Robust Inference of a LASSO-type Estimator for Correlated Factors *

Chuanping Sun †

Abstract

Using LASSO related methods to solve high-dimensional financial problems has become the new mainstream in recent years thanks to efficient computational algorithms and their capacity in dealing with very large dimensions. Nonetheless, empirical work has demonstrated strong evidence that high dimensional datasets are usually highly correlated and unfortunately the LASSO estimator performs poorly under such circumstance.[1] Subsequently, we consider the correlation-robust Ordered-Weighted-LASSO (OWL) estimator (Figueiredo and Nowak, 2016) which is structure-free and data-driven. This paper extends the theoretical work of Sun (2019) and focuses on developing its asymptotic properties under *less restrictive* assumptions (i.e., mixing condition and fatter tails) on random variables. Then we further develop the *de-biased* version of the OWL estimator and show that it is asymptotically normally distributed. Using simulated data, we find that the de-biased OWL estimator can greatly reduce the estimation error of the OWL estimator for various sample sizes, while the true value of parameter is included in the 95% confidence interval with satisfying coverage rate. Empirically, we apply the de-biased OWL estimator on factor investing using 15 large stocks in the Dow Jones industrial average index.

JEL classification: C52, C55, C58, G11

Keywords: LASSO, Inference, Weak Dependence, De-biased estimator, Factor Investing, Factor Zoo

*Preliminary draft.

†School of Economics and Finance, Queen Mary University of London, chuanping.sun@qmul.ac.uk

[1]For a detailed discussion, see Figueiredo and Nowak (2016); Zou and Hastie (2005); Asness et al. (2013); Kleibergen (2009).

1 Introduction

Economic and financial research topics related to the LASSO (Tibshirani, 1996) estimator have burgeoned and evolved rapidly in the past decade as high-dimensional big datasets become more available. For some examples, see Feng et al. (2020), Freyberger et al. (2019), Kozak et al. (2020) among others. However, as pointed out by Babii et al. (2019): *["...the bulk of machine learning methods assume i.i.d. regressors and residuals."]*. They further argue that time series data are usually correlated and, as a remedy, they utilize a structured group-LASSO estimator using mixed frequency time series data.[2] Nonetheless, empirical evidence has suggested that correlations are also commonly observed in the cross-sectional dimension,[3] yet we often encounter insufficient information to impose structural restrictions on cross-sectional covariates. Consequently, it is not straightforward to implement the group-LASSO method while the cross-sectional dimension being large and potentially highly correlated. Conversely, we resort to a newly developed machine learning tool, the Ordered-Weighted-LASSO (OWL) estimator, which is structure-free (needless to define group structures ex ante) and entirely data-driven to exploit cross-sectional correlations. Figueiredo and Nowak (2016) demonstrated that the OWL estimator explicitly permits correlations among covariates and achieves correlation identification and sparsity shrinkage simultaneously. Sun (2019) further established the consistency property of the OWL estimator under i.i.d Gaussian assumptions and applied the OWL estimator to dissect the factor zoo.

This paper focuses on developing robust inference of the OWL estimator under more general conditions. First, we relax the usual i.i.d. assumption for regressors and instead impose less restrictive weak dependence conditions among high dimensional covariates before we derive the non-asymptotic bounds for the prediction error and the parameter estimation error. In particular, we assume $\alpha-$mixing conditions and potentially fatter (than sub-Gaussian) tails on variables and their distributions. We leave a free parameter q that

[2] In particular, each group consists of lagged values of either the dependent variable or a single explanatory variable, which means in effect, correlations on the time-series dimension are all retained in separate groups.

[3] Asness et al. (2013) find negative correlation between value and momentum factors which can be utilized to achieve superior portfolio performance. Kleibergen (2009) cautions about the collinearity between factor loadings when implementing a Fama-MacBeth regression.

controls the fatness of the tail distribution and we derive the probability measure of the validity of the oracle inequality in relation to q. Furthermore, we do not rely on an upper bound assumption for any random variable, which is usually required before implementing a Bernstein type inequality. Instead, we follow Dendramis et al. (2019) to truncate random variables at a level which will be specified later to bring together a refined bound for Bernstein type inequality under strong mixing conditions. In this respect, our theoretical framework requires much less restrictive assumptions and explicitly allows researchers to investigate cross-sectional correlations.

Second, following recent development of the de-sparsified LASSO estimator, for instance see Van De Geer et al. (2014), Belloni and Chernozhukov (2012), Kock (2016), Caner and Kock (2018), Kock and Tang (2019) among others, we extend Figueiredo and Nowak (2016) and Sun (2019) to develop the de-biased OWL estimator using the nodewise LASSO technique. The OWL estimator has appealing properties of grouping together highly correlated variables without pre-specifying any factor structures. Although Sun (2019) shows that the OWL estimator is consistent under some regularity conditions, it is biased in small samples. The de-biased OWL estimator bridges that gap. We show that after bias-correction, it is asymptotically normal and we derive the confidence intervals for each parameter.

Empirically, we apply the de-biased OWL estimator on 15 large stocks in the Dow Jones industrial average index with 80 factors constructed using accounting data. We implement a portfolio sorting method to obtain our factor zoo library.[4] It is worth stressing that we are not implementing a two-pass Fama-MacBeth type of regression or a stochastic discount factor (SDF) method[5] to identify true factors that drive asset prices, which are most commonly studied in the cross-sectional asset pricing literature. Instead, this exercise focuses on forecasting. We implement a simple one-pass time series regression to predict stock returns directly from lagged values of factors, which are high dimensional and potentially correlated.[6] We are interested in whether the de-biased OWL estimator can outperform

[4] In particular, we sort stocks (after removing micro-stocks) from the NYSE, NASDAQ and AMEX into decile portfolios according to a large number of firm characteristics at each point of time. For each characteristic we compute the spread returns between the top and bottom decile portfolios at each point of time.

[5] See Sun (2019) for an example of implementing the SDF method to find pervasive factors on the cross-section of stock returns.

[6] Nonetheless, the de-biased OWL estimator can also be implemented for the Fama-Macbeth regression

other benchmarks in an out-of-sample framework in terms of predicting asset returns given a set of test assets. Empirical evidence suggests that the de-biased OWL estimator yields higher out-of-sample Sharpe ratios compared to standard LASSO and OLS methods. In addition, the de-biased OWL estimator illustrates a clear pattern of time-varying nature of factor selections during different periods, while LASSO and OLS do not show strong evidence of such pattern.

This paper builds naturally on the active and expanding literature pertaining to the LASSO estimator, in both the machine learning and empirical asset pricing literature. Tibshirani (1996) proposes the LASSO estimator that achieves efficient dimension reduction within a convex optimization problem, which enjoys huge success. Since then voluminous research has evolved to broaden the scope of the LASSO estimator. Yuan and Lin (2006) allow covariates sharing similar characteristics to be grouped together as a unit and propose the group LASSO estimator that performs sparse selection among groups. Freyberger et al. (2019) apply the adaptive group LASSO method to find pervasive firm characteristics to predict stock returns while Babii et al. (2019) implement the group LASSO estimator with mixed-frequency time series data for nowcasting GDP growth. Belloni and Chernozhukov (2012) and Belloni et al. (2014) propose the three-pass double LASSO estimation method to de-bias LASSO coefficients of a set of factors that of primary interest to researchers. Feng et al. (2020) adopt the double LASSO selection procedure to "tame" the factor zoo. Zou and Hastie (2005) combine the ℓ_1 and ℓ_2 norm regularization and propose the elastic net (EN), which stabilizes LASSO coefficients especially when covariates exhibit correlations. Kozak et al. (2020) employ EN in a Bayesian framework and find that sparse components can largely explain the cross section of average returns. Bondell and Reich (2008) propose the octagonal shrinkage and clustering algorithm for regression (OSCAR) method by exploring the ℓ_∞ norm of parameters pair-wisely to achieve clustered selections when covariates are highly correlated. Zeng and Figueiredo (2015) and Figueiredo and Nowak (2016) promote the Ordered-Weighted-LASSO (OWL) estimator, which is closely related to the SLOPE (Sorted ℓ_1 Penalized Estimator) by Bogdan et al. (2015): both assign a fixed and decreasing weighting vector to penalized coefficients (by contrast, LASSO estimator assigns the same

(or SDF method) to identify pricing factors for a universe of stocks.

penalty to all coefficients), with the larger coefficients (absolute value) receiving larger penalty. Bogdan et al. (2015) continue to specify a normal CDF based (non-linear) design for the decreasingly ordered weighting vector, before using the false discovery rate (FDR) to infer significance in the multi-testing framework assuming i.i.d. covariates. On the other hand, the OWL estimator, although having the same design in the regularization as the SLOPE, differs substantially in the weighting vector specification. Figueiredo and Nowak (2016) specify a *linear* weighting vector, and they further find that, by adopting a linear weighting vector, the OWL estimator encompasses the OSCAR regularization, which has appealing properties to group together highly correlated variables without imposing any structural restrictions ex ante. Van De Geer et al. (2014) developed the de-sparsified LASSO estimator using the nodewise LASSO technique, which enables them to find a way to approximate the usually un-invertible scaled Gram matrix to identify and quantify the bias of the LASSO estimator. The de-sparsified LASSO estimator enjoys asymptotic normality. Kock (2016), Caner and Kock (2018) and Kock and Tang (2019) expand the de-sparsified LASSO estimator on panel data and develop statistical properties under sub-Gaussian assumption. Babii et al. (2019) extend the nodewise LASSO technique to group-LASSO estimator using mixed frequency time-series data. This paper marries the OWL estimator and the nodewise LASSO technique to propose the de-biased version of the OWL estimator. Meanwhile, this paper relaxes the usual i.i.d. and sub-Gaussian assumptions to derive (non)asymptotic properties of the estimator. In particular, we allow for weak dependence (α-mixing) between covariates and fatter (than sub-Gaussian) tails.

In the remainder of this paper, Section 2 outlines the OWL estimation framework and we study its (non)asymptotic properties and further discuss a de-biased version of the OWL estimator and its asymptotic normality property. Section 3 studies Monte Carlo experiments with various settings in dimensions and correlations. Section 4 applies the de-biased OWL estimator on 15 large stocks to find the best predictors from a factor zoo library constructed from accounting data.

2 Model

In this section, we define the Ordered-Weighted-LASSO (OWL) estimator and derive its theoretical properties under mixing and some other regularity assumptions. Then we develop the de-biased OWL estimator, and show that it has asymptotically normal distribution.

Notation

Throughout this paper, X is a $n \times p$ matrix, and y is a $n \times 1$ vector. We denote by $\hat{\Sigma} = \frac{1}{n}X'X$ the scaled Gram Matrix of X, while $\Sigma = E(\hat{\Sigma})$ is the expected (true) value of the scaled Gram matrix. For any $x, y \in R^n$, we denote $\|x\|_2 = (\sum_{i=1}^{n} x_i^2)^{1/2}$, $\|x\|_1 = \sum_{i=1}^{n} |x_i|$, $\|x\|_\infty = \max_{1\le i\le n} |x_i|$, $\|x\|_0$ the cardinality of x, and $x \odot y$ the Hadamard (pointwise) production of two vectors. For matrix $\mathbb{M} \in R^{n\times n}$, Λ_{min} and Λ_{max} denotes the smallest and largest eigenvalues of $\mathbb{M}$. For two sequences x_n and y_n, we write $x_n \asymp y_n$ if there exist $0 < a \le b < \infty$, such that $ay_n \le x_n \le by_n$ and we write $x_n \lesssim y_n$ if $x_n \le by_n$ for some $0 < b < \infty$. For any set s, s^c denotes the complimentary set. For two scalers p and q, $p \vee q := \max(p, q)$ and $p \wedge q := \min(p, q)$. For any $\beta = \{\beta_1, \cdots, \beta_p\} \in R^p$, we denote $|\beta|_\downarrow := (|\beta|_{[1]}, |\beta|_{[2]}, \cdots, |\beta|_{[p]})'$, where $|\beta|_{[1]} \ge |\beta|_{[2]} \ge \cdots \ge |\beta|_{[p]}$ and $|\beta|_{[j]}$ is the j^{th} element of $|\beta|_\downarrow$.

2.1 OWL estimator and the oracle inequality

Consider a linear model

$$y = X\beta^0 + \epsilon, \tag{1}$$

where $X := (X_1, \cdots, X_p)$ and $\beta^0 = (\beta_1^0, \cdots, \beta_p^0)'$. Note that in the high-dimensional case, we allow $p \gg n$ and $X_j's$ can be correlated for $j = 1, \cdots, p$. The OWL estimator $\hat{\beta}$ minimizes the objective function

$$\hat{\beta} = \arg\min_{\beta} \quad \left[\frac{1}{n}||y - X\beta||_2^2 + \frac{1}{n}\omega'|\beta|_\downarrow\right], \qquad \omega'|\beta|_\downarrow = \sum_{j=1}^{p} \omega_j |\beta|_{[j]}, \tag{2}$$

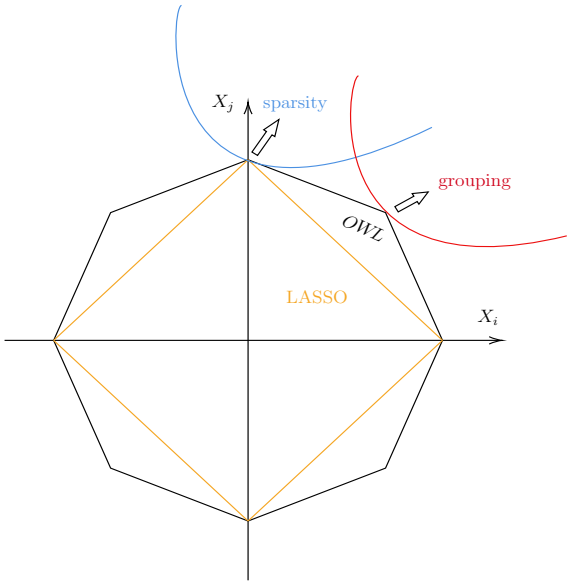

Figure 1. Geometric illustration for the atomic norm of OWL and LASSO penalty

where $\omega = (\omega_1, \cdots, \omega_p)'$, $\omega_j = \lambda_1 + \lambda_2(p-j)$, $j = 1, \cdots, p$ and $\lambda_1, \lambda_2 \geq 0$ are tuning parameters.

Zeng and Figueiredo (2015) have shown that the OWL estimator has sparsity selection and correlation identification properties. Figure 1 shows the geometric interpretation of the OWL penalty ($\omega'|\beta|_{\downarrow}$) and the LASSO penalty ($\|\beta\|_1$).[7] In particular, the tuning parameter λ_1 controls the overall level of penalty while λ_2 influences the grouping property: large (small) λ_2 encourages (discourages) correlated variables to be grouped together by assigning them with similar coefficients, see Sun (2019) and Figueiredo and Nowak (2016) for a detailed discussion. We want to stress here that we do not impose any factor structure restrictions in our model, for instance defining groups ex ante to encapsulate correlated variables. Correlation identification is entirely data-driven. On the other hand, the OWL penalty term encompasses the LASSO setup. Setting $\lambda_2 = 0$, the OWL estimator will collapse to the standard LASSO estimator. A gradient proximal algorithm can be implemented to solve the optimization problem in (2), see Sun (2019) for technical details.

Before we derive the statistical properties for the OWL estimator $\hat{\beta}$, we make the following assumptions which are the foundation for building the theoretical framework and

[7] This figure shows the atomic norm of the OWL penalty and the LASSO penalty and explains why the OWL estimator achieves Sparsity selection and correlation identification simultaneously, while LASSO only having the sparsity selection property.

add novelty to our contributions. Assumption 1 states restriction on random variables, including cross-sectional dependence and on tails of their distributions. Assumption 2 is a standard requirement for developing asymptotic theory for LASSO type estimators in high dimensions. Assumption 3 specifies some rates on s, n and p required to obtain consistent estimators.

Assumption 1 (Random variables, Dendramis et al. (2019))**.**

(a) *For all* $j = 1, \cdots, p$, $\{X_{i,j}\}_{i=1}^{n}$ *and* $\{X_{i,j}\epsilon_i\}_{i=1}^{n}$ *are* $\alpha-$*mixing sequences, which are not necessarily stationary. The mixing coefficients have property* $\alpha_k \leq c\phi^k$, $c > 0$, $0 < \phi < 1$, $k \geq 1$*;*

(b) $\sup_{i,j} \mathbb{P}(|X_{i,j}| > a) \leq c_1 \exp[-c_2 a^{q_1}]$ *and* $\sup_i \mathbb{P}(|\epsilon_i| > a) \leq c_1 \exp[-c_2 a^{q_2}]$ *for all* $a > 0$*, for some* $q_1, q_2 > 0$ *and* $c_1, c_2 > 0$ *which do not depend on* a, i, j*;*

(c) $\mathrm{E}(\epsilon_i | X_{i,j}) = 0$, *and* $\max_{i,j} \mathrm{E}(X_{i,j}^4) < \infty$.

Assumption 1(a) relaxes the i.i.d condition which is usually assumed on X_j in the bulk of LASSO related literature, for instance see Kock (2016), Van De Geer et al. (2014) and Belloni and Chernozhukov (2012). Instead, we allow variables X_j to be weakly dependent, i.e. $\alpha-$mixing. Furthermore, mixing condition permits heteroscedasticity which is typically exhibited in empirical data. Assumption 1(b) further specifies tail bounds of distributions of X_j and ϵ. Although we use an exponential type of bound, it allows tails to be fatter than in the sub-Gaussian case. The tail parameter q controls the fatness of the tails, and it encompasses the sub-Gaussian tail ($q = 2$) as a special case. Assumption 1(c) is a standard assumption stating that the error term is orthogonal to covariates, in other words $\{X_{i,j}\epsilon_i\}$ is a zero mean sequence. Note that we do not assume random variables to be bounded which is typically assumed when implementing a Bernstein type inequalities. To this end, our assumptions are more general and less restrictive than many of those in the literature which typically consider sub-Gaussian i.i.d. random variables.

Assumption 2 (Restricted eigenvalue condition on $\hat{\Sigma}$, Bickel et al. (2009))**.**
Let $s_0 \subset \{1, \cdots, p\}$ *be a subset and* $s := |s_0|$ *the cardinality of* s_0*. For* $\beta = \{\beta_1, \cdots, \beta_p\}$*, denote* $\beta_{s_0} := \beta_i \mathbf{1}\{i \in s_0, i = 1, \cdots, p\}$, $\beta_{s_0^c} := \beta_i \mathbf{1}\{i \notin s_0, i = 1, \cdots, p\}$,

so that $\beta = \beta_{s_0} + \beta_{s_0^c}$. *We suppose that for all* β *such that* $\|\beta_{s_0^c}\|_1 \leq 3\|\beta_{s_0}\|_1$, $\hat{\Sigma}$ *satisfies the restricted eigenvalue condition*

$$\phi_0^2 = \min_{\substack{s_0 \subset \{1,\cdots,p\} \\ s<p}} \min_{\substack{\beta \in R^p \setminus \{0\} \\ \|\beta_{s_0^c}\|_1 \leq 3\|\beta_{s_0}\|_1}} \frac{\beta' \hat{\Sigma} \beta}{\|\beta_{s_0}\|_2^2} > 0. \tag{3}$$

Assumption 2 is a cornerstone to many theoretical results related to LASSO estimation. First of all, it allows to specify the approximate sparsity condition as follows: only for a subset s_0, the true parameter vector has non-zero values ($\beta_i^0 \neq 0 : \forall i \in s_0$), while the complement contains only zeros ($\beta_i^0 = 0 : \forall i \notin s_0$). The cardinality s of such subset s_0 does not need to be known ex ante nor its elements, though we restrict it so that $s \ll p$. The restricted eigenvalue condition implies the compatibility condition of Buhlmann and Van de Geer (2011) (see below Lemma 1), which is an essential element in the proof of Theorem 2.1.

Lemma 1 (Compatibility condition for $\hat{\Sigma}$, Buhlmann and Van de Geer (2011))**.** *If the scaled Gram matrix* $\hat{\Sigma}$ *satisfies the restricted eigenvalue condition in* (3)*, then for any* β

$$\|\beta_{s_0}\|_1^2 \leq (\beta' \hat{\Sigma} \beta) s / \phi_0^2.$$

Proof: see Appendix A.4

Assumption 3 (Rates on n, p and s)**.** *Denote by* $s := |s_0|$ *the sparsity parameter indicating the number of non-zero elements in* $\hat{\beta}$ *as in* (2) *and* s_j *the sparsity parameter in* (15) *by regressing the* j^{th} *column of* X *on the remaining columns of* X*. For any* $j \in \{1, \cdots, p\}$*, we assume*

(a) $(s \vee s_j)\sqrt{\dfrac{\log p}{n}} = o(1)$,

(b) $s_j \sqrt{\dfrac{\log^2 p}{n}} = o(1)$.

Assumption 3 specifies some rates on n, p, s and s_j which lead to consistent estimators. The rate that is required in 3(a) is rather standard and similar to that used in Kock (2016) and Van De Geer et al. (2014). The other requirement in 3(b) is typically weaker than in Kock (2016).

2.1.1 Statistical properties

Next, we investigate some statistical properties for the OWL estimator. Theorem 2.1 establishes oracle inequality for the prediction error and parameter estimation error. The probability we obtained is based on the assumption of weak dependence. Its proof uses Bernstein type inequalities for $\alpha-$mixing variables obtained in Dendramis et al. (2019).

Theorem 2.1 (Oracle inequality). *Suppose Assumption 1 and 2 hold. Set* $\lambda_0 = \kappa\sqrt{\frac{\log p}{n}}$, *where* κ *is a positive constant. Let* $\frac{\lambda_1}{n} = 2\lambda_0$ *and assume* $\frac{\lambda_2}{n} = O_p(\frac{s\log p}{np})$. *Suppose that for some* $\delta > 0$, $p \lesssim n^{\delta}$.

1. *Let* $n, p \to \infty$. *Then for sufficiently large* κ,

$$\frac{1}{n}||X(\hat{\beta} - \beta^0)||_2 \lesssim 4\lambda_0\sqrt{s}/\phi_0 + \lambda_0\sqrt{2s\|\beta^0\|_1} \tag{4}$$

$$||\hat{\beta} - \beta^0||_1 \lesssim 8\lambda_0 s/\phi_0^2 + \lambda_0 s\|\beta^0\|_1, \tag{5}$$

   with probability at least $1 - c_0' p^{-\epsilon} \to 1$, *for some* $\epsilon > 0$, *where* c_0' *is a positive constant which is independent on* n *and* p.

2. *Let* p *be bounded. Then* (4) *and* (5) *hold with probability at least*

$$1 - pc_0\left[\exp(-\frac{c_1'}{4}\kappa^2\log p) + \exp\left(-c_2'\left(\frac{\kappa\sqrt{n\log p}}{2\log^2 n}\right)^{\zeta}\right)\right], \tag{6}$$

   where $\zeta = q/(q+1), q = q_1q_2/(q_1+q_2)$ *and* c_0, c_1', c_2' *are some positive constants which are independent on* n *and* p.

Proof: see Appendix A.1.

Remark 1 Theorem 2.1 offers bounds for the prediction error $||X(\hat{\beta} - \beta^0)||_2/n$ and parameter estimation error $||\hat{\beta} - \beta^0||_1$ for the OWL estimator under strong mixing conditions. Once we further incorporate Assumption 3, we will derive consistency and the convergence rate for the OWL estimator. See Corollary 2.1.

Remark 2 We analyze the probability of (4) and (5) to hold under two scenarios. First, when $n, p \to \infty$, we find that those inequalities hold with probability tending to one once

a sufficiently large κ is chosen. Second, when p is fixed, we find that the probability of (4) and (5) to hold converges to $1 - pc_0 \exp(-c_1'' \kappa^2 \log p)$ as $n \to \infty$, where c_0 and $c_1'' = c_1'/4$ are some positive constants which depend only on the mixing coefficient α_k in Assumption 1. Then we need to select κ sufficiently large to ensure that $pc_0 \exp(-c_1'' \kappa^2 \log p)$ is close to zero.

Remark 3 Our results on the probability measures are obtained under general assumption of exponential decaying tails on random variable $z_{i,j} := X_{i,j}\epsilon_i$. If $q_1, q_2 = 2$, equation (6) encompasses the sub-Gaussian case, which is a popular assumption in related literature, see Kock (2016) and Kock and Tang (2019) for example. In addition, it also accommodates for fatter tails, i.e. $0 < q_1, q_2 < 2$. However, when both p and n are bounded, the probability of (4) and (5) to hold depends also on the tail parameters. The thinner is the tail of the distribution of the random variable $z_{i,j}$ (i.e., large q, where $q = q_1 q_2/(q_1 + q_2)$), the closer of the probability in (6) is to one.

To this end, we want to emphasize that our results in Theorem 2.1 are based on less restrictive assumptions, where we allow for weak dependence between random variables $X_j's$ and we further relax the sub-Gaussian tail restriction where we leave a parameter q that controls the fatness of the tail distributions.

Corollary 2.1 (Convergence rate)**.** *Suppose Assumption 3 is satisfied and assume $n, p \to \infty$. Then for sufficiently large κ, with probability tending to one,*

$$\|\hat{\beta} - \beta^0\|_2 = O_p\left(\sqrt{\frac{s \log p}{n}}\right) = o_p(1), \qquad \|\hat{\beta} - \beta^0\|_1 = O_p\left(s\sqrt{\frac{\log p}{n}}\right) = o_p(1). \tag{7}$$

Proof: see Appendix A.2.

Corollary 2.1 establishes the convergence rate in ℓ_1 and ℓ_2 norm of the OWL estimator $\hat{\beta}$. After specifying some growth rate for n and p in Assumption 3, we show that the OWL estimator is consistent.

2.1.2 Choice of penalty parameters

It is well recognized that the choice of penalty level has huge impact on the performance of LASSO type estimators. In the machine learning literature, cross-validation is the

most commonly implemented method for choosing penalty parameters. However, cross-validation can be computationally expensive to implement, for instance, in a recursively estimated application.[8] Hence, it would be useful if we can infer an appropriate penalty level based on the statistical properties of the estimator. Belloni and Chernozhukov (2012) argue that we should choose a penalty level that is sufficiently large to cancel noises coming from estimation errors (i.e. $\mathbb{P}(\lambda_0 > 2\|X'\epsilon\|_\infty/n)$ is large), yet not overly large to write off signals from variables. To achieve that, we propose the rule of thumb about penalty choice below based on a similar argument to Belloni et al. (2012) but incorporating our unique setting for random variables (weak dependence and exponential tails).

Proposition 2.1. *Let Assumption 1 be satisfied, $\Phi^{-1}(\cdot)$ denote the inverse of the standard normal distribution function. We propose the following values for turning parameters λ_1 and λ_2 in (2).*

$$\frac{\lambda_1}{n} = \frac{4}{\sqrt{n}}\sigma^*(1+\frac{1}{\log n})^{1/2}\Phi^{-1}(1-\frac{\alpha}{2p}), \qquad \frac{\lambda_2}{n} = \frac{\lambda_1}{n}\frac{\sqrt{\log p}}{\sqrt{n}\,p}, \tag{8}$$

where we evaluate σ^* recursively similar to Algorithm A.1 in Belloni et al. (2012) and α is a significance level.

Proof: see Appendix A.6.

Note that α is selected to ensure the probability that the penalty is large enough to cancel out noises is close to one, that is $\mathbb{P}(\lambda_0 > 2\|X'\epsilon\|_\infty/n) \geq 1-\alpha$. So a smaller value of α will result in larger penalty level. Proposition 2.1 offers a guideline for penalty choices when cross-validation is too expensive to implement. Equation (8) suggests that the penalty level depends on four elements. First, the noise level σ^* affects penalty level. Large variance of the error term requires a higher penalty level to cancel out noises. We evaluate σ^* recursively: we first evaluate the model and obtain the residuals while setting $\sigma^* = 1$, then update σ^* with the empirical residual variance and re-evaluate the model.

[8] Taking the commonly used 10-fold cross-validation as an example, at each step of the recursive exercise (for instance, a rolling window estimation procedure), we need to split the sample into 10 folds, while holding one tenth of the sample as testing sample and the remaining as estimation sample to evaluate and test the model, then swap positions of testing/estimation samples to re-evaluate the model (10 times). Suppose we have two tuning parameters and we want to search for a best fit in a 5×5 grid, and suppose the rolling window requires T recursive estimations. Then the 10-fold cross-validation method would require to run the model $5^2 * 10 * T$ times.

Second, large n reduces the penalty level. Note that the total penalty is determined by λ_1/n and λ_2/n in (2), so large n commands smaller values for λ_1/n and λ_2/n. From a different perspective, we can view that large n leads to smaller variance σ^2, which requires less penalty on parameters. Third, the dimension of covariates p dictates the optimal penalty level. Large p requires higher level of penalty to shrink off more irrelevant variables. Fourth, the significant parameter α.

2.2 De-biased OWL estimator

Although Theorem 2.1 shows that the OWL estimator is consistent under some regularity conditions, it is biased in small samples. In this section, we discuss a bias-corrected version of the OWL estimator using the nodewise LASSO method introduced in Van De Geer et al. (2014). Then we develop the asymptotic normal approximation result for the de-biased OWL estimator.

2.2.1 Identifying the bias of the OWL estimator

For the convenience of expression, the OWL estimator defined in (2) can be written as

$$\hat{\beta} = \arg\min_{\beta} \left[\|y - X\beta\|_2^2/n + 2\omega'|\beta|_{\downarrow}/n \right], \tag{9}$$

where we extract 2 out of the weighting vector ω.[9] The first order condition of minimization of (9) gives

$$-X'(y - X\hat{\beta})/n + \omega \odot \hat{\tau}/n = 0, \qquad \hat{\tau} = \begin{cases} 1 & \text{if} \quad \hat{\beta} > 0 \\ [-1, 1] & \text{if} \quad \hat{\beta} = 0 \\ -1 & \text{if} \quad \hat{\beta} < 0. \end{cases} \tag{10}$$

where $\odot$ denotes point-wise product of two vectors, and $\hat{\tau}$ is the definition of sub-gradient of $|\hat{\beta}|_{\downarrow}$. We further utilize the equality $y = X\beta^0 + \epsilon$ and $\hat{\Sigma} = X'X/n$. Then (10) can be

[9] Note that ω is exactly pinned down by λ_1 and λ_2 which can be determined according to (8). So for the convenience of expression, we keep the same notation here for ω.

written as

$$\hat{\Sigma}(\hat{\beta} - \beta^0) + \omega \odot \hat{\tau}/n = X'\epsilon/n. \tag{11}$$

Since $\hat{\Sigma}$ is not invertible when $p > n$, we are using a relaxed form $\hat{\Theta}$ suggested by Van De Geer et al. (2014) to approximate the unobservable Σ^{-1}, where Σ is the population value of $\hat{\Sigma}$. Suppose such $\hat{\Theta}$ exists. Then we can write

$$\hat{\beta} - \beta^0 + \hat{\Theta}\omega \odot \hat{\tau}/n = \hat{\Theta}X'\epsilon/n - \Delta/\sqrt{n}, \tag{12}$$

$$\Delta = \sqrt{n}(\hat{\Theta}\hat{\Sigma} - I)(\hat{\beta} - \beta^0), \tag{13}$$

where we will show later that $\hat{\Theta}X'\epsilon/n$ is asymptotically normal and the approximation error, Δ, is negligible. Then we obtain the de-biased OWL estimator

$$\hat{b} = \hat{\beta} + \hat{\Theta}\omega \odot \hat{\tau}/n = \hat{\beta} + \hat{\Theta}X'(Y - X\hat{\beta})/n, \tag{14}$$

where the second equation holds in view of (10). So the bias is identified as $\widehat{bias} = \hat{\Theta}\omega \odot \hat{\tau}/n = \hat{\Theta}X'(Y - X\hat{\beta})/n$. In the next subsection, we construct required approximation $\hat{\Theta}$.

2.2.2 Construction of $\hat{\Theta}$

We follow Van De Geer et al. (2014) and Kock (2016) and use the nodewise LASSO technique to obtain $\hat{\Theta}$. First, the nodewise LASSO estimator is defined as

$$\hat{\gamma}_j = \underset{\gamma \in R^{p-1}}{\arg\min} \left(\|X_j - X_{-j}\gamma\|_2^2/n + 2\lambda_j \|\gamma_j\|_1 \right), \tag{15}$$

where $\hat{\gamma}_j := \{\hat{\gamma}_{j,k} : j, k = 1, \cdots, p, k \neq j\} \in R^{p-1}$ is a row vector of the nodewise LASSO estimator by regressing X_j (the j^{th} column of matrix X) on X_{-j} (which denotes the remaining columns of X) with LASSO penalty λ_j. Define a $p \times p$ matrix $\hat{C}$ and a $p \times p$

diagonal matrix $\hat{T}^2$ as

$$\hat{C} := \begin{pmatrix} 1 & -\hat{\gamma}_{1,2} & \cdots & -\hat{\gamma}_{1,p} \\ -\hat{\gamma}_{2,1} & 1 & \cdots & -\hat{\gamma}_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ -\hat{\gamma}_{p,1} & -\hat{\gamma}_{p,2} & \cdots & 1 \end{pmatrix}, \qquad \hat{T}^2 := \text{diag}(\hat{\delta}_1^2, \hat{\delta}_2^2, \cdots, \hat{\delta}_p^2,), \tag{16}$$

where for $j = 1, \cdots, p$,

$$\hat{\delta}_j^2 = \|X_j - X_{-j}\hat{\gamma}_j\|_2^2/n + \lambda\|\hat{\gamma}_j\|_1. \tag{17}$$

Then $\hat{\Theta}$ is constructed by setting

$$\hat{\Theta} := \hat{T}^{-2}\hat{C}. \tag{18}$$

For a close consideration of whether $\hat{\Theta}$ is a good approximation of Σ^{-1}, see Appendix A.5.

2.2.3 Inference on the de-biased OWL estimator

Denote $\Sigma_{X\epsilon} := \text{E}[\frac{1}{n}\sum_{i=1}^{n}(X_i'\epsilon_i)(X_i'\epsilon_i)']$, $\hat{\Sigma}_{X\epsilon} := \frac{1}{n}\sum_{i=1}^{n}[(X_i'\hat{\epsilon}_i)(X_i'\hat{\epsilon}_i)']$ and $\Theta := \Sigma^{-1}$. For any $l \in \{1, ..., p\}$, let $\hat{\Theta}_l$ (Θ_l) be the l^{th} row of the $\hat{\Theta}$ (Θ) matrix, written as a column vector.

Theorem 2.2. *Let $\hat{b}$ and $\hat{\Theta}$ be defined as in (14) and (18), respectively. Then the following hold:*

$$\sqrt{n}(\hat{b} - \beta^0) = \hat{\Theta}X'\epsilon/\sqrt{n} + o_p(1), \tag{19}$$

$$\hat{\Theta}_l'X'\epsilon/\sqrt{n} \to \mathbb{N}(0, \Theta_l'\Sigma_{X\epsilon}\Theta_l), \tag{20}$$

Furthermore, a uniformly valid point-wise confidence interval based on the t-statistics for β_l^0 where $l = 1, \cdots, p$ is given by

$$[\hat{b}_l - C(\alpha, \hat{\Theta}_l, \hat{\Sigma}_{X\epsilon}), \quad \hat{b}_l + C(\alpha, \hat{\Theta}_l, \hat{\Sigma}_{X\epsilon})], \tag{21}$$

where $C(\alpha, \hat{\Theta}_l, \hat{\Sigma}_{X\epsilon}) = \Phi^{-1}(1 - \alpha/2)\sqrt{\hat{\Theta}_l'\hat{\Sigma}_{X\epsilon}\hat{\Theta}_l/n}$ and α is the confidence level.
Proof: see Appendix A.3.

Theorem 2.2 arrives at the asymptotic normality property for the de-biased OWL es-

timator $\hat{b}$ and allows uniformly valid test for β_l^0 (i.e. the confidence interval applies to all $l = 1, \cdots, p$). The confidence interval is derived through the t-statistics based on the asymptotically normal property of the de-biased OWL estimator $\hat{b}$. Alternatively, a related Wald test can be subsequently developed. However, in this paper, we focus on the t-statistics and using (21) for testing the significance of the de-biased OWL estimator in our empirical exercises (i.e. we exclude covariates from a set that only contains useful covariates if their estimated coefficients are tested not significantly different from zero).

Next, we investigate the performance of the de-biased OWL estimator using simulated data.

3 Simulation

This section reports results on the performance of the de-biased OWL estimator alongside other benchmark estimations using simulated data. First, let us consider a toy example of 300 test assets ($N = 300$) and 20 covariates ($K = 20$). The oracle (true) values of the first six coefficient parameters of covariates are non-zeros and the rest are all zeros. Specifically, we set $\beta_0 = \{10, \frac{10}{2}, \frac{10}{3}, ..., \frac{10}{6}, 0, 0, ..., 0\} \in R^{20}$. Variables are not correlated.

Figure 2 displays the plots of estimated coefficients using various methods, alongside the true values ($b0$, blue line). The shaded area is the 95% confidence interval for the de-biased OWL estimator. First of all, we find the OWL estimator (red/circle) exhibits good sparse-selection property: it shrinks the coefficients of all useless factors to zeros. Meanwhile, we also find that the OWL estimates for the non-zero coefficients are all biased towards zero, which is a common pitfall of many LASSO related estimators in small samples. On the other hand, we find that the de-biased OWL estimator (yellow/asterisk) *corrects* the bias: the bias-corrected estimates are much closer to the oracle values (blue line), with the oracle values lying inside the confidence interval (shaded area). On the flip side, the de-biased OWL estimates lose the sparse-selection property: all those useless factors now have non-zero coefficients using the de-biased OWL estimator. However, this incorrect de-biasing is bounded by the confidence intervals. We find that the true values (zeros) of the coefficients of these useless factors lie inside the confidence interval. Hence, we can easily remove those useless factors by running a t-test. This simple toy example illustrates the nice properties

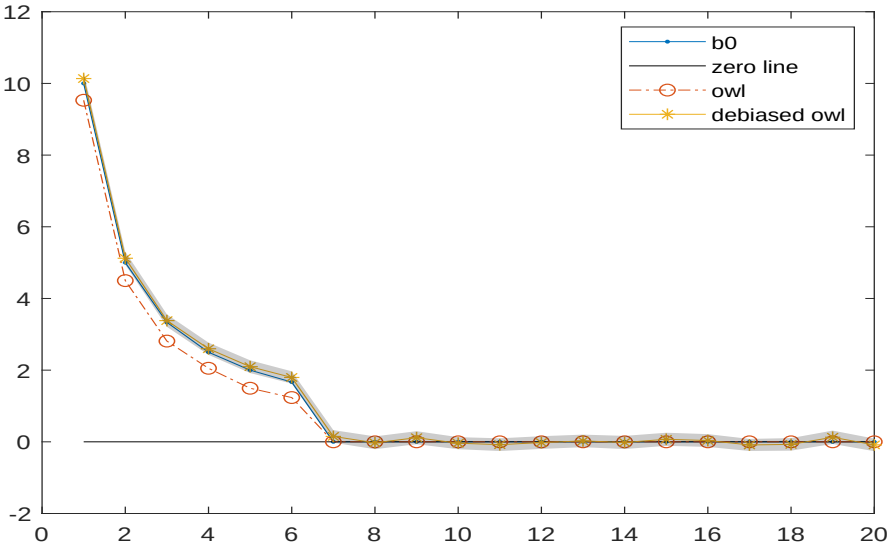

Figure 2. A toy example

This graph plots the estimated coefficients using OWL estimator and its de-biased version, along side with the true values ($b0$, blue line). There are total 20 covariates, the first six (true value) are non-zeros, while the remaining are zeros. The shaded area is confidence interval for de-biased OWL estimator. Variables are uncorrelated.

of the de-biased OWL estimator. Next, we run a sequences of Monte Carlo experiments to investigate how dimensions of data-set, correlations and other aspects would affect the performance of the de-biased OWL estimation.

We set the dimension of covariates X such that $K = dim(X) \in \{100, 1000\}$ and the number of observations $N \in \{60, 800, 1000\}$. We allow covariates in X to be correlated, and their covariance structure is defined as

$$\text{Corr}_{i,j}(X) = \Sigma_{i,j}(\rho) = \rho^{|i-j|}, \qquad i, j \in \{1, 2, ..., K\}, \qquad \rho \in \{0, 0.3, 0.5, 0.7\},$$

where $\text{Corr}_{i,j}$ is the element in the i^{th} row and j^{th} column of a correlation coefficient matrix. The true oracle value for β is set to be

$$\beta_0 = \{10, \frac{10}{2}, \frac{10}{3}, ..., \frac{10}{6}, 0, 0, ..., 0\} \in R^K.$$

The first six elements are non-zeros, and the rest are zeros. The covariates matrix X and

the response y are generated through the following distribution

$$X = Z * chol(\Sigma), \quad Z \sim \mathbf{N}(0,1) \in R^{N\times K},$$
$$y = X\beta_0 + \epsilon, \quad \epsilon \sim \mathbf{N}(0, 0.01) \in R^{N\times 1},$$

where $chol(.)$ is the lower triangle matrix of the Cholesky decomposition. We use the de-biased OWL estimator to obtain estimated coefficients. The penalty hyper-parameters of λ_1 and λ_2 are chosen according to the optimal level discussed in Proposition 2.1 in Section 2.1.2.

$$\lambda_1/N = \tilde{\sigma}(1 + \frac{1}{\log N})^{1/2}\Phi^{-1}(1 - \frac{\alpha}{2K})/\sqrt{N},$$
$$\lambda_2/N = (\lambda_1/N)\sqrt{\log K}/(\sqrt{N}K),$$

where $\Phi^{-1}(\cdot)$ is the inverse of a normal cumulative distribution function and $\alpha = 5\%$. We set $\tilde{\sigma} = 4\sigma^* = 0.01$ to gain computational speed.[10] We compare the de-biased OWL estimator with other benchmarks, including the OLS (when it is feasible) and the LASSO estimators. The number of the Monte Carlo repetition is 500 ($rep = 500$) for all set-ups. We report four estimated coefficients of $\hat{\beta}$, of which two have the true value of non-zeros: $\{\beta_3, \beta_6\}$, the other two have true values of zeros: $\{\beta_{12}, \beta_{20}\}$. We report the performance of $\hat{\beta}$ in Table 1 using the following criteria:

1. Coverage rate for de-biased OWL. We compute the confidence interval of de-biased OWL according to (21). The coverage rate is the rate of the true value of the parameter included in the confidence interval throughout all Monte Carlo repetitions. We compute the coverage rate for each of these four parameters.

2. The width of confidence intervals (CI) for the de-biased OWL estimates. We compute the average width of confidence intervals of de-biased OWL throughout all Monte Carlo repetitions.

[10] We opt to this easy choice of σ^* to gain computation speed, especially in high-dimensional cases. The de-biased OWL estimates may be sub-optimal, and a carefully cross-validated choice of σ^* can potentially improve the de-biased OWL estimates.

3. MAE (Mean Absolute Errors). We compare the mean absolute estimation errors between the de-biased OWL, LASSO and OLS estimates. The MAE for each coefficient $j \in \{3, 6, 12, 20\}$ is defined as $\text{MAE}^j_{benchmark} = \sum_{i=1}^{rep} |\beta^i_{j,0} - \hat{\beta}^{benchmark,i}_j|/rep$, and the average MAE across all coefficients of $j \in \{3, 6, 12, 20\}$ for each benchmark is defined as $\text{MAE}_{benchmark} = \sum_{i=1}^{rep} \sum_j |\beta^i_{j,0} - \hat{\beta}^{benchmark,i}_j|/(4rep)$.

Table 1. Simulation result

Panel A: Coverage rate, CI width and MAE comparison between benchmarks

	Coverage rate of dowl				Width of CI of dowl				Average MAE			
	β_3	β_6	β_{12}	β_{20}	β_3	β_6	β_{12}	β_{20}	dowl	ols	lasso	lasso_cv
K = 50, N = 60												
$\rho = 0$	0.9360	0.9350	0.9600	0.9360	0.1016	0.0665	0.0820	0.0942	0.0112	0.0263	0.0819	0.0819
$\rho = 0.3$	0.9560	0.9300	0.9280	0.9320	0.1316	0.0948	0.1138	0.1424	0.0143	0.0347	0.0657	0.0657
$\rho = 0.5$	0.9560	0.9320	0.9420	0.9560	0.1209	0.1271	0.2372	0.1142	0.0154	0.0396	0.0894	0.0894
$\rho = 0.7$	0.9780	0.9780	0.9620	0.9500	0.1857	0.1782	0.1897	0.1504	0.0185	0.0495	0.0663	0.0663
K = 50, N = 1000												
$\rho = 0$	0.9420	0.9480	0.9380	0.9600	0.0129	0.0123	0.0127	0.0121	0.0015	0.0026	0.0689	0.0689
$\rho = 0.3$	0.9480	0.9600	0.9480	0.9540	0.0139	0.0139	0.0137	0.0137	0.0016	0.0028	0.0813	0.0813
$\rho = 0.5$	0.9640	0.9380	0.9280	0.9520	0.0158	0.0170	0.0162	0.0161	0.0019	0.0033	0.0758	0.0758
$\rho = 0.7$	0.9380	0.9600	0.9420	0.9400	0.0214	0.0207	0.0210	0.0211	0.0025	0.0044	0.0755	0.0755
K = 1000, N = 800												
$\rho = 0$	0.9080	0.9340	0.9400	0.9300	0.0939	0.1000	0.0907	0.0726	0.0131	N/A	0.0738	0.0738
$\rho = 0.3$	0.9460	0.9360	0.9280	0.9460	0.0823	0.0804	0.0996	0.0925	0.0105	N/A	0.0777	0.0777
$\rho = 0.5$	0.9620	0.9580	0.9460	0.9420	0.0878	0.0889	0.0832	0.0762	0.0096	N/A	0.0806	0.0806
$\rho = 0.7$	0.9720	0.9400	0.9400	0.9680	0.0756	0.0837	0.0882	0.0840	0.0089	N/A	0.0776	0.0776

Panel B: MAE comparison of each coefficient

	MAE_dowl				MAE_ols				MAE_lasso			
	β_3	β_6	β_{12}	β_{20}	β_3	β_6	β_{12}	β_{20}	β_3	β_6	β_{12}	β_{20}
K = 50, N = 60												
$\rho = 0$	0.0219	0.0173	0.0019	0.0036	0.0257	0.0243	0.0328	0.0223	0.1645	0.1629	0.0000	0.0000
$\rho = 0.3$	0.0266	0.0198	0.0048	0.0060	0.0405	0.0277	0.0276	0.0428	0.0562	0.2064	0.0000	0.0000
$\rho = 0.5$	0.0238	0.0266	0.0082	0.0029	0.0292	0.0313	0.0678	0.0299	0.1857	0.1721	0.0000	0.0000
$\rho = 0.7$	0.0337	0.0315	0.0043	0.0045	0.0488	0.0572	0.0492	0.0429	0.0576	0.2075	0.0000	0.0000
K = 50, N = 1000												
$\rho = 0$	0.0026	0.0026	0.0005	0.0003	0.0026	0.0026	0.0026	0.0024	0.1403	0.1352	0.0000	0.0000
$\rho = 0.3$	0.0028	0.0027	0.0004	0.0004	0.0028	0.0027	0.0028	0.0028	0.1445	0.1807	0.0000	0.0000
$\rho = 0.5$	0.0031	0.0036	0.0007	0.0004	0.0031	0.0036	0.0034	0.0032	0.1017	0.2014	0.0000	0.0000
$\rho = 0.7$	0.0045	0.0041	0.0007	0.0008	0.0045	0.0041	0.0044	0.0046	0.0603	0.2417	0.0000	0.0000
K = 1000, N = 800												
$\rho = 0$	0.0228	0.0231	0.0033	0.0030	N/A	N/A	N/A	N/A	0.1465	0.1488	0.0000	0.0000
$\rho = 0.3$	0.0164	0.0182	0.0044	0.0030	N/A	N/A	N/A	N/A	0.1392	0.1717	0.0000	0.0000
$\rho = 0.5$	0.0157	0.0174	0.0026	0.0027	N/A	N/A	N/A	N/A	0.0995	0.2231	0.0000	0.0000
$\rho = 0.7$	0.0130	0.0180	0.0032	0.0016	N/A	N/A	N/A	N/A	0.0783	0.2319	0.0000	0.0000

Panel A of Table 1 shows the results of coverage rate and the confidence interval (CI) width of the de-biased OWL estimator, as well as the average MAE (mean absolute error)

of each method. For LASSO estimator, we consider two methods for tuning the penalty parameter: one is by a ten-fold cross-validation (lasso_cv), which is widely used in machine learning literature; another one is by specifying the maximum number of non-zero coefficients we want to obtain.[11] We consider three settings in our experiment about the dimension of the dataset. First, we consider the case where $K = 50, N = 60$ ($N \approx K$) . Second, we look into the near asymptotic case where $K = 50, N = 1000$ ($N \gg K$). Third, we investigate the high-dimensional case where $K = 1000, N = 800$ ($K > N$). First of all, we find that the coverage rates of the de-biased OWL estimates for all cases are above 90%. In particular, the coverage rate for the near asymptotic case is near the correct size (95%) when correlation is not too high ($\rho < 0.5$). Comparing coverage rates with different correlation profile within each settings suggests that the coverage rate is typically higher when correlation is high ($\rho = 0.7$). However, we find that this is a result of enlarged confidence interval width rather than improved estimation accuracy. The width of confidence interval at the near asymptotic case suggests that when the correlation coefficient increases (ρ increases from 0 to 0.7), the width of confidence interval enlarges, particularly when ρ changes from 0.5 to 0.7. Meanwhile, an increase in ρ also associates with a decrease in estimation accuracy: the average MAE for the de-biased OWL estimate increases steadily when ρ increases. Also, comparing the average MAE of four coefficients ($\beta_3, \beta_6, \beta_{12}, \beta_{20}$) between the de-biased OWL, OLS and LASSO estimators, we find that the de-biased OWL estimate yields the lowest estimation errors in all cases.

Panel B of Table 1 gives a detailed illustration of MAE comparison between benchmarks for each coefficient. We find that the OLS estimator is good at estimating β_3 and β_6 because the OLS estimator is unbiased. However, the OLS estimation error is large when estimating β_{12} and β_{20} when their true values are zeros. The performance of the LASSO estimator is the opposite: it correctly shrinks β_{12} and β_{20} to zeros (in which case there is no estimation error for β_{12} and β_{20}) but the LASSO estimates for β_3 and β_6 are biased, and the estimation errors are large compared to the OLS estimates. The de-biased OWL estimate combines the merits of the OLS and LASSO estimators: it achieves unbiased estimation for the non-zero coefficients but also shrinks zero coefficients. In the cases where $K = 50$ (the

[11] We specify the maximum number of non-zero coefficients as ten to ensure sparse selection. After evaluation, we find both methods for choosing LASSO penalty parameter tend to yield the same result.

OLS estimator is feasible), the de-biased OWL estimates for β_3 and β_6 are very close to the OLS estimates, especially in the near asymptotic case. Meanwhile, the de-biased OWL estimates for β_{12} and β_{20} are close to LASSO estimates, performing sparsity shrinkage for useless covariates (whose true coefficients are zeros). In the high-dimensional case where $K = 1000$, we find that the MAE of the de-biased OWL estimates are substantially smaller than that of the LASSO estimates while the OLS estimates becoming infeasible.

This Monte Carlo experiment shows that, in both the low- and high-dimensional cases, the de-biased OWL estimator delivers unbiased estimation for useful covariates (whose true coefficients are non-zeros) as good as the OLS estimator while shrinking off useless covariates almost as good as the LASSO estimator.

4 Empirical application on factor investing

In this section, we apply the de-biased OWL method to predict stock returns using firm-characteristic based factors. We first introduce the dataset and the empirical method before conducting the empirical analysis.

4.1 Data and empirical method

We use the U.S. stock data from the Center for Research in Security Prices (CRSP) and Compustat database, both downloaded from the Wharton Research Data Service. The data spans between January 1980 and December 2017, totalling 456 months on all NYSE, AMEX and NASDAQ listed common stocks. Risk-free rate and market returns are downloaded from Kenneth French's on-line data library.[12] For predicting stock returns, we use a factor library which contains 80 anomaly factors constructed using characteristics sorted portfolios. More details of constructing those anomaly factors can be found in Sun (2019). We consider 30 stocks in the Dow Jones Industrial Average index as test assets while deleting stocks having any missing data between January 1980 and December 2017, which leaves 15 stocks as test assets. We then use these characteristic-based factors to predict stock returns for each of those 15 stocks.

[12] $https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html$

Suppose we use the lagged factor returns to predict individual stock return. The predicted return of any stock i at time t is

$$\hat{R}^i_{t+1} = f_t\tilde{\beta}_t, \qquad \tilde{\beta}_t \in \{\tilde{\beta}^{dOWL}_t, \tilde{\beta}^{OWL}_t, \tilde{\beta}^{LASSO}_t, \tilde{\beta}^{OLS}_t\}, \tag{22}$$

$$\tilde{\beta}^{dOWL}_t = \tilde{\beta}^{OWL}_t + \hat{\Theta}(R^i_t - f_{t-1}\tilde{\beta}^{OWL}_t)/n, \tag{23}$$

$$\tilde{\beta}^{OWL}_t = \arg\min_{\beta} \|R^i_t - f_{t-1}\beta\|_2^2 + \Omega(\beta), \tag{24}$$

$$\tilde{\beta}^{LASSO}_t = \arg\min_{\beta} \|R^i_t - f_{t-1}\beta\|_2^2 + \lambda\|\beta\|_1, \tag{25}$$

$$\tilde{\beta}^{OLS}_t = \arg\min_{\beta} \|R^i_t - f_{t-1}\beta\|_2^2, \tag{26}$$

where $\tilde{\beta}_t$ includes the de-biased OWL ('dOWL') estimator as well as benchmarks such as the OWL, OLS and LASSO estimators. $\hat{\Theta}$ is constructed in (18). $\tilde{\beta}^{OWL}_t$ is the OWL estimator in (2) and $\tilde{\beta}^{dOWL}_t$ is the de-biased OWL estimator in (14). λ is a hyper parameter for LASSO estimator and we use two methods to determine its value: either by a 10-fold cross-validation (CV) method or restricting its maximum non-zero coefficients to ten (DFmax = 10) to ensure sparsity.

4.2 A stock-by-stock analysis

In this subsection, we look at each stock and find which factors are the best predictors using the full sample estimation. However, it is worth stressing that our target here is to predict stock returns using a potentially large number of predictors. Since our approach is stock-specific, the selected factors for each stock should not be interpreted as a cross-sectionally valid true factors. Cross-sectional stock returns are typically investigated through the Fama-MacBeth regression method or the SDF method, see Sun (2019) for more details on dissecting the factor zoo for cross-sectional asset returns.

Figure 3 shows the contour plot of the estimated de-biased OWL coefficients (absolute value).[13] The vertical axis lists all the factors considered in the factor library and the horizontal axis shows 15 stocks as test assets. The left panel displays the estimated coefficient

[13]Note that we excluded 'betasq' in the factor library because the correlation coefficient between 'beta' and 'betasq' is more than 0.9. Including both of them in the factor zoo leads to serious estimation problems for OLS and LASSO estimators. For that reason, we exclude 'betasq' from the factor library.

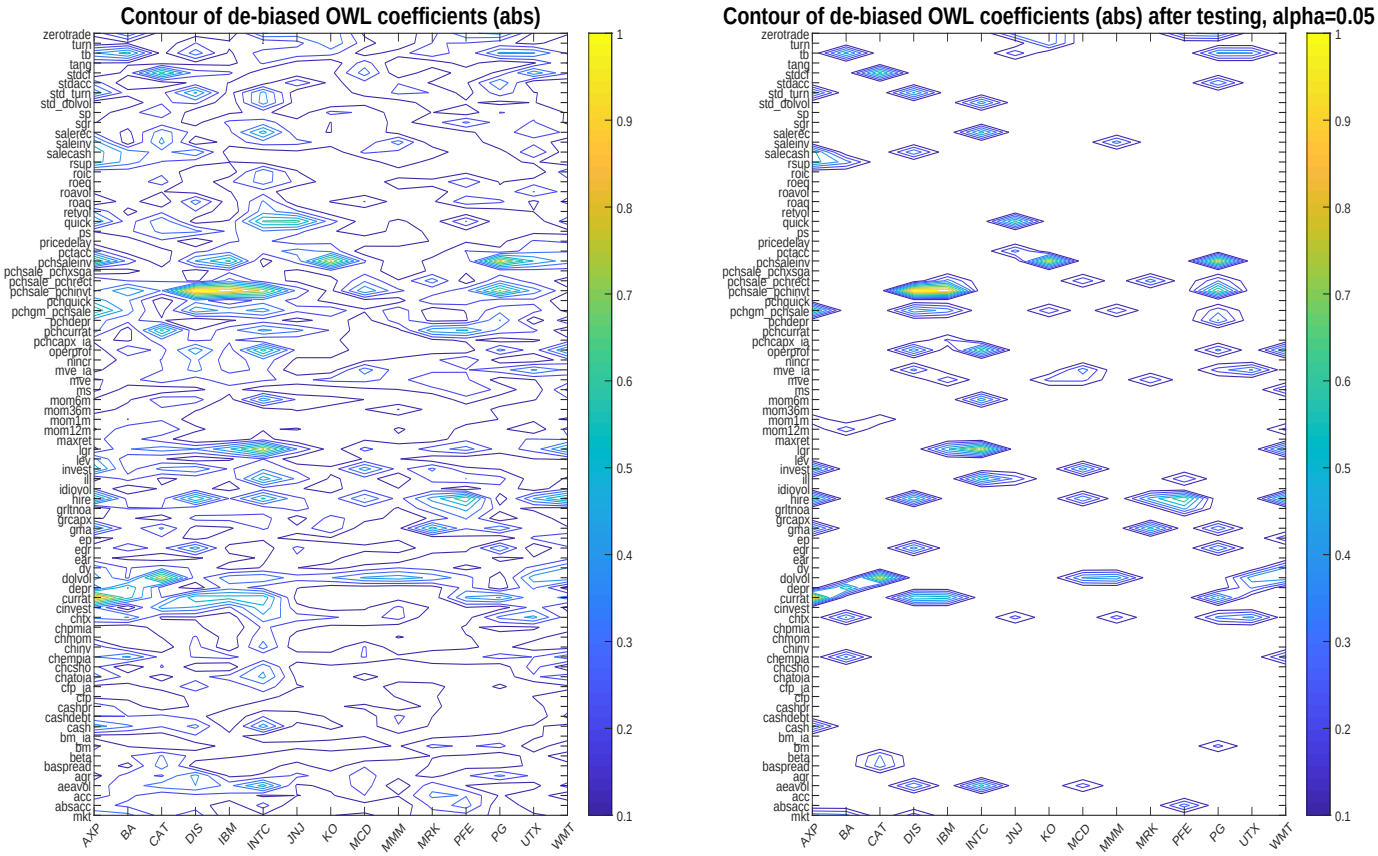

Figure 3. Contour plot of the de-biased OWL estimator of factor loadings. Yellow (blue) areas indicate large (small) estimated coefficients in terms of absolute values. The blank area indicates the estimated coefficients are close to zeros. The left panel is before testing and the right panel is after testing.

before testing, while the right panel displaying the contour plot of the de-biased OWL estimate after removing insignificant ones by applying the confidence interval in (21) at a significance level $\alpha = 5\%$. We first find that 'sales' related factors are typically selected as strong predictors for many stocks, while 'profitability' and 'investment' related factors form the second tier of strong predictors for stocks returns. The right panel confirms that most of those strong predictors are tested significant while many other minor predictors are removed after applying the confidence interval. Meanwhile, it also suggests that some stocks, for instance 'KO' and 'MMM' are sensitive to only very few (less than five) factors in our factor library, while others like 'J&J' and 'DIS' having many (more than ten)

significant predictors.

Next we choose a random stock, for instance 'DIS' to compare the estimation results using different methods. Figure 4 shows the plot of estimated coefficients using the de-biased OWL (blue), OWL (red), LASSO (yellow, with DFmax = 10), LASSO_CV (purple, with 10-fold cross validation) and OLS (green) estimators. The grey area displays the 95% confidence intervals for the de-biased OWL estimator.

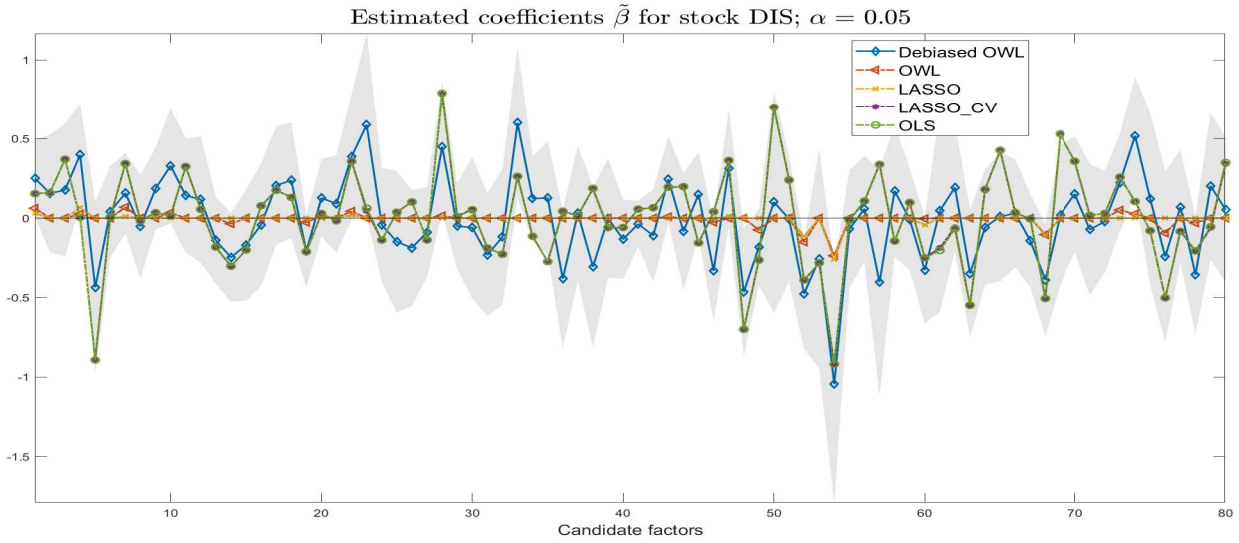

Figure 4. Estimated factor loadings of 'DIS'

This figure plots the estimated factor loadings using the de-biased OWL, OWL, LASSO and OLS estimators. The shaded area is the 95% confidence interval for the de-biased OWL estimator.

Figure 4 shows that the OWL estimator yields very similar result to the LASSO estimator (with maximum number of non-zero coefficients restricted to ten to ensure sparsity) for the sparsity property, i.e., they both shrink many factors' coefficients to zeros, yet they differs in some of the survival factors (i.e., factors having non-zero estimated coefficients). Meanwhile, the estimated coefficients of survival factors of both the OWL and the LASSO (yellow) estimators are very close zero, which are caused by an inward bias pulling the coefficients towards zeros. The cross validated LASSO estimator yields very similar result to the OLS estimator. Cross validation method suggests all factors are useful to predict stock returns and thus shrinks no factors and yields almost the same result as the OLS estimator. The de-biased OWL estimator corrects that bias for the OWL estimator. We find that after bias-correction, the de-biased OWL estimate displays a similar trend to the

OLS estimator, although the magnitude of estimated coefficients varies on some factors compared to the OLS estimator. Meanwhile, the de-biased OWL estimator loses the sparsity property (i.e. no factors receive zero coefficients for the de-biased OWL estimator), but we find that many of those factors receiving zero coefficients in the OWL estimation are insignificant in the de-biased OWL estimation after applying the confidence intervals. In addition, to preserve the sparsity property of the OWL estimator while correcting the bias for survival factors, we can selectively de-bias these estimated non-zero coefficients of the OWL estimator.

5 Conclusion

In high dimensional datasets where covariates exhibit high correlations, Zou and Hastie (2005) and Figueiredo and Nowak (2016) have shown that the LASSO estimator performs poorly. Figueiredo and Nowak (2016) introduced the Ordered-Weighted-LASSO (OWL) estimator which is specifically tailored to deal with correlations between covariates. Sun (2019) introduced the OWL estimator to dissect the factor zoo for the cross sectional asset returns and further developed asymptotic properties for the OWL estimator. Although Sun (2019) shows that the OWL estimator is consistent, it is biased in small samples. This paper extends Figueiredo and Nowak (2016) and Sun (2019) to study the (non)asymptotic properties of the OWL estimator with *less restrictive* assumptions and further propose a bias-corrected version of the OWL estimator. Monte Carlo experiments show that, in both the low- and high- dimensional settings, the de-biased OWL estimator delivers unbiased estimation for useful covariates as good as the OLS estimator while shrinking off useless covariates almost as good as the LASSO estimator. In the empirical analysis, we implement the de-biased OWL estimation to predict returns for 15 stocks from the Dow Jones Industrial Average index using 80 factors. We find some 'sales', 'profitability' and 'investment' related factors are strong predictors for many stock returns.

Appendix

A Technical proofs

A.1 Proof of Theorem 2.1

Proof. The proof of Theorem 2.1 consists of two parts. In the first part we derive the oracle inequality (4) and (5) under the event E, which will be specified below in (A.4). The second part we will derive the probability of this event $\mathbb{P}(E)$ to be true.

Part I. According to the "argmin" property,

$$\frac{1}{n}||y - X\hat{\beta}||_2^2 + \frac{1}{n}\sum_{j=1}^{p}\omega_j|\hat{\beta}|_{[j]} \leq \frac{1}{n}||y - X\beta^0||_2^2 + \frac{1}{n}\sum_{j=1}^{p}\omega_j|\beta^0|_{[j]}. \tag{A.1}$$

Since $(\omega_1, \cdots, \omega_p)'$ where $\omega_j = \lambda_1 + \lambda_2(p-j),\ j = \{1, \cdots, p\}$ is in a monotone non-negative cone, so $\omega_1 \geq \omega_2 \geq ... \geq \omega_p$. Then we have

$$\sum_{j=1}^{p}\omega_j|\hat{\beta}|_{[j]} \geq \omega_p||\hat{\beta}||_1 = \lambda_1||\hat{\beta}||_1,$$
$$\sum_{j=1}^{p}\omega_j|\beta^0|_{[j]} \leq \omega_1||\beta^0||_1 = [\lambda_1 + \lambda_2(p-1)]||\beta^0||_1.$$

Together with $y = X\beta^0 + \epsilon$, this implies that (A.1) can be simplified as follow:

$$\frac{1}{n}||\epsilon - X(\hat{\beta} - \beta^0)||_2^2 + \frac{1}{n}\omega_p||\hat{\beta}||_1 \leq \frac{1}{n}||\epsilon||_2^2 + \frac{1}{n}\omega_1||\beta^0||_1 \tag{A.2}$$
$$\frac{1}{n}||X(\hat{\beta} - \beta^0)||_2^2 + \frac{\lambda_1}{n}||\hat{\beta}||_1 \leq \frac{2}{n}\epsilon' X(\hat{\beta} - \beta^0) + \frac{1}{n}[\lambda_1 + \lambda_2(p-1)]||\beta^0||_1. \tag{A.3}$$

Note that $\epsilon' X(\hat{\beta} - \beta^0) \leq \|\epsilon' X\|_\infty ||\hat{\beta} - \beta^0||_1$. Let $\lambda_0 > 0$ and consider an event

$$E := \left\{\frac{1}{n}\|\epsilon' X\|_\infty \leq \frac{\lambda_0}{2}\right\}, \tag{A.4}$$

where $\lambda_0 = \kappa\sqrt{\frac{\log p}{n}}$, where $\kappa > 0$ is a constant. Then in view of (A.4), (A.3) can be bounded as

$$\frac{1}{n}||X(\hat{\beta}-\beta^0)||_2^2 + \frac{\lambda_1}{n}||\hat{\beta}||_1 \leq \lambda_0||\hat{\beta}-\beta^0||_1 + \frac{1}{n}[\lambda_1 + \lambda_2(p-1)]||\beta^0||_1. \tag{A.5}$$

By assumption of the theorem, $\frac{\lambda_1}{n} = 2\lambda_0$. So we obtain

$$\frac{1}{n}||X(\hat{\beta}-\beta^0)||_2^2 + \frac{\lambda_1}{n}||\hat{\beta}||_1 \leq \frac{\lambda_1}{2n}||\hat{\beta}-\beta^0||_1 + \frac{1}{n}[\lambda_1 + \lambda_2(p-1)]||\beta^0||_1. \tag{A.6}$$

By the definition of s_0, $\hat{\beta} = \hat{\beta}_{s_0} + \hat{\beta}_{s_0^c}$. Utilizing the triangle inequality $\|a\|_1 + \|b\|_1 \geq \|a+b\|_1$ for any vector a and b, we obtain

$$||\hat{\beta}||_1 = ||\hat{\beta}_{s_0}||_1 + ||\hat{\beta}_{s_0^c}||_1 \geq ||\beta^0_{s_0}||_1 - ||\hat{\beta}_{s_0} - \beta^0_{s_0}||_1 + ||\hat{\beta}_{s_0^c}||_1, \tag{A.7}$$

$$||\hat{\beta}-\beta^0||_1 = ||\hat{\beta}_{s_0} - \beta^0_{s_0}||_1 + ||\hat{\beta}_{s_0^c}||_1. \tag{A.8}$$

Therefore, using (A.7) and (A.8), (A.6) can be written as

$$\frac{2}{n}||X(\hat{\beta}-\beta^0)||_2^2 + \frac{2\lambda_1}{n}(||\beta^0_{s_0}||_1 - ||\hat{\beta}_{s_0} - \beta^0_{s_0}||_1 + ||\hat{\beta}_{s_0^c}||_1)$$
$$\leq \frac{\lambda_1}{n}(||\hat{\beta}_{s_0} - \beta^0_{s_0}||_1 + ||\hat{\beta}_{s_0^c}||_1) + \frac{2}{n}[\lambda_1 + \lambda_2(p-1)]||\beta^0||_1. \tag{A.9}$$

Note that $\|\beta^0_{s_0}\|_1 = \|\beta^0\|_1$, so (A.9) can be written as

$$\frac{2}{n}||X(\hat{\beta}-\beta^0)||_2^2 + \frac{\lambda_1}{n}||\hat{\beta}_{s_0^c}||_1 \leq \frac{3\lambda_1}{n}||\hat{\beta}_{s_0} - \beta^0_{s_0}||_1 + \frac{2\lambda_2(p-1)}{n}||\beta^0||_1. \tag{A.10}$$

By (A.8), $||\hat{\beta}_{s_0^c}||_1 = ||\hat{\beta}-\beta^0||_1 - ||\hat{\beta}_{s_0} - \beta^0_{s_0}||_1$. Utilizing this in (A.10), we obtain

$$\frac{2}{n}||X(\hat{\beta}-\beta^0)||_2^2 + \frac{\lambda_1}{n}||\hat{\beta}-\beta^0||_1 \leq \frac{4\lambda_1}{n}||\hat{\beta}_{s_0} - \beta^0_{s_0}||_1 + \frac{2\lambda_2(p-1)}{n}||\beta^0||_1. \tag{A.11}$$

Utilizing the compatibility condition $\|\beta_{s_0}\|_1^2 \leq (\beta'\hat{\Sigma}\beta)s/\phi_0^2$ given in Lemma 1 on $||\hat{\beta}_{s_0} - \beta^0_{s_0}||_1$

and using definition $\hat{\Sigma} = \dfrac{X'X}{n}$, we obtain

$$\begin{aligned} ||\hat{\beta}_{s_0} - \beta^0_{s_0}||_1^2 \leq (\hat{\beta} - \beta^0)'\hat{\Sigma}(\hat{\beta} - \beta^0)s/\Phi_0^2 = \|X(\hat{\beta} - \beta^0)\|_2^2 s/(n\Phi_0^2), \\ ||\hat{\beta}_{s_0} - \beta^0_{s_0}||_1 \leq \|X(\hat{\beta} - \beta^0)\|_2 \sqrt{s}/(\sqrt{n}\Phi_0). \end{aligned} \tag{A.12}$$

Therefore, applying inequality $4ab \leq a^2 + 4b^2$, we obtain

$$\begin{aligned} \frac{4\lambda_1}{n}||\hat{\beta}_{s_0} - \beta^0_{s_0}||_1 &\leq 4\left(\frac{\|X(\hat{\beta} - \beta^0)\|_2}{\sqrt{n}}\right)\left(\frac{\lambda_1}{n}\frac{\sqrt{s}}{\Phi_0}\right) \\ &\leq \frac{1}{n}\|X(\hat{\beta} - \beta^0)\|_2^2 + 4(\frac{\lambda_1}{n})^2\frac{s}{\Phi_0^2}. \end{aligned}$$

So (A.11) can be written as

$$\frac{1}{n}||X(\hat{\beta} - \beta^0)||_2^2 + \frac{\lambda_1}{n}||\hat{\beta} - \beta^0||_1 \leq 4(\frac{\lambda_1}{n})^2\frac{s}{\Phi_0^2} + \frac{2\lambda_2(p-1)}{n}||\beta^0||_1. \tag{A.13}$$

By assumption of the theorem, $\dfrac{\lambda_1}{n} = 2\lambda_0 \asymp \sqrt{\dfrac{\log p}{n}}$, and $\dfrac{\lambda_2}{n} \lesssim \dfrac{s\log p}{np} \asymp \dfrac{s\lambda_0^2}{p}$. Therefore, (A.13) can be written as

$$\frac{1}{n}||X(\hat{\beta} - \beta^0)||_2^2 + 2\lambda_0||\hat{\beta} - \beta^0||_1 \lesssim 16\lambda_0^2 s/\Phi_0^2 + 2\lambda_0^2 s\|\beta^0\|_1. \tag{A.14}$$

Using $\sqrt{a^2 + b^2} \leq a + b$, for all $a, b > 0$, (A.14) implies

$$\frac{1}{n}||X(\hat{\beta} - \beta^0)||_2 \lesssim 4\lambda_0\sqrt{s}/\Phi_0 + \lambda_0\sqrt{2s\|\beta^0\|_1}, \tag{A.15}$$

$$||\hat{\beta} - \beta^0||_1 \lesssim 8\lambda_0 s/\Phi_0^2 + \lambda_0 s\|\beta^0\|_1. \tag{A.16}$$

This shows that (4) and (5) in Theorem 2.1 are valid, assuming that (A.4) holds.

Part II. Next we calculate $\mathbb{P}(E)$. We have

$$\begin{aligned}\mathbb{P}(E^c) &= \mathbb{P}(\frac{1}{n}\|X'\epsilon\|_\infty > \frac{\lambda_0}{2}) = \mathbb{P}(\frac{1}{n}\max_{j=1,\cdots,p}|\sum_{i=1}^n X_{i,j}\epsilon_i| > \frac{\lambda_0}{2}) \\ &\leq \sum_{j=1}^p \mathbb{P}(\frac{1}{\sqrt{n}}\sum_{i=1}^n |X_{i,j}\epsilon_i| > \frac{\lambda_0\sqrt{n}}{2}) = p\max_{j=1,\cdots,p}\mathbb{P}(\frac{1}{\sqrt{n}}\sum_{i=1}^n |X_{i,j}\epsilon_i| > \frac{\lambda_0\sqrt{n}}{2}).\end{aligned} \tag{A.17}$$

By Assumption 1, for $j = 1, \cdots, p$, sequence $\{z_{i,j}\}_{i=1}^n := \{X_{i,j}\epsilon_i\}_{i=1}^n$ is $\alpha-$mixing with exponential decaying mixing coefficients, and by Lemma A4 in Dendramis et al. (2019), we have

$$\mathbb{P}(|z_{i,j}| \geq a) \leq c_1 \exp(-c_2 a^q),$$

where $a > 0, q = q_1q_2/(q_1+q_2) > 0$. It also has zero-mean, i.e. $\mathrm{E}(z_{i,j}) = 0$. Thus, by Lemma 1 in Dendramis et al. (2019), for all $j = 1, \cdots, p$,

$$\mathbb{P}\left(\frac{1}{\sqrt{n}}\left|\sum_{i=1}^n z_{i,j}\right| \geq \xi\right) \leq c_0\left[\exp(-c_1'\xi^2) + \exp\left(-c_2'\left(\frac{\xi\sqrt{n}}{\log^2 n}\right)^\zeta\right)\right],$$

where $\zeta = q/(q+1)$ and constants c_0, c_1', c_2' do not depend on ξ, i and j.
Note that $\lambda_0 = \kappa\sqrt{\log p/n}$. Setting $\xi = \lambda_0\sqrt{n}/2 = \kappa\sqrt{\log p}/2$, we obtain

$$\begin{aligned}p\mathbb{P}(\frac{1}{\sqrt{n}}|z_{i,j}| > \frac{\lambda_0\sqrt{n}}{2}) &\leq pc_0\exp(-c_1'(\frac{\kappa}{2})^2\log p) + pc_0\exp(-c_2'(\frac{\kappa\sqrt{n\log p}}{2\log^2 n})^\zeta) \\ &:= r_p + r'_{p,n}.\end{aligned} \tag{A.18}$$

Now we consider two cases of different rates of p and n.

Case 1: $n, p \to \infty$.

Selecting $\kappa > 0$, such that $c_1'(\kappa/2)^2 > 1+\epsilon$ for some small number $\epsilon > 0$, we obtain

$$r_p \leq pc_0\exp[-(1+\epsilon)\log p] = c_0 p^{-\epsilon} \to 0, \quad \text{as } p \to \infty. \tag{A.19}$$

By Assumption $p = O(n^\delta)$ for some $\delta > 0$, we have $n^{1/4} \geq p^{1/(4\delta)}$. Also, $n^{1/4} > 2\log^2 n$ as

$n \to \infty$. Then

$$c_2'(\frac{\kappa\sqrt{n\log p}}{2\log^2 n})^\zeta \geq c_2'(\kappa p^{1/(4\delta)}\sqrt{\log p})^\zeta > (1+\epsilon)\log p, \quad \text{as } p \to \infty. \tag{A.20}$$

Therefore, equation (A.19) and (A.20) imply that

$$r'_{p,n} \leq r_p \to 0, \quad \text{as } n, p \to \infty.$$

Then by (A.17) and (A.18), we obtain

$$\begin{aligned} \mathbb{P}(E^c) &= r_p + r'_{p,n} \leq 2r_p \leq 2c_0 p^{-\epsilon}, \\ \mathbb{P}(E) &= 1 - \mathbb{P}(E^c) \geq 1 - c_0' p^{-\epsilon} \to 1, \quad \text{as } n, p \to \infty, \end{aligned} \tag{A.21}$$

where $c_0' = 2c_0$. This proves the first probability claim in part one of Theorem 2.1.

Case 2: p is bounded.

In this case, $\log p$ is also bounded, then r_p and $r'_{p,n}$ in (A.18) can be bounded as

$$r_p = pc_0 \exp(-\frac{c_1'}{4}\kappa^2 \log p), \qquad r'_{p,n} = pc_0 \exp\left(-c_2'\left(\frac{\kappa\sqrt{n\log p}}{2\log^2 n}\right)^\zeta\right).$$

Therefore,

$$\mathbb{P}(E) = 1 - \mathbb{P}(E^c) = 1 - pc_0\left[\exp(-\frac{c_1'}{4}\kappa^2\log p) + \exp\left(-c_2'\left(\frac{\kappa\sqrt{n\log p}}{2\log^2 n}\right)^\zeta\right)\right], \tag{A.22}$$

which complete the proof of Theorem 2.1.

□

A.2 Proof of corollary 2.1

Proof. Note that $\lambda_0 = \kappa\sqrt{\log p/n}$, where $\kappa > 0$ is a tuning parameter. By (5) in Theorem 2.1 and Assumption 3(a), it follows naturally that

$$\|\hat{\beta} - \beta^0\|_1 = O_p(s\sqrt{\frac{\log p}{n}}) = o_p(1), \tag{A.23}$$

which proves the second claim of (7). Utilizing $\hat{\Sigma} = X'X/n$, we obtain

$$\begin{aligned} \|X(\hat{\beta} - \beta^0)\|_2^2/n &= (\hat{\beta} - \beta^0)'\hat{\Sigma}(\hat{\beta} - \beta^0) \\ &= (\hat{\beta} - \beta^0)'(\hat{\Sigma} - \Sigma)(\hat{\beta} - \beta^0) + (\hat{\beta} - \beta^0)'\Sigma(\hat{\beta} - \beta^0). \end{aligned} \tag{A.24}$$

Note that $\Sigma = \mathrm{E}(\hat{\Sigma})$ is non-singular, so

$$(\hat{\beta} - \beta^0)'\Sigma(\hat{\beta} - \beta^0) \geq \Lambda_{min}^2 \|\hat{\beta} - \beta^0\|_2^2,$$

where Λ_{min} is the smallest eigenvalue of Σ, and $\Lambda_{min} > 0$. Moreover, the first part of the r.h.s of (A.24) has the following property:

$$(\hat{\beta} - \beta^0)'(\hat{\Sigma} - \Sigma)(\hat{\beta} - \beta^0) \geq -\|\hat{\Sigma} - \Sigma\|_\infty \|\hat{\beta} - \beta^0\|_1^2,$$

where $\|\hat{\Sigma} - \Sigma\|_\infty := \max_{1 \leq i,j \leq p} |\hat{\Sigma}_{i,j} - \Sigma_{i,j}|$. Using lemma 14.12 in Buhlmann and Van de Geer (2011), we have $\max_{1 \leq i,j \leq p} |\hat{\Sigma}_{i,j} - \Sigma_{i,j}| = O_p(\sqrt{\log p/n})$. Together with $\|\hat{\beta} - \beta^0\|_1 = O_p(s\sqrt{\log p/n})$ obtained in (A.23), this implies that (A.24) can be bounded as

$$\begin{aligned} \frac{1}{n}\|X(\hat{\beta} - \beta^0)\|_2^2 &= (\hat{\beta} - \beta^0)'\Sigma(\hat{\beta} - \beta^0) + (\hat{\beta} - \beta^0)'(\hat{\Sigma} - \Sigma)(\hat{\beta} - \beta^0) \\ &\geq \Lambda_{min}^2 \|\hat{\beta} - \beta^0\|_2^2 - \|\hat{\Sigma} - \Sigma^0\|_\infty \|\hat{\beta} - \beta^0\|_1^2 \\ &\geq \Lambda_{min}^2 \|\hat{\beta} - \beta^0\|_2^2 - O_p\left(s^2 \left(\frac{\log p}{n}\right)^{3/2}\right). \end{aligned} \tag{A.25}$$

Note that $\lambda_0 \asymp \sqrt{\log p/n}$. So by (4) we obtain

$$\frac{1}{n}\|X(\hat{\beta} - \beta^0)\|_2^2 = O_p\left(\frac{s \log p}{n}\right). \tag{A.26}$$

Plugging (A.26) into (A.25) and rearranging (A.25), we obtain

$$\|\hat{\beta} - \beta^0\|_2^2 \leq \frac{1}{\Lambda_{min}^2} O_p\left(\frac{s \log p}{n}\right) + \frac{1}{\Lambda_{min}^2} O_p\left(s^2 \left(\frac{\log p}{n}\right)^{3/2}\right),$$

where $O_p\left(s^2\left(\frac{\log p}{n}\right)^{3/2}\right) = O_p\left(\frac{s\log p}{n}\right) O_p\left(s\sqrt{\frac{\log p}{n}}\right)$. Note that $\Lambda_{min} \geq a > 0$ where a is a constant, hence $\frac{1}{\Lambda_{min}^2} = O(1)$. Then by Assumption 3, we obtain

$$\|\hat{\beta} - \beta^0\|_2^2 = o_p(1), \tag{A.27}$$

which proves the first claim of (7). Also, by Theorem 2.1 part one, (A.23) and (A.27) hold with probability tending to one. This completes the proof. □

A.3 Proof of Theorem 2.2

Proof. By the definition of $\hat{b}$ in (14) and by extracting $\sqrt{n}$ from (12), it is easy to show that

$$\sqrt{n}(\hat{b} - \beta^0) = \hat{\Theta}X'\epsilon/\sqrt{n} - \Delta,$$

where Δ is defined in (13). Then to prove (19), it suffices to show that

$$\Delta = o_p(1). \tag{A.28}$$

Let X_i be a $1 \times p$ vector and denote

$$\hat{\Sigma}_{X\epsilon} = \frac{1}{n}\sum_{i=1}^{n}\left[(X_i'\hat{\epsilon}_i)(X_i'\hat{\epsilon}_i)'\right]. \tag{A.29}$$

To show (20) and (A.28), it suffices to prove that for any $l = 1, 2, \cdots, p$ such that

$$t = \frac{\sqrt{n}(\hat{b}_l - \beta_l^0)}{\sqrt{\hat{\Theta}_l'\hat{\Sigma}_{X\epsilon}\hat{\Theta}_l}} = \frac{\hat{\Theta}_l X'\epsilon/\sqrt{n}}{\sqrt{\hat{\Theta}_l'\hat{\Sigma}_{X\epsilon}\hat{\Theta}_l}} + \frac{-\Delta}{\sqrt{\hat{\Theta}_l'\hat{\Sigma}_{X\epsilon}\hat{\Theta}_l}} := t_1 + t_2,$$

where t_1 is asymptotically normal and $t_2 = o_p(1)$.

Step 1: we will show that t_1 is asymptotically normal. Let

$$t_1^* = \frac{\Theta_l' X'\epsilon/\sqrt{n}}{\sqrt{\Theta_l'\Sigma_{X\epsilon}\Theta_l}} = \frac{\Theta_l'\sum_{i=1}^{n} X_i'\epsilon_i/\sqrt{n}}{\sqrt{\Theta_l'\Sigma_{X\epsilon}\Theta_l}},$$

where $\Sigma_{X\epsilon} = E[\frac{1}{n}\sum_{i=1}^{n}(X_i'\epsilon_i)(X_i'\epsilon_i)']$. We assume in Theorem 2.2 that $X_i'\epsilon_i$ is a stationary sequence, then $\Sigma_{X\epsilon} = \mathrm{E}[(X_1'\epsilon_1)(X_1'\epsilon_1)'] = \mathrm{Var}(X_1'\epsilon_1) > 0$. By Assumption 1 and the definition of $\Sigma_{X\epsilon}$, we have

$$E\left[\frac{\Theta_l'X'\epsilon/\sqrt{n}}{\sqrt{\Theta_l'\Sigma_{X\epsilon}\Theta_l}}\right] = E\left[\frac{\Theta_l'\sum_{i=1}^{n}X_i'\epsilon_i/\sqrt{n}}{\sqrt{\Theta_l'\Sigma_{X\epsilon}\Theta_l}}\right] = 0,$$

and

$$E\left[\frac{\Theta_l'X'\epsilon/\sqrt{n}}{\sqrt{\Theta_l'\Sigma_{X\epsilon}\Theta_l}}\right]^2 = E\left[\frac{\Theta_l'\frac{1}{n}\sum_{i=1}^{n}(X_i'\epsilon_i)(X_i'\epsilon_i)'\Theta_l}{\Theta_l'\Sigma_{X\epsilon}\Theta_l}\right] = 1,$$

where $\Theta_l'\Sigma_{X\epsilon}\Theta_l$ is bounded away from zero. Indeed, since $\Sigma_{X\epsilon}$ is a symmetric positive definite matrix, it can be decomposed such that

$$\Theta_l'\Sigma_{X\epsilon}\Theta_l = \Theta_l'P'\mathrm{eig}(\Sigma_{X\epsilon})P\Theta_l \geq \Lambda_{min}(\Sigma_{X\epsilon})\|\Theta_l\|_2^2 > 0, \tag{A.30}$$

where $\mathrm{eig}(\Sigma_{X\epsilon})$ is the diagonal matrix that collects the eigenvalues of $\Sigma_{X\epsilon}$, and P is an orthonormal matrix. Because $\Lambda_{min}(\Sigma_{X\epsilon}) \geq a > 0$ where a is a constant and $\|\Theta_l\|_2^2 > 0$, so $\Theta_l'\Sigma_{X\epsilon}\Theta_l > 0$. Then by Theorem 24.6 and Corollary 24.7 in Davidson (1994), $\Theta_l'X'\epsilon/\sqrt{n} \to \mathbb{N}(0, \Theta_l\Sigma_{X\epsilon}\Theta_l')$, or $t_1^* \to \mathbb{N}(0,1)$.

Next we will show that

$$|\hat{\Theta}_l'\hat{\Sigma}_{X\epsilon}\hat{\Theta}_l - \Theta_l'\Sigma_{X\epsilon}\Theta_l| = o_p(1). \tag{A.31}$$

Set

$$\tilde{\Sigma}_{X\epsilon} = \frac{1}{n}\sum_{i=1}^{n}\left[(X_i'\epsilon_i)(X_i'\epsilon_i)'\right]. \tag{A.32}$$

Then

$$
\begin{aligned}
|\hat{\Theta}_l'\hat{\Sigma}_{X\epsilon}\hat{\Theta}_l - \Theta_l'\Sigma_{X\epsilon}\Theta_l| &\leq |\hat{\Theta}_l'\hat{\Sigma}_{X\epsilon}\hat{\Theta}_l - \hat{\Theta}_l'\Sigma_{X\epsilon}\hat{\Theta}_l| + |\hat{\Theta}_l'\Sigma_{X\epsilon}\hat{\Theta}_l - \Theta_l'\Sigma_{X\epsilon}\Theta_l| \\
&\leq |\hat{\Theta}_l'\hat{\Sigma}_{X\epsilon}\hat{\Theta}_l - \hat{\Theta}_l'\tilde{\Sigma}_{X\epsilon}\hat{\Theta}_l| + |\hat{\Theta}_l'\tilde{\Sigma}_{X\epsilon}\hat{\Theta}_l - \hat{\Theta}_l'\Sigma_{X\epsilon}\hat{\Theta}_l| + |\hat{\Theta}_l'\Sigma_{X\epsilon}\hat{\Theta}_l - \Theta_l'\Sigma_{X\epsilon}\Theta_l| \\
&= (I) + (II) + (III).
\end{aligned} \tag{A.33}
$$

For (I), we have

$$
|\hat{\Theta}_l'\hat{\Sigma}_{X\epsilon}\hat{\Theta}_l - \hat{\Theta}_l'\tilde{\Sigma}_{X\epsilon}\hat{\Theta}_l| \leq \|\hat{\Sigma}_{X\epsilon} - \tilde{\Sigma}_{X\epsilon}\|_\infty \|\hat{\Theta}_l\|_1^2.
$$

Note that $\hat{\epsilon}_i = \epsilon_i + X_i(\beta^0 - \hat{\beta})$. Plugging $\hat{\epsilon}_i$ into $\hat{\Sigma}_{X\epsilon} - \tilde{\Sigma}_{X\epsilon}$, we obtain

$$
\begin{aligned}
\hat{\Sigma}_{X\epsilon} - \tilde{\Sigma}_{X\epsilon} =& \frac{1}{n}\sum_{i=1}^n \left[[X_i'(\epsilon_i + X_i(\beta^0 - \hat{\beta})][X_i'(\epsilon_i + X_i(\beta^0 - \hat{\beta})]'\right] - \frac{1}{n}\sum_{i=1}^n [(X_i'\epsilon_i)(X_i'\epsilon_i)'] \\
=& \frac{1}{n}\sum_{i=1}^n X_i'X_i(\beta^0 - \hat{\beta})[X_i'X_i(\beta^0 - \hat{\beta})]' + \frac{1}{n}\sum_{i=1}^n X_i'\epsilon_i[X_i'X_i(\beta^0 - \hat{\beta})]' \\
&+ \frac{1}{n}\sum_{i=1}^n [X_i'X_i(\beta^0 - \hat{\beta})](X_i'\epsilon_i)' \\
=& (i) + (ii) + (iii).
\end{aligned}
$$

Next, we will show that $\|(i)\|_\infty = O_p(s\sqrt{\log p/n})$, $\|(ii)\|_\infty = O_p(\sqrt{s\log p/n})$ and $\|(iii)\|_\infty = O_p(\sqrt{s\log p/n})$. First of all, for (i), we have

$$
\begin{aligned}
\|\frac{1}{n}\sum_{i=1}^n X_i'X_i(\hat{\beta} - \beta^0)[X_i'X_i(\hat{\beta} - \beta^0)]'\|_\infty =& \|\frac{1}{n}\sum_{i=1}^n X_i'X_i(\hat{\beta} - \beta^0)(\hat{\beta} - \beta^0)'X_i'X_i\|_\infty \\
\leq& \frac{1}{n}\sum_{i=1}^n \|X_i'X_iX_i'X_i\|_\infty \|\hat{\beta} - \beta^0\|_1^2 \\
\leq& \max_j \frac{1}{n}\sum_{i=1}^n X_{i,j}^4 \ \|\hat{\beta} - \beta^0\|_1^2,
\end{aligned} \tag{A.34}
$$

where $j = 1, \cdots, p$. By Assumption 1, $\mathbb{P}(|X_{i,j}| > a) \leq c_1\exp(-c_2a^{q_1})$. Set $Y_{i,j} = X_{i,j}^4$, then $\mathbb{P}(|Y_{i,j}| > a) = \mathbb{P}(|X_{i,j}| > a^{1/4}) \leq c_1\exp(-c_2a^{q_1/4})$. So $X_{i,j}^4$ also has exponential tail bound (with a different parameter). Then by (A.17) and (A.21), for all $j = 1, \cdots, p$, we have

$\mathbb{P}(|n^{-1}\sum_{i=1}^{n} X_{i,j}^4 - E(X_{i,j}^4)| > \lambda_0/2) \leq c_0' p^{-\epsilon} \to 0$ as $n, p \to \infty$, where c_0' is a positive constant and $\epsilon > 0$ is a small number. Note that $\lambda_0 \asymp \sqrt{\log p/n}$. Hence $|n^{-1}\sum_{i=1}^{n} X_{i,j}^4 - E(X_{i,j}^4)| = O_p(\sqrt{\log p/n})$ with probability tending to one. Then by Assumption 1, $E(X_{i,j}^4) < \infty$, and by Assumption 3, $\sqrt{\log p/n} = o_p(1)$, so we have $|n^{-1}\sum_{i=1}^{n} X_i^4| \leq |n^{-1}\sum_{i=1}^{n} X_{i,j}^4 - E(X_{i,j}^4)| + |E(X_{i,j}^4)| = o_p(1) + O_p(1) = O_p(1)$. By (7) we have $\|\hat{\beta} - \beta^0\|_1 = O_p(s\sqrt{\log p/n})$ with probability tending to one. Therefore, we have $\|(i)\|_\infty = O_p(s\sqrt{\log p/n})$ with probability tending to one.

For (ii), we have

$$\begin{aligned}
\|\frac{1}{n}\sum_{i=1}^{n} X_i'\epsilon_i[X_i'X_i(\beta^0 - \hat{\beta})]'\|_\infty &= \|\frac{1}{n}\sum_{i=1}^{n} X_i'\epsilon_i X_i\|_\infty [X_i(\beta^0 - \hat{\beta})]' \\
&\leq \frac{1}{n}\|\sum_{i=1}^{n} X_i'X_iX_i'X_i\epsilon_i^2\|_\infty^{1/2}(\sum_{i=1}^{n}[X_i(\beta^0 - \hat{\beta})]^2)^{1/2} \\
&\leq (\frac{1}{n}\max_j \sum_{i=1}^{n} X_{i,j}^4\epsilon_i^2)^{1/2}(\frac{1}{n}\|X_i(\beta^0 - \hat{\beta})\|_2^2)^{1/2} \\
&\leq (\frac{1}{n}\max_j \sum_{i=1}^{n} X_{i,j}^4\epsilon_i^2)^{1/2}(\frac{1}{n}\|X(\beta^0 - \hat{\beta})\|_2).
\end{aligned}$$

By (4) and Assumption 3, we have $\|n^{-1}X(\hat{\beta} - \beta^0)\|_2 = O_p(\sqrt{(s\log p)/n})$. Since both $X_{i,j}$ and ϵ_i are $\alpha-$mixing and have exponential tail distributions, then following a similar argument as in (i) and using (A.17) and (A.21), for any $j = 1, \cdots, p$, we have $n^{-1}\sum_{i=1}^{n} X_{i,j}^4\epsilon_i^2 = O_p(\sqrt{\log p/n}) + O_p(1) = o_p(1) + O_p(1) = O_p(1)$. Therefore, $\|(ii)\|_\infty = O_p(\sqrt{s\log p/n})$. For (iii), it is easy to show that $(iii) = (ii)'$, so $\|(iii)\|_\infty = O_p(\sqrt{s\log p/n})$.

Then by Lemma 2 below, $\|\hat{\Theta}_l\|_1 = O_p(\sqrt{s_l})$ and by Assumption 3, we obtain

$$(I) = O_p\left(s\sqrt{\frac{\log p}{n}}\right)O_p(s_l) + O_p\left(\sqrt{\frac{s\log p}{n}}\right)O_p(s_l) = o_p(1).$$

For (II), we have

$$|\hat{\Theta}_l'\tilde{\Sigma}_{X\epsilon}\hat{\Theta}_l - \hat{\Theta}_l'\Sigma_{X\epsilon}\hat{\Theta}_l| \leq \|\tilde{\Sigma}_{X\epsilon} - \Sigma_{X\epsilon}\|_\infty\|\hat{\Theta}_l\|_1^2,$$

where

$$\|\tilde{\Sigma}_{X\epsilon} - \Sigma_{X\epsilon}\|_\infty = \left\| \frac{1}{n}\sum_{i=1}^{n}(X_i'\epsilon_i)(X_i'\epsilon_i)' - E[\frac{1}{n}\sum_{i=1}^{n}(X_i'\epsilon_i)(X_i'\epsilon_i)'] \right\|_\infty .$$

Since $X_i'\epsilon_i$ is $\alpha-$mixing and has exponential tail distribution, by (A.17) and (A.21), $\|\tilde{\Sigma}_{X\epsilon} - \Sigma_{X\epsilon}\|_\infty = O_p(\sqrt{\log p/n})$ with probability tending to one. Therefore, by assumption 3, we obtain $(II) = O_p(\sqrt{\log p/n})\ O_p(s_l) = o_p(1)$.

For (III), by Lemma 3.1 in the supplement material of Van De Geer et al. (2014),

$$|\hat{\Theta}_l'\Sigma_{X\epsilon}\hat{\Theta}_l - \Theta_l'\Sigma_{X\epsilon}\Theta_l| \leq \|\Sigma_{X\epsilon}\|_\infty \|\hat{\Theta}_l - \Theta_l\|_1^2 + 2\|\Sigma_{X\epsilon}\Theta_l\|_2 \|\hat{\Theta}_l - \Theta_l\|_2,$$

where, by Lemma 2, $\|\hat{\Theta}_l - \Theta_l\|_1 = O_p(s_l\sqrt{\log p/n})$ and $\|\hat{\Theta}_l - \Theta_l\|_2 = O_p(\sqrt{s_l \log p/n})$. Furthermore, note that Σ and $\Theta := \Sigma^{-1}$ are symmetric positive definite matrices, and their smallest eigenvalues are strictly greater than zero and their largest eigenvalues are bounded above. Denote $\text{Var}(\epsilon) := \sigma^2$ which is a scalar and $0 < \sigma^2 < \infty$. Then $\Sigma_{X\epsilon} := \Sigma\sigma^2$. Therefore,

$$\begin{aligned}
\|\Sigma_{X\epsilon}\|_\infty \leq & \|\Sigma_{X\epsilon}\|_2 = \Lambda_{\max}(\Sigma_{X\epsilon}) = \sigma^2\Lambda_{\max}(\Sigma) = O_p(1), \\
\|\Sigma_{X\epsilon}\Theta_l\|_\infty \leq & \|\Sigma_{X\epsilon}\|_\infty \|\Theta_l\|_\infty \leq O_p(1)\|\Theta\|_2 = O_p(1)\Lambda_{max}(\Theta) = O_p(1)/\Lambda_{min}(\Sigma) = O_p(1).
\end{aligned}$$

Thus, by Assumption 3, we obtain that $(III) = O_p(s_l^2 \log p/n) + O_p(\sqrt{s_l \log p/n}) = o_p(1)$. Therefore, in equation (A.33) we have

$$|\hat{\Theta}_l'\hat{\Sigma}_{X\epsilon}\hat{\Theta}_l - \Theta_l'\Sigma_{X\epsilon}\Theta_l| \leq (I) + (II) + (III) = o_p(1).$$

Next, we will show that

$$|\hat{\Theta}_l'X'\epsilon/\sqrt{n} - \Theta_l'X'\epsilon/\sqrt{n}| = o_p(1). \tag{A.35}$$

By Lemma 2, $\|\hat{\Theta}_l - \Theta_l\|_1 = O_p(s_l\sqrt{\log p/n})$ and by (A.17) and (A.21), $\|X'\epsilon/n\|_\infty =$

$O_p(\sqrt{\log p/n})$. Then by Assumption 3, equation (A.35) can be written as

$$\begin{aligned}|\hat{\Theta}_l' X'\epsilon/\sqrt{n} - \Theta_l' X'\epsilon/\sqrt{n}| \leq & \|\hat{\Theta}_l - \Theta_l\|_1 \|\frac{X'\epsilon}{n}\|_\infty \sqrt{n} \\ = & O_p(s_l\sqrt{\frac{\log p}{n}}) O_p(\sqrt{\frac{\log p}{n}})\sqrt{n} = O_p(\frac{s_l \log p}{\sqrt{n}}) = o_p(1),\end{aligned}$$

with probability tending to one, which completes the proof of (20).

Step 2: now we will show that $t_2 = o_p(1)$. Note that for any $l = 1, \cdots, p$,

$$\|\Delta\|_\infty = \|\sqrt{n}(\hat{\Theta}\hat{\Sigma} - I)(\hat{\beta} - \beta^0)\|_\infty \leq \sqrt{n} \max_l \|\hat{\Sigma}\hat{\Theta}_l - e_l\|_\infty \|\hat{\beta} - \beta^0\|_1,$$

where $\hat{\Theta}_l$ is the l^{th} row of $\hat{\Theta}$ written as a column vector and e_l is a $p \times 1$ column vector where the l^{th} element is one, while elsewhere being zeros. By Lemma 5.3 in Van De Geer et al. (2014), $1/\hat{\delta}_l^2 = O_p(1)$ where $\hat{\delta}_l^2$ is defined as in (17), and by (A.41), we obtain

$$\|\Delta\|_\infty \leq \sqrt{n}\, \frac{\lambda_l}{\hat{\delta}_l^2} O_p(s\sqrt{\frac{\log p}{n}}) = \sqrt{n} O_p(\sqrt{\frac{\log p}{n}}) O_p(s\sqrt{\frac{\log p}{n}}) = O_p(\frac{s \log p}{\sqrt{n}}) = o_p(1).$$

We have shown that $|\hat{\Theta}_l' \hat{\Sigma}_{X\epsilon} \hat{\Theta}_l - \Theta_l' \Sigma_{X\epsilon} \Theta_l| = o_p(1)$ and by (A.30), $|\Theta_l' \Sigma_{X\epsilon} \Theta_l| \geq a > 0$ where a is a constant. Using triangle inequality $|\hat{\Theta}_l' \hat{\Sigma}_{X\epsilon} \hat{\Theta}_l - \Theta_l' \Sigma_{X\epsilon} \Theta_l| \geq |\Theta_l' \Sigma_{X\epsilon} \Theta_l| - |\hat{\Theta}_l' \hat{\Sigma}_{X\epsilon} \hat{\Theta}_l|$, we obtain $|\hat{\Theta}_l' \hat{\Sigma}_{X\epsilon} \hat{\Theta}_l| \geq a - o_p(1) > 0$. Therefore,

$$t_2 = \frac{-\Delta}{\sqrt{\hat{\Theta}_l' \hat{\Sigma}_{X\epsilon} \hat{\Theta}_l}} = \frac{-\sqrt{n}(\hat{\Theta}_l' \hat{\Sigma}_{X\epsilon} - e_l)(\hat{\beta} - \beta^0)}{\sqrt{\hat{\Theta}_l' \hat{\Sigma}_{X\epsilon} \hat{\Theta}_l}} = o_p(1),$$

which proves (A.28). □

A.4 Proof of Lemma 1

Proof. The restricted eigenvalue condition for $\hat{\Sigma}$ in (3) implies that

$$0 < \phi_0^2 \leq \frac{\beta' \hat{\Sigma} \beta}{\|\beta_{s_0}\|_2^2} \leq \frac{\beta' \hat{\Sigma} \beta s}{\|\beta_{s_0}\|_1^2},$$

where for the second inequality we utilize the norm inequality $\sqrt{s}\|\beta_{s_0}\|_2 \geq \|\beta_{s_0}\|_1$. Rearranging the above inequality, we have

$$\|\beta_{s_0}\|_1^2 \leq (\beta'\hat{\Sigma}\beta)s/\phi_0^2,$$

which completes the proof. □

A.5 $\hat{\Theta}$ as approximation of Σ^{-1}

In this section, we closely follow Van De Geer et al. (2014) and Kock (2016) to check whether $\hat{\Theta}$ is a good approximation of Σ^{-1}. The first order condition of (15) implies

$$X'_{-j}(X_j - X_{-j}\hat{\gamma}_j)/n = \lambda_j\hat{\tau}_j. \tag{A.36}$$

Note that $\hat{\gamma}'_j\lambda_j\hat{\tau}_j = \lambda_j\|\hat{\gamma}_j\|_1$. Then left-multiplying $\hat{\gamma}'_j$ on both sides of (A.36) implies

$$\hat{\gamma}'_j X'_{-j}(X_j - X_{-j}\hat{\gamma}_j)/n = \lambda_j\|\hat{\gamma}_j\|_1. \tag{A.37}$$

Therefore, plugging the above equation into (17), we have

$$\begin{aligned}\hat{\delta}_j^2 &= \frac{1}{n}(X_j - X_{-j}\hat{\gamma}_j)'(X_j - X_{-j}\hat{\gamma}_j) + \frac{1}{n}\hat{\gamma}'_j X'_{-j}(X_j - X_{-j}\hat{\gamma}_j) \\ &= \frac{1}{n}[(X_j - X_{-j}\hat{\gamma}_j)' + \hat{\gamma}'_j X'_{-j}](X_j - X_{-j}\hat{\gamma}_j) \\ &= \frac{1}{n}X'_j(X_j - X_{-j}\hat{\gamma}_j).\end{aligned} \tag{A.38}$$

By definition of $\hat{C}_j$ (j^{th} row of matrix $\hat{C}$) in (16), we have $X_j - X_{-j}\hat{\gamma}_j = X\hat{C}_j$, and by the definition of $\hat{\Theta}_j = \hat{C}_j/\hat{\delta}_j^2$ in (18), equation (A.38) becomes

$$\hat{\delta}_j^2 = \frac{1}{n}X'_j X\hat{C}_j, \qquad \text{or} \qquad \frac{1}{n}X'_j X\hat{\Theta}_j = 1. \tag{A.39}$$

where $\hat{\Theta}_j$ is the j^{th} row of $\hat{\Theta}$ written as a column vector. Thus we can see that $\hat{\Theta}$ is a good approximation of the inverse of the Gram matrix $\hat{\Sigma} := X'X/n$.

Next, we look into the approximation error $\|\hat{\Theta}\hat{\Sigma} - I\|_\infty$, or specifically the j^{th} column

of the approximation error, which is $\|\hat{\Sigma}\hat{\Theta}_j - e_j\|_\infty$ for all $j = 1, \cdots, p$, where e_j is the j^{th} column of the identity matrix. By the definition of $\hat{\tau}$ in (10), $\|\hat{\tau}\|_\infty \leq 1$. Taking the norm on both sides of (A.36) and using $\hat{\Theta}_j = \hat{C}_j/\hat{\delta}_j^2$, we obtain

$$\begin{aligned} \|X'_{-j}(X_j - X_{-j}\hat{\gamma}_j)\|_\infty/n = \|X'_{-j}X\hat{C}_j\|_\infty = \|\lambda_j\hat{\tau}_j\|_\infty, \\ \|X'_{-j}X\hat{\Theta}_j\|_\infty/n = \lambda_j\|\hat{\tau}_j\|_\infty/\hat{\delta}^2 \leq \lambda_j/\hat{\delta}_j^2. \end{aligned} \tag{A.40}$$

By the definition of X_{-j} and $\hat{\Sigma} := X'X/n$ and by (A.39), we have $\|X'_{-j}X\hat{\Theta}_j\|_\infty = \|\hat{\Sigma}\hat{\Theta}_j - e_j\|_\infty$. Thus (A.40) can be written as

$$\|\hat{\Sigma}\hat{\Theta}_j - e_j\|_\infty \leq \lambda_j/\hat{\delta}_j^2. \tag{A.41}$$

Next, we formally investigate the asymptotic properties of $\hat{\Theta}$.

Asymptotic properties of $\hat{\Theta}$

Let Θ denote the population value of $\hat{\Theta}$ such that $\Theta := E(\hat{\Theta}) := \Sigma^{-1}$. First, partitioning Σ^{-1} into the first element and the remaining ones gives

$$\begin{pmatrix} \Sigma_{1,1} & \Sigma_{1,-1} \\ \Sigma_{-1,1} & \Sigma_{-1,-1} \end{pmatrix}^{-1} = \begin{pmatrix} \overbrace{(\Sigma_{1,1} - \Sigma_{1,-1}\Sigma_{-1,-1}^{-1}\Sigma_{-1,1})^{-1}}^{\Theta_{1,1}} & \overbrace{-\Theta_{1,1}\Sigma_{1,-1}\Sigma_{-1,-1}^{-1}}^{\Theta_{1,-1}} \\ -\Sigma_{-1,-1}^{-1}\Sigma_{-1,1}\Theta_{1,1} & (\Sigma_{-1,-1} - \Sigma_{-1,1}\Sigma_{1,1}^{-1}\Sigma_{1,-1})^{-1} \end{pmatrix},$$

where '-1' indicates all the rows (columns) excluding the first row (column). More generally, for the j^{th} row and column of Θ, we can write

$$\Theta_{j,j} = (\Sigma_{j,j} - \Sigma_{j,-j}\Sigma_{-j,-j}^{-1}\Sigma_{-j,j})^{-1}, \qquad \Theta_{j,-j} = -\Theta_{j,j}\Sigma_{j,-j}\Sigma_{-j,-j}^{-1}. \tag{A.42}$$

Denote γ_j the population value of $\hat{\gamma}_j$. Then

$$\gamma_j := \arg\min_\gamma \frac{1}{n}\sum_{i=1}^{n} E(X_{i,j} - X'_{i,-j}\gamma)^2.$$

Then the first order condition of the above equation implies,

$$\gamma_j = [\frac{1}{n}\sum_{i=1}^{n} E(X'_{i,-j}X_{i,-j})]^{-1}[\frac{1}{n}\sum_{i=1}^{n} E(X'_{i,-j}X_{i,j})] = \Sigma^{-1}_{-j,-j}\Sigma_{-j,j}. \tag{A.43}$$

Thus, (A.42) and (A.43) implies that $\Theta_{j,-j} = -\Theta_{j,j}\gamma'_j$. Denoting δ_j^2 the population value of $\hat{\delta}_j^2$ and utilizing (A.43), we obtain

$$\begin{aligned}
\delta_j^2 &= E[\frac{1}{n}\sum_{i=1}^{n} E(X_{i,j} - X'_{i,-j}\gamma_j)^2] \\
&= \Sigma_{j,j} + \Sigma_{j,-j}\Sigma^{-1}_{-j,-j}\Sigma_{-j,j} - 2\Sigma_{j,-j}\Sigma^{-1}_{-j,-j}\Sigma_{-j,j} \\
&= \Sigma_{j,j} - \Sigma_{j,-j}\Sigma^{-1}_{-j,-j}\Sigma_{-j,j} = \frac{1}{\Theta_{j,j}},
\end{aligned}$$

where the last equality comes from (A.42). Therefore, $\Theta_{j,j} = 1/\delta_j^2$ and $\Theta_{j,-j} = -\gamma'_j/\hat{\delta_j}^2$. Then it follows that $\Theta = T^{-2}C$, where C is the population value of $\hat{C}$ in (16) (by replacing $\hat{\gamma}_j$ with γ_j) and T^2 is the population value of $\hat{T}^2$ in (16) (by replacing $\hat{\delta}_j^2$ with δ_j^2).

Formally, the following lemma derives the rate of the approximation $\hat{\Theta}_j$ and the true value Θ_j.

Lemma 2. *Suppose Assumption 1 and 2 hold, then*

$$\begin{aligned}
\|\hat{\Theta}_j - \Theta_j\|_1 &= O_p(s_j\sqrt{\frac{\log p}{n}}), \\
\|\hat{\Theta}_j - \Theta_j\|_2 &= O_p(\sqrt{\frac{s_j \log p}{n}}), \\
\|\Theta_j\|_1 &= O(\sqrt{s_j}), \\
\|\hat{\Theta}_j\|_1 &= O_p(\sqrt{s_j}).
\end{aligned}$$

Proof of Lemma 2. First, we consider $|\hat{\delta}_j^2 - \delta_j^2|$. From (A.38) we have $\hat{\delta}_j^2 = X'_j(X_j - X_{-j}\hat{\gamma}_j)/n$. Suppose $X_j = X_{-j}\gamma_j + \eta_j$ and $X_j = X_{-j}\hat{\gamma}_j + \hat{\eta}_j$, where η_j and $\hat{\eta}_j$ are residuals. Then we obtain that $\hat{\delta}_j^2 = X'_j\hat{\eta}_j/n$ and $\hat{\eta}_j = X_{-j}(\gamma_j - \hat{\gamma}_j) + \eta_j$. Plugging X_j and $\hat{\eta}_j$ into $\hat{\delta}_j^2$

gives

$$\begin{aligned}\hat{\delta}_j^2 &= \frac{1}{n}(X_{-j}\hat{\gamma}_j + \hat{\eta}_j)'[X_{-j}(\gamma_j - \hat{\gamma}_j) + \eta_j] \\ &= \frac{1}{n}\gamma_j' X_{-j}' X_{-j}(\gamma_j - \hat{\gamma}_j) + \frac{1}{n}\gamma_j' X_{-j}\eta_j + \frac{1}{n}\eta_j' X_{-j}'(\gamma_j - \hat{\gamma}_j) + \frac{1}{n}\eta_j'\eta_j.\end{aligned} \tag{A.44}$$

Therefore, we obtain

$$\begin{aligned}|\hat{\delta}_j^2 - \delta_j^2| &\leq |\frac{1}{n}\eta_j'\eta_j - \delta_j^2| + |\frac{1}{n}\eta_j' X_{-j}'(\gamma_j - \hat{\gamma}_j)| + |\frac{1}{n}\gamma_j' X_{-j}\eta_j| + |\frac{1}{n}\gamma_j' X_{-j}' X_{-j}(\gamma_j - \hat{\gamma}_j)| \\ &:= I + II + III + IV.\end{aligned} \tag{A.45}$$

For (I), note that $\delta_j = \mathrm{E}(X_j - X_{-j}\gamma_j) = \mathrm{E}(\eta_j)$. We assume η_j^2 is $\alpha-$mixing with exponential decaying mixing coefficients as in Assumption 1. Then by (A.17) and (A.21), we obtain $|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\eta_{i,j}^2 - \mathrm{E}\eta_{i,j}^2| = O_p(1)$. Therefore,

$$|\frac{1}{n}\eta_j'\eta_j - \delta_j^2| = |\frac{1}{n}\sum_{i=1}^{n}\eta_{i,j}^2 - \mathrm{E}\eta_{i,j}^2| = O_p(\frac{1}{\sqrt{n}}). \tag{A.46}$$

For (II), we have

$$|\frac{1}{n}\eta_j' X_{-j}'(\gamma_j - \hat{\gamma}_j)| \leq \frac{1}{n}\|\eta_j' X_{-j}\|_\infty \|\gamma_j - \hat{\gamma}_j\|_1, \tag{A.47}$$

where $\frac{1}{n}\|\eta_j' X_{-j}\|_\infty = \max_{k\in\{1,\cdots,p\}\setminus\{j\}}|\frac{1}{n}\sum_{i=1}^{n}X_{i,k}\eta_{i,j}|$. Note that $X_{i,k}\eta_{i,j}$ is $\alpha-$mixing with exponential decaying tail distribution. Then by (A.17) and (A.21), we obtain

$$\frac{1}{n}\|\eta_j' X_{-j}\|_\infty = O_p(\sqrt{\log p/n}). \tag{A.48}$$

Together with $\|\gamma_j - \hat{\gamma}_j\|_1 = O_p(s_j\sqrt{\log p/n})$, (A.47) can be bounded

$$|\frac{1}{n}\eta_j' X_{-j}'(\gamma_j - \hat{\gamma}_j)| = O_p(\sqrt{\frac{\log p}{n}})O_p(s_j\sqrt{\frac{\log p}{n}}) = O_p(\frac{s_j\log p}{n}). \tag{A.49}$$

For (III), we have

$$|\frac{1}{n}\gamma_j' X_{-j}\eta_j| \leq \|\frac{1}{n}X_{-j}'\eta_j\|_\infty\|\gamma_j\|_1. \tag{A.50}$$

Note that $X_j = X_{-j}\gamma_j + \eta_j$. we can bound $\Sigma_{j,j}$ as

$$E(X_j'X_j) = \Sigma_{j,j} \geq E[(X_{-j}\gamma_j)'X_{-j}\gamma_j] = \gamma_j'\Sigma_{-j,-j}\gamma_j \geq \Lambda_{min}^2\|\gamma_j\|_2^2, \tag{A.51}$$

where Λ_{min} is the smallest eigenvalue of $\Sigma_{-j,-j}$ (i.e., removing j^{th} row and column from Σ gives $\Sigma_{-j,-j}$). Since Σ is a symmetric positive definite matrix, so $\Lambda_{min} \geq a > 0$, thus $1/\Lambda_{min}^2 = O(1)$. Then the above inequality implies that $\|\gamma_j\|_2 \leq \sqrt{\Sigma_{j,j}}/\Lambda_{min}$. Further utilizing the norm inequality $\|\gamma_j\|_1 \leq \sqrt{s_j}\|\gamma_j\|_2$, we obtain $\|\gamma_j\|_1 \leq \sqrt{s_j\Sigma_{j,j}}/\Lambda_{min}$. Therefore, by (A.48), inequality (A.50) can be bounded as

$$|\frac{1}{n}\gamma_j'X_{-j}\eta_j| = O_p(\sqrt{\frac{\log p}{n}})O_p(\sqrt{s_j}) = O_p(\sqrt{\frac{s_j\log p}{n}}).$$

For (IV), the first order condition of nodewise LASSO in (A.36) implies

$$\lambda_j\hat{\tau}_j + \frac{1}{n}X_{-j}'X_{-j}\hat{\gamma}_j - \frac{1}{n}X_{-j}'X_j = 0.$$

Plugging $X_j = X_{-j}\gamma_j + \eta_j$ into the above equation gives

$$\frac{1}{n}X_{-j}'X_{-j}(\gamma_j - \hat{\gamma}_j) = \lambda_j\hat{\tau}_j - \frac{1}{n}X_{-j}'\eta_j.$$

By (A.48) and $\lambda_j \asymp \sqrt{\log p/n}$, $\|\hat{\tau}_j\|_\infty \leq 1$, we obtain

$$\|\frac{1}{n}X_{-j}'X_{-j}(\gamma_j - \hat{\gamma}_j)\|_\infty \leq \|\frac{1}{n}X_{-j}'\eta_j\|_\infty + \lambda_j\|\hat{\tau}_j\|_\infty = O_p(\sqrt{\frac{\log p}{n}}).$$

Note that by (A.51), $\|\gamma_j\|_2 = O(1)$. Then using the norm inequality, we have $\|\gamma_j\|_1 \leq \sqrt{s_j}\|\gamma_j\|_2 = O(\sqrt{s_j})$. Therefore, (IV) can be bounded as

$$|\frac{1}{n}\gamma_j'X_{-j}'X_{-j}(\gamma_j - \hat{\gamma}_j)| \leq \|\frac{1}{n}X_{-j}'X_{-j}(\gamma_j - \hat{\gamma}_j)\|_\infty\|\gamma_j\|_1 = O_p(\sqrt{\frac{s_j\log p}{n}}). \tag{A.52}$$

Note that $\max_j(s_j\log p/n) = o(1)$, thus for any $j = 1, \cdots, p$, $s_j\log p/n \leq \sqrt{s_j\log p/n}$.

Therefore, we have

$$|\hat{\delta}_j^2 - \delta_j^2| = O_p(\sqrt{\frac{s_j \log p}{n}}).$$

By Lemma 5.3 in Van De Geer et al. (2014), we have $\frac{1}{\hat{\delta}_j^2} = O_p(1)$ and $\frac{1}{\delta_j^2} = O(1)$. Then it follows

$$\left|\frac{1}{\hat{\delta}_j^2} - \frac{1}{\delta_j^2}\right| \leq \frac{|\hat{\delta}_j^2 - \delta_j^2|}{\hat{\delta}_j^2 \delta_j^2} = O_p(\sqrt{\frac{s_j \log p}{n}}).$$

Then, by the definition of $\hat{\Theta}$ and $\hat{C}$ in (18) and (16), we obtain

$$\begin{aligned}
\|\hat{\Theta}_j - \Theta_j\|_1 &= \|\frac{\hat{C}_j}{\hat{\delta}_j^2} - \frac{C_j}{\delta_j^2}\|_1 = \|\frac{1-\hat{\gamma}_j}{\hat{\delta}_j^2} - \frac{1-\gamma_j}{\delta_j^2}\|_1 \\
&\leq \|\frac{1}{\hat{\delta}_j^2} - \frac{1}{\delta_j^2}\|_1 + \|\frac{\hat{\gamma}_j}{\hat{\delta}_j^2} - \frac{\gamma_j}{\hat{\delta}_j^2} + \frac{\gamma_j}{\hat{\delta}_j^2} - \frac{\gamma_j}{\delta_j^2}\|_1 \\
&\leq \|\frac{1}{\hat{\delta}_j^2} - \frac{1}{\delta_j^2}\|_1 + \|\frac{1}{\hat{\delta}_j^2}\|_1 \|\hat{\gamma}_j - \gamma_j\|_1 + \|\gamma_j\|_1 \|\frac{1}{\hat{\delta}_j^2} - \frac{1}{\delta_j^2}\|_1 \qquad \text{(A.53)} \\
&= O_p(\sqrt{\frac{s_j \log p}{n}}) + O_p(1) O_p(s_j \sqrt{\frac{\log p}{n}}) + O_p(\sqrt{s_j}) O_p(\sqrt{\frac{s_j \log p}{n}}) \\
&= O_p(s_j \sqrt{\frac{\log p}{n}}).
\end{aligned}$$

Next, we will bound $\|\hat{\Theta}_j - \Theta_j\|_2$. Note that $\|\hat{\gamma}_j - \gamma_j\|_2 = O_p(\sqrt{s_j \log p / n})$ and $\|\frac{1}{\hat{\delta}_j^2} - \frac{1}{\delta_j^2}\|_2 = \|\frac{1}{\hat{\delta}_j^2} - \frac{1}{\delta_j^2}\|_1$ and $\|\frac{1}{\hat{\delta}_j^2}\|_2 = \|\frac{1}{\hat{\delta}_j^2}\|_1$ since they are both scalars. Similarly to (A.53) we have

$$\begin{aligned}
\|\hat{\Theta}_j - \Theta_j\|_2 &\leq \|\frac{1}{\hat{\delta}_j^2} - \frac{1}{\delta_j^2}\|_2 + \|\frac{1}{\hat{\delta}_j^2}\|_2 \|\hat{\gamma}_j - \gamma_j\|_2 + \|\gamma_j\|_2 \|\frac{1}{\hat{\delta}_j^2} - \frac{1}{\delta_j^2}\|_2 \\
&= O_p(\sqrt{\frac{s_j \log p}{n}}) + O_p(1) O_p(\sqrt{\frac{s_j \log p}{n}}) + O_p(1) O_p(\sqrt{\frac{s_j \log p}{n}}) \\
&= O_p(\sqrt{\frac{s_j \log p}{n}}).
\end{aligned}$$

Next, by the definition of Θ and $\sqrt{\log p/n} = o_p(1)$, we obtain

$$\|\Theta_j\|_1 \leq \|\frac{1}{\delta_j^2}\|_1 \|C_j\|_1 \leq \|\frac{1}{\delta_j^2}\|_1 + \|\frac{1}{\delta_j^2}\|_1 \|\gamma_j\|_1 = O(\sqrt{s_j}),$$
$$\|\hat{\Theta}_j\|_1 \leq \|\hat{\Theta}_j - \Theta_j\|_1 + \|\Theta_j\|_1 = O_p(s_j\sqrt{\frac{\log p}{n}}) + O(\sqrt{s_j}) = O_p(\sqrt{s_j}),$$

which completes the proof of Lemma 2. □

A.6 Proof of Proposition 2.1

Proof. We utilize the self-normalized sums properties in Lemma 4 under weak dependence to bound tuning parameters λ_1 and λ_2. To choose appropriate values for turning parameters such that the penalty level is large enough to cancel out noises from estimation errors, we need to ensure that $\mathbb{P}(\|X'\epsilon\|_\infty/n \leq \lambda_0/2)$ is close to one. Or equivalently we want to show that

$$\mathbb{P}(\|X'\epsilon\|_\infty/n > \lambda_0/2) \leq \alpha, \tag{A.54}$$

where α is a small positive number. First, suppose that all $X'_{i,j}s$ are normalized, such that for all $j = 1, \ldots, p$, $\frac{1}{n}\sum_{i=1}^n X_{i,j}^2\epsilon_i^2 \to \sigma^2$ as $n \to \infty$. Let G denote an event such that $G = \left\{\max_{j\leq p}|\frac{1}{n}\sum_{i=1}^n X_{i,j}^2\epsilon_i^2 - \sigma^2| \leq \frac{\sigma^2}{\log n}\right\}$. Suppose that when $n \to \infty$, $\mathbb{P}(G) \to 1$, and on G, $\frac{1}{n}\sum_{i=1}^n X_{i,j}^2\epsilon_i^2 \leq (1 + 1/\log n)\sigma^2$. The definition of G ensures that $\frac{1}{n}\sum_{i=1}^n X_i^2\epsilon_i^2$ converges to σ^2 at the rate of $\log n$. Then, utilizing the union bound in (A.54), we have

$$\mathbb{P}(\|X'\epsilon\|_\infty/n > \lambda_0/2) \leq \mathbb{P}(\|X'\epsilon\|_\infty/n > \lambda_0/2, G) + \mathbb{P}(G^C) \tag{A.55}$$
$$= \mathbb{P}(\max_j|\frac{1}{n}\sum_{i=1}^n X_{i,j}\epsilon_i| > \frac{\lambda_0}{2}, G) + \mathbb{P}(G^C) \tag{A.56}$$
$$\leq p\ \mathbb{P}\left(|\frac{1}{n}\sum_{i=1}^n X_{i,j}\epsilon_i| > \lambda_0/2, G\right) + \mathbb{P}(G^C) \leq \alpha. \tag{A.57}$$

Note that on G, we have $\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\epsilon_i^2\right)^{1/2} \geq (1+1/\log n)^{1/2}\sigma$. So (A.57) can be written as

$$\mathbb{P}(\|X'\epsilon\|_\infty/n > \lambda_0/2) \leq p\ \mathbb{P}\left(\left\{\frac{|\frac{1}{n}\sum_{i=1}^{n} X_{i,j}\epsilon_i|}{\left(\frac{1}{n}\sum_{i=1}^{n} X_{i,j}^2\epsilon_i^2\right)^{1/2}} > \frac{\lambda_0}{2\sigma(1+1/\log n)^{1/2}}\right\} \cap G\right) + \mathbb{P}(G^C) \tag{A.58}$$

$$\leq 2p\ \mathbb{P}\left(\frac{\sum_{i=1}^{n} X_{i,j}\epsilon_i/\sqrt{n}}{\left(\sum_{i=1}^{n} X_{i,j}^2\epsilon_i^2/n\right)^{1/2}} > \frac{\lambda_0\sqrt{n}}{2\sigma(1+1/\log n)^{1/2}}\right) + o(1) \tag{A.59}$$

$$\leq \alpha. \tag{A.60}$$

Applying the self-normalization theorem of Chen et al. (2016) given in Lemma 4 below on (A.59) gives

$$\mathbb{P}\left(\frac{\sum_{i=1}^{n} X_{i,j}\epsilon_i/\sqrt{n}}{\left(\sum_{i=1}^{n} X_{i,j}^2\epsilon_i^2/n\right)^{1/2}} > \frac{\lambda_0\sqrt{n}}{2\sigma(1+1/\log n)^{1/2}}\right) \to 1 - \Phi\left(\frac{\lambda_0\sqrt{n}}{2\sigma(1+1/\log n)^{1/2}}\right).$$

Together with (A.60), this implies

$$\begin{aligned} 2p\left[1-\Phi\left(\frac{\lambda_0\sqrt{n}}{2\sigma(1+1/\log n)^{1/2}}\right)\right] &\leq \alpha - o(1), \\ \lambda_0 &\geq \frac{2\sigma}{\sqrt{n}}(1+\frac{1}{\log n})^{1/2}\Phi^{-1}(1-\frac{\alpha}{2p}). \end{aligned} \tag{A.61}$$

Since $\lambda_1/n = 2\lambda_0$, we obtain the first part of (8). Also, since $\lambda_2/n = O_p(s\log p/(np))$ and $\frac{\sqrt{\log p}}{\sqrt{n}} \asymp \lambda_1/n$, and we assume that s is small relative to p, so we approximate $\frac{s\sqrt{\log p}}{\sqrt{n}} \approx \lambda_1/n$, which gives the second part of (8). However, σ is unknown. We implement a recursive procedure to evaluate the unknown variance following Algorithm A.1 in Belloni et al. (2012). In particular, we first set $\sigma = 1$ to evaluate the penalized regression and get a preliminary empirical variance $\hat{\sigma}^2$. Then we refine the estimation result using the updated

empirical variance for σ. We repeat this exercise K times to get the final estimate.[14] $\square$

A.7 Auxiliary lemmas

Lemma 3 (Dendramis et al. (2019), Lemma 1)**.** *Let $\{X\}_n$ be a sequence that satisfies Assumption 1. Then*

$$\mathbb{P}\left(\frac{1}{\sqrt{n}}\left|\sum_{i=1}^{n} X_i\right| \geq \xi\right) \leq c_0\left[\exp(-c_1\xi^2) + \exp\left(-c_2\left(\frac{\xi\sqrt{n}}{\log^2 n}\right)^s\right)\right],$$

where $s = q/(q+1)$, and constants c_0, c_1, c_2 do not depend on ξ and i.

Proof: see Dendramis et al. (2019).

Lemma 4 (Chen et al. (2016), Theorem 4.1)**.** *Let weekly dependent random variable X_i be zero-mean, $E(X_i) = 0$. Write $S_{k,m} = \sum_{i=k+1}^{k+m} X_i$. Suppose for a positive constant c, $E(S_{k,m}^2) \geq c^2 m$ for any $k \geq 0$, $m \geq 1$. Let $m_1 > m_2 > 0$, $m^* = m_1 + m_2$, $k = \lfloor n/m^* \rfloor$.* [15] *For $1 \leq j \leq k$, denote $H_{j,1} = \{i : (j-1)m^* + 1 \leq i \leq (j-1)m^* + m_1\}$ and $H_{j,2} = \{i : (j-1)m^* + m_1 + 1 \leq i \leq jm^*\}$. Define $Y_j := \sum_{i \in H_{j,1}} X_i$ and $W_n := \sum_{j=1}^{k} Y_j / (\sum_{j=1}^{k} Y_j^2)^{1/2}$. Then*

$$\frac{\mathbb{P}(W_n \geq t)}{1 - \Phi(t)} \to 1,$$

uniformly in $0 \leq t \leq o(n^{1/8})$.

Proof: see Chen et al. (2016).

[14] For instance in Belloni et al. (2012), $K = 15$.

[15] We use $\lfloor \cdot \rfloor$ to denote the integer part of a floating number.

References

Asness, C. S., T. J. Moskowitz, and L. H. Pedersen (2013): "Value and Momentum Everywhere," *Journal of Finance*, 68, 929–985.

Babii, A., E. Ghysels, and J. Striaukas (2019): "Estimation and HAC-based Inference for Machine Learning Time Series Regressions," *SSRN Electronic Journal*.

Belloni, A., D. Chen, V. Chernozhukov, and C. Hansen (2012): "Sparse Models and Methods for Optimal Instruments With an Application to Eminent Domain," *Econometrica*, 80, 2369–2429.

Belloni, A. and V. Chernozhukov (2012): "High Dimensional Sparse Econometric Models: An Introduction," *SSRN Electronic Journal*.

Belloni, A., V. Chernozhukov, and C. Hansen (2014): "Inference on Treatment Effects After Selection Among High-dimensional Controls," *Review of Economic Studies*, 81, 608–650.

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

Billingsley, P. (1995): *Probability and Measure*, Wiley series in probability and mathematical statistics, third edit ed.

Bogdan, M., E. van den Berg, C. Sabatti, W. Su, and E. J. Candès (2015): "SLOPE - Adaptive Variable Selection via Convex Optimization," *Annals of Applied Statistics*, 9, 1103–1140.

Bondell, H. D. and B. J. Reich (2008): "Simultaneous Regression Shrinkage, Variable Selection, and Supervised Clustering of Predictors with OSCAR," *Biometrics*, 64, 115–123.

Buhlmann, P. and S. Van de Geer (2011): *Statistics for High-Dimensional Data: Methods, Theory and Applications*.

Caner, M. and A. B. Kock (2018): "Asymptotically honest confidence regions for high dimensional parameters by the desparsified conservative Lasso," *Journal of Econometrics*, 203, 143–168.

Chen, X., Q. M. Shao, W. B. Wu, and L. Xu (2016): "Self-normalized cramér-type moderate deviations under dependence," *Annals of Statistics*, 44, 1593–1617.

Davidson, J. (1994): *Stochastic Limit Theory: An Introduction for Econometricians*, Oxford University Press.

Dendramis, Y., L. Giraitis, and G. Kapetanios (2019): "Estimation of Time-Varying Covariance Matrices for Large Datasets," .

Feng, G., S. Giglio, and D. Xiu (2020): "Taming the Factor Zoo: A Test of New Factors," *The Journal of Finance*, 1–76.

Figueiredo, M. A. T. and R. D. Nowak (2016): "Ordered Weighted L1 Regularized Regression with Strongly Correlated Covariates: Theoretical Aspects," *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, 41, 930–938.

Freyberger, J., A. Neuhierl, M. Weber, A. Beber, J. Berk, O. Bondarenko, S. Bryzgalova, J. Campbell, J. Chen, J. Coval, K. Daniel, V. Demiguel, A. Donangelo, G. Fama, K. French, E. Hansen, L. Hansen, B. Holcblat, A. Karolyi, B. Kelly, and L. Kogan (2019): "Dissecting Characteristics Nonparametrically," *Review of Financial Studies*.

Kleibergen, F. (2009): "Tests of risk premia in linear factor models," *Journal of Econometrics*, 149, 149–173.

Kock, A. B. (2016): "Oracle inequalities, variable selection and uniform inference in high-dimensional correlated random effects panel data models," *Journal of Econometrics*, 195, 71–85.

Kock, A. B. and H. Tang (2019): "Uniform inference in high-dimensional dynamic panel data models with approximately sparse fixed effects," *Econometric Theory*, 35, 295–359.

Kozak, S., S. Nagel, and S. Santosh (2020): "Shrinking the cross-section," *Journal of Financial Economics*, 135, 271–292.

Sun, C. (2019): "Dissecting the Factor Zoo : A Correlation-Robust Approach," *Working Paper*.

Tibshirani, R. (1996): "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society*, 58, 267–288.

Van De Geer, S., P. Bühlmann, Y. Ritov, and R. Dezeure (2014): "On asymptotically optimal confidence regions and tests for high-dimensional models," *Annals of Statistics*, 42, 1166–1202.

Yuan, M. and Y. Lin (2006): "Model Selection and Estimation in Regression with Grouped Variables," *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 68, 49–67.

Zeng, X. and M. A. T. Figueiredo (2015): "The Ordered Weighted L1 Norm: Atomic Formulation, Projections, and Algorithms," .

Zou, H. and T. Hastie (2005): "Regularization and Variable Selection via the Elastic-Net," *Journal of the Royal Statistical Society*, 67, 301–320.