

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

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).

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 $E_{\mathbf{t}}(\boldsymbol{\beta}) = \lim_{n \rightarrow \infty} \mathbb{E} \mathbf{t}(\mathbf{Z}, \boldsymbol{\beta})$ ¹ denote the limiting expected value of the function $\mathbf{t}(\mathbf{Z}, \boldsymbol{\beta})$ as $n \rightarrow \infty$. 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

¹Here 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):

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

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 $\hat{\boldsymbol{\beta}} = (\hat{\theta}, \hat{\boldsymbol{\gamma}})$ 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 $\mathbb{E}(Y|X) = X^T \boldsymbol{\beta}^*$. Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ be the design matrix stacking the i.i.d. covariates $\{X_i\}_{i=1}^n$ and $\mathbf{Y} \in \mathbb{R}^n$ be the response vector with independent entries Y_i . Given the moment condition, we can easily construct the estimating function as $\mathbf{t}((Y, \mathbf{X}), \boldsymbol{\beta}) = n^{-1} \mathbf{X}^T (\mathbf{X}\boldsymbol{\beta} - \mathbf{Y})$ and $E_t(\boldsymbol{\beta}) = \mathbb{E} \mathbf{t}((Y, \mathbf{X}), \boldsymbol{\beta})$. In addition, $E_t(\boldsymbol{\beta}) = 0$ has the true value $\boldsymbol{\beta}^*$ as its unique root, provided that the second moment matrix $\boldsymbol{\Sigma}_X := n^{-1} \mathbb{E} \mathbf{X}^T \mathbf{X}$ is positive definite. In the high-dimensional setting, Candes and Tao (2007) estimated $\boldsymbol{\beta}$ by the following Dantzig selector:

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

Instrumental Variables Regression (IVR). Similar to the previous example, consider the linear model $Y = X^T \boldsymbol{\beta}^* + \varepsilon$. In economics' applications, it is not always reasonable to believe that the error and the design variables are uncorrelated, that is, $\mathbb{E}[X\varepsilon] = 0$, 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* $\mathbf{W} \in \mathbb{R}^d$ which are correlated with \mathbf{X} but satisfy $\mathbb{E}[\mathbf{W}\varepsilon] = 0$ and $\mathbb{E}[\varepsilon^2|\mathbf{W}] = \sigma^2$. 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 $\{X_i\}_{i=1}^n$ and instrumental variables $\{W_i\}_{i=1}^n$, respectively, and $\mathbf{Y} \in \mathbb{R}^n$ be the response vector with independent entries Y_i . Using the instrumental variables, one can construct the estimating function $\mathbf{t}((Y, \mathbf{X}, \mathbf{W}), \boldsymbol{\beta}) = n^{-1} \mathbf{W}^T (\mathbf{X}\boldsymbol{\beta} - \mathbf{Y})$ with $E_t(\boldsymbol{\beta}) = \mathbb{E} \mathbf{t}((Y, \mathbf{X}, \mathbf{W}), \boldsymbol{\beta})$. In addition, $E_t(\boldsymbol{\beta})$ has $\boldsymbol{\beta}^*$ as its unique root, provided that the second moment matrix $\boldsymbol{\Sigma}_{\mathbf{W}\mathbf{X}} := n^{-1} \mathbb{E} \mathbf{W}^T \mathbf{X}$ is of full rank. Inspired by Gautier and Tsybakov (2011), we consider the following estimator $\hat{\boldsymbol{\beta}}$:

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

Graphical Models via CLIME/SKEPTIC (Cai, Liu and Luo, 2011, Liu, Han and Zhang, 2012). Let X_1, \dots, X_n be i.i.d. copies of $X \in \mathbb{R}^d$ with $\mathbb{E}(X) = 0$ and $\operatorname{Cov}(X) = \boldsymbol{\Sigma}_X$. It is well known that in the case when X are Gaussian, the precision matrix $\boldsymbol{\Omega}^* = (\boldsymbol{\Sigma}_X)^{-1}$ induces a graph, encoding conditional independencies of the variables X . More generally, this observation can be extended to transelliptical distributions (Liu, Han and Zhang, 2012).

Let $\boldsymbol{\Sigma}_n = n^{-1} \sum_{i=1}^n X_i X_i^T$ be the sample covariance of X_1, \dots, X_n [recall $\mathbb{E}(X_i) = 0$]. Based on the second moment condition $\boldsymbol{\Sigma}_X \boldsymbol{\Omega}^* = \mathbf{I}_d$, Cai, Liu and Luo (2011) proposed the CLIME estimator of $\boldsymbol{\Omega}^*$:

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

In this case, we have $\mathbf{t}(\mathbf{X}, \boldsymbol{\Omega}) = \boldsymbol{\Sigma}_n \boldsymbol{\Omega} - \mathbf{I}_d$, and $E_t(\boldsymbol{\Omega}) = \boldsymbol{\Sigma}_X \boldsymbol{\Omega} - \mathbf{I}_d$. 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 \mathbf{X} and \mathbf{Y} be d -dimensional random vectors, coming from two populations with different means $\boldsymbol{\mu}_1 = \mathbb{E}(\mathbf{X})$, $\boldsymbol{\mu}_2 = \mathbb{E}(\mathbf{Y})$, and a common covariance matrix $\boldsymbol{\Sigma} = \text{Cov}(\mathbf{X}) = \text{Cov}(\mathbf{Y})$. Given some training samples, we are interested in classifying a new observation \mathbf{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

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

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 \mathbf{O} is classified into population 1 if and only if $\psi(\mathbf{O}) = 1$.

To implement $\psi(\mathbf{O})$ 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 $\mathbf{X}_1, \dots, \mathbf{X}_{n_1} \in \mathbb{R}^d$ and $\mathbf{Y}_1, \dots, \mathbf{Y}_{n_2} \in \mathbb{R}^d$. We assume that

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

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

In the high-dimensional setting with $d \gg n$, we cannot directly estimate $\boldsymbol{\Omega}$ by $\hat{\boldsymbol{\Sigma}}_n^{-1}$, 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:

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

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

Vector Autoregressive Models (Han, Lu and Liu, 2015). Let $\{\mathbf{X}_t\}_{t=-\infty}^\infty$ be a stationary sequence of mean

0 random vectors in \mathbb{R}^d with covariance matrix $\boldsymbol{\Sigma}$. The sequence $\{\mathbf{X}_t\}_{t=-\infty}^\infty$ is said to follow a lag-1 autoregressive model if

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

where \mathbf{A} is a $d \times d$ transition matrix, and the noise vectors \mathbf{W}_t are i.i.d. with $\mathbf{W}_t \sim N(0, \boldsymbol{\Psi})$ and independent of the history $\{\mathbf{X}_s\}_{s < t}$. Under the additional assumption that $\det(\mathbf{I}_d - \mathbf{A}^T z) \neq 0$ for all $z \in \mathbb{C}$ with $|z| \leq 1$, it can be shown that $\boldsymbol{\Psi}$ can be selected so that the process is stationary, i.e. for all t : $\mathbf{X}_t \sim N(0, \boldsymbol{\Sigma})$. Let $\boldsymbol{\Sigma}_i := \text{Cov}(\mathbf{X}_0, \mathbf{X}_i)$, 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

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

Assume that the data $(\mathbf{X}_1, \dots, \mathbf{X}_T)$ 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

$$(1.6) \quad \begin{aligned} \hat{\mathbf{A}} &= \argmin_{\mathbf{M} \in \mathbb{R}^{d \times d}} \sum_{1 \leq j, k \leq d} |M_{jk}| \\ &\text{subject to } \|\mathbf{S}_0 \mathbf{M} - \mathbf{S}_1\|_{\max} \leq \lambda, \end{aligned}$$

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

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 β , 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 $\mathbf{w}^T(\mathbf{Y} - \theta\mathbf{X}_{*1}) = 0$. 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 $n^{-1} \sum_{i=1}^n [\partial \ell(\mathbf{Z}_i, \beta) / \partial \theta - \mathbf{w}^T \partial \ell(\mathbf{Z}_i, \beta) / \partial \gamma]$, where $\ell(\mathbf{Z}_i, \beta)$ is the log-likelihood for data \mathbf{Z}_i , and $\mathbf{w}^T \partial \ell(\mathbf{Z}_i, \beta) / \partial \gamma$ is the sparse projection of the θ -score function $\partial \ell(\mathbf{Z}_i, \beta) / \partial \theta$ to the $(d - 1)$ -dimensional nuisance score space $\text{span}\{\partial \ell(\mathbf{Z}_i, \beta) / \partial \gamma\}$. 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 $\tilde{\theta}$ 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 $\tilde{\theta}$ is uniformly asymptotically normal for β^* 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 $\tilde{\theta}$ 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 $\mathbf{v} = (v_1, \dots, v_d)^T \in \mathbb{R}^d$, let $\|\mathbf{v}\|_q = (\sum_{i=1}^d |v_i|^q)^{1/q}$, $1 \leq q < \infty$, $\|\mathbf{v}\|_0 = |\text{supp}(\mathbf{v})|$, where $\text{supp}(\mathbf{v}) = \{j : v_j \neq 0\}$, 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 j th 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 $p \geq 1$. 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 $S \subset \{1, \dots, d\}$ let $\mathbf{v}_S = \{v_j : j \in S\}$ and S^c be the complement of S . We denote with ϕ , Φ , $\bar{\Phi}$ 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 $\ell \geq 1$ as

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

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 $\mathbb{P}(|X| > t) \leq \exp(1 - t/K_1)$ for all $t \geq 0$. Similarly, a random variable is sub-Gaussian, if there exists a $K_2 > 0$ such that $\mathbb{P}(|X| > t) \leq \exp(1 - t^2/K_2^2)$ for all $t \geq 0$. 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 $\limsup_n a_n/b_n \leq C$ and $\liminf_n a_n/b_n \geq c$.

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):

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

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

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

In the preceding display, we assume $E_{\mathbf{T}}(\boldsymbol{\beta}) := \lim_{n \rightarrow \infty} \mathbb{E} \mathbf{T}(\mathbf{Z}, \boldsymbol{\beta})^2$ 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 $\hat{\theta}$.

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):

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

which is the first stage of our algorithm. Due to the constraint in (1.1), the limiting distribution of the estimator $\hat{\boldsymbol{\beta}}$, and $\hat{\theta}$ 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 $\hat{\mathbf{v}}$:

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

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

$$(2.3) \quad \begin{aligned} \hat{\mathbf{v}} &= \operatorname{argmin} \|\mathbf{v}\|_1 \\ \text{such that } &\|\mathbf{v}^T \mathbf{T}(\mathbf{Z}, \hat{\boldsymbol{\beta}}) - \mathbf{e}_1\|_{\infty} \leq \lambda'. \end{aligned}$$

In (2.3), λ' is an additional tuning parameter, and \mathbf{e}_1 is a d -dimensional row vector $(1, 0, \dots, 0)$, where the position of 1 corresponds to that of θ among $\boldsymbol{\beta}$. It is easily seen that $\hat{\mathbf{v}}^T$ 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, $\hat{S}(\boldsymbol{\beta})$ 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

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

where

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

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

We emphasize that the construction of $\hat{\mathbf{v}}$ 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 $\hat{\boldsymbol{\beta}} = (\hat{\theta}, \hat{\boldsymbol{\gamma}})$. By plugging in the estimator $\hat{\boldsymbol{\gamma}}$, we obtain the projected estimating equation $\hat{S}(\theta, \hat{\boldsymbol{\gamma}})$ for the parameter of interest θ . Similar to the classical estimating equation approach, we propose to estimate θ by a Z-estimator $\tilde{\theta}$, which is the root of $\hat{S}(\theta, \hat{\boldsymbol{\gamma}}) = 0$. In practice, we can solve $\tilde{\theta}$ 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 $\tilde{\theta}$ is asymptotically normal.

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

²Recall 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 $\hat{\beta}$:

$$\begin{aligned} \hat{\beta} &= \operatorname{argmin} \|\beta\|_1 \\ &\text{subject to } \|\mathbf{t}(\mathbf{Z}, \beta)\|_\infty \leq \lambda; \end{aligned}$$

2. Calculate the projection direction $\hat{\mathbf{v}}^T$ through the following optimization based on (2.3):

$$\begin{aligned} \hat{\mathbf{v}} &= \operatorname{argmin} \|\mathbf{v}\|_1 \\ &\text{such that } \|\mathbf{v}^T \mathbf{T}(\mathbf{Z}, \hat{\beta}) - \mathbf{e}_1\|_\infty \leq \lambda'; \end{aligned}$$

3. Output the sparse projected test function $\hat{S}(\beta) = \hat{\mathbf{v}}^T \mathbf{t}(\mathbf{Z}, \beta)$. Solve

$$\hat{S}(\theta, \hat{\gamma}) = 0$$

to obtain the corrected estimate $\tilde{\theta}$. ($\hat{\gamma}$ is directly obtained from the first step estimate $\hat{\beta}$.)

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 $\tilde{\theta}$, which is the root of

$$\hat{S}(\theta, \hat{\gamma}) = 0,$$

as defined in Algorithm 1. Here, $\hat{\gamma}$ is directly obtained from the $\hat{\beta}$ 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 $\mathbf{t}(\mathbf{Z}, \beta)$ is twice differentiable in β . Recall that we further require β^* to be the unique solution to $E_{\mathbf{t}}(\beta) = 0$, where $E_{\mathbf{t}}(\beta) = \lim_{n \rightarrow \infty} \mathbb{E} \mathbf{t}(\mathbf{Z}, \beta)$ is

the limiting value of $\mathbb{E} \mathbf{t}(\mathbf{Z}, \beta)$ as we hold d fixed to its present value. For any β , we let $S(\beta) := \mathbf{v}^{*T} \mathbf{t}(\mathbf{Z}, \beta)$, where $\mathbf{v}^{*T} := [E_{\mathbf{T}}(\beta^*)]_{1*}^{-1}$. Let \mathbb{P}_β be the probability measure under the parameter β . We use the shorthand notation $\mathbb{P}^* = \mathbb{P}_{\beta^*}$, to indicate the measure under the true parameter β^* . For any vector $\beta = (\theta, \gamma)$, we use the following shorthand notation $\beta_{\check{\theta}} = (\check{\theta}, \gamma)$ 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 $\hat{\beta}$ and $\hat{\mathbf{v}}$ can be shown to be L_1 consistent, that is, $\|\hat{\beta} - \beta^*\|_1 = o_p(1)$ and $\|\hat{\mathbf{v}} - \mathbf{v}^*\|_1 = o_p(1)$. 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 $\hat{S}(\theta, \hat{\gamma}) = \hat{S}(\hat{\beta}_\theta)$ about the equation $S(\theta, \gamma^*) = S(\beta_\theta^*)$ [recall $S(\beta_\theta^*) := \mathbf{v}^{*T} \mathbf{t}(\mathbf{Z}, \beta_\theta^*)$], that is, we aim to establish $\hat{S}(\hat{\beta}_\theta) = S(\beta_\theta^*) + o_p(1)$. By the mean value theorem,

$$\begin{aligned} \hat{S}(\hat{\beta}_\theta) &= S(\beta_\theta^*) + \hat{\mathbf{v}}^T \mathbf{T}(\mathbf{Z}, \tilde{\beta}_v) (\hat{\beta}_\theta - \beta_\theta^*) \\ &\quad + (\hat{\mathbf{v}} - \mathbf{v}^*)^T \mathbf{t}(\mathbf{Z}, \beta_\theta^*), \end{aligned} \quad (3.1)$$

where $\tilde{\beta}_v$ is a point on the line segment joining $\hat{\beta}_\theta$ with β_θ^* . Owing to the L_1 consistency of $\hat{\mathbf{v}}$ and $\hat{\beta}$, (3.1) can indeed be rewritten in the form $\hat{S}(\hat{\beta}_\theta) = S(\beta_\theta^*) + o_p(1)$, provided that $\|\hat{\mathbf{v}}^T \mathbf{T}(\mathbf{Z}, \tilde{\beta}_v)\|_\infty = O_p(1)$ and $\|\mathbf{t}(\mathbf{Z}, \beta_\theta^*)\|_\infty = O_p(1)$. A sufficient and also sensible condition for these bounds, is to desire $\|\mathbf{t}(\mathbf{Z}, \beta_\theta^*) - E_{\mathbf{t}}(\beta_\theta^*)\|_\infty = o_p(1)$, and $\|\hat{\mathbf{v}}^T \mathbf{T}(\mathbf{Z}, \tilde{\beta}_v) - \mathbf{v}^{*T} \times E_{\mathbf{T}}(\beta_\theta^*)\|_\infty = o_p(1)$, where $E_{\mathbf{t}}(\beta_\theta^*)$ and $E_{\mathbf{T}}(\beta_\theta^*)$ are the limiting expected values of $\mathbf{t}(\mathbf{Z}, \beta_\theta^*)$ and $\hat{\mathbf{v}}^T \mathbf{T}(\mathbf{Z}, \tilde{\beta}_v)$, 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}_{θ^*} of θ^* , such that, for all $\theta \in \mathcal{N}_{\theta^*}$,

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

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

$$(3.4) \quad \lim_{n \rightarrow \infty} \mathbb{P}^*\left(\sup_{\nu \in [0, 1]} \|\hat{\mathbf{v}}^T \mathbf{T}(\mathbf{Z}, \tilde{\boldsymbol{\beta}}_{\nu}) - \mathbf{v}^{*T} E_{\mathbf{T}}(\boldsymbol{\beta}_{\theta}^*)\|_{\infty} \leq r_3(n, \theta)\right) = 1,$$

where $\tilde{\boldsymbol{\beta}}_{\nu} = \nu \hat{\boldsymbol{\beta}}_{\theta} + (1 - \nu) \boldsymbol{\beta}_{\theta}^*$, $\sup_{\theta \in \mathcal{N}_{\theta^*}} \max(r_1(n, \theta), r_2(n, \theta), r_3(n, \theta)) = o(1)$, 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 $\mathbf{T}(\mathbf{Z}, \tilde{\boldsymbol{\beta}}_{\nu})$ on $\hat{\mathbf{v}}$ 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 $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{v}}$ satisfy

$$(3.5) \quad \lim_{n \rightarrow \infty} \mathbb{P}^*(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \leq r_4(n)) = 1,$$

$$\lim_{n \rightarrow \infty} \mathbb{P}^*(\|\hat{\mathbf{v}} - \mathbf{v}^*\|_1 \leq r_5(n)) = 1,$$

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 $\theta \mapsto \hat{S}(\hat{\boldsymbol{\beta}}_{\theta})$ be either continuous or nondecreasing, and has a single root $\tilde{\theta}$. Furthermore, suppose that, for any $\varepsilon > 0$,*

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

Under Assumptions 1 and 2, we have that

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

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 $\hat{S}(\hat{\boldsymbol{\beta}}_{\theta})$ 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 $\hat{S}(\hat{\boldsymbol{\beta}}_{\hat{\theta}}) = o_p(1)$. To establish the asymptotic normality of $\hat{\theta}$, 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 $\boldsymbol{\Sigma} = \lim_{n \rightarrow \infty} n \text{Cov}(\mathbf{t}(\mathbf{Z}, \boldsymbol{\beta}^*))$, and assume that $\sigma^2 \geq C > 0$ 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

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

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

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

Assumption 5 is a technical condition, which says that the multiplication of the estimation errors of $\hat{\boldsymbol{\gamma}}$ (or $\hat{\mathbf{v}}$) by the error of the concentration inequalities (Assumption 1) is negligible in the bias of the final estimate $\tilde{\theta}$. This assumption is crucial for the $n^{1/2}$ -consistency of $\tilde{\theta}$, 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 $\hat{\sigma}^2$ is a consistent estimator of σ^2 , then for any $t \in \mathbb{R}$, we have*

$$\lim_{n \rightarrow \infty} |\mathbb{P}^*(\hat{U}_n \leq t) - \Phi(t)| = 0$$

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

Some generic sufficient conditions for the consistency of $\hat{\sigma}$ 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 $\hat{\sigma}^2$, Theorem 2 implies that we can construct a $(1 - \alpha)\%$ confidence interval of θ^* in the following way:

$$(3.9) \quad \begin{aligned} & \lim_{n \rightarrow \infty} \mathbb{P}^*(\theta^* \in [\tilde{\theta} - \Phi^{-1}(1 - \alpha/2)\hat{\sigma}/\sqrt{n}, \\ & \tilde{\theta} + \Phi^{-1}(1 - \alpha/2)\hat{\sigma}/\sqrt{n}]) \\ & = 1 - \alpha. \end{aligned}$$

We now note a property of our estimator $\tilde{\theta}$ 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 $-\mathbb{E} \mathbf{H}(\mathbf{Z}, \boldsymbol{\beta}^*) = \text{Cov} \mathbf{h}(\mathbf{Z}, \boldsymbol{\beta}^*)$, we have $\mathbf{v}^{*T} \boldsymbol{\Sigma} \mathbf{v}^* = (\boldsymbol{\Sigma}^{-1})_{11}$. In this case, the Z-estimator $\tilde{\theta}$ 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 $\mathbf{W} \equiv \mathbf{X}$. 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 $\boldsymbol{\Sigma}_X$. Our goal is to construct confidence intervals for the parameter θ . In the linear regression case, we can easily show that $\hat{S}(\boldsymbol{\beta})$ reduces to

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

where

$$(4.1) \quad \begin{aligned} & \hat{\mathbf{v}} = \underset{\mathbf{v}}{\text{argmin}} \|\mathbf{v}\|_1 \\ & \text{subject to } \|\mathbf{v}^T \boldsymbol{\Sigma}_n - \mathbf{e}_1\|_\infty \leq \lambda', \end{aligned}$$

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

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

$$\|\varepsilon\|_{\psi_2} := K < \infty, \quad \sup_{j \in \{1, \dots, d\}} \|\mathbf{X}_j\|_{\psi_2} := K_X < \infty,$$

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

While assumption that the smallest eigenvalue of Σ_X is bounded away from 0 could be somewhat restrictive given that the dimension of Σ_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 $\hat{U}_n = \hat{\Delta}^{-1} n^{1/2}(\tilde{\theta} - \theta^*)$, where $\tilde{\theta}$ is defined as the solution to $\hat{S}(\theta, \hat{\gamma}) = 0$, and

$$(4.2) \quad \hat{\Delta} := \hat{\mathbf{v}}^T \Sigma_n \hat{\mathbf{v}} n^{-1} \sum_{i=1}^n (Y_i - X_i^T \hat{\beta})^2,$$

is an estimator of the asymptotic variance $\Delta := \mathbf{v}^{*T} \Sigma_X \mathbf{v}^* \text{Var}(\varepsilon)$. In high-dimensional models, it is often reasonable to assume that the vector β^* is sparse. Additionally, if we are in a setting where X_1 is expected to be conditionally uncorrelated with many entries of the vector X_{-1} , it is also reasonable to postulate that \mathbf{v}^* is sparse. Let s and s_v denote the sparsity of β^* and \mathbf{v}^* correspondingly, that is, $\|\beta^*\|_0 = s$ and $\|\mathbf{v}^*\|_0 = s_v$. The next corollary of the general Theorem 2 shows the asymptotic normality of \hat{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*

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

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

$$\lim_{n \rightarrow \infty} |\mathbb{P}^*(\hat{U}_n \leq t) - \Phi(t)| = 0.$$

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 $s_v^3/n = o(1)$, we can show that \hat{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 Σ_X satisfy $\lambda_{\min}(\Sigma_X) \geq \delta > 0$, where δ is some fixed constant. Similar to Section 4.1, we assume that X is coordinatewise sub-Gaussian, that is,

$$(4.3) \quad K_X := \max_{j \in \{1, \dots, d\}} \|X_j\|_{\psi_2} < \infty,$$

for some fixed constant $K_X > 0$. Our goal is to construct confidence intervals for a component of Ω^* , where $\Omega^* = (\Sigma_X)^{-1}$. Without loss of generality, we focus on the parameter Ω_{1m}^* for some $m \in \{1, \dots, d\}$. When 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 X is sub-Gaussian, whereas the majority of the existing methods require the data to be sampled from Gaussian or Gaussian copula distributions.

Let $\beta^* := \Omega_{*m}^*$, be the m th column of Ω^* . Then the CLIME estimator of β^* given by (1.2) reduces to

$$\hat{\beta} = \text{argmin} \|\beta\|_1 \quad \text{subject to } \|\Sigma_n \beta - \mathbf{e}_m^T\|_{\infty} \leq \lambda,$$

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

$$\hat{S}(\beta) = \hat{\mathbf{v}}^T (\Sigma_n \beta - \mathbf{e}_m^T),$$

where

$$(4.4) \quad \begin{aligned} \hat{\mathbf{v}} &= \text{argmin} \|\mathbf{v}\|_1 \\ \text{such that } \|\mathbf{v}^T \Sigma_n - \mathbf{e}_1\|_{\infty} &\leq \lambda'. \end{aligned}$$

Here, $\hat{\mathbf{v}}$ is an estimate of $\mathbf{v}^* := (\Sigma_X)_{*1}^{-1} = \Omega_{*1}^*$. Notice that, due to the symmetry of $\hat{\beta}$ and $\hat{\mathbf{v}}$, if we take $\lambda = \lambda'$, it suffices to simply solve the CLIME optimization (1.2) once in order to evaluate $\hat{S}(\hat{\beta})$, as $\hat{\beta} = \hat{\Omega}_{*m}$ and $\hat{\mathbf{v}} = \hat{\Omega}_{*1}$. 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 $\tilde{\theta}$ to the equation $\hat{S}(\theta, \hat{\gamma}) = 0$ has the following closed form expression:

$$(4.5) \quad \tilde{\theta} = \hat{\theta} - \frac{\hat{\mathbf{v}}^T (\Sigma_n \hat{\beta} - \mathbf{e}_m^T)}{\hat{\mathbf{v}}^T \Sigma_{n,*1}}.$$

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

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

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

We note that Assumption 7 is natural. For example, when $\mathbf{X} \sim N(0, \Sigma_X)$, Isserlis' theorem yields that for any two vectors ξ and θ ,

$$\begin{aligned} \text{Var}(\xi^T \mathbf{X} \mathbf{X}^T \theta) &= (\xi^T \Sigma_X \xi)(\theta^T \Sigma_X \theta) + (\xi^T \Sigma_X \theta)^2 \\ &\geq \lambda_{\min}^2(\Sigma_X) \|\xi\|_2^2 \|\theta\|_2^2, \end{aligned}$$

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

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

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

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

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

$$\lim_{n \rightarrow \infty} |\mathbb{P}^*(\hat{U}_n \leq t) - \Phi(t)| = 0.$$

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 $\tilde{\Delta} := \hat{v}_1 \hat{\beta}_m + \hat{v}_m \hat{\beta}_1$ of Δ , which can also be shown to be consistent under the assumption $\max(s_v, s) \sqrt{\log d/n} = o(1)$. The second part of condition (4.6) is mild, since it is only slightly stronger than $n^{-1} s_v s = o(1)$. 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 Σ_n in (1.2) and (4.4) by

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

where

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

Similar to Corollary 2, the asymptotic normality of the estimator $\tilde{\theta}$ 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 $\hat{\psi}(\mathbf{O})$ (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 β , that is, $\beta = (\theta, \gamma)$. Note that if $\theta = 0$, then it implies that the first feature of \mathbf{O} is not needed in the Bayes' rule $\psi(\mathbf{O})$. Hence, our procedure can be used to assess whether a certain feature is significant in the classification.

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

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

where

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

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

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

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

where V_{\min} is a positive constant. In addition, let $K_U = \max_{j \in \{1, \dots, d\}} \|\mathbf{U}_j\|_{\psi_2} < \infty$.

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

$$(4.7) \quad \Delta := \alpha V_1 + (1 - \alpha) V_2,$$

and $\hat{U}_n := \hat{\Delta}^{-1/2} n^{1/2} (\tilde{\theta} - \theta^*)$, where $\hat{\Delta}$ is some consistent estimator of Δ . The explicit form of $\hat{\Delta}$ is complicated, and we defer its expression to Appendix I of the Supplementary Material. Denote $\|\boldsymbol{\beta}^*\|_0 = s$ and $\|\mathbf{v}^*\|_0 = s_v$. 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 $\lambda_{\min}(\Sigma) > \delta$ for some constant $\delta > 0$, and let Assumption 8 hold. If*

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

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

$$\lim_{n \rightarrow \infty} |\mathbb{P}^*(\hat{U}_n < t) - \Phi(t)| = 0.$$

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 $\{\mathbf{X}_t\}_{t=-\infty}^{\infty}$ be a stationary sequence of mean 0 random vectors in \mathbb{R}^d with covariance matrix Σ which is assumed to follow a lag-1 autoregressive model

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

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

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

$$(4.9) \quad \begin{aligned} & \hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^d}{\text{argmin}} \|\boldsymbol{\beta}\|_1 \\ & \text{subject to } \|\mathbf{S}_0 \boldsymbol{\beta} - \mathbf{S}_{1,*m}\|_{\infty} \leq \lambda, \end{aligned}$$

where $\lambda > 0$ is a tuning parameter, $\mathbf{S}_0 = T^{-1} \times \sum_{t=1}^T \mathbf{X}_t \mathbf{X}_t^T$ and $\mathbf{S}_1 = (T - 1)^{-1} \sum_{t=1}^{T-1} \mathbf{X}_t \mathbf{X}_{t+1}^T$. In terms of our notation, we have that $\mathbf{t}(\{\mathbf{X}_t\}_{t=1}^T, \boldsymbol{\beta}) = \mathbf{S}_0 \boldsymbol{\beta} - \mathbf{S}_{1,*m}$, and $E_{\mathbf{t}}(\mathbf{M}) = \Sigma_0 \boldsymbol{\beta} - \Sigma_{1,*m}$,³ where recall that $\Sigma_0 = \text{Cov}(\mathbf{X}_0, \mathbf{X}_0) = \Sigma$ and $\Sigma_1 = \text{Cov}(\mathbf{X}_0, \mathbf{X}_1)$.

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

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

where

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

is an estimator of $\mathbf{v}^{*T} = (\Sigma_0^{-1})_{1*}$. Define $\tilde{\theta}$ to be the solution to $\hat{S}(\theta, \hat{\boldsymbol{\gamma}}) = 0$. Note that in this framework the estimating equation $\mathbf{t}(\mathbf{X}, \boldsymbol{\beta}) = \mathbf{S}_0 \boldsymbol{\beta} - \mathbf{S}_{1,*m}$ 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 $T^{1/2}(\tilde{\theta} - \theta^*)$ converges to $N(0, \Delta)$ in distribution, where

$$(4.10) \quad \Delta := \Psi_{mm} \mathbf{v}^{*T} \Sigma_0 \mathbf{v}^*.$$

Recall that Ψ 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 $\mathbf{S}_{0,mm} - \hat{\boldsymbol{\beta}}^T \mathbf{S}_0 \hat{\boldsymbol{\beta}}$. Hence let $\hat{\Delta} = (\mathbf{S}_{0,mm} - \hat{\boldsymbol{\beta}}^T \mathbf{S}_0 \hat{\boldsymbol{\beta}})(\hat{\mathbf{v}}^T \mathbf{S}_0 \hat{\mathbf{v}})$ be an estimator of the asymptotic variance Δ , and define

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

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

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

$$\mathcal{L} := \left\{ \mathbf{M} \in \mathbb{R}^{d \times d} : \|\mathbf{M}^{-1}\|_1 \leq M, \|\mathbf{M}\|_2 \leq M \right\},$$

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

³Recall that subindexing a matrix with $_{*m}$ indicates the m th column of this matrix.

COROLLARY 4. *Suppose $\Sigma_0 \in \mathcal{L}$, $\mathbf{A} \in \mathcal{M}(s)$, $\min_j \Psi_{jj} \geq C > 0$ and $\|\mathbf{v}^*\|_0 = s_v$. Then there exist $\lambda \asymp \sqrt{\log d/T}$ and $\lambda' \asymp \sqrt{\log d/T}$ such that if $\max(s_v, s) \log d = o(\sqrt{T})$, we have for all $t \in \mathbb{R}$*

$$\lim_{T \rightarrow \infty} |\mathbb{P}^*(\hat{U}_n \leq t) - \Phi(t)| \rightarrow 0.$$

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 $\mathbf{X} \sim N(0, \Sigma_X)$, where Σ_X is a Toeplitz matrix with $\Sigma_{X,ij} = \rho^{|i-j|}$, $i, j = 1, \dots, d$. 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 β^* under two settings. In the first setting, β^* is held fixed, i.e., $\beta^* = (1, 1, 1, 0, \dots, 0)^T$, and in the second setting we take $\beta^* = (U_1, U_2, U_3, 0, \dots, 0)^T$, 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., $\|\beta^*\|_0 = 3$. The outcome is generated by $Y = \mathbf{X}^T \beta^* + \varepsilon$, 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 β .

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 Ω with $\Omega_{ii} = 1$, $i = 1, \dots, d$ and $\Omega_{i,i+1} = \Omega_{i+1,i} = \rho \in \{0.3, 0.4\}$ for $i = 1, \dots, d-1$. Then we generate data from the Gaussian graphical model $\mathbf{X} \sim N(0, \Omega^{-1})$. 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 Ω is generated in the same way as in the previous scenario, and then is normalized so that $\Sigma = \Omega^{-1}$, satisfies $\text{diag}(\Sigma) = 1$. Next, a normally distributed random vector \mathbf{Z} is generated through $\mathbf{Z} \sim N(0, \Sigma)$, and is transformed to a new random vector $\mathbf{X} = (X_1, \dots, X_d)$, where

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

and $f(t) := \text{sign}(t)|t|^\alpha$ is a symmetric power transformation with $\alpha = 5$ and $\phi(t)$ is the p.d.f. of a standard normal distribution. Then \mathbf{X} follows from the transelliptical graphical model with the latent generalized concentration matrix Ω . 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

d	Method	Uniform			Dirac		
		$\rho = 0.25$	$\rho = 0.4$	$\rho = 0.6$	$\rho = 0.25$	$\rho = 0.4$	$\rho = 0.6$
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

d	Method	Gaussian		Transelliptical	
		$\rho = 0.3$	$\rho = 0.4$	$\rho = 0.3$	$\rho = 0.4$
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)

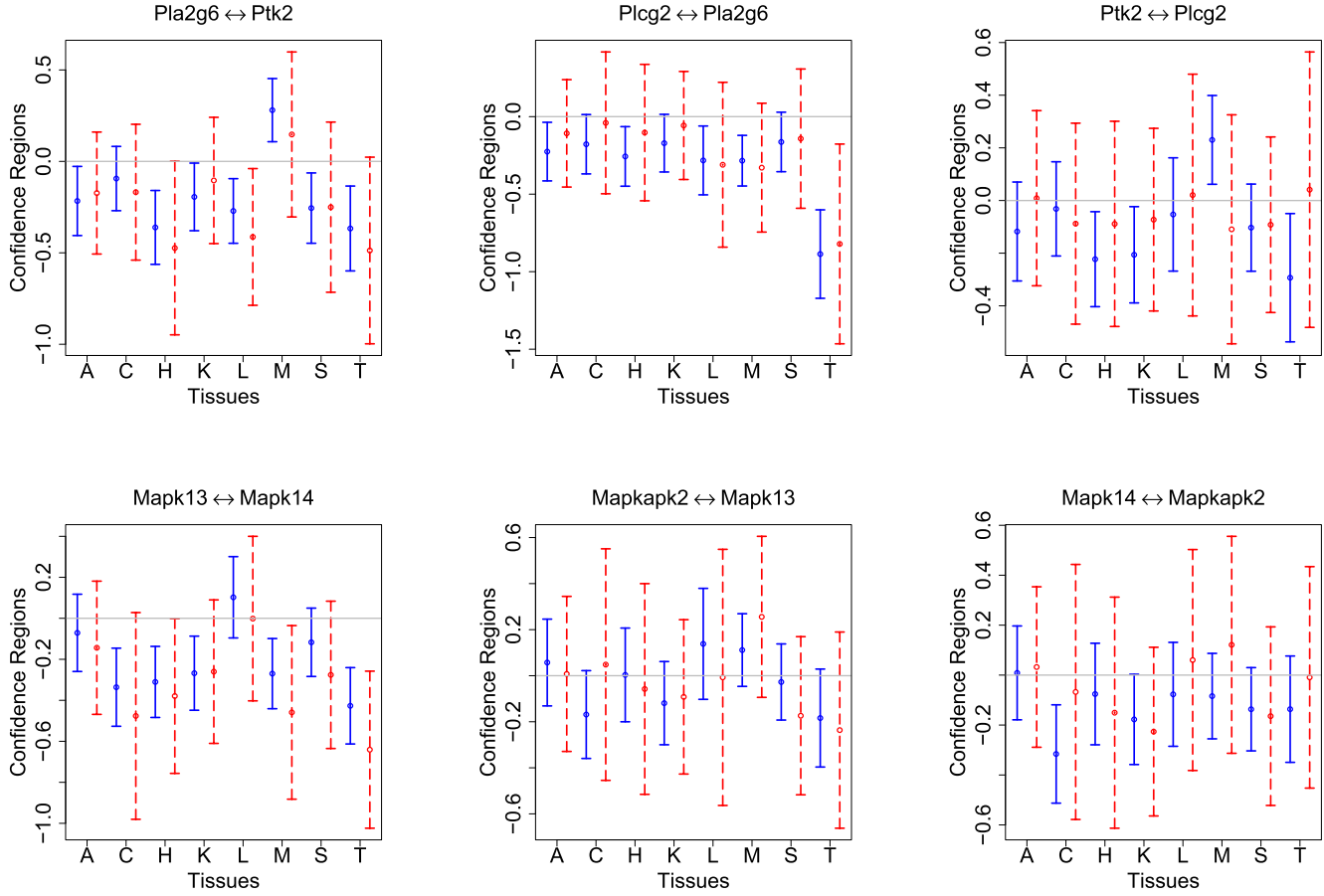


FIG. 1. 95% confidence intervals for the edges among the two sets of genes—*Plcg2*, *Pla2g6* and *Pla2g6* (first row) and *Mapk13*, *Mapk14* and *Mapkapk2* (second row), within each of 8 tissues indicated by their first letter. The confidence intervals based on our EE method are displayed in solid lines, 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 β . It is an interesting problem to explore whether we can carry over certain results in the framework of honest confidence regions for β 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 $\hat{S}(\theta, \hat{\mathcal{P}}) = 0$. 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 $\hat{\beta}$. 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 $\hat{S}(\theta, \hat{\mathcal{P}}) = 0$.

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](https://doi.org/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](https://arxiv.org/abs/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](https://arxiv.org/abs/1302.07983)
- 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](https://arxiv.org/abs/1304.3969)
- 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](https://arxiv.org/abs/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](https://arxiv.org/abs/1404.4408)
- CAI, T. T., LIANG, T. and RAKHLIN, A. (2014). Geometrizing local rates of convergence for linear inverse problems. Preprint. Available at [arXiv:1404.4408](https://arxiv.org/abs/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](https://arxiv.org/abs/1006.1566)
- 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](https://arxiv.org/abs/1006.1566)
- CANDES, E. and TAO, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n . *Ann. Statist.* **35** 2313–2351. [MR2382644](https://arxiv.org/abs/0606001)
- 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](https://arxiv.org/abs/1502.02347)
- CHERNOZHUKOV, V., CHETVERIKOV, D. and KATO, K. (2014). Gaussian approximation of suprema of empirical processes. *Ann. Statist.* **42** 1564–1597. [MR3262461](https://arxiv.org/abs/1304.3969)
- FAN, J. and LV, J. (2011). Nonconcave penalized likelihood with NP-dimensionality. *IEEE Trans. Inform. Theory* **57** 5467–5484. [MR2849368](https://arxiv.org/abs/1006.1566)
- GAUTIER, E. and TSYBAKOV, A. (2011). High-dimensional instrumental variables regression and confidence sets. Preprint. Available at [arXiv:1105.2454](https://arxiv.org/abs/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](https://arxiv.org/abs/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](https://arxiv.org/abs/1502.02347)
- 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](https://arxiv.org/abs/1304.3969)
- JAVANMARD, A. and MONTANARI, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *J. Mach. Learn. Res.* **15** 2869–2909. [MR3277152](https://arxiv.org/abs/1304.3969)
- 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](https://arxiv.org/abs/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](#)
- NEWBY, 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](#).

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

Matey Neykov Yang Ning Jun S. Liu Han Liu

APPENDIX A: THEORY ON UNIFORMLY VALID CONFIDENCE INTERVALS

In Section 3, we showed that if β^* is fixed, the solution to the equation $\widehat{S}(\theta, \widehat{\gamma}) = 0$ can be used to construct asymptotically valid confidence regions for the parameter θ . In this Section we prove a stronger result which guarantees that the confidence interval is valid uniformly over the following parameter space:

$$\Omega = \{\beta^* : \|\beta^*\|_0 \leq s^*\}.$$

We restrict our attention to Ω since we need the parameter β^* to be sufficiently sparse in order for us to consistently estimate it. In the following, we introduce some assumptions guaranteeing uniform convergence.

Assumption A.1 (Uniform Consistent Estimation).

$$\lim_{n \rightarrow \infty} \sup_{\beta \in \Omega} \mathbb{P}_\beta(\|\widehat{\beta} - \beta\|_1 \leq r_1(n)) = 1, \quad \lim_{n \rightarrow \infty} \sup_{\beta \in \Omega} \mathbb{P}_\beta(\|\widehat{\mathbf{v}} - \mathbf{v}\|_1 \leq r_2(n)) = 1,$$

where $r_1(n), r_2(n) = o(1)$.

Assumption A.2. Assume that there exists an $\eta > 0$ such that:

$$(A.1) \quad \lim_{n \rightarrow \infty} \inf_{\beta \in \Omega} \inf_{\tilde{\theta} \in \mathcal{N}_\theta} \mathbb{P}_\beta(\|\mathbf{t}(\mathbf{Z}, \beta_{\tilde{\theta}}) - E_{\mathbf{t}}(\beta_{\tilde{\theta}})\|_\infty \leq r_3(n)) = 1,$$

$$(A.2) \quad \lim_{n \rightarrow \infty} \inf_{\beta \in \Omega} \inf_{\tilde{\theta} \in \mathcal{N}_\theta} \mathbb{P}_\beta(|\mathbf{v}^T \mathbf{t}(\mathbf{Z}, \beta_{\tilde{\theta}}) - \mathbf{v}^T E_{\mathbf{t}}(\beta_{\tilde{\theta}})| \leq r_4(n)) = 1,$$

$$(A.3) \quad \lim_{n \rightarrow \infty} \inf_{\beta \in \Omega} \inf_{\tilde{\theta} \in \mathcal{N}_\theta} \mathbb{P}_\beta\left(\sup_{\nu \in [0,1]} \|\widehat{\mathbf{v}}^T \mathbf{T}(\mathbf{Z}, \widetilde{\beta}_\nu) - \mathbf{v}^T E_{\mathbf{T}}(\beta_{\tilde{\theta}})\|_\infty \leq r_5(n)\right) = 1,$$

where $\mathcal{N}_\theta = (\theta - \eta, \theta + \eta)$ and $\max(r_3(n), r_4(n), r_5(n)) = o(1)$, $\widetilde{\beta}_\nu = \nu \widehat{\beta}_{\tilde{\theta}} + (1 - \nu) \beta_{\tilde{\theta}}$. We also assume

$$\sup_{\beta \in \Omega} \sup_{\tilde{\theta} \in \mathcal{N}_\theta} \|E_{\mathbf{t}}(\beta_{\tilde{\theta}})\|_\infty < \infty, \quad \sup_{\beta \in \Omega} \sup_{\tilde{\theta} \in \mathcal{N}_\theta} \|\mathbf{v}^T [E_{\mathbf{T}}(\beta_{\tilde{\theta}})]_{-1}\|_\infty < \infty.$$

We next prove the uniform consistency of the Z-estimator $\widetilde{\theta}$, which is a uniform analogue of Theorem 1.

Proposition A.1. Assume that the (stochastic) map $\check{\theta} \mapsto \widehat{S}(\widehat{\beta}_{\check{\theta}})$ is continuous with a single root or nondecreasing. In addition, assume that $\mathbf{v}^T E_{\mathbf{t}}(\beta_{\theta-\epsilon}) \times \mathbf{v}^T E_{\mathbf{t}}(\beta_{\theta+\epsilon}) < 0$ for any $\epsilon > 0$. Under conditions A.1 and A.2 we have that for any $\epsilon > 0$: $\sup_{\beta \in \Omega} \mathbb{P}_{\beta}(|\check{\theta} - \theta| > \epsilon) = o(1)$.

Assumption A.3 (Uniform CLT). For $\sigma^2 = \mathbf{v}^T \Sigma \mathbf{v}$ we have:

$$(A.4) \quad \lim_{n \rightarrow \infty} \sup_{\beta \in \Omega} \sup_{t \in \mathbb{R}} |\mathbb{P}_{\beta}(\sigma^{-1} n^{1/2} S(\beta) \leq t) - \Phi(t)| = 0,$$

where $\Sigma = \lim_{n \rightarrow \infty} n \text{Cov } \mathbf{t}(\mathbf{Z}, \beta)$, and it is assumed that $\inf_{\beta \in \Omega} \sigma^2 \geq C > 0$.

Assumption A.4. Assume that there exists a $\gamma > 0$ such that:

$$(A.5) \quad \sup_{\beta \in \Omega} \sup_{\max(\|\check{\mathbf{v}} - \mathbf{v}\|_1, \|\check{\beta} - \beta\|_1) < \gamma} \left| \check{\mathbf{v}}^T \frac{\partial}{\partial \theta} [\mathbf{T}(\mathbf{Z}, \check{\beta})]_{*1} \right| \leq \psi(\mathbf{Z}),$$

where ψ is an integrable function such that $\sup_{\beta \in \Omega} \mathbb{E} \psi(\mathbf{Z}) < \infty$.

Assumption A.5 (Uniform Consistency of Variance). Assume there exists an estimator $\widehat{\sigma}^2$ of σ^2 , such that $\lim_{n \rightarrow \infty} \inf_{\beta \in \Omega} \mathbb{P}_{\beta}(|\widehat{\sigma}^2 - \sigma^2| \leq r_6(n)) = 1$, where $r_6(n) = o(1)$.

In Section B.2 we provide sufficient conditions to obtain $\widehat{\sigma}$ satisfying Assumption A.5. Finally, we present a uniform weak convergence result for the Z-estimator which strengthens Theorem 2. Its proof can be found in Appendix E.

Theorem A.1. Under Assumptions A.1 – A.5, the assumptions in Proposition A.1 and $n^{1/2}(r_1(n)r_5(n) + r_2(n)r_3(n)) = o(1)$, we have $\lim_{n \rightarrow \infty} \sup_{\beta \in \Omega} \sup_{t \in \mathbb{R}} |\mathbb{P}_{\beta}(\widehat{U}_n \leq t) - \Phi(t)| = 0$, where $\widehat{U}_n = \frac{n^{1/2}}{\widehat{\sigma}}(\check{\theta} - \theta)$.

Remark A.1. Notice that Theorem A.1 immediately implies that

$$\lim_{n \rightarrow \infty} \sup_{\beta \in \Omega} \sup_{t \in \mathbb{R}_+} |\mathbb{P}_{\beta}(|\widehat{U}_n| \leq t) - \Phi(t) + \Phi(-t)| = 0.$$

The latter can be equivalently expressed as

$$\lim_{n \rightarrow \infty} \sup_{\beta \in \Omega} \sup_{t \in \mathbb{R}_+} |\mathbb{P}_{\beta}(\theta \in (\check{\theta} - \widehat{\sigma}t/\sqrt{n}, \check{\theta} + \widehat{\sigma}t/\sqrt{n})) - \Phi(t) + \Phi(-t)| = 0,$$

which implies that the confidence region $(\check{\theta} - \widehat{\sigma}t/\sqrt{n}, \check{\theta} + \widehat{\sigma}t/\sqrt{n})$ is uniformly valid over the parameter space $\beta \in \Omega$ provided that the assumptions we discussed in this section hold.

APPENDIX B: SUFFICIENT CONDITIONS FOR VARIANCE CONSISTENCY

In this brief Section we present sufficient conditions for having consistent variance estimators.

B.1 Consistent Estimators of the Variance

We now provide generic sufficient conditions for constructing a consistent estimate of the variance. Let $\widehat{\Sigma}$ be a consistent estimator of $\text{Cov} \mathbf{t}(\mathbf{Z}, \boldsymbol{\beta})$. In the case when $\mathbf{t}(\mathbf{Z}, \boldsymbol{\beta}) = n^{-1} \sum_{i=1}^n \mathbf{h}(\mathbf{Z}_i, \boldsymbol{\beta})$ one can use $\widehat{\Sigma} := \frac{1}{n} \sum_{i=1}^n \mathbf{h}(\mathbf{Z}_i, \widehat{\boldsymbol{\beta}}) \mathbf{h}(\mathbf{Z}_i, \widehat{\boldsymbol{\beta}})^T$. We consider the “plugin” estimator: $\widehat{\sigma}^2 = \widehat{\mathbf{v}}^T \widehat{\Sigma} \widehat{\mathbf{v}}$ of σ^2 . Define: $\widehat{U}_n = \widehat{\sigma}^{-1} n^{1/2} (\widehat{\theta} - \theta^*)$. We are interested in showing that \widehat{U}_n converges weakly to a standard normal distribution. To this end we define the following assumption:

Assumption B.1 (Variance Consistency). Assume that the following holds:

$$\lim_{n \rightarrow \infty} \mathbb{P}^*(\|\widehat{\Sigma} - \Sigma\|_{\max} \leq r_7(n)) = 1,$$

where $r_7(n) = o(1)$.

Proposition B.1. Assume the same assumptions as in Proposition 2 plus Assumption B.1. Let furthermore $\|\Sigma\|_{\max} = O(1)$, $\|\mathbf{v}^{*T} \Sigma\|_{\infty} r_2(n) = o(1)$ and $\|\mathbf{v}^*\|_1^2 r_7(n) = o(1)$, then for any $t \in \mathbb{R}$ we have:

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

The proof of Proposition B.1 can be found in Section E.

B.2 Uniformly Consistent Estimators of the Variance

Assumption B.2 (Plugin Variance Consistency). Assume that the following holds:

$$\lim_{n \rightarrow \infty} \inf_{\boldsymbol{\beta} \in \Omega} \mathbb{P}_{\boldsymbol{\beta}}(\|\widehat{\Sigma} - \Sigma\|_{\max} \leq r_8(n)) = 1,$$

where $r_8(n) = o(1)$.

We then have the following:

Proposition B.2. Assume that

$$\sup_{\boldsymbol{\beta} \in \Omega} \|\Sigma\|_{\max} = O(1), \quad \sup_{\boldsymbol{\beta} \in \Omega} \|\mathbf{v}^T \Sigma\|_{\infty} r_2(n) = o(1), \quad \sup_{\boldsymbol{\beta} \in \Omega} \|\mathbf{v}\|_1^2 r_8(n) = o(1),$$

and $\widehat{\Sigma}$ satisfies Assumption B.2. Then $\widehat{\sigma}^2 = \widehat{\mathbf{v}}^T \widehat{\Sigma} \widehat{\mathbf{v}}$ satisfies Assumption A.5.

The proof of this proposition is omitted as it follows easily from the proof of Proposition B.1 and Lemma E.2 in the Appendix.

APPENDIX C: CONFIDENCE INTERVALS FOR INSTRUMENTAL VARIABLES REGRESSION

Recall that $\boldsymbol{\beta} := (\theta, \boldsymbol{\gamma})$, and let $\Sigma_n = n^{-1} [\mathbf{X}; \mathbf{W}]^T [\mathbf{X}; \mathbf{W}]$ be the empirical estimator of $\Sigma = \mathbb{E}(\mathbf{X}^T, \mathbf{W}^T)^T (\mathbf{X}^T, \mathbf{W}^T)$. Conformally we decompose the matrices Σ_n and Σ into four blocks corresponding to ordered pairings of \mathbf{X} and \mathbf{W} indicated by subscripting with the pair, e.g. $\Sigma_{\mathbf{W}\mathbf{X},n}$ and $\Sigma_{\mathbf{W}\mathbf{X}}$ correspond to $n^{-1} \mathbf{W}^T \mathbf{X}$ and $\mathbb{E} \mathbf{W} \mathbf{X}^T$ respectively. Our goal is to construct

confidence intervals for the parameter θ . It is easy to show that $\widehat{S}(\beta)$ reduces to

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

where

$$(C.1) \quad \widehat{\mathbf{v}} = \operatorname{argmin} \|\mathbf{v}\|_1, \quad \text{subject to} \quad \|\mathbf{v}^T \Sigma_{\mathbf{W}\mathbf{X},n} - \mathbf{e}_1\|_\infty \leq \lambda',$$

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

Assumption C.1. Let the error $\varepsilon := Y - \mathbf{X}^T \beta^*$, the predictors \mathbf{X} and instruments \mathbf{W} be coordinate-wise sub-Gaussian, i.e.,

$$\|\varepsilon\|_{\psi_2} := K < \infty, \quad \sup_{j \in \{1, \dots, d\}} \max(\|X_j\|_{\psi_2}, \|W_j\|_{\psi_2}) := K_{\mathbf{W}\mathbf{X}} < \infty,$$

for some fixed constants $K, K_{\mathbf{W}\mathbf{X}} > 0$ and $\operatorname{Var}(\varepsilon) \geq C_\varepsilon > 0$. Furthermore, assume $\lambda_{\min}(\Sigma_{\mathbf{W}\mathbf{X}} \Sigma_{\mathbf{X}\mathbf{W}}) \geq \delta^2 > 0$, where δ is some fixed constant. Additionally recall that $\mathbb{E}[\varepsilon^2 | \mathbf{Z}] = \sigma^2$, $\mathbb{E}[\mathbf{Z}\varepsilon] = 0$.

Assumption C.2. We impose the following assumptions on the covariance Σ . Assume:

$$\|\Sigma_{\mathbf{X}\mathbf{W}}^{-1} \Sigma_{\mathbf{W}\mathbf{W}} \Sigma_{\mathbf{W}\mathbf{X}}^{-1}\|_{\max} \leq D_{\max}, \quad \inf_{s \in \{1, \dots, d\}, s \log d < \sqrt{n}\omega} \operatorname{CS}_{\Sigma_{\mathbf{W}\mathbf{X}}}(s, 1) \geq \kappa^* > 0,$$

for some sufficiently large $D_{\max} > 0$, $\omega > 0$ is a sufficiently small fixed constant, and the quantity $\operatorname{CS}_{\Sigma_{\mathbf{W}\mathbf{X}}}(s, 1)$ is defined in Definition G.1.

Assumption C.1 is mild and ensures that the random variables $\varepsilon, \mathbf{X}, \mathbf{W}$ are not heavy-tailed, the matrix $\Sigma_{\mathbf{W}\mathbf{X}}$ is invertible, the instrumental variables \mathbf{W} are uncorrelated with ε , and that ε is homoscedastic given the instrumental variables. Assumption C.2 is a technical assumption. The first condition ensures that the random variable $\mathbf{v}^{*T} \mathbf{W}$ has a finite second moment, i.e. $\mathbb{E}(\mathbf{v}^{*T} \mathbf{W})^2 \leq D_{\max} < \infty$. The second condition implies that the matrix $\Sigma_{\mathbf{W}\mathbf{X}}$ is “coordinate-wise sensitive” with respect to the L_1 norm. Such a condition is first proposed by Gautier and Tsybakov (2011), and can be viewed as an extension of the commonly used restricted eigenvalue (RE) condition. It is not hard to show that this condition holds (for all $s \in \{1, \dots, d\}$) if for example $\lambda_{\min}(\Sigma_{\mathbf{X}\mathbf{W}} + \Sigma_{\mathbf{W}\mathbf{X}}) \geq 4\kappa^*$, where the inequality is in the sense of eigenvalues comparison.

To construct confidence intervals for θ , we consider $\widehat{U}_n = \widehat{\Delta}^{-1} n^{1/2} (\widehat{\theta} - \theta^*)$, where $\widehat{\theta}$ is defined as the solution to $\widehat{S}(\theta, \widehat{\gamma}) = 0$, and

$$(C.2) \quad \widehat{\Delta} := n^{-1} \sum_{i=1}^n ((\widehat{\mathbf{v}}^T \mathbf{W}_i)(Y_i - \mathbf{X}_i^T \widehat{\beta}))^2,$$

is an estimator of the asymptotic variance $\Delta := \operatorname{Var}[\mathbf{v}^{*T} \mathbf{W}\varepsilon]$. As in the linear model we will assume that \mathbf{v}^* and β^* are sparse. Let s and $s_{\mathbf{v}}$ denote the sparsity of β^* and \mathbf{v}^* correspondingly, i.e., $\|\beta^*\|_0 = s$ and $\|\mathbf{v}^*\|_0 = s_{\mathbf{v}}$. The next corollary of the general Theorem 2 shows the asymptotic normality of \widehat{U}_n in instrumental variable regressions. 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 C.1. Assume that condition C.1 and C.2 hold, 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 $\lambda' \asymp \sqrt{\log d / n}$, \hat{U}_n satisfies for any $t \in \mathbb{R}$:

$$\lim_{n \rightarrow \infty} |\mathbb{P}^*(\hat{U}_n \leq t) - \Phi(t)| = 0.$$

The proof of Corollary C.1 can be found in Appendix G. The conditions in Corollary C.1 agree with the existing conditions in Zhang and Zhang (2014); van de Geer et al. (2014) for the simple linear model. In fact, under the additional assumption $s_{\mathbf{v}}^3/n = o(1)$, we can show that \hat{U}_n is uniformly asymptotically normal; see Remark G.1.

APPENDIX D: CONFIDENCE INTERVALS FOR TRANSELLIPTICAL MODELS

In this subsection we consider the transelliptical graphical models (TGM), proposed by Liu et al. (2012b). We recall several definitions before we proceed.

Definition D.1 (Elliptical distribution Fang et al. (1990)). Let $\boldsymbol{\mu} \in \mathbb{R}^d$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$. We say that the d -dimensional vector \mathbf{X} has an elliptical distribution, and we denote it with $\mathbf{X} \sim EC_d(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \xi)$ if $\mathbf{X} \stackrel{d}{=} \boldsymbol{\mu} + \xi \mathbf{A} \mathbf{U}$, where \mathbf{U} is a random vector uniformly distributed on the unit sphere in \mathbb{R}^q , $\xi \geq 0$ is a scalar random variable independent of \mathbf{U} , $\mathbf{A} \in \mathbb{R}^{d \times q}$ is a deterministic matrix such that $\mathbf{A} \mathbf{A}^T = \boldsymbol{\Sigma}$.

Definition D.2 (Transelliptical distribution Liu et al. (2012b)). We call the continuous random vector $\mathbf{X} = (X_1, \dots, X_d)^T$ transelliptically distributed, and we denote it with $\mathbf{X} \sim TE_d(\boldsymbol{\Sigma}, \xi; f_1, \dots, f_d)$, if there exists a set of monotone univariate functions f_1, \dots, f_d and a non-negative random variable ξ , with $\mathbb{P}(\xi = 0) = 0$, such that:

$$(f_1(X_1), \dots, f_d(X_d))^T \sim EC_d(0, \boldsymbol{\Sigma}, \xi),$$

where $\boldsymbol{\Sigma}$ is symmetric with $\text{diag}(\boldsymbol{\Sigma}) = 1$ and $\boldsymbol{\Sigma} > 0$ in a positive-definite sense. Here $\boldsymbol{\Sigma}$ is called the “latent generalized correlation matrix”.

The graphical structure in TGMs can then be defined through the notion of the “latent generalized concentration matrix” — $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$, i.e. an edge is present between two variables X_j, X_k if and only if $\Omega_{jk} \neq 0$. To construct an estimate of $\boldsymbol{\Omega}$, Liu et al. (2012b) suggested estimating the correlation matrix $\boldsymbol{\Sigma}$ first. This can be done by using a non-parametric estimate of the correlation such as Kendall’s tau, and transforming it back, to obtain an estimate of $\boldsymbol{\Sigma}$.

Assume that $\mathbf{X}_1, \dots, \mathbf{X}_n$ are i.i.d. copies of \mathbf{X} . Recall the definitions of $\hat{\tau}_{jk}$ and $\hat{\mathbf{S}}_{jk}^\tau$ given in Remark 2. Note that it is clear from the definition of $\hat{\tau}_{jk}$, that it is an unbiased estimator of:

$$\tau_{jk} = \mathbb{P}((Y_j - Y'_j)(Y_k - Y'_k) > 0) - \mathbb{P}((Y_j - Y'_j)(Y_k - Y'_k) < 0),$$

where $\mathbf{Y}, \mathbf{Y}' \sim \mathbf{X}$ are i.i.d. random variables. It can be seen that $\hat{\mathbf{S}}_{jk}^\tau$ consistently estimates Σ (see e.g. (Liu et al., 2012a)). Let $\Omega^* = \Sigma^{-1}$. The TGM estimator with CLIME is given by:

$$\hat{\Omega} = \operatorname{argmin} \|\Omega\|_1, \text{ such that } \|\hat{\mathbf{S}}^\tau \Omega - \mathbf{I}_d\|_{\max} \leq \lambda.$$

To derive confidence intervals for the parameter Ω_{1m}^* , we can apply a similar approach to the graphical models. Denote with $\beta = \Omega_{*m}$. Then the CLIME with TGM reduces to

$$\hat{\beta} = \operatorname{argmin} \|\beta\|_1, \text{ such that } \|\hat{\mathbf{S}}^\tau \beta - \mathbf{e}_m^T\|_\infty \leq \lambda.$$

According to our formulation of the test statistic we have $\hat{S}(\beta) = \hat{\mathbf{v}}^T(\hat{\mathbf{S}}^\tau \beta - \mathbf{e}_m^T)$, where

$$(D.1) \quad \hat{\mathbf{v}} = \operatorname{argmin} \|\mathbf{v}\|_1, \text{ such that } \|\mathbf{v}^T \hat{\mathbf{S}}^\tau - \mathbf{e}_1\|_\infty \leq \lambda'.$$

The solution $\tilde{\theta}$ to equation $\hat{S}(\theta, \hat{\gamma}) = 0$, has a closed form expression in this example, and it is given below:

$$(D.2) \quad \tilde{\theta} = \hat{\theta} - \frac{\hat{\mathbf{v}}^T(\hat{\mathbf{S}}^\tau \hat{\beta} - \mathbf{e}_m^T)}{\hat{\mathbf{v}}^T \hat{\mathbf{S}}_{*1}^\tau}.$$

Next we argue that $\tilde{\theta}$ can be used to construct confidence intervals for the parameter θ . We note that the estimating equation in the TGM with CLIME is not a sum of i.i.d. statistics, due to the U -statistic structure of \mathbf{S}^τ as compared to estimating equations we considered in our previous examples. Nevertheless, the assumptions in Section 2 are general enough to handle such a case. Let Θ be a $d \times d$ random matrix with entries $\Theta_{jk} := \pi \cos\left(\frac{\pi}{2} \tau_{jk}\right) \tau_{jk}^{\mathbf{Y}}$, where:

$$\tau_{jk}^{\mathbf{Y}} = [\mathbb{P}((Y_j - Y_j')(Y_k - Y_k') > 0 | \mathbf{Y}) - \mathbb{P}((Y_j - Y_j')(Y_k - Y_k') < 0 | \mathbf{Y}) - \tau_{jk}],$$

with \mathbf{Y}, \mathbf{Y}' are i.i.d. copies of \mathbf{X} (and all $\tau_{jk}^{\mathbf{Y}}$ being a random variable depending on \mathbf{Y}). Define

$$(D.3) \quad \Delta := \operatorname{Var}(\mathbf{v}^{*T} \Theta \beta^*).$$

Assumption D.1. Let \mathbf{X} satisfy the following distributional assumption — there exists $\alpha_{\min} > 0$ such that $\Delta \geq \alpha_{\min} \|\mathbf{v}^*\|_2^2 \|\beta^*\|_2^2$.

Remark D.1. As in the CLIME case, we can show that the condition $\operatorname{Var}(\mathbf{v}^{*T} \Theta \beta^*) \geq \alpha_{\min} \|\mathbf{v}^*\|_2^2 \|\beta^*\|_2^2$ is equivalent to $\operatorname{Var}(\mathbf{v}^{*T} \Theta \beta^*) \geq V_{\min}$, assuming that the matrix $\Sigma \geq \delta > 0$.

Next, we proceed to define $\hat{\Delta}$ an estimate for Δ . To this end define the matrices $\hat{\Theta}^i$:

$$\hat{\tau}_{jk}^i := \frac{1}{n-1} \sum_{i' \neq i} \operatorname{sign}((X_{ij} - X_{i'j})(X_{ik} - X_{i'k})) - \hat{\tau}_{jk}, \quad \hat{\Theta}_{jk}^i := \pi \cos\left(\frac{\pi}{2} \hat{\tau}_{jk}^i\right) \hat{\tau}_{jk}^i.$$

Note that $\{\hat{\Theta}_{jk}^i\}_{jk}$ is symmetric by definition. Define the estimator $\hat{\Delta} := n^{-1} \sum_{i=1}^n (\hat{\mathbf{v}}^T \hat{\Theta}^i \hat{\beta})^2$. We have the following:

Corollary D.1. Assume the data follow from the transelliptical graphical models and Assumption D.1 holds. Furthermore, assume that the smallest eigenvalue of the correlation matrix satisfies $\lambda_{\min}(\Sigma) \geq \delta > 0$ is bounded away from 0, and $\text{diag}(\Sigma) = 1$. Let $\|\beta^*\|_0 = s$ and $\|\mathbf{v}^*\|_0 = s_{\mathbf{v}}$ and $\hat{\Delta}$ is a consistent estimate of Δ . Under the following additional assumptions

$$(D.4)$$

$$\max(s_{\mathbf{v}}, s) \|\mathbf{v}^*\|_1 \|\beta^*\|_1 \log d/n^{1/2} = o(1), \quad \|\mathbf{v}^*\|_1^2 \|\beta^*\|_1^2 \max(s_{\mathbf{v}}, s) \sqrt{\log d/n} = o(1),$$

$$(D.5) \quad \exists k > 2 : (s_{\mathbf{v}} s)^k / n^{k-1} = o(1),$$

we can take $\lambda \asymp \|\beta^*\|_1 \sqrt{\log d/n}$ and $\lambda' \asymp \|\mathbf{v}^*\|_1 \sqrt{\log d/n}$ so that the statistic $\hat{U}_n = \hat{\Delta}^{-1/2} n^{1/2} (\hat{\theta} - \theta)$ satisfies for all t : $\lim_{n \rightarrow \infty} |\mathbb{P}(\hat{U}_n \leq t) - \Phi(t)| = 0$.

Similarly, we can show the following result regarding uniformly valid confidence intervals. Before that we define a class of correlation matrices, in analogy to the CLIME case:

$$\tilde{\mathcal{S}}(L, s) = \{\Sigma : \Sigma = \Sigma^T, 0 < \delta \leq \lambda_{\min}(\Sigma), \text{diag}(\Sigma) = 1, \|\Sigma^{-1}\|_1 \leq L, \max_i \|\Sigma_{*i}^{-1}\|_0 \leq s\},$$

Corollary D.2. Let $\mathbf{X} \sim TE_d(\mu, \Sigma, \xi)$ with $\Sigma \in \tilde{\mathcal{S}}(L, s)$. Assume furthermore that \mathbf{X} satisfies Assumption D.1, and

$$(D.6) \quad s^3 n^{-1/2} = o(1), \quad s L^2 \log(d) n^{-1/2} = o(1), \quad s L^4 \sqrt{\log d/n} = o(1).$$

Then we have $\lim_{n \rightarrow \infty} \sup_{\Sigma \in \tilde{\mathcal{S}}(L, s)} \sup_t |\mathbb{P}_{\beta}(\hat{U}_n \leq t) - \Phi(t)| = 0$.

APPENDIX E: PROOFS OF THE GENERAL THEORY

Recall that $S(\beta) := \mathbf{v}^{*T} \mathbf{t}(\mathbf{Z}, \beta)$, and $\hat{S}(\beta) := \hat{\mathbf{v}}^T \mathbf{t}(\mathbf{Z}, \beta)$. We start by deriving an asymptotic expansion of the projected estimating equation $\hat{S}(\hat{\beta}_{\theta^*})$.

Lemma E.1. Suppose Assumptions (3.5), 1 and

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

hold. Then we have the following asymptotic expansion:

$$n^{1/2} \hat{S}(\hat{\beta}_{\theta^*}) = n^{1/2} S(\beta^*) + o_p(1).$$

PROOF OF LEMMA E.1. Let for brevity $r_1(n) = \sup_{\theta \in \mathcal{N}_{\theta^*}} r_1(n, \theta)$ and $r_3(n) = \sup_{\theta \in \mathcal{N}_{\theta^*}} r_3(n, \theta)$. We start by showing that for all $\theta \in \mathcal{N}_{\theta^*}$:

$$(E.2) \quad \hat{S}(\hat{\beta}_{\theta}) = S(\beta_{\theta}^*) + o_p(1) = \mathbf{v}^{*T} [E_{\mathbf{t}}(\beta_{\theta}^*)] + o_p(1).$$

This fact will be used in the proof of Theorem 1. By the mean value theorem we have:

$$\hat{S}(\hat{\beta}_{\theta}) = \mathbf{v}^{*T} \mathbf{t}(\mathbf{Z}, \beta_{\theta}^*) + \underbrace{\hat{\mathbf{v}}^T \mathbf{T}(\mathbf{Z}, \tilde{\beta}_{\nu}) (\hat{\beta}_{\theta} - \beta_{\theta}^*)}_{I_1} + \underbrace{(\hat{\mathbf{v}} - \mathbf{v}^*)^T \mathbf{t}(\mathbf{Z}, \beta_{\theta}^*)}_{I_2}$$

Next we control I_1 :

$$(E.3) \quad |I_1| \leq \|[\hat{\mathbf{v}}^T \mathbf{T}(\mathbf{Z}, \tilde{\boldsymbol{\beta}}_\nu)]_{-1}\|_\infty \|\hat{\boldsymbol{\beta}}_\theta - \boldsymbol{\beta}_\theta^*\|_1 \leq O_p(r_3(n) + \underbrace{\|\mathbf{v}^{*T} [E_{\mathbf{T}}(\boldsymbol{\beta}_\theta^*)]_{-1}\|_\infty}_{O(1)}) O_p(r_4(n))$$

where by $[\cdot]_{-1}$ we mean discarding the first entry (corresponding to θ) of the vector. We proceed to bound I_2 :

$$(E.4) \quad |I_2| \leq \|\hat{\mathbf{v}} - \mathbf{v}^*\|_1 \|\mathbf{t}(\mathbf{Z}, \boldsymbol{\beta}_\theta^*)\|_\infty = O_p(r_5(n)) O_p(r_1(n) + \underbrace{\|E_{\mathbf{t}}(\boldsymbol{\beta}_\theta^*)\|_\infty}_{O(1)}).$$

This combined with (3.3) concludes the our initial statement. For the influence function expansion stated in Lemma E.1 observe that $E_{\mathbf{t}}(\boldsymbol{\beta}_{\theta^*}^*) = 0$ and $\mathbf{v}^{*T} [E_{\mathbf{T}}(\boldsymbol{\beta}_{\theta^*}^*)]_{-1} = 0$ and hence using (E.1) we can modify bounds (E.3) and (E.4) to:

$$n^{1/2}(|I_1| + |I_2|) \leq n^{1/2} O_p(r_4(n) r_3(n, \theta^*) + r_5(n) r_1(n, \theta^*)) = o_p(1),$$

and we are done. \square

PROOF OF THEOREM 1. First assume that the map $\theta \mapsto \hat{S}(\hat{\boldsymbol{\beta}}_\theta)$ is continuous and has a unique 0. Take $\epsilon > 0$ so that both $\theta^* - \epsilon, \theta^* + \epsilon \in \mathcal{N}_{\theta^*}$. Without loss of generality let $\mathbf{v}^{*T} E_{\mathbf{t}}(\boldsymbol{\beta}_{\theta^* - \epsilon}) < 0$ and $\mathbf{v}^{*T} E_{\mathbf{t}}(\boldsymbol{\beta}_{\theta^* + \epsilon}) > 0$ for all $\epsilon > 0$. Note that by the continuity of $\theta \mapsto \hat{S}(\hat{\boldsymbol{\beta}}_\theta)$, and the fact that $\tilde{\theta}$ is the unique root $\mathbb{P}^*(\hat{S}(\hat{\boldsymbol{\beta}}_{\theta^* - \epsilon}) < 0, \hat{S}(\hat{\boldsymbol{\beta}}_{\theta^* + \epsilon}) > 0) \leq \mathbb{P}^*(\theta^* - \epsilon < \tilde{\theta} < \theta^* + \epsilon)$. Now by (E.2) from Lemma E.1 we have that the left hand side converges to 1 and this concludes the proof in the first case. The same argument goes through in the second case. \square

PROOF OF THEOREM 2. Let $U_n = \frac{n^{1/2}}{\sqrt{\mathbf{v}^{*T} \boldsymbol{\Sigma} \mathbf{v}^*}} (\tilde{\theta} - \theta^*)$. It suffices to show that the statement holds for U_n , as the statement for \hat{U}_n is a corollary after an application of Slutsky's theorem. By the mean value theorem:

$$(E.5) \quad \hat{S}(\hat{\boldsymbol{\beta}}_{\tilde{\theta}}) = \hat{S}(\hat{\boldsymbol{\beta}}_{\theta^*}) + \hat{\mathbf{v}}^T [\mathbf{T}(\mathbf{Z}, \hat{\boldsymbol{\beta}}_{\theta^*})]_{*1} (\tilde{\theta} - \theta^*) + \frac{1}{2} \hat{\mathbf{v}}^T \frac{\partial}{\partial \theta} [\mathbf{T}(\mathbf{Z}, \tilde{\boldsymbol{\beta}}_\nu)]_{*1} (\tilde{\theta} - \theta^*)^2,$$

where $\tilde{\boldsymbol{\beta}}_\nu = \nu \hat{\boldsymbol{\beta}}_{\tilde{\theta}} + (1 - \nu) \hat{\boldsymbol{\beta}}_{\theta^*}$ for some $\nu \in [0, 1]$. By (3.7) and the fact that $\hat{\mathbf{v}}$ and $\tilde{\boldsymbol{\beta}}_\nu$ are consistent (by (3.5) and Theorem 1) we have that:

$$\left| \frac{1}{2} \hat{\mathbf{v}}^T \frac{\partial}{\partial \theta} [\mathbf{T}(\mathbf{Z}, \tilde{\boldsymbol{\beta}}_\nu)]_{*1} (\tilde{\theta} - \theta^*)^2 \right| \leq (\tilde{\theta} - \theta^*)^2 O_p(1) = |\tilde{\theta} - \theta^*| o_p(1).$$

Observe that by Lemma E.1 and Assumption 3 we have $\frac{n^{1/2}}{\sqrt{\mathbf{v}^{*T} \boldsymbol{\Sigma} \mathbf{v}^*}} \hat{S}(\hat{\boldsymbol{\beta}}_{\theta^*}) = O_p(1)$, and more precisely $\frac{n^{1/2}}{\sqrt{\mathbf{v}^{*T} \boldsymbol{\Sigma} \mathbf{v}^*}} \hat{S}(\hat{\boldsymbol{\beta}}_{\theta^*}) \rightsquigarrow N(0, 1)$. Next by Assumption 1,

$$n^{1/2} |\hat{\theta} - \theta^*| (\underbrace{|\hat{\mathbf{v}}^T [\mathbf{T}(\mathbf{Z}, \hat{\boldsymbol{\beta}}_{\theta^*})]_{*1}|}_{1 + o_p(1)} + o_p(1)) = O_p(1),$$

and hence we conclude that $\hat{\theta} - \theta^* = O_p(n^{-1/2})$. Thus by Assumption 1 and Slutsky's:

$$\frac{n^{1/2}(\hat{\theta} - \theta^*)}{\sqrt{\mathbf{v}^{*T} \Sigma \mathbf{v}^*}} = -\frac{n^{1/2} \hat{S}(\hat{\beta}_{\theta^*})}{\sqrt{\mathbf{v}^{*T} \Sigma \mathbf{v}^*} \hat{\mathbf{v}}^T [\mathbf{T}(\mathbf{Z}, \hat{\beta}_{\theta^*})]_{*1}} + o_p(1) \rightsquigarrow N(0, 1),$$

which concludes the proof. \square

PROOF OF PROPOSITION A.1. We start by showing that for all $\epsilon > 0$:

$$\sup_{\beta \in \Omega} \sup_{\check{\theta} \in \mathcal{N}_\theta} \mathbb{P}_\beta \left(|\hat{S}(\hat{\beta}_{\check{\theta}}) - \mathbf{v}^T E_{\mathbf{t}}(\beta_{\check{\theta}})| > \epsilon \right) = o(1).$$

We have:

$$\begin{aligned} & \sup_{\beta \in \Omega} \sup_{\check{\theta} \in \mathcal{N}_\theta} \mathbb{P}_\beta \left(|\hat{S}(\hat{\beta}_{\check{\theta}}) - \mathbf{v}^T E_{\mathbf{t}}(\beta_{\check{\theta}})| > \epsilon \right) \\ & \leq \sup_{\beta \in \Omega} \sup_{\check{\theta} \in \mathcal{N}_\theta} \mathbb{P}_\beta \left(\left| \mathbf{v}^T \mathbf{t}(\mathbf{Z}, \beta_{\check{\theta}}) - \mathbf{v}^T E_{\mathbf{t}}(\beta_{\check{\theta}}) \right| > \frac{\epsilon}{3} \right) \\ & \quad + \sup_{\beta \in \Omega} \sup_{\check{\theta} \in \mathcal{N}_\theta} \mathbb{P}_\beta \left(\sup_{\nu \in [0,1]} \left\| \hat{\mathbf{v}}^T [\mathbf{T}(\mathbf{Z}, \tilde{\beta}_\nu)]_{-1} \right\|_\infty \|\hat{\beta}_{\check{\theta}} - \beta_{\check{\theta}}\|_1 > \frac{\epsilon}{3} \right) \\ \text{(E.6)} \quad & + \sup_{\beta \in \Omega} \sup_{\check{\theta} \in \mathcal{N}_\theta} \mathbb{P}_\beta \left(\|\hat{\mathbf{v}} - \mathbf{v}\|_1 \|\mathbf{t}(\mathbf{Z}, \beta_{\check{\theta}})\|_\infty > \frac{\epsilon}{3} \right), \end{aligned}$$

where $\tilde{\beta}_\nu = \nu \hat{\beta}_{\check{\theta}} + (1 - \nu) \beta_{\check{\theta}}$. By Assumptions A.1 and A.2 it follows that the RHS is $o(1)$, as we claimed. First let us assume that the maps $\check{\theta} \mapsto \hat{S}(\hat{\beta}_{\check{\theta}})$ are continuous. To shorten the notation in the remaining of the proof we will use the following notation: If A, B are random variables we write $\mathbb{P}(A \geq c_1, B \leq c_2) := \mathbb{P}((A > c_1 \cap B < c_2) \cup (A < c_1 \cap B > c_2))$. Then following inequality holds $\inf_{\beta \in \Omega} \mathbb{P}_\beta(\hat{S}(\hat{\beta}_{\theta-\epsilon}) \leq 0, \hat{S}(\hat{\beta}_{\theta+\epsilon}) \geq 0) \leq \inf_{\beta \in \Omega} \mathbb{P}_\beta(\theta - \epsilon < \tilde{\theta} < \theta + \epsilon)$. Next for small enough $\epsilon > 0$ such that $\theta \pm \epsilon \in \mathcal{N}_\theta$, we have:

$$\begin{aligned} & \inf_{\beta \in \Omega} \mathbb{P}_\beta(\mathbf{v}^T E_{\mathbf{t}}(\beta_{\theta-\epsilon}) \leq 0, \mathbf{v}^T E_{\mathbf{t}}(\beta_{\theta+\epsilon}) \geq 0) \\ & - \sup_{\beta \in \Omega} \mathbb{P}_\beta(\hat{S}(\hat{\beta}_{\theta-\epsilon}) \geq 2\mathbf{v}^T E_{\mathbf{t}}(\beta_{\theta-\epsilon}), \hat{S}(\hat{\beta}_{\theta+\epsilon}) \leq 2\mathbf{v}^T E_{\mathbf{t}}(\beta_{\theta+\epsilon})) \\ & \leq \inf_{\beta \in \Omega} \mathbb{P}_\beta(\hat{S}(\hat{\beta}_{\theta-\epsilon}) \leq 0, \hat{S}(\hat{\beta}_{\theta+\epsilon}) \geq 0). \end{aligned}$$

The LHS goes to 1 by (E.6), and hence the proof is complete in this case. In the case when $\check{\theta} \mapsto \hat{S}(\hat{\beta}_{\check{\theta}})$ are non-decreasing, exactly the same argument goes through. \square

Lemma E.2. Let $X_n(\mathbf{r})$ and $\xi_n(\mathbf{r})$ be two sequences of random variables, depending on a parameter $\mathbf{r} \in \mathbf{R}$. Suppose that $\lim_n \sup_{\mathbf{r} \in \mathbf{R}} \sup_t |\mathbb{P}_{\mathbf{r}}(X_n(\mathbf{r}) \leq t) - F(t)| = 0$ where F is a continuous cdf, and $\lim_n \inf_{\mathbf{r} \in \mathbf{R}} \mathbb{P}_{\mathbf{r}}(|1 - \xi_n(\mathbf{r})| \leq \tau(n)) = 1$ for $\tau(n) = o(1)$. Assume in addition that F is Lipschitz, i.e. there

exist $\kappa > 0$, such that for any $t, s \in \mathbb{R}$: $|F(t) - F(s)| \leq \kappa|t - s|$. Then we have:

$$\limsup_n \sup_{\mathbf{r} \in \mathbf{R}} \sup_t \left| \mathbb{P}_{\mathbf{r}} \left(\frac{X_n(\mathbf{r})}{\xi_n(\mathbf{r})} \leq t \right) - F(t) \right| = 0.$$

PROOF OF LEMMA E.2. The proof follows by a direct calculation, and we omit the details. \square

Lemma E.3. Under Assumptions A.1 — A.4, we have:

$$(E.7) \quad \lim_{n \rightarrow \infty} \inf_{\beta \in \Omega} \mathbb{P}_{\beta} \left(\left| \widehat{S}(\widehat{\beta}_{\theta}) - S(\beta) \right| \leq r_1(n)r_5(n) + r_2(n)r_3(n) \right) = 1.$$

If in addition $n^{1/2}(r_1(n)r_5(n) + r_2(n)r_3(n)) = o(1)$, we have:

$$(E.8) \quad \lim_{n \rightarrow \infty} \sup_{\beta \in \Omega} \sup_t |\mathbb{P}_{\beta}(\sigma^{-1}n^{1/2}\widehat{S}(\widehat{\beta}_{\theta}) \leq t) - \Phi(t)| = 0.$$

PROOF OF LEMMA E.3. Let \mathcal{G}_i^{β} be the event inside the probability measures in assumptions A.1 — A.4 corresponding to the rate $r_i(n)$ for $i = 1, \dots, 5$. It is clear that $\inf_{\beta \in \Omega} \mathbb{P}_{\beta}(\mathcal{G}^{\beta}) \geq 1 - \sum_{i=1}^5 \sup_{\beta \in \Omega} \mathbb{P}_{\beta}[(\mathcal{G}_i^{\beta})^c] \rightarrow 1$. Next, the proof of (E.7) can be done through the same argument as in the proof of Theorem E.1, but using the uniform convergence assumptions. Note that the bounds (E.3) and (E.4) continue to hold on the event $\mathcal{G}^{\beta} = \mathcal{G}_1^{\beta} \cap \dots \cap \mathcal{G}_5^{\beta}$. Hence by Assumptions A.1 and A.2 the proof of (E.7) is complete.

Next we show (E.8). Let $\kappa(n) = n^{1/2}C^{-1/2}(r_1(n)r_5(n) + r_2(n)r_3(n))$, where we recall the definition of C : $\inf_{\beta \in \Omega} \mathbf{v}^T \Sigma \mathbf{v} \geq C > 0$. Then we have:

$$\begin{aligned} \mathbb{P}_{\beta}(\sigma^{-1}n^{1/2}\widehat{S}(\widehat{\beta}_{\theta}) \leq t) &\leq \mathbb{P}_{\beta}(\sigma^{-1}n^{1/2}\widehat{S}(\widehat{\beta}_{\theta}) \leq t, \mathcal{G}^{\beta}) + \mathbb{P}_{\beta}((\mathcal{G}^{\beta})^c) \\ &\leq \mathbb{P}_{\beta}(\sigma^{-1}n^{1/2}S(\beta) \leq t + \kappa(n)) + \mathbb{P}_{\beta}((\mathcal{G}^{\beta})^c). \end{aligned}$$

The above implies the following inequality:

$$\begin{aligned} &\mathbb{P}_{\beta}(\sigma^{-1}n^{1/2}\widehat{S}(\widehat{\beta}_{\theta}) \leq t) - \Phi(t) \\ &\leq \mathbb{P}_{\beta}(\sigma^{-1}n^{1/2}S(\beta) \leq t + \kappa(n)) - \Phi(t + \kappa(n)) + (\Phi(t + \kappa(n)) - \Phi(t)) + \mathbb{P}_{\beta}((\mathcal{G}^{\beta})^c) \\ &\leq \mathbb{P}_{\beta}(\sigma^{-1}n^{1/2}S(\beta) \leq t + \kappa(n)) - \Phi(t + \kappa(n)) + \frac{\kappa(n)}{\sqrt{2\pi}} + \mathbb{P}_{\beta}((\mathcal{G}^{\beta})^c), \end{aligned}$$

where we took into account the fact that Φ is Lipschitz with constant $\leq \frac{1}{\sqrt{2\pi}}$. Now using Assumption A.3, the fact that $\kappa(n) = o(1)$ and $\mathbb{P}_{\beta}((\mathcal{G}^{\beta})^c) = o(1)$ we conclude that:

$$\limsup_{n \rightarrow \infty} \sup_{\beta \in \Omega} \sup_t \mathbb{P}_{\beta}(\sigma^{-1}n^{1/2}\widehat{S}(\widehat{\beta}_{\theta}) \leq t) - \Phi(t) \leq 0.$$

With a similar argument one can show the converse, namely:

$$\liminf_{n \rightarrow \infty} \inf_{\beta \in \Omega} \inf_t \mathbb{P}_{\beta}(\sigma^{-1}n^{1/2}\widehat{S}(\widehat{\beta}_{\theta}) \leq t) - \Phi(t) \geq 0.$$

This concludes the proof of (E.8). \square

PROOF OF THEOREM A.1. By Assumption A.5, we have:

$$\inf_{\beta} \mathbb{P}_{\beta} \left(\left| 1 - \widehat{\sigma} \sigma^{-1} \right| \leq \frac{r_6(n)}{C} \right) \geq \mathbb{P}_{\beta} \left(|\widehat{\sigma} - \sigma| \leq \frac{r_6(n)}{\sqrt{C}} \right) \geq \inf_{\beta} \mathbb{P}_{\beta} (|\widehat{\sigma}^2 - \sigma^2| \leq r_6(n)) \\ \text{(E.9)} \quad \quad \quad = 1 - o(1),$$

where recall that $\inf_{\beta \in \Omega} \mathbf{v}^T \Sigma \mathbf{v} \geq C > 0$. The last expression implies that:
Next define:

$$\zeta_n(\beta) = \widehat{\sigma} \sigma^{-1} \left(\widehat{\mathbf{v}}^T [\mathbf{T}(\mathbf{Z}, \widetilde{\beta}_{\theta})]_{*1} \right), \quad \eta_n(\beta, \nu) = \frac{\widehat{\sigma} \sigma^{-1}}{2} \widehat{\mathbf{v}}^T \frac{\partial}{\partial \theta} [\mathbf{T}(\mathbf{Z}, \widetilde{\beta}_{\nu})]_{*1} (\widetilde{\theta} - \theta),$$

where $\widetilde{\beta}_{\nu} = \nu \widehat{\beta}_{\widetilde{\theta}} + (1 - \nu) \widehat{\beta}_{\theta}$ and let $\xi_n(\beta, \nu) = \zeta_n(\beta) + \eta_n(\beta, \nu)$.

We will show that

$$\text{(E.10)} \quad \lim_n \inf_{\beta \in \Omega} \inf_{\nu \in [0,1]} \mathbb{P}_{\beta} (|1 - \xi_n(\beta, \nu)| \leq \tau(n)) = 1,$$

for some $\tau(n) = o(1)$, or equivalently — for every $\epsilon > 0$: $\sup_{\beta \in \Omega} \sup_{\nu \in [0,1]} \mathbb{P}_{\beta} (|1 - \xi_n(\beta, \nu)| > \epsilon) = o(1)$. We proceed with the following:

$$\sup_{\beta \in \Omega} \sup_{\nu \in [0,1]} \mathbb{P}_{\beta} (|1 - \xi_n(\beta, \nu)| > \epsilon) \leq \underbrace{\sup_{\beta \in \Omega} \mathbb{P}_{\beta} (|1 - \zeta_n(\beta)| > \epsilon/2)}_{I_1} + \underbrace{\sup_{\beta \in \Omega} \sup_{\nu \in [0,1]} \mathbb{P}_{\beta} (|\eta_n(\beta, \nu)| > \epsilon/2)}_{I_2}.$$

First we tackle I_1 . We have:

$$I_1 \leq \sup_{\beta \in \Omega} \mathbb{P}_{\beta} (|1 - \widehat{\sigma} \sigma^{-1}| > \epsilon/4) + \sup_{\beta \in \Omega} \mathbb{P}_{\beta} (|\widehat{\sigma} \sigma^{-1}| > 2) + \sup_{\beta \in \Omega} \mathbb{P}_{\beta} (|1 - \widehat{\mathbf{v}}^T [\mathbf{T}(\mathbf{Z}, \widetilde{\beta}_{\theta})]_{*1}| > \epsilon/8),$$

and all of the terms on the RHS are $o(1)$ due to (E.9) and Assumption A.4 respectively.

Next we handle I_2 term. Let $E = \sup_{\beta \in \Omega} \mathbb{E}_{\beta} \psi(\mathbf{Z})$, where the function ψ is defined in Assumption A.4. Fix an $\alpha > 0$, and proceed as:

$$I_2 \leq \sup_{\beta \in \Omega} \mathbb{P}_{\beta} (|\widehat{\sigma} \sigma^{-1}| \geq 2) + \sup_{\beta \in \Omega} \mathbb{P}_{\beta} (|\widetilde{\theta} - \theta^*| > E^{-1} \epsilon \alpha / 2) + \sup_{\beta \in \Omega} \mathbb{P}_{\beta} (\psi(\mathbf{Z}) \geq \alpha^{-1} E) \leq o(1) + \alpha,$$

where the last inequality follows from (E.9), Proposition A.1 and Markov's inequality. Taking the limit $n \rightarrow \infty$ shows that $\lim_n \sup_{\beta \in \Omega} \sup_{\nu \in [0,1]} \mathbb{P}_{\beta} (|\eta_n(\beta, \nu)| > \epsilon/2) \leq \alpha$. Taking $\alpha \rightarrow 0$ shows (E.10). Next observe that by (E.5) we have the following identity:

$$\text{(E.11)} \quad \frac{n^{1/2}(\widetilde{\theta} - \theta)}{\widehat{\sigma}} = - \frac{n^{1/2} \widehat{S}(\widehat{\beta}_{\theta})}{\sigma \xi_n(\beta, \nu)}.$$

Now combining the fact that (E.10) with our results from Lemmas E.2 and E.3 in addition to (E.11) completes the proof. \square

PROOF OF PROPOSITION B.1. Note that the only thing left to show is the consistency of the plugin estimate $\widehat{\mathbf{v}}^T \widehat{\Sigma} \widehat{\mathbf{v}}$ for $\mathbf{v}^{*T} \Sigma \mathbf{v}^*$, with the rest of the

argument following from Theorem 2 and Slutsky's theorem. By the triangle inequality we have:

$$|\widehat{\mathbf{v}}^T \widehat{\Sigma} \widehat{\mathbf{v}} - \mathbf{v}^{*T} \Sigma \mathbf{v}^*| \leq \underbrace{\|\widehat{\mathbf{v}}^T - \mathbf{v}^{*T}\|_1 \|\widehat{\Sigma}(\widehat{\mathbf{v}} - \mathbf{v}^*)\|_\infty}_{I_1} + 2 \underbrace{\|\mathbf{v}^{*T} \widehat{\Sigma}\|_\infty \|\widehat{\mathbf{v}} - \mathbf{v}^*\|_1}_{I_2} + \underbrace{\|\mathbf{v}^*\|_1^2 \|\widehat{\Sigma} - \Sigma\|_{\max}}_{I_3}$$

Next we control $|I_1| \leq O_p(r_2(n)^2 r_7(n) + \|\Sigma\|_{\max} r_2^2(n)) = o_p(1)$. Next

$$|I_2| \leq 2O_p(\|\mathbf{v}^*\|_1 r_2(n) r_7(n) + \|\mathbf{v}^{*T} \Sigma\|_\infty r_2(n)) = o_p(1).$$

Finally, for I_3 we have $|I_3| \leq \|\mathbf{v}^*\|_1^2 O_p(r_7(n)) = o_p(1)$. \square

APPENDIX F: PROOFS FOR THE DANTZIG SELECTOR

We recall the definition of *restricted eigenvalue* (RE) assumption (Bickel et al., 2009).

Definition F.1 (RE). We say that the symmetric positive semi-definite matrix $\mathbf{M}_{k \times k}$ possesses the restricted eigenvalue property if:

$$\text{RE}_{\mathbf{M}}(s, \xi) = \min_{S \subset \{1, \dots, k\}, |S| \leq s} \min \left\{ \frac{\mathbf{u}^T \mathbf{M} \mathbf{u}}{\|\mathbf{u}_S\|_2^2} : \mathbf{u} \in \mathbb{R}^d \setminus \{0\}, \|\mathbf{u}_{S^c}\|_1 \leq \xi \|\mathbf{u}_S\|_1 \right\} > 0.$$

Definition F.2. Denote with \mathbf{X} the $n \times d$ matrix whose rows are the \mathbf{X}_i^T vectors stacked together. Let \mathbf{Y} be the an $n \times 1$ vector stacking the observations Y_i for $i = 1, \dots, n$ and let $\boldsymbol{\varepsilon} = \mathbf{Y} - \mathbf{X}\boldsymbol{\beta}^*$.

PROOF OF COROLLARY 1. We will prove this result by validating all of the assumptions required of the general Theorem 2. Regarding Assumption (3.5), observe that from Lemma F.4, we have that $\|\mathbf{v}^* - \widehat{\mathbf{v}}\|_1 = O_p(\|\mathbf{v}^*\|_1 s_{\mathbf{v}} \sqrt{\log d/n})$ and by Lemma F.6 — $\|\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}\|_1 = O_p(s \sqrt{\log d/n})$.

Assumption 1 can be verified as follows. To see (3.2), fix a $|\theta - \theta^*| < \epsilon$, for some $\epsilon > 0$. Next by the triangle inequality:

$$\left\| \Sigma_n \boldsymbol{\beta}_\theta^* - n^{-1} \mathbf{X}^T \mathbf{Y} - \Sigma_{\mathbf{X}}(\boldsymbol{\beta}_\theta^* - \boldsymbol{\beta}^*) \right\|_\infty \leq \|n^{-1} \mathbf{X}^T \boldsymbol{\varepsilon}\|_\infty + \|\Sigma_{n,*1} - \Sigma_{\mathbf{X},*1}\|_\infty \epsilon.$$

The two terms on the RHS are $O_p(\sqrt{\log d/n})$, by the proof of Lemma F.5 (see F.6) and Lemma F.2. The same logic shows that $|\mathbf{v}^{*T} \Sigma_n \boldsymbol{\beta}_\theta^* - n^{-1} \mathbf{v}^{*T} \mathbf{X}^T \mathbf{Y} - \mathbf{v}^{*T} \Sigma_{\mathbf{X},*1}(\theta - \theta^*)| = O_p(\|\mathbf{v}^*\|_1 \sqrt{\log d/n})$, which implies (3.3). Since the Hessian \mathbf{T} in (3.4) is free of $\boldsymbol{\beta}$ we are allowed to set $r_3(n) = \lambda' \asymp \|\mathbf{v}^*\|_1 \sqrt{\log d/n} = o(1)$ (by Lemma F.4). Finally the two expectations in Assumption 1, are bounded as we see below:

$$\|\Sigma_{\mathbf{X}}(\boldsymbol{\beta}_\theta^* - \boldsymbol{\beta}^*)\|_\infty \leq \|\Sigma_{\mathbf{X},*1}\|_\infty \epsilon \leq 2K_{\mathbf{X}}^2 \epsilon, \quad \|\mathbf{v}^{*T} \Sigma_{\mathbf{X},-1}\|_\infty = 0.$$

By adding up the following two identities:

$$\begin{aligned} \sqrt{n} O_p(\|\mathbf{v}^*\|_1 \sqrt{\log d/n}) O_p(s \sqrt{\log d/n}) &= o_p(1), \\ \sqrt{n} O_p(\|\mathbf{v}^*\|_1 s_{\mathbf{v}} \sqrt{\log d/n}) O_p(\sqrt{\log d/n}) &= O_p(\|\mathbf{v}^*\|_1 s_{\mathbf{v}} \log d / \sqrt{n}) = o_p(1), \end{aligned}$$

we get that (3.8) is also valid in this case.

To verify the consistency of $\tilde{\theta}$ we check the assumptions in Theorem 1. Clearly the map $\theta \mapsto \mathbf{v}^{*T} \Sigma_{\mathbf{X}} (\beta_{\theta}^* - \beta^*) = (\theta - \theta^*)$ has a unique 0 when $\theta = \theta^*$. Moreover, the map $\theta \mapsto \hat{\mathbf{v}}^T (\Sigma_n \hat{\beta}_{\theta} - n^{-1} \mathbf{X}^T \mathbf{Y})$ is continuous as it is linear. In addition, it has a unique zero except in cases when $\hat{\mathbf{v}}^T \Sigma_{n,*1} = 0$. However note that $|\hat{\mathbf{v}}^T \Sigma_{n,*1} - 1| \leq \lambda'$ by (4.1), and hence for small enough values of λ' a unique zero will always exist.

Assumption 3 is verified in Lemma F.1. We move on to show (3.7). Clearly (3.7) is trivial as its LHS $\equiv 0$ in this case. Finally Proposition F.1 checks that $\hat{\Delta}$ is a consistent estimate of Δ . This completes the proof. \square

Remark F.1. In fact, under the additional assumption $s_{\mathbf{v}}^3/n = o(1)$, the proof of Corollary 1 implies that the uniform types of assumptions in Section A are satisfied, and hence under the same assumptions as in Corollary 1, we have:

$$\lim_{n \rightarrow \infty} \sup_{\|\beta\|_0 \leq s} \sup_{t \in \mathbb{R}} |\mathbb{P}_{\beta}(\hat{U}_n \leq t) - \Phi(t)| = 0.$$

By Remark A.1 the above equality readily translates from estimator uniform consistency to confidence region uniform consistency. It is noteworthy to mention that the space of uniformity is $\Omega = \{\beta : \|\beta\|_0 \leq s\}$, provided that the conditions of Corollary 1 hold, which includes that the covariates \mathbf{X} satisfy $s_{\mathbf{v}} = \|\mathbf{v}^*\|_0$ and $\max(s_{\mathbf{v}}, s) \|\mathbf{v}^*\|_1 \log d / \sqrt{n} = o(1)$, $\|\mathbf{v}^*\|_1^2 \sqrt{\log d / n} = o(1)$.

Remark F.2. Note here that it is implied that $\lambda' = o(1)$ and hence since $\|\mathbf{v}^*\|_1 \geq 2K_{\mathbf{X}}^{-2}$ it follows that $\lambda = o(1)$ as well.

Remark F.3. Observe that $\|\mathbf{v}^*\|_1 \leq \sqrt{s_{\mathbf{v}}} \|\mathbf{v}^*\|_2 \leq \sqrt{s_{\mathbf{v}}} \delta$. This yields sufficient conditions by substituting $\|\mathbf{v}^*\|_1$ with $\sqrt{s_{\mathbf{v}}}$. Moreover, under the assumption that $\mathbf{v}^{*T} \mathbf{X}$ is sub-Gaussian, we can further relax the requirements on sparsity $s_{\mathbf{v}}$ dimension d and number of observations n .

Lemma F.1. Assume that condition 6 holds and $\max(s_{\mathbf{v}}, s) \|\mathbf{v}^*\|_1 \log d / \sqrt{n} = o(1)$. Then $\Delta^{-1/2} n^{1/2} S(\beta^*) \rightsquigarrow N(0, 1)$.

PROOF OF LEMMA F.1. To show the weak convergence we verify Lyapunov's condition for the CLT. We need to show that $\frac{n^{-2}}{\Delta^2} \sum_{i=1}^n \mathbb{E} |\mathbf{v}^{*T} \mathbf{X}_i (\mathbf{X}_i^T \beta^* - Y_i)|^4$ converges to 0. Note that we have $\Delta^2 \geq \lambda_{\min}(\Sigma_{\mathbf{X}}) \|\mathbf{v}^*\|_2^4 \text{Var}(\varepsilon)^2 = O(1) \|\mathbf{v}^*\|_2^4$. Therefore it suffices to consider the following expression:

$$(F.1) \quad \frac{n^{-2}}{\|\mathbf{v}^*\|_2^4} \sum_{i=1}^n \mathbb{E} |\mathbf{v}^{*T} \mathbf{X}_i (\mathbf{X}_i^T \beta^* - Y_i)|^4 \leq n^{-2} \sum_{i=1}^n \mathbb{E} \|(\mathbf{X}_i \varepsilon_i)_{S_{\mathbf{v}}}\|_2^4 \leq n^{-1} s_{\mathbf{v}}^2 M,$$

where $M = 2^8 (KK_{\mathbf{X}})^4$, and the last inequality holding from Lemma F.9. Using the fact that $\|\mathbf{v}^*\|_1 \geq 2K_{\mathbf{X}}^{-2}$ as seen in Remark F.2, we have that $\max(s_{\mathbf{v}}, s) \|\mathbf{v}^*\|_1 \log d / \sqrt{n} = o(1)$ implies $s_{\mathbf{v}}^2/n = o(1)$ and completes the proof. \square

Remark F.4. Using the Berry-Esseen theorem for non-identical random variables in combination with Lemma F.9 we can further show:

$$\sup_t \left| \mathbb{P}^* \left(\frac{n^{1/2}}{\sqrt{\Delta}} S(\beta^*) \leq t \right) - \Phi(t) \right| \leq C_{BE} (6KK_{\mathbf{X}})^3 n^{-1/2} s_{\mathbf{v}}^3 = o(1),$$

where M and C_{BE} are absolute constants.

Proposition F.1. Under Assumption 6, and the following additional assumption $\|\mathbf{v}^*\|_1^2 \sqrt{\frac{\log d}{n}} = o(1)$, we have that $\hat{\Delta} \rightarrow_p \Delta$.

PROOF OF PROPOSITION F.1. We show that each of the two sums is corresponding to its population counterpart, and then the proof follows upon an application of Slutsky's theorem. We start with the first term:

$$\begin{aligned} \left| \frac{1}{n} \sum_{i=1}^n (\hat{\mathbf{v}}^T \mathbf{X}_i)^2 - \mathbf{v}^{*T} \Sigma_{\mathbf{X}} \mathbf{v}^* \right| &\leq \underbrace{\left| \frac{1}{n} \sum_{i=1}^n [(\hat{\mathbf{v}}^T \mathbf{X}_i)^2 - (\mathbf{v}^{*T} \mathbf{X}_i)^2] \right|}_{I_1} + \underbrace{|\mathbf{v}^{*T} \Sigma_n \mathbf{v}^* - \mathbf{v}^{*T} \Sigma_{\mathbf{X}} \mathbf{v}^*|}_{I_2}, \\ |I_1| &\leq \|\hat{\mathbf{v}} - \mathbf{v}^*\|_1 (\|\Sigma_n \hat{\mathbf{v}}\|_{\infty} + \|\Sigma_n \mathbf{v}^*\|_{\infty}). \end{aligned}$$

We know from Lemma F.4, that $\|\mathbf{v}^* - \hat{\mathbf{v}}\|_1 = O_p(\|\mathbf{v}^*\|_1 s_{\mathbf{v}} \sqrt{\log d/n})$, and by definition $\|\Sigma_n \hat{\mathbf{v}}\|_{\infty} \leq 1 + \lambda'$. In the proof of Lemma F.4 we also show that

$$\|\Sigma_n \mathbf{v}^*\|_{\infty} = 1 + O_p(\|\mathbf{v}^*\|_1 \sqrt{\log d/n}),$$

upon appropriately choosing $\lambda' \asymp \|\mathbf{v}^*\|_1 \sqrt{\log d/n}$, with a large enough proportionality constant. Thus since $O_p(\|\mathbf{v}^*\|_1 s_{\mathbf{v}} \sqrt{\log d/n}) (2 + O_p(\|\mathbf{v}^*\|_1 \sqrt{\log d/n})) = o_p(1)$ we have shown $|I_1| = o_p(1)$. We next tackle $|I_2| \leq \|\mathbf{v}^*\|_1^2 \|\Sigma_n - \Sigma_{\mathbf{X}}\|_{\max}$. Lemma F.2 gives us that $\|\Sigma_n - \Sigma_{\mathbf{X}}\|_{\max} = O_p(\sqrt{\log d/n})$, and thus because of our extra assumption we have $|I_2| = O_p(\|\mathbf{v}^*\|_1^2 \sqrt{\log d/n}) = o_p(1)$.

Now we turn to the second part of the proof:

$$\begin{aligned} &\left| \frac{1}{n} \sum_{i=1}^n (Y_i - \mathbf{X}_i^T \hat{\beta})^2 - \text{Var}(\varepsilon) \right| \\ &\leq \underbrace{\left| \frac{1}{n} \sum_{i=1}^n (Y_i - \mathbf{X}_i^T \hat{\beta})^2 - \frac{1}{n} \sum_{i=1}^n (Y_i - \mathbf{X}_i^T \beta^*)^2 \right|}_{I_3} + \underbrace{\left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i^2 - \text{Var}(\varepsilon) \right|}_{I_4}. \end{aligned}$$

The term I_4 is clearly $o_p(1)$ because of the LLN (ε_i are centered and have finite variance as sub-Gaussian random variables). Thus we are left to deal with I_3 :

$$\begin{aligned} |I_3| &\leq \frac{1}{n} \|\mathbf{X}(\hat{\beta} - \beta^*)\|_2^2 + \frac{2}{n} \sum_{i=1}^n |\mathbf{X}_i^T (\hat{\beta} - \beta^*)| |\varepsilon_i| \\ &\leq \frac{1}{n} \|\mathbf{X}(\hat{\beta} - \beta^*)\|_2^2 + \frac{2}{n} \|\mathbf{X}(\hat{\beta} - \beta^*)\|_2 \sqrt{\sum_{i=1}^n \varepsilon_i^2}, \end{aligned}$$

where $\mathbf{X}_{n \times d}$ is a matrix, with rows \mathbf{X}_i^T stacked together. (F.8) in Lemma F.6 gives us that $\frac{1}{n} \|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2^2 = O_p(\frac{s \log d}{n}) = o_p(1)$, and since by LLN $\sqrt{\frac{1}{n} \sum_{i=1}^n \varepsilon_i^2} = O_p(1)$, we have $|I_3| = o_p(1)$, which shows the consistency of the second estimator and concludes the proof. \square

Lemma F.2. We have that with probability at least $1 - 2d^{2-\bar{c}A_X^2}$:

$$\|\boldsymbol{\Sigma}_n - \boldsymbol{\Sigma}_{\mathbf{X}}\|_{\max} \leq 4A_X K_{\mathbf{X}}^2 \sqrt{\log d/n}.$$

Note. The constant \bar{c} is a universal constant independent of the \mathbf{X} distribution, $K_{\mathbf{X}}$ is as defined in the main text, and $A_X > 0$ is an arbitrarily chosen constant satisfying $A_X \sqrt{\log d/n} \leq 1$.

PROOF OF LEMMA F.2. First we note that the elements of the matrix $-\mathbf{X}\mathbf{X}^T$ are sub-exponential random variables. This fact can be seen since by Cauchy-Schwartz, one can easily obtain that:

$$(F.2) \quad \|X_i X_j\|_{\psi_1} \leq 2\|X_i\|_{\psi_2} \|X_j\|_{\psi_2} \leq 2K_{\mathbf{X}}^2.$$

Next by the triangle inequality it is clear that $\|X_i X_j - \mathbb{E}X_i X_j\|_{\psi_1} \leq 4K_{\mathbf{X}}^2$. The proof is completed by applying a standard Bernstein tail bound (see Proposition 5.16 in Vershynin (2010) e.g.). \square

Lemma F.3. Assume the same conditions as in Lemma F.2, and assume further that the minimum eigenvalue $\lambda_{\min}(\boldsymbol{\Sigma}_{\mathbf{X}}) > 0$ and $s\sqrt{\log d/n} \leq (1 - \kappa) \frac{\lambda_{\min}(\boldsymbol{\Sigma}_{\mathbf{X}})}{(1+\xi)^2 4A_X K_{\mathbf{X}}^2}$, where $0 < \kappa < 1$. We then have that $\boldsymbol{\Sigma}_n$ satisfies the RE property with $\text{RE}_{\boldsymbol{\Sigma}_n}(s, \xi) \geq \kappa \text{RE}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(s, \xi) \geq \kappa \lambda_{\min}(\boldsymbol{\Sigma}_{\mathbf{X}}) > 0$ with probability at least $1 - 2d^{2-\bar{c}A_X^2}$.

PROOF OF LEMMA F.3. The proof follows a standard argument so omit the details. \square

Definition F.3. For a fixed $0 < \kappa < 1$, let $\text{RE}_{\kappa}(s, \xi) = \kappa \text{RE}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(s, \xi)$.

Lemma F.4. Assume that — $\lambda_{\min}(\boldsymbol{\Sigma}_{\mathbf{X}}) > \delta > 0$, $s_{\mathbf{v}}\sqrt{\log d/n} \leq (1 - \kappa) \frac{\lambda_{\min}(\boldsymbol{\Sigma}_{\mathbf{X}})}{(1+1)^2 4A_X K_{\mathbf{X}}^2}$, where $0 < \kappa < 1$ and $\lambda' \geq \|\mathbf{v}^*\|_1 4A_X K_{\mathbf{X}}^2 \sqrt{\log d/n}$. Then we have that $\|\hat{\mathbf{v}} - \mathbf{v}^*\|_1 \leq \frac{8\lambda' s_{\mathbf{v}}}{\text{RE}_{\kappa}(s_{\mathbf{v}}, 1)}$ with probability at least $1 - 2d^{2-\bar{c}A_X^2}$. Additionally we have:

$$(F.3) \quad \|\hat{\mathbf{v}}_{S_{\mathbf{v}}^c} - \mathbf{v}_{S_{\mathbf{v}}^c}^*\|_1 \leq \|\hat{\mathbf{v}}_{S_{\mathbf{v}}} - \mathbf{v}_{S_{\mathbf{v}}}^*\|_1.$$

PROOF OF LEMMA F.4. The proof follows a standard argument so omit the details. \square

Lemma F.5. Assume the same conditions as in Lemma F.2 and that $\sqrt{\log d/n} \leq C$ for some constant C . Let $S = \text{supp}(\boldsymbol{\beta}^*)$, and let $\lambda = AK\sqrt{\frac{\log d}{n}}$. Then, with probability at least $1 - ed^{1-\frac{cA^2}{2(1+2CA_X)K_{\mathbf{X}}^2}} - 2d^{2-\bar{c}A_X^2}$

(where c is a universal constant independent of the distribution of ε , $K = \|\varepsilon\|_{\psi_2}$, and the other constants are defined in Lemma F.2) we have:

$$(F.4) \quad \|\widehat{\beta}_{S^c} - \beta_{S^c}^*\|_1 \leq \|\widehat{\beta}_S - \beta_S^*\|_1, \quad \text{and:}$$

$$(F.5) \quad \left\| \Sigma_n(\beta^* - \widehat{\beta}) \right\|_\infty \leq 2\lambda.$$

PROOF OF LEMMA F.5. Note that by a Hoeffding's type of inequality for sub-Gaussian random variables (see Proposition 5.10 (Vershynin, 2010)) and the union bound, we have:

$$(F.6) \quad \mathbb{P} \left(\left\| \frac{1}{n} \mathbf{X}^T \varepsilon \right\|_\infty \geq t \mid \mathbf{X} \right) \leq ed \exp \left(-\frac{cnt^2}{K^2 \|\Sigma_n\|_{\max}} \right),$$

where c is a universal constant. The remainder of the proof follows by standard arguments so we omit the details. \square

Lemma F.6. Assume the same conditions as in Lemmas F.2, F.3 (with $\xi = 1$), and F.5, so that Σ_n satisfies the RE assumption with $\text{RE}_\kappa(s, 1)$ with high probability. Set $\lambda = AK\sqrt{\frac{\log d}{n}}$, as in Lemma F.5. Then with probability at least $1 - ed^{1 - \frac{cA^2}{2(1+2CA_X)K^2}} - 2d^{2-\bar{c}A_X^2}$ we have:

$$(F.7) \quad \|\widehat{\beta} - \beta^*\|_1 \leq \frac{8AK}{\text{RE}_\kappa(s, 1)} s \sqrt{\log d/n},$$

$$(F.8) \quad \|\mathbf{X}(\widehat{\beta} - \beta^*)\|_2^2 \leq \frac{16A^2K^2}{\text{RE}_\kappa(s, 1)} s \log d.$$

PROOF OF LEMMA F.6. The proof follows by standard arguments and we omit the details. \square

Lemma F.7. Let $\{\mathbf{X}_i\}_{i=1}^n$ are identical (not necessarily independent), d -dimensional sub-Gaussian vectors with $\max_{1 \leq i \leq n, 1 \leq j \leq d} \|X_{ij}\|_{\psi_2} = K$. Then we have:

$$\max_{i=1, \dots, n} \|\mathbf{X}_i \mathbf{X}_i^T\|_{\max} = O_p(\log(nd)).$$

PROOF OF LEMMA F.7. The proof follows after an application of a Bernstein type of tail bound (Vershynin, 2010, see e.g.) and we omit the details. \square

Lemma F.8. Let $X_i, i = 1 \dots k$ are sub-exponential with $\|X_i\|_{\psi_\ell} \leq U$, for some $\ell \geq 1$ and denote by \mathbf{X} the vector with entries X_i . Then for any $p, q \geq 1$ we have:

$$\mathbb{E} \|\mathbf{X}\|_q^p \leq [\mathbb{E} [\|\mathbf{X}\|_q^{pq}]^{1/(qp)}] \leq (pq)^{p/\ell} U^p k^{p/q}$$

PROOF OF LEMMA F.8. We apply Jensen's followed by Minkowski's inequality to obtain the following:

$$[\mathbb{E}[\|\mathbf{X}\|_q^p]]^{1/p} \leq [\mathbb{E}[\|\mathbf{X}\|_q^{pq}]]^{1/(qp)} \leq [k(\mathbb{E}(X_i)^{pq})^{1/q}]^{1/p} \leq (pq)^{1/\ell} U k^{1/p},$$

where the last inequality follows by the definition of ψ_ℓ norm. Raising this inequality to the power of p finishes the proof. \square

Lemma F.9. Let $R \subset \{1, \dots, d\}$ with $|R| = r$. Then we have the following:

$$\mathbb{E}\|(\mathbf{X}\varepsilon)_R\|_2^4 \leq r^2 2^8 (K K_{\mathbf{X}})^4.$$

PROOF OF LEMMA F.9. Simply observe that $\mathbb{E}\|(\mathbf{X}\varepsilon)_R\|_2^4 \leq \sqrt{\mathbb{E}|\varepsilon|^8} \sqrt{\mathbb{E}\|(\mathbf{X})_R\|_2^8}$, and apply Lemma F.8 for ψ_2 . \square

APPENDIX G: PROOFS FOR IVR

Definition G.1 (CS). For the (not necessarily symmetric) matrix $\mathbf{M}_{k \times k}$ we define its coordinate-wise sensitivity with respect to the L_1 norm by:

$$\text{CS}_{\mathbf{M}}(s, \xi) = \min_{S \subset \{1, \dots, k\}, |S| \leq s} \min \{s \|\mathbf{M}\mathbf{u}\|_\infty : \mathbf{u} \in \mathbb{R}^d \setminus \{0\}, \|\mathbf{u}_{S^c}\|_1 \leq \xi \|\mathbf{u}_S\|_1, \|\mathbf{u}_S\|_1 = 1\} > 0.$$

This definition is inspired by [Gautier and Tsybakov \(2011\)](#).

Definition G.2. Denote with \mathbf{X} and \mathbf{W} the $n \times d$ matrices whose rows are the \mathbf{X}_i^T and \mathbf{W}_i^T vectors stacked together respectively. Let \mathbf{Y} be the an $n \times 1$ vector stacking the observations Y_i for $i = 1, \dots, n$ and let $\varepsilon = \mathbf{Y} - \mathbf{X}\beta^*$.

PROOF OF COROLLARY C.1. The proof is the same as the proof of Corollary 1 upon usages of the Lemmas developed in this section. We omit the details. \square

Remark G.1. In fact, the proof of Corollary C.1 implies that the uniform types of assumptions in Section A are satisfied, and hence under the same assumptions as in Corollary C.1, we have:

$$\lim_{n \rightarrow \infty} \sup_{\|\beta\|_0 \leq s} \sup_{t \in \mathbb{R}} |\mathbb{P}_\beta(\hat{U}_n \leq t) - \Phi(t)| = 0.$$

Lemma G.1. Assume that condition 6 holds and $\max(s_{\mathbf{v}}, s)(\|\mathbf{v}^*\|_1 \vee 1) \log d / \sqrt{n} = o(1)$. Then:

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

PROOF OF LEMMA G.1. The proof is the same as that of Lemma F.1 after using the Lemmas developed below. We omit the details. \square

Remark G.2. Using the Berry-Esseen theorem for non-identical random variables in combination with Lemma F.9 we can further show:

$$\sup_t \left| \mathbb{P}^* \left(\frac{n^{1/2}}{\sqrt{\Delta}} S(\beta^*) \leq t \right) - \Phi(t) \right| \leq C_{BE} (6K K_{\mathbf{W}\mathbf{X}})^3 n^{-1/2} s_{\mathbf{v}}^{3/2} = o(1),$$

where M and C_{BE} are absolute constants.

Proposition G.1. Under assumption 6, and the following additional assumption:

$$\|\mathbf{v}^*\|_1^2 \sqrt{\log d/n} = o(1),$$

we have that $\hat{\Delta} \rightarrow_p \Delta$.

PROOF OF PROPOSITION G.1. By the triangle inequality, for any two vectors \mathbf{a} and \mathbf{b} we have $|\|\mathbf{a}\|_2 - \|\mathbf{b}\|_2| \leq \|\mathbf{a} - \mathbf{b}\|_2$. Making multiple usages of this inequality one realizes that it suffices to show:

$$\begin{aligned} n^{-1} \sum_{i=1}^n ((\mathbf{v}^* - \hat{\mathbf{v}})^T \mathbf{W}_i)^2 ((\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}})^T \mathbf{X}_i)^2 &= o_p(1), n^{-1} \sum_{i=1}^n (\mathbf{v}^{*T} \mathbf{W}_i)^2 ((\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}})^T \mathbf{X}_i)^2 = o_p(1), \\ n^{-1} \sum_{i=1}^n ((\mathbf{v}^* - \hat{\mathbf{v}})^T \mathbf{W}_i)^2 \varepsilon_i^2 &= o_p(1), n^{-1} \sum_{i=1}^n (\mathbf{v}^{*T} \mathbf{W}_i)^2 \varepsilon_i^2 - \mathbb{E}(\mathbf{v}^{*T} \mathbf{W}_i)^2 \varepsilon_i^2 = o_p(1), \end{aligned}$$

and $\mathbb{E}(\mathbf{v}^{*T} \mathbf{W}_i)^2 \varepsilon_i^2 < \infty$. We show these convergences in turn. For the first term we have:

$$n^{-1} \sum_{i=1}^n ((\mathbf{v}^* - \hat{\mathbf{v}})^T \mathbf{W}_i)^2 ((\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}})^T \mathbf{X}_i)^2 \leq \|\mathbf{v}^* - \hat{\mathbf{v}}\|_1^2 \|\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}\|_1^2 O_p(\log(nd)^2) = o_p(1),$$

with high probability, where we used Lemma F.7 and Lemmas G.6 and G.4. For the second term:

$$\begin{aligned} n^{-1} \sum_{i=1}^n (\mathbf{v}^{*T} \mathbf{W}_i)^2 ((\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}})^T \mathbf{X}_i)^2 &\leq \|\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}\|_1^2 \max_{i=1, \dots, n} \|\mathbf{X}_i\|_\infty^2 n^{-1} \sum_{i=1}^n (\mathbf{v}^{*T} \mathbf{W}_i)^2 \\ &\leq O_p\left(\frac{s^2 \log d \log(nd)}{n}\right) n^{-1} \sum_{i=1}^n (\mathbf{v}^{*T} \mathbf{W}_i)^2, \end{aligned}$$

with high probability. By Lemma G.2 we have:

$$n^{-1} \sum_{i=1}^n (\mathbf{v}^{*T} \mathbf{W}_i)^2 \leq \underbrace{\|\mathbf{v}^*\|_1^2 \left\| n^{-1} \sum_{i=1}^n \mathbf{W}_i \mathbf{W}_i^T - \Sigma_{\mathbf{W}\mathbf{W}} \right\|_{\max}}_{O_p(\sqrt{\log d/n})} + \underbrace{\mathbf{v}^{*T} \Sigma_{\mathbf{W}\mathbf{W}} \mathbf{v}^*}_{\leq D_{\max}} = O_p(1),$$

which combined with the bound in the previous display completes the proof for the second term. The third term bound follows upon noticing:

$$\begin{aligned} n^{-1} \sum_{i=1}^n ((\mathbf{v}^* - \hat{\mathbf{v}})^T \mathbf{W}_i)^2 \varepsilon_i^2 &\leq \|\hat{\mathbf{v}} - \mathbf{v}^*\|_1^2 \max \|\mathbf{W}_i\|_\infty^2 \underbrace{n^{-1} \sum_{i=1}^n \varepsilon_i^2}_{O_p(1)} \leq O_p\left(\frac{s^2 \log d \log(nd)}{n}\right) O_p(1) \\ &= o_p(1). \end{aligned}$$

Moving to the last term we have:

$$\left| n^{-1} \sum_{i=1}^n (\mathbf{v}^{*T} \mathbf{W}_i)^2 \varepsilon_i^2 - \mathbb{E}(\mathbf{v}^{*T} \mathbf{W}_i)^2 \varepsilon_i^2 \right| \leq \|\mathbf{v}^*\|_1^2 \left\| n^{-1} \sum_{i=1}^n \mathbf{W}_{i, S_{\mathbf{v}}} \mathbf{W}_{i, S_{\mathbf{v}}}^T \varepsilon_i^2 - \Sigma_{\mathbf{W}\mathbf{W}, S_{\mathbf{v}} S_{\mathbf{v}}} \sigma^2 \right\|_\infty,$$

where $S_{\mathbf{v}} = \text{supp}(\mathbf{v}^*)$. The final concentration is handled in Lemma G.7. Applying this lemma in conjunction with the union bound gives us the existence of a constant $C_{K\mathbf{W}\mathbf{X}}$ depending on K and $K_{\mathbf{W}\mathbf{X}}$ such that:

$$\mathbb{P}(\|n^{-1} \sum_{i=1}^n \mathbf{W}_{i,S_{\mathbf{v}}} \mathbf{W}_{i,S_{\mathbf{v}}}^T \varepsilon_i^2 - \Sigma_{\mathbf{W}\mathbf{W},S_{\mathbf{v}}S_{\mathbf{v}}} \sigma^2\|_{\infty} \geq t) \leq s_{\mathbf{v}}^2 \frac{C_{K\mathbf{W}\mathbf{X}}^k [\sqrt{k/n} + k^2/n]^k}{t^k},$$

for all $k \in \mathbb{N}$. Selecting $t = 2e^4 C_{K\mathbf{W}\mathbf{X}} \sqrt{\frac{\log d}{n}}$, $k = \lceil \min(\log d, (n \log d)^{1/4}) \rceil$ brings the above bound of the order $O(s_{\mathbf{v}}^2 \exp(-4 \lceil \min(\log d, (n \log d)^{1/4}) \rceil)) = o(1)$, and shows that:

$$\left| n^{-1} \sum_{i=1}^n (\mathbf{v}^{*T} \mathbf{W}_i)^2 \varepsilon_i^2 - \mathbb{E}(\mathbf{v}^{*T} \mathbf{W}_i)^2 \varepsilon_i^2 \right| \leq O\left(\|\mathbf{v}^*\|_1^2 \sqrt{\frac{\log d}{n}}\right) = o(1).$$

We conclude the proof with noticing that $\mathbb{E}[\mathbb{E}[(\mathbf{v}^{*T} \mathbf{W}_i)^2 \varepsilon_i^2 | \mathbf{W}]] = \sigma^2 \mathbb{E}[\mathbf{v}^{*T} \mathbf{W}_i \mathbf{W}_i^T \mathbf{v}^*] < \infty$.

□

Lemma G.2. We have that with probability at least $1 - 2d^{2-\bar{c}A_{\mathbf{X}}^2}$:

$$\left\| \frac{1}{n} \sum_{i=1}^n [\mathbf{W}_i^T, \mathbf{X}_i^T]^T [\mathbf{W}_i^T, \mathbf{X}_i^T] - \Sigma \right\|_{\max} = \|\Sigma_n - \Sigma\|_{\max} \leq 4A_{\mathbf{W}\mathbf{X}} K_{\mathbf{W}\mathbf{X}}^2 \sqrt{\log d/n}.$$

Note. The constant \bar{c} is a universal constant independent of the \mathbf{X} and \mathbf{W} distributions, $K_{\mathbf{W}\mathbf{X}}$ is as defined in the main text, and $A_{\mathbf{W}\mathbf{X}} > 0$ is an arbitrarily chosen constant satisfying $A_{\mathbf{W}\mathbf{X}} \sqrt{\log d/n} \leq 1$.

PROOF OF LEMMA G.2. Proof is follows by the same argument as in Lemma F.2, so we omit the details. □

Lemma G.3. Assume the same conditions as in Lemma G.2, and assume further that the matrix $\Sigma_{\mathbf{W}\mathbf{X}}$ satisfies $\text{CS}_{\Sigma_{\mathbf{W}\mathbf{X}}}(s, \xi) > \kappa^*$ and that s is sufficiently small so that $s\sqrt{\log d/n} \leq (1 - \kappa) \frac{\kappa^*}{(1+\xi)4A_{\mathbf{W}\mathbf{X}}K_{\mathbf{W}\mathbf{X}}^2}$, where $0 < \kappa < 1$. We then have that $\Sigma_{\mathbf{W}\mathbf{X},n}$ satisfies the CS property with $\text{CS}_{\Sigma_{\mathbf{W}\mathbf{X},n}}(s, \xi) \geq \kappa \text{CS}_{\Sigma_{\mathbf{W}\mathbf{X}}}(s, \xi) > 0$ with probability at least $1 - 2d^{2-\bar{c}A_{\mathbf{W}\mathbf{X}}^2}$.

PROOF OF LEMMA G.3. This proof is simply using Definition G.1 and Lemma G.2. □

Definition G.3. For a fixed $0 < \kappa < 1$, let $\text{CS}_{\kappa}(s, \xi) = \kappa \text{CS}_{\Sigma_{\mathbf{W}\mathbf{X}}}(s, \xi)$.

Lemma G.4. Assume that $\text{CS}_{\Sigma_{\mathbf{W}\mathbf{X}}}(s_{\mathbf{v}}, 1) \geq \kappa^* > 0$, and that further $s_{\mathbf{v}}$ is small enough so that $s_{\mathbf{v}}\sqrt{\log d/n} \leq (1 - \kappa) \frac{\kappa^*}{(1+1)4A_{\mathbf{W}\mathbf{X}}K_{\mathbf{W}\mathbf{X}}^2}$, where $0 < \kappa < 1$ and $\lambda' \geq \|\mathbf{v}^*\|_1 4A_{\mathbf{W}\mathbf{X}}K_{\mathbf{W}\mathbf{X}}^2 \sqrt{\frac{\log d}{n}}$. Then we have that $\|\hat{\mathbf{v}} - \mathbf{v}^*\|_1 \leq \frac{8\lambda's_{\mathbf{v}}}{\text{CS}_{\Sigma_{\mathbf{W}\mathbf{X}}}(s_{\mathbf{v}}, 1)}$ with probability at least $1 - 2d^{2-\bar{c}A_{\mathbf{W}\mathbf{X}}^2}$.

PROOF OF LEMMA G.4. Using a standard argument and Lemma G.2 we can show that with probability at least $1 - 2d^{2-2\bar{c}A_{WX}^2}$, \mathbf{v}^* satisfies (C.1) and consequently

$$(G.1) \quad \left\| \frac{1}{n} \sum_{i=1}^n (\hat{\mathbf{v}} - \mathbf{v}^*)^T \mathbf{W}_i \mathbf{X}_i^T \right\|_\infty \leq \left\| \frac{1}{n} \sum_{i=1}^n \hat{\mathbf{v}}^T \mathbf{W}_i \mathbf{X}_i^T - \mathbf{e}_1 \right\|_\infty + \left\| \frac{1}{n} \sum_{i=1}^n \mathbf{v}^{*T} \mathbf{W}_i \mathbf{X}_i^T - \mathbf{e}_1 \right\|_\infty \leq 2\lambda'.$$

Let $S_{\mathbf{v}} = \text{supp}(\mathbf{v}^*)$, with $s_{\mathbf{v}} = |S_{\mathbf{v}}|$. Using the formulation of program (C.1) it is not hard to show that:

$$(G.2) \quad \|\hat{\mathbf{v}}_{S_{\mathbf{v}}^c} - \mathbf{v}_{S_{\mathbf{v}}^c}^*\|_1 \leq \|\hat{\mathbf{v}}_{S_{\mathbf{v}}} - \mathbf{v}_{S_{\mathbf{v}}}^*\|_1.$$

By Lemma G.3, $\Sigma_{\mathbf{W}\mathbf{X},n}$ satisfies the CS assumption under our conditions and hence

$$\frac{\|\hat{\mathbf{v}}_{S_{\mathbf{v}}} - \mathbf{v}_{S_{\mathbf{v}}}^*\|_1 \text{CS}_{\Sigma_{\mathbf{W}\mathbf{X}}}(s_{\mathbf{v}}, 1)}{s_{\mathbf{v}}} \leq \|\Sigma_{\mathbf{W}\mathbf{X},n}(\hat{\mathbf{v}} - \mathbf{v}^*)\|_\infty \stackrel{\text{by (G.1)}}{\leq} 2\lambda'$$

Hence by (G.2) we conclude $\|\hat{\mathbf{v}} - \mathbf{v}^*\|_1 \leq 4s_{\mathbf{v}}\lambda' / \text{CS}_{\Sigma_{\mathbf{W}\mathbf{X}}}(s_{\mathbf{v}}, 1)$, which is what we wanted to show. \square

Lemma G.5. Assume the same conditions as in Lemma G.2 and that $\sqrt{\log d/n} \leq C$ for some constant C . Let $S = \text{supp}(\beta^*)$, and let $\lambda = AK\sqrt{\frac{\log d}{n}}$. Then, with probability at least $1 - 2d^{1-(cA^2)/(4K_{\mathbf{X}\mathbf{X}}^2)}$ (where c is a universal constant Lemma G.2) we have:

$$(G.3) \quad \|\hat{\beta}_{S^c} - \beta_{S^c}^*\|_1 \leq \|\hat{\beta}_S - \beta_S^*\|_1,$$

and:

$$(G.4) \quad \left\| \Sigma_{\mathbf{W}\mathbf{X},n}(\beta^* - \hat{\beta}) \right\|_\infty \leq 2\lambda.$$

PROOF OF LEMMA G.5. Note that by a Bernstein type of inequality for sub-exponential random variables (see Proposition 5.16 (Vershynin, 2010)) and the union bound, we have:

$$(G.5) \quad \mathbb{P} \left(\left\| \frac{1}{n} \mathbf{W}^T \boldsymbol{\varepsilon} \right\|_\infty \geq t \right) \leq 2d \exp \left(-c \min \left(\frac{nt^2}{4K^2K_{\mathbf{W}\mathbf{X}}^2}, \frac{nt}{2KK_{\mathbf{W}\mathbf{X}}} \right) \right),$$

where c is a universal constant, and we used the fact that $\max_{i \in \{1, \dots, d\}} \|\varepsilon W_i\|_{\psi_1} \leq 2KK_{\mathbf{Z}\mathbf{X}}$. Set $t = \lambda$. Provided that $A \leq 2KK_{\mathbf{W}\mathbf{X}}$ the above probability bounds gives us that the event $E := \left\{ \left\| \frac{1}{n} \mathbf{W}^T \boldsymbol{\varepsilon} \right\|_\infty \leq \lambda \right\}$ holds with probability at least $1 - 2d^{1-(cA^2)/(4K_{\mathbf{X}\mathbf{X}}^2)}$.

Note that when E holds, the true parameter satisfies the Dantzig selector constraint and thus we can obtain (G.3) in the same manner as in Lemma G.4. Using the triangle inequality on E shows (G.4). \square

Lemma G.6. Assume the same conditions in Lemmas G.2, G.3 (with $\xi = 1$), and G.5, so that $\Sigma_{\mathbf{W}\mathbf{X},n}$ satisfies the CS assumption with $\text{CS}_\kappa(s, 1)$ with high probability. Set $\lambda = AK\sqrt{\frac{\log d}{n}}$, as in Lemma G.5. Then with probability at least $1 - 2d^{1-(cA^2)/(4K_{\mathbf{X}\mathbf{X}}^2)}$ we have:

$$(G.6) \quad \|\hat{\beta} - \beta^*\|_1 \leq \frac{4AK}{\text{CS}_\kappa(s, 1)} s \sqrt{\log d/n}.$$

PROOF. Recall from Lemma G.5, that on the event E we have that (G.3) and (G.4) hold, and furthermore $\frac{1}{n}\mathbf{W}^T\mathbf{W} = \Sigma_{\mathbf{W}\mathbf{X},n}$ satisfies the CS condition. Thus on the event E , we have:

$$\frac{\|\hat{\beta}_S - \beta_S^*\|_1 \text{CS}_\kappa(s, 1)}{s} \leq \|\Sigma_{\mathbf{W}\mathbf{X},n}(\beta^* - \hat{\beta})\|_\infty \leq 2\lambda.$$

To get (G.6), note that by (G.3) we have: $\|\hat{\beta} - \beta^*\|_1 \leq 2\|\hat{\beta}_S - \beta_S^*\|_1 \leq \frac{4AK\lambda s}{\text{CS}_\kappa(s, 1)}$, and we are done. \square

Lemma G.7. Let $\{X_i\}_{i=1}^n$ be an i.i.d. collection of sub-Gaussian random variables satisfying $\|X_i\|_{\psi_2} \leq K$. Then we have:

$$\mathbb{P}\left(|n^{-1} \sum_{i=1}^n X_i^4 - \mathbb{E}X_i^4| \geq t\right) \leq \frac{\tilde{K}_4^k [\sqrt{k/n} + k^2/n]^k}{t^k},$$

for any $k \in \mathbb{N}$ and some fixed constant \tilde{K}_4 depending solely on K .

PROOF. We make usage of Theorem 1.4 of Adamczak and Wolff (2015), which provides a convenient concentration bound for higher moments of sub-Gaussian random variables. Using this result it is not hard to check that: $[\mathbb{E}|n^{-1} \sum_{i=1}^n X_i^4 - \mathbb{E}X_i^4|^{1/k}] \leq \tilde{K}_4[\sqrt{k/n} + k^2/n]$, where \tilde{K}_4 is a constant depending solely on K . Consequently, applying Markov's inequality we obtain the final conclusion. \square

APPENDIX H: PROOFS FOR GRAPHICAL MODELS

H.1 Proofs for Graphical Models with CLIME

PROOF OF COROLLARY 2. Before we proceed with the proof note that we are guaranteed to have $\|\mathbf{v}^*\|_1 \geq (\Sigma_{\mathbf{X}}^{-1})_{11} \geq (\Sigma_{\mathbf{X},11})^{-1} \geq (2K_{\mathbf{X}}^2)^{-1} > 0$, and similarly $\|\beta^*\|_1 \geq (2K_{\mathbf{X}}^2)^{-1} > 0$. Hence, $\max(s_{\mathbf{v}}\|\mathbf{v}^*\|_1, s\|\beta^*\|_1)\|\mathbf{v}^*\|_1\|\beta^*\|_1 \log d \log(nd)/n = o(1)$ implies that

$$\max(s_{\mathbf{v}}, s)\|\mathbf{v}^*\|_1\|\beta^*\|_1 \log d/\sqrt{n} = o(1).$$

We show this result by verifying the conditions of Section 3. To see Assumption (3.5), we can use Lemma F.4 to argue that $\|\hat{\beta} - \beta^*\|_1 = O_p\left(\|\beta^*\|_1 s \sqrt{\log d/n}\right)$, $\|\hat{\mathbf{v}} - \mathbf{v}^*\|_1 = O_p\left(\|\mathbf{v}^*\|_1 s_{\mathbf{v}} \sqrt{\log d/n}\right)$ provided that λ and λ' are large enough.

Next we check Assumption 1. To see (3.2), fix a $|\theta - \theta^*| < \epsilon$, for some $\epsilon > 0$. By the triangle inequality:

$$\|\Sigma_n \beta_\theta^* - \Sigma_X \beta_\theta^*\|_\infty \leq \|\Sigma_n - \Sigma_X\|_{\max} (\|\beta\|_1^* + \epsilon).$$

The RHS is $O_p\left(\|\beta\|_1^* \sqrt{\log d/n}\right)$, by Lemma F.2. The same logic shows that $\left|\mathbf{v}^{*T} \Sigma_n \beta_\theta^* - \mathbf{v}^{*T} \Sigma_X \beta_\theta^*\right| = O_p\left((\|\beta^*\|_1 + \epsilon) \|\mathbf{v}^*\|_1 \sqrt{\log d/n}\right)$, which implies (3.3). Since the Hessian \mathbf{T} in (3.4) is free of β we are allowed to set $r_3(n) = \lambda' \asymp \|\mathbf{v}^*\|_1 \sqrt{\log d/n} = o(1)$ (by Lemma F.4). Finally the two expectations in Assumption 1, are bounded as we see below:

$$\|\Sigma_X \beta_\theta^* - \mathbf{e}_m^T\|_\infty = \|\Sigma_X (\beta_\theta^* - \beta^*)\|_\infty \leq \|\Sigma_{X,*1}\|_\infty \epsilon \leq 2K_X^2 \epsilon, \quad \|\mathbf{v}^{*T} \Sigma_{X,-1}\|_\infty = 0.$$

By adding up the following two identities:

$$\begin{aligned} \sqrt{n} O_p\left(\|\mathbf{v}^*\|_1 \sqrt{\log d/n}\right) O_p\left(\|\beta^*\|_1 s \sqrt{\log d/n}\right) &= o_p(1), \\ \sqrt{n} O_p\left(\|\mathbf{v}^*\|_1 s_{\mathbf{v}} \sqrt{\log d/n}\right) O_p\left(\|\beta^*\|_1 \sqrt{\log d/n}\right) &= o_p(1), \end{aligned}$$

we get that (3.8) is also valid in this case after a usage of (4.6).

To verify the consistency of $\tilde{\theta}$ we check the assumptions in Theorem 1. Clearly the map $\mathbf{v}^{*T} \Sigma_X (\beta_\theta^* - \beta^*) = (\theta - \theta^*)$ has a unique 0 when $\theta = \theta^*$. Moreover, the map $\theta \mapsto \hat{\mathbf{v}}^T (\Sigma_n \hat{\beta}_\theta - \mathbf{e}_m^T)$ is continuous as it is linear. In addition, it has a unique zero except in cases when $\hat{\mathbf{v}}^T \Sigma_{n,*1} = 0$. However note that $|\hat{\mathbf{v}}^T \Sigma_{n,*1} - 1| \leq \lambda'$ by (4.4), and hence for small enough values of λ' there will exist a unique zero.

Assumption 3 is verified in Lemma H.1. Observe that (3.7) is trivial as its LHS $\equiv 0$ in this case. Finally, the fact that $\hat{\Delta}$ is consistent for Δ is verified in Lemma H.2. \square

Next, we proceed to formulate a uniform weak convergence result. To this end, for fixed $M > \delta > 0$, define the following parameter space of covariance matrices:

$$\mathcal{S}(L, s) = \{\Sigma : \Sigma = \Sigma^T, 0 < \delta \leq \lambda_{\min}(\Sigma), \|\Sigma\|_{\max} \leq M, \|\Sigma^{-1}\|_1 \leq L, \max_{j=1,\dots,d} \|\Sigma_{*j}^{-1}\|_0 \leq s\}.$$

We have the following result in terms of uniform convergence:

Corollary H.1. Let (4.3) holds, and $\text{Cov}(\mathbf{X}) = \Sigma_X \in \mathcal{S}(L, s)$. Let $\Omega = (\Sigma_X)^{-1}$ and denote $\beta = \Omega_{*m}$, and $\mathbf{v} = \Omega_{*1}$. Assume there exist two constants V_{\min} and V_{\max} such that:

$$(H.1) \quad \text{Var}(\mathbf{v}^T \mathbf{X} \mathbf{X}^T \beta) \geq V_{\min} > 0, \quad \mathbb{E}(\mathbf{v}^T \mathbf{X} \mathbf{X}^T \beta)^4 \leq V_{\max} < \infty.$$

Then under the following conditions:

$$(H.2) \quad sL^3 \log(d) \log(nd)/n = o(1), \quad s^3/\sqrt{n} = o(1),$$

we have $\lim_{n \rightarrow \infty} \sup_{\Sigma_X \in \mathcal{S}(L, s)} \sup_{t \in \mathbb{R}} |\mathbb{P}_\beta(\hat{U}_n \leq t) - \Phi(t)| = 0$.

The conditions in this Corollary are essentially the same as those in Corollary 2. Condition $\mathbb{E}(\mathbf{v}^T \mathbf{X} \mathbf{X}^T \boldsymbol{\beta})^4 \leq V_{\max}$ ensures $\mathbb{E}(\mathbf{v}^T \mathbf{X} \mathbf{X}^T \boldsymbol{\beta})^2 < \infty$ and $\text{Var}((\mathbf{v}^T \mathbf{X} \mathbf{X}^T \boldsymbol{\beta})^2) = o(n)$. The only other difference is that the second condition in (H.2) is stronger than the counterpart in Corollary 2. We need this condition to apply the Berry-Esseen theorem to control the normal approximation error uniformly.

Lemma H.1. Under the assumptions of Corollary 2 we have that: $\Delta^{-1/2} n^{1/2} S(\boldsymbol{\beta}^*) \rightsquigarrow N(0, 1)$.

PROOF OF LEMMA H.1. Similarly to Lemma F.1 we will verify Lyapunov's condition for the CLT. It suffices to bound the quantity for some $k > 2$:

$$\frac{n^{-k/2}}{\|\mathbf{v}^*\|_2^k \|\boldsymbol{\beta}^*\|_2^k \alpha_{\min}^{k/2}} \sum_{i=1}^n \mathbb{E} |\mathbf{v}^{*T} \mathbf{X}_i \mathbf{X}_i^T \boldsymbol{\beta}^* - \mathbf{v}^{*T} \boldsymbol{\Sigma}_{\mathbf{X}} \boldsymbol{\beta}^*|^k,$$

where we used assumption Assumption 7. By Cauchy-Schwartz we can bound the above expression by following (up to a constant factor): $n^{-k/2} \sum_{i=1}^n \mathbb{E} \|(\mathbf{X}_i \mathbf{X}_i^T - \boldsymbol{\Sigma}_{\mathbf{X}})_{S_{\mathbf{v}}, S}\|^k$ where by subscripting the matrix we mean setting all elements not in the supports of \mathbf{v}^* or $\boldsymbol{\beta}^*$ ($S_{\mathbf{v}}$, and S correspondingly) to 0, and $\|\cdot\|_F$ is the Frobenius norm of the matrix. Finally using Lemma H.3 we conclude that we can control the expression above by $n^{-k/2+1} (s_{\mathbf{v}} s)^{k/2} (8k K_{\mathbf{X}})^k$, and hence the conclusion follows. \square

Remark H.1. Using the Berry-Esseen theorem for non-identical random variables in combination with the bound we derived above, we can further show:

$$\sup_t |\mathbb{P}^* (\Delta^{-1/2} n^{1/2} S(\boldsymbol{\beta}^*) \leq t) - \Phi(t)| \leq C_{BE} n^{-1/2} (24 K_{\mathbf{X}})^3 (s_{\mathbf{v}} s)^{3/2} = o(1),$$

where C_{BE} is an absolute constant.

Lemma H.2. Under the assumptions from Corollary 2, we have that the plugin estimator $\hat{\Delta} \xrightarrow{P} \Delta$.

PROOF OF LEMMA H.2. Note that $\hat{\Delta} = \frac{1}{n} \sum_{i=1}^n (\hat{\mathbf{v}}^T \mathbf{X}_i \mathbf{X}_i^T \hat{\boldsymbol{\beta}})^2 - (\hat{\mathbf{v}}^T \boldsymbol{\Sigma}_n \hat{\boldsymbol{\beta}})^2$. Similarly to the analysis of the first term in Proposition F.1 one can show that $\hat{\mathbf{v}}^T \boldsymbol{\Sigma}_n \hat{\boldsymbol{\beta}}$ is consistent for $\mathbf{v}^{*T} \boldsymbol{\Sigma}_{\mathbf{X}} \boldsymbol{\beta}^*$ under our assumptions. Hence it suffices to show that $|\frac{1}{n} \sum_{i=1}^n (\hat{\mathbf{v}}^T \mathbf{X}_i \mathbf{X}_i^T \hat{\boldsymbol{\beta}})^2 - \mathbb{E}(\mathbf{v}^{*T} \mathbf{X} \mathbf{X}^T \boldsymbol{\beta}^*)^2| = o_p(1)$. We start by arguing that $n^{-1} \underbrace{\left[\sum_{i=1}^n (\hat{\mathbf{v}}^T \mathbf{X}_i \mathbf{X}_i^T \hat{\boldsymbol{\beta}} - \mathbf{v}^{*T} \mathbf{X}_i \mathbf{X}_i^T \boldsymbol{\beta}^*)^2 \right]}_I$, is asymptotically negligible.

$$I^{1/2} \leq \underbrace{\left[n^{-1} \sum_{i=1}^n (\hat{\mathbf{v}}^T \mathbf{X}_i \mathbf{X}_i^T (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*))^2 \right]^{1/2}}_{I_1} + \underbrace{\left[n^{-1} \sum_{i=1}^n ((\hat{\mathbf{v}} - \mathbf{v}^*)^T \mathbf{X}_i \mathbf{X}_i^T \boldsymbol{\beta}^*)^2 \right]^{1/2}}_{I_2}.$$

We first handle $I_1^2 = (\hat{\beta} - \beta^*)^T n^{-1} \underbrace{\sum_{i=1}^n \mathbf{X}_i \mathbf{X}_i^T \hat{\mathbf{v}} \hat{\mathbf{v}}^T \mathbf{X}_i \mathbf{X}_i^T}_{\mathbf{M}} (\hat{\beta} - \beta^*)$. Using

Lemma F.7, we can bound \mathbf{M} in the following way:

(H.3)

$$\|\mathbf{M}\|_{\max} \leq \max \|\mathbf{X}_i \mathbf{X}_i^T\|_{\max} n^{-1} \sum_{i=1}^n \hat{\mathbf{v}}^T \mathbf{X}_i \mathbf{X}_i^T \hat{\mathbf{v}} \leq O_p(\log(nd)) \|\hat{\mathbf{v}}\|_1 \|\hat{\Sigma} \hat{\mathbf{v}}\|_{\infty}.$$

By the definition of $\hat{\mathbf{v}}$ we have: $\|\hat{\Sigma} \hat{\mathbf{v}}\|_{\infty} \leq (1 + \lambda')$. Hence:

$$\|\mathbf{M}\|_{\max} \leq O_p(\log(nd)) (\|\mathbf{v}^*\|_1 + \|\hat{\mathbf{v}} - \mathbf{v}^*\|_1) (1 + \lambda') = O_p(\log(nd)) \|\mathbf{v}^*\|_1,$$

where we used that $\lambda' \asymp \|\mathbf{v}^*\|_1 \sqrt{\log d/n}$ and $\|\hat{\mathbf{v}} - \mathbf{v}^*\|_1 = O_p(\|\mathbf{v}^*\|_1 s_{\mathbf{v}} \sqrt{\log d/n})$ which are quantities going to 0 under our assumptions. Thus:

$$I_1^2 \leq \|\hat{\beta} - \beta^*\|_1^2 \|\mathbf{v}^*\|_1 O_p(\log(nd)) = O_p(\|\mathbf{v}^*\|_1 \|\beta^*\|_1^2 s^2 \log d/n \log(nd)) = o_p(1),$$

By a similar argument we can show that I_2 is of similar order. Putting everything together we conclude:

(H.4)

$$I = O_p(\max(s_{\mathbf{v}}^2 \|\mathbf{v}^*\|_1, s^2 \|\beta^*\|_1) \|\mathbf{v}^*\|_1 \|\beta^*\|_1 \log d \log(nd)/n) = o_p(1).$$

Next, we argue that $n^{-1} \sum_{i=1}^n (\mathbf{v}^{*T} \mathbf{X}_i \mathbf{X}_i^T \beta^*)^2 - \mathbb{E}(\mathbf{v}^{*T} \mathbf{X} \mathbf{X}^T \beta^*)^2$ is small. Recall that we are assuming $\text{Var}((\mathbf{v}^{*T} \mathbf{X} \mathbf{X}^T \beta^*)^2) = o(n)$. A usage of Chebyshev's inequality shows that $n^{-1} \sum_{i=1}^n (\mathbf{v}^{*T} \mathbf{X}_i \mathbf{X}_i^T \beta^*)^2 - \mathbb{E}(\mathbf{v}^{*T} \mathbf{X} \mathbf{X}^T \beta^*)^2 = o_p(1)$. Finally note that by the triangle inequality the following two inequalities hold:

$$\begin{aligned} \left[n^{-1} \sum_{i=1}^n (\hat{\mathbf{v}}^T \mathbf{X}_i \mathbf{X}_i^T \hat{\beta})^2 \right]^{1/2} &\leq \left[n^{-1} \sum_{i=1}^n (\mathbf{v}^{*T} \mathbf{X}_i \mathbf{X}_i^T \beta^*)^2 \right]^{1/2} + I^{1/2}, \\ \left[n^{-1} \sum_{i=1}^n (\mathbf{v}^{*T} \mathbf{X}_i \mathbf{X}_i^T \beta^*)^2 \right]^{1/2} &\leq \left[n^{-1} \sum_{i=1}^n (\hat{\mathbf{v}}^T \mathbf{X}_i \mathbf{X}_i^T \hat{\beta})^2 \right]^{1/2} + I^{1/2}. \end{aligned}$$

Observe that $n^{-1} \sum_{i=1}^n (\mathbf{v}^{*T} \mathbf{X}_i \mathbf{X}_i^T \beta^*)^2 = \mathbb{E}(\mathbf{v}^{*T} \mathbf{X} \mathbf{X}^T \beta^*)^2 + o_p(1) = O_p(1)$. This completes the proof. \square

PROOF OF COROLLARY H.1. To prove this corollary note that all bounds we showed in the proof of Corollary 2 hold uniformly in the parameter set $\mathcal{S}(L, s)$. Note that as both \mathbf{v} and β are columns of Ω we have that $\|\mathbf{v}\|_0, \|\beta\|_0 \leq s$, $\|\mathbf{v}\|_1, \|\beta\|_1 \leq L$ and $M^{-1} \leq \|\mathbf{v}\|_2, \|\beta\|_2 \leq \delta^{-1}$. These conditions in conjunction with the assumptions of the present result, can be seen to imply the conditions from Section A and this completes the proof. \square

Lemma H.3. Let $R_{\mathbf{v}}, R \subset \{1, \dots, d\}$ with $|R_{\mathbf{v}}| = r_{\mathbf{v}}, |R| = r$. Then we have the following $\mathbb{E} \|(\mathbf{X} \mathbf{X}^T - \Sigma)_{R_{\mathbf{v}}, R}\|_F^k \leq (r_{\mathbf{v}} r)^{k/2} (8k K_{\mathbf{X}}^2)^k$.

PROOF. Apply (F.2) to get $\|X_i X_j\|_{\psi_1} \leq 2K_{\mathbf{X}}^2$. Combined with the triangle inequality it gives us $\|X_i X_j - \sigma_{ij}\|_{\psi_1} \leq 4K_{\mathbf{X}}^2$. Next simply use Lemma F.8 for ψ_1 to complete the proof. \square

H.2 Proofs for Transelliptical Graphical Models

In the transelliptical case the lemmas from the CLIME case are no longer applicable, as the estimator of Σ is constructed in a completely different manner. Furthermore, the vector \mathbf{X} is coming from a nonparanormal family and need not be sub-Gaussian. Fortunately, [Liu et al. \(2012a\)](#) provide a concentration result which we state below:

Theorem H.1 (Liu 2012). For any $n > 1$ with probability at least $1 - 1/d$, we have

$$(H.5) \quad \|\hat{\mathbf{S}}^\tau - \Sigma\|_{\max} \leq 2.45\pi\sqrt{\log d/n}.$$

While this theorem is defined within the framework of nonparanormal models, the proof doesn't utilize the fact that the family is nonparanormal, and thus extends to the transelliptical case. As we can see from the theorem, the rate of Kendall's tau estimate (H.5), is no different than the one using the sample covariance matrix, provided in Lemma F.2.

PROOF OF COROLLARY D.1. The proof of this result is the same as Corollary 2, except we use Σ in place of $\Sigma_{\mathbf{X}}$, $\hat{\mathbf{S}}^\tau$ in place of Σ_n , Lemma H.6 in place of Lemma F.4 and Theorem H.1 in place of Lemma F.2. The only different step is the verification of Assumption 3 which we provide in Lemma H.4. We omit the rest of the details. \square

Lemma H.4. Under the assumptions of Corollary D.1 we have that $\Delta^{-1/2}n^{1/2}S(\beta^*) \rightsquigarrow N(0, 1)$, where Δ is defined as in (D.3).

PROOF OF LEMMA H.4. Note that by the mean value theorem we have the following representation:

$$\begin{aligned} n^{1/2}\mathbf{v}^{*T} \left(\hat{\mathbf{S}}^\tau \beta^* - \mathbf{e}_m^T \right) &= n^{1/2}\mathbf{v}^{*T} \left(\hat{\mathbf{S}}^\tau - \Sigma \right) \beta^* = n^{1/2} \sum_{\substack{j \in S_{\mathbf{v}}, k \in S \\ j \neq k}} v_j^* \beta_k^* \left(\sin \left(\hat{\tau}_{jk} \frac{\pi}{2} \right) - \sin \left(\tau_{jk} \frac{\pi}{2} \right) \right) \\ &= n^{1/2} \sum_{\substack{j \in S_{\mathbf{v}}, k \in S \\ j \neq k}} v_j^* \beta_k^* \cos \left(\tau_{jk} \frac{\pi}{2} \right) \frac{\pi}{2} (\hat{\tau}_{jk} - \tau_{jk}) \\ &\quad - \frac{n^{1/2}}{2} \sum_{\substack{j \in S_{\mathbf{v}}, k \in S \\ j \neq k}} v_j^* \beta_k^* \sin \left(\tilde{\tau}_{jk} \frac{\pi}{2} \right) \left(\frac{\pi}{2} (\hat{\tau}_{jk} - \tau_{jk}) \right)^2, \end{aligned}$$

where $\tilde{\tau}_{ij}$ is a number between $\hat{\tau}_{ij}$ and τ_{ij} . We will first deal with the first term in the sum above. Since this term is a linear combination of second order (dependent) U -statistics, we will make usage of Hájek's projection method. A similar approach was used in the celebrated paper of [Hoeffding \(1948\)](#). To this end we define the following notations:

$$\begin{aligned} \tau_{jk}^{ii'} &= \text{sign}((X_{ij} - X_{i'j})(X_{ik} - X_{i'k})) - \tau_{jk}, & \tau_{jk}^{ii'|i} &= \mathbb{E}[\tau_{jk}^{ii'} | \mathbf{X}_i], \\ \tau_{jk}^i &= \frac{1}{n-1} \sum_{i' \neq i} \tau_{jk}^{ii'|i}, & w_{jk}^{ii'} &= \tau_{jk}^{ii'} - \tau_{jk}^{ii'|i} - \tau_{jk}^{ii'|i'}, \end{aligned}$$

In terms of these notations we therefore have $\widehat{\tau}_{jk} - \tau_{jk} = \frac{2}{n} \sum_{i=1}^n \tau_{jk}^i + \frac{2}{n(n-1)} \sum_{1 \leq i < i' \leq n} w_{jk}^{ii'}$. This gives us the following identity:

$$\begin{aligned} n^{1/2} \sum_{\substack{j \in S_{\mathbf{v}}, k \in S \\ j \neq k}} v_j^* \beta_k^* \cos\left(\tau_{jk} \frac{\pi}{2}\right) \frac{\pi}{2} (\widehat{\tau}_{jk} - \tau_{jk}) &= \underbrace{\pi n^{-1/2} \sum_{\substack{j \in S_{\mathbf{v}}, k \in S \\ j \neq k}} v_j^* \beta_k^* \cos\left(\tau_{jk} \frac{\pi}{2}\right) \sum_{i=1}^n \tau_{jk}^i}_{I_1} \\ &+ \underbrace{\frac{\pi}{n^{1/2}(n-1)} \sum_{\substack{j \in S_{\mathbf{v}}, k \in S \\ j \neq k}} v_j^* \beta_k^* \cos\left(\tau_{jk} \frac{\pi}{2}\right) \sum_{1 \leq i < i' \leq n} w_{jk}^{ii'}}_{I_2}. \end{aligned}$$

We first deal with I_1 which can clearly be represented as a sum of iid mean 0 terms, by verifying Lyapunov's condition for the CLT. I_1 can be rewritten as:

$$(H.6) \quad I_1 = n^{-1/2} \sum_{i=1}^n \underbrace{\sum_{\substack{j \in S_{\mathbf{v}}, k \in S \\ j \neq k}} v_j^* \beta_k^* \pi \cos\left(\tau_{jk} \frac{\pi}{2}\right) \tau_{jk}^i}_{M_i}.$$

Construct the matrix $\Theta^i \in \mathbb{R}^{d \times d}$ given entrywise by

$$(H.7) \quad \Theta_{jk}^i = \pi \cos\left(\tau_{jk} \frac{\pi}{2}\right) \tau_{jk}^i, \text{ hence } \Theta_{jj}^i = 0.$$

We can then rewrite $M_i = \mathbf{v}^{*T} \Theta^i \boldsymbol{\beta}^* = \mathbf{v}_{S_{\mathbf{v}}, S}^{*T} \Theta_{S_{\mathbf{v}}, S}^i \boldsymbol{\beta}_S^*$. Calculating the variance of M_i gives $\text{Var}(M_i) = \mathbb{E}(\mathbf{v}^{*T} \Theta^i \boldsymbol{\beta}^*)^2 \geq \alpha_{\min} \|\mathbf{v}^*\|_2^2 \|\boldsymbol{\beta}^*\|_2^2$, where the inequality follows by assumption. We proceed to verify Lyapunov's condition for some $k > 2$ (where we ignore the constant $\alpha_{\min} > 0$):

$$\frac{n^{-k/2}}{\|\mathbf{v}^*\|_2^k \|\boldsymbol{\beta}^*\|_2^k} \sum_{i=1}^n \mathbb{E}|M_i|^k \leq n^{-k/2} \sum_{i=1}^n \mathbb{E}\|\Theta_{S_{\mathbf{v}}, S}^i\|_F^k \leq \frac{(s_{\mathbf{v}} s)^{k/2} (2\pi)^k}{n^{k/2-1}} = o(1),$$

where the first inequality follows from Cauchy-Schwartz, and to see the second one notice that each element of Θ^i is bounded $|\Theta_{jk}^i| \leq 2\pi$, and hence $\|\Theta_{S_{\mathbf{v}}, S}^i\|_F^k \leq (s_{\mathbf{v}} s)^{k/2} (2\pi)^k$. This implies that $I_1 \rightsquigarrow N(0, \Delta)$, with $\Delta = \mathbb{E}(\mathbf{v}^{*T} \Theta^i \boldsymbol{\beta}^*)^2$.

Next we deal with the second term I_2 , which is mean 0, by showing that its (standardized) variance goes to 0 asymptotically. Before we compute its variance we make several preliminary calculations:

$$\begin{aligned} \mathbb{E}(w_{jk}^{ii'} w_{ls}^{rr'}) &= \mathbb{E}(\tau_{jk}^{ii'} \tau_{ls}^{rr'}) - \mathbb{E}(\tau_{jk}^{ii'} \tau_{ls}^{rr'|r}) - \mathbb{E}(\tau_{jk}^{ii'} \tau_{ls}^{rr'|r'}) - \mathbb{E}(\tau_{jk}^{ii'|i} \tau_{ls}^{rr'}) + \mathbb{E}(\tau_{jk}^{ii'|i} \tau_{ls}^{rr'|r}) \\ &+ \mathbb{E}(\tau_{jk}^{ii'|i} \tau_{ls}^{rr'|r'}) - \mathbb{E}(\tau_{jk}^{ii'|i'} \tau_{ls}^{rr'}) + \mathbb{E}(\tau_{jk}^{ii'|i'} \tau_{ls}^{rr'|r}) + \mathbb{E}(\tau_{jk}^{ii'|i'} \tau_{ls}^{rr'|r'}). \end{aligned}$$

In the expression above we have taken $j \neq k, l \neq s, r \neq i \neq i' \neq r \neq r' \neq i$. Notice now that all elements above are independent and since $\mathbb{E}(w_{jk}^{ii'}) =$

$\mathbb{E}(w_{ls}^{rr'}) = 0$, we conclude that $\mathbb{E}(w_{jk}^{ii'} w_{ls}^{rr'}) = 0$. Following, the same logic, for $j \neq k$, $l \neq s$, $i \neq i' \neq r' \neq i$:

$$\mathbb{E}(w_{jk}^{ii'} w_{ls}^{ir'}) = \mathbb{E}(\tau_{jk}^{ii'} \tau_{ls}^{ir'}) - \mathbb{E}(\tau_{jk}^{ii'} \tau_{ls}^{ir'|i}) - \mathbb{E}(\tau_{jk}^{ii'|i} \tau_{ls}^{ir'}) + \mathbb{E}(\tau_{jk}^{ii'|i} \tau_{ls}^{ir'|i})$$

where all the rest terms are 0, by the same argument as in the first case. Using iterated expectation by conditioning on \mathbf{X}_i it can be easily seen that all terms become equal to $-\mathbb{E}(\tau_{jk}^{ii'|i} \tau_{jk}^{ir'|i})$, and we can conclude that $\mathbb{E}(w_{jk}^{ii'} w_{ls}^{ir'}) = 0$. Since $\mathbb{E}I_2 = 0$, we have:

$$\begin{aligned} \frac{\text{Var}(I_2)}{\text{Var}(M_i)} &\leq \frac{\mathbb{E}(I_2^2)}{\alpha_{\min} \|\mathbf{v}^*\|_2^2 \|\boldsymbol{\beta}^*\|_2^2} \\ &= \frac{\pi^2}{\alpha_{\min} \|\mathbf{v}^*\|_2^2 \|\boldsymbol{\beta}^*\|_2^2 n(n-1)^2} \sum_{1 \leq i < i' \leq n} \mathbb{E} \left(\sum_{\substack{j \in S_{\mathbf{v}}, k \in S \\ j \neq k}} v_j^* \beta_k^* \cos \left(\tau_{jk} \frac{\pi}{2} \right) w_{jk}^{ii'} \right)^2 \\ &\leq \frac{\pi^2 \binom{n}{2} 36 \left(\sum_{j \in S_{\mathbf{v}}} |v_j^*| \right)^2 \left(\sum_{k \in S} |\beta_k^*| \right)^2}{n(n-1)^2 \alpha_{\min} \|\mathbf{v}^*\|_2^2 \|\boldsymbol{\beta}^*\|_2^2} \leq \frac{\pi^2 18 s_{\mathbf{v}} s}{(n-1) \alpha_{\min}} = o(1), \end{aligned}$$

where in the next to last inequality we used the trivial bound $|w_{jk}^{ii'}| \leq |\tau_{jk}^{ii'}| + |\tau_{jk}^{ii'|i}| + |\tau_{jk}^{ii'|i'}| \leq 6$. Thus the term $\frac{\text{Var}(I_2)}{\text{Var}(M_i)} = o(1)$ and therefore, Chebyshev's inequality gives us that $\frac{I_2}{\sqrt{\text{Var}(M_i)}} = o_p(1)$.

Finally we deal with the standardized version of the last term:

$$(H.8) \quad \frac{1}{\sqrt{\text{Var}(\mathbf{v}^{*T} \boldsymbol{\Theta} \boldsymbol{\beta}^*)}} \frac{n^{1/2}}{2} \sum_{\substack{j \in S_{\mathbf{v}}, k \in S \\ j \neq k}} v_j^* \beta_k^* \sin \left(\tilde{\tau}_{jk} \frac{\pi}{2} \right) \left(\frac{\pi}{2} (\hat{\tau}_{jk} - \tau_{jk}) \right)^2.$$

As we mentioned previously it's clear that $\hat{\tau}_{jk}$ is a U -statistic, and its kernel is a bounded function (between -1 and 1). Furthermore, we have that $\mathbb{E} \hat{\tau}_{jk} = \tau_{jk}$. Thus, we can apply Hoeffding's inequality for U -statistics (see [Hoeffding \(1963\)](#) equation (5.7)), to obtain that:

$$(H.9) \quad \mathbb{P}(\sup_{jk} |\hat{\tau}_{jk} - \tau_{jk}| > t) \leq 2d^2 \exp \left(-\frac{nt^2}{4} \right).$$

It follows that selecting $t = 9\sqrt{\log d/n}$ suffices to keep the probability going to 0. Notice that the (H.8) can be controlled by:

$$\frac{n^{1/2} \pi^2 \sqrt{s_{\mathbf{v}} s}}{8\theta_{\min} \|\mathbf{v}^*\|_2 \|\boldsymbol{\beta}^*\|_2} \sup_{jk} (\hat{\tau}_{jk} - \tau_{jk})^2 = O_p \left(\frac{\sqrt{s_{\mathbf{v}} s} \log d}{n^{1/2}} \right) = o_p(1).$$

The last equation is implied by our assumption. This concludes the proof. \square

Remark H.2. Using the Berry-Esseen theorem for non-identical random variables we can strengthen weak convergence statement to $\sup_t \left| \mathbb{P}^* \left(\frac{I_1}{\sqrt{\Delta}} \leq t \right) - \Phi(t) \right| \leq C_{BE} n^{-1/2} (s_{\mathbf{v}} s)^{3/2} = o(1)$, where C_{BE} is an absolute constant. Note that we decomposed our test into $\frac{I_1}{\sqrt{\Delta}} + o_p(1)$, and hence this statement is valid more generally for Corollary D.1.

Proposition H.1. Under the same assumptions as in Corollary D.1, we have that $\hat{\Delta} \rightarrow_p \Delta$.

PROOF OF PROPOSITION H.1. Before we go to the main proof, recall the definition of Θ^i (H.7), where $\Theta_{jk}^i = \pi \cos\left(\tau_{jk} \frac{\pi}{2}\right) \tau_{jk}^i$. Note that in fact $\mathbb{E}\Theta^i = 0$, since $\mathbb{E}\tau_{jk}^i = 0$, and thus $\text{Var}(\mathbf{v}^{*T} \Theta^i \beta^*) = \mathbb{E}(\mathbf{v}^{*T} \Theta^i \beta^*)^2$. Similarly one can note the simple identity: $\frac{1}{n} \sum_{i=1}^n \hat{\Theta}^i = 0$. Thus we will in fact focus on showing that $\frac{1}{n} \sum_{i=1}^n (\hat{\mathbf{v}}^T \hat{\Theta}^i \hat{\beta})^2$ is consistent for $\mathbb{E}(\mathbf{v}^{*T} \Theta^i \beta^*)^2$.

Consider the following decomposition:

$$\frac{1}{n} \sum_{i=1}^n (\hat{\mathbf{v}}^T \hat{\Theta}^i \hat{\beta})^2 = \underbrace{\frac{1}{n} \sum_{i=1}^n [(\hat{\mathbf{v}}^T \hat{\Theta}^i \hat{\beta})^2 - (\mathbf{v}^{*T} \hat{\Theta}^i \hat{\beta})^2]}_{I_1} + \underbrace{\frac{1}{n} \sum_{i=1}^n [(\mathbf{v}^{*T} \hat{\Theta}^i \hat{\beta})^2 - (\mathbf{v}^{*T} \Theta^i \beta^*)^2]}_{I_2}.$$

Below we show that I_1 is asymptotically negligible. We have:

$$|I_1| = \left| (\hat{\mathbf{v}} - \mathbf{v}^*)^T \frac{1}{n} \sum_{i=1}^n \hat{\Theta}^i \hat{\beta} \hat{\beta}^T \hat{\Theta}^i (\hat{\mathbf{v}} + \mathbf{v}^*)^T \right| \leq \|\hat{\mathbf{v}} - \mathbf{v}^*\|_1 \|\hat{\mathbf{v}} + \mathbf{v}^*\|_1 \|\hat{\beta}\|_1^2 (2\pi)^2,$$

where we made use of $\|\hat{\Theta}^i\|_{\max} \leq 2\pi$. Thus by Lemma H.6

$$|I_1| = O_p \left(\|\mathbf{v}^*\|_1^2 \|\beta^*\|_1^2 s_{\mathbf{v}} \sqrt{\log d/n} \right) = o_p(1),$$

by assumption. Similarly we obtain $|I_2| = O_p \left(\|\mathbf{v}^*\|_1^2 \|\beta^*\|_1^2 s \sqrt{\log d/n} \right) = o_p(1)$. Next, we inspect the following difference $I_3 = \frac{1}{n} \sum_{i=1}^n [(\mathbf{v}^{*T} \hat{\Theta}^i \beta^*)^2 - (\mathbf{v}^{*T} \Theta^i \beta^*)^2]$. Before we bound this term recall that we have the following useful inequality $\|\Theta^i\|_{\max} \leq 2\pi$. Thus:

$$(H.10) \quad |I_3| \leq \|\mathbf{v}^*\|_1^2 \|\beta^*\|_1^2 4\pi \max_{i=1, \dots, n} \|\hat{\Theta}^i - \Theta^i\|_{\max}.$$

To bound the difference $\max_{i=1, \dots, n} \|\hat{\Theta}^i - \Theta^i\|_{\max}$ we will use some concentration inequalities. First, since \cos is Lipschitz with constant 1 — $|\cos(\frac{\pi}{2} \hat{\tau}_{jk}) - \cos(\frac{\pi}{2} \tau_{jk})| \leq \frac{\pi}{2} |\hat{\tau}_{jk} - \tau_{jk}|$, we have:

$$\begin{aligned} |\hat{\Theta}_{jk}^i - \Theta_{jk}^i| &\leq \pi \left| \cos\left(\frac{\pi}{2} \hat{\tau}_{jk}^i\right) - \cos\left(\frac{\pi}{2} \tau_{jk}^i\right) \right| |\hat{\tau}_{jk}^i| + \pi \left| \cos\left(\frac{\pi}{2} \tau_{jk}^i\right) \right| |\hat{\tau}_{jk}^i - \tau_{jk}^i| \\ &\leq \pi^2 |\hat{\tau}_{jk}^i - \tau_{jk}^i| + \pi |\hat{\tau}_{jk}^i - \tau_{jk}^i|. \end{aligned}$$

where we used the simple observation that $|\widehat{\tau}_{jk}^i| \leq 2$. Next we have:

$$\begin{aligned} & |\widehat{\tau}_{jk}^i - \tau_{jk}^i| \leq |\widehat{\tau}_{jk} - \tau_{jk}| \\ & + \left| \underbrace{\frac{1}{n-1} \sum_{i' \neq i} \text{sign}((X_{ij} - X_{i'j})(X_{ik} - X_{i'k}))}_{\widehat{\theta}_{jk}^i} - \underbrace{\mathbb{E}[\text{sign}((X_{ij} - X_{i'j})(X_{ik} - X_{i'k})) | \mathbf{X}_i]}_{\theta_{jk}^i} \right| \end{aligned}$$

This gives us:

$$(H.11) \quad |\widehat{\Theta}_{jk}^i - \Theta_{jk}^i| \leq (\pi^2 + \pi) |\widehat{\tau}_{jk} - \tau_{jk}| + \pi |\widehat{\theta}_{jk}^i - \theta_{jk}^i|.$$

Next, note since the terms in $\widehat{\theta}_{jk}^i$ are iid conditional on \mathbf{X}_i , and they are in the set $\{-1, 1\}$ by Hoeffding's inequality, integrating \mathbf{X}_i out and the union bound we obtain:

$$\mathbb{P}(\max_{i,j,k} |\widehat{\theta}_{jk}^i - \theta_{jk}^i| > t) \leq 2nd^2 \exp\left(-\frac{(n-1)t^2}{2}\right).$$

This implies that selecting $t = 4\sqrt{\log(nd)/n}$, would keep the probability converging to 0. Combining this result with (H.9) and (H.11) gives us $\max_i \|\widehat{\Theta}^i - \Theta^i\|_{\max} = O_p\left(\sqrt{\log(nd)/n}\right)$. Hence using (H.10), we get $|I_3| =$

$\|\mathbf{v}^*\|_1^2 \|\boldsymbol{\beta}^*\|_1^2 O_p\left(\sqrt{\log(nd)/n}\right) = o_p(1)$, which follows since $\|\mathbf{v}^*\|_1^2 \|\boldsymbol{\beta}^*\|_1^2 \max(s_{\mathbf{v}}, s) \sqrt{\log d/n} = o(1)$, implies $\|\mathbf{v}^*\|_1^2 \|\boldsymbol{\beta}^*\|_1^2 \sqrt{\log(nd)/n} = o(1)$. To see the last implication it is sufficient to observe that $\|\mathbf{v}^*\|_1 \geq \Omega_{11} \geq 1$, $\|\boldsymbol{\beta}^*\|_1 \geq \Omega_{mm} \geq 1$. Finally we assess the difference $\frac{1}{n} \sum_{i=1}^n (\mathbf{v}^{*T} \Theta^i \boldsymbol{\beta}^*)^2 - \mathbb{E}(\mathbf{v}^{*T} \Theta^i \boldsymbol{\beta}^*)^2$. By Markov's inequality in much the same way as in the last part of the proof of Proposition H.2, we can show that the expression above is $o_p(1)$ if $\text{Var}((\mathbf{v}^{*T} \Theta \boldsymbol{\beta}^*)^2) = o(n)$. Since the elements of Θ are bounded by 2π we have $|\mathbf{v}^{*T} \Theta \boldsymbol{\beta}^*| \leq \|\mathbf{v}^*\|_1 \|\boldsymbol{\beta}^*\|_1 2\pi$. Hence since $\|\mathbf{v}^*\|_1^2 \|\boldsymbol{\beta}^*\|_1^2 \max(s_{\mathbf{v}}, s) \sqrt{\log d/n} = o(1)$, implies $\|\mathbf{v}^*\|_1^4 \|\boldsymbol{\beta}^*\|_1^4 = o(n)$ the proof is complete. \square

PROOF OF COROLLARY D.2. Similarly to the proof of Corollary H.1 we simply need to note that our conditions imply the conditions required by Corollary D.1 and also note that the bounds in the proofs hold uniformly. \square

Lemma H.5. Assume that the minimum eigenvalue $\lambda_{\min}(\boldsymbol{\Sigma}) > 0$ and $s\sqrt{\log d/n} \leq (1 - \kappa) \frac{\lambda_{\min}(\boldsymbol{\Sigma}_{\mathbf{X}})}{(1+\xi)^{2.45\pi}}$, where $0 < \kappa < 1$. We then have that $\widehat{\mathbf{S}}^\tau$ satisfies the RE property with $\text{RE}_{\widehat{\mathbf{S}}^\tau}(s, \xi) \geq \kappa \lambda_{\min}(\boldsymbol{\Sigma})$ with probability at least $1 - 1/d$.

PROOF OF LEMMA H.5. Proof is the same as in Lemma F.3, but we use Theorem H.1 instead of Lemma F.2. We omit the details. \square

Definition H.1. Define $\text{RE}_\kappa(s, \xi) := \kappa \text{RE}_{\boldsymbol{\Sigma}}(s, \xi) \geq \kappa \lambda_{\min}(\boldsymbol{\Sigma})$.

Lemma H.6. Assume that $-\lambda_{\min}(\Sigma) > 0$, $s_{\mathbf{v}}\sqrt{\log d/n} \leq (1-\kappa)\frac{\lambda_{\min}(\Sigma)}{(1+1)^{2.45\pi}}$, where $0 < \kappa < 1$ and $\lambda' \geq \|\mathbf{v}^*\|_1 2.45\pi\sqrt{\log d/n}$. Then we have that $\|\hat{\mathbf{v}} - \mathbf{v}^*\|_1 \leq \frac{8\lambda's_{\mathbf{v}}}{\text{RE}_{\kappa}(s_{\mathbf{v}},1)}$ with probability at least $1 - 1/d$.

PROOF OF LEMMA H.6. Proof is the same as in Lemma F.4, but we use Theorem H.1 instead of Lemma F.2 and we use Lemma H.5 instead of Lemma F.3. We omit the details. \square

APPENDIX I: PROOFS FOR THE LINEAR DISCRIMINANT ANALYSIS

Lemma I.1. Under Assumption 8 we have that $\alpha V_1 + (1-\alpha)V_2 \geq V'_{\min}(\|\beta^*\|_2^2 \|\mathbf{v}^*\|_2^2 + \|\mathbf{v}\|_2^2)$.

PROOF OF LEMMA I.1. The proof follows by an elementary calculation so we omit the details. \square

Remark I.1. This also shows that $\Delta \geq \delta(\alpha^{-1} + (1-\alpha)^{-1})\|\mathbf{v}^*\|_2^2 \geq \delta(\alpha^{-1} + (1-\alpha)^{-1})4K_U^{-4} > 0$.

PROOF OF COROLLARY 3. We verify the conditions of Section 3. To see Assumption (3.5), we can use Lemmas I.5 and I.6 to get $\|\hat{\beta} - \beta^*\|_1 = O_p\left((\|\beta^*\|_1 \vee 1)s\sqrt{\log d/n}\right)$, $\|\hat{\mathbf{v}} - \mathbf{v}^*\|_1 = O_p\left(\|\mathbf{v}^*\|_1 s_{\mathbf{v}}\sqrt{\log d/n}\right)$ provided that λ and λ' are large enough, and we used the fact that $n_1 \asymp n_2$.

Next we check Assumption 1. To see (3.2), fix a $|\theta - \theta^*| < \epsilon$, for some $\epsilon > 0$. By the triangle inequality:

$$\left\| \hat{\Sigma}_n \beta_{\theta}^* - \Sigma \beta_{\theta}^* + (\bar{\mathbf{X}} - \bar{\mathbf{Y}}) - \boldsymbol{\mu}_1 + \boldsymbol{\mu}_2 \right\|_{\infty} \leq \|\hat{\Sigma}_n - \Sigma\|_{\max}(\|\beta^*\|_1 + \epsilon) + \|\bar{\mathbf{X}} - \boldsymbol{\mu}_1\|_{\infty} + \|\bar{\mathbf{Y}} - \boldsymbol{\mu}_2\|_{\infty}.$$

The RHS is $O_p\left((\|\beta^*\|_1 \vee 1)\sqrt{\log d/n}\right)$, by Lemma I.3 and bound (I.3). The same logic shows that $r_2(n) \asymp \|\mathbf{v}^*\|_1(\|\beta^*\|_1 \vee 1)\sqrt{\log d/n}$, which implies (3.3). Since the Hessian \mathbf{T} in (3.4) is free of β we are allowed to set $r_3(n) = \lambda' \asymp \|\mathbf{v}^*\|_1\sqrt{\log d/n} = o(1)$ (by Lemma I.6). Finally the two expectations in Assumption 1, are bounded as we see below:

$$\|\Sigma \beta_{\theta}^* - \Sigma \beta^*\|_{\infty} = \|\Sigma(\beta_{\theta}^* - \beta^*)\|_{\infty} \leq \|\Sigma_{*1}\|_{\infty} \epsilon \leq 2K_U^2 \epsilon, \quad \|\mathbf{v}^{*T} \Sigma_{-1}\|_{\infty} = 0.$$

By adding up the following two identities:

$$\begin{aligned} \sqrt{n} O_p\left(\|\mathbf{v}^*\|_1 \sqrt{\log d/n}\right) O_p\left((\|\beta^*\|_1 \vee 1)s\sqrt{\log d/n}\right) &= o_p(1), \\ \sqrt{n} O_p\left(\|\mathbf{v}^*\|_1 s_{\mathbf{v}} \sqrt{\log d/n}\right) O_p\left((\|\beta^*\|_1 \vee 1)\sqrt{\log d/n}\right) &= o_p(1), \end{aligned}$$

we get that (3.8) is also valid in this case by assumption.

To verify the consistency of $\tilde{\theta}$ we check the assumptions in Theorem 1. Clearly the map $\mathbf{v}^{*T} \Sigma(\beta_{\theta}^* - \beta^*) = (\theta - \theta^*)$ has a unique 0 when $\theta = \theta^*$.

Moreover, the map $\theta \mapsto \hat{\mathbf{v}}^T(\Sigma_n \hat{\boldsymbol{\beta}}_\theta - (\bar{\mathbf{X}} - \bar{\mathbf{Y}}))$ is continuous as it is linear. In addition, it has a unique zero except in cases when $\hat{\mathbf{v}}^T \Sigma_{n,*1} = 0$. However note that $|\hat{\mathbf{v}}^T \Sigma_{n,*1} - 1| \leq \lambda'$ by (4.4), and hence for small enough values of λ' there will exist a unique zero.

Next we verify Assumption 3 in Lemma I.2. Finally we move on to show (3.7). Observe that (3.7) is trivial as its LHS $\equiv 0$ in this case. \square

Lemma I.2. Under the conditions of Corollary 3 we have the following $\Delta^{-1/2} \sqrt{n} S(\boldsymbol{\beta}^*) \rightsquigarrow N(0, 1)$, where Δ is defined as in (4.7).

PROOF OF LEMMA I.2. We start by defining the following quantity:

$$(I.1) \quad \tilde{\Sigma}_n = \frac{1}{n} \left[\sum_{i=1}^{n_1} (\mathbf{X}_i - \boldsymbol{\mu}_1)(\mathbf{X}_i - \boldsymbol{\mu}_1)^T + \sum_{i=1}^{n_2} (\mathbf{Y}_i - \boldsymbol{\mu}_2)(\mathbf{Y}_i - \boldsymbol{\mu}_2)^T \right].$$

$$\text{We have } n^{1/2} \mathbf{v}^{*T} (\hat{\Sigma}_n \boldsymbol{\beta}^* - (\bar{\mathbf{X}} - \bar{\mathbf{Y}})) = \underbrace{n^{1/2} \mathbf{v}^{*T} (\tilde{\Sigma}_n \boldsymbol{\beta}^* - (\bar{\mathbf{X}} - \bar{\mathbf{Y}}))}_{I_1} + \underbrace{n^{1/2} \mathbf{v}^{*T} (\hat{\Sigma}_n - \tilde{\Sigma}_n) \boldsymbol{\beta}^*}_{I_2}.$$

We proceed with showing that the term I_2 is small:

$$|I_2| \leq n^{1/2} \|\mathbf{v}^*\|_1 \|\boldsymbol{\beta}^*\|_1 \|\hat{\Sigma}_n - \tilde{\Sigma}_n\|_{\max} = \|\mathbf{v}^*\|_1 \|\boldsymbol{\beta}^*\|_1 O_p(\log d/n^{1/2}) = o_p(1),$$

where we used (I.4) from Lemma I.3 (and made usage of the fact that $n_1 \asymp n_2$). Next we take a closer look at the term I_1 :

$$\begin{aligned} I_1 &= n^{1/2} \mathbf{v}^{*T} (\tilde{\Sigma}_n \boldsymbol{\beta}^* - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)) + n^{1/2} \mathbf{v}^{*T} (\bar{\mathbf{X}} - \boldsymbol{\mu}_1 - \bar{\mathbf{Y}} + \boldsymbol{\mu}_2) \\ &= n^{1/2} \mathbf{v}^{*T} \frac{1}{n} \sum_{i=1}^n \left(\mathbf{U}_i \mathbf{U}_i^T \boldsymbol{\beta}^* - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) + \left[\frac{n}{n_1} I(i \leq n_1) - \frac{n}{n_2} I(i > n_1) \right] \mathbf{U}_i \right). \end{aligned}$$

Next, by $n_1/n + o_p(1) = \alpha$, it is clear that:

$$\begin{aligned} I_1 &= n^{-1/2} \mathbf{v}^{*T} \sum_{i=1}^{n_1} (\mathbf{U}_i \mathbf{U}_i^T \boldsymbol{\beta}^* - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) + \alpha^{-1} \mathbf{U}_i) \\ &\quad + n^{-1/2} \mathbf{v}^{*T} \sum_{i=n_1+1}^n (\mathbf{U}_i \mathbf{U}_i^T \boldsymbol{\beta}^* - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) - (1 - \alpha)^{-1} \mathbf{U}_i) \\ &\quad + \underbrace{\left(\frac{n}{n_1} - \alpha^{-1} \right)}_{o_p(1)} \underbrace{n^{-1/2} \sum_{i=1}^{n_1} \mathbf{v}^{*T} \mathbf{U}_i}_{O_p(1)} + \underbrace{\left(\frac{n}{n_2} - (1 - \alpha)^{-1} \right)}_{o_p(1)} \underbrace{n^{-1/2} \sum_{i=n_1+1}^n \mathbf{v}^{*T} \mathbf{U}_i}_{O_p(1)}, \end{aligned}$$

where we implicitly used Chebyshev's inequality and the fact that $\text{Var}(\mathbf{v}^{*T} \mathbf{U}) \leq 2 \mathbf{v}^{*T} \Sigma \mathbf{v}^* \leq 2\delta^{-1}$. Next we verify Lyapunov's condition. The sum of variances of the terms above equals:

$$n_1 V_1 + n_2 V_2 = n(\alpha V_1 + (1 - \alpha) V_2)(1 + o(1)) \geq n V'_{\min} (\|\boldsymbol{\beta}^*\|_2^2 \|\mathbf{v}^*\|_2^2 + \|\mathbf{v}^*\|_2^2)(1 + o(1)),$$

by Lemma I.1. Without loss of generality let's assume that $\alpha^{-1} > (1 - \alpha)^{-1}$.

It follows then from Lemma I.7, that for any $k > 2$:

$$\mathbb{E} \left| \mathbf{v}^{*T} \mathbf{U}_i \mathbf{U}_i^T \boldsymbol{\beta}^* - \mathbf{v}^{*T} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) + \alpha^{-1} \mathbf{v}^{*T} \mathbf{U}_i \right|^k \leq \|\mathbf{v}^*\|_2^k (C_1 (s_{\mathbf{v}} s)^{k/2} \|\boldsymbol{\beta}^*\|_2^k + C_2 \alpha^{-k} s_{\mathbf{v}}^{k/2}),$$

and similarly:

$$\mathbb{E} \left| \mathbf{v}^{*T} \mathbf{U}_i \mathbf{U}_i^T \boldsymbol{\beta}^* - \mathbf{v}^{*T} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) - (1 - \alpha)^{-1} \mathbf{v}^{*T} \mathbf{U}_i \right|^k \leq \|\mathbf{v}^*\|_2^k (C_1 (s_{\mathbf{v}} s)^{k/2} \|\boldsymbol{\beta}^*\|_2^k + C_2 \alpha^{-k} s_{\mathbf{v}}^{k/2}),$$

where C_1 and C_2 are some absolute constants depending on k (see the Lemma for details). Therefore we conclude that the sum in Lyapunov's condition, is bounded by:

$$\frac{(s_{\mathbf{v}} s)^{k/2}}{(1 + o(1)) n^{k/2-1}} \underbrace{\frac{C_1 \|\boldsymbol{\beta}^*\|_2^k + \frac{C_2 \alpha^{-k}}{s^{k/2}}}{(V'_{\min})^{k/2} (\|\boldsymbol{\beta}^*\|_2^2 + 1)^{k/2}}}_{O(1)} = o(1).$$

This completes the proof. \square

Remark I.2. We propose the following consistent estimator of Δ , and prove its consistency in Proposition I.1.

$$\begin{aligned} \hat{\Delta} := & \frac{1}{n} \sum_{i=1}^{n_1} \left(\hat{\mathbf{v}}^T (\mathbf{X}_i - \bar{\mathbf{X}}) (\mathbf{X}_i - \bar{\mathbf{X}})^T \hat{\boldsymbol{\beta}} \right)^2 + \frac{1}{n} \sum_{i=1}^{n_1} \left(\frac{n}{n_1} \hat{\mathbf{v}}^T (\mathbf{X}_i - \bar{\mathbf{X}}) \right)^2 \\ & + \frac{1}{n} \sum_{i=n_1+1}^n \left(\hat{\mathbf{v}}^T (\mathbf{Y}_i - \bar{\mathbf{Y}}) (\mathbf{Y}_i - \bar{\mathbf{Y}})^T \hat{\boldsymbol{\beta}} \right)^2 \\ & + \frac{1}{n} \sum_{i=n_1+1}^n \left(\frac{n}{n_2} \hat{\mathbf{v}}^T (\mathbf{Y}_i - \bar{\mathbf{Y}}) \right)^2 - (\hat{\mathbf{v}}^T (\bar{\mathbf{X}} - \bar{\mathbf{Y}}))^2. \end{aligned}$$

Proposition I.1. Under the same conditions as in Corollary 3, $\max(\|\boldsymbol{\mu}_1\|_{\infty}, \|\boldsymbol{\mu}_2\|_{\infty}) = O(1)$, and the following additional assumptions:

$$\begin{aligned} \max(\lambda' s_{\mathbf{v}}, \lambda s) \|\mathbf{v}^*\|_1 \|\boldsymbol{\beta}^*\|_1 \sqrt{\log(nd)} &= o(1), \quad \text{Var}((\mathbf{v}^{*T} \mathbf{U})^2) = o(n), \quad \text{Var}(\mathbf{v}^{*T} \mathbf{U} \mathbf{U}^T \boldsymbol{\beta}^*) = o(n), \\ \mathbb{E}(\mathbf{v}^{*T} \mathbf{U} \mathbf{U}^T \boldsymbol{\beta}^*)^2 &= O(1), \end{aligned}$$

we have that $\hat{\Delta} \rightarrow_p \Delta$.

PROOF OF PROPOSITION I.1. Note that Δ can be decomposed as:

$$\begin{aligned} \Delta = & \alpha \mathbb{E}(\mathbf{v}^{*T} \mathbf{U} \mathbf{U}^T \boldsymbol{\beta}^*)^2 + \alpha^{-1} \mathbb{E}(\mathbf{v}^{*T} \mathbf{U})^2 + (1 - \alpha) \mathbb{E}(\mathbf{v}^{*T} \mathbf{U} \mathbf{U}^T \boldsymbol{\beta}^*)^2 \\ & + (1 - \alpha)^{-1} \mathbb{E}(\mathbf{v}^{*T} \mathbf{U})^2 - (\mathbf{v}^{*T} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2))^2. \end{aligned}$$

We start from the last term:

$$\underbrace{(\hat{\mathbf{v}}^T (\bar{\mathbf{X}} - \bar{\mathbf{Y}}))^2}_I = \underbrace{[(\hat{\mathbf{v}}^T (\bar{\mathbf{X}} - \bar{\mathbf{Y}}))^2 - (\mathbf{v}^{*T} (\bar{\mathbf{X}} - \bar{\mathbf{Y}}))^2]}_{I_1} + \underbrace{(\mathbf{v}^{*T} (\bar{\mathbf{X}} - \bar{\mathbf{Y}}))^2}_{I_2}.$$

We have $|I_1| \leq \|\hat{\mathbf{v}} - \mathbf{v}^*\|_1 \|\hat{\mathbf{v}} + \mathbf{v}^*\|_1 \|(\bar{\mathbf{X}} - \bar{\mathbf{Y}})(\bar{\mathbf{X}} - \bar{\mathbf{Y}})^T\|_{\max}$. Using Lemma I.6, we know $\|\hat{\mathbf{v}} - \mathbf{v}^*\|_1 = O_p\left(\|\mathbf{v}^*\|_1 s_{\mathbf{v}} \sqrt{\log d/n}\right)$. We can apply the concentration inequality (I.3) provided in Lemma I.3 to claim that $\|(\bar{\mathbf{X}} - \bar{\mathbf{Y}})(\bar{\mathbf{X}} -$

$\bar{\mathbf{Y}})^T\|_{\max} \leq \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|_{\infty}^2 + \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|_{\infty} O_p\left(\sqrt{\log d/n}\right)$, where we used the triangle inequality $\|\bar{\mathbf{X}} - \bar{\mathbf{Y}}\|_{\infty} \leq \|\bar{\mathbf{X}} - \boldsymbol{\mu}_1\|_{\infty} + \|\bar{\mathbf{Y}} - \boldsymbol{\mu}_2\|_{\infty} + \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|_{\infty}$. Finally due to our assumptions we have $|I_1| = \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|_{\infty}^2 O_p\left(\|\mathbf{v}^*\|_1^2 s_{\mathbf{v}} \sqrt{\log d/n}\right) = o_p(1)$. Next we tackle $I_2 = \underbrace{(\mathbf{v}^{*T}(\bar{\mathbf{X}} - \bar{\mathbf{Y}}))^2}_{I_{21}} - \underbrace{(\mathbf{v}^{*T}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2))^2}_{I_{22}} + \underbrace{(\mathbf{v}^{*T}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2))^2}_{I_{22}}$.

In a similar fashion to before, applying inequality (I.3), we can get $|I_{21}| \leq \|\mathbf{v}^*\|_1^2 O_p\left(\sqrt{\log d/n}\right) \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|_{\infty} = o_p(1)$. Thus we have shown $I = (\mathbf{v}^{*T}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2))^2 + o_p(1)$. To this end define the following shorthand notations:

$$I_X(\mathbf{v}, \boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n_1} (\mathbf{v}^T(\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^T \boldsymbol{\beta})^2, \quad I_Y(\mathbf{v}, \boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n_2} (\mathbf{v}^T(\mathbf{Y}_i - \bar{\mathbf{Y}})(\mathbf{Y}_i - \bar{\mathbf{Y}})^T \boldsymbol{\beta})^2$$

Next we show that $I_X(\hat{\mathbf{v}}, \hat{\boldsymbol{\beta}}) + I_Y(\hat{\mathbf{v}}, \hat{\boldsymbol{\beta}})$ is consistent for $\mathbb{E}(\mathbf{v}^* \mathbf{U} \mathbf{U}^T \boldsymbol{\beta}^*)^2$. We begin with the following bound:

$$\begin{aligned} & \frac{1}{n} \sum_{i=1}^{n_1} \left((\hat{\mathbf{v}} - \mathbf{v}^*)^T (\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^T \hat{\boldsymbol{\beta}} \right)^2 + \frac{1}{n} \sum_{i=1}^{n_2} \left((\hat{\mathbf{v}} - \mathbf{v}^*)^T (\mathbf{Y}_i - \bar{\mathbf{Y}})(\mathbf{Y}_i - \bar{\mathbf{Y}})^T \hat{\boldsymbol{\beta}} \right)^2 \\ & \leq \|\hat{\mathbf{v}} - \mathbf{v}^*\|_1^2 M \|\hat{\boldsymbol{\Sigma}}_n \hat{\boldsymbol{\beta}}\|_{\infty} \|\hat{\boldsymbol{\beta}}\|_1, \end{aligned}$$

where $M = \max \left\{ \max_{i=1, \dots, n_1} \|(\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^T\|_{\max}, \max_{i=1, \dots, n_2} \|(\mathbf{Y}_i - \bar{\mathbf{Y}})(\mathbf{Y}_i - \bar{\mathbf{Y}})^T\|_{\max} \right\}$. Note that the random variables $\mathbf{X}_i - \bar{\mathbf{X}}$ and $\mathbf{Y}_i - \bar{\mathbf{Y}}$ are in fact mean 0 sub-Gaussian variables since e.g. $\|\mathbf{X}_i - \bar{\mathbf{X}}\|_{\psi_2} \leq \|\mathbf{X}_i - \boldsymbol{\mu}_1\|_{\psi_2} + \|\bar{\mathbf{X}} - \boldsymbol{\mu}_1\|_{\psi_2} \leq 2K_{\mathbf{U}}$. Thus an application of Lemma F.7, and the fact that $n_1 \asymp n_2 \asymp n$, gives us that $M = O(\log(nd))$. Furthermore we have:

$$\|\hat{\boldsymbol{\Sigma}}_n \hat{\boldsymbol{\beta}}\|_{\infty} \leq \lambda + \|\bar{\mathbf{X}} - \boldsymbol{\mu}_1\|_{\infty} + \|\bar{\mathbf{Y}} - \boldsymbol{\mu}_2\|_{\infty} + \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|_{\infty} = O_p(1),$$

by application of (I.3), and the way we select λ . Putting the last several inequalities together with Lemma I.5, Lemma I.6 and the triangle inequality, we obtain:

$$|\sqrt{I_X(\hat{\mathbf{v}}, \hat{\boldsymbol{\beta}}) + I_Y(\hat{\mathbf{v}}, \hat{\boldsymbol{\beta}})} - \sqrt{I_X(\mathbf{v}^*, \hat{\boldsymbol{\beta}}) + I_Y(\mathbf{v}^*, \hat{\boldsymbol{\beta}})}| \leq \|\mathbf{v}\|_1 \sqrt{\|\boldsymbol{\beta}\|_1} s_{\mathbf{v}} \sqrt{\log(nd) \log d/n} O_p(1).$$

Applying the same technique we can further show that in fact:

$$|\sqrt{I_X(\hat{\mathbf{v}}, \hat{\boldsymbol{\beta}}) + I_Y(\hat{\mathbf{v}}, \hat{\boldsymbol{\beta}})} - \sqrt{I_X(\mathbf{v}^*, \boldsymbol{\beta}^*) + I_Y(\mathbf{v}^*, \boldsymbol{\beta}^*)}| = o_p(1).$$

We proceed to show that $I_X(\mathbf{v}^*, \boldsymbol{\beta}^*) + I_Y(\mathbf{v}^*, \boldsymbol{\beta}^*)$ is consistent for $\mathbb{E}(\mathbf{v}^* \mathbf{U} \mathbf{U}^T \boldsymbol{\beta}^*)^2$ and since $\mathbb{E}(\mathbf{v}^* \mathbf{U} \mathbf{U}^T \boldsymbol{\beta}^*)^2 = O(1)$, the latter inequality also shows that $I_X(\hat{\mathbf{v}}, \hat{\boldsymbol{\beta}}) + I_Y(\hat{\mathbf{v}}, \hat{\boldsymbol{\beta}})$ is consistent for $\mathbb{E}(\mathbf{v}^* \mathbf{U} \mathbf{U}^T \boldsymbol{\beta}^*)^2$. Define the following notation $\widetilde{M} := \max_{i=1, \dots, n} \|\mathbf{U}_i \mathbf{U}_i^T\|_{\max}$. For exactly the same reasons as for

M we have $\widetilde{M} = O_p(\log(nd))$. Next we consider the difference:

$$\begin{aligned} & |I_X(\mathbf{v}^*, \boldsymbol{\beta}^*) + I_Y(\mathbf{v}^*, \boldsymbol{\beta}^*) - n^{-1} \sum_{i=1}^n (\mathbf{v}^{*T} \mathbf{U}_i \mathbf{U}_i^T \boldsymbol{\beta}^*)^2| \\ & \leq \|\mathbf{v}^*\|_1 \|\boldsymbol{\beta}^*\|_1 V \left(\frac{1}{n} \sum_{i=1}^{n_1} |\mathbf{v}^{*T} (\mathbf{X}_i - \bar{\mathbf{X}})| |(\mathbf{X}_i - \bar{\mathbf{X}})^T \boldsymbol{\beta}^*| + \frac{1}{n} \sum_{i=1}^{n_2} |\mathbf{v}^{*T} (\mathbf{Y}_i - \bar{\mathbf{Y}})| |(\mathbf{Y}_i - \bar{\mathbf{Y}})^T \boldsymbol{\beta}^*| \right. \\ & \quad \left. + \frac{1}{n} \sum_{i=1}^n |\mathbf{v}^{*T} \mathbf{U}_i| |\mathbf{U}_i^T \boldsymbol{\beta}^*| \right), \end{aligned}$$

where

$$V = \max \left\{ \max_{i=1, \dots, n_1} \|(\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^T - \mathbf{U}_i \mathbf{U}_i^T\|_{\max}, \max_{i=1, \dots, n_2} \|(\mathbf{Y}_i - \bar{\mathbf{Y}})(\mathbf{Y}_i - \bar{\mathbf{Y}})^T - \mathbf{U}_{i+n_1} \mathbf{U}_{i+n_1}^T\|_{\max} \right\}$$

Note that by the simple inequality $|ab| \leq (a^2 + b^2)/2$, we have, that the expression in the brackets is bounded by:

$$\leq \mathbf{v}^{*T} (\widehat{\boldsymbol{\Sigma}}_n + \widetilde{\boldsymbol{\Sigma}}_n) \mathbf{v}^* / 2 + \boldsymbol{\beta}^{*T} (\widehat{\boldsymbol{\Sigma}}_n + \widetilde{\boldsymbol{\Sigma}}_n) \boldsymbol{\beta}^* / 2.$$

We have that $\mathbf{v}^{*T} \widehat{\boldsymbol{\Sigma}}_n \mathbf{v}^* \leq \|\mathbf{v}^*\|_1 \|\mathbf{v}^{*T} \widehat{\boldsymbol{\Sigma}}_n\|_{\infty} = \|\mathbf{v}^*\|_1 + \|\mathbf{v}^*\|_1^2 O_p(\sqrt{\log d/n})$. Similarly since by (I.4) $\|\widehat{\boldsymbol{\Sigma}}_n - \widetilde{\boldsymbol{\Sigma}}_n\|_{\max} = O_p(\frac{\log d}{n})$ we have that $\mathbf{v}^{*T} \widetilde{\boldsymbol{\Sigma}}_n \mathbf{v}^* \leq \|\mathbf{v}^*\|_1 + \|\mathbf{v}^*\|_1^2 O_p(\sqrt{\log d/n})$. Similarly one can show that $\boldsymbol{\beta}^{*T} \widehat{\boldsymbol{\Sigma}}_n \boldsymbol{\beta}^* \leq \|\boldsymbol{\beta}^*\|_1 \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|_{\infty} + \|\boldsymbol{\beta}^*\|_1 (\|\boldsymbol{\beta}^*\|_1 \vee 1) O_p(\sqrt{\log d/n})$, and a similar inequality for $\boldsymbol{\beta}^{*T} \widetilde{\boldsymbol{\Sigma}}_n \boldsymbol{\beta}^*$. We next inspect V :

$$\begin{aligned} \max_{i=1, \dots, n_1} \|(\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^T - \mathbf{U}_i \mathbf{U}_i^T\|_{\max} & \leq \max_{i=1, \dots, n_1} 2\|\mathbf{X}_i\|_{\infty} \|\bar{\mathbf{X}} - \boldsymbol{\mu}_1\|_{\infty} \\ & \quad + \|\bar{\mathbf{X}} - \boldsymbol{\mu}_1\|_{\infty} (\|\bar{\mathbf{X}} - \boldsymbol{\mu}_1\|_{\infty} + 2\|\boldsymbol{\mu}_1\|_{\infty}), \end{aligned}$$

and we can similarly bound the other term in V . Note that in Lemma F.7 we showed that $\max_{i=1, \dots, n_1} \|\mathbf{X}_i\|_{\infty} = O_p(\sqrt{\log(nd)})$, and as we argue in (I.3), we have $\|\bar{\mathbf{X}} - \boldsymbol{\mu}_1\|_{\infty} = O_p(\sqrt{\log d/n})$, and thus $V = O_p(\sqrt{\log d/n}) (\sqrt{\log(nd)} + \|\boldsymbol{\mu}_1\|_{\infty} + \|\boldsymbol{\mu}_2\|_{\infty})$. Hence under our assumptions, we have:

$$|I_X(\mathbf{v}^*, \boldsymbol{\beta}^*) + I_Y(\mathbf{v}^*, \boldsymbol{\beta}^*) - n^{-1} \sum_{i=1}^n (\mathbf{v}^{*T} \mathbf{U}_i \mathbf{U}_i^T \boldsymbol{\beta}^*)^2| = o_p(1).$$

Finally we finish this part upon noting that $\left| \frac{1}{n} \sum_{i=1}^n (\mathbf{v}^{*T} \mathbf{U}_i \mathbf{U}_i^T \boldsymbol{\beta}^*)^2 - \mathbb{E} (\mathbf{v}^{*T} \mathbf{U}_i \mathbf{U}_i^T \boldsymbol{\beta}^*)^2 \right| = o_p(1)$, and that under the assumption $\text{Var}((\mathbf{v}^{*T} \mathbf{U} \mathbf{U}^T \boldsymbol{\beta}^*)^2) = o(n)$ by Chebyshev's inequality.

Next we turn our attention to the term $\frac{n}{n_1} \frac{1}{n_1} \sum_{i=1}^{n_1} (\widehat{\mathbf{v}}^T (\mathbf{X}_i - \bar{\mathbf{X}}))^2$, and show it's consistent for $\alpha^{-1} \mathbb{E}(\mathbf{v}^{*T} \mathbf{U}_i)^2$. First note that since $\frac{n}{n_1} = \alpha^{-1} + o(\frac{1}{n})$,

and we will show the rest of the expression is $O_p(1)$, we will just focus on the average term. We first show the following difference is small:

$$\begin{aligned} \left| \frac{1}{n_1} \sum_{i=1}^{n_1} [(\hat{\mathbf{v}}^T(\mathbf{X}_i - \bar{\mathbf{X}}))^2 - (\mathbf{v}^{*T}(\mathbf{X}_i - \bar{\mathbf{X}}))^2] \right| &= |(\hat{\mathbf{v}} - \mathbf{v}^*)^T \hat{\Sigma}_{\mathbf{X}} (\hat{\mathbf{v}} + \mathbf{v}^*)^T| \\ &\leq \|\hat{\mathbf{v}} - \mathbf{v}^*\|_1 (\|\hat{\mathbf{v}} - \mathbf{v}^*\|_1 + 2\|\mathbf{v}^*\|_1) \|\hat{\Sigma}_{\mathbf{X}}\|_{\max}. \end{aligned}$$

Using the same technique as in the proof of Lemma I.3, one can show that $\|\hat{\Sigma}_{\mathbf{X}}\|_{\max} \leq \|\Sigma\|_{\max} + O_p(\sqrt{\log d/n})$. By Lemma I.6 we have $\|\hat{\mathbf{v}} - \mathbf{v}^*\|_1 = \|\mathbf{v}^*\|_1 s_{\mathbf{v}} O_p(\sqrt{\log d/n})$, and hence:

$$\frac{1}{n_1} \left| \sum_{i=1}^{n_1} [(\hat{\mathbf{v}}^T(\mathbf{X}_i - \bar{\mathbf{X}}))^2 - (\mathbf{v}^{*T}(\mathbf{X}_i - \bar{\mathbf{X}}))^2] \right| \leq \|\mathbf{v}^*\|_1^2 s_{\mathbf{v}} O_p(\sqrt{\log d/n}) = o_p(1),$$

by assumption. Next we control:

$$\begin{aligned} \left| \frac{1}{n_1} \sum_{i=1}^{n_1} [(\mathbf{v}^{*T}(\mathbf{X}_i - \bar{\mathbf{X}}))^2 - (\mathbf{v}^{*T}\mathbf{U}_i)^2] \right| &= \left| \mathbf{v}^{*T}(\boldsymbol{\mu}_1 - \bar{\mathbf{X}}) \frac{1}{n_1} \sum_{i=1}^{n_1} (2\mathbf{X}_i - \bar{\mathbf{X}} - \boldsymbol{\mu}_1) \mathbf{v}^* \right| \\ &\leq \|\mathbf{v}^*\|_1^2 \|\boldsymbol{\mu}_1 - \bar{\mathbf{X}}\|_{\infty}^2 = \|\mathbf{v}^*\|_1^2 O_p\left(\frac{\log d}{n}\right) = o_p(1). \end{aligned}$$

Thus after using Chebyshev's inequality upon observing that $\text{Var}((\mathbf{v}^{*T}\mathbf{U})^2) = o(n)$, we have shown the desired consistency. Similarly we can also show that $\frac{n}{n_2} \frac{1}{n_2} \sum_{i=1}^{n_2} (\hat{\mathbf{v}}^T(\mathbf{Y}_i - \bar{\mathbf{Y}}))^2$ is consistent for $(\alpha - 1)^{-1} \mathbb{E}(\mathbf{v}^{*T}\mathbf{U}_i)^2$. This concludes the proof. \square

Lemma I.3. The following inequality holds $\|\hat{\Sigma}_n - \Sigma\|_{\max} \leq \tilde{t}_{\mathbf{U}}(d, n) + t_{\mathbf{U}}^2(d, n)$, with probability at least $1 - 2d^{2-\bar{c}A_{\mathbf{U}}^2} - 2ed^{1-cA_{\mathbf{U}}^2}$, where:

$$(I.2) \quad t_{\mathbf{U}}(d, n) = A_{\mathbf{U}} K_{\mathbf{U}} \sqrt{\log d / \min(n_1, n_2)}; \quad \tilde{t}_{\mathbf{U}}(d, n) = 4A_{\mathbf{U}} K_{\mathbf{U}}^2 \sqrt{\log d / n}.$$

and $A_{\mathbf{U}} > 0$ is an arbitrary positive constant, \bar{c} and c are absolute contents independent of the distribution of \mathbf{U} , and $K_{\mathbf{U}}$ is as defined in the main section of the text.

PROOF OF LEMMA I.3. We start by showing a concentration bound on $\|\bar{\mathbf{X}} - \boldsymbol{\mu}_1\|_{\infty}$ and $\|\bar{\mathbf{Y}} - \boldsymbol{\mu}_2\|_{\infty}$. By proposition 5.10 in Vershynin (2010) and the union bound, we have:

$$(I.3) \quad \mathbb{P}(\|\bar{\mathbf{X}} - \boldsymbol{\mu}_1\|_{\infty} > t) \leq ed \exp\left(-\frac{cn_1 t^2}{K_{\mathbf{U}}^2}\right).$$

A similar inequality holds for $\|\bar{\mathbf{Y}} - \boldsymbol{\mu}_2\|_{\infty}$. Select $t_{\mathbf{U}}(d, n) = A_{\mathbf{U}} K_{\mathbf{U}} \sqrt{\log d / \min(n_1, n_2)}$, where $A_{\mathbf{U}} > 0$ is some large constant. The triangle inequality yields $\|\hat{\Sigma}_n -$

$\Sigma\|_{\max} \leq \|\widehat{\Sigma}_n - \widetilde{\Sigma}_n\|_{\max} + \|\widetilde{\Sigma}_n - \Sigma\|_{\max}$, where $\widetilde{\Sigma}_n$ is defined as in (I.1). Next, we have that:

$$\begin{aligned} \|\widehat{\Sigma}_n - \widetilde{\Sigma}_n\|_{\max} &\leq \frac{n_1}{n} \|(\bar{\mathbf{X}} - \boldsymbol{\mu}_1)(\bar{\mathbf{X}} - \boldsymbol{\mu}_1)^T\|_{\max} + \frac{n_2}{n} \|(\bar{\mathbf{Y}} - \boldsymbol{\mu}_2)(\bar{\mathbf{Y}} - \boldsymbol{\mu}_2)^T\|_{\max} \\ (I.4) \quad &\leq \frac{n_1}{n} (\|\bar{\mathbf{X}} - \boldsymbol{\mu}_1\|_{\infty})^2 + \frac{n_2}{n} (\|\bar{\mathbf{Y}} - \boldsymbol{\mu}_2\|_{\infty})^2 \leq t_U^2(d, n). \end{aligned}$$

where the last inequality holds with high probability. Note that by Lemma F.2 we have:

$$\|\widetilde{\Sigma}_n - \Sigma\|_{\max} \leq 4A_U K_U^2 \sqrt{\log d/n} =: \tilde{t}_U(d, n),$$

with probability at least $1 - 2d^{2-\bar{c}A_U^2}$. Adding the last two inequalities completes the proof. \square

Lemma I.4. Assume the same conditions as in Lemma I.3, and assume further that the minimum eigenvalue $\lambda_{\min}(\Sigma) > 0$ and $s(\tilde{t}_U(d, n) + t_U^2(d, n)) \leq (1 - \kappa) \frac{\lambda_{\min}(\Sigma)}{(1+\xi)^2}$, where $0 < \kappa < 1$. We then have that $\widehat{\Sigma}_n$ satisfies the RE property with $\text{RE}_{\widehat{\Sigma}_n}(s, \xi) \geq \kappa \lambda_{\min}(\Sigma)$ with probability at least $1 - 2d^{2-\bar{c}A_U^2} - 2ed^{1-cA_U^2}$.

Remark I.3. In fact this event happens on the same event as in Lemma I.3.

PROOF OF LEMMA I.4. The proof follows the proof of Lemma F.3, but uses Lemma I.3 instead of Lemma F.2, hence we omit it. \square

Lemma I.5. Assume that $\lambda_{\min}(\Sigma) > 0$, $s(\tilde{t}_U(d, n) + t_U^2(d, n)) \leq (1 - \kappa) \frac{\lambda_{\min}(\Sigma)}{(1+\xi)^2}$, where $0 < \kappa < 1$ and $\lambda \geq (\tilde{t}_U(d, n) + t_U^2(d, n)) \|\boldsymbol{\beta}^*\|_1 + 2t_U(d, n)$. Then we have that $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \leq \frac{8\lambda s}{\text{RE}(s, 1)}$ with probability at least $1 - 2d^{2-\bar{c}A_U^2} - 2ed^{1-cA_U^2}$. (see (I.2) for definition of t_U and \tilde{t}_U)

Remark I.4. In fact this event happens on the same event as in Lemma I.3.

PROOF OF LEMMA I.5. We start by showing the true parameter $\Omega\boldsymbol{\delta} = \boldsymbol{\beta}^*$ satisfies the sparse LDA constraint — $\|\widehat{\Sigma}_n \boldsymbol{\beta}^* - (\bar{\mathbf{X}} - \bar{\mathbf{Y}})\|_{\infty} \leq \lambda$ with probability at least $1 - 2d^{2-\bar{c}A_U^2} - 2ed^{1-cA_U^2}$. We have that:

$$\begin{aligned} \|\widehat{\Sigma}_n \boldsymbol{\beta}^* - (\bar{\mathbf{X}} - \bar{\mathbf{Y}})\|_{\infty} &\leq \underbrace{\|\Sigma \boldsymbol{\beta}^* - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\|_{\infty}}_0 + \|\widehat{\Sigma}_n - \Sigma\|_{\max} \|\boldsymbol{\beta}^*\|_1 \\ &\quad + \|\bar{\mathbf{X}} - \boldsymbol{\mu}_1\|_{\infty} + \|\bar{\mathbf{Y}} - \boldsymbol{\mu}_2\|_{\infty}. \end{aligned}$$

Collecting the bounds we derived in Lemma I.3 we get:

$$\|\widehat{\Sigma}_n \boldsymbol{\beta}^* - (\bar{\mathbf{X}} - \bar{\mathbf{Y}})\|_{\infty} \leq (\tilde{t}_U(d, n) + t_U^2(d, n)) \|\boldsymbol{\beta}^*\|_1 + 2t_U(d, n).$$

The last inequality implies that if we select $\lambda \geq (\tilde{t}_U(d, n) + t_U^2(d, n)) \|\boldsymbol{\beta}^*\|_1 + 2t_U(d, n)$, it will follow that $\boldsymbol{\beta}^*$ satisfies the constraint with probability at least $1 - 2d^{2-cA_U^2} - 2ed^{1-cA_U^2}$.

The rest of the proof is identical to the proof of Lemma F.4 but instead of using Lemma F.3 we use Lemma I.4. Thus we omit the proof. \square

Lemma I.6. Assume that — $\lambda_{\min}(\Sigma) > 0$, $s_{\mathbf{v}}(\tilde{t}_{\mathbf{U}}(d, n) + t_{\mathbf{U}}^2(d, n)) \leq (1 - \kappa) \frac{\lambda_{\min}(\Sigma)}{(1+\xi)^2}$, where $0 < \kappa < 1$ and $\lambda' \geq \|\mathbf{v}^*\|_1(\tilde{t}_{\mathbf{U}}(d, n) + t_{\mathbf{U}}^2(d, n))$. Then we have that $\|\hat{\mathbf{v}} - \mathbf{v}^*\|_1 \leq \frac{8\lambda' s_{\mathbf{v}}}{\text{RE}_{\kappa}(s_{\mathbf{v}}, 1)}$ with probability at least $1 - 2d^{2-\bar{c}A_{\mathbf{U}}^2} - 2ed^{1-cA_{\mathbf{U}}^2}$. (see (I.2) for definition of $t_{\mathbf{U}}$ and $\tilde{t}_{\mathbf{U}}$)

Remark I.5. In fact this event happens on the same event as in Lemma I.3.

PROOF OF LEMMA I.6. The proof is identical to the one of Lemma F.4 but instead of using Lemma F.3 we use Lemma I.4, and we use Lemma I.3 instead of using Lemma F.2. We omit the proof. \square

Lemma I.7. We have the following inequality:

$$\mathbb{E}|\mathbf{v}^{*T} \mathbf{U} \mathbf{U}^T \beta^* - \mathbf{v}^{*T}(\mu_1 - \mu_2) + c \mathbf{v}^{*T} \mathbf{U}|^k \leq 2^{k-1} \|\mathbf{v}^*\|_2^k (\|\beta^*\|_2^k (s_{\mathbf{v}} s)^{k/2} (8k K_{\mathbf{U}}^2)^k + |c|^k s_{\mathbf{v}}^{k/2} (\sqrt{k} K_{\mathbf{U}})^k).$$

PROOF OF LEMMA I.7. The argument follows applying standard inequalities, and the details are omitted. \square

APPENDIX J: PROOFS FOR VECTOR AUTOREGRESSIVE MODELS

Define the following quantities which will be used throughout. Let:

$$K_d(\Sigma_0, \mathbf{A}) := \frac{32 \|\Sigma_0\|_2 \max_j (\Sigma_{0,jj})}{\min_j (\Sigma_{0,jj}) (1 - \|\mathbf{A}\|_2)}, \quad \tilde{K}_d(\Sigma_0, \mathbf{A}) := K_d(\Sigma_0, \mathbf{A}) (2M + 3).$$

We set

$$(J.1) \quad \lambda := \tilde{K}_d(\Sigma_0, \mathbf{A}) \sqrt{\log d/T}, \quad \lambda' := \frac{K_d(\Sigma_0, \mathbf{A})}{2} \|\Sigma_0^{-1}\|_1 \left(\sqrt{6 \log d/T} + 2\sqrt{1/T} \right).$$

Lemma J.1. Assume that $\Sigma_0 \in \mathcal{L}$, $\mathbf{A} \in \mathcal{M}(s)$, $\min_j \Psi_{jj} \geq C > 0$ and $\max(s_{\mathbf{v}}, s) \log d = o(\sqrt{T})$. Then the following relationships hold:

$$\lambda = o(1), \quad \lambda' = o(1), \quad \sqrt{T} \max(s_{\mathbf{v}}, s) \|\Sigma_0^{-1}\|_1 \lambda' \lambda = o(1),$$

$$\Delta \geq C' > 0, \quad \frac{\beta^{*T} \Sigma_0 \beta^*}{\Psi_{mm}} = o(T), \quad \frac{\|\mathbf{v}^*\|_1^2}{\mathbf{v}^{*T} \Sigma_0 \mathbf{v}^*} \frac{\lambda'}{\|\Sigma_0^{-1}\|_1} = o(1),$$

where C' is some positive constant.

PROOF OF LEMMA J.1. Clearly, since $M = O(1)$, $\|\Sigma_0^{-1}\|_1 = O(1)$, $K_d(\Sigma_0, \mathbf{A}) = O(1)$ and $\max(s_{\mathbf{v}}, s) \log d = o(\sqrt{T})$, it follows that $\lambda = o(1)$, $\lambda' = o(1)$, $\lambda' M = o(1)$ and in addition $\sqrt{T} \max(s_{\mathbf{v}}, s) \|\Sigma_0^{-1}\|_1 \lambda' \lambda = o(1)$.

By the inequality $(\Sigma_0^{-1})_{jj} \Sigma_{0,jj} \geq 1$ it follows that $(\Sigma_0^{-1})_{0,jj} \geq (\max_j \Sigma_{0,jj})^{-1} \geq \|\Sigma_0\|_2^{-1} \geq M^{-1}$. Hence $\Delta = \Psi_{mm} \mathbf{v}^{*T} \Sigma_0 \mathbf{v}^* \geq \Psi_{mm} \min_{jj} \Sigma_{0,jj}^{-1} \geq C M^{-1} > 0$.

Next, to show that $\frac{\beta^{*T} \Sigma_0 \beta^*}{\Psi_{mm}} = o(T)$, it suffices to see that $\beta^{*T} \Sigma_0 \beta^* = O(1)$. To this end note that $|\beta^{*T} \Sigma_0 \beta^*| \leq \|\Sigma_{1,*k}\|_\infty \|\mathbf{A}\|_1 \leq \|\Sigma_{1,*k}\|_\infty M$. Next since $\Psi = \Sigma_0 - \Sigma_1$, we have $\|\Sigma_{1,*k}\|_\infty \leq \max_j \Sigma_{0,jj} - \min_j \Psi_{jj} \leq \|\Sigma_0\|_2$, which shows that $|\beta^{*T} \Sigma_0 \beta^*| = O(1)$.

Finally we check $\frac{\|\mathbf{v}^*\|_1^2}{\mathbf{v}^{*T} \Sigma_0 \mathbf{v}^*} \frac{\lambda'}{\|\Sigma_0^{-1}\|_1} = o(1)$. Note that $\|\mathbf{v}^*\|_1 \leq \|\Sigma_0^{-1}\|_1$, and also that $\mathbf{v}^{*T} \Sigma_0 \mathbf{v}^* \geq \min_j \Sigma_{0,jj} \geq \min_j \Psi_{jj} \geq C$, hence it suffices to show that $\|\mathbf{v}^*\|_1 \lambda' = o(1)$. However, evidently $\|\mathbf{v}^*\|_1 \leq \|\Sigma_0^{-1}\|_1 = O(1)$ and $\lambda' = o(1)$, which shows what we wanted. \square

Next we summarize several results by [Han et al. \(2014\)](#), which we use in the later development.

Theorem J.1 (Theorem 4.1. [Han et al. \(2014\)](#)). Suppose that $(\mathbf{X}_t)_{t=1}^T$ from a lag 1 vector autoregressive process $(\mathbf{X}_t)_{t=-\infty}^\infty$. Assume that $\mathbf{A} \in \mathcal{M}(s)$. Let $\hat{\mathbf{A}}$ be the optimizer of (1.6) with the tuning parameter:

$$\lambda = \tilde{K}_d(\Sigma_0, \mathbf{A}) \sqrt{\log d/T}.$$

For $T \geq 6 \log d + 1$ and $d \geq 8$, we have, with probability at least $1 - 14d^{-1}$: $\|\hat{\mathbf{A}} - \mathbf{A}\|_1 \leq 4s \|\Sigma_0^{-1}\|_1 \lambda$. In fact on the same event (see Lemmas A.1. and A.2. ([Han et al., 2014](#))), we have:

$$\begin{aligned} \|\mathbf{S}_0 - \Sigma_0\|_{\max} &\leq K_d(\Sigma_0, \mathbf{A})/2 \left(\sqrt{6 \log d/T} + 2\sqrt{1/T} \right), \\ \|\mathbf{S}_1 - \Sigma_1\|_{\max} &\leq K_d(\Sigma_0, \mathbf{A}) \left(\sqrt{3 \log d/T} + 2/T \right). \end{aligned}$$

PROOF OF COROLLARY 4. We verify the conditions of Section 3. To see Assumption (3.5), note that by Theorem J.1 we have $\|\hat{\beta} - \beta^*\|_1 \leq 4s \|\Sigma_0^{-1}\|_1 \lambda$. Next we inspect $\|\hat{\mathbf{v}} - \mathbf{v}^*\|_1 = O_p(s_{\mathbf{v}} \|\Sigma_0^{-1}\|_1 \lambda')$ according to Lemma J.3. Next we check Assumption 1. To see (3.2), fix a $|\theta - \theta^*| < \epsilon$, for some $\epsilon > 0$. By the triangle inequality:

$$\|\mathbf{S}_0 \beta_\theta^* - \Sigma_0 \beta_\theta^* - \mathbf{S}_{1,*m} + \Sigma_{1,*m}\|_\infty \leq \|\mathbf{S}_0 \beta^* - \mathbf{S}_{1,*m}\|_\infty + \|\mathbf{S}_{0,*1} - \Sigma_{0,*1}\|_\infty \epsilon$$

The RHS is $O_p(\lambda)$, by Theorem J.1 and by the fact that $\|\mathbf{S}_0 \beta^* - \mathbf{S}_{1,*m}\|_\infty \leq \lambda$ with probability at least $1 - 14d^{-1}$, as is seen from the proof of Theorem J.1 (see [Han et al. \(2014\)](#) for details). The same logic shows that $r_2(n) \asymp \|\mathbf{v}^*\|_1 \lambda$, which implies (3.3). Since the Hessian \mathbf{T} in (3.4) is free of β we are allowed to set $r_3(n) = \lambda' = o(1)$. Finally the two expectations in Assumption 1, are bounded as we see below:

$$\|\Sigma_0 \beta_\theta^* - \Sigma_0 \beta^*\|_\infty \leq \|\Sigma_{0,*1}\|_\infty \epsilon = \Sigma_{0,11} \epsilon, \quad \|\mathbf{v}^{*T} \Sigma_{0,-1}\|_\infty = 0.$$

By adding up the following two identities:

$$\sqrt{T} O_p(4s \|\Sigma_0^{-1}\|_1 \lambda) O_p(\lambda') = o_p(1), \quad \sqrt{T} O_p(s_{\mathbf{v}} \|\Sigma_0^{-1}\|_1 \lambda') O_p(\lambda) = o_p(1),$$

we get that (3.8) is also valid in this case by assumption.

To verify the consistency of $\tilde{\theta}$ we check the assumptions in Theorem 1. Clearly the map $\mathbf{v}^{*T}\Sigma_0(\beta_\theta^* - \beta^*) = (\theta - \theta^*)$ has a unique 0 when $\theta = \theta^*$. Moreover, the map $\theta \mapsto \hat{\mathbf{v}}^T(\Sigma_0\hat{\beta}_\theta - \mathbf{S}_1)$ is continuous as it is linear. In addition, it has a unique zero except in cases when $\hat{\mathbf{v}}^T\Sigma_{0,*1} = 0$. However note that $|\hat{\mathbf{v}}^T\Sigma_{0,*1} - 1| \leq \lambda'$ by (4.4), and hence for small enough values of λ' there will exist a unique zero.

Next we verify Assumption 3 in Lemma J.2. In addition the fact that $\hat{\Delta}$ is consistent for Δ is checked in Proposition J.1. Finally we move on to show (3.7). Observe that (3.7) is trivial as its LHS $\equiv 0$ in this case. \square

Lemma J.2. Under the conditions of Corollary 4, we have that:

$$\Delta^{-1/2}T^{1/2}S(\beta^*) = \Delta^{-1/2}T^{1/2}\mathbf{v}^{*T}(\mathbf{S}_0\beta^* - \mathbf{S}_1) \rightsquigarrow N(0, 1),$$

where the definition of Δ is given in (4.10).

PROOF OF LEMMA J.2. First, construct the sequence $\xi_1 = 0$, $\xi_{t+1} = \frac{\mathbf{v}^{*T}\mathbf{X}_t\mathbf{X}_t^T\beta^* - \mathbf{v}^{*T}\mathbf{X}_t\mathbf{X}_{t+1}^T\mathbf{e}_m^T}{\sqrt{(T-1)\Delta}}$ for $t = 1, \dots, T-1$. We start by showing that the difference between the sequence $\sum_{t=1}^T \xi_t$ and $\frac{\sqrt{T-1}\mathbf{v}^{*T}(\mathbf{S}_0\beta^* - \mathbf{S}_{1,*m})}{\sqrt{\Delta}}$ is asymptotically negligible. We have:

$$\sum_{t=1}^T \xi_t - \frac{\sqrt{T-1}\mathbf{v}^{*T}(\mathbf{S}_0\beta^* - \mathbf{S}_{1,*m})}{\sqrt{\Delta}} = \underbrace{\frac{(\sqrt{T-1}T)^{-1}}{\sqrt{\Delta}} \sum_{t=1}^T \mathbf{v}^{*T}\mathbf{X}_t\mathbf{X}_t^T\beta^*}_{I_1} - \underbrace{\frac{\mathbf{v}^{*T}\mathbf{X}_T\mathbf{X}_T^T\beta^*}{\sqrt{T-1}\sqrt{\Delta}}}_{I_2}.$$

By Lemma J.3, $\|\mathbf{v}^{*T}\mathbf{S}_0 - \mathbf{e}_1\|_\infty \leq \lambda'$ with probability not smaller than $1 - 14d^{-1}$. Thus we have:

$$|I_1| \leq \frac{\lambda'\|\beta^*\|_1 + |\beta_1^*|}{\sqrt{T-1}\sqrt{\Delta}} \leq \frac{\lambda'M + |\beta_1^*|}{\sqrt{T-1}\sqrt{\Delta}} = o(1),$$

with probability at least $1 - 14d^{-1}$, where we used the fact that $|\beta_1^*| = O(1)$.

Next observe that $\mathbb{E}I_2 = 0$, and using Isserlis' theorem and Cauchy-Schwartz we have:

$$\text{Var}(I_2) = \frac{1}{T-1} \left(\frac{\beta^{*T}\Sigma_0\beta^*}{\Psi_{mm}} + \frac{(\mathbf{v}^{*T}\Sigma_0\beta^*)^2}{\mathbf{v}^{*T}\Sigma_0\mathbf{v}^*\Psi_{mm}} \right) \leq \frac{2}{T-1} \frac{\beta^{*T}\Sigma_0\beta^*}{\Psi_{mm}} = o(1).$$

and hence $I_2 = o_p(1)$. The last shows that, $\sum_{t=1}^T \xi_t - \frac{\sqrt{T}\mathbf{v}^{*T}(\mathbf{S}_0\beta^* - \mathbf{S}_{1,*m})}{\sqrt{\Delta}} = o_p(1)$, provided that $\sum_{t=1}^T \xi_t = O_p(1)$, which we show next. Observe that the sequence $(\xi_t)_{t=1}^T$ forms a martingale difference sequence with respect to the filtration $\mathcal{F}_t = \sigma(\mathbf{X}_1, \dots, \mathbf{X}_t)$ for $t = 1, \dots, T$, as we clearly have $\mathbb{E}[\xi_t|\mathcal{F}_{t-1}] = 0$. Furthermore a simple calculation yields that for $t \geq 2$ we

have $\mathbb{E}[\xi_t^2 | \mathcal{F}_{t-1}] = \frac{(\mathbf{v}^{*T} \mathbf{X}_{t-1})^2}{(T-1)\mathbf{v}^{*T} \Sigma_0 \mathbf{v}^*}$. Thus:

$$\begin{aligned} \left| \sum_{t=1}^T \mathbb{E}[\xi_t^2 | \mathcal{F}_{t-1}] - 1 \right| &= \left| \frac{\mathbf{v}^{*T}}{(T-1)\mathbf{v}^{*T} \Sigma_0 \mathbf{v}^*} \sum_{t=1}^{T-1} [\mathbf{X}_t \mathbf{X}_t^T - \Sigma_0] \mathbf{v}^* \right| \\ &\leq \frac{\|\mathbf{v}^*\|_1^2}{\mathbf{v}^{*T} \Sigma_0 \mathbf{v}^*} \underbrace{\left\| \frac{1}{T-1} \sum_{t=1}^{T-1} [\mathbf{X}_t \mathbf{X}_t^T - \Sigma_0] \right\|_{\max}}_I. \end{aligned}$$

Using Theorem J.1, it is evident that $I \leq K_d(\Sigma_0, \mathbf{A})/2 \left(\sqrt{\frac{6 \log d}{T-1}} + 2\sqrt{\frac{1}{T-1}} \right)$ with probability at least $1 - 14d^{-1}$, and hence the above quantity converges to 0 in probability.

Having noted these facts, we want to show that $\sum_{t=1}^T \xi_t$ converges weakly to a $N(0, 1)$ with the help of a version of the martingale central limit theorem (MCLT) (Hall and Heyde, 1980). Next we show the Lindeberg condition for the MCLT. For $t \geq 2$ and a fixed $\delta > 0$ we have:

$$\mathbb{E}[\xi_t^2 1(|\xi_t| \geq \delta) | \mathcal{F}_{t-1}] = \frac{(\mathbf{v}^{*T} \mathbf{X}_{t-1})^2 \mathbb{E}[Z^2 1(|Z| > \delta C)]}{(T-1)\mathbf{v}^{*T} \Sigma_0 \mathbf{v}^*},$$

where $Z \sim N(0, 1)$ and $C = \left\{ \frac{(\mathbf{v}^{*T} \mathbf{X}_{t-1})^2}{(T-1)\mathbf{v}^{*T} \Sigma_0 \mathbf{v}^*} \right\}^{-\frac{1}{2}}$. Using the properties of the truncated standard normal distribution we have that $\mathbb{E}[Z^2 | Z > c] = 1 + \frac{\phi(c)}{\Phi(c)}c$, and hence

$$\mathbb{E}[Z^2 1(|Z| > c)] = 2\bar{\Phi}(c) \left(1 + \frac{\phi(c)}{\bar{\Phi}(c)}c \right) = 2\bar{\Phi}(c) + 2\phi(c)c \leq 2\phi(c)(c^{-1} + c),$$

where the last inequality follows from a standard tail bound for the normal distribution.

Now notice that by the union bound and a standard bound on the normal cdf we have

$$\mathbb{P}(\max_{t=1, \dots, T} |\mathbf{v}^{*T} \mathbf{X}_t| > u) \leq 2T \exp(-u^2 / (2\mathbf{v}^{*T} \Sigma_0 \mathbf{v}^*)).$$

Selecting $u = 2\sqrt{\log(T)\mathbf{v}^{*T} \Sigma_0 \mathbf{v}^*}$ gives $\max_t |\mathbf{v}^{*T} \mathbf{X}_t| \leq 2\sqrt{\log(T)\mathbf{v}^{*T} \Sigma_0 \mathbf{v}^*}$ with probability at least $1 - \frac{2}{T}$. Hence on this event we have:

$$\mathbb{E}[\xi_t^2 1(|\xi_t| \geq \delta) | \mathcal{F}_{t-1}] \leq \frac{8 \log T \phi(\delta \tilde{C}) ((\delta \tilde{C})^{-1} + \delta \tilde{C})}{(T-1)},$$

where $\tilde{C} = \sqrt{\frac{T-1}{4 \log T}}$, and we used the fact that the function $\phi(x)(x^{-1} + x)$ is decreasing. Summing up over t yields:

$$\sum_{t=1}^T \mathbb{E}[\xi_t^2 1(|\xi_t| \geq \delta) | \mathcal{F}_{t-1}] \leq 8 \log T \phi(\delta \tilde{C}) ((\delta \tilde{C})^{-1} + \delta \tilde{C}) \rightarrow 0.$$

This shows that the Lindeberg condition holds with probability 1. Hence by the MCLT we can claim $\sum_{t=1}^T \xi_t \rightsquigarrow N(0, 1)$. \square

Proposition J.1. Under the assumptions of Corollary 4 we have $\widehat{\Delta} \rightarrow_p \Delta$.

PROOF OF PROPOSITION J.1. We begin with showing the consistency of $\widehat{\Psi}_{mm} = \mathbf{S}_{0,mm} - \widehat{\beta}^T \mathbf{S}_0 \widehat{\beta}$ is consistent for Ψ_{mm} . First note that $\Psi = \Sigma_0 - \mathbf{A}^T \Sigma_0 \mathbf{A}$, and thus $\Psi_{mm} = \Sigma_{0,mm} - \beta^{*T} \Sigma_0 \beta^*$. Then we have:

$$|\widehat{\Psi}_{mm} - \Psi_{mm}| \leq |\mathbf{S}_{0,mm} - \Sigma_{0,mm}| + |(\widehat{\beta} - \beta^*)^T \mathbf{S}_0 \widehat{\beta}| + |\beta^{*T} (\mathbf{S}_0 \widehat{\beta} - \Sigma_0 \beta^*)|.$$

Firstly, by Theorem J.1, we have with probability at least $1 - 14d^{-1}$:

$$|\mathbf{S}_{0,mm} - \Sigma_{0,mm}| \leq \|\mathbf{S}_0 - \Sigma_0\|_{\max} \leq \frac{K_d(\Sigma_0, \mathbf{A})}{2} \left(\sqrt{6 \log d/T} + 2\sqrt{1/T} \right) = \lambda' \|\Sigma_0^{-1}\|_1^{-1} = o(1).$$

Secondly:

$$\begin{aligned} |(\widehat{\beta} - \beta^*)^T \mathbf{S}_0 \widehat{\beta}| &\leq \|\widehat{\beta} - \beta^*\|_1 (\|\mathbf{S}_0 \beta^*\|_{\infty} + \|\mathbf{S}_0 \widehat{\beta} - \mathbf{S}_0 \beta^*\|_{\infty}) \\ &\leq \|\widehat{\beta} - \beta^*\|_1 (\|\mathbf{S}_0\|_{\max} \|\beta^*\|_1 + \|\mathbf{S}_0 \widehat{\beta} - \mathbf{S}_{1,*m}\|_{\infty} + \|\mathbf{S}_0 \beta^* - \mathbf{S}_{1,*m}\|_{\infty}). \end{aligned}$$

On the event of Theorem J.1 we further have:

$$\|\beta - \beta^*\|_1 \|\mathbf{S}_0\|_{\max} \|\beta^*\|_1 \leq 4s \|\Sigma_0^{-1}\|_1 \lambda [\|\Sigma_0\|_{\max} + \|\mathbf{S}_0 - \Sigma_0\|_{\max}] M = o(1).$$

Furthermore within the proof of Theorem J.1, it can be seen that on the event of interest we have $\|\mathbf{S}_0 \beta^* - \mathbf{S}_{1,*m}\|_{\infty} \leq \lambda$, and hence:

$$\|\widehat{\beta} - \beta^*\|_1 (\|\mathbf{S}_0 \widehat{\beta} - \mathbf{S}_{1,*m}\|_{\infty} + \|\mathbf{S}_0 \beta^* - \mathbf{S}_{1,*m}\|_{\infty}) \leq 2\lambda \|\widehat{\beta} - \beta^*\|_1 = o_p(1).$$

Lastly,

$$\begin{aligned} |\beta^{*T} (\mathbf{S}_0 \widehat{\beta} - \Sigma_0 \beta^*)| &\leq |\beta^{*T} \mathbf{S}_0 (\widehat{\beta} - \beta^*)| + |\beta^{*T} (\mathbf{S}_0 - \Sigma_0) \beta^*| \\ &\leq \|\beta^*\|_1 2\lambda + \|\beta^*\|_1 [\|\mathbf{S}_0 \beta^* - \mathbf{S}_{1,*m}\|_{\max} + \|\mathbf{S}_{1,*m} - \Sigma_{1,*m}\|_{\max}] \\ &\leq M3\lambda + MK_d(\Sigma_0, \mathbf{A}) \left(\sqrt{3 \log d/T} + 2/T \right) = o(1), \end{aligned}$$

where the last two inequalities hold on the event of Theorem J.1, and we used the fact that $\|\beta^*\|_1 \leq M$ since $\mathbf{A} \in \mathcal{M}(s)$.

Next, we show that $\widehat{\mathbf{v}}^T \mathbf{S}_0 \widehat{\mathbf{v}} \rightarrow_p \mathbf{v}^{*T} \Sigma_0 \mathbf{v}^*$. Similarly to before we have:

$$|\widehat{\mathbf{v}}^T \mathbf{S}_0 \widehat{\mathbf{v}} - \mathbf{v}^{*T} \Sigma_0 \mathbf{v}^*| \leq |(\widehat{\mathbf{v}} - \mathbf{v}^*)^T \mathbf{S}_0 \widehat{\mathbf{v}}| + |\mathbf{v}^{*T} (\mathbf{S}_0 \widehat{\mathbf{v}} - \Sigma_0 \mathbf{v}^*)|$$

For the first term we have:

$$\begin{aligned} |(\widehat{\mathbf{v}} - \mathbf{v}^*)^T \mathbf{S}_0 \widehat{\mathbf{v}}| &\leq \|\widehat{\mathbf{v}} - \mathbf{v}^*\|_1 \|\widehat{\mathbf{v}}^T \mathbf{S}_0 - \mathbf{e}_1\|_{\infty} + |\mathbf{e}_1 (\widehat{\mathbf{v}} - \mathbf{v}^*)| \leq 4s_{\mathbf{v}} \|\Sigma_0^{-1}\|_1 (\lambda')^2 + \|\widehat{\mathbf{v}} - \mathbf{v}^*\|_{\infty} \\ &\leq 4s_{\mathbf{v}} \|\Sigma_0^{-1}\|_1 (\lambda')^2 + \|\Sigma_0^{-1}\|_1 2\lambda' = o(1), \end{aligned}$$

with the last two inequalities following from Lemma J.3 and holding on the event from Theorem J.1. Recall that \mathbf{e} is a unit row vector.

Finally, for the second term we have:

$$|\mathbf{v}^{*T} (\mathbf{S}_0 \widehat{\mathbf{v}} - \Sigma_0 \mathbf{v}^*)| \leq \|\mathbf{v}^*\|_1 \lambda' \leq \|\Sigma_0^{-1}\|_1 \lambda' = o(1),$$

and this concludes the proof. \square

Lemma J.3. Assume the assumptions in Theorem J.1. Let

$$\lambda' = \|\Sigma_0^{-1}\|_1 \frac{K_d(\Sigma_0, \mathbf{A})}{2} \left(\sqrt{6 \log d/T} + 2\sqrt{1/T} \right).$$

Then on the same event as in Theorem J.1, we have $\|\hat{\mathbf{v}} - \mathbf{v}^*\|_1 \leq 4s_v \|\Sigma_0^{-1}\|_1 \lambda'$.

PROOF OF LEMMA J.3. We first start by showing that \mathbf{v}^* satisfies the constraint in the $\hat{\mathbf{v}}$ optimization problem with high probability. According to Theorem J.1, we have with probability not smaller than $1 - 14d^{-1}$:

$$\begin{aligned} \|\mathbf{v}^{*T} \mathbf{S}_0 - \mathbf{e}_1\|_\infty &= \|\mathbf{v}^{*T} (\mathbf{S}_0 - \Sigma_0)\|_\infty \leq \|\mathbf{v}^*\|_1 \|\mathbf{S}_0 - \Sigma_0\|_{\max} \\ &\leq \|\mathbf{v}^*\|_1 \frac{K_d(\Sigma_0, \mathbf{A})}{2} \left(\sqrt{6 \log d/T} + 2\sqrt{1/T} \right) \leq \lambda'. \end{aligned}$$

This implies that $\|\hat{\mathbf{v}}\|_1 \leq \|\mathbf{v}^*\|_1 \leq \|\Sigma_0^{-1}\|_1$, and hence similarly to (F.3) in Lemma F.4 in the Supplementary Material we can conclude:

$$(J.2) \quad \|\hat{\mathbf{v}}_{S_v^c} - \mathbf{v}_{S_v^c}^*\|_1 \leq \|\hat{\mathbf{v}}_{S_v} - \mathbf{v}_{S_v}^*\|_1$$

Next we control $\|\hat{\mathbf{v}} - \mathbf{v}^*\|_\infty$. We have:

$$\begin{aligned} \|\hat{\mathbf{v}} - \mathbf{v}^*\|_\infty &= \|(\hat{\mathbf{v}}^T \Sigma_0 - \mathbf{e}_1) \Sigma_0^{-1}\|_\infty \leq \|\Sigma_0^{-1}\|_1 (\|\hat{\mathbf{v}}^T \Sigma_0 - \mathbf{e}_1\|_\infty + \|\hat{\mathbf{v}}\|_1 \|\Sigma_0 - \Sigma_0\|_{\max}) \\ &\leq \|\Sigma_0^{-1}\|_1 2\lambda'. \end{aligned}$$

Combining the last bound with (J.2), we get:

$$\|\hat{\mathbf{v}} - \mathbf{v}^*\|_1 \leq 4s_v \|\Sigma_0^{-1}\|_1 \lambda',$$

which is what we wanted to show. \square

REFERENCES

- ADAMCZAK, R. and WOLFF, P. (2015). Concentration inequalities for non-lipschitz functions with bounded derivatives of higher order. *Probability Theory and Related Fields* **162** 531–586.
- BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. B. (2009). Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics* 1705–1732.
- FANG, K.-T., KOTZ, S. and NG, K. W. (1990). *Symmetric multivariate and related distributions*. Chapman and Hall.
- GAUTIER, E. and TSYBAKOV, A. (2011). High-dimensional instrumental variables regression and confidence sets. *arXiv preprint arXiv:1105.2454*.
- HALL, P. and HEYDE, C. C. (1980). *Martingale limit theory and its application*. Academic press New York.
- HAN, F., LU, H. and LIU, H. (2014). A direct estimation of high dimensional a direct estimation of high dimensional stationary vector autoregressions. *arXiv preprint arXiv:1307.0293v3*.
- HOEFFDING, W. (1948). A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics* 293–325.
- HOEFFDING, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American statistical association* **58** 13–30.
- LIU, H., HAN, F., YUAN, M., LAFFERTY, J., WASSERMAN, L. ET AL. (2012a). High-dimensional semiparametric gaussian copula graphical models. *The Annals of Statistics* **40** 2293–2326.
- LIU, H., HAN, F. and ZHANG, C.-H. (2012b). Transelliptical graphical models. In *Advances in Neural Information Processing Systems*.
- VAN DE GEER, S., BÜHLMANN, P. and RITOV, Y. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics* **42** 1166–1202.

- VERSHYNIN, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027* .
- ZHANG, C.-H. and ZHANG, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **76** 217–242.