Estimating the Block-Diagonal Idiosyncratic Covariance in High-Dimensional Factor Models

Lucija Žignić*, Stjepan Begušić†, and Zvonko Kostanjčar†

*Department for Strategy and Operations, PricewaterhouseCoopers
†University of Zagreb, Faculty of Electrical Engineering and Computing
Laboratory for Financial and Risk Analytics, Unska 3, 10000 Zagreb, Croatia
lucija.zignic@fer.hr, stjepan.begusic@fer.hr, zvonko.kostanjcar@fer.hr.

***Abstract*—Factor models are often used to infer lower-dimensional correlation structures in data, especially when the number of variables grows close to or beyond the number of data points. The data covariance under a factor model structure is a combination of a low-rank component due to common factors and a diagonal or sparse idiosyncratic component. In this paper we consider the estimation of the idiosyncratic component under the assumption of grouped variables, which result in a block-diagonal matrix. We propose a shrinkage approach which ensures the positive definiteness of the estimated matrix, using either known group structures or clustering algorithms to determine them. The proposed methods are tested in a portfolio optimization scenario using simulations and historical data. The results show that the cluster based estimators yield improved performance in terms of out-of-sample portfolio variance, as well as remarkable stability in terms of resilience to the error in the estimated number of latent factors.**

***Index Terms*—high-dimensional factor model, idiosyncratic component, thresholding, shrinkage, clustering**

I. Introduction

With the growth of available data, high-dimensional covariance matrices are becoming increasingly important in many areas [1]. When the number of variables p is similar to, or even greater than the number of data points or the length of the time series T, the classical sample covariance estimator is driven by noise and suffers from invertibility and stability issues [2]–[4]. To address the problem, different approaches have been proposed – some model-free [5]–[7], some based on a factor model interpretation [8]–[10].

In this paper we are motivated by the correlation structures in high-dimensional financial time series. It is a well documented fact that a small set of common factors (such as macroeconomic shocks, interest rates, or the market factor itself) drive the observed dynamics of a large number of time series [11]. These factors form a low-rank covariance structure which only depends on the estimated factor loadings [8]. These loadings can be estimated via maximum likelihood methods, principal components or even shrinkage approaches [4], [12], [13]. Conditional on the pervasive factors, the idiosyncratic components are either uncorrelated or affected by specific factors (such as sectors, countries, or asset classes) [9], [14]. These factors affect only a smaller number of variables (securities), resulting in a sparse covariance component. Consequently, the covariance matrix in such a model is the sum of a of low-rank component and a sparse component. This type of covariance matrices can be estimated using a principal components approach for the low-rank component and a thresholding approach for the sparse component [15].

However, the thresholding approach does not take into account the narrow factor interpretation, by which the idiosyncratic components are grouped depending on the specific factors affecting them. As a consequence of these grouped structures, the security returns may exhibit clustering, which has been observed in previous studies [16]. Some approaches include modelling these as cluster-specific factors [17]–[19], or simply clustering the original security returns and using a block structure as a shrinkage target [20]. On the other hand, a simple approach which uses industry classification groups for estimating the block-diagonal idiosyncratic component has recently been considered, where the number of variables in a single group does not exceed the number of data points (time series length) [9], [14]. This approach uses a binary mask which leaves the elements of the covariance matrix belonging to the same group equal to the sample estimate, and reduces all others to zero (resulting in a block-diagonal matrix). However, in a high-dimensional case the number of variables in a particular cluster is not guaranteed to be lower than T – therefore such a simple approach does not guarantee positive definiteness and will not be appropriate for all high-dimensional scenarios.

To estimate the block-diagonal sparse component, we propose a method which introduces shrinkage within each estimated block to ensure the positive-definiteness of the estimate, even in very high-dimensional settings [5], [21]. The clusters of the idiosyncratic component may either be known in advance (for instance, using industry classification in security return time series [4], [9]), or estimated using a clustering procedure on the residual time series from the latent factor model estimate. In such a setting, the proposed estimator can be used without prior knowledge of the variable group structures. The estimator can also absorb the pervasive factor influence in case of misspecification of the number of factors – for instance, when some latent factors are not taken into account by the low-rank component, they will be absorbed by the block-diagonal component. This makes the proposed method much more resilient to the problem of choosing the number of

latent factors than the classical thresholding estimators [15]. The proposed approach is tested in a simulation scenario and applied to a portfolio optimization problem with historical security return data. The results demonstrate the validity of the approach and suggest that the block-diagonal idiosyncratic component estimation improves the out-of-sample portfolio risk in comparison to other covariance matrix estimators.

II. Methodology

In this paper we consider the observable random vector $\mathbf{Y} = [Y_1, \ldots, Y_p]'$, with realizations $\mathbf{Y}_t = [Y_{1t}, \ldots, Y_{pt}]'$ at time t. The sample covariance matrix estimator is

$$\widehat{\mathbf{\Sigma}} = \frac{1}{T-1}\sum_{i=1}^{T}(\mathbf{Y}_t - \widehat{\mu})(\mathbf{Y}_t - \widehat{\mu})', \tag{1}$$

where $\widehat{\boldsymbol{\mu}} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{Y}_t$ is the sample mean vector and $\{\mathbf{Y}_1, ..., \mathbf{Y}_T\}$ are the observed realizations of the sample size T. When the dimension p is allowed to grow at same, or even higher rate than sample size T, the sample covariance matrix estimates are driven by noise [7]. Moreover, for $p > T$, the estimates are singular, meaning that the matrix becomes positive semidefinite and thus not invertible. Furthermore, the eigenvalues of the estimates may greatly deviate from their true values, according to the Marčenko-Pastur law [3]. To obtain numerically stable estimates and reduce estimation noise, two main approaches are commonly used: (i) model free estimators and (ii) estimators based on factor models. Some of the most commonly employed methods in both of these categories are discussed below.

A. Model free estimators

1) Linear Shrinkage: One of the most commonly used class of estimators are shrinkage estimators, especially the linear shrinkage, which can be viewed as a weighted average of the variance part and bias part of the covariance estimates, where weights should optimize the bias-variance trade-off [1]. A common form of the estimator is a linear combination of the sample covariance matrix $\widehat{\mathbf{\Sigma}}$ and the shrinkage target matrix $\tilde{\mathbf{\Sigma}}$, with sample variances $\hat{\sigma} = [\widehat{\sigma}_{11}, \ldots, \widehat{\sigma}_{pp}]'$ of the variables on the diagonal:

$$\widehat{\mathbf{\Sigma}}_s = \alpha\widehat{\mathbf{\Sigma}} + (1-\alpha)\tilde{\mathbf{\Sigma}}, \tag{2}$$

where α is a scalar parameter between 0 and 1. The method recognizes extremely high or low coefficients in sample covariance matrix and pulls them downward or upward, respectively, to compensate [5]. The optimal shrinkage intensity α, should minimize the expected value of the quadratic loss function

$$L(\alpha) = \|\alpha\widehat{\mathbf{\Sigma}} + (1-\alpha)\tilde{\mathbf{\Sigma}} - \mathbf{\Sigma}\|^2, \tag{3}$$

where $\mathbf{\Sigma}$ is the unknown population covariance. To estimate α from sample data, we follow the well-established Ledoit and Wolf [5], [21] procedure, the detailed explanation of which is beyond the scope of this paper.

2) Thresholding: Covariance thresholding estimators are permutation invariant methods encouraging sparsity with unknown zero patterns [22]. For any $\tau \geq 0$, the generalized thresholding operator is a function $s_\tau : \mathbb{R} \to \mathbb{R}$ which, for all $z \in \mathbb{R}$ satisfies the following conditions [6]:

(i) $|s_\tau(z)| \leq |z|$,
(ii) $s_\tau(z) = 0$ for $|z| \leq \tau$,
(iii) $|s_\tau(z) - z| \leq \tau$.

These conditions are satisfied by several popular thresholding functions. In this paper we consider two more advanced functions:

- adaptive lasso [23]:

$$s_\tau^{AL}(z) = \text{sign}(z)(|z| - \tau_{ij}^{a+1}|z|^{-a})_+, \tag{4}$$

- SCAD [24]:

$$s_\tau^{SCAD}(z) = \begin{cases} \text{sign}(z)(|z| - \tau_{ij})_+, & |z| \leq 2\tau_{ij} \\ \dfrac{[(a-1)z - \text{sign}(z)a\tau_{ij}]}{(a-2)}, & 2\tau_{ij} < |z| \leq a\tau_{ij} \\ z, & |z| > a\tau_{ij}. \end{cases} \tag{5}$$

The estimated covariance matrix is then given by

$$\widehat{\mathbf{\Sigma}}_\tau = \begin{cases} \widehat{\sigma}_{ii} & i = j \\ s_\tau(\widehat{\sigma}_{ij}) & i \neq j. \end{cases} \tag{6}$$

The adaptive thresholding parameter [25] is of the form

$$\tau_{ij} = \tau\sqrt{\frac{\widehat{\theta}_{ij}\log p}{T}}, \tag{7}$$

where τ is a tuning parameter and $\widehat{\theta}_{ij}$ are estimates of $\theta_{ij} = \text{Var}[(Y_i - \mu_i)(Y_j - \mu_j)]$. The value of τ can be fixed or can be data-driven and chosen through cross-validation [25]. To ensure the positive definiteness, the value τ should be chosen on the space where the estimator is satisfying the condition $\lambda_{\min}(\widehat{\mathbf{\Sigma}}_\tau) > 0$. However, when τ is sufficiently large, the estimator becomes diagonal. Therefore, the desired interval is $(\tau_{\min} + \epsilon, \tau_{\max})$ [15].

B. Factor models

Factor models assume that a small number of underlying broad factors drives observed variable dynamics. In addition, conditional on the common factors, the variables are either uncorrelated or affected by narrow factors which affect only some subsets of variables, forming the sparse idiosyncratic covariance component [1].

For the random vector $\mathbf{Y} = [Y_1, ...Y_p]'$ with mean $\boldsymbol{\mu} = \mathbf{0}$, the linear factor model is defined as

$$Y_{it} = \mathbf{b}_i'\mathbf{f}_t + e_{it}, \tag{8}$$

where Y_{it} is the realization of the i-th variable, for $i = 1, 2, ..., p$, at the time t, for $t = 1, 2, ...T$. Here $\mathbf{b}_i$ is $K \times 1$ a vector of factor loadings, $\mathbf{f}_t$ is $K \times 1$ vector of K common factors and e_{it} idiosyncratic component, assumed to be uncorrelated with common factors.

Under the model, the covariance matrix $\mathbf{\Sigma}$ has the following decomposition:

$$\mathbf{\Sigma} = \mathbf{B}\mathbf{\Sigma}_f\mathbf{B}' + \mathbf{\Sigma}_e, \tag{9}$$

where $\mathbf{B} = [\mathbf{b}_1, \ldots, \mathbf{b}_p]'$ is the $p \times K$ factor loadings matrix, $\mathbf{\Sigma}_f$ is the $K \times p$ covariance matrix of observed or estimated factors and $\mathbf{\Sigma}_e$ is the $p \times p$ covariance matrix of idiosyncratic components. Here the first term $\mathbf{L} = \mathbf{B}\mathbf{\Sigma}_f\mathbf{B}'$ is the low-rank component, while $\mathbf{S} = \mathbf{\Sigma}_e$ is the sparse component.

Some of the most influential work on estimating these types of models includes the *POET* estimtor by Fan et al. [15]. The estimator is defined as

$$\widehat{\mathbf{\Sigma}}_{POET} = \sum_{i=1}^{K} \widehat{\lambda}_i \widehat{\mathbf{b}}_i \widehat{\mathbf{b}}_i' + \widehat{\mathbf{S}}_\tau, \tag{10}$$

where $\widehat{\lambda}_1 \geq \widehat{\lambda}_2 \geq ... \geq \widehat{\lambda}_p$ are the eigenvalues and $\widehat{\mathbf{b}}_i, i = 1, ...p$ the corresponding eigenvectors of the sample covariance matrix $\widehat{\mathbf{\Sigma}}$. The main assumption is that the first K eigenvalues of $\mathbf{\Sigma}$ are spiked and grow at the same rate as number of variables p, where p is not bounded. This ensures that the estimated low-rank component is a valid approximation [15]. Theresholding is applied to the residual matrix $\widehat{\mathbf{S}}$ in order to obtain a sparse component $\widehat{\mathbf{S}}_\tau$. The thresholding constant τ is obtained via cross-validation, by a grid search over the thresholding constants and choosing the one which minimizes the out-of-sample Frobenius norm. In the original work of Fan et al [15], the *SCAD* thresholding method was used – in this paper we also consider the *adaptive lasso*, and denote the resulting estimates $\widehat{\mathbf{\Sigma}}_{AL}$ and $\widehat{\mathbf{\Sigma}}_{SCAD}$.

Motivated by the intuition that the firms within the same industries have higher correlations beyond common factors [9], [14], we assume that the sparse component is a *block-diagonal* matrix. To estimate such idiosyncratic correlation structures in a high-dimensional setting where the cluster sizes may outnumber the time series length, we propose a combined shrinkage and clustering approach in the following section.

III. Estimating the block-diagonal idiosyncratic component

The simplest approach to estimating a block-diagonal idiosyncratic covariance is based on a pre-determined clustering, for instance using industry classifications of the securities [9]. However, as mentioned before, two issues emerge. Firstly, after extraction of common factors, the observed idiosyncratic component may be driven by some other narrow factors which are not known up front. For instance, the industry classifications may not be the optimal grouping for the security returns. Secondly, this simple approach does not guarantee positive definiteness of the estimates, especially when the sizes of particular clusters may outnumber the length of the time series they are estimated from. For the purpose of resolving these issues, we propose a combined approach including a clustering procedure on the factor model residuals and a shrinkage estimator on the blocks of the estimated block-diagonal component.

A. Estimator definition

1) Clustering: To estimate the clusters when the group structure is unknown, firstly the common part of the pervasive factors is estimated via principal components, as in (10):

$$\widehat{\mathbf{L}} = \sum_{i=1}^{K} \widehat{\lambda}_i \widehat{\mathbf{b}}_i \widehat{\mathbf{b}}_i', \tag{11}$$

and the residuals are calculated:

$$\widehat{e}_{it} = Y_{it} - \widehat{\mathbf{b}}_i' \widehat{\mathbf{f}}_t, \tag{12}$$

where $\widehat{\mathbf{b}}_i, i = 1, ...K$ are the first K eigenvectors of the sample covariance matrix, and $\widehat{\mathbf{f}}_t$ can be calculated as the principal component realizations.

In our analysis, we apply the *k-means* procedure to estimate the unknown clustering from the residuals, mainly because of its simplicity and robustness – exploration of other potential algorithms is out of the scope of this paper. The approach differs from the POET estimator only in the idiosyncratic covariance estimation, and thus its complexity is fairly similar. In the clustering procedure we use a correlation-based distance measure rather than the usual Euclidean distance, due to the heteroscedasticity of the idiosyncratic components:

$$d(\widehat{\mathbf{e}}_i, \widehat{\mathbf{e}}_j) = 1 - r_{ij}, \tag{13}$$

where r_{ij} is the Pearson correlation coefficient between pairs of idiosyncratic components $\widehat{\mathbf{e}}_i$ and $\widehat{\mathbf{e}}_j$.

We denote the group membership information as a zero-one $p \times p$ indicator matrix $\mathbf{C}$ (also known as a *mask*), where the element $C_{ij} = 1$ if i and j are in the same group, for $i, j \in 1, ..., p$. If the rows and columns of $\mathbf{C}$ (and consequently, the p variables in the factor model (8)) are sorted according to their cluster membership, then $\mathbf{C}$ is a block-diagonal matrix. Without loss of generality, in the following notation we assume that the variables are sorted according to their cluster membership (this can also be done after clustering) and that $\mathbf{C}$ is block-diagonal.

Let matrix $\mathbf{C}$ contain $\mathbf{C}_1$, $\mathbf{C}_2$,...,$\mathbf{C}_M$ cluster blocks for each of the M clusters. The imposed block-diagonal covariance $\widehat{\mathbf{S}}^C$ obtained from the initial idiosyncratic estimate $\widehat{\mathbf{S}}$ is:

$$\widehat{\mathbf{S}}^C = (\widehat{S}_{ij}\mathbf{1}_{(ij)\in\mathbf{C}}) = \widehat{\mathbf{S}} \circ \mathbf{C} = \\ = \widehat{\mathbf{S}} \circ \begin{bmatrix} \mathbf{C}_1 & 0 & \ldots & 0 \\ 0 & \mathbf{C}_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \mathbf{C}_M \end{bmatrix} = \begin{bmatrix} \widehat{\mathbf{S}}^{C_1} & 0 & \ldots & 0 \\ 0 & \widehat{\mathbf{S}}^{C_2} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \widehat{\mathbf{S}}^{C_M} \end{bmatrix}, \tag{14}$$

where each sub-block is defined as $(\widehat{\mathbf{S}}^{C_m} = \widehat{S}_{ij}\mathbf{1}_{(ij)\in\mathbf{C}_m})$, $m \in 1, ...M$, and $\circ$ denotes the Hadamard element-wise product.

2) Shrinkage: Clustering itself does not ensure positive definitenes, as the cluster block size can be higher than the sample size. Therefore, to ensure positive definitenes of the estimator, we propose a shrinkage estimator which individually treats each cluster block component $\widehat{\mathbf{S}}^{C_m}$, $m \in 1, ...M$.

For each block component C_m, we search for the optimal shrinkage constant α_m, using the Ledoit and Wolf procedure [5]. The block component is then defined as

$$\widehat{\mathbf{S}}_s^{C_m} = \alpha_m \widehat{\mathbf{S}}^{C_m} + (1-\alpha_m)\tilde{\mathbf{S}}^{C_m}, \tag{15}$$

where $\tilde{\mathbf{S}}^{C_m}$ is the diagonal shrinkage target with sample variances on the diagonal. The resulting sparse component estimate is

$$\widehat{\mathbf{S}}_s^{C} = \begin{bmatrix} \widehat{\mathbf{S}}_s^{C_1} & 0 & \dots & 0 \\ 0 & \widehat{\mathbf{S}}_s^{C_2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \widehat{\mathbf{S}}_s^{C_M} \end{bmatrix}, \tag{16}$$

and the final covariance estimate (including the low-rank and sparse components):

$$\widehat{\mathbf{\Sigma}}_{CS} = \widehat{\mathbf{L}} + \widehat{\mathbf{S}}_s^{C}. \tag{17}$$

In this paper, depending on the clustering information we consider two *cluster-shrinkage (CS)* estimators:

- Industry based block-diagonal shrinkage estimator $\widehat{\mathbf{\Sigma}}_{CSI}$ (denoted *CS-I*), which uses the industry classification for obtaining cluster membership,
- Clustering based block-diagonal shrinkage estimator $\widehat{\mathbf{\Sigma}}_{CSC}$ (denoted *CS-C*), which uses the (*k*-means) clustering algorithm for obtaining cluster membership.

To be comparable with the *industry based block-diagonal* estimator and for the sake of simplicity, in this paper we use a fixed number of clusters equal to $M = 11$. The problem of estimating the number of clusters in financial data is an important issue, but is not within the scope of this paper.

IV. Experimental results

To test the proposed methods we construct a simulation scenario which follows the block-diagonal sparse component assumption, and generate random realizations for the time series on which the estimators are applied. We also use a collection of historical data to verify the validity of the approach in a real-world scenario.

A. Simulation procedure

We construct the low-rank component and the sparse component independently and randomly. To generate the low-rank component, we generate random factor loadings $\mathbf{B} = [\mathbf{b}_1, ..., \mathbf{b}_p]'$ from a uniform distribution with mean $\mathbf{0}$ and variance $\mathbf{1}$, and scale with a random $p \times 1$ vector of individual time series variances, multiplied by the the predefined percentage of explained variance equal to 0.7 (meaning that the common factors explain 70% of the total variance in the data). Finally, we calculate the low-rank component as $\mathbf{L} = \mathbf{B}'\mathbf{B}$.

Independently we construct the sparse component reflecting the cluster structure. We generate random non-integer cluster sizes lying in the predefined interval $[c_{min}, c_{max}]$ with the total sum equal to p from the uniform distribution. The cluster sizes are then rounded to integer values. We use $c_{min} = 10$ and $c_{max} = 300$ to ensure that there are no singleton clusters, and that there are large enough clusters to potentially cause invertibility issues in the sample estimate. With the obtained cluster vector, we generate the zero-one cluster representation matrix $\mathbf{C}$ as described in III. The elements themselves are generated by simulating a one-factor model within each cluster and using the sample covariance estimates obtained that way.

For the simulation we fix the dimension to $p = 800$, number of factors to $K = 5$, and simulate time series of length , $T = 200$ using the Student's t-distribution with 5 degrees of freedom, in order to replicate the heavy tailed property of scurity returns. The simulations are repeated a total of 1000 times.

B. Historical data

We consider a collection of weekly observations[1] of the MSCI World Index constituents from Jan-2005 to Sep-2020. There are $p = 1015$ stocks in the dataset and all the constituents have returns during the observed period. We also collect the Global Industrial Classification standard (GICS) sector codes for index constituents, to be used for determining the group membership in the *CS-I* estimator.

C. Measuring performance

1) Minimum variance portfolios: To evaluate our results we construct minimum variance portfolios and assess the obtained risks by using different covariance estimators. Specifically, minimum variance portfolios are used in this context since they only depend on the estimated covariance matrix, and are commonly used to measure the performance of covariance estimators [13], [26], [27].

The minimum variance portfolio $\boldsymbol{w} = [w_1, ..., w_p]'$ is obtained by solving a quadratic minimization problem:

$$\min_{\mathbf{w}} \mathbf{w}'\mathbf{\Sigma}\mathbf{w} \quad \text{s.t. } \mathbf{1}'\mathbf{w} = \mathbf{1}, \tag{18}$$

where $\mathbf{\Sigma}$ is the covariance of the security returns. In this paper we estimate the covariance matrix using the low-rank and sparse covariance estimation procedure, where the sparse component is estimated by different estimators: (i) Adaptive lasso: $\widehat{\mathbf{\Sigma}}_{AL}$; (ii) SCAD: $\widehat{\mathbf{\Sigma}}_{SCAD}$; (iii) Industry based block-diagonal shrinkage: $\widehat{\mathbf{\Sigma}}_{CSI}$; and (iv) Clustering based block-diagonal shrinkage: $\widehat{\mathbf{\Sigma}}_{CSC}$.

To evaluate portfolio performance we calculate the out-of-sample *portfolio risk* as the standard deviation of the portfolio returns (also known as volatility):

$$\sigma_p := \sqrt{\widehat{\mathbf{w}}'\mathbf{\Sigma}\widehat{\mathbf{w}}}, \tag{19}$$

where the portfolio weights $\widehat{\mathbf{w}}$ are obtained by solving (18) with different covariance estimates.

For simulation data, portfolios are optimized using the generated sample time series, and volatilities are calculated using the population covariance $\mathbf{\Sigma}$. When using historical data, the population covariance is unknown, so a backtest approach is applied with rolling time windows. At each time step,

[1]Weekly returns are used to avoid any synchronization issues in daily data due to different time zones of the exchanges.

the portfolios are constructed using the covariance estimated during the past 4 years of returns (a total of $4{\cdot}52 = 208$ weekly data points). Then, the optimal portfolios are held for the next year (52 weeks) and the portfolio volatility is calculated on this out-of-sample future holding period. The portfolios are rebalanced once a year.

2) Classification measures for non-zero elements: Since the true sparse components are known in the simulation scenario, we also measure the accuracy of identifying the true non-zero and zero elements in the population idiosyncratic component. We denote the classes of each element of the population sparse matrix with 0 if the element is zero and 1 if the element is non-zero. In our results we report accuracy, true positive rate (TPR), true negative rate (TNR) and the F1 score, some of the most commonly used classification performance measures [28]. Since the idiosyncratic covariance is sparse (and the classes are thus heavily imbalanced in favor of the 0-class), accuracy, TPR and TNR will be affected by the class imbalance – we nevertheless consider them since they reveal some specific traits of the estimators and the confusion matrices they yield. To compare the overall performance of the estimator, the F1 score is most reliable in an imbalanced data setting. For more details on the classification performance measures, see Sokolova and Lapalme [28].

3) Statistical inference: In addition to reporting the average results for the above mentioned performance measures, for the simulation study we also consider the number of experiments in which the clustering shrinkage estimator ($\widehat{\mathbf{\Sigma}}_{CSC}$) outperforms (smaller out-of-sample volatility, higher classification measures) the thresholding-based estimators ($\widehat{\mathbf{\Sigma}}_{SCAD}$ and $\widehat{\mathbf{\Sigma}}_{AL}$). Let the number of outcomes in which the *CS-C* estimator outperforms a benchmark for a given performance measure be denoted by n_+, in a total number of n experiments. For the proportion n_+/n we apply a non-parametric paired sign test, for the null-hypothesis that the probability of *CS-C* outperforming the given benchmark for a given measure is 0.5, and a one-sided alternative that the probability is greater than 0.5. Under the null hypothesis, n_+ follows a binomial distribution $B(n, 0.5)$, which is directly used to calculate the corresponding p-value. In this case, two tests are applied for each performance measure: (i) *CS-C* vs. *AL*, and (ii) *CS-C* vs. *SCAD*, in order to test whether the proposed approach yields statistically significant improvements over the benchmark methods [19].

D. Results

Firstly, we consider the simulation results, the summary of which is given in Table I. We observe significantly lower portfolio risk for the *CS-C* estimator. For both benchmark methods, the observed improvement is statistically significant (with p-values effectively equal to zero) – specifically, we report that proposed *CS-C* estimator outperformed the benchmark methods in all simulated cases.

The same pattern in the results is observed for the accuracy of classifying non-zero off-diagonal entries in the idiosyncratic covariance. However, these results are also affected by the imbalance of class labels, due to the sparsity of the idiosyncratic covariance. This is especially evident in the results for TPR and TNR. Specifically, the adaptive lasso (AL) estimator has a higher TNR (with a higher results in almost all of the simulated cases), but a much lower TPR. This is due to the fact that AL yields a much more sparse estimate. In an analogy, the most sparse estimate would be a diagonal covariance, which would have a TPR of 0 and TNR of 1. The results for the benchmark methods quantify this sparsity effect, as documented by a relatively high TNR and lower TPR. The F1 score, which is a better measure of the classification performance for imbalanced data, demonstrates these shortcomings in the benchmark methods. The proposed *CS-C* estimator outperforms them in all of the simulated cases. This means that the proposed estimator has the best balance in predicting non-zero and zero entries in the idiosyncratic covariance.

TABLE I
Portfolio risk σ_p and classification results for the considered estimators on simulation data. The p-values of the paired sign test in comparison with the proposed *CS-C* method are shown in parentheses below the results.

Estimator	σ_p	Accuracy	TPR	TNR	F1
SCAD	3.62% (< 0.01)	90.1% (< 0.01)	43.5% (< 0.01)	97.9% (< 0.01)	54.1% (< 0.01)
AL	3.82% (< 0.01)	87.6% (< 0.01)	13.6% (< 0.01)	**99.8%** (~ 1)	23.5% (< 0.01)
CS-C	**1.13%**	**96.3%**	**79.2%**	99.4%	**85.9%**

We also apply the considered estimators to the historical data, described in Section IV-B. Different numbers of latent factors are considered for all estimators, and the procedure was repeated for all considered numbers of latent factors $K = 1, 2, \ldots, 20$. The average out-of-sample portfolio risks for each of these are shown in Figure 1.

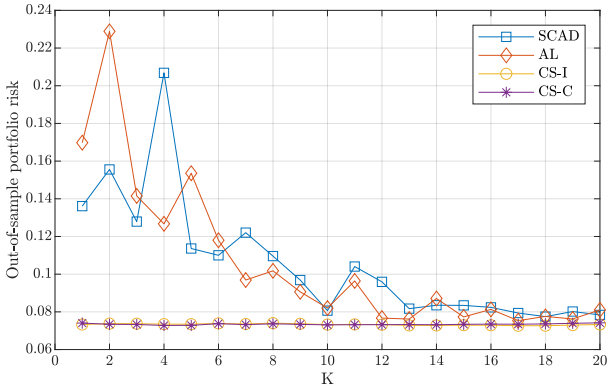

Fig. 1. Out-of-sample portfolio risk for *cluster based* estimators $\widehat{\mathbf{\Sigma}}_{CSC}$ and $\widehat{\mathbf{\Sigma}}_{CSI}$, in comparison with the *thresholding based* estimators $\widehat{\mathbf{\Sigma}}_{SCAD}$ and $\widehat{\mathbf{\Sigma}}_{AL}$, all obtained on historical market data and displayed for different numbers of latent factors K.

Evidently, the volatilities are lower for both cluster-shrinkage estimators than the considered thresholding bench-

marks. Moreover, the results demonstrate the remarkable stability of the out-of-sample volatilities for different estimates of K, as opposed to those of the *AL* and *SCAD* estimators, which deteriorate when the number of latent factors is underestimated. There seems to be no meaningful distinction between the *cluster based* estimators *CS-I* and *CS-C* as both perform similarly well in all out-of-sample historical scenarios. This means that both the industry classification and the unsupervised *k*-means clustering capture relevant security groups in the observed data.

V. Conclusion

In this paper the performance of high-dimensional covariance matrix estimators with low-rank and sparse components are considered. The low-rank component is estimated using principal components, while special focus is given to the different sparse component estimation methods. A new approach based on the combination of clustering and shrinkage is proposed, with a block-diagonal structure of the idiosyncratic covariance component. Depending on the approach for determining the groups, we derive two estimators – the *CS-I* method based on known industry classification and the *CS-C* method based on a k-means clustering procedure. A vital part of the estimator is the addition of a shrinkage procedure on each of the blocks in the idiosyncratic component, which enables scaling such estimators to high-dimensional scenarios in which the number of variables in some clusters may exceed the number of data samples. The proposed approach was tested in a simulation scenario, in which the results demonstrate the ability of the block-diagonal estimators to identify the true sparsity patterns in the idiosyncratic components, and reduce the out-of-sample portfolio risk. Moreover, an empirical backtesting study was performed on historical security returns, where the proposed estimators were shown to exhibit lower out-of-sample portfolio volatilities than the considered benchmark thresholding methods. Moreover, the proposed approach demonstrated a remarkable degree of stability and resilience to the potential misspecification of the number of latent factors.

Acknowledgment

This work was supported in part by the Croatian Science Foundation under Project 5241, and in part by the European Regional Development Fund under Grant KK.01.1.1.01.0009 (DATACROSS).

References

[1] Mohsen Pourahmadi, *High-Dimensional Covariance Estimation*, ser. Wiley Series in Probability and Statistics. Hoboken, NJ, USA: John Wiley & Sons, Inc., jun 2013.

[2] J. Bai and S. Shi, "Estimating high dimensional covariance matrices and its applications," pp. 199–215, nov 2011.

[3] J. Bun, J. P. Bouchaud, and M. Potters, "Cleaning large correlation matrices: Tools from Random Matrix Theory," *Physics Reports*, vol. 666, pp. 1–109, 2017.

[4] J. Fan, Y. Liao, and H. Liu, "An overview of the estimation of large covariance and precision matrices," *The Econometrics Journal*, vol. 19, no. 1, pp. C1–C32, feb 2016.

[5] O. Ledoit and M. Wolf, "A well-conditioned estimator for large-dimensional covariance matrices," *Journal of Multivariate Analysis*, 2004.

[6] A. J. Rothman, E. Levina, and J. Zhu, "Generalized thresholding of large covariance matrices," *Journal of the American Statistical Association*, vol. 104, no. 485, pp. 177–186, mar 2009.

[7] J. Bun, J.-P. Bouchaud, and M. Potters, "Cleaning correlation matrices," *Risk Magazine*, vol. 2015, no. April, 2015.

[8] J. Fan, Y. Liao, and M. Mincheva, "High-dimensional covariance matrix estimation in approximate factor models," *The Annals of Statistics*, vol. 39, no. 6, pp. 3320–3356, dec 2011.

[9] Y. Aït-Sahalia and D. Xiu, "Using principal component analysis to estimate a high dimensional factor model with high-frequency data," *Journal of Econometrics*, vol. 201, no. 2, pp. 384–399, dec 2017.

[10] L. R. Goldberg, A. Papanicolaou, and A. Shkolnik, "The Dispersion Bias," *SIAM Journal on Financial Mathematics*, vol. 13, no. 2, pp. 521–550, jun 2022.

[11] G. Connor, "The Three Types of Factor Models: A Comparison of Their Explanatory Power," *Financial Analysts Journal*, vol. 51, no. 3, pp. 42–46, may 1995.

[12] J. Bai and K. Li, "Statistical analysis of factor models of high dimension," *The Annals of Statistics*, vol. 40, no. 1, pp. 436–465, feb 2012.

[13] L. R. Goldberg, A. Papanicolaou, A. Shkolnik, and S. Ulucam, "Better Betas," *The Journal of Portfolio Management*, vol. 47, no. 1, pp. 119–136, oct 2020.

[14] J. Fan, A. Furger, and D. Xiu, "Incorporating Global Industrial Classification Standard Into Portfolio Allocation: A Simple Factor-Based Large Covariance Matrix Estimator With High-Frequency Data," *Journal of Business and Economic Statistics*, vol. 34, no. 4, pp. 489–503, oct 2016.

[15] J. Fan, Y. Liao, and M. Mincheva, "Large covariance estimation by thresholding principal orthogonal complements," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 75, no. 4, pp. 603–680, sep 2013.

[16] M. Tumminello, F. Lillo, and R. N. Mantegna, "Correlation, hierarchies, and networks in financial markets," *Journal of Economic Behavior and Organization*, 2010.

[17] T. Ando and J. Bai, "Panel Data Models with Grouped Factor Structure Under Unknown Group Membership," *Journal of Applied Econometrics*, vol. 31, no. 1, pp. 163–191, jan 2016.

[18] ——, "Clustering Huge Number of Financial Time Series: A Panel Data Approach With High-Dimensional Predictors and Factor Structures," *Journal of the American Statistical Association*, vol. 112, no. 519, pp. 1182–1198, 2017.

[19] S. Begušić and Z. Kostanjčar, "Cluster-Specific Latent Factor Estimation in High-Dimensional Financial Time Series," *IEEE Access*, vol. 8, pp. 164 365–164 379, sep 2020.

[20] ——, "Cluster-Based Shrinkage of Correlation Matrices for Portfolio Optimization," in *2019 11th International Symposium on Image and Signal Processing and Analysis (ISPA)*. IEEE, sep 2019, pp. 301–305.

[21] O. Ledoit and M. Wolf, "Honey, I shrunk the sample covariance matrix," pp. 110–119+7, jul 2004.

[22] P. J. Bickel and E. Levina, "Regularized estimation of large covariance matrices," *Annals of Statistics*, 2008.

[23] H. Zou, "The adaptive lasso and its oracle properties," *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1418–1429, dec 2006.

[24] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American Statistical Association*, vol. 96, no. 456, pp. 1348–1360, dec 2001.

[25] T. Cai and W. Liu, "Adaptive Thresholding for Sparse Covariance Matrix Estimation," *Journal of the American Statistical Association*, 2011.

[26] E. Pantaleo, M. Tumminello, F. Lillo, and R. N. Mantegna, "When do improved covariance matrix estimators enhance portfolio optimization? An empirical comparative study of nine estimators," *Quantitative Finance*, vol. 11, no. 7, pp. 1067–1080, jul 2011.

[27] Y. G. Choi, J. Lim, and S. Choi, "High-dimensional Markowitz portfolio optimization problem: empirical comparison of covariance matrix estimators," *Journal of Statistical Computation and Simulation*, vol. 89, no. 7, pp. 1278–1300, may 2019.

[28] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing & Management*, vol. 45, no. 4, pp. 427–437, jul 2009.