On the Inference of a LASSO-type Estimator with Highly Correlated Variables *

Chuanping Sun †

Abstract

The LASSO-type shrinkage methods have become increasingly popular in the big data era. However, variable correlations can significantly compromise the stability and validity of such estimators. This paper advances the development of a correlation-robust LASSO-type estimator.We develop the (non)asymptotic properties of the this estimator under less restrictive conditions, including the $\alpha-$mixing condition and accommodating heavier tails than the standard *i.i.d.* sub-Gaussian setting. Furthermore, we propose a de-biased version of this estimator and establish its asymptotic normality. Through simulated data, we demonstrate that the de-biased estimator significantly reduces estimation errors. Empirically, we apply it to identify crucial factors from the factor zoo, revealing that, despite high correlation with numerous other factors, the 'market' factor is the most influential in driving cross-sectional asset returns. Our findings also highlight the significant impact of 'liquidity,' 'profitability,' and 'momentum'-related factors on cross-sectional asset returns.

JEL classification: C52, C55, C58
Keywords: LASSO, Inference, Weak Dependence, De-biased estimator

*Preliminary version.

†Faculty of Finance, Bayes Business School (formerly Cass), City University of London. 106 Bunhill Row, London EC1Y 8TZ, United Kingdom, chuanping.sun@city.ac.uk

1 Introduction

The LASSO (Tibshirani, 1996) and related shrinkage methods have gained tremendous popularity in recent research agendas related to economic research, see Callot et al. (2021), Kock (2016), Caner and Kock (2018), Medeiros and Mendes (2016), Babii et al. (2021), Freyberger et al. (2020), Chinco et al. (2019), Belloni and Chernozhukov (2012) and among others for example. However, it is widely recognized that high correlations among explanatory variables can significantly undermine the robustness of the LASSO shrinkage method, leading to unstable variable selections. For instance, Zou and Hastie (2005) point out ["if there is a group of variables among which the pairwise correlations are very high, the LASSO tends to select only one variable from the group and does not care which one is selected..."]. In response to this challenge, the group LASSO estimator (Yuan and Lin, 2006) has gained prominence as a strategy to mitigate the issues arising from correlated variables. The group LASSO shrinkage method organizes variables into groups based on shared characteristics, allowing the simultaneous shrinking of variables within certain groups. However, it's important to note that implementing the group LASSO shrinkage method requires prior knowledge regarding how to effectively group variables. This task is far from trivial, and the underlying structural assumptions often necessitate rigorous justification. In contrast, Figueiredo and Nowak (2016) propose the Ordered-Weighted-LASSO (OWL hereafter) estimator, which is robust to variable correlations and free from structural assumptions on variables.

This paper focuses on developing robust inference for the OWL estimator under more general assumptions. First, we relax the i.i.d. Gaussian assumption made in Figueiredo and Nowak (2016) and derive the non-asymptotic error bounds (oracle inequality) for the OWL estimator under $\alpha-$mixing conditions. We also permit fatter tails in the distribution of random variables. We show how and to what extent fat tails affect the probability of the oracle inequality. It is worth stressing that a few recent papers have studied the error bound of the LASSO estimator under i.i.d variables with sub-Gaussian tail bounds, see Kock (2016) for example. Our assumptions in this paper are more general: first, we assume $\alpha-$mixing condition instead of i.i.d for random variables. Second, although we

restrict that the tail behaviour of random variables is exponentially decaying, we permit fatter-than-sub-Gaussian tails.

Then, we further devise a bias-corrected version of the OWL estimator following the recent development in the nodewise LASSO technique, see Callot et al. (2021) , Van De Geer et al. (2014) and Kock (2016) for example. We show that the de-biased OWL estimator is asymptotically normally distributed and we derive the confidence interval for the estimate given a significance level.

Empirically, we employ the de-biased OWL estimator to infer significant factors from the factor zoo, while many factors are highly correlated which often dampens the robustness of inference in standard econometric frameworks. We obtain the data from CRSP and Compustat datasets between January 1980 and March 2022 for the US stock returns and firm characteristics. We use portfolio sorting to construct firm characteristic based factors (a.k.a anomaly factors) as well as test portfolios. Specifically, at each point of time, we sort stock in descending order according to each firm characteristic. Then, we form decile portfolios and compute the spread returns between the top and bottom decile (high-minus-low) portfolios, which is used to approximate the firm characteristic based factor returns. The pool of all decile portfolios (with respect to all firm characteristics considered) are used as the test portfolios. For robustness, we also consider the bi-variate sorted portfolios as the test assets following Feng et al. (2020).[1] After obtaining the test assets and anomaly factors, we employ the Stochastic Discount Factor (SDF) method, along with the de-biased OWL estimator, to infer risk prices for all factors in the factor zoo. First, we find excessively high correlation between some factors. For example, the 'beta' related factors are often highly correlated with 'liquidity' and 'volatility' related factors. Also, many factors are close cousins, for example several different measurements of liquidity are highly correlated. This motivates us to employ the OWL shrinkage method due to its correlation-robust properties. Our empirical findings suggest that the 'market' factor along with some 'liquidity' and 'profitability' related factors are among the strong factors driving

[1]To obtain the bi-variate sorted test portfolios, we use the 'size' firm characteristic ('mve') to form 5 by 5 sorted portfolios with any other firm characteristic in the factor zoo, before pooling all these portfolios together as the grand set of test portfolios.

asset prices. It is worth noting that, when we use the pooled decile portfolios as test assets, the 'market' factor has estimated coefficient double that of the second most important factor.[2] This finding suggests that the 'market' factor is a predominant factor to drive cross-sectional asset prices, and its importance is not compromised by its high correlation with other factors in the factor zoo. In addition, we find that 'liquidity', 'profitability' and 'momentum' related factors are also important factors to drive asset prices.

Related literature This paper builds naturally on the active and expanding literature pertaining to the LASSO estimator, in both the machine learning and 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. (2020) apply the adaptive group LASSO method to find pervasive firm characteristics to predict stock returns, while Babii et al. (2021) implement the sparse group LASSO estimator with mixed-frequency time series data for nowcasting GDP growth. Belloni and Chernozhukov (2012), Belloni et al. (2014) and Chernozhukov et al. (2018) propose the double LASSO selection procedure, also known as the double machine learning method, to remove bias from the LASSO estimates for a set of factors that are of primary interest to researchers, while having a large number of controlling factors. 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) estimator, which stabilizes LASSO selections when covariates are correlated. Although the EN estimator improves the LASSO shrinkage method under highly correlated settings, Hiraki and Sun (2022) shows its performance is substantially outperformed by the OWL estimator. Kozak and Santosh (2020) employ the 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

[2]Note that all factors are scaled to have zero mean and the same variance.

by exploring the ℓ_∞ norm of parameters pair-wisely to achieve clustered selections when covariates are highly correlated. Zeng and Figueiredo (2014) and Figueiredo and Nowak (2016) propose the Ordered-Weighted-LASSO (OWL) estimator, which is a generalization of the OSCAR estimator. Figueiredo and Nowak (2016) also show that the OWL estimator is closely related to the SLOPE (Sorted ℓ_1 Penalized Estimator) by Bogdan et al. (2015): both estimators assign a fixed and decreasing weighting vector to penalize factors, while having different weighting schemes for the weighting vectors.[3] On the statistical advances for the LASSO estimator, Van De Geer et al. (2014) developed the de-sparsified LASSO estimator using the nodewise LASSO technique, which enables a bias-corrected LASSO estimator. The de-biased LASSO estimator enjoys asymptotic normality and is therefore suitable for conducting tests to infer significance. Callot et al. (2021) utilise the nodewise LASSO technique to estimate large portfolios and apply it for mean-variance portfolio optimization problems. Kock (2016) expands the de-biased LASSO estimator on panel data and develops statistical properties under i.i.d sub-Gaussian assumption. In this paper, we utilize the nodewise LASSO technique to identify the bias in the OWL estimator and therefore propose a de-biased version before deriving its asymptotic normality.

In the remainder of this paper, Section 2 outlines the OWL estimation framework and we study its (non)asymptotic properties. We also derive a de-biased version of the OWL estimator and show 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 the factor zoo to infer significant factors in the factor zoo.

2 Model

In this section, we define the Ordered-Weighted-LASSO (OWL) estimator and comprehensively compare it with the well-known LASSO estimator before deriving its asymptotic properties. Then we further develop the de-biased OWL estimator and show that it is asymptotically normally distributed.

[3]The decreasing weighting vector for the OWL estimator is linear, while it is non-linear for the SLOPE estimator.

Notation Throughout this paper, X is a $n \times p$ matrix, and y is a $n \times 1$ vector. Denote by $\zeta_j := \epsilon' X^{(j)} := \sum_{i=1}^{n} \epsilon_i X_i^{(j)} := \sum_{i=1}^{n} \zeta_{i,j}$, where ϵ is defined in (1) and $X^{(j)}$ is the j^{th} column of X, $j \in \{1, \cdots, p\}$. We denote $\hat{\Sigma} = \frac{1}{n} X'X$ as the scaled Gram Matrix of X, while $\Sigma = \mathrm{E}(\hat{\Sigma})$ 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 (point-wise) 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 $a y_n \le x_n \le b y_n$ and we write $x_n \lesssim y_n$ if $x_n \le b y_n$ for some $0 < b < \infty$. For any set s, s^c denotes the complimentary set. For two scalars 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 The OWL estimator and the comparison with the LASSO

Consider a linear regression model

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

where β^0 is a $p-$dimensional vector representing the true coefficients, with possibility that $p \gg n$. Furthermore, the columns of the design matrix X can be highly correlated. This relaxed condition on the columns of X differentiates this model from other standard high-dimensional linear regression models in the literature. It is widely acknowledged that significant collinearity in the design matrix can cause instability in the variable selection process for the LASSO estimator. To address the issues, Figueiredo and Nowak (2016) proposed the Ordered-Weighted-LASSO (OWL) estimator, which can be defined as

$$\hat{\beta} = \arg\min_{\beta} \quad \left[\frac{1}{n} ||y - X\beta||_2^2 + \frac{1}{n} \Omega_{OWL}(\beta) \right], \tag{2}$$

where the penalty term

$$\Omega_{OWL}(\beta) = \omega'|\beta|_{\downarrow} = \sum_{j=1}^{p} \omega_j |\beta|_{[j]}, \tag{3}$$

and $\omega_1 \geq \omega_2 \geq \cdots \geq \omega_p \geq 0$. This formulation for the OWL estimator is quite general; by specifying the weighting vector ω, it encompasses other shrinkage estimators.

Example 1:

Set $\omega_j = z(1 - j.q/2p)$ *where* $q \in (0, 1)$ *and* $z(\alpha)$ *is the quantile of a standard normal distribution.*

With this non-linear decreasing function with respect to j, the OWL estimator is equivalent to the SLOPE (Sorted L-One Penalized Estimator) from Bogdan et al. (2015). It was utilized for multiple hypothesis testing with false discovery rates under an orthogonal design matrix.

Example 2:

Set

$$\omega_j = \lambda_1 + \lambda_2(p - j), \quad j \in \{1, \cdots, p\}, \tag{4}$$

and $\lambda_1, \lambda_2 \geq 0$ *are two tuning parameters.*

By defining ω_j as a linear decreasing function of j, the OWL estimator encompasses the OSCAR (Octagonal Shrinkage and Clustering Algorithm for Regression) estimator by Bondell and Reich (2008). The authors show that the OSCAR estimator can mitigate multicollinearity issues in regression problems and simultaneously group highly correlated covariates into clusters without imposing structural assumptions on covariates.

Example 3:

Set $\lambda_2 = 0$ *and* $\lambda_1 > 0$ *in* (4).

In this case, $\omega_1 = \omega_2 = \cdots = \omega_j$. The weighting vector ω is a constant, corresponding to the LASSO estimator.

Note that although Examples 1 and 2 have similar settings on the penalty term (i.e. both ω_j and $\beta_{[j]}$ are ordered decreasingly), they are based on contrasting assumptions about the design matrix. The SLOPE estimator is developed under an orthogonal design,

whereas the OWL estimator in (4) accommodates and groups highly correlated covariates. Given the main focus of this paper, we concentrate sorely on the setting in Example 2, which is robust to highly correlated variables.

Next, we investigate the differences (and connections) between the OWL and the LASSO shrinkage methods. Consider a simple two dimensional case, where $p = 2$. In this case, by (4), we have $\omega_1 = \lambda_1 + \lambda_2$ and $\omega_2 = \lambda_1$. Similarly, $|\beta|_{\downarrow} = (|\beta|_{[1]}, |\beta|_{[2]})'$ and

$$|\beta|_{[1]} = \max(|\beta_1|, |\beta_2|) = \frac{1}{2}(|\beta_1| + |\beta_2| + \Big||\beta_1| - |\beta_2|\Big|),$$
$$|\beta|_{[2]} = \min(|\beta_1|, |\beta_2|) = \frac{1}{2}(|\beta_1| + |\beta_2| - \Big||\beta_1| - |\beta_2|\Big|).$$

Then the OWL penalty can be written as

$$\Omega_{OWL}(\beta) = \omega_1|\beta|_{[1]} + \omega_2|\beta|_{[2]} = \frac{\omega_1 + \omega_2}{2}(|\beta_1| + |\beta_2|) + \frac{\overbrace{\omega_1 - \omega_2}^{\lambda_2}}{2}\Big||\beta_1| - |\beta_2|\Big|, \tag{5}$$

which suggests that the OWL penalty term can be decomposed into two components: first, $|\beta_1| + |\beta_2|$, which is the same as the LASSO shrinkage method. It produces sparse variable selection; second, $\Big||\beta_1| - |\beta_2|\Big|$, which shrinks the distance between $|\beta_1|$ and $|\beta_2|$ if $|\beta_1| \neq |\beta_2|$, where the shrinkage intensity is controlled by $(\omega_1 - \omega_2)/2$. The second component in (5) encourages assigning similar values to $|\beta_1|$ and $|\beta_2|$. We regard this property as grouping. Note that, by the definition of ω, we have $\omega_1 - \omega_2 = \lambda_2$. Therefore, the turning parameter λ_2 has a direct impact on the grouping property of the OWL estimator and thus can be controlled to achieve desirable grouping intensity. Figueiredo and Nowak (2016) (Theorem 1) shows that the grouping property is influenced by the correlation between factors as well as the tuning parameter λ_2. From this decomposition analysis, we can clearly see how λ_2 affect the grouping property.

Furthermore, a detailed examination of the grouping property is provided in Appendix C. This examination employs a geometric interpretation of the penalty terms, consistent with the typical analytical approaches found in the machine learning literature.

2.2 Asymptotic properties

In this section, we extend the asymptotic framework of the OWL estimator. Figueiredo and Nowak (2016) derived the error bound of the OWL estimator under *i.i.d.* Gaussianity assumption. In this paper, we derive the error bounds of the OWL estimator under the strong mixing condition and allowing for fatter-than-sub-Gaussian tails. First, to derive the next theorem we make the following assumptions.

Assumption 1 (Random variables)**.**

(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 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 random variables to be weakly dependent, i.e. $\alpha-$mixing. This assumption is relevant to Dendramis et al. (2021), in which they derive a Bernstein-type inequality for variables under strong-mixing condition, which will be used for the proof of Theorem 2.1. Assumption 1(b) specifies tail bounds for the distributions of X_j and ϵ. Although we use an exponential type of bound, we allow tails to be fatter than 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. If $0 < q < 2$, then the random variable has fatter tails than the sub-Gaussian case. Assumption 1(c) requires X to have at least four finite moments. Note that we do not assume random variables to be bounded which is typically assumed when implementing a Bernstein type inequalities.

Let s_0 denote a subset, $s_0 \subset \{1, \cdots, p\}$ collects the index where its corresponding elements in a vector are non-zeros. Let $|s_0|$ be the cardinality of s_0. For $\beta = \{\beta_1, \cdots, \beta_p\} \in$

$\mathbf{R}^p$, denote $\beta_{s_0} := \beta_i \mathbf{1}\{i \in s_0, i = 1, \cdots, p\}$, $\beta_{s_0^c} := \beta_i \mathbf{1}\{i \notin s_0, i = 1, \cdots, p\}$. Then $\beta = \beta_{s_0} + \beta_{s_0^c}$.

Assumption 2 (Compatibility condition, Van de Geer and Bühlmann (2009)). *For all β such that $||\beta_{s_0^c}||_1 \leq 3||\beta_{s_0}||_1$, $\hat{\Sigma}$ satisfies the compatibility condition if*

$$\phi_0^2 := \min_{\substack{s_0 \subset \{1,\ldots,p\} \\ |s_0| < p}} \min_{\substack{\beta \in R^p \setminus \{0\} \\ ||\beta_{s_0^c}||_1 \leq 3||\beta_{s_0}||_1}} \frac{s\beta' \hat{\Sigma} \beta}{||\beta_{s_0}||_1^2} > 0. \tag{6}$$

Van de Geer and Bühlmann (2009) shows that Assumption 2 is a sufficient, yet weaker restriction than the commonly used restricted eigenvalue condition (Bickel et al. (2009)) for deriving oracle properties for LASSO-type estimators. In addition, the irrepresentable condition is often used to derive the variable selection consistency properties, see Fan et al. (2020) for example. Van de Geer and Bühlmann (2009) show that both the restricted eigenvalue condition and the irrepresentable condition imply the compatibility condition under mild conditions. We refer to Van de Geer and Bühlmann (2009) for a detailed discussion on various conditions and their relations for deriving oracle properties.

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). We assume $s\sqrt{\frac{\log p}{n}} = o(1)$.*

Assumption 3 specifies the growing rate of n, p and s. Next, Theorem 2.1 establishes error bounds for the prediction error and parameter estimation error of the OWL estimator under mixing condition and fat tails.

Theorem 2.1 (Error bounds). *Suppose Assumptions 1, 2 and 3 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,*

$$\frac{1}{n}||X(\hat{\beta} - \beta^0)||_2 \lesssim 4\lambda_0\sqrt{s}/\phi_0 + \lambda_0\sqrt{2s\|\beta^0\|_1} \tag{7}$$

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

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* (7) *and* (8) *hold with probability at least*

$$1 - pc_0 \left[\exp(-\frac{c_1'}{4}\kappa^2 \log p) + \exp\left(-c_2' \left(\frac{\kappa\sqrt{n \log p}}{2 \log^2 n} \right)^{\zeta} \right) \right], \tag{9}$$

where $\zeta = q/(q+1), q = q_1 q_2/(q_1 + q_2)$ *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. Note that the restriction we imposed on the growth rate of λ_2 is to ensure the linear weighting scheme for ω_j does not diverge. We utilize a Bernstein type of inequality derived for mixingales in Dendramis et al. (2021) for obtaining the probability of the oracle inequalities.

Remark 2 We analyze the probability of (7) and (8) that 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 and κ is a constant which determines the level of penalty imposed on the shrinkage estimator. Second, when p is fixed, we decompose the probability of (7) and (8) to hold into two components. The component includes n vanishes as as $n \to \infty$. Then (9) becomes $1 - pc_0 \exp(-\frac{c_1'}{4}\kappa^2 \log p)$, where c_0 and c_1' are some positive constants which depend only on the mixing coefficient α_k in Assumption 1. Therefore, to ensure high probability of these inequalities hold, we need to select κ sufficiently large such that $pc_0 \exp(-c_1''\kappa^2 \log p)$ is close to zero. This result also offers some insights on the choice of the tuning parameters λ_1 and λ_2, which are monotonic functions of κ.

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 (9) 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. $0 < q_1, q_2 < 2$. 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 (9) is to one, given other parameters holding the same.

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 and we further relax the sub-Gaussian tail restriction where we leave a parameter q that controls the fatness of the tail distribution.

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,*

$$\|\hat{\beta} - \beta^0\|_2 = O_p\left(\sqrt{\frac{s \log p}{n}}\right) = o_p(1), \qquad \|\hat{\beta} - \beta^0\|_1 = O_p\left(s\sqrt{\frac{\log p}{n}}\right) = o_p(1). \tag{10}$$

Proof: see Appendix A.2.

Corollary 2.1 establishes the convergence rate in ℓ_1 and ℓ_2 norm of the OWL estimator $\hat{\beta}$. We find that the OWL estimator achieves the same rate of convergence as the LASSO estimator. With Assumption 3, we show that the OWL estimator is a consistent estimator.

2.3 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.[4] 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

[4]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.

from estimation errors (i.e. $\mathbb{P}(\lambda_0 > 2\|X'\epsilon\|_\infty/n)$ is large), yet not too large to write off signals from variables. To achieve that, we propose a rule of thumb for hyper-parameter choice based on a similar argument to Belloni et al. (2012) but incorporating our unique setting for random variables (weak dependence and our tail restrictions).

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

$$\frac{\lambda_1}{n} = \frac{4}{\sqrt{n}}\sigma^*(1+\frac{1}{\log n})^{1/2}\Phi^{-1}(1-\frac{\alpha}{2p}), \qquad \frac{\lambda_2}{n} = \frac{\lambda_1}{n}\frac{\sqrt{\log p}}{\sqrt{n}\ p}, \tag{11}$$

where we evaluate σ^* recursively similar to Algorithm A.1 in Belloni et al. (2012) and $0 < \alpha < 1$ is a significance level.

Proof: see Appendix A.5.

The proof of Proposition 2.1 relies on a self-normalization technique from Chen et al. (2016). First, 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.[5] Equation (11) 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. 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 significance parameter α, as discussed earlier.

[5]For example, in our Monte Carlo experiments, we specify multiple choices for n and p, implementing a cross validation method to determine hyper parameters would be too time-consuming.

2.4 De-biased OWL estimator

Corollary 2.1 shows that the OWL estimator is consistent under some regularity conditions. It is, however, biased in small samples. In this section, we discuss a bias-corrected version of the OWL estimator using the nodewise LASSO technique from Van De Geer et al. (2014). Then we develop the asymptotic normality for the de-biased OWL estimator.

2.4.1 Identifying the bias of the OWL estimator

For the convenience of expression, the OWL estimator defined in (2) can be written as[6]

$$\hat{\beta} = \arg\min_{\beta} \left[\|y - X\beta\|_2^2/n + 2\omega'|\beta|_{\downarrow}/n \right]. \tag{12}$$

The first order condition of the minimization of (12) gives

$$-X'(y - X\hat{\beta})/n + \omega \odot \hat{\tau}/n = 0, \qquad \hat{\tau}_j = \begin{cases} 1 & \text{if} \quad \hat{\beta}_{[j]} > 0 \\ [-1, 1] & \text{if} \quad \hat{\beta}_{[j]} = 0 \\ -1 & \text{if} \quad \hat{\beta}_{[j]} < 0. \end{cases} \tag{13}$$

where $\odot$ denotes point-wise product of two vectors, and $\hat{\tau} = (\hat{\tau}_1, \cdots, \hat{\tau}_p)$ is the sub-gradient of $|\hat{\beta}|_{\downarrow}$. We further utilize the equality $y = X\beta^0 + \epsilon$ and $\hat{\Sigma} = X'X/n$. Then (13) can be written as

$$\hat{\Sigma}(\hat{\beta} - \beta^0) + \omega \odot \hat{\tau}/n = X'\epsilon/n. \tag{14}$$

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

$$\hat{\beta} - \beta^0 + \hat{\Theta}\omega \odot \hat{\tau}/n = \hat{\Theta}X'\epsilon/n - \Delta/\sqrt{n}, \tag{15}$$

$$\Delta = \sqrt{n}(\hat{\Theta}\hat{\Sigma} - I)(\hat{\beta} - \beta^0), \tag{16}$$

[6]Without loss of generality, we extract 2 out of the weighting vector ω. Note that ω is dependent only on tuning parameters λ_1 and λ_2 according to (11). So, for the convenience of expression, we keep the same notation here for ω.

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

$$\hat{b} = \hat{\beta} + \hat{\Theta}\omega \odot \hat{\tau}/n = \hat{\beta} + \hat{\Theta}X'(Y - X\hat{\beta})/n, \tag{17}$$

where the second equation holds in view of (13). So the bias is identified as $\widehat{bias} = \hat{\Theta}\omega \odot \hat{\tau}/n = \hat{\Theta}X'(Y - X\hat{\beta})/n$. Therefore, the construction of the $\hat{\Theta}$ is the key to recover the bias item of the OWL estimator. We follow Van De Geer et al. (2014) and Kock (2016) using the nodewise LASSO technique to obtain $\hat{\Theta}$. First, the nodewise LASSO estimator is defined as

$$\hat{\gamma}_j = \underset{\gamma \in R^{p-1}}{\arg\min} \left(\|X_j - X_{-j}\gamma\|_2^2/n + 2\lambda_j\|\gamma_j\|_1 \right), \tag{18}$$

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

$$\hat{C} := \begin{pmatrix} 1 & -\hat{\gamma}_{1,2} & \cdots & -\hat{\gamma}_{1,p} \\ -\hat{\gamma}_{2,1} & 1 & \cdots & -\hat{\gamma}_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ -\hat{\gamma}_{p,1} & -\hat{\gamma}_{p,2} & \cdots & 1 \end{pmatrix}, \qquad \hat{T}^2 := \mathrm{diag}(\hat{\delta}_1^2, \hat{\delta}_2^2, \cdots, \hat{\delta}_p^2,), \tag{19}$$

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

$$\hat{\delta}_j^2 = \|X_j - X_{-j}\hat{\gamma}_j\|_2^2/n + \lambda\|\hat{\gamma}_j\|_1. \tag{20}$$

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

$$\hat{\Theta} := \hat{T}^{-2}\hat{C}. \tag{21}$$

For a close consideration of whether $\hat{\Theta}$ is a good approximation of Σ^{-1} as well as some statistical properties of $\hat{\Theta}$, see Appendix A.4.

2.4.2 Inference on the de-biased OWL estimator

Denote $\Sigma_{X\epsilon} := \mathrm{E}[\frac{1}{n}\sum_{i=1}^{n}(X_i'\epsilon_i)(X_i'\epsilon_i)']$, $\hat{\Sigma}_{X\epsilon} := \frac{1}{n}\sum_{i=1}^{n}[(X_i'\hat{\epsilon}_i)(X_i'\hat{\epsilon}_i)']$ 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.

Assumption 4. *Denote s_j the sparsity parameter in (18) by regressing the j^{th} column of X on the remaining columns of X. For any $j \in \{1, \cdots, p\}$, we assume that $n^{-1/2}s_j \log p = o(1)$.*

Theorem 2.2. *Let $\hat{b}$ and $\hat{\Theta}$ be defined as in (17) and (21), respectively. Let Assumptions 1, 2, 3 and 4 be satisfied and further assume $X_i'\epsilon_i$ for $i \in \{1, \cdots, n\}$ is stationary. Then the following hold:*

$$\sqrt{n}(\hat{b} - \beta^0) = \hat{\Theta}X'\epsilon/\sqrt{n} + o_p(1), \tag{22}$$

$$\hat{\Theta}_l'X'\epsilon/\sqrt{n} \to \mathbb{N}(0, \Theta_l'\Sigma_{X\epsilon}\Theta_l). \tag{23}$$

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

$$[\hat{b}_l - C(\alpha, \hat{\Theta}_l, \hat{\Sigma}_{X\epsilon}), \quad \hat{b}_l + C(\alpha, \hat{\Theta}_l, \hat{\Sigma}_{X\epsilon})], \tag{24}$$

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.

The proof follows closely to Kock (2016) and Van De Geer et al. (2014) but with specifically identified bias term for the OWL estimator. Theorem 2.2 arrives at the asymptotic normality property for the de-biased OWL estimator $\hat{b}$ and it enables uniformly valid test for β_l^0 for all $l = 1, \cdots, p$. We use (24) to test for significant factors from the factor zoo.

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.

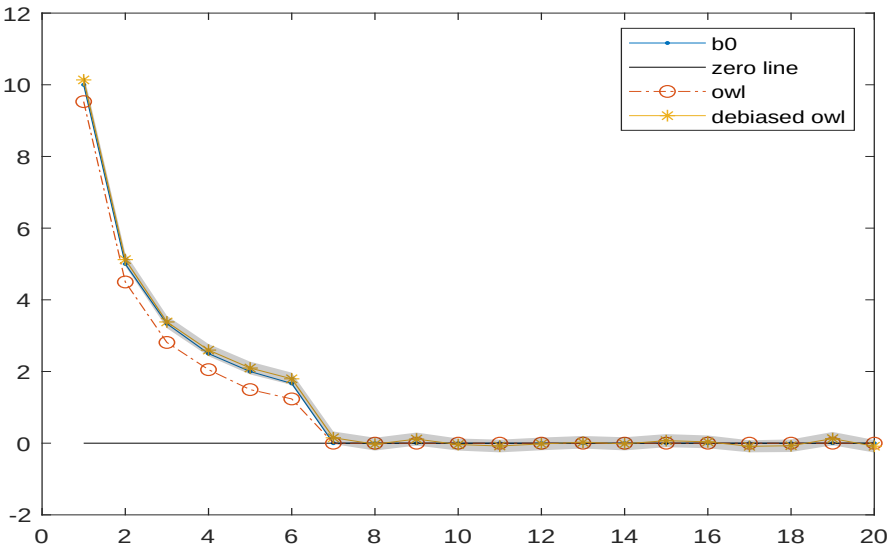

Figure 1. A toy example

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.

Figure 1 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 (i.e. factors with true zero coefficients) now have non-zero coefficients using the de-biased OWL estimator. However, this incorrect de-biasing for those useless factors is bounded by the confidence intervals. We find that the true values (zeros) of the coefficients of these useless factors lie within the shaded area. Hence, we can remove those useless factors by applying the confidence interval. This simple toy example illustrates the nice properties 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

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

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

2.3.

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

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.[7] We compare the de-biased OWL estimator with other benchmarks, including the OLS (when it is feasible) and the LASSO estimators. The reason for choosing these two benchmarks is that the OLS estimator yields unbiased estimation for the useful factors (i.e., factors with true non-zero coefficients) while the LASSO estimator (under proper choice of penalty level) would yield zero coefficients for useless factors. We will check how well the debaised OWL estimator performs in both scenarios (i.e., useful and useless factors) compared to an ideal estimator. 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 (24). 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.

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

[7] 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.

$j \in \{3, 6, 12, 20\}$ 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 ma-

chine learning literature; another one is by specifying the maximum number of non-zero coefficients we want to obtain.[8] 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 OLS estimator is feasible), the de-biased OWL estimates for β_3 and β_6 are very close to

[8] 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.

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

In this section, we apply the debiased OWL estimator to select factors from the factor zoo for the cross-sectional asset returns. Harvey et al. (2015) and Hou et al. (2020) show that the increasing number of factors that have been proposed in the finance literature that claims to explain the cross-sectional asset returns often fails the intense scrutiny of data-snooping and p-hacking. Cochrane (2011) dubs this phenomenon the "factor zoo". Feng et al. (2020) suggest to use the double LASSO selection method devised by Belloni et al. (2014) to select factors recursively and chronologically.[9] In a similar vein, we utilize the nodewise LASSO method in Van De Geer et al. (2014) to obtain a debiased OWL estimator. It is worth stressing that our de-biasing method is different from Chernozhukov et al. (2018) and therefore, our empirical analysis is more general than that in Feng et al. (2020). The nodewise LASSO technique enables us to make inference for any factor in the factor zoo, without splitting factors into (high-dimensional) controlling variables and a small set of factors that to be tested. The empirical application also deviates from Hiraki and Sun (2022), in which the author focuses on correlation-robust selection of factors while the OWL estimator is biased. Hiraki and Sun (2022) vigorously compares the selected

[9]The LASSO estimator is biased and therefore unfit to conduct statistical testing to infer significance of factors. Belloni et al. (2014) and Chernozhukov et al. (2018) proposed the double-LASSO shrinkage method, also known as the double machine learning method, to enable a bias-corrected estimator and therefore, capable of conducting statistical testing and inferring significance of factors.

factors via OWL shrinkage method with other popular benchmarks and demonstrates that OWL estimator selects factors outperforming other benchmarks in an out-of-sample setting.

4.1 Data

We construct 80 firm-characteristics[10] based factors using portfolio sorting. We consider all stocks traded on the New York stock exchange (NYSE), NASDAQ and AMEX between January 1980 and March 2022. For each firm-characteristic, we sort stocks into decile portfolios in each month, then we compute the spread return of the top decile portfolio and the bottom decile portfolio, which is approximated as this firm-characteristic based factor return.[11] We use the pooled decile portfolios for all firm characteristics as the grand set of test portfolios.[12] Then, we use stochastic discount factor (SDF) method to infer priced factors.

It is worth stressing that the selection of factors for the cross-sectional stock returns should be looked together with the test assets - the choice of test asset can directly affect the selection of factors. For instance, if we choose individual stocks as test assets, the factors selected will be most influential to small stocks, since small stocks constitute a substantially large proportion of test assets; on the other hand, if we choose value weighted sorted portfolios (as we do in this paper) as test portfolios, selected factors would reflect the most influential factors for the aggregated market fluctuations.

4.2 Inference with risk prices

We follow Cochrane (2005) to introduce the Stochastic Discount Factor (SDF) framework to infer priced factors. It is worth noting that the risk price and the risk premium of a factor are closely related, yet they differs in their implications especially when factors are correlated - see Hiraki and Sun (2022) for a detailed discussion. Let m denote the SDF

[10] See Appendix D for a detailed description of those firm characteristics.

[11] All factors are scaled to have zero mean and the same variance as the 'mkt' factor.

[12] For robustness check, we also consider the bi-varite sorted five-by-five test portfolios following Feng et al. (2020). Specifically, we use the 'size' characteristic to form bi-variate sorted portfolios with other firm characteristics before pooling them together as the grand set of test portfolios. For robust results, we remove firm characteristics that exhibit more than 40% missing data. We follow the convention in the finance literature to remove micro stocks which are smaller than 20 percentile of the NYSE listed stocks.

and we assume that the SDF is linear to factors.

$$m = 1 - b'(f - \mathrm{E}(f)), \tag{25}$$

where f is a $K \times 1$ vector of K factor returns. b is a $K \times 1$ vector of the SDF coefficient, referred to as the *risk price*; a non-zero (zero) entry of b means the corresponding factor is (not) priced. Denote by R the excess returns of a vector of N test assets. Define $Y = (f', R')'$, so $\mathrm{Var}(Y) = \begin{pmatrix} \mathrm{Var}(f) & \mathrm{Cov}(R, f)' \\ \mathrm{Cov}(R, f) & \mathrm{Var}(R) \end{pmatrix}$, where $\mathrm{Var}(f)$ and $\mathrm{Var}(R)$ are the $K \times K$ and $N \times N$ variance-covariance matrices of factors f and test asset returns R, respectively. $\mathrm{Cov}(R, f)$ is the $N \times K$ covariance matrix of returns and factors. The fundamental asset pricing equation states that $\mathrm{E}(Rm) = \mathbf{0}$ for any admissible SDF. The deviation from zero of the fundamental asset pricing equation is regarded as the pricing error. Let $m(b)$ denote the (unknown) SDF which depends on the (unknown) risk price b. Pricing error $e(b)$ can be written and simplified as

$$\begin{aligned} e(b) &= \mathrm{E}[Rm(b)] = \mathrm{E}(R)\mathrm{E}(m(b)) + \mathrm{Cov}(R, m(b)) \\ &= \mathrm{E}(R)\mathrm{E}(1 - b'(f - E(f))) + \mathrm{Cov}(R, 1 - b'(f - E(f))) \\ &= \mathrm{E}(R) - \mathrm{Cov}(R, f)b \\ &= \mu_R - Cb, \end{aligned} \tag{26}$$

where $\mu_R := \mathrm{E}(R)$ is the $N \times 1$ vector of the expectation of excess returns of test assets and $C := \mathrm{Cov}(R, f)$. Define a quadratic form of the pricing error such that $Q(b) = e(b)'\, W\, e(b)$, where W is a $N \times N$ weighting matrix. Following the suggestion in Ludvigson (2013), when the test assets are abundant we can set $W = I$, where I is an identity matrix. Then, we can estimate b by minimizing $Q(b)$.

$$\hat{b} = \arg\min_b Q(b) = \arg\min_b (\mu_R - Cb)'(\mu_R - Cb), \tag{27}$$

which gives $\hat{b} = (C'C)^{-1}C'\mu_R$. Since both C and μ_R are unobservable, we adopt the sample analogs such that $\hat{C} = \widehat{\mathrm{Cov}(R, f)} = \frac{1}{T}\sum_{t=1}^{T}(R_t - \hat{\mu}_R)(f_t - \hat{\mu}_f)'$, $\hat{\mu}_f = \frac{1}{T}\sum_{t=1}^{T} f_t$ and $\hat{\mu}_R =$

$\frac{1}{T}\sum_{t=1}^{T} R_t$ in our empirical applications.[13] The finance literature provides ample evidence and suggests that the true factors that drive cross-sectional asset prices are sparse, see Harvey et al. (2015) for example. More specifically, only a small number of those factors are priced, commanding a small number of elements in b are non-zeros. With this sparsity condition, we consider the OWL shrinkage method given high correlations between factors in the factor zoo. Then, (27) becomes

$$\check{b} = \arg\min_{b} \ (\hat{\mu}_R - \hat{C}b)'(\hat{\mu}_R - \hat{C}b) + \Omega_\omega(b), \qquad \Omega_\omega(b) = \omega'|b|_\downarrow, \tag{28}$$

where $\Omega_\omega(b)$ is defined similarly as in (2) and $\check{b}$ denotes the penalised estimator for risk prices. Note that $\check{b}$ is a biased estimator. Then, after utilizing the de-biased OWL estimator specified in (17), we obtain

$$\check{b}_{debiased} = \check{b} + \check{\Theta}\hat{C}'(\hat{\mu}_R - \hat{C}\check{b})/N, \tag{29}$$

where $\check{\Theta}$ is an approximation of the inverse of $\hat{C}'\hat{C}/N$ using the nodewise LASSO regression method. We make inference on the significance of risk prices based on the 95% confidence interval in (24).

4.3 Estimation results

Figure 2 shows the correlation coefficient matrix for all columns in $\hat{C}$. We find that many factors are close cousins and therefore are highly correlated - a substantial number of pairs exhibit correlation coefficients greater than 0.9 (absolute value). Traditionally, an ad-hoc screening of factors are required before modelling and drawing inferences. For example, Green et al. (2017) delete beta-related factors before implementing a Fama-MacBeth regression model to infer risk premiums of factors, as they are highly correlated with many other factors. On the other hand, since the OWL estimator is robust with factor correlations, we can include all factors for consideration before estimation.

[13] In our empirical applications, we have $T = 519$ and $K = 81$, that consists of 80 anomaly factors and the 'mkt' factor.

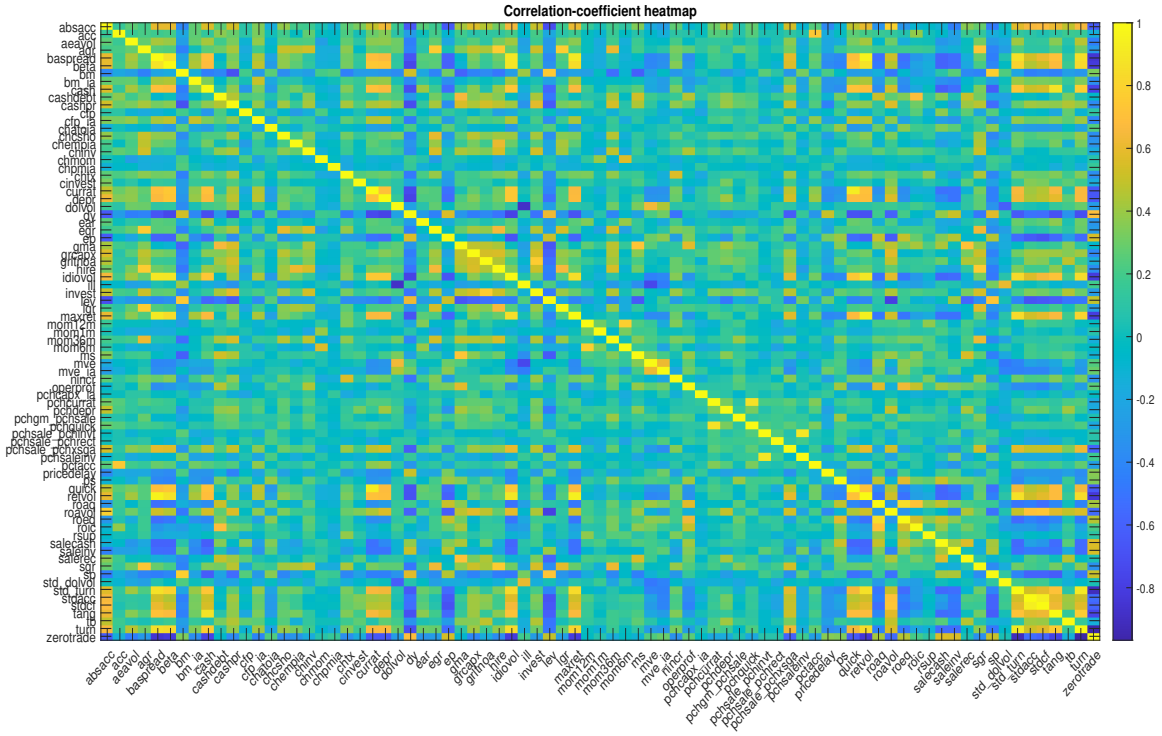

Figure 2. Factor correlation

Figure 3 shows the estimation result using uni-variate sorted decile portfolios for all firm characteristics as the test assets. The null hypothesis is that for any l, $l \in \{1, \cdots, K\}$, $\check{b}_{l,debiased} = 0$. Therefore, for each l, if zero is contained within the confidence interval then we do not reject the null hypothesis. Note that we also add the 'mkt' factor into the factor zoo, because the 'mkt' factor not only has been an indispensable factor in many celebrated asset pricing models such as in Fama and French (2016) and Hou et al. (2014), it also offers a key element to test whether a candidate estimator can select the 'mkt' factor from the factor zoo and draw inference on its significance. We find that the 'mkt' factor is the strongest factors in the debiased OWL estimation while we use uni-variate sorted portfolios as test assets. This finding is consistent with mainstream asset pricing models. Then, 'liquidity', 'momentum' and 'profitability' related factors are other strong factors to drive asset prices. We also offer a robustness check in Appendix E while using alternative test assets and various criteria to screen out micro stocks.[14]

[14]It is worth mention that factor selection also depends on the time-window we select. Different periods command time-varying dominant factors for the cross-sectional asset returns. This can be demonstrated

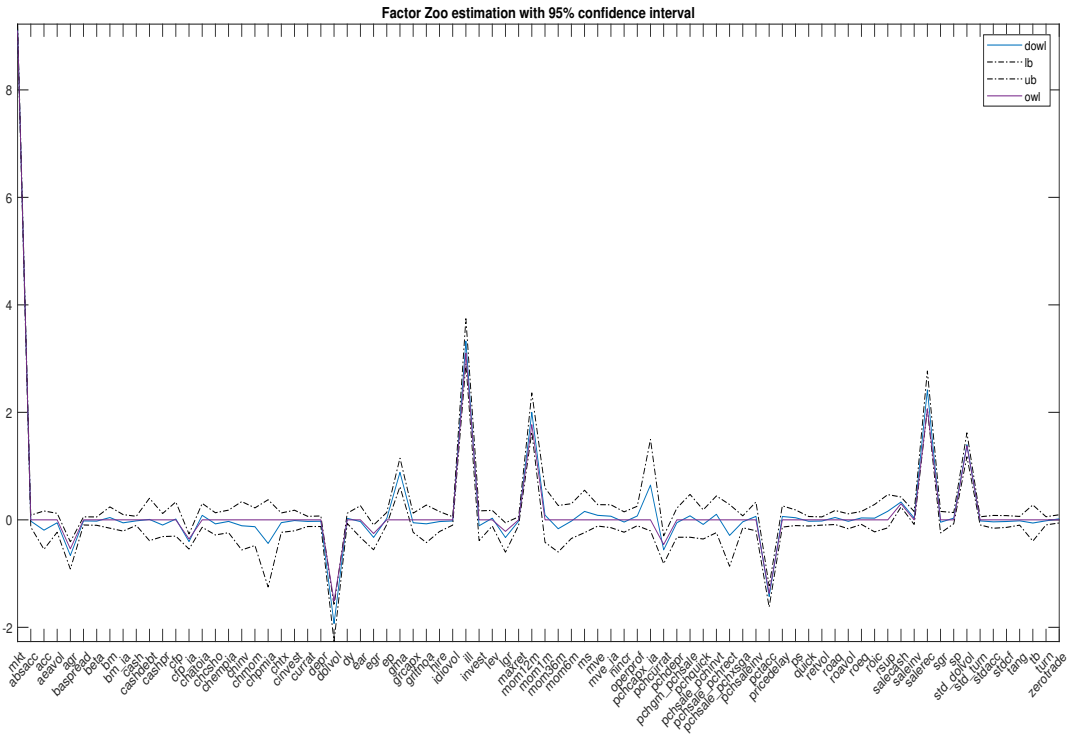

Figure 3. Estimation of factor zoo coefficients using uni-variate sorted decile portfolios as test assets. 'lb' and 'ub' are the lower and upper bounds for the 95% confidence interval. 'owl' is the biased OWL estimate while 'dowl' is the debiased OWL estimate.

Our empirical analysis is closely related to Feng et al. (2020), who also employ the SDF method to select factors from the factor zoo using sorted portfolios as test assets. Feng et al. (2020) utilize the double LASSO selection procedure (a.k.a double machine learning method) put forward by Belloni et al. (2014) and Chernozhukov et al. (2018). They test for significant factors recursively based on the calendar year when factors are proposed, while using all factors proposed before that year as controlling factors. They repeat such test on each calendar year. Our debiased OWL estimator differentiates itself in two distinctive ways: first, the debiased OWL estimator is focused on correlation-robust estimation - it is an important question that needs to be addressed in high dimensional setting, as high correlation between explanatory variables often dampens standard estimators, including the LASSO shrinkage method. Second, our approach is more general without the necessity to separate factors into controlling factors (on which the double machine learning method does not draw inference) and (a small number of) factors to be tested. We can make inference on any factors in the high dimensional factor zoo. Therefore, our method and empirical analysis on the factor zoo complements the existing literature and shed new light on the correlation related problems in the high dimensional setting.

5 Concluding remarks

In this paper, we study the statistical properties of a correlation-robust shrinkage estimator. Specifically, we develop the error bound (oracle inequality) for the ordered-weighted-LASSO (OWL) estimator under less restrictive assumptions. We derive the probability of the oracle inequality holds, and numerically dissect the elements that affect this probability. We also show that the OWL estimator is consistent under some regularity conditions. We then further give guidance on the choice of penalty hyper-parameters using a self-normalization tool. In addition, we devise a bias-corrected OWL estimator and show that it is asymptotically normally distributed. We also derive a point-wise confidence interval for the estimate given a significance level. Monte Carlo simulation shows that the de-biased OWL estimator reduces estimation error substantially and achieves satisfying coverage

by sub-sample estimations, see Hiraki and Sun (2022) for example.

rate. Empirically, we employ the de-biased OWL estimator to choose factors which have significant risk prices under the SDF framework. We find that liquidity and profitability related factors, along with the market factor, are among the strong factors that drive the cross-sectional asset prices.

References

Babii, A., E. Ghysels, and J. Striaukas (2021): "Machine Learning Time Series Regressions With an Application to Nowcasting," *Journal of Business and Economic Statistics*, 1–23.

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.

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.

Callot, L., M. Caner, A. Önder, and E. Ulaşan (2021): "A Nodewise Regression Approach to Estimating Large Portfolios," *Journal of Business and Economic Statistics*.

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.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018): "Double/debiased machine learning for treatment and structural parameters," *Econometrics Journal*, 21, C1–C68.

Chinco, A., A. D. Clark-Joseph, and M. Ye (2019): "Sparse Signals in the Cross-Section of Returns," *Journal of Finance*, 74, 449–492.

Cochrane, J. H. (2005): *Asset Pricing*, Princeton University Press.

——— (2011): "Presidential Address: Discount Rates," *Journal of Finance*, 66.

Dendramis, Y., L. Giraitis, and G. Kapetanios (2021): "Estimation of Time-Varying Covariance Matrices for Large Datasets," *Econometric Theory*, 0, 1–35.

Fama, E. F. and K. R. French (2016): "Dissecting Anomalies with a Five-Factor Model," *Review of Financial Studies*, 29.

Fan, J., Y. Ke, and K. Wang (2020): "Factor-adjusted regularized model selection," *Journal of Econometrics*, 216.

Feng, G., S. Giglio, and D. Xiu (2020): "Taming the Factor Zoo: A Test of New Factors," *Journal of Finance*, 75, 1327–1370.

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

Freyberger, J., A. Neuhierl, and M. Weber (2020): "Dissecting Characteristics Nonparametrically," *The Review of Financial Studies*, 33, 2326–2377.

Green, J., J. R. M. Hand, and X. F. Zhang (2017): "The Characteristics that Provide Independent Information about Average U.S. Monthly Stock Returns," *The Review of Financial Studies*, 30, 4389–4436.

Harvey, C. R., Y. Liu, and H. Zhu (2015): "... and the Cross-Section of Expected Returns," *The Review of Financial Studies*, 29, 5–68.

Hiraki, K. and C. Sun (2022): "A toolkit for exploiting contemporaneous stock correlations," *Journal of Empirical Finance*, 65.

Hou, K., C. Xue, and L. Zhang (2014): "Digesting anomalies: An investment approach," *Review of Financial Studies*, 28.

——— (2020): "Replicating Anomalies," *The Review of Financial Studies*, 33, 2019–2133.

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 Fixted Effects," *Econometric Theory*, 35, 295–359.

Kozak, S. and S. Santosh (2020): "Why do discount rates vary?" *Journal of Financial Economics*, 137, 740–751.

Ludvigson, S. C. (2013): "Advances in Consumption-Based Asset Pricing: Empirical Tests," in *Handbook of the Economics of Finance*, vol. 2.

Medeiros, M. C. and E. F. Mendes (2016): "L1 -regularization of high-dimensional time-series models with non-Gaussian and heteroskedastic errors," *Journal of Econometrics*, 191, 255–271.

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

Van De Geer, S., P. Buhlmann, Y. Ritov, and R. Dezeure (2014): "ON ASYMPTOTICALLY OPTIMAL CONFIDENCE REGIONS AND TESTS FOR HIGH-DIMENSIONAL MODELS," *The Annals of Statistics*, 42, 1166–1202.

Van de Geer, S. A. and P. Bühlmann (2009): "On the conditions used to prove oracle results for the lasso," *Electronic Journal of Statistics*, 3.

Yuan, M. and Y. Lin (2006): "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society. Series B*, 68, 49–67.

Zeng, X. and M. A. T. Figueiredo (2014): "The Ordered Weighted L1 Norm: Atomic Formulation, Projections, and Algorithms," *arXiv: Data Structures and Algorithms*.

Zou, H. and T. Hastie (2005): "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society. Series B*, 67, 301–320.

Online Appendix for "On the Inference of a LASSO-type Estimator with Highly Correlated Variables "

Chuanping Sun*

This online appendix provides technical proofs for Theorems, Proposition and Corollaries. We also offer detailed accounts of the algorithm used for solving the OWL optimization problem, additional comparison between the OWL estimator and other popular shrinkage estimators, and conduct robustness check for the empirical application.

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 (7) and (8) under the event E, which is specified in (A.5). The second part we will derive the probability of this event $\mathbb{P}(E)$ to be true under mixing condition and fatter tails. For the derivation of the second part, we take advantage of the following auxiliary lemma and the proof of this lemma can be found in Dendramis et al. (2021).

Lemma 1 (Dendramis et al. (2021))**.** *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.

Part I.

By definition the OWL estimator is minimizing the function

$$\hat{\beta} = \hat{\beta}_{OWL} = \arg\min_{\beta} \quad \frac{1}{n}||y - X\beta||_2^2 + \frac{1}{n}\sum_{i=1}^{p}[\lambda_1 + \lambda_2(p-i)]|\beta|_{[i]},$$

where $|\beta|_{[\cdot]}$ denotes the element of the decreasingly ordered vector of $|\beta|$, such that $|\beta|_{[1]} \geq |\beta|_{[2]} \geq ... \geq |\beta|_{[p]}$. Let β^0 denote the true values of β and $y = X\beta^0 + \epsilon$. According to the

*Faculty of Finance, Bayes Business School (formerly Cass), City University of London. 106 Bunhill Row, London EC1Y 8TZ, United Kingdom, chuanping.sun@city.ac.uk

"argmin" property, definition of $\hat{\beta}$ implies

$$\frac{1}{n}||y-X\hat{\beta}||_2^2+\frac{1}{n}\sum_i[\lambda_1+\lambda_2(p-i)]|\hat{\beta}|_{[i]} \leq \frac{1}{n}||y-X\beta^0||_2^2+\frac{1}{n}\sum_i[\lambda_1+\lambda_2(p-i)]|\beta^0|_{[i]}. \quad \text{(A.1)}$$

Since $\omega_i = \lambda_1 + \lambda_2(p-i)$ is in a monotone non-negative cone and $\omega_1 \geq \omega_2 \geq ... \geq \omega_p$, we have

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

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

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

Note that

$$2|\epsilon' X(\hat{\beta}-\beta^0)| \leq \left(\max_{1\leq j\leq p} 2|\epsilon' X^{(j)}|\right)||\hat{\beta}-\beta^0||_1. \quad \text{(A.3)}$$

Hence, (A.2) can be written as

$$\frac{1}{n}||X(\hat{\beta}-\beta^0)||_2^2+\frac{\lambda_1}{n}||\hat{\beta}||_1 \leq \left(\frac{1}{n}\max_{1\leq j\leq p} 2|\epsilon' X^{(j)}|\right)||\hat{\beta}-\beta^0||_1+\frac{1}{n}[\lambda_1+\lambda_2(p-1)]||\beta^0||_1. \quad \text{(A.4)}$$

Consider the event

$$E := \left\{\frac{1}{n}\max_{1\leq j\leq p} 2|\epsilon' X^{(j)}| \leq \lambda_0\right\}, \quad \text{(A.5)}$$

where $\lambda_0 = \kappa\sqrt{\frac{\log p}{n}}$ and κ is a positive constant. Then, in view of (A.5), inequation (A.4) can be bounded as

$$\frac{1}{n}||X(\hat{\beta}-\beta^0)||_2^2 + \frac{1}{n}\lambda_1||\hat{\beta}||_1 \leq \lambda_0||\hat{\beta}-\beta^0||_1 + \frac{1}{n}[\lambda_1 + \lambda_2(p-1)]||\beta^0||_1. \quad \text{(A.6)}$$

By assumption, $\frac{\lambda_1}{n} = 2\lambda_0$. Therefore, (A.6) can be written as

$$\frac{2}{n}||X(\hat{\beta}-\beta^0)||_2^2 + \frac{2}{n}\lambda_1||\hat{\beta}||_1 \leq \frac{\lambda_1}{n}||\hat{\beta}-\beta^0||_1 + \frac{2}{n}[\lambda_1 + \lambda_2(p-1)]||\beta^0||_1. \quad \text{(A.7)}$$

Note that

$$||\hat{\beta}||_1 = ||\hat{\beta}_{s_0}||_1 + ||\hat{\beta}_{s_0^c}||_1 \geq ||\beta^0_{s_0}||_1 - ||\hat{\beta}_{s_0} - \beta^0_{s_0}||_1 + ||\hat{\beta}_{s_0^c}||_1, \quad \text{(A.8)}$$
$$||\hat{\beta}-\beta^0||_1 = ||\hat{\beta}_{s_0} - \beta^0_{s_0}||_1 + ||\hat{\beta}_{s_0^c}||_1. \quad \text{(A.9)}$$

Therefore, using (A.8) and (A.9), inequation (A.7) can be written as

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

Note that $\|\beta^0_{s_0}\|_1=\|\beta^0\|_1$, so (A.10) can be written as

$$\frac{2}{n}||X(\hat{\beta}-\beta^0)||_2^2+\frac{\lambda_1}{n}||\hat{\beta}_{s_0^c}||_1\leq 3\frac{\lambda_1}{n}||\hat{\beta}_{s_0}-\beta^0_{s_0}||_1+\frac{2\lambda_2(p-1)}{n}||\beta^0||_1. \quad \text{(A.11)}$$

By (A.9), $||\hat{\beta}_{s_0^c}||_1=||\hat{\beta}-\beta^0||_1-||\hat{\beta}_{s_0}-\beta^0_{s_0}||_1$. Utilizing this in (A.11), we obtain

$$\frac{2}{n}||X(\hat{\beta}-\beta^0)||_2^2+\frac{\lambda_1}{n}||\hat{\beta}-\beta^0||_1\leq 4\frac{\lambda_1}{n}||\hat{\beta}_{s_0}-\beta^0_{s_0}||_1+\frac{2\lambda_2(p-1)}{n}||\beta^0||_1. \quad \text{(A.12)}$$

By (6) in Assumption 2, it is easy to show that for any β,

$$||\beta_{s_0}||_1^2\leq \beta'\hat{\Sigma}\beta s/\phi_0^2. \quad \text{(A.13)}$$

Applying (A.13) on $||\hat{\beta}_{s_0}-\beta^0_{s_0}||_1$ and using $\hat{\Sigma}=\frac{X'X}{n}$, we have

$$||\hat{\beta}_{s_0}-\beta^0_{s_0}||_1^2\leq(\hat{\beta}-\beta^0)'\hat{\Sigma}(\hat{\beta}-\beta^0)s/\phi_0^2=||X(\hat{\beta}-\beta^0)||_2^2 s/(n\phi_0^2),$$
$$||\hat{\beta}_{s_0}-\beta^0_{s_0}||_1\leq||X(\hat{\beta}-\beta^0)||_2\sqrt{s}/(\sqrt{n}\phi_0).$$

Therefore, using inequality $4ab\leq a^2+4\beta^2$, we obtain

$$4\frac{\lambda_1}{n}||\hat{\beta}_{s_0}-\beta^0_{s_0}||_1\leq 4\left(\frac{||X(\hat{\beta}-\beta^0)||_2}{\sqrt{n}}\right)\left(\frac{\lambda_1}{n}\frac{\sqrt{s}}{\phi_0}\right)$$
$$\leq\frac{1}{n}||X(\hat{\beta}-\beta^0)||_2^2+4(\frac{\lambda_1}{n})^2\frac{s}{\phi_0^2}.$$

So (A.12) can be written as

$$\frac{1}{n}||X(\hat{\beta}-\beta^0)||_2^2+\frac{\lambda_1}{n}||\hat{\beta}-\beta^0||_1\leq 4(\frac{\lambda_1}{n})^2\frac{s}{\phi_0^2}+\frac{2\lambda_2(p-1)}{n}||\beta^0||_1. \quad \text{(A.14)}$$

Then, by assumption, $\frac{\lambda_1}{n}=2\lambda_0\asymp\sqrt{\frac{\log p}{n}}$, and $\frac{\lambda_2}{n}\lesssim\frac{s\log p}{np}\asymp\frac{s\lambda_0^2}{p}$. Therefore, (A.14) can be written as

$$\frac{1}{n}||X(\hat{\beta}-\beta^0)||_2^2+2\lambda_0||\hat{\beta}-\beta^0||_1\lesssim 16\lambda_0^2 s/\phi_0^2+2\lambda_0^2 s\|\beta^0\|_1. \quad \text{(A.15)}$$

Using $\sqrt{a^2+b^2} \le a+b$, for all $a, b > 0$, (A.15) implies

$$\frac{1}{n}||X(\hat{\beta}-\beta^0)||_2 \lesssim 4\lambda_0\sqrt{s}/\phi_0 + \lambda_0\sqrt{2s\|\beta^0\|_1}, \tag{A.16}$$

$$||\hat{\beta}-\beta^0||_1 \lesssim 8\lambda_0 s/\phi_0^2 + \lambda_0 s\|\beta^0\|_1. \tag{A.17}$$

This shows that (7) and (8) in Theorem 2.1 are valid, assuming that (A.5) holds.

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

$$\begin{aligned}\mathbb{P}(E^c) &= \mathbb{P}(\frac{1}{n}\|X'\epsilon\|_\infty > \frac{\lambda_0}{2}) = \mathbb{P}(\frac{1}{n}\max_{j=1,\cdots,p}|\sum_{i=1}^n X_{i,j}\epsilon_i| > \frac{\lambda_0}{2}) \\ &\le \sum_{j=1}^p \mathbb{P}(\frac{1}{\sqrt{n}}\sum_{i=1}^n |X_{i,j}\epsilon_i| > \frac{\lambda_0\sqrt{n}}{2}) = p\max_{j=1,\cdots,p}\mathbb{P}(\frac{1}{\sqrt{n}}\sum_{i=1}^n |X_{i,j}\epsilon_i| > \frac{\lambda_0\sqrt{n}}{2}).\end{aligned} \tag{A.18}$$

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. (2021), we have

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

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

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

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

$$\begin{aligned}p\mathbb{P}(\frac{1}{\sqrt{n}}|z_{i,j}| > \frac{\lambda_0\sqrt{n}}{2}) &\le pc_0\exp(-c_1'(\frac{\kappa}{2})^2\log p) + pc_0\exp(-c_2'(\frac{\kappa\sqrt{n\log p}}{2\log^2 n})^\zeta) \\ &:= r_p + r'_{p,n}.\end{aligned} \tag{A.19}$$

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

$$r_p \le pc_0\exp[-(1+\epsilon)\log p] = c_0 p^{-\epsilon} \to 0, \quad \text{as } p \to \infty. \tag{A.20}$$

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

$$c_2'(\frac{\kappa\sqrt{n\log p}}{2\log^2 n})^\zeta \ge c_2'(\kappa p^{1/(4\delta)}\sqrt{\log p})^\zeta > (1+\epsilon)\log p, \quad \text{as } p \to \infty. \tag{A.21}$$

Therefore, equation (A.20) and (A.21) imply that

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

Then by (A.18) and (A.19), we obtain

$$\begin{aligned} \mathbb{P}(E^c) &= r_p + r'_{p,n} \leq 2r_p \leq 2c_0 p^{-\epsilon}, \\ \mathbb{P}(E) &= 1 - \mathbb{P}(E^c) \geq 1 - c'_0 p^{-\epsilon} \to 1, \quad \text{as } n, p \to \infty, \end{aligned} \tag{A.22}$$

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.19) 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,

$$\mathbb{P}(E) = 1 - \mathbb{P}(E^c) = 1 - pc_0 \left[\exp(-\frac{c'_1}{4}\kappa^2 \log p) + \exp\left(-c'_2 \left(\frac{\kappa\sqrt{n \log p}}{2 \log^2 n}\right)^{\zeta}\right)\right], \tag{A.23}$$

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 (8) in Theorem 2.1 and Assumption 3(a), it follows naturally that

$$\|\hat{\beta} - \beta^0\|_1 = O_p(s\sqrt{\frac{\log p}{n}}) = o_p(1), \tag{A.24}$$

which proves the second claim of (10). Utilizing $\hat{\Sigma} = X'X/n$, we obtain

$$\begin{aligned} \|X(\hat{\beta} - \beta^0)\|_2^2/n &= (\hat{\beta} - \beta^0)'\hat{\Sigma}(\hat{\beta} - \beta^0) \\ &= (\hat{\beta} - \beta^0)'(\hat{\Sigma} - \Sigma)(\hat{\beta} - \beta^0) + (\hat{\beta} - \beta^0)'\Sigma(\hat{\beta} - \beta^0). \end{aligned} \tag{A.25}$$

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.25) 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 $\|\hat{\Sigma} - \Sigma\|_\infty := \max_{1\le i,j\le p} |\hat{\Sigma}_{i,j} - \Sigma_{i,j}|$. Using lemma 14.12 in Buhlmann and Van De Geer (2011), we have $\max_{1\le i,j\le p} |\hat{\Sigma}_{i,j} - \Sigma_{i,j}| = O_p(\sqrt{\log p/n})$. Together with $\|\hat{\beta} - \beta^0\|_1 = O_p(s\sqrt{\log p/n})$ obtained in (A.24), this implies that (A.25) can be bounded as

$$\begin{aligned}\frac{1}{n}\|X(\hat{\beta} - \beta^0)\|_2^2 &= (\hat{\beta} - \beta^0)'\Sigma(\hat{\beta} - \beta^0) + (\hat{\beta} - \beta^0)'(\hat{\Sigma} - \Sigma)(\hat{\beta} - \beta^0) \\ &\ge \Lambda_{min}^2\|\hat{\beta} - \beta^0\|_2^2 - \|\hat{\Sigma} - \Sigma^0\|_\infty\|\hat{\beta} - \beta^0\|_1^2 \\ &\ge \Lambda_{min}^2\|\hat{\beta} - \beta^0\|_2^2 - O_p\left(s^2\left(\frac{\log p}{n}\right)^{3/2}\right).\end{aligned} \tag{A.26}$$

Note that $\lambda_0 \asymp \sqrt{\log p/n}$. So by (7) we obtain

$$\frac{1}{n}\|X(\hat{\beta} - \beta^0)\|_2^2 = O_p\left(\frac{s\log p}{n}\right). \tag{A.27}$$

Plugging (A.27) into (A.26) and rearranging (A.26), we obtain

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

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 $\Lambda_{min} \ge a > 0$ where a is a constant, hence $\frac{1}{\Lambda_{min}^2} = O(1)$. Then by Assumption 3, we obtain

$$\|\hat{\beta} - \beta^0\|_2^2 = o_p(1), \tag{A.28}$$

which proves the first claim of (10). Also, by Theorem 2.1 part one, (A.24) and (A.28) 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 (17) and by extracting $\sqrt{n}$ from (15), it is easy to show that

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

where Δ is defined in (16). Then to prove (22), it suffices to show that

$$\Delta = o_p(1). \tag{A.29}$$

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

$$\hat{\Sigma}_{X\epsilon} = \frac{1}{n}\sum_{i=1}^{n}\left[(X_i'\hat{\epsilon}_i)(X_i'\hat{\epsilon}_i)'\right]. \tag{A.30}$$

To show (23) and (A.29), 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

$$\Theta_l' \Sigma_{X\epsilon} \Theta_l = \Theta_l' P' \mathrm{eig}(\Sigma_{X\epsilon}) P \Theta_l \geq \Lambda_{min}(\Sigma_{X\epsilon}) \|\Theta_l\|_2^2 > 0, \tag{A.31}$$

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

$$|\hat{\Theta}_l' \hat{\Sigma}_{X\epsilon} \hat{\Theta}_l - \Theta_l' \Sigma_{X\epsilon} \Theta_l| = o_p(1). \tag{A.32}$$

Set

$$\tilde{\Sigma}_{X\epsilon} = \frac{1}{n} \sum_{i=1}^n \left[(X_i' \epsilon_i)(X_i' \epsilon_i)'\right]. \tag{A.33}$$

Then

$$
\begin{aligned}
|\hat{\Theta}_l'\hat{\Sigma}_{X\epsilon}\hat{\Theta}_l - \Theta_l'\Sigma_{X\epsilon}\Theta_l| &\le |\hat{\Theta}_l'\hat{\Sigma}_{X\epsilon}\hat{\Theta}_l - \hat{\Theta}_l'\Sigma_{X\epsilon}\hat{\Theta}_l| + |\hat{\Theta}_l'\Sigma_{X\epsilon}\hat{\Theta}_l - \Theta_l'\Sigma_{X\epsilon}\Theta_l| \\
&\le |\hat{\Theta}_l'\hat{\Sigma}_{X\epsilon}\hat{\Theta}_l - \hat{\Theta}_l'\tilde{\Sigma}_{X\epsilon}\hat{\Theta}_l| + |\hat{\Theta}_l'\tilde{\Sigma}_{X\epsilon}\hat{\Theta}_l - \hat{\Theta}_l'\Sigma_{X\epsilon}\hat{\Theta}_l| + |\hat{\Theta}_l'\Sigma_{X\epsilon}\hat{\Theta}_l - \Theta_l'\Sigma_{X\epsilon}\Theta_l| \\
&= (I) + (II) + (III).
\end{aligned} \tag{A.34}
$$

For (I), we have

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

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

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

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

$$
\begin{aligned}
\|\frac{1}{n}\sum_{i=1}^n X_i'X_i(\hat{\beta} - \beta^0)[X_i'X_i(\hat{\beta} - \beta^0)]'\|_\infty =& \|\frac{1}{n}\sum_{i=1}^n X_i'X_i(\hat{\beta} - \beta^0)(\hat{\beta} - \beta^0)'X_i'X_i\|_\infty \\
\le& \frac{1}{n}\sum_{i=1}^n \|X_i'X_iX_i'X_i\|_\infty \|\hat{\beta} - \beta^0\|_1^2 \\
\le& \max_j \frac{1}{n}\sum_{i=1}^n X_{i,j}^4 \ \|\hat{\beta} - \beta^0\|_1^2,
\end{aligned} \tag{A.35}
$$

where $j = 1, \cdots, p$. By Assumption 1, $\mathbb{P}(|X_{i,j}| > a) \le c_1\exp(-c_2a^{q_1})$. Set $Y_{i,j} = X_{i,j}^4$, then $\mathbb{P}(|Y_{i,j}| > a) = \mathbb{P}(|X_{i,j}| > a^{1/4}) \le c_1\exp(-c_2a^{q_1/4})$. So $X_{i,j}^4$ also has exponential tail bound (with a different parameter). Then by (A.18) and (A.22), for all $j = 1, \cdots, p$, we have $\mathbb{P}(|n^{-1}\sum_{i=1}^n X_{i,j}^4 - E(X_{i,j}^4)| > \lambda_0/2) \le c_0'p^{-\epsilon} \to 0$ 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 $|n^{-1}\sum_{i=1}^n X_i^4| \le |n^{-1}\sum_{i=1}^n X_{i,j}^4 - E(X_{i,j}^4)| +$

$|E(X_{i,j}^4)| = o_p(1) + O_p(1) = O_p(1)$. By (10) 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 (7) 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.18) and (A.22), 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.18) and (A.22), $\|\tilde{\Sigma}_{X\epsilon} - \Sigma_{X\epsilon}\|_\infty = O_p(\sqrt{\log p/n})$ 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.34) 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

$$|\hat{\Theta}_l'X'\epsilon/\sqrt{n} - \Theta_l'X'\epsilon/\sqrt{n}| = o_p(1). \tag{A.36}$$

By Lemma 2, $\|\hat{\Theta}_l - \Theta_l\|_1 = O_p(s_l\sqrt{\log p/n})$ and by (A.18) and (A.22), $\|X'\epsilon/n\|_\infty = O_p(\sqrt{\log p/n})$. Then by Assumption 3, equation (A.36) 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 (23).
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 (20), and by (A.42), 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.31), $|\Theta_l'\Sigma_{X\epsilon}\Theta_l| \geq a > 0$ 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.29). □

A.4 $\hat{\Theta}$ 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 (18) implies

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

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.37) implies

$$\hat{\gamma}'_j X'_{-j}(X_j - X_{-j}\hat{\gamma}_j)/n = \lambda_j \|\hat{\gamma}_j\|_1. \tag{A.38}$$

Therefore, plugging the above equation into (20), we have

$$\begin{aligned} \hat{\delta}_j^2 &= \frac{1}{n}(X_j - X_{-j}\hat{\gamma}_j)'(X_j - X_{-j}\hat{\gamma}_j) + \frac{1}{n}\hat{\gamma}'_j X'_{-j}(X_j - X_{-j}\hat{\gamma}_j) \\ &= \frac{1}{n}[(X_j - X_{-j}\hat{\gamma}_j)' + \hat{\gamma}'_j X'_{-j}](X_j - X_{-j}\hat{\gamma}_j) \\ &= \frac{1}{n} X'_j (X_j - X_{-j}\hat{\gamma}_j). \end{aligned} \tag{A.39}$$

By definition of $\hat{C}_j$ (j^{th} row of matrix $\hat{C}$) in (19), 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 (21), equation (A.39) becomes

$$\hat{\delta}_j^2 = \frac{1}{n} X'_j X \hat{C}_j, \qquad \text{or} \qquad \frac{1}{n} X'_j X \hat{\Theta}_j = 1. \tag{A.40}$$

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 (13), $\|\hat{\tau}\|_\infty \leq 1$. Taking the norm on both sides of (A.37) and using $\hat{\Theta}_j = \hat{C}_j/\hat{\delta}_j^2$, we obtain

$$\begin{aligned} \|X'_{-j}(X_j - X_{-j}\hat{\gamma}_j)\|_\infty/n = \|X'_{-j}X\hat{C}_j\|_\infty = \|\lambda_j \hat{\tau}_j\|_\infty, \\ \|X'_{-j}X\hat{\Theta}_j\|_\infty/n = \lambda_j \|\hat{\tau}_j\|_\infty/\hat{\delta}^2 \leq \lambda_j/\hat{\delta}_j^2. \end{aligned} \tag{A.41}$$

By the definition of X_{-j} and $\hat{\Sigma} := X'X/n$ and by (A.40), we have $\|X'_{-j}X\hat{\Theta}_j\|_\infty = \|\hat{\Sigma}\hat{\Theta}_j - e_j\|_\infty$. Thus (A.41) can be written as

$$\|\hat{\Sigma}\hat{\Theta}_j - e_j\|_\infty \leq \lambda_j/\hat{\delta}_j^2. \tag{A.42}$$

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

$$\Theta_{j,j} = (\Sigma_{j,j} - \Sigma_{j,-j}\Sigma_{-j,-j}^{-1}\Sigma_{-j,j})^{-1}, \qquad \Theta_{j,-j} = -\Theta_{j,j}\Sigma_{j,-j}\Sigma_{-j,-j}^{-1}. \tag{A.43}$$

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,

$$\gamma_j = [\frac{1}{n}\sum_{i=1}^{n} E(X_{i,-j}'X_{i,-j})]^{-1}[\frac{1}{n}\sum_{i=1}^{n} E(X_{i,-j}'X_{i,j})] = \Sigma_{-j,-j}^{-1}\Sigma_{-j,j}. \tag{A.44}$$

Thus, (A.43) and (A.44) 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.44), we obtain

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

where the last equality comes from (A.43). 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 (19) (by replacing $\hat{\gamma}_j$ with γ_j) and T^2 is the population value of $\hat{T}^2$ in (19) (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.39) 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

$$\begin{aligned}\hat{\delta}_j^2 &= \frac{1}{n}(X_{-j}\hat{\gamma}_j + \hat{\eta}_j)'[X_{-j}(\gamma_j - \hat{\gamma}_j) + \eta_j]\\ &= \frac{1}{n}\gamma_j' X_{-j}' X_{-j}(\gamma_j - \hat{\gamma}_j) + \frac{1}{n}\gamma_j' X_{-j}\eta_j + \frac{1}{n}\eta_j' X_{-j}'(\gamma_j - \hat{\gamma}_j) + \frac{1}{n}\eta_j'\eta_j.\end{aligned} \tag{A.45}$$

Therefore, we obtain

$$\begin{aligned}|\hat{\delta}_j^2 - \delta_j^2| &\le |\frac{1}{n}\eta_j'\eta_j - \delta_j^2| + |\frac{1}{n}\eta_j' X_{-j}'(\gamma_j - \hat{\gamma}_j)| + |\frac{1}{n}\gamma_j' X_{-j}\eta_j| + |\frac{1}{n}\gamma_j' X_{-j}' X_{-j}(\gamma_j - \hat{\gamma}_j)|\\ &:= I + II + III + IV.\end{aligned} \tag{A.46}$$

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.18) and (A.22), we obtain $|\frac{1}{\sqrt{n}}\sum_{i=1}^n \eta_{i,j}^2 - \mathrm{E}\eta_{i,j}^2| = O_p(1)$. Therefore,

$$|\frac{1}{n}\eta_j'\eta_j - \delta_j^2| = |\frac{1}{n}\sum_{i=1}^n \eta_{i,j}^2 - \mathrm{E}\eta_{i,j}^2| = O_p(\frac{1}{\sqrt{n}}). \tag{A.47}$$

For (II), we have

$$|\frac{1}{n}\eta_j' X_{-j}'(\gamma_j - \hat{\gamma}_j)| \le \frac{1}{n}\|\eta_j' X_{-j}\|_\infty \|\gamma_j - \hat{\gamma}_j\|_1, \tag{A.48}$$

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.18) and (A.22), we obtain

$$\frac{1}{n}\|\eta_j' X_{-j}\|_\infty = O_p(\sqrt{\log p/n}). \tag{A.49}$$

Together with $\|\gamma_j - \hat{\gamma}_j\|_1 = O_p(s_j\sqrt{\log p/n})$, (A.48) can be bounded

$$|\frac{1}{n}\eta_j' X_{-j}'(\gamma_j - \hat{\gamma}_j)| = O_p(\sqrt{\frac{\log p}{n}})O_p(s_j\sqrt{\frac{\log p}{n}}) = O_p(\frac{s_j \log p}{n}). \tag{A.50}$$

For (III), we have

$$|\frac{1}{n}\gamma_j' X_{-j}\eta_j| \leq \|\frac{1}{n}X_{-j}'\eta_j\|_\infty \|\gamma_j\|_1. \tag{A.51}$$

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

$$E(X_j'X_j) = \Sigma_{j,j} \geq E[(X_{-j}\gamma_j)'X_{-j}\gamma_j] = \gamma_j'\Sigma_{-j,-j}\gamma_j \geq \Lambda_{min}^2\|\gamma_j\|_2^2, \tag{A.52}$$

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.49), inequality (A.51) 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.37) 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.49) 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.52), $\|\gamma_j\|_2 = O(1)$. 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

$$|\frac{1}{n}\gamma_j' X_{-j}'X_{-j}(\gamma_j - \hat{\gamma}_j)| \leq \|\frac{1}{n}X_{-j}'X_{-j}(\gamma_j - \hat{\gamma}_j)\|_\infty\|\gamma_j\|_1 = O_p(\sqrt{\frac{s_j \log p}{n}}). \tag{A.53}$$

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 (21) and (19), we obtain

$$\begin{aligned}
\|\hat{\Theta}_j - \Theta_j\|_1 &= \|\frac{\hat{C}_j}{\hat{\delta}_j^2} - \frac{C_j}{\delta_j^2}\|_1 = \|\frac{1-\hat{\gamma}_j}{\hat{\delta}_j^2} - \frac{1-\gamma_j}{\delta_j^2}\|_1 \\
&\leq \|\frac{1}{\hat{\delta}_j^2} - \frac{1}{\delta_j^2}\|_1 + \|\frac{\hat{\gamma}_j}{\hat{\delta}_j^2} - \frac{\gamma_j}{\hat{\delta}_j^2} + \frac{\gamma_j}{\hat{\delta}_j^2} - \frac{\gamma_j}{\delta_j^2}\|_1 \\
&\leq \|\frac{1}{\hat{\delta}_j^2} - \frac{1}{\delta_j^2}\|_1 + \|\frac{1}{\hat{\delta}_j^2}\|_1 \|\hat{\gamma}_j - \gamma_j\|_1 + \|\gamma_j\|_1 \|\frac{1}{\hat{\delta}_j^2} - \frac{1}{\delta_j^2}\|_1 \qquad \text{(A.54)} \\
&= O_p(\sqrt{\frac{s_j \log p}{n}}) + O_p(1) O_p(s_j \sqrt{\frac{\log p}{n}}) + O_p(\sqrt{s_j}) O_p(\sqrt{\frac{s_j \log p}{n}}) \\
&= O_p(s_j \sqrt{\frac{\log p}{n}}).
\end{aligned}$$

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.54) 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

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

which completes the proof of Lemma 2. □

A.5 Proof of Proposition 2.1

Proof. We begin our proof by stating the following auxiliary lemma.

Lemma 3 (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 $m_1 > m_2 > 0$, $m^* = m_1 + m_2$, $k = \lfloor n/m^* \rfloor$.* [1] *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 $W_n := \sum_{j=1}^{p} Y_j / (\sum_{j=1}^{k} Y_j^2)^{1/2}$. 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).

We utilize the self-normalized sums properties in Lemma 3 under weak dependence to bound tuning parameters λ_1 and λ_2. To choose appropriate values for tuning 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

$$\mathbb{P}(\|X'\epsilon\|_\infty/n > \lambda_0/2) \leq \alpha, \tag{A.55}$$

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.55), we have

$$\begin{aligned}
\mathbb{P}(\|X'\epsilon\|_\infty/n > \lambda_0/2) &\leq \mathbb{P}(\|X'\epsilon\|_\infty/n > \lambda_0/2, G) + \mathbb{P}(G^C) && \text{(A.56)} \\
&= \mathbb{P}(\max_j |\frac{1}{n}\sum_{i=1}^{n} X_{i,j}\epsilon_i| > \frac{\lambda_0}{2}, G) + \mathbb{P}(G^C) && \text{(A.57)} \\
&\leq p\, \mathbb{P}\left(|\frac{1}{n}\sum_{i=1}^{n} X_{i,j}\epsilon_i| > \lambda_0/2, G \right) + \mathbb{P}(G^C) \leq \alpha. && \text{(A.58)}
\end{aligned}$$

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

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.58) can be written as

$$\mathbb{P}(\|X'\epsilon\|_\infty/n > \lambda_0/2) \leq p\ \mathbb{P}\left(\left\{\frac{|\frac{1}{n}\sum_{i=1}^{n} X_{i,j}\epsilon_i|}{\left(\frac{1}{n}\sum_{i=1}^{n} X_{i,j}^2\epsilon_i^2\right)^{1/2}} > \frac{\lambda_0}{2\sigma(1+1/\log n)^{1/2}}\right\} \cap G\right) + \mathbb{P}(G^C) \tag{A.59}$$

$$\leq 2p\ \mathbb{P}\left(\frac{\sum_{i=1}^{n} X_{i,j}\epsilon_i/\sqrt{n}}{\left(\sum_{i=1}^{n} X_{i,j}^2\epsilon_i^2/n\right)^{1/2}} > \frac{\lambda_0\sqrt{n}}{2\sigma(1+1/\log n)^{1/2}}\right) + o(1) \tag{A.60}$$

$$\leq \alpha. \tag{A.61}$$

Applying the self-normalization theorem of Chen et al. (2016) given in Lemma 3 on (A.60) 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.61), this implies

$$\begin{aligned} 2p\left[1-\Phi\left(\frac{\lambda_0\sqrt{n}}{2\sigma(1+1/\log n)^{1/2}}\right)\right] &\leq \alpha - o(1), \\ \lambda_0 &\geq \frac{2\sigma}{\sqrt{n}}(1+\frac{1}{\log n})^{1/2}\Phi^{-1}(1-\frac{\alpha}{2p}). \end{aligned} \tag{A.62}$$

Since $\lambda_1/n = 2\lambda_0$, we obtain the first part of (11). 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 (11). 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.[2] □

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

B Solving the OWL optimization problem

This section follows similar arguments in Zeng and Figueiredo (2014) and explains how to use the proximal gradient descent algorithm to solve the optimization problem of the OWL estimator. The first subsection introduces the OWL proximal function which is used to compute the optimizer at each step. The second subsection outlines the fast-iterative-soft-thresholding-algorithm (FISTA) used to find the global optimizer, together with a backtracking line search condition which speeds up computation greatly.

B.1 OWL proximal function

Denote by $b = (\beta_1, \cdots, \beta_n)'$, $x = (x_1, \cdots, x_n)'$ column vectors. First we define the proximal function as

$$Prox_{\Omega_\omega}(\beta) = \arg\min_x [\frac{1}{2}||x-\beta||_2^2 + \Omega_\omega(x)], \quad \Omega_\omega(x) = \omega'|x|_\downarrow \tag{B.63}$$

where $\omega \in \kappa$, takes values from a monotone non-negative cone, defined as $\kappa := \{v \in R^n : v_1 \geq v_2 \geq \cdots \geq v_n \geq 0\}$, $|x|_\downarrow = (|x|_{[1]}, |x|_{[2]}, \cdots, |x|_{[n]})'$ and $|x|_{[1]} \geq |x|_{[2]} \geq \cdots \geq |x|_{[n]}$, is the vector of absolute values of elements of vector x, decreasingly ordered. By the definition of $\Omega_\omega(\beta)$, we have

$$\Omega_\omega(\beta) = \Omega_\omega(|\beta|), \tag{B.64}$$

where $|\beta| = (|\beta_1|, \cdots, |\beta_n|)'$. It is easy to show that

$$||\beta - \text{sign}(\beta) \odot |x|||_2^2 \leq ||\beta - x||_2^2, \tag{B.65}$$

where $\text{sign}(\beta) = (\text{sign}(\beta_1), \cdots, \text{sign}(\beta_n))'$ is a function that retrieves signs from a vector, with elements in $\{1, -1, 0\}$ and $\odot$ is a point-wise production operator. Therefore, (B.64) and (B.65) imply

$$Prox_{\Omega_\omega}(\beta) = \text{sign}(\beta) \odot Prox_{\Omega_\omega}(|\beta|). \tag{B.66}$$

Let P be a permutation matrix that orders elements of a vector in decreasing order. Then permutation matrix has property

$$||P(x-b)||_2^2 = ||x-\beta||_2^2, \tag{B.67}$$

and by the definition of $\Omega_\omega(\beta)$,

$$\Omega_\omega(\beta) = \Omega_\omega(Pb). \tag{B.68}$$

So (B.67) and (B.68) imply that (B.66) can be written as

$$Prox_{\Omega_\omega}(\beta) = sign(\beta) \odot P' \, Prox_{\Omega_\omega}(|\beta|_\downarrow), \tag{B.69}$$

where $|\beta|_\downarrow$ is defined as $|x|_\downarrow$, and P' is the transpose of the permutation matrix, which recovers the order of $|\beta|_\downarrow$, i.e. $P|\beta| = |\beta|_\downarrow$, $P'|\beta|_\downarrow = |\beta|$ and $P'P = I$, where I is an identity matrix.

Since $|\beta|_\downarrow \in \kappa$, for any $x^* \in \kappa$ and any $x \in R^n$, we have $|\beta|'_\downarrow x \leq |\beta|'_\downarrow x^*$. Therefore,

$$\begin{aligned}\frac{1}{2}||x-|\beta|_\downarrow||_2^2+\Omega_\omega(x) &= \frac{1}{2}||x||_2^2+\frac{1}{2}|||\beta|_\downarrow||_2^2-|\beta|'_\downarrow x+\Omega_\omega(x)\\ &\geq \frac{1}{2}||x^*||_2^2+\frac{1}{2}|||\beta|_\downarrow||_2^2-|\beta|'_\downarrow x^*+\Omega_\omega(x^*)\\ &= \frac{1}{2}||x^*-|\beta|_\downarrow||_2^2+\Omega_\omega(x^*).\end{aligned}$$

Note that $Prox_{\Omega_\omega}(|\beta|_\downarrow) = \arg\min_x[\frac{1}{2}||x-|\beta|_\downarrow||_2^2+\Omega_\omega(x)]$, and $\frac{1}{2}||x^*-|\beta|_\downarrow||_2^2+\Omega_\omega(x^*) \leq \frac{1}{2}||x-|\beta|_\downarrow||_2^2+\Omega_\omega(x)$. It implies that $Prox_{\Omega_\omega}(|\beta|_\downarrow) \in \kappa$, and $Prox_{\Omega_\omega}(|\beta|_\downarrow) = \arg\min_{x\in\kappa}[\frac{1}{2}||x-|\beta|_\downarrow||_2^2+\omega' x]$. Completing the square, we have

$$Prox_{\Omega_\omega}(|\beta|_\downarrow) = \arg\min_{x\in\kappa}(\frac{1}{2}\|x-|\beta|_\downarrow\|_2^2+\omega' x) = \arg\min_{x\in\kappa}\frac{1}{2}\|x-(|\beta|_\downarrow-\omega)\|_2^2,$$

which is the projection of $(|\beta|_\downarrow - \omega)$ onto κ [3]. Then equation (B.69) can be written as

$$Prox_{\Omega_\omega}(\beta) = \text{sign}(\beta)\odot P'\, Proj_\kappa(|\beta|_\downarrow-\omega)), \tag{B.70}$$

where $Proj_\kappa(.)$ is the projection operator onto κ.

After solving the proximal function, we can employ the iterative soft-thresholding algorithm to find the global optimizer. First, we initialize $\beta^{(0)}$, [4] then repeat

$$\beta^{(k+1)} = prox_{\Omega_\omega}(\beta^{(k)} - sz_k \bigtriangledown g(\beta^{(k)})) \tag{B.71}$$

until a stopping criterion is met, where $k = 1, 2, 3, ...$ are steps of each iteration, $g(\beta) = \frac{1}{2}(y-X\beta)'W(y-X\beta)$ and sz_k is the step size at the k^{th} iteration.

B.2 FISTA algorithm

Algorithm 1 is based on Zeng and Figueiredo (2014) and fast computation is achieved by using the backtracking line condition (step 7) and the acceleration in u (step 12). The backtracking line condition allows large step sizes if optimizer stays in the right direction, otherwise shrinks step sizes. Steps 11 to 12 accelerate computation by moving the optimizer further towards the global optimizer at early iterations, while this acceleration diminishes when approaching the global optimizer.

[3] The projection onto κ is an isotonic optimization problem and can be obtained by using the Pool-Adjacent-Violators algorithm in de Leeuw et al. (2009).

[4] For instance, we use the OLS estimate as initialization in our application but it can be any random vector, which will results in the same global minimizer for β since it is a convex minimization problem. However, a good choice of initialization can reduce computation time greatly.

Algorithm 1: FISTA-OWL

Input: y, C, ω

Output: **OWL estimator** $\hat{\beta}$

Initialisation: $\beta_0 = \hat{\beta}_{OLS}, t_0 = t_1 = 1, u_1 = \beta_0, k = 1, \eta \in (0,1), \tau_0 \in (0, 1/L)$ [a]

while *some stopping criterion not met* **do**

  $\tau_k = \tau_{k-1}$;

  $\beta_k = Prox_{\Omega_\omega}(u_k + \tau * X' * (y - X\beta))$

  while $\frac{1}{2}||y - X\beta_k||_2^2 > Q(\beta_k, u_k)$ [b] **do**

    $\tau_k = \eta * \tau_k$;

    $\beta_k = Prox_{\Omega_\omega}(u_k + \tau * X' * (y - X\beta))$

  end

  $t_{k+1} = (1 + \sqrt{1 + 4t_k^2})/2$

  $u_{k+1} = \beta_k + \frac{t_{k-1}}{t_{k+1}}(\beta_k - \beta_{k-1})$

  $k \leftarrow k + 1$

end

Return: β_{k-1}

[a] L is a Lipschitz constant.

[b] $Q(\beta_k, u_k) := \frac{1}{2}||y - Xu_k||_2^2 - (\beta_k - u_k)'X'(y - Cu_k) + \frac{1}{2\tau_k}||\beta_k - u_k||_2^2$ is the backtracking line condition.

C Geometric interpretation of the OWL penalty

In this section, we follow Zeng and Figueiredo (2014) to show that the OWL estimator achieves sparsity selection and correlation identification simultaneously by inspecting the atomic norm of the OWL penalty term. Next, we will compare the LASSO and OWL penalty terms following a geometric argument, typically illustrated in the machine learning literature. Consider a simple two dimensional case, where $p = 2$. Then, the atomic norm of the LASSO and the OWL penalty terms can be written as

$$\Omega_{LASSO}(\beta) = \lambda|\beta_1| + \lambda|\beta_2| \leq 1, \tag{C.72}$$

$$\Omega_{OWL}(\beta) = \omega_1|\beta|_{[1]} + \omega_2|\beta|_{[2]} \leq 1, \tag{C.73}$$

respectively. Recall that $|\beta|_{[1]} = \max(|\beta_1|, |\beta_2|)$, $|\beta|_{[2]} = \min(|\beta_1|, |\beta_2|)$ and $\omega_1 \geq \omega_2$. Therefore, equation (C.73) can be written as

$$\Omega_{\omega,OWL}(\beta) = \begin{cases} \omega_1|\beta_1| + \omega_2|\beta_2| \leq 1, & \text{if} \quad |\beta_1| \geq |\beta_2|, \\ \omega_1|\beta_2| + \omega_2|\beta_1| \leq 1, & \text{if} \quad |\beta_1| < |\beta_2|, \end{cases} \tag{C.74}$$

which implies that the geometric interpretation of the atomic norm of the LASSO and OWL penalties can be shown as in Figure 4.

From Figure 4, we can see that the LASSO norm has vertexes on both axes, which makes the LASSO estimator enjoy the sparse selection property (i.e., it shrinks one variable to zero while keeping the other non-zero). During estimation, the tangent point between

the penalty norm and the contour coming from the un-regularized solution determines the estimation results. However, when two variables are highly correlated, the frontier of the contour coming from the un-regularized solution is flat. Given the shapes of the LASSO norm and the contour under correlated factors, it is very unstable in determining which variable to shrink. A slight estimation error from the un-regularized solution can easily produce opposite inferences on factors selections. On the other hand, the OWL norm not only has vertexes on both axes, it also has vertexes on the ± 45 degree lines. Those vertexes on the axes produce sparse selection like the LASSO estimator, while those on the ± 45 degree lines yield grouping property which ensures robust factor selection while factors are correlated. When factors are highly correlated, they will be assigned with similar coefficients. Figueiredo and Nowak (2016) (Theorem 2.1) gives more details regarding the grouping property, from which we find that the turning parameter λ_2 and the correlation between factors are crucial factors that influence the grouping property.

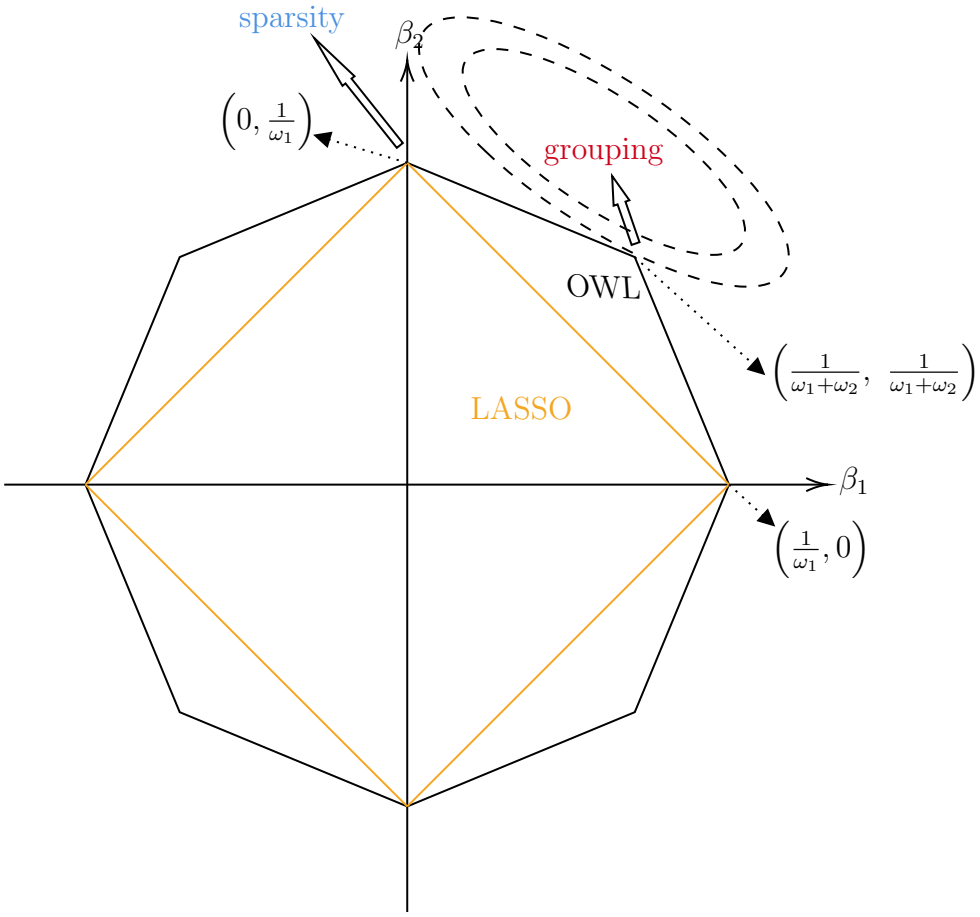

Figure 4. Geometric interpretation of OWL and LASSO penalties
This figure shows the geometric interpretation of the OWL penalty ($\Omega_\omega(\beta) := \omega'|\beta|_\downarrow$) and the LASSO penalty ($\Omega_{LASSO}(\beta) := \lambda\|\beta\|_1$).

D List of Firm Characteristics

Table 2 shows all firm characteristics considered in the factor library.

E Robustness check

In this section, we do some robustness check for the empirical applications by using alternative test assets. We also consider various criteria to screen out micro stocks before portfolio sorting, and inspect its impact to factor selections.

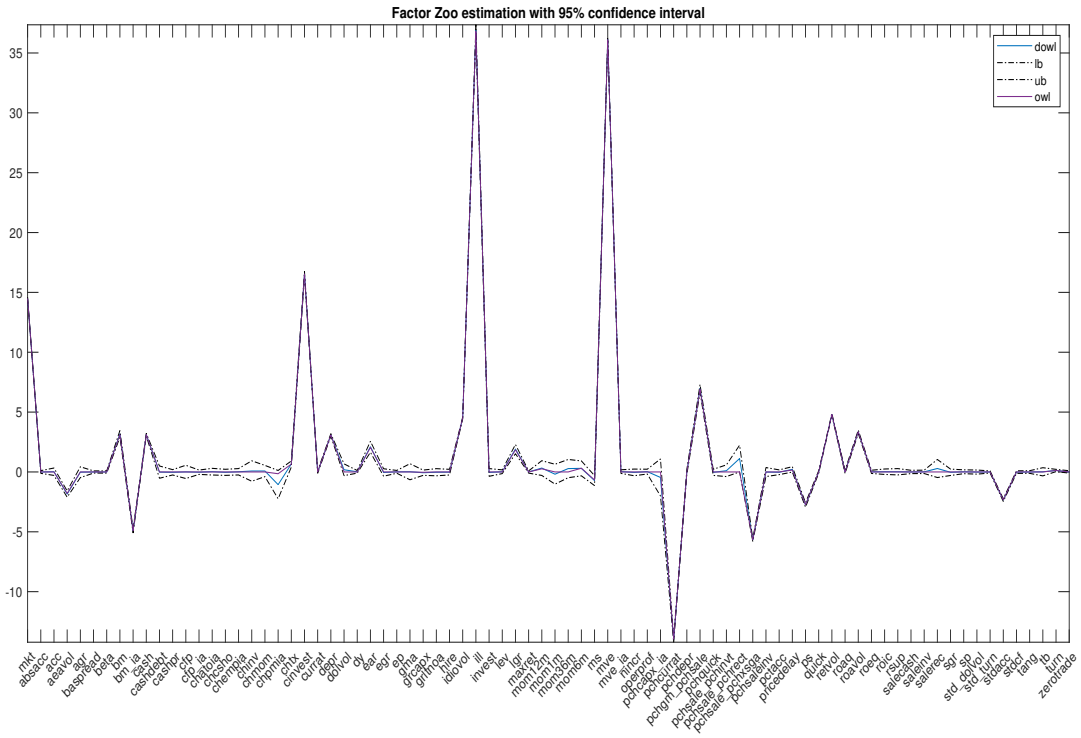

Figure 5. Estimation of factor zoo coefficients using bi-variate sorted portfolios as test assets.

First, for the bi-variate sorted test portfolios, we find that the size factor ('mve') is a dominant factor while insignificant for the uni-variate sorted portfolios. This is the caused by the way how we sort bi-variate sorted portfolios, where 'mve' firm characteristic are used multiple times to form test portfolios. We find the liquidity related factors is also a strong factor in the bi-variate sorted case, followed by other profitability related factors. Comparing the estimated risk prices using the OWL estimator and the debiased OWL estimator, we find some weak factors (i.e., estimated risk price is close to zero) are dropped after applying the 95% confidence interval.

Next, we are going to investigate the impact of the treatment of micro stocks on estimated risk prices. We use the bi-variate sorted test portfolios to conduct our experiment.

Table 2. List of Firm Characteristics

This table lists all 80 firm characteristics considered in our factor library. The abbreviation is consistent with Green et al. (2017). For a more detailed description of each factor, including the original paper where it is proposed, please refer to Green et al. (2017).

Abbreviation	Firm Characteristics	Abbreviation	Firm Characteristics
'absacc'	absolute accruals	'mom1m'	1 month momentum
'acc'	working capital accruals	'mom36m'	36 month momentum
'aeavol'	abnormal earnings announcement volume	'mom6m'	6 month momentum
'agr'	asset growth	'ms'	financial statement score
'baspread'	bid-ask spread	'mve'	size
'beta'	beta	'mve_ia'	industry adjusted size
'betasq'	beta squared	'nincr'	number of earnings increases
'bm'	book-to-market	'operprof'	operating profitability
'bm_ia'	industry adjusted book-to-market	'pchcapx_ia'	i.a. %change in capital expenditures
'cash'	cash holding	'pchcurrat'	% change in current ratio
'cashdebt'	cash flow to debt	'pchdepr'	% change in depreciation
'cashpr'	cash productivity	'pchgm_pchsale'	% change in gross margin - %change in sales
'cfp'	cash flow to price ratio	'pchquick'	%change in quick ratio
'cfp_ia'	industry adjusted cfp	'pchsale_pchinvt'	% change in sale - % change in inventory
'chatoia'	industry adjusted change in asset turnover	'pchsale_pchrect'	% change in sale - % change in A/R
'chcsho'	change in share outstanding	'pchsale_pchxsga'	% change in sale - % change in SG&A
'chempia'	industry adjusted change in employees	'pchsaleinv'	% change in sales-to-inventory
'chinv'	change in inventory	'pctacc'	percent accruals
'chmom'	change in 6-month momentum	'pricedelay'	price delay
'chpmia'	industry adjusted change in profit margin	'ps'	financial statement score
'chtx'	change in tax expense	'quick'	quick ratio
'cinvest'	corporate investment	'retvol'	return volatility
'currat'	current ratio	'roaq'	return on assets
'depr'	depreciation	'roavol'	earning volatility
'dolvol'	dollar trading volume	'roeq'	return on equity
'dy'	dividend to price	'roic'	return on invested capital
'ear'	earnings announcement return	'rsup'	revenue surprise
'egr'	growth in common shareholder equity	'salecash'	sales to cash
'ep'	earnings to price	'saleinv'	sales to inventory
'gma'	gross profitability	'salerec'	sales to receivables
'grcapx'	growth in capital expenditure	'sgr'	sales growth
'grltnoa'	growth in long term net operating assets	'sp'	sales to price
'hire'	employee growth rate	'std_dolvol'	volatility of liquidity (dollar trading volume)
'idiovol'	idiosyncratic return volatility	'std_turn'	volatility of liquidity (share turnover)
'ill'	illiquidity	'stdacc'	accrual volatility
'invest'	capital expenditure and inventory	'stdcf'	cash flow volatility
'lev'	leverage	'tang'	debt capacity/firm tangibility
'lgr'	growth in long term debt	'tb'	Tax income to book income
'maxret'	max daily return	'turn'	share turnover
'mom12m'	12 month momentum	'zerotrade'	zero trading days

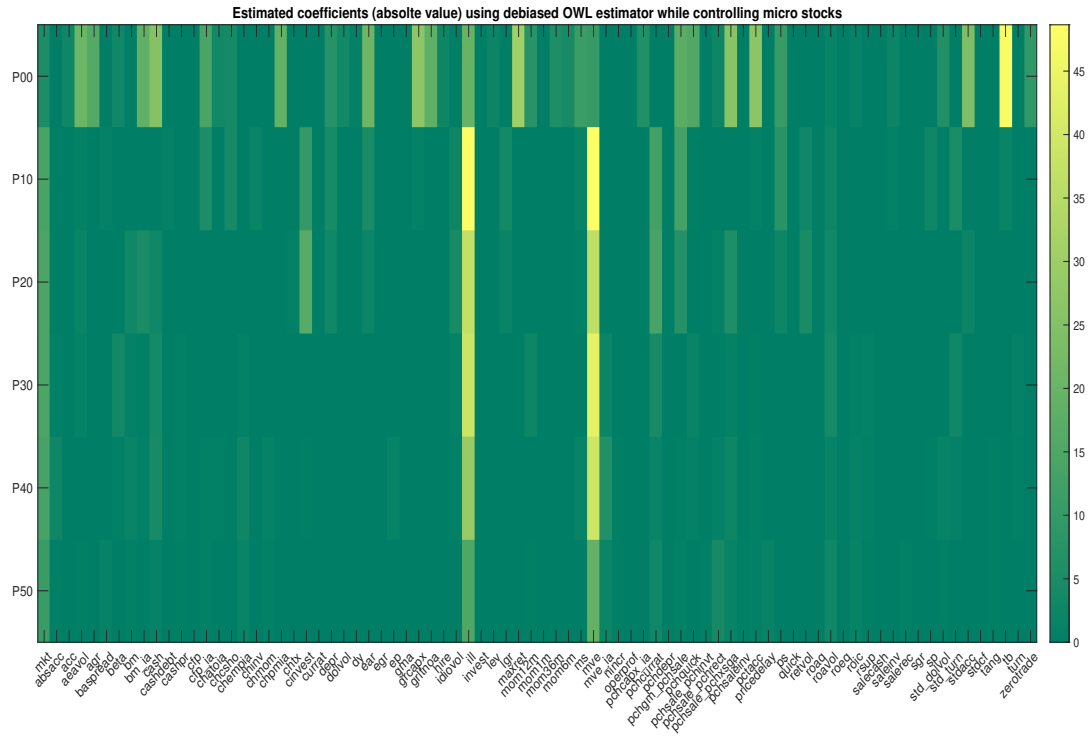

Figure 6. Estimated risk prices (absolute value) while controlling micro stocks

We separately construct the bi-variate sorted portfolios while using different threshold to remove micro stocks. We set 6 levels for controlling micro-stocks: P00 to P50 (i.e., P00 keeps all stocks, while Px0 removes stocks whose market capitalization are smaller than x0 percentile of NYSE listed stocks, where $x \in \{1, 2, 3, 4, 5\}$). Figure 6 shows the estimated risk prices while using different thresholds to remove micro stocks.

References

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.

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

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.

de Leeuw, J., K. Hornik, and P. Mair (2009): "Isotone optimization in R: Pool-adjacent-violators algorithm (PAVA) and active set methods," *Journal of Statistical Software*, 32.

Dendramis, Y., L. Giraitis, and G. Kapetanios (2021): "Estimation of Time-Varying Covariance Matrices for Large Datasets," *Econometric Theory*, 0, 1–35.

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

Green, J., J. R. M. Hand, and X. F. Zhang (2017): "The Characteristics that Provide Independent Information about Average U.S. Monthly Stock Returns," *The Review of Financial Studies*, 30, 4389–4436.

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.

Van De Geer, S., P. Buhlmann, Y. Ritov, and R. Dezeure (2014): "ON ASYMPTOTICALLY OPTIMAL CONFIDENCE REGIONS AND TESTS FOR HIGH-DIMENSIONAL MODELS," *The Annals of Statistics*, 42, 1166–1202.

Zeng, X. and M. A. T. Figueiredo (2014): "The Ordered Weighted L1 Norm: Atomic Formulation, Projections, and Algorithms," *arXiv: Data Structures and Algorithms*.