsymbol{\beta}^*) \geq V_{\min}\|\mathbf{v}^*\|_2^2\|\boldsymbol{\beta}^*\|_2^2,$$

where $V_{\min}$ is a positive constant. In addition, let $K_{\boldsymbol{U}} = \max_{j\in\{1,\ldots,d\}}\|U_j\|_{\psi_2} < \infty$.

As seen in the comments on Assumption 7, we can similarly show that Assumption 8 holds if $\boldsymbol{U} \sim N(0, \boldsymbol{\Sigma})$ and $\lambda^2_{\min}(\boldsymbol{\Sigma})$ is lower bounded by a positive constant. We define $V_1 := \mathrm{Var}(\mathbf{v}^{*T}\boldsymbol{U}\boldsymbol{U}^T\boldsymbol{\beta}^* + \alpha^{-1}\mathbf{v}^{*T}\boldsymbol{U})$, $V_2 := \mathrm{Var}(\mathbf{v}^{*T}\boldsymbol{U}\boldsymbol{U}^T\boldsymbol{\beta}^* - (1-\alpha)^{-1}\mathbf{v}^{*T}\boldsymbol{U})$, where $\frac{n_1}{n} = \alpha + o(1)$ for some $0 < \alpha < 1$. Denote

$$\Delta := \alpha V_1 + (1-\alpha)V_2, \tag{4.7}$$

and $\widehat{U}_n := \widehat{\Delta}^{-1/2} n^{1/2}(\widetilde{\theta} - \theta^*)$, where $\widehat{\Delta}$ is some consistent estimator of Δ. The explicit form of $\widehat{\Delta}$ is complicated, and we defer its expression to Appendix I of the Supplementary Material. Denote $\|\boldsymbol{\beta}^*\|_0 = s$ and $\|\mathbf{v}^*\|_0 = s_{\mathbf{v}}$. Once again for simplicity of the presentation we assume that $\|\mathbf{v}^*\|_1$ and $\|\boldsymbol{\beta}^*\|_1$ are bounded. We obtain the following asymptotic normality result.

COROLLARY 3. *Assume that* $\lambda_{\min}(\boldsymbol{\Sigma}) > \delta$ *for some constant* $\delta > 0$, *and let Assumption 8 hold. If*

$$\begin{aligned} &\max(s_{\mathbf{v}}, s)\log d/\sqrt{n} = o(1) \\ &\exists k > 2 : (s_{\mathbf{v}} s)^k / n^{k-1} = o(1), \end{aligned} \tag{4.8}$$

holds and $\lambda \asymp \sqrt{\log d/n}$ *and* $\lambda' \asymp \sqrt{\log d/n}$, *then for each* $t \in \mathbb{R}$:

$$\lim_{n\to\infty} \left|\mathbb{P}^*(\widehat{U}_n < t) - \Phi(t)\right| = 0.$$

The second part of (4.8) is similar to that in Corollary 2, which is used to establish the Lyapunov's condition for central limit theorem. The proof of Corollary 3 can be found in Appendix I of the Supplementary Material.

4.4 Stationary Vector Autoregressive Models

In this section, we develop inferential methods for the lag-1 vector autoregressive models considered in the Introduction. To this end, we remind the reader some of the notation; for the full notation, please refer to page 4. Let $\{\boldsymbol{X}_t\}_{t=-\infty}^{\infty}$ be a stationary sequence of mean 0 random vectors in $\mathbb{R}^d$ with covariance matrix $\boldsymbol{\Sigma}$ which is assumed to follow a lag-1 autoregressive model

$$\boldsymbol{X}_t = \mathbf{A}^T\boldsymbol{X}_{t-1} + \boldsymbol{W}_t, \quad t \in \mathbb{Z} := \{\ldots, -1, 0, 1, \ldots.\},$$

where $\mathbf{A}$ is a $d \times d$ transition matrix, and the noise vectors $\boldsymbol{W}_t$ are i.i.d. with $\boldsymbol{W}_t \sim N(0, \boldsymbol{\Psi})$ and independent of the history $\{\boldsymbol{X}_s\}_{s<t}$. Let $\boldsymbol{\beta}^* = \mathbf{A}_{*m}$, that is, the mth column of $\mathbf{A}$, be the parameter of interest.

The estimator (1.6) of $\boldsymbol{\beta}^*$ reduces to

$$\begin{aligned} \widehat{\boldsymbol{\beta}} = &\operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^d} \|\boldsymbol{\beta}\|_1 \\ &\text{subject to } \|\boldsymbol{S}_0\boldsymbol{\beta} - \boldsymbol{S}_{1,*m}\|_\infty \le \lambda, \end{aligned} \tag{4.9}$$

where $\lambda > 0$ is a tuning parameter, $\boldsymbol{S}_0 = T^{-1} \times \sum_{t=1}^{T} \boldsymbol{X}_t\boldsymbol{X}_t^T$ and $\boldsymbol{S}_1 = (T-1)^{-1}\sum_{t=1}^{T-1}\boldsymbol{X}_t\boldsymbol{X}_{t+1}^T$. In terms of our notation, we have that $\mathbf{t}(\{\boldsymbol{X}_t\}_{t=1}^T, \boldsymbol{\beta}) = \boldsymbol{S}_0\boldsymbol{\beta} - \boldsymbol{S}_{1,*m}$, and $E_{\mathbf{t}}(\mathbf{M}) = \boldsymbol{\Sigma}_0\boldsymbol{\beta} - \boldsymbol{\Sigma}_{1,*m}$,[3] where recall that $\boldsymbol{\Sigma}_0 = \mathrm{Cov}(\boldsymbol{X}_0, \boldsymbol{X}_0) = \boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}_1 = \mathrm{Cov}(\boldsymbol{X}_0, \boldsymbol{X}_1)$.

Han, Lu and Liu (2015) showed that procedure (4.9) consistently estimates $\boldsymbol{\beta}$ under certain sparsity assumptions. In the following, we apply our method to construct confidence intervals for θ, where θ is the first component of $\boldsymbol{\beta}$, that is, $\boldsymbol{\beta} = (\theta, \boldsymbol{\gamma})$. Following Algorithm 1, the projected estimating equation for θ is given by

$$\widehat{S}(\boldsymbol{\beta}) = \widehat{\mathbf{v}}^T(\boldsymbol{S}_0\boldsymbol{\beta} - \boldsymbol{S}_{1,*m}),$$

where

$$\widehat{\mathbf{v}} = \min_{\mathbf{v}\in\mathbb{R}^d} \|\mathbf{v}\|_1 \quad \text{subject to } \left\|\mathbf{v}^T\boldsymbol{S}_0 - \mathbf{e}_1\right\|_\infty \le \lambda',$$

is an estimator of $\mathbf{v}^{*T} = (\boldsymbol{\Sigma}_0^{-1})_{1*}$. Define $\widetilde{\theta}$ to be the solution to $\widehat{S}(\theta, \widehat{\boldsymbol{\gamma}}) = 0$. Note that in this framework the estimating equation $\mathbf{t}(\mathbf{X}, \boldsymbol{\beta}) = \boldsymbol{S}_0\boldsymbol{\beta} - \boldsymbol{S}_{1,*m}$ decomposes into a sum of dependent random variables. To handle this challenge, our main technical tool is the martingale central limit theorem and concentration inequalities for dependent random variables.

In the following, we will show that $T^{1/2}(\widetilde{\theta} - \theta^*)$ converges to $N(0, \Delta)$ in distribution, where

$$\Delta := \Psi_{mm}\mathbf{v}^{*T}\boldsymbol{\Sigma}_0\mathbf{v}^*. \tag{4.10}$$

Recall that $\boldsymbol{\Psi}$ is the covariance of the noise vectors $\mathbf{W}_t$ as introduced in the beginning of the section. In the Appendix, we argue that Ψ_{mm} is well estimated by $\boldsymbol{S}_{0,mm} - \widehat{\boldsymbol{\beta}}^T\boldsymbol{S}_0\widehat{\boldsymbol{\beta}}$. Hence let $\widehat{\Delta} = (\boldsymbol{S}_{0,mm} - \widehat{\boldsymbol{\beta}}^T\boldsymbol{S}_0\widehat{\boldsymbol{\beta}})(\widehat{\mathbf{v}}^T\boldsymbol{S}_0\widehat{\mathbf{v}})$ be an estimator of the asymptotic variance Δ, and define

$$\widehat{U}_n := \widehat{\Delta}^{-1/2}T^{1/2}(\widetilde{\theta} - \theta^*).$$

To establish the asymptotic normality of $\widetilde{\theta}$ (or equivalently of $\widehat{U}_n$), we define the following classes of matrices:

$$\begin{aligned} \mathcal{M}(s) := \Big\{&\mathbf{M} \in \mathbb{R}^{d\times d} : \max_{1\le j\le d} \|\mathbf{M}_{*j}\|_0 \le s, \|\mathbf{M}\|_1 \le M, \\ &\|\mathbf{M}\|_2 \le 1 - \varepsilon\Big\}, \\ \mathcal{L} := \big\{&\mathbf{M} \in \mathbb{R}^{d\times d} : \left\|\mathbf{M}^{-1}\right\|_1 \le M, \|\mathbf{M}\|_2 \le M\big\}, \end{aligned}$$

where M and $1 > \varepsilon > 0$ are some fixed constants. We have the following asymptotic normality result.

[3] Recall that subindexing a matrix with $*m$ indicates the mth column of this matrix.

COROLLARY 4. *Suppose* $\mathbf{\Sigma}_0 \in \mathcal{L}, \mathbf{A} \in \mathcal{M}(s)$, $\min_j \Psi_{jj} \geq C > 0$ *and* $\|\mathbf{v}^*\|_0 = s_{\mathbf{v}}$. *Then there exist* $\lambda \asymp \sqrt{\log d/T}$ *and* $\lambda' \asymp \sqrt{\log d/T}$ *such that if* $\max(s_{\mathbf{v}}, s) \log d = o(\sqrt{T})$, *we have for all* $t \in \mathbb{R}$

$$\lim_{T\to\infty} \left|\mathbb{P}^*(\widehat{U}_n \leq t) - \Phi(t)\right| \to 0.$$

Similar to Han, Lu and Liu (2015), we assume that the matrix $\mathbf{A}$ belongs to $\mathcal{M}(s)$, for the estimation purpose. The proof of Corollary 4 is given in Appendix J of the Supplementary material. In this section, we only discussed the lag-1 autoregressive model. As mentioned in Han, Lu and Liu (2015), lag-p models can be accommodated in the current lag-1 model framework. Thus, similar methods can be applied to construct confidence intervals under the lag-p model.

5. NUMERICAL RESULTS

In this section, we present numerical results to support our theoretical claims. Numerical studies on hypothesis testing are available from the authors upon request.

5.1 Linear Model

In this section, we compare our estimating equation (EE) based procedure with two existing methods: the desparsity (van de Geer et al., 2014) and the debias (Javanmard and Montanari, 2014) methods in linear models. Note that in their methods the LASSO estimator is used as an initial estimator.

Our simulation setup is as follows. We first generate $n = 150$ observations $\boldsymbol{X} \sim N(0, \mathbf{\Sigma}_{\boldsymbol{X}})$, where $\mathbf{\Sigma}_{\boldsymbol{X}}$ is a Toeplitz matrix with $\mathbf{\Sigma}_{\boldsymbol{X},ij} = \rho^{|i-j|}, i, j = 1, \ldots, d$. We consider three scenarios for the correlation parameter $\rho = 0.25, 0.4, 0.6$ and three possible values of the dimension $d = 100, 200, 500$. We generate $\boldsymbol{\beta}^*$ under two settings. In the first setting, $\boldsymbol{\beta}^*$ is held fixed, i.e., $\boldsymbol{\beta}^* = (1, 1, 1, 0, \ldots, 0)^T$, and in the second setting we take $\boldsymbol{\beta}^* = (U_1, U_2, U_3, 0, \ldots, 0)^T$, where U_i follows a uniform distribution on the interval $[0, 2]$ for $i = 1, 2, 3$. The former setting is labeled as "Dirac" and the latter as "Uniform" in Table 1 below. Both settings have three nonzero values, i.e., $\|\boldsymbol{\beta}^*\|_0 = 3$. The outcome is generated by $Y = \boldsymbol{X}^T\boldsymbol{\beta}^* + \varepsilon$, where $\varepsilon \sim N(0, 1)$. The simulations are repeated 500 times. The tuning parameter λ is selected by a 10-fold cross validation. The parameter λ' is manually set to $\frac{1}{2}\sqrt{\log d/n}$. Although its theoretical validity has not been formally proved we observed that the result is robust with respect to the choice of λ and λ'. Based on the selected λ and λ', we construct the confidence intervals for the first component of $\boldsymbol{\beta}$.

In Table 1, we summarize the empirical coverage probability of 95% confidence intervals and their average lengths of our estimating equation (EE) based method, desparsity and debias methods. We find that the empirical coverage probability of our method is very close to the desired nominal level. In particular, our method tends to have shorter confidence intervals than the existing two methods, when the dimension is large (e.g. $d = 500$).

5.2 Graphical Models

In this section we compare our estimating equation (EE) based procedure to the desparsity method proposed by Janková and van de Geer (2015) based on the graphical LASSO. We consider two scenarios. In the first scenario, our data generating process is similar to Janková and van de Geer (2015). Specifically, we consider a tridiagonal precision matrix $\mathbf{\Omega}$ with $\Omega_{ii} = 1, i = 1, \ldots, d$ and $\Omega_{i,i+1} = \Omega_{i+1,i} = \rho \in \{0.3, 0.4\}$ for $i = 1, \ldots, d-1$. Then we generate data from the Gaussian graphical model $\boldsymbol{X} \sim N(0, \mathbf{\Omega}^{-1})$. We have three settings for $d = 60, 70, 80$, and we fix the sample size at $n = 250$, which is comparable to Janková and van de Geer (2015). In the second scenario, we generate data from the transelliptical graphical model. Specifically, the latent generalized concentration matrix $\mathbf{\Omega}$ is generated in the same way as in the previous scenario, and then is normalized so that $\mathbf{\Sigma} = \mathbf{\Omega}^{-1}$, satisfies $\operatorname{diag}(\mathbf{\Sigma}) = 1$. Next, a normally distributed random vector $\boldsymbol{Z}$ is generated through $\boldsymbol{Z} \sim N(0, \mathbf{\Sigma})$, and is transformed to a new random vector $\boldsymbol{X} = (X_1, \ldots, X_d)$, where

$$X_j = \frac{f(Z_j)}{\sqrt{\int f^2(t)\phi(t)\,dt}},$$

and $f(t) := \operatorname{sign}(t)|t|^\alpha$ is a symmetric power transformation with $\alpha = 5$ and $\phi(t)$ is the p.d.f. of a standard normal distribution. Then $\boldsymbol{X}$ follows from the transelliptical graphical model with the latent generalized concentration matrix $\mathbf{\Omega}$. Similarly, we consider $d = 60, 70, 80$, and fix the sample size at $n = 250$. The simulations are repeated 500 times. The tuning parameters $\lambda = \lambda'$ are set equal to $0.5\sqrt{\log d/n}$. In the following, we construct confidence intervals for the parameter Ω_{12}.

In Table 2, we present the empirical coverage probability of 95% confidence intervals and their average lengths of our estimating equation (EE) based method, and the desparsity method. As expected, under the

TABLE 1

The empirical coverage percentages of 95% *confidence intervals constructed by our estimating equation* (*EE*) *based method*, *desparsity and debias methods under the linear model. The average lengths* (*multiplied by* 100) *of confidence intervals is shown in parenthesis*

		Uniform			Dirac		
d	Method	$\rho = 0.25$	$\rho = 0.4$	$\rho = 0.6$	$\rho = 0.25$	$\rho = 0.4$	$\rho = 0.6$
100	EE	94 (34)	95 (38)	95 (43)	95 (33)	96 (38)	95 (43)
	desparsity	96 (37)	96 (39)	95 (44)	96 (38)	95 (39)	94 (44)
	debias	95 (34)	95 (36)	94 (41)	95 (37)	94 (42)	95 (42)
200	EE	95 (33)	96 (41)	95 (45)	95 (35)	94 (41)	95 (45)
	desparsity	94 (32)	95 (38)	95 (47)	95 (39)	96 (44)	96 (47)
	debias	95 (38)	95 (38)	95 (43)	96 (38)	96 (42)	95 (45)
500	EE	96 (39)	96 (40)	95 (42)	95 (39)	96 (40)	95 (42)
	desparsity	96 (39)	95 (42)	95 (48)	96 (43)	96 (45)	96 (48)
	debias	95 (44)	95 (44)	94 (50)	95 (45)	95 (45)	94 (52)

Gaussian graphical model, the confidence intervals of both methods have accurate empirical coverage probability and similar lengths. However, the desparsity method which imposes the Gaussian assumption shows significant under-coverage for the transelliptical graphical model. In contrast, the proposed method preserves the nominal coverage probability, which demonstrates the numerical advantage of our method.

5.3 Real Data Analysis

In this section, we construct confidence regions for the gene network from the atlas of gene expression in the mouse aging project dataset (Zahn et al., 2007). The same dataset has been previously analyzed in Ning and Liu (2013), where the authors focus on a subset of $d = 37$ genes belonging to the mouse vascular endothelial growth factor signaling pathway in 8 tissues. The number of replicates within each tissue is $n = 40$.

Our analysis proceeds conditionally on each of the 8 tissue types—Adrenal (A), Cerebrum (C), Hippocampus (H), Kidney (K), Lung (L), Muscle (M), Spinal (S), Thymus (T). Namely, for each type of tissue, we construct the confidence intervals of each edges in the gene network by using our method and the procedure proposed by Janková and van de Geer (2015). In particular, our inference is based on the approach developed in Section 4.2 with the sample covariance matrix replaced by the rank covariance matrix defined in Remark 2; see also Appendix D in the Supplementary Material for details. The tuning parameter λ is determined by the 5-fold cross-validation, under the Gaussian likelihood function, for a grid of values in the interval $[0.3, 0.8]$, which is selected based on the fact that $\sqrt{\log d/n} \approx 0.3$. The tuning parameter λ' is set to be the same as λ. The tuning parameter in Janková and van de Geer (2015) is selected by the same cross-validation method.

TABLE 2

The empirical coverage probability of 95% *confidence intervals constructed by our estimating equation* (*EE*) *based method and the desparsity method under the Gaussian graphical model and transelliptical graphical model. The average length of confidence intervals is shown in parenthesis*

		Gaussian		Transelliptical	
d	Method	$\rho = 0.3$	$\rho = 0.4$	$\rho = 0.3$	$\rho = 0.4$
60	EE	0.95 (0.3)	0.94 (0.2)	0.93 (0.3)	0.94 (0.3)
	desparsity	0.95 (0.3)	0.95 (0.3)	0.80 (0.3)	0.44 (0.3)
70	EE	0.95 (0.3)	0.94 (0.2)	0.92 (0.3)	0.94 (0.3)
	desparsity	0.95 (0.3)	0.96 (0.3)	0.74 (0.3)	0.47 (0.3)
80	EE	0.95 (0.3)	0.95 (0.2)	0.93 (0.3)	0.94 (0.4)
	desparsity	0.94 (0.3)	0.94 (0.3)	0.70 (0.3)	0.44 (0.3)

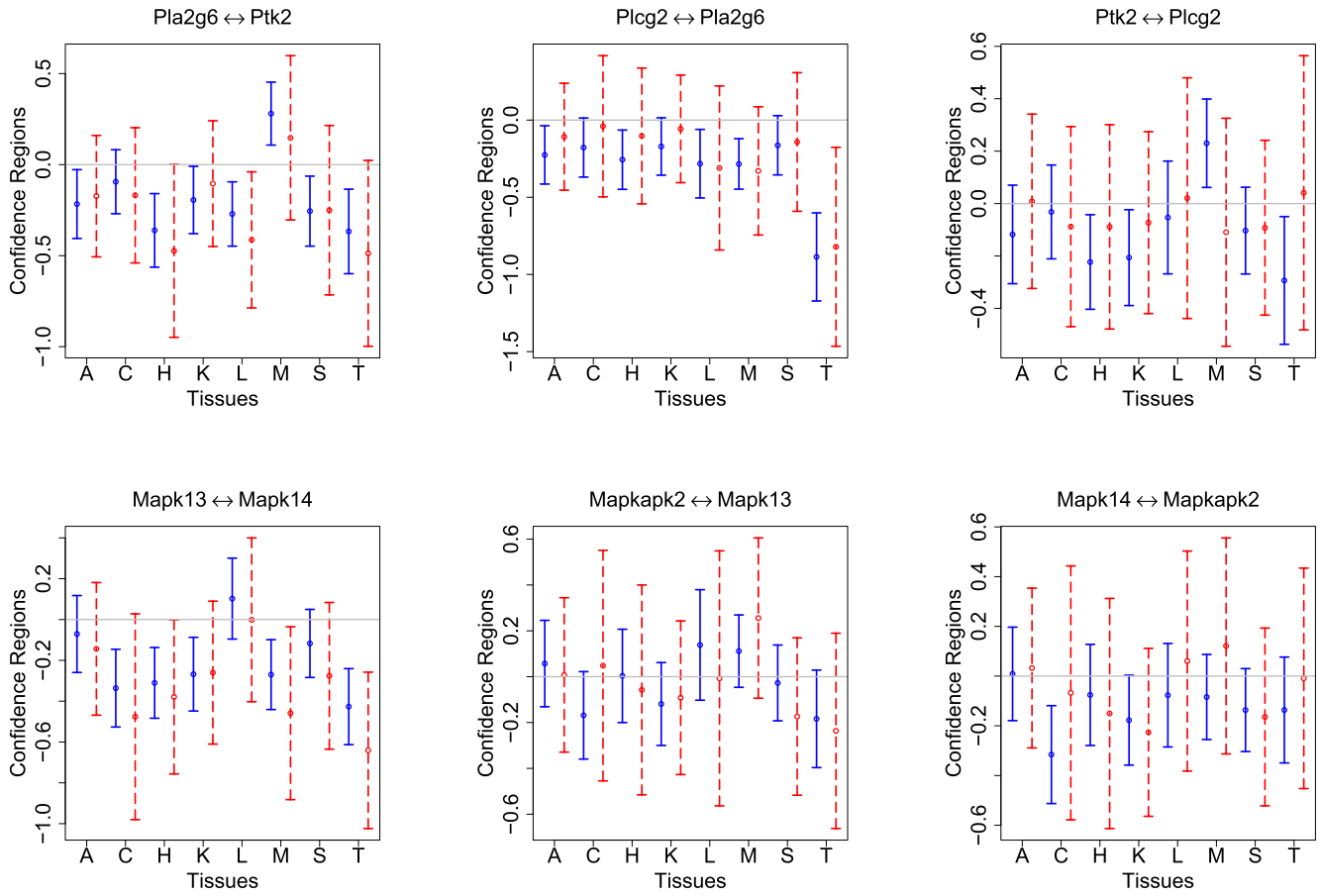

FIG. 1. *95% confidence intervals for the edges among the two sets of genes—Plcg2, Pla2g6 and Pla2g6 (first row) and Mapk13, Mapk14 and Mapkapk2 (second row), within each of 8 tissues indicated by their first letter. The confidence intervals based on our EE method are displayed in solid lines, while the intervals of Janková and van de Geer (2015) are displayed in dashed lines.*

To perform the comparison, we consider 2 sets of genes which have been shown to be associated by biologists. The first set of genes—Pla2g6, Ptk2 and Plcg2, comes from the group of PLC-γ genes in the PKC-dependent pathway, and is crucial for ERK phosphorylation and proliferation (Holmes et al., 2007). The second set of genes is comprised of Mapk13, Mapk14 and Mapkapk2, which are related to the migration of endothelial cells. Instead of plotting confidence intervals for all the edges in the gene network, in Figure 1 we only plot confidence intervals for the 3 edges connecting genes Pla2g6, Ptk2 and Plcg2, and genes Mapk13, Mapk14 and Mapkapk2, within each of the 8 tissues. As we see from the plot, while most of the point estimates of our method and Janková and van de Geer (2015) are close, their variances differ drastically. The main reason is that in this dataset the gene expression values are highly non-Gaussian; see Ning and Liu (2013) for demonstration. Thus, the inference procedure based on the Gaussian assumption (Janková and van de Geer, 2015) seems to provide inaccurate results with very wide confidence intervals. In contrast, the proposed method which relaxes the Gaussian assumption, produces confidence intervals with shorter length. In fact, most of the 95% confidence intervals by the proposed method do not cover 0, which concludes that these genes are statistically dependent. This result is consistent with the biological findings that genes Pla2g6, Ptk2 and Plcg2, and genes Mapk13, Mapk14 and Mapkapk2 are associated.

6. DISCUSSION

In this paper, we propose a generic procedure to construct confidence intervals for Z-estimators in a high-dimensional setting. We establish a general theoretical framework, and illustrate it with several important applications including linear models, instrumental variables regression, graphical models, classification and time series models. Our framework has better numerical performance than previously suggested algorithms,

and has the advantage of having a broader scope. In particular, it covers many applications (e.g., instrumental variables regression, linear discriminant analysis and vector autoregressive models) for which the inferential procedure is previously unexplored.

Additionally, our results can be easily extended to cases with multidimensional parameters of interest. We would like to mention that unlike approaches such as the ones developed by Nickl and van de Geer (2013), our methodology cannot be immediately extended to find a global *honest* confidence region for the entire parameter $\boldsymbol{\beta}$. It is an interesting problem to explore whether we can carry over certain results in the framework of honest confidence regions for $\boldsymbol{\beta}$ under the linear regression considered by Nickl and van de Geer (2013), to the general estimating equations that we consider. We leave this question for future investigation. Finally, we would like to discuss one caveat in the proposed method. If the equation $\mathbf{t}$ is nonconvex, it is less clear how one can find the global minimizer of the first step optimization (1.1) and there may exist multiple solutions of $\widehat{S}(\theta, \widehat{\boldsymbol{\gamma}}) = 0$. Although our theory continues to hold in such cases, the practical implementation requires extra attention. To this end, we make the following two comments. First, Chapter 1.2 of Zhao (2012) provided an alternative minimization approach, which can be used to define the first step estimator $\widehat{\boldsymbol{\beta}}$. Second, Small and Yang (1999) discussed how to choose roots when estimating equations have multiple roots. Their approach can be potentially applied to select the root of $\widehat{S}(\theta, \widehat{\boldsymbol{\gamma}}) = 0$.

ACKNOWLEDGMENTS

The authors are grateful to the Editor, Associate Editor and the anonymous referees for their suggestions which led to substantial improvements in the presentation of this work.

SUPPLEMENTARY MATERIAL

Supplement to "A Unified Theory of Confidence Regions and Testing for High-Dimensional Estimating Equations" (DOI: 10.1214/18-STS661SUPP; .pdf). This is the supplementary material to "A Unified Theory of Confidence Regions and Testing for High-Dimensional Estimating Equations" by M. Neykov, Y. Ning, H. Liu and J. Liu.

REFERENCES

BARBER, R. F. and KOLAR, M. (2015). Rocket: Robust confidence intervals via Kendall's tau for transelliptical graphical models. Preprint. Available at arXiv:1502.07641.

BELLONI, A., CHERNOZHUKOV, V. and HANSEN, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *Rev. Econ. Stud.* **81** 608–650. MR3207983

BELLONI, A., CHERNOZHUKOV, V. and KATO, K. (2015). Uniform post-selection inference for least absolute deviation regression and other Z-estimation problems. *Biometrika* **102** 77–94. MR3335097

BELLONI, A., CHERNOZHUKOV, V. and WEI, Y. (2013). Honest confidence regions for logistic regression with a large number of controls. Preprint. Available at arXiv:1304.3969.

CAI, T. T. and GUO, Z. (2017). Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. *Ann. Statist.* **45** 615–646. MR3650395

CAI, T. T., LIANG, T. and RAKHLIN, A. (2014). Geometrizing local rates of convergence for linear inverse problems. Preprint. Available at arXiv:1404.4408.

CAI, T. and LIU, W. (2011). A direct estimation approach to sparse linear discriminant analysis. *J. Amer. Statist. Assoc.* **106** 1566–1577. MR2896857

CAI, T., LIU, W. and LUO, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *J. Amer. Statist. Assoc.* **106** 594–607. MR2847973

CANDES, E. and TAO, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist.* **35** 2313–2351. MR2382644

CHEN, M., REN, Z., ZHAO, H. and ZHOU, H. (2016). Asymptotically normal and efficient estimation of covariate-adjusted Gaussian graphical model. *J. Amer. Statist. Assoc.* **111** 394–406. MR3494667

CHERNOZHUKOV, V., CHETVERIKOV, D. and KATO, K. (2014). Gaussian approximation of suprema of empirical processes. *Ann. Statist.* **42** 1564–1597. MR3262461

FAN, J. and LV, J. (2011). Nonconcave penalized likelihood with NP-dimensionality. *IEEE Trans. Inform. Theory* **57** 5467–5484. MR2849368

GAUTIER, E. and TSYBAKOV, A. (2011). High-dimensional instrumental variables regression and confidence sets. Preprint. Available at arXiv:1105.2454.

GODAMBE, V. P. (1991). *Estimating functions*. Clarendon Press, Oxford.

GU, Q., CAO, Y., NING, Y. and LIU, H. (2015). Local and global inference for high dimensional gaussian copula graphical models. Preprint. Available at arXiv:1502.02347.

HAN, F., LU, H. and LIU, H. (2015). A direct estimation of high dimensional stationary vector autoregressions. *J. Mach. Learn. Res.* **16** 3115–3150. MR3450535

HOLMES, K., ROBERTS, O. L., THOMAS, A. M. and CROSS, M. J. (2007). Vascular endothelial growth factor receptor-2: Structure, function, intracellular signalling and therapeutic inhibition. *Cellular Signalling* **19** 2003–2012.

JANKOVÁ, J. and VAN DE GEER, S. (2015). Confidence intervals for high-dimensional inverse covariance estimation. *Electron. J. Stat.* **9** 1205–1229. MR3354336

JAVANMARD, A. and MONTANARI, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *J. Mach. Learn. Res.* **15** 2869–2909. MR3277152

LEE, J. D., SUN, D. L., SUN, Y. and TAYLOR, J. E. (2013). Exact inference after model selection via the lasso. Preprint. Available at arXiv:1311.6238.

LIU, W. (2013). Gaussian graphical model estimation with false discovery rate control. *Ann. Statist.* **41** 2948–2978. MR3161453

LIU, H., HAN, F. and ZHANG, C.-H. (2012). Transelliptical graphical models. In *Advances in Neural Information Processing Systems*.

LOCKHART, R., TAYLOR, J., TIBSHIRANI, R. J. and TIBSHIRANI, R. (2014). A significance test for the lasso. *Ann. Statist.* **42** 413–468. MR3210970

LOH, P.-L. (2017). Statistical consistency and asymptotic normality for high-dimensional robust M-estimators. *Ann. Statist.* **45** 866–896. MR3650403

LU, S., LIU, Y., YIN, L. and ZHANG, K. (2015). Confidence intervals and regions for the lasso using stochastic variational inequality techniques in optimization. Technical report.

MARDIA, K. V., KENT, J. T. and BIBBY, J. M. (1979). *Multivariate Analysis*: *Probability and Mathematical Statistics*. Academic Press, London. MR0560319

MEINSHAUSEN, N. and BÜHLMANN, P. (2010). Stability selection. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **72** 417–473. MR2758523

MEINSHAUSEN, N., MEIER, L. and BÜHLMANN, P. (2009). p-values for high-dimensional regression. *J. Amer. Statist. Assoc.* **104** 1671–1681. MR2750584

NEWEY, W. K. and MCFADDEN, D. (1994). Large sample estimation and hypothesis testing. In *Handbook of Econometrics, Vol. IV. Handbooks in Econom.* **2** 2111–2245. North-Holland, Amsterdam. MR1315971

NEYKOV, M., NING, Y., LIU, J. S and LIU, H. (2018). Supplement to "A Unified Theory of Confidence Regions and Testing for High Dimensional Estimating Equations." DOI:10.1214/18-STS661SUPP.

NICKL, R. and VAN DE GEER, S. (2013). Confidence sets in sparse regression. *Ann. Statist.* **41** 2852–2876. MR3161450

NING, Y. and LIU, H. (2013). High-dimensional semiparametric bigraphical models. *Biometrika* **100** 655–670. MR3094443

NING, Y. and LIU, H. (2014). Sparc: Optimal estimation and asymptotic inference under semiparametric sparsity. Preprint. Available at arXiv:1412.2295.

NING, Y. and LIU, H. (2017). A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *Ann. Statist.* **45** 158–195. MR3611489

REN, Z., SUN, T., ZHANG, C.-H. and ZHOU, H. H. (2015). Asymptotic normality and optimalities in estimation of large Gaussian graphical models. *Ann. Statist.* **43** 991–1026. MR3346695

SHAH, R. D. and SAMWORTH, R. J. (2013). Variable selection with error control: Another look at stability selection. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **75** 55–80. MR3008271

SMALL, C. G. and YANG, Z. (1999). Multiple roots of estimating functions. *Canad. J. Statist.* **27** 585–598. MR1745824

TAYLOR, J., LOCKHART, R., TIBSHIRANI, R. J. and TIBSHIRANI, R. (2014). Post-selection adaptive inference for least angle regression and the lasso. Preprint. Available at arXiv:1401.3889.

TIAN, X. and TAYLOR, J. (2018). Selective inference with a randomized response. *Ann. Statist.* **46** 679–710. MR3782381

VAN DE GEER, S., BÜHLMANN, P., RITOV, Y. and DEZEURE, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.* **42** 1166–1202. MR3224285

VAN DER VAART, A. W. (1998). *Asymptotic Statistics*. *Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge Univ. Press, Cambridge. MR1652247

VERSHYNIN, R. (2012). Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing* 210–268. Cambridge Univ. Press, Cambridge. MR2963170

VOORMAN, A., SHOJAIE, A. and WITTEN, D. (2014). Inference in high dimensions with the penalized score test. Preprint. Available at arXiv:1401.2678.

WASSERMAN, L. and ROEDER, K. (2009). High-dimensional variable selection. *Ann. Statist.* **37** 2178–2201. MR2543689

ZAHN, J. M., POOSALA, S., OWEN, A. B., INGRAM, D. K., LUSTIG, A., CARTER, A., WEERARATNA, A. T., TAUB, D. D., GOROSPE, M., MAZAN-MAMCZARZ, K. et al. (2007). AGEMAP: A gene expression database for aging in mice. *PLoS Genet.* **3** e201.

ZHANG, C.-H. and ZHANG, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 217–242. MR3153940

ZHAO, S. D. (2012). Survival analysis with high-dimensional covariates, with applications to cancer genomics. Ph.D. thesis, Harvard Univ.

ZHU, Y. and BRADIC, J. (2016). Linear hypothesis testing in dense high-dimensional linear models. Preprint. Available at arXiv:1610.02987.