Contents lists available at ScienceDirect

Economics Letters

journal homepage: www.elsevier.com/locate/ecolet

A correlation-robust shrinkage estimator: Oracle inequality and an application on out-of-sample factor selection

Chuanping Sun

Faculty of Finance, Bayes Business School (formerly Cass), City St George's, University of London, 106 Bunhill Row, London EC1Y 8TZ, United Kingdom

ARTICLE INFO

JEL classification:
C52
C55
G12

Keywords:
Correlation-robust shrinkage
Ordered-weighted LASSO
Oracle inequality

ABSTRACT

Shrinkage methods are widely used in big data to achieve sparse variable selection and reduce overfitting. However, these methods, such as LASSO (Tibshirani, 1996), often struggle when faced with highly correlated predictors. In this paper, we examine a recently developed machine learning estimator that is robust to highly correlated variables, providing superior out-of-sample performance compared to traditional shrinkage techniques. We establish the asymptotic properties of this estimator under general conditions, including i.i.d. sub-Gaussianity. Empirically, we demonstrate the practical benefits of this approach in selecting factors to construct hedged portfolios, achieving significantly higher Sharpe ratios compared to benchmarks such as LASSO, Ridge, and Elastic Net in an out-of-sample context.

1. Introduction

Recent advancements in big data analytics have significantly propelled high-dimensional statistical research, offering new approaches to tackle the curse of dimensionality commonly encountered across various fields. In particular, the Least Absolute Shrinkage and Selection Operator (LASSO) (Tibshirani, 1996) and related shrinkage methods have gained prominence in economic applications due to their ability to perform sparse selection and their well-established statistical properties. However, traditional LASSO estimators often struggle when faced with highly correlated covariates. As Zou and Hastie (2005) noted, LASSO tends to arbitrarily shrink one variable from a highly correlated pair while retaining the other, leading to unstable variable selection results. In the past few decades, hundreds of factors (a.k.a. the factor zoo) have been proposed to explain cross-sectional asset returns. We find that some of these factors are highly correlated, with correlation coefficients exceeding 0.9.[1] This can cause severe complications when using standard methods to select factors. Various methods have since been proposed to address this limitation. One prominent approach is the group LASSO (Yuan and Lin, 2006), which mitigates issues related to highly correlated variables by shrinking groups of variables with similar characteristics together. However, this method requires prior knowledge of group structures—an often non-trivial task that demands rigorous justification. Alternatively, the FARM selection procedure (Fan et al., 2020) addresses the problem by using Principal Component Analysis (PCA) regression to extract common factors from highly correlated covariates and utilizing the residuals for adjusted factors, which typically exhibit low correlations. However, PCA-adjusted factors can compromise the economic interpretability of the original factors.

This paper investigates a recent development in shrinkage methods–the Ordered-Weighted LASSO (OWL) estimator (Figueiredo and Nowak, 2016). The OWL estimator is robust to highly correlated covariates and does not rely on assumptions about factor structures (i.e., how factors should be grouped). We extend the analysis of the OWL estimator by deriving its oracle inequality property under a more general i.i.d. sub-Gaussian framework, relaxing assumptions to accommodate datasets with heavier-tailed variables. Empirically, we apply the OWL shrinkage method to select the most relevant factors driving cross-sectional asset prices in an out-of-sample framework. These selected factors are then used to construct hedge portfolios. Comparisons of OWL-hedged portfolios with benchmarks such as LASSO, Ridge, and Elastic Net reveal that the OWL-hedged portfolios achieve the highest Sharpe ratio. This paper contributes to a growing body of literature that applies shrinkage methods to economic and financial research. For instance, Chinco et al. (2019) employ LASSO to predict stock returns, demonstrating significant improvements in out-of-sample R^2. Feng et al. (2020) use double-LASSO to select factors from the "factor zoo" over time, while Babii et al. (2021) use sparse group LASSO for nowcasting GDP. This paper differs from existing work by focusing on the sparse selection properties of highly correlated predictors and empirically demonstrating the usefulness of the OWL shrinkage method in this context, outperforming other shrinkage techniques.

E-mail address: chuanping.sun@city.ac.uk.

[1] See Fig. 3 in Appendix D for an illustration of the correlations within the factor zoo.

https://doi.org/10.1016/j.econlet.2025.112480
Received 10 October 2024; Received in revised form 26 June 2025; Accepted 30 June 2025
Available online 28 July 2025

The rest of this paper is organized as follows: Section 2 introduces the shrinkage method for factor selection in asset pricing and derives its oracle inequality under more general assumptions. Section 3 applies this method to select factors and highlights its superior out-of-sample performance.

2. Method

2.1. Baseline model

Consider a stochastic discount factor (SDF) asset pricing model where the SDF is defined as a linear function of factors: $m_t = 1 - b'(f_t - \mu)$, where μ represents the expected factor returns and $f_t - \mu$ is the $K \times 1$ factor innovation, and b is a vector of SDF loadings for K factors. The SDF loadings can be estimated using the Generalized Method of Moments (GMM) method with the following moment conditions: $\mathrm{E}(m_t r_t) = 0_{N\times 1}$, and $\mathrm{E}(f_t - \mu) = 0_{K\times 1}$, where r_t is $N \times 1$ test asset returns in excess of risk free rate. The GMM estimator is therefore defined as the minimizer of $\hat{b} = \arg\min_b \hat{g}_T'(b)\hat{W}\hat{g}_T(b)$, where $\hat{g}_T = \begin{pmatrix} \frac{1}{T}\sum_{t=1}^T r_t(1 - b'(f_t - \mu)) \\ \frac{1}{T}\sum_{t=1}^T f_t - \mu \end{pmatrix}_{(N+K)\times 1}$. Solving the above optimization problem gives $\hat{b} = (\hat{C}'\hat{W}\hat{C})'\hat{C}'\hat{W}\bar{r}$, where $\bar{r}$ is the $N \times 1$ average excess returns of test assets and $\hat{C} := \widehat{\mathrm{Cov}}(r_t, f_t)$ is the $N\times K$ sample covariance matrix between r_t and f_t. Following Ludvigson (2013), it is optimal to set $\hat{W} = I$ when the number of test assets is large, and the goal is to infer which factors drive cross-sectional asset returns. Thus, $\hat{b}$ can be interpreted as the solution to a linear regression of $\bar{r}$ on $\hat{C}$. Although recent finance literature argues that the set of factors *explaining* cross-sectional asset returns is dense rather than sparse, e.g., Bryzgalova et al. (2023), fitting too many factors in an out-of-sample forecasting context can lead to overfitting and poor performance, as demonstrated by Freyberger et al. (2020). To address this, we introduce a penalty term to regularize $\hat{b}$. The penalized estimate, $\hat{b}_{penalized}$, is given by: $\hat{b}_{penalized} = \arg\min_b \left(\|\bar{r} - \hat{C}b\|_2^2 + Penalty(\lambda, b)\right)$. By choosing different forms for the penalty function, we can apply various shrinkage methods, such as Ridge, LASSO, and Elastic Net, as shown in (1).

$$Penalty(\lambda, b) = \begin{cases} \lambda\|b\|_2^2, & \text{Ridge} \\ \lambda\|b\|_1, & \text{LASSO} \\ \lambda(\alpha\|b\|_1 + (1-\alpha)\|b\|_2^2), & \text{Elastic Net,} \end{cases} \tag{1}$$

where $\|b\|_2 = \sqrt{\sum_{i=1}^K b_i^2}$ and $\|b\|_1 = \sum_{i=1}^K |b_i|$ denote the ℓ_2 and ℓ_1 norm of the $K \times 1$ vector b, respectively. λ is the shrinkage tuning parameter. Figueiredo and Nowak (2016) introduced the Ordered-Weighted LASSO (OWL) estimator, a shrinkage method that is robust to variable correlations, which can challenge many traditional shrinkage estimators. The OWL estimator is defined as the following:

$$\hat{b}_{OWL} = \arg\min_b \quad \left(\|\bar{r} - \hat{C}b\|_2^2 + \omega'|b|_\downarrow\right), \tag{2}$$

where $|b|_\downarrow = (|b|_{[1]}, |b|_{[2]}, \ldots, |b|_{[j]}, \ldots, |b|_{[K]})$ and $|b|_{[1]} \geq |b|_{[2]} \geq \cdots \geq |b|_{[j]} \geq \cdots \geq |b|_{[K]}$, $\omega = [\omega_1, \omega_2, \ldots, \omega_K]'$ is a $K\times 1$ weighting vector, and $\omega_1 \geq \omega_2 \geq \cdots \geq \omega_K \geq 0$. More specifically, ω is defined as

$$\omega_j = \lambda_1 + \lambda_2(K - j), \quad j = 1, \cdots, K, \tag{3}$$

where λ_1 and λ_2 are two tuning parameters. In accordance with the machine learning literature, we often employ cross-validation to determine the appropriate tuning parameters. The key contribution of the OWL estimator, compared to other shrinkage methods such as LASSO, lies in its ability to handle highly correlated variables, which often pose significant challenges. The OWL estimator possesses two important properties simultaneously: the shrinkage property, which shrinks unimportant variables to zero, and the grouping property, which identifies highly correlated variables and assigns them similar coefficients, without requiring any structural assumptions about the factors. In contrast, the LASSO method only possesses the shrinkage property. When faced with highly correlated variables, LASSO tends to inconsistently shrink some variables to zero while keeping others non-zero, leading to instability in variable selection. See Appendix A for a detailed comparison between these shrinkage estimators. In the next section, we establish the statistical properties of the OWL estimator under more general assumptions.[2]

2.2. Statistical properties

Without loss of generality and for simplicity of notations, (2) can be written as:

$$\hat{b}_{OWL} = \arg\min_b \quad \left(\frac{1}{N}\|y - Xb\|_2^2 + \frac{1}{N}\omega'|b|_\downarrow\right), \tag{4}$$

where y and X are $N \times 1$ vector and $N \times K$ matrix, respectively. To facilitate the analysis, consider the following linear regression model:

$$y = Xb^0 + \epsilon, \tag{5}$$

where b^0 represents the true $K \times 1$ vector of coefficients and ϵ is the error term. In a high-dimensional setting, K can exceed N. We now discusses the asymptotic properties of the OWL estimator, extending the framework of Figueiredo and Nowak (2016) to more general assumptions. We begin by introducing some notations and assumptions in Appendix B. It is worth noting that these assumptions are more general in the statistical literature and more relaxed than those in Figueiredo and Nowak (2016).

Theorem 2.1 (*Oracle Inequality*). *Let Assumptions 1, 2 and 3 be satisfied. Suppose that* $\lambda_0 = \kappa\sqrt{\dfrac{\log K}{N}} = o(1)$*, where* κ *is a positive constant. Let* $\dfrac{\lambda_1}{N} = 2\lambda_0$ *and* $\dfrac{\lambda_2}{N} = O(\dfrac{S\log K}{NK})$*. Then, by selecting a sufficiently large* κ*, as* $N, K \to \infty$*, with probability tending to one,* $\hat{b}$ *satisfies*

$$(\hat{b} - b^0)'\hat{\Sigma}(\hat{b} - b^0) + \frac{\lambda_1}{N}\|\hat{b} - b^0\|_1 \leq 4(\frac{\lambda_1}{N})^2\frac{S}{\phi_0^2} + 2\frac{\lambda_2}{N}(K-1)\|b^0\|_1. \tag{6}$$

Proof. see Appendix C.

The oracle inequality in (6) can be further developed to offer upper bounds separately for the prediction error $(\hat{b} - b^0)'\hat{\Sigma}(\hat{b} - b^0) := \|X(\hat{b} - b^0)\|_2^2/N$ and the estimation error $\|\hat{b} - b^0\|_1$. These bounds are crucial in determining the convergence rate of the OWL estimator.

3. Empirical application

In this section, we apply several shrinkage methods, including LASSO, Ridge, Elastic Net, and OWL, to select factors from the "factor zoo" to explain cross-sectional asset prices in an out-of-sample framework.[3] We use the Open Source Asset Pricing dataset from Chen and Zimmermann (2022) for our empirical analysis. We discuss the data and how we clean the data in Appendix D. To evaluate out-of-sample performance, we implement a rolling window approach to select factors from the factor zoo, as provided by the Open Asset Pricing factor library.

[2] Note that we develop the asymptotic properties while relaxing the *i.i.d.* normality assumption on variables made in Figueiredo and Nowak (2016).

[3] We use the 'cvxpy' package and the 'mosek' solver for optimization problems for LASSO, Ridge and Elastic Net shrinkage methods, while we develop our own optimization method and code for solving the OWL shrinkage problem. A detailed explanation of the algorithm can be found in Sun (2024). The tuning parameters for those shrinkage methods are determined using the cross-validation, by searching for the best values, given a grid of candidate values, that produce the smallest out-of-sample mean squared errors using multiple splittings for training and testing samples. The optimal values for tuning parameters are between 10^{-5} and 10^{-6}.

Table 1
Sharpe ratio comparison.

Out-of-sample sharpe ratio of hedged portfolios				
	LASSO	Ridge	EN	OWL
win=240	0.2359	0.1132	0.3161	0.3673
win=360	0.5410	0.3222	0.5001	0.7208
win=480	0.3667	0.2865	0.4154	0.9526

Note: this Table reports the Sharpe ratio of the hedged portfolios using top 5 selected factors via four competing methods, including LASSO, Ridge, elastic net and OWL.

Specifically, we define a window size $win = \{240, 360, 480\}$ months. At each time t, we estimate our four competing models using data from $t-\text{win}+1$ to t. Within each rolling window, we first run a cross-sectional regression of average test portfolio returns on the covariance between test portfolios and all factors.[4]

The estimated model is then used to identify a small set of important factors for predicting each test portfolio's return in the next period. For simplicity and comparability, we select the five most important factors to forecast returns for all test portfolios at time $t + 1$ by running a predictive linear regression of each test portfolio on the selected factors.

After obtaining the predicted returns, we sort test portfolios into deciles based on their predicted returns and construct a hedging strategy by going long on the top decile portfolio and short on the bottom decile portfolio.

Table 1 reports the out-of-sample Sharpe ratios for hedged portfolios formed using the top 5 selected factors from each of the four competing methods: LASSO, Ridge, Elastic Net, and OWL. We conduct robustness checks by using three different rolling window sizes (240, 360, and 480 months). The results show that OWL-hedged portfolios consistently deliver the highest Sharpe ratios compared to other benchmarks. Moreover, the OWL-hedged portfolio performs best when the rolling window is set to 480 months, providing sufficient historical data to estimate the primary factors driving cross-sectional asset returns. In this case, the OWL achieves a Sharpe ratio more than doubles of other benchmarks.

Next, we examine the most frequently selected factors in the out-of-sample period by each of the four methods. Fig. 1 shows that all methods agree on the importance of the 'Market' factor in driving asset prices. Additionally, reversal factors, volatility related factors are commonly selected by all methods, although with different interpretations. LASSO frequently selects 'short-term reversal' and 'return on asset' as key factors after the 'market' factor. Notably, LASSO shows a shift in importance from 'short-term reversal' to 'return on assets' in the second half of the out-of-sample period. In contrast, OWL identifies 'long-term reversal' and 'idiosyncratic volatility' as key factors following the 'market' factor, with a shift towards 'idiosyncratic volatility' in the late 2010s. However, in the most recent two years, 'idiosyncratic volatility' loses its importance, while factors such as 'Analyst Valuation' and 'Asset Growth' become more significant in forecasting asset prices.

4. Conclusion

We extend the statistical properties of a correlation-robust shrinkage method under relaxed assumptions, a framework commonly used in economic research. This method is applied alongside other benchmarks to select factors for predicting cross-sectional asset returns and constructing hedged portfolios. Our empirical results show that the OWL-hedged portfolio consistently achieves the highest Sharpe ratios compared to other methods.

For future research, further development of the statistical properties of this correlation-robust shrinkage estimator could involve relaxing assumptions, such as removing the i.i.d. assumption on variables, and investigating the consistency of variable selection. Additionally, developing a debiased version of the estimator would enable statistical inference and further enhance its practical application.

Overall, this shrinkage method is particularly valuable when dealing with highly correlated variables in high-dimensional settings, making it a robust and practical tool for financial modeling.

Appendix A. The geometric interpretation of the OWL penalty and its comparison with the LASSO and the Elastic Net penalties

Fig. 2 shows the geometric representation of the penalty terms of LASSO, Elastic Net and OWL shrinkage methods. The LASSO penalty is demonstrated as the diamond-shaped rectangular, where the OWL penalty is the octagonal-shaped one. The tangent point between the penalty term and the contour from the un-regularized least square estimator determines the shrinkage estimator. 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 penalty 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 EN penalty is curved by combining the LASSO and Ridge penalties together. The curved edge stabilizes the tangent point with a flat contour of the un-regularized solution. Therefore, it avoids randomly shrinking one highly correlated variable to zero while keeping the other as non-zero, alleviating the unstable solutions from the LASSO shrinkage method. Finally, the OWL penalty is octagonal shaped, it 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 encourage assigning similar coefficients for highly correlated variables, as these vertexes on the ±45 degree lines are most likely to have the tangent point with a flat contour from the un-regularized solutions. This is regarded as the grouping property which ensures robust factor selection while factors are correlated. When factors are highly correlated, they will be assigned with similar coefficients.

Appendix B. Notations and assumptions

Let $\zeta_j := \epsilon' X^{(j)} := \sum_{i=1}^N \epsilon_i X_i^{(j)} := \sum_{i=1}^N \zeta_{i,j}$, where $X^{(j)}$ is the jth column of X and ϵ is defined in (5). We denote $\hat{\Sigma} = \frac{1}{N} X'X$ as the scaled Gram Matrix of X. For any vector $x \in R^N$, we denote $\|x\|_2 = (\sum_{i=1}^N x_i^2)^{1/2}$, $\|x\|_1 = \sum_{i=1}^N |x_i|$ and $\|x\|_\infty = \max_{1\le i\le N} |x_i|$. Let s_0 denote a subset, $s_0 \subset \{1, \dots, K\}$, and $|s_0|$ the cardinality of s_0. For $b = \{b_1, \dots, b_K\} \in \mathbf{R}^K$, denote $b_{s_0} := b_i \mathbf{1}\{i \in s_0, i = 1, \dots, K\}$, $b_{s_0^c} := b_i \mathbf{1}\{i \notin s_0, i = 1, \dots, K\}$. Then $b = b_{s_0} + b_{s_0^c}$. We establish the following assumptions.

Assumption 1 (*Random Variables*). $\{\zeta_{i,j}\}_{i=1}^N$ are identically and independently distributed and $\mathrm{E}(\zeta_{i,j}) = 0$ for $i = 1, \dots, N$ and $j = 1, \dots, K$. The distributions of variable $X_i^{(j)}$ and ϵ_i for all $i = 1, \dots, N$ are uniformly subgaussian such that $\sup_{i,j} \mathbb{P}(|X_i^{(j)}| > a) \le c_1 \exp[-c_2 a^2]$ and $\sup_i \mathbb{P}(|\epsilon_i| > a) \le c_1 \exp[-c_2 a^2]$ for all $i = 1, \dots, N$, $a > 0$ and some $c_1, c_2 > 0$ which do not depend on a, i, j.

Assumption 2 (*Sparsity*). Denote by S the number of non-zero parameters in $b^0 = \{b_1^0, b_2^0, \dots, b_K^0\}$. We assume that $S\sqrt{\frac{\log K}{N}} = o(1)$ when $N, K \to \infty$.

[4] This step is similar to the Fama–MacBeth regression for inferring risk premiums. The difference here is that we infer risk prices. See Sun (2024) for a detailed discussion on the relationship between these methods.

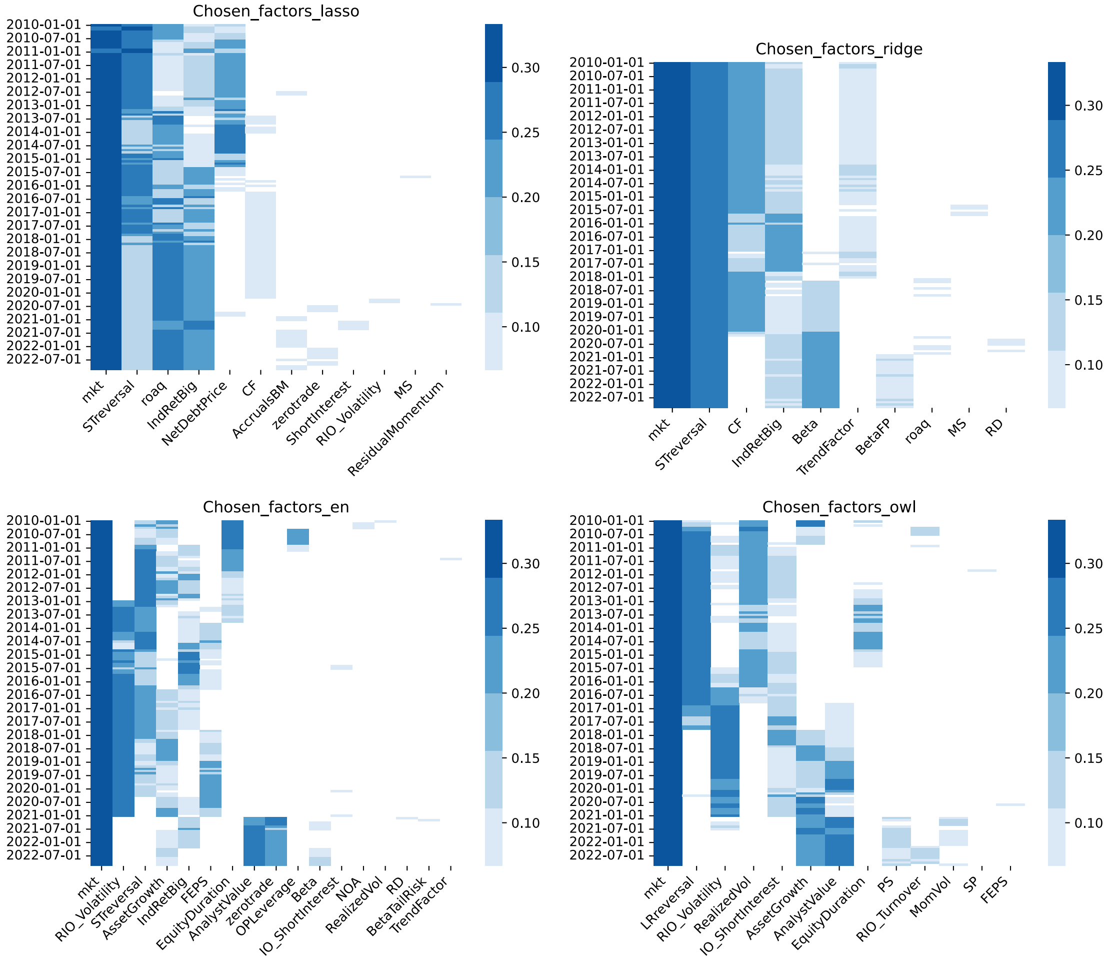

Fig. 1. Out-of-sample factor selections with win=360.
Note: This table list the most selected factors during the out-of-sample period, ordered by frequency that they are selected by each candidate model. The color indicate how important the selected factor is for driving cross-sectional asset returns. The darker the color the more importance of the factor.

Assumption 3 (*Compatibility Condition, Buhlmann and Van De Geer, 2011*)**.** For all b such that $\|b_{s_0^c}\|_1 \leq 3\|b_{s_0}\|_1$, we have

$$\phi_0^2 := \min_{\substack{s_0\subset\{1,\ldots,K\}\\ |s_0|<K}} \; \min_{\substack{b\in R^K\setminus\{0\}\\ \|b_{s_0^c}\|_1\leq 3\|b_{s_0}\|_1}} \frac{b'\hat{\Sigma}bS}{\|b_{s_0}\|_1^2} > 0. \tag{B.1}$$

Assumption 1 specifies that the distributions of the random variables are *i.i.d.* and sub-Gaussian. This assumption is more general than the *i.i.d.* normality assumption made by Figueiredo and Nowak (2016), allowing for heavier tails in the variables. The adoption of *i.i.d.* sub-Gaussian assumptions is consistent with standard practices in high-dimensional econometrics, as noted by Kock (2016), Kock and Tang (2019). Assumption 2, governs the growth rate of the dimension of X, the sparsity parameter S, and the number of observations N. Importantly, the exact sparsity level S is not predetermined. Assumption 3, the compatibility condition, addresses challenges posed by a degenerate scaled Gram matrix in high-dimensional factor models. This condition is less restrictive than the commonly used restricted eigenvalue or irrepresentable conditions in high-dimensional statistics, as discussed by Van de Geer and Bühlmann (2009).

Appendix C. Proof of Theorem 2.1

Proof. By definition the OWL estimator is minimizing the function

$$\hat{b} = \hat{b}_{OWL} = \arg\min_{b} \quad \frac{1}{N}\|y - Xb\|_2^2 + \frac{1}{N}\sum_{i=1}^{K}[\lambda_1 + \lambda_2(K-i)]|b|_{[i]},$$

where $|b|_{[\cdot]}$ denotes the element of the decreasingly ordered vector of $|\mathbf{b}|$, such that $|b|_{[1]} \geq |b|_{[2]} \geq \cdots \geq |b|_{[K]}$. Let b^0 be the vector of true values of risk prices, and $y = Xb^0 + \epsilon$. According to the "argmin" property, definition of $\hat{b}$ implies

$$\frac{1}{N}\|y - X\hat{b}\|_2^2 + \frac{1}{N}\sum_i[\lambda_1 + \lambda_2(K-i)]|\hat{b}|_{[i]} \leq \frac{1}{N}\|y - Xb^0\|_2^2 + \frac{1}{N}\sum_i[\lambda_1 + \lambda_2(K-i)]|b^0|_{[i]}. \tag{C.1}$$

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

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

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

$$\frac{1}{N}\|X(\hat{b}-b^0)\|_2^2 + \frac{\lambda_1}{N}\|\hat{b}\|_1 \leq \frac{2}{N}\epsilon' X_j(\hat{b}-b^0) + \frac{1}{N}[\lambda_1+\lambda_2(K-1)]\|b^0\|_1. \tag{C.2}$$

Note that

$$2|\epsilon' X(\hat{b} - b^0)| \leq \left(\max_{1\leq j\leq K} 2|\epsilon' X^{(j)}|\right)\|\hat{b} - b^0\|_1. \tag{C.3}$$

Hence, (C.2) can be written as

$$\frac{1}{N}\|X(\hat{b} - b^0)\|_2^2 + \frac{\lambda_1}{N}\|\hat{b}\|_1 \leq \left(\frac{1}{N}\max_{1\leq j\leq K} 2|\epsilon' X^{(j)}|\right)\|\hat{b} - b^0\|_1$$

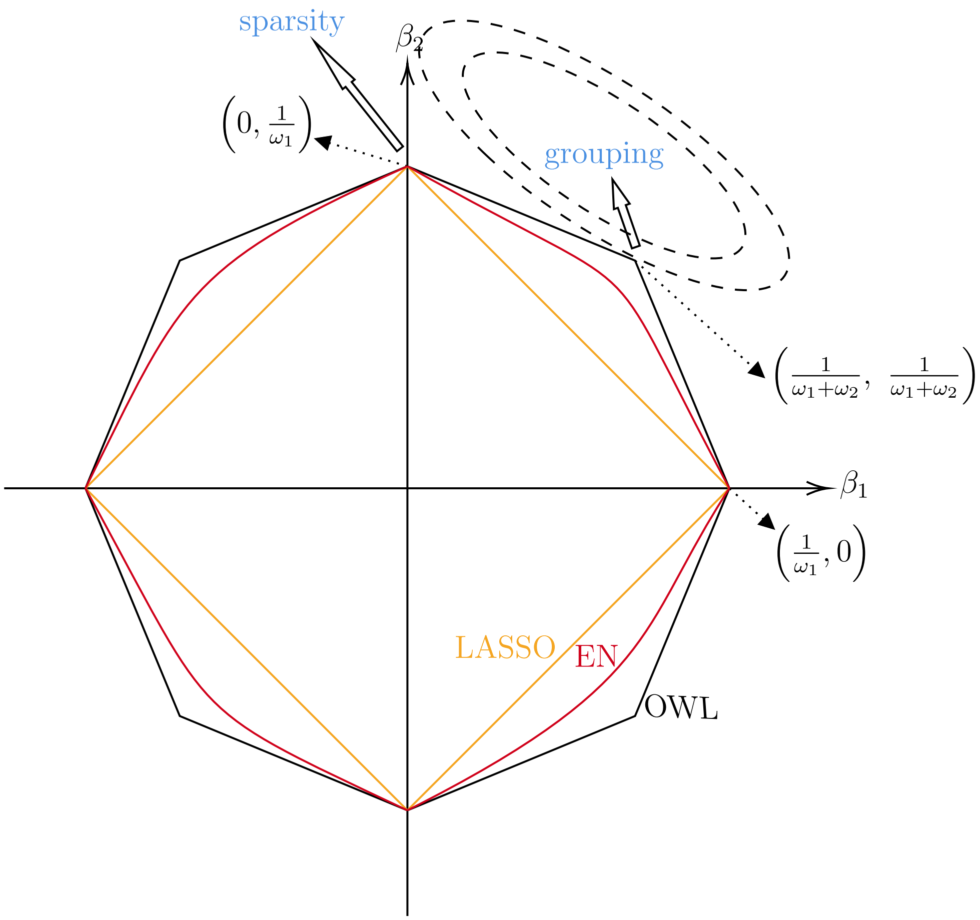
sparsity
β_2
$\left(0, \frac{1}{\omega_1}\right)$
grouping
$\left(\frac{1}{\omega_1+\omega_2}, \frac{1}{\omega_1+\omega_2}\right)$
β_1
$\left(\frac{1}{\omega_1}, 0\right)$
LASSO
EN
OWL

Fig. 2. Geometric interpretation of LASSO, EN, and OWL penalties.

$$+\frac{1}{N}[\lambda_1+\lambda_2(K-1)]\|b^0\|_1. \tag{C.4}$$

Consider the event

$$E := \left\{\frac{1}{N}\max_{1\le j\le K} 2|\epsilon' X^{(j)}| \le \lambda_0\right\}, \tag{C.5}$$

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

$$\frac{1}{N}\|X(\hat{b}-b^0)\|_2^2+\frac{1}{N}\lambda_1\|\hat{b}\|_1 \le \lambda_0\|\hat{b}-b^0\|_1+\frac{1}{N}[\lambda_1+\lambda_2(K-1)]\|b^0\|_1. \tag{C.6}$$

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

$$\frac{2}{N}\|X(\hat{b}-b^0)\|_2^2+\frac{2}{N}\lambda_1\|\hat{b}\|_1 \le \frac{\lambda_1}{N}\|\hat{b}-b^0\|_1+\frac{2}{N}[\lambda_1+\lambda_2(K-1)]\|b^0\|_1. \tag{C.7}$$

Note that

$$\|\hat{b}\|_1 = \|\hat{b}_{s_0}\|_1 + \|\hat{b}_{s_0^c}\|_1 \ge \|b^0_{s_0}\|_1 - \|\hat{b}_{s_0} - b^0_{s_0}\|_1 + \|\hat{b}_{s_0^c}\|_1, \tag{C.8}$$

$$\|\hat{b}-b^0\|_1 = \|\hat{b}_{s_0} - b^0_{s_0}\|_1 + \|\hat{b}_{s_0^c}\|_1. \tag{C.9}$$

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

$$\begin{aligned}&\frac{2}{N}\|X(\hat{b}-b^0)\|_2^2+\frac{2\lambda_1}{N}(\|b^0_{s_0}\|_1-\|\hat{b}_{s_0}-b^0_{s_0}\|_1+\|\hat{b}_{s_0^c}\|_1)\\ &\le \frac{\lambda_1}{N}(\|\hat{b}_{s_0}-b^0_{s_0}\|_1+\|\hat{b}_{s_0^c}\|_1)+\frac{2\lambda_1}{N}\|b^0\|_1+\frac{2\lambda_2(K-1)}{N}\|b^0\|_1.\end{aligned} \tag{C.10}$$

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

$$\frac{2}{N}\|X(\hat{b}-b^0)\|_2^2+\frac{\lambda_1}{N}\|\hat{b}_{s_0^c}\|_1 \le 3\frac{\lambda_1}{N}\|\hat{b}_{s_0}-b^0_{s_0}\|_1+\frac{2\lambda_2(K-1)}{N}\|b^0\|_1. \tag{C.11}$$

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

$$\frac{2}{N}\|X(\hat{b}-b^0)\|_2^2+\frac{\lambda_1}{N}\|\hat{b}-b^0\|_1 \le 4\frac{\lambda_1}{N}\|\hat{b}_{s_0}-b^0_{s_0}\|_1+\frac{2\lambda_2(K-1)}{N}\|b^0\|_1. \tag{C.12}$$

By Assumption 3, we have

$$\|b_{s_0}\|_1^2 \le b'\hat{\Sigma}bS/\phi_0^2. \tag{C.13}$$

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

$$\|\hat{b}_{s_0}-b^0_{s_0}\|_1^2 \le (\hat{b}-b^0)'\hat{\Sigma}(\hat{b}-b^0)S/\phi_0^2 = \|X(\hat{b}-b^0)\|_2^2 S/(N\phi_0^2),$$
$$\|\hat{b}_{s_0}-b^0_{s_0}\|_1 \le \|X(\hat{b}-b^0)\|_2\sqrt{S}/(\sqrt{N}\phi_0).$$

Therefore, using inequality $4ab \le a^2+4b^2$, we obtain

$$\begin{aligned}4\frac{\lambda_1}{N}\|\hat{b}_{s_0}-b^0_{s_0}\|_1 &\le 4\left(\frac{\|X(\hat{b}-b^0)\|_2}{\sqrt{N}}\right)\left(\frac{\lambda_1}{N}\frac{\sqrt{S}}{\phi_0}\right)\\ &\le \frac{1}{N}\|X(\hat{b}-b^0)\|_2^2+4(\frac{\lambda_1}{N})^2\frac{S}{\phi_0^2}.\end{aligned}$$

So (C.12) can be written as

$$\frac{1}{N}\|X(\hat{b}-b^0)\|_2^2+\frac{\lambda_1}{N}\|\hat{b}-b^0\|_1 \le 4(\frac{\lambda_1}{N})^2\frac{S}{\phi_0^2}+\frac{2\lambda_2(K-1)}{N}\|b^0\|_1. \tag{C.14}$$

Note that $\frac{1}{N}\|X(\hat{b}-b^0)\|_2^2 = (\hat{b}-b^0)'\hat{\Sigma}(\hat{b}-b^0)$, so (C.14) completes the proof of (6).

Now we have obtained (6) assuming (C.5). In the next step we want to evaluate the probability of the inequality (C.5) to be true, i.e. $\mathbb{P}(E)$. By a union bound and using the notation $\zeta_j = \epsilon' X^{(j)} = \sum_{i=1}^N \epsilon_i X_i^{(j)} =$

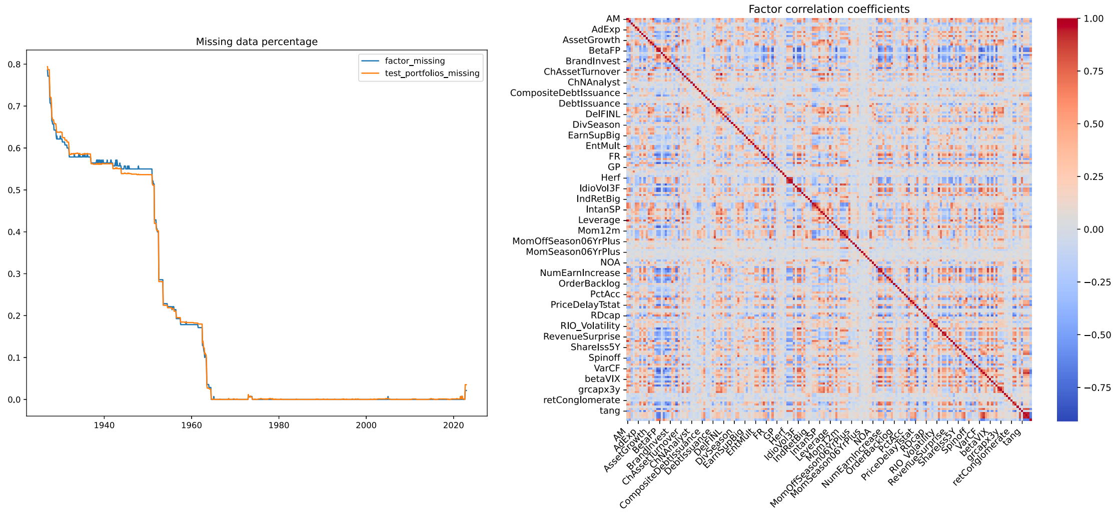

Fig. 3. Preliminary analysis on data.

$\sum_{i=1}^N \zeta_{i,j}$, we obtain

$$\mathbb{P}(E^C) = \mathbb{P}(\frac{1}{N}\max_{1\le j\le K} 2|\epsilon' X^{(j)}|) \ge \lambda_0 \le \sum_{j=1}^{K} \mathbb{P}(\frac{1}{N}|\zeta_j| \ge \frac{\lambda_0}{2}). \tag{C.15}$$

Note that both $\{\epsilon_i\}_{i=1}^N$ and $\{X_i^{(j)}\}_{i=1}^N$ for all $i = 1, \dots, N$ and $j = 1, \dots, K$ are uniformly subgaussian variables. Therefore, variables $\{\zeta_{i,j}\}_{i=1}^N$ are uniformly subexponentially distributed. Hence, applying Corollary 5.17 in Vershynin (2012) and utilizing $\lambda_0 = \kappa\sqrt{\frac{\log K}{N}}$, we obtain

$$\mathbb{P}(E^C) \le K \max_{1\le j\le K} \mathbb{P}(\frac{1}{N}|\zeta_j| \ge \frac{\lambda_0}{2}) = K \max_{1\le j\le K} \mathbb{P}(\frac{1}{N}\left|\sum_{i=1}^N \zeta_{i,j}\right| \ge \frac{\lambda_0}{2})$$
$$\le 2K\exp[-c\kappa^2 \log K] = 2K^{1-c\kappa^2}.$$

where c and κ are positive constants. Therefore, selecting κ such that $c\kappa^2 > 1$, we have the following property for (C.5):

$$\mathbb{P}(E) = 1 - \mathbb{P}(E^C) \ge 1 - 2K^{1-c\kappa^2} \to 1, \tag{C.16}$$

as $N, K \to \infty$. This completes the proof of Theorem 2.1. □

Appendix D. Data

To study firm characteristics and their ability to predict stock returns, we use data from the Open Source Asset Pricing dataset from Chen and Zimmermann (2022).[5] This dataset includes 212 firm characteristic-based factors with monthly returns from January 1926 to December 2022. Factors are constructed by sorting stocks into decile portfolios according to a given characteristic at each point in time. The spread return between the top and bottom deciles represents the factor return associated with that characteristic. The decile portfolios for all characteristics are used as test portfolios. However, missing data presents a challenge in training and validating models. The left panel of Fig. 3 shows the percentage of missing data over time for factors and for test portfolios. Before the mid-1960s, the missing data for both factors and portfolios is substantial, exceeding 50% before 1950. After the mid-1960s, the missing data percentage falls below 5%, and drops to less than 1% after 1980. For this reason, we restrict our analysis to data from January 1980 to December 2022 and exclude columns containing any missing values.[6] Additionally, we obtain market returns ('Mkt') and the risk-free rate from Kenneth French's website.[7]

After addressing the missing data, we conduct a preliminary analysis to examine the correlations between factors. The right panel of Fig. 3 reveals that some factors exhibit very high correlations, indicated by dark colors on the heatmap (some correlation coefficients exceed 0.9). This raises concerns about the reliability and stability of LASSO in selecting variables, as Zou and Hastie (2005) highlight that high correlations can make LASSO variable selection unreliable. In contrast, the OWL estimator is well-suited for handling highly correlated variables.

Data availability

I have shared the link to the data in the appendix.

References

Babii, A., Ghysels, E., Striaukas, J., 2021. Machine learning time series regressions with an application to nowcasting. J. Bus. Econom. Statist. 1–23.

Bryzgalova, S., Huang, J., Julliard, C., 2023. Bayesian solutions for the factor zoo: We just ran two quadrillion models. J. Financ. 78 (1).

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

Chen, A.Y., Zimmermann, T., 2022. Open source cross-sectional asset pricing. Crit. Financ. Rev. 11 (2).

Chinco, A., Clark-Joseph, A.D., Ye, M., 2019. Sparse signals in the cross-section of returns. J. Financ. 74 (1), 449–492.

Fan, J., Ke, Y., Wang, K., 2020. Factor-adjusted regularized model selection. J. Econom. 216 (1).

Feng, G., Giglio, S., Xiu, D., 2020. Taming the factor zoo: A test of new factors. J. Financ. 75 (3), 1327–1370.

Figueiredo, M., Nowak, R., 2016. Ordered weighted L1 regularized regression with strongly correlated covariates: Theoretical aspects. In: Proceedings of the 19th International Conference on Artificial Intelligence and Statistics. pp. 930–938.

Freyberger, J., Neuhierl, A., Weber, M., 2020. Dissecting characteristics nonparametrically. Rev. Financ. Stud. 33 (1), 2326–2377.

Kock, A.B., 2016. Oracle inequalities , variable selection and uniform inference in high-dimensional correlated random effects panel data models. J. Econom. 195 (1), 71–85.

Kock, A.B., Tang, H., 2019. Uniform inference in high-dimensional dynamic panel data models with approximately sparse fixted effects. Econometric Theory 35, 295–359.

[5] The data is downloaded from https://www.openassetpricing.com/.

[6] For the optimization problem with the shrinkage methods, containing missing values in the data may trigger non-convex errors. Therefore, we delete variables that contains any missing values. In the end, we obtain 170 factors with no missing values.

[7] Market data and risk free rates are downloaded from https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html.

Ludvigson, S.C., 2013. Advances in consumption-based asset pricing: Empirical tests. In: Handbook of the Economics of Finance, vol. 2, (PB).

Sun, C., 2024. Factor correlation and the cross section of asset returns: A correlation-robust machine learning approach. J. Empir. Financ. 77.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B 58 (1), 267–288.

Van de Geer, S.A., Bühlmann, P., 2009. On the conditions used to prove oracle results for the lasso. Electron. J. Stat. 3.

Vershynin, R., 2012. Introduction to the non-asymptotic analysis of random matrices. In: Compressed Sensing: Theory and Applications.

Yuan, M., Lin, Y., 2006. Model selection and estimation in regression with grouped variables. J. R. Stat. Soc. Ser. B 68 (1), 49–67.

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B 67 (2), 301–320.