JUSTC

Article

http://justc.ustc.edu.cn

Received: January 15, 2025; Accepted: February 25, 2025

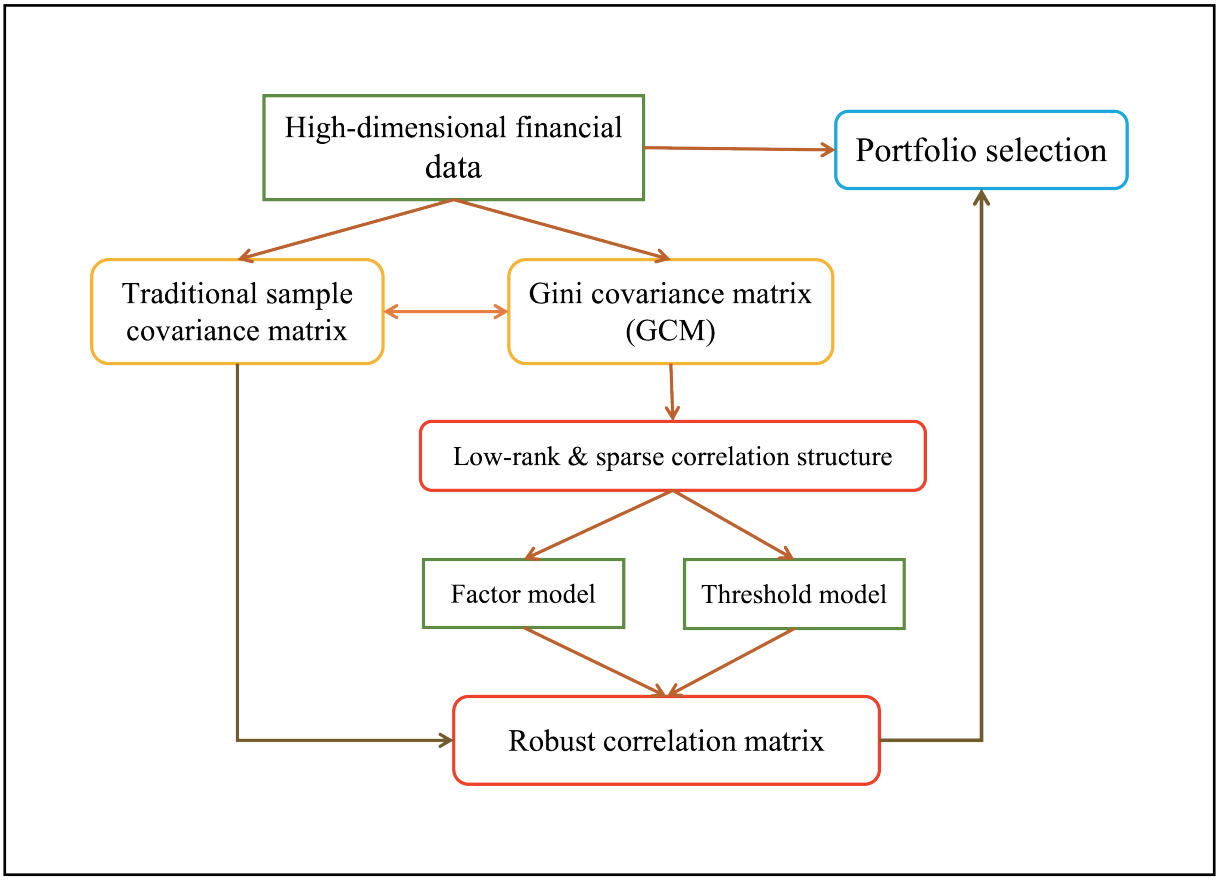
Robust Gini covariance matrix estimation for portfolio selection based on a factor model

Yongda Zhu, and Lei Shu ✉

Department of Statistics and Finance, School of Management, University of Science and Technology of China, Hefei 230026, China

✉Correspondence: Lei Shu, E-mail: sl2018@ustc.edu.cn

Graphical abstract



A robust portfolio selection method based on a low-rank and sparse structured Gini covariance matrix.

Public summary

- This article proposes a Gini covariance matrix estimation method based on factor models, which demonstrates greater robustness than traditional sample covariance matrices when handling high-dimensional financial data.
- By integrating factor models to capture low-rank structures and employing thresholding rules to achieve the final estimation, the study establishes the consistency of its estimator.
- Through extensive simulation experiments and empirical analysis of S&P 100 stock data, this study illustrates the practical application value of the proposed method in constructing high-performance portfolios.

Citation: Zhu Y D, Shu L. Robust Gini covariance matrix estimation for portfolio selection based on a factor model. *JUSTC*, **2025**, 55(8): 0806. CSTR: 32290.14.JUSTC-2025-0003. DOI: 10.52396/JUSTC-2025-0003

JUSTC

Article

http://justc.ustc.edu.cn

Received: January 15, 2025; Accepted: February 25, 2025

Robust Gini covariance matrix estimation for portfolio selection based on a factor model

Yongda Zhu, and Lei Shu ✉

Department of Statistics and Finance, School of Management, University of Science and Technology of China, Hefei 230026, China

✉Correspondence: Lei Shu, E-mail: sl2018@ustc.edu.cn

Cite This: *JUSTC*, **2025**, 55(8): 0806 (8pp) Read Online SI Supporting Information

Abstract: Portfolio theory has been extensively studied and applied in finance. To determine the optimal portfolio weight under the global minimum variance strategy, it is necessary to estimate both the covariance matrix and its inverse. However, the high dimensionality and heavy-tailed nature of financial data pose significant challenges to this estimation. In this study, we propose a method to estimate the Gini covariance matrix by introducing a low-rank and sparse correlation structure, as an alternative to the traditional sample covariance matrix. Our approach employs a factor model to capture the low-rank structure, combined with thresholding rules to achieve the final estimation. We demonstrate the consistency of our estimators and validate our approach through simulation experiments and empirical portfolio analyses. Simulation results show that our method is highly applicable across a variety of distributional scenarios. Furthermore, empirical portfolio analysis indicates that our method can construct portfolios with superior performance.

Keywords: elliptical distribution; factor model; Gini covariance matrix; portfolio selection

CLC number: F064.1 **Document code:** A

2020 Mathematics Subject Classification: 91G10

1 Introduction

Portfolio theory, encompassing portfolio selection and asset allocation, has become a cornerstone of modern financial analysis with extensive practical applications. The foundation of modern portfolio theory was established by Markowitz[1], who pioneered the integration of mathematical statistics into portfolio analysis. This seminal work has subsequently led to the development of numerous portfolio selection methodologies, including the minimum variance approach[1], equal weight strategy[2], and maximum diversification technique[3]. Among these, the minimum variance portfolio, which focuses on minimizing volatility risk, represents a fundamental concept in portfolio theory and serves as the primary focus of our investigation. The minimum variance portfolio can be mathematically formulated via the covariance matrix of stock returns ($\boldsymbol{\Sigma}$) as follows:

$$\boldsymbol{w}_{\text{opt}} = \arg\min_{\mathbf{1}^\top \boldsymbol{w}=1} \boldsymbol{w}^\top \boldsymbol{\Sigma} \boldsymbol{w} = \frac{\boldsymbol{\Sigma}^{-1}\mathbf{1}}{\mathbf{1}^\top \boldsymbol{\Sigma}^{-1}\mathbf{1}}, \tag{1}$$

where $\mathbf{1}$ represents a vector of ones with a dimension of $N \times 1$ and $\boldsymbol{w}_{\text{opt}}$ denotes the optimal weights of the stocks. Thus, accurately estimating the covariance matrix and its inverse is vital for deriving optimal portfolio weights as Eq. (1).

Traditional portfolio selection methods typically rely on the sample covariance matrix estimator. However, this approach faces two fundamental limitations in modern financial applications: ① sensitivity to heavy-tailed return distributions and ② dimensionality challenges in high-frequency trading environments. Our proposed solution using the Gini covariance matrix (GCM) addresses both limitations through robust statistical estimation and structural decomposition. The conventional assumption of Gaussian returns proves inadequate for modeling financial data, where heavy-tailed distributions and extreme events occur frequently[4,5]. While researchers[6–8] have adopted the elliptical distribution family (encompassing both normal and Student's t-distributions) as a more flexible framework, practical implementation remains a challenge due to outlier sensitivity. The GCM, introduced by Sang et al.[9] and Dang et al.[10], resolves this through its rank-based formulation. This transformation achieves outlier resistance by down-weighting extreme observations through its bounded influence function while maintaining validity across all elliptical distributions.

When the dimensionality N (number of assets) approaches the sample size T (historical returns), traditional covariance estimation becomes ill-posed: The number of parameters ($O(N^2)$) exceeds the number of data points ($O(NT)$); and although there is no singularity theoretically, the singularity occurs when $N > T$. Our solution decomposes the GCM within a factor model structure: "Low-rank" + "Sparse", where latent common factors ("Low-rank") capture systematic risk through rank constraints, whereas idiosyncratic components ("Sparse") admit sparse estimation. This dual structure enables consistent estimation even when $N/T \to c \in (0, \infty)$, overcoming the curse of dimensionality through three mechanisms: The factor structure reduces the effective parameter space, the sparsity constraint stabilizes the idiosyncratic

CSTR: 32290.14.JUSTC-2025-0003
DOI: 10.52396/JUSTC-2025-0003
JUSTC, **2025**, 55(8): 0806

estimates, and matrix invertibility is maintained to optimize the portfolio.

Factor models, which are widely utilized in economics and finance[11–13], describe the latent common component of a low-rank structure. The latent common return x_i in these models is expressed as:

$$x_i = l_{i1}f_1 + \cdots + l_{im}f_m, \quad i = 1, \cdots, N, \tag{2}$$

where $f_1, \cdots, f_m$ denote the excess returns of m latent factors, with l_{ij} $(i = 1, \cdots, N,\ j = 1, \cdots, m)$ as the unknown factor loadings. The factor model (2) significantly reduces the GCM estimation parameters if a few factors effectively capture cross-sectional risk. Additionally, under the assumption of a sparse structure for the GCM of the idiosyncratic errors, the removal of common components substantially reduces parameter dimensionality. Consequently, this study estimates the GCM and analyzes portfolios in terms of a low-rank factor structure and a sparse idiosyncratic error structure, aligning with previous studies such as Refs. [14–17].

To underscore the practical value of our methods, we initiate an extensive simulation study, evaluating the empirical performance of the methods across a range of scenarios. This simulation study provides valuable insights into the effectiveness of the GCM in contrast to the sample covariance matrix and the sample Kendall τ matrix within a factor structure framework. Next, in our empirical analysis, we turn to S&P 100 stock data spanning from 2020 to 2022 to examine the performance of portfolios constructed via the Gini method compared with the PCA method and the Kendall τ method. Our findings consistently demonstrate that the Gini method yields superior portfolio performance, reinforcing the merits of our approach. We also establish the consistency property of factor estimation in the elliptical distribution framework under some mild assumptions, providing theoretical guarantees of the methodology.

The rest of this paper is organized as follows. Section 2 introduces the portfolio optimization problem on the basis of global minimum variance, for which we propose a multi-factor Gini covariance strategy and present the corresponding algorithm. In Section 3, we outline the theoretical guarantees of the strategy and present the consistency of the latent common component estimates under some basic assumptions. Section 4 presents the results of our simulation study. We then delve into an analysis of a real financial dataset, exploring optimal portfolio allocation and portfolio returns in Section 5. Finally, Section 6 provides our concluding remarks. Supporting information presents several lemmas and proofs of Proposition 1 and Theorem 1.

Notations. For any vector $\boldsymbol{\mu} = (\mu_1, \cdots, \mu_n)^\top \in \mathbb{R}^n$, let $\|\boldsymbol{\mu}\| = \left(\sum_{i=1}^{n} \mu_i^2\right)^{1/2}$, $\|\boldsymbol{\mu}\|_\infty = \max_i |\mu_i|$. For a matrix $\boldsymbol{M}$, let M_{ij} be the entry ij of $\boldsymbol{M}$, $\boldsymbol{M}^\top$ the transpose of $\boldsymbol{M}$, diag($\boldsymbol{M}$) a vector consisting of the diagonal elements of $\boldsymbol{M}$ if $\boldsymbol{M}$ is a square matrix. Let $\|\boldsymbol{M}\|_\infty = \max\left\{\left|M_{ij}\right|\right\}$ be the matrix entry-wise maximum value and $\|\boldsymbol{M}\|_F = (\sum M_{ij}^2)^{1/2}$ be the Frobenius norm of $\boldsymbol{M}$. Denote $\lambda_j(\boldsymbol{M})$ as the jth largest eigenvalue of a non-negative definite matrix $\boldsymbol{M}$.

2 Methodology

2.1 Background on portfolio optimization and elliptical distribution

The minimum variance portfolio is an important concept in modern portfolio theory. Consider the following estimation problem for a global minimum variance portfolio and assume that there are no short-selling restrictions. The formulation of this problem is as follows:

$$\min_{\boldsymbol{w}} \boldsymbol{w}^\top \boldsymbol{\Sigma} \boldsymbol{w} \quad \text{subject to } \boldsymbol{w}^\top \mathbf{1} = 1, \tag{3}$$

where $\boldsymbol{\Sigma} \in \mathbb{R}^{N\times N}$ denotes the covariance matrix of the N stock returns. The primary objective of this optimization problem (3) is to determine an asset allocation strategy that minimizes the overall variance, subject to the constraint that the total weight of the portfolio is fixed at 1. This problem can be solved by Eq. (1). In practical applications, it is common to replace the unknown $\boldsymbol{\Sigma}$ with the estimate $\widehat{\boldsymbol{\Sigma}}$, yielding a feasible portfolio.

$$\widehat{\boldsymbol{w}}_{\text{opt}} = \frac{\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{1}}{\mathbf{1}^\top \widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{1}}.$$

Effective covariance matrix estimation is thus crucial in solving the portfolio problem. However, as previously noted, directly estimating the covariance matrix poses a significant challenge. In the following section, we explore a multi-factor strategy in which the GCM is used to address this issue.

Moreover, the elliptical distribution is crucial in portfolio theory, as it is a generalization of the normal distribution, including the heavy-tailed distributions common in finance, such as the t distribution. The elliptical distribution shares several properties with the normal distribution. Specifically, within the elliptical distribution, both the marginal and conditional distributions are elliptical, which is similar to the normal distribution. The definition of the elliptical distribution is as follows:

Definition 1. (Elliptical distribution) Let $\boldsymbol{\mu} \in \mathbb{R}^p$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$ with $\text{rank}(\boldsymbol{\Sigma}) = q \le p$. A p-dimensional random vector $\boldsymbol{Y}$ is said to have an elliptical distribution, denoted as $\boldsymbol{Y} \sim \text{ED}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \zeta)$, if it can be represented stochastically as:

$$\boldsymbol{Y} \overset{\text{d}}{=} \boldsymbol{\mu} + \zeta \boldsymbol{A}\boldsymbol{U},$$

where $\boldsymbol{U}$ is a random vector uniformly sampled from the unit sphere $\mathcal{S}^{q-1}$ in $\mathbb{R}^q$ and $\zeta \ge 0$ is a non-negative random variable independent of $\boldsymbol{U}$. Additionally, $\boldsymbol{A} \in \mathbb{R}^{p\times q}$ is a deterministic matrix satisfying $\boldsymbol{A}\boldsymbol{A}^\top = \boldsymbol{\Sigma}$, where $\boldsymbol{\Sigma}$ is known as the scatter matrix.

The elliptical distribution allows portfolios to be characterized entirely by location and scale when asset returns follow these distributions. This means that portfolios with the same location and scale parameters have identical return distributions. This property is consistent across portfolio analysis aspects, including mutual fund separation theorems and the capital asset pricing model, as demonstrated in Refs. [6, 8, 18].

2.2 Gini covariance matrix estimation

Denote by T the number of historical returns for each stock

CSTR: 32290.14.JUSTC-2025-0003
DOI: 10.52396/JUSTC-2025-0003
JUSTC, **2025**, 55(8): 0806

and by N the number of stocks, with observations $\boldsymbol{y}_t = (y_{1t},\cdots,y_{Nt})^\top$, $1 \le t \le T$. As mentioned earlier, traditional sample covariance matrix estimation may not be robust in an elliptical distribution framework, particularly with heavy-tailed effects and outliers. Therefore, we suggest using the Gini covariance matrix (GCM) as a robust alternative to the covariance matrix. Our initial step is to use the observations $\boldsymbol{Y} = (\boldsymbol{y}_1,\cdots,\boldsymbol{y}_T)$ to estimate the GCM, which was pioneered by Dang et al.[10] and is a spatial rank-based measure of covariance among random vectors, as shown in Definition 2.

Definition 2. (Gini covariance matrix (GCM)) Given a random vector $\boldsymbol{Y} \in \mathbb{R}^p$, the GCM of $\boldsymbol{Y}$ was defined as

$$\boldsymbol{G} = 2\mathbb{E}[\boldsymbol{Y}\boldsymbol{r}^\top(\boldsymbol{Y})],$$

where $\boldsymbol{r}(\boldsymbol{Y})$ represents the spatial rank function defined as $\boldsymbol{r}(\boldsymbol{Y}) = \mathbb{E}_{\widetilde{\boldsymbol{Y}}}[\boldsymbol{s}(\boldsymbol{Y}-\widetilde{\boldsymbol{Y}})]$. $\widetilde{\boldsymbol{Y}}$ is an independent copy of $\boldsymbol{Y}$ and $\boldsymbol{s}(\boldsymbol{y}) = \boldsymbol{y}/\|\boldsymbol{y}\|$ denotes the spatial sign function. For straightforward algebraic operations, we have

$$\boldsymbol{G} = 2\mathbb{E}\boldsymbol{Y}\mathbb{E}[\boldsymbol{s}^\top(\boldsymbol{Y}-\widetilde{\boldsymbol{Y}}) \mid \boldsymbol{Y}] = 2\mathbb{E}[\boldsymbol{Y}\boldsymbol{s}^\top(\boldsymbol{Y}-\widetilde{\boldsymbol{Y}})] = \mathbb{E}\frac{(\boldsymbol{Y}-\widetilde{\boldsymbol{Y}})(\boldsymbol{Y}-\widetilde{\boldsymbol{Y}})^\top}{\|\boldsymbol{Y}-\widetilde{\boldsymbol{Y}}\|}.$$

Examining the GCM definition reveals its incorporation of a rank function, which is a characteristic that enhances its robustness, particularly against outliers. In essence, GCM exhibits greater resilience in the face of individual data points that exhibit substantial deviations from the majority of the data. This attribute renders it a valuable choice for exploring cross-sectional correlations, especially in scenarios characterized by heavy-tailed distributions. The estimation of $\boldsymbol{G}$ can be achieved through a second-order U-statistic:

$$\widetilde{\boldsymbol{G}}_y = \frac{2}{T(T-1)}\sum_{t<t^*}\frac{(\boldsymbol{y}_t-\boldsymbol{y}_{t^*})(\boldsymbol{y}_t-\boldsymbol{y}_{t^*})^\top}{\|\boldsymbol{y}_t-\boldsymbol{y}_{t^*}\|}. \tag{4}$$

Furthermore, to improve the accuracy of the estimation in the high-dimensional case, we consider a high-dimensional latent component model for the dataset $\{y_{it}, 1 \le i \le N, 1 \le t \le T\}$,

$$y_{it} = x_{it} + \epsilon_{it},\ i \le N,\ t \le T,\ \text{ or }\ \boldsymbol{y}_t = \boldsymbol{x}_t + \boldsymbol{\epsilon}_t \ \text{ in vector form,} \tag{5}$$

where $\boldsymbol{y}_t = (y_{1t},\cdots,y_{Nt})^\top$ represents the observed stock returns at time t, $\boldsymbol{x}_t = (x_{1t},\cdots,x_{Nt})^\top$ is the common component of $\boldsymbol{y}_t$, and $\boldsymbol{\epsilon}_t = (\epsilon_{1t},\cdots,\epsilon_{Nt})^\top$ represents the idiosyncratic errors. We maintain the assumption that $\mathbb{E}(\boldsymbol{\epsilon}_t|\boldsymbol{x}_t) = 0$, and that $\boldsymbol{\epsilon}_t$ and $\boldsymbol{x}_t$ are independent.

Under the model (5), we can see that $\boldsymbol{G}_y$, which represents the correlation of returns between different assets, is calculated as:

$$\boldsymbol{G}_y = \mathbb{E}\frac{(\boldsymbol{y}_t-\widetilde{\boldsymbol{y}}_t)(\boldsymbol{y}_t-\widetilde{\boldsymbol{y}}_t)^\top}{\|\boldsymbol{y}_t-\widetilde{\boldsymbol{y}}_t\|} = \mathbb{E}\frac{(\boldsymbol{x}_t-\widetilde{\boldsymbol{x}}_t)(\boldsymbol{x}_t-\widetilde{\boldsymbol{x}}_t)^\top}{\|\boldsymbol{y}_t-\widetilde{\boldsymbol{y}}_t\|} + \mathbb{E}\frac{(\boldsymbol{\epsilon}_t-\widetilde{\boldsymbol{\epsilon}}_t)(\boldsymbol{\epsilon}_t-\widetilde{\boldsymbol{\epsilon}}_t)^\top}{\|\boldsymbol{y}_t-\widetilde{\boldsymbol{y}}_t\|} = \boldsymbol{G}_x + \boldsymbol{G}_\epsilon,$$

where $\boldsymbol{G}_x$ and $\boldsymbol{G}_\epsilon$ denote the correlations embedded in the latent common component and the idiosyncratic error, respectively. Within the framework (5), the return GCM is divided into two separate unknown components. The combined estimates of these components may serve as an alternative to directly estimating the return GCM. However, this division still poses the challenge of dimensional catastrophe in high-dimensional contexts. To mitigate the issue of excessive parameters, we propose imposing a structure on $\boldsymbol{G}_x$ and $\boldsymbol{G}_\epsilon$. Specifically, the latent common components are assumed to conform to a low-rank structure, essentially a factor model, as follows:

$$x_{it} = \boldsymbol{l}_i^\top \boldsymbol{f}_t,\quad i \le N,\ t \le T,\ \text{ or }\ \boldsymbol{x}_t = \boldsymbol{L}\boldsymbol{f}_t, \tag{6}$$

where $\boldsymbol{f}_t \in \mathbb{R}^m$ (m is the number of the factors) is the latent factor, $\boldsymbol{l}_i \in \mathbb{R}^m$ represents the factor loading of the ith stock and $\boldsymbol{L} = (\boldsymbol{l}_1,\cdots,\boldsymbol{l}_N)^\top$ is the factor loading matrix. Therefore, we will split into two components, estimating $\boldsymbol{G}_x$ and $\boldsymbol{G}_\epsilon$ separately to obtain the final $\boldsymbol{G}_y$ estimate.

Component 1: Estimation of $\boldsymbol{G}_x$. In practice, we can use the estimation of factors and factor loadings to derive the common components $\widehat{\boldsymbol{x}}_t = \widehat{\boldsymbol{L}}\widehat{\boldsymbol{f}}_t$, which gives us an estimate of $\boldsymbol{G}_x$ as

$$\widehat{\boldsymbol{G}}_x = \frac{2}{T(T-1)}\sum_{t<t^*}\frac{(\widehat{\boldsymbol{x}}_t-\widehat{\boldsymbol{x}}_{t^*})(\widehat{\boldsymbol{x}}_t-\widehat{\boldsymbol{x}}_{t^*})^\top}{\|\boldsymbol{y}_t-\boldsymbol{y}_{t^*}\|}. \tag{7}$$

Due to the complexity of simultaneously estimating $\boldsymbol{l}_i$ and $\boldsymbol{f}_t$, as described in Remark 1, we adopt a sequential estimation approach.

Initially, we use an eigen decomposition-based approach to estimate the factor loadings. Following the GCM estimation $\widetilde{\boldsymbol{G}}_y$ in (4), we identify its leading m eigenvectors, $\widehat{\boldsymbol{\xi}}_1,\cdots,\widehat{\boldsymbol{\xi}}_m$, to form the matrix $\widehat{\boldsymbol{\Gamma}} = (\widehat{\boldsymbol{\xi}}_1,\cdots,\widehat{\boldsymbol{\xi}}_m)$. We use the scaled matrix $\widehat{\boldsymbol{L}} = (\widehat{\boldsymbol{l}}_1,\cdots,\widehat{\boldsymbol{l}}_N)^\top = \sqrt{N}\widehat{\boldsymbol{\Gamma}}$ as the estimated factor loading matrix $\boldsymbol{L}$. Remark 2 elaborates that this approach is theoretically sound within the elliptical distribution framework.

A least squares regression approach is subsequently used to derive the corresponding latent factors. After obtaining $\widehat{\boldsymbol{L}}$, we estimate $\boldsymbol{f}_t$ by regressing $\boldsymbol{y}_t$ on $\widehat{\boldsymbol{L}}$. The estimation is formulated as:

$$\widehat{\boldsymbol{f}}_t = \underset{\boldsymbol{\beta}_t \in R^m}{\arg\min}\sum_{i=1}^{N}(y_{it}-\widehat{\boldsymbol{l}}_i^\top\boldsymbol{\beta}_t)^2,\quad \text{for each } t = 1,\cdots,T. \tag{8}$$

Let $\boldsymbol{F} = (\boldsymbol{f}_1,\cdots,\boldsymbol{f}_T)^\top$. Then, the explicit solution for $\widehat{\boldsymbol{F}}$ is $\widehat{\boldsymbol{F}} = (\widehat{\boldsymbol{L}}^\top\widehat{\boldsymbol{L}})^{-1}\widehat{\boldsymbol{L}}^\top\boldsymbol{Y}$. Throughout this study, we treat m as a predetermined and fixed parameter. For cases where m requires estimation, the methodologies introduced by Refs. [19, 20] provide reliable estimation frameworks.

Remark 1. The factor loading matrix $\boldsymbol{L}$ and the factor matrix $\boldsymbol{F}$ remain inherently unobservable, thereby rendering them nonseparately identifiable. In fact, for any arbitrary invertible $m \times m$ matrix $\boldsymbol{H}$, alternative representations can be easily constructed as $\boldsymbol{L}^* = \boldsymbol{L}\boldsymbol{H}^\top$ and $\boldsymbol{F}^* = \boldsymbol{F}\boldsymbol{H}^{-1}$, while still satisfying the equality $\boldsymbol{L}^*\boldsymbol{F}^{*\top} = \boldsymbol{L}\boldsymbol{F}^\top$. To ensure identifiability in light of this, we impose specific constraints. Specifically, we stipulate that:

$$\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\boldsymbol{l}_i\boldsymbol{l}_j^\top = \boldsymbol{I}_m.$$

CSTR: 32290.14.JUSTC-2025-0003
DOI: 10.52396/JUSTC-2025-0003
JUSTC, **2025**, 55(8): 0806

This is the rationale behind our utilization of the eigenvectors associated with the first m eigenvalues in the formation of a matrix, which serves as an estimation of the factor loadings.

Remark 2. The procedure for estimating factor loadings utilizing the GCM enjoys theoretical guarantees. The assertion is supported by Theorem 2.1 in Ref. [10], which establishes a fundamental property concerning the eigenvector space. If $\boldsymbol{Y}$ is elliptically distributed with the first moment and the scatter parameter $\boldsymbol{\Sigma}$ has the spectral decomposition $\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^{\top}$, then we have $\boldsymbol{G}=\boldsymbol{V}\boldsymbol{\Lambda}_g\boldsymbol{V}^{\top}$ with

$$\boldsymbol{\Lambda}_g=\operatorname{diag}(\lambda_{g,1},\cdots,\lambda_{g,q})=c\cdot\mathbb{E}\left[\frac{\boldsymbol{\Lambda}^{1/2}\boldsymbol{U}\boldsymbol{U}^{T}\boldsymbol{\Lambda}^{1/2}}{\sqrt{\boldsymbol{U}^{\top}\boldsymbol{\Lambda}\boldsymbol{U}}}\right].$$

Consequently, we obtain the expressions for $\lambda_{g,i}$ as follows:

$$\lambda_{g,i}=c\cdot\mathbb{E}[\frac{\lambda_i u_i^2}{\sqrt{\sum_{j=1}^{q}\lambda_j u_j^2}}].$$

Here, $\boldsymbol{U}=(u_1,\cdots,u_q)^{\top}$ is uniformly distributed on the unit sphere and λ_i represents the eigenvalues of $\boldsymbol{\Sigma}$, where $\operatorname{rank}(\boldsymbol{\Sigma})=q$ and c is a constant associated with the distribution.

The foremost revelation derived from this analysis attests to the alignment between the matrices $\boldsymbol{\Sigma}$ and $\boldsymbol{G}$, both of which share a congruent orthogonal matrix $\boldsymbol{V}$. In essence, this assertion signifies that the eigenvectors are consistently projected onto the identical subspace, a fundamental tenet underpinning the factor model. This establishes the theoretical groundwork for ensuring consistency in both the estimated factors and their corresponding loadings.

Component 2: Estimation of $\boldsymbol{G}_{\epsilon}$. To address idiosyncratic errors, we apply a specific regularization structure to the $\boldsymbol{G}_{\epsilon}$ matrix, utilizing sparsity constraints, diagonal matrix settings, and related methods. Assuming a sparse structure for $\boldsymbol{G}_{\epsilon}$, we employ a thresholding technique to estimate this matrix as follows:

$$\begin{aligned}&\widehat{\boldsymbol{G}}_{\epsilon}=\operatorname{Threshold}(\boldsymbol{M},\tau),\\&\boldsymbol{M}=\frac{2}{T(T-1)}\sum_{t<t^*}\frac{(\widehat{\boldsymbol{\epsilon}}_t-\widehat{\boldsymbol{\epsilon}}_{t^*})(\widehat{\boldsymbol{\epsilon}}_t-\widehat{\boldsymbol{\epsilon}}_{t^*})^{\top}}{\|\boldsymbol{y}_t-\boldsymbol{y}_{t^*}\|},\end{aligned}\tag{9}$$

where $\operatorname{Threshold}(\boldsymbol{M},\tau)$ is a truncation operator for a matrix $\boldsymbol{M}$ and a constant τ to obtain the sparse structure. It is specifically defined as

$$[\operatorname{Threshold}(\boldsymbol{M},\tau)]_{ij}=\begin{cases}M_{ij}, & \text{if } |M_{ij}|\geq\tau;\\ 0, & \text{otherwise}.\end{cases}$$

This definition was put forward by Bickel and Levina[21] for the sparse estimation of the idiosyncratic error covariance matrix. The approach involves compressing the elements of the matrix with absolute values less than a certain threshold τ to 0.

Finally, using the estimated common component covariance matrix $\widehat{\boldsymbol{G}}_x$ and the idiosyncratic error covariance matrix $\widehat{\boldsymbol{G}}_{\epsilon}$, an estimate of $\boldsymbol{G}_y$ was derived as $\widehat{\boldsymbol{G}}_y=\widehat{\boldsymbol{G}}_x+\widehat{\boldsymbol{G}}_{\epsilon}$. Algorithm 1 summarizes the entire process of solving for optimal portfolio weights as described above.

Algorithm 1: Portfolio optimization.

Input: The sample return data $\boldsymbol{Y}=(\boldsymbol{y}_1,\cdots,\boldsymbol{y}_T)$, the number of factors m.

Output: The optimal portfolio weight $\widehat{\boldsymbol{w}}_{\text{opt}}$.

1: Calculate the GCM based on (4).

2: Calculate the leading m eigenvectors of $\widehat{\boldsymbol{G}}_y$ and obtain the estimated factor loading matrix $\widehat{\boldsymbol{L}}=(\widehat{\boldsymbol{l}}_1,\cdots,\widehat{\boldsymbol{l}}_N)^{\top}$.

3: **for** t in 1:T **do**

4: ⌊ Use the least squares regression to obtain $\widehat{\boldsymbol{f}}_t$ based on (8).

5: Estimate the common component as $\widehat{x}_{it}=\widehat{\boldsymbol{l}}_i^{\top}\widehat{\boldsymbol{f}}_t$ and the idiosyncratic error as $\widehat{\epsilon}_{it}=y_{it}-\widehat{x}_{it}$.

6: Estimate the correlations embedded in the latent common component and the idiosyncratic error based on (7) and (9).

7: Calculate the final return GCM as $\widehat{\boldsymbol{G}}_y=\widehat{\boldsymbol{G}}_x+\widehat{\boldsymbol{G}}_{\epsilon}$.

8: Output the optimal portfolio weight

$\widehat{\boldsymbol{w}}_{\text{opt}}=\frac{\widehat{\boldsymbol{G}}_y^{-1}\mathbf{1}}{\mathbf{1}^{\top}\widehat{\boldsymbol{G}}_y^{-1}\mathbf{1}}$.

3 Theoretical analysis

This section discusses theoretical guarantees within the covariance matrix estimation process, which are primarily set in the elliptical distribution framework. Initially, we assume a low-rank structure for latent common components and apply factor modeling for estimation. During this, we extract the eigenvectors of the sample covariance matrix to estimate factor loadings. We now analyze the theoretical properties of estimating latent common components on the basis of certain fundamental assumptions.

Assumption 1. We assume that

$$\begin{pmatrix}\boldsymbol{f}_t\\ \boldsymbol{\epsilon}_t\end{pmatrix}=\zeta_t\begin{pmatrix}\boldsymbol{I}_m & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{A}\end{pmatrix}\frac{\boldsymbol{g}_t}{\|\boldsymbol{g}_t\|},$$

where ζ_t's represent independent samples from ζ, and $\boldsymbol{g}_t$'s are independent realizations of $\boldsymbol{g}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I}_{m+N})$ with fixed m. Additionally, ζ and $\boldsymbol{g}$ are mutually independent, and the condition $\zeta/\sqrt{N}=O_p(1)$ holds as $N\to\infty$. This implies that the pairs $(\boldsymbol{f}_t^{\top},\boldsymbol{\epsilon}_t^{\top})^{\top}$ are independent realizations of $\mathrm{ED}\left(\boldsymbol{0},\boldsymbol{\Sigma}_0,\zeta\right)$. Here we use ED to represent the elliptical distribution. Elliptical distributions are crucial for $t=1,\cdots,T$, with $\boldsymbol{\Sigma}_0=\begin{pmatrix}\boldsymbol{I}_m & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{\Sigma}_{\epsilon}\end{pmatrix}$ and $\boldsymbol{\Sigma}_{\epsilon}=\boldsymbol{A}\boldsymbol{A}^{\top}$. To guarantee the identifiability of the model, we stipulate that $\left\|\operatorname{diag}(\boldsymbol{\Sigma}_0)\right\|_{\infty}=1$.

Assumption 2. We assume that $\boldsymbol{L}^{\top}\boldsymbol{L}/N\to\boldsymbol{V}$ as $N\to\infty$, with $\boldsymbol{V}$ being positive definite. Additionally, we posit the existence of positive constants c_1 and c_2, bounding the eigenvalues of $\boldsymbol{V}$: $c_1\leq\lambda_m(\boldsymbol{V})<\cdots<\lambda_1(\boldsymbol{V})\leq c_2$.

Assumption 3. We assume that the eigenvalues of $\boldsymbol{\Sigma}_{\epsilon}$ are bounded by two positive constants c_3 and c_4: $c_3\leq\lambda_{\min}(\boldsymbol{\Sigma}_{\epsilon})\leq\lambda_{\max}(\boldsymbol{\Sigma}_{\epsilon})\leq c_4$.

Assumption 1 states that the vectors $(\boldsymbol{f}_t^{\top},\boldsymbol{\epsilon}_t^{\top})^{\top}$ are identically and independently distributed (i.i.d.) and adhere to an elliptical distribution. Consequently, the vectors $\boldsymbol{y}_t$ are also i.i.d. and elliptically distributed, with a scatter matrix given by $\Sigma_y=\boldsymbol{L}\boldsymbol{L}^{\top}+\Sigma_{\epsilon}$. We assume the scatter matrix for $\boldsymbol{f}_t$ is the identity matrix; however, this can extend to any symmetric

CSTR: 32290.14.JUSTC-2025-0003
DOI: 10.52396/JUSTC-2025-0003
JUSTC, **2025**, 55(8): 0806

positive definite matrix $\boldsymbol{\Sigma}_f$. Specifically, with $\boldsymbol{L}^{\text{new}} = \boldsymbol{L}\boldsymbol{\Sigma}_f^{1/2}$ and $\boldsymbol{f}_t^{\text{new}} = \boldsymbol{\Sigma}_f^{-1/2}\boldsymbol{f}_t$, the model adapts this generalization. Assumption 2 assumes that $\boldsymbol{L}^\top\boldsymbol{L}/N$ converges to a fixed positive definite matrix with bounded eigenvalues. The distinct eigenvalues assumption in Assumption 2 ensures the identifiability of the corresponding eigenvectors. Lastly, Assumption 3 requires that the eigenvalues of $\boldsymbol{\Sigma}_\epsilon$ are bounded within specific limits. This ensures that the idiosyncratic error is negligible compared with common components.

Next, We introduce a key property encompassing the consistency of factor and factor loading estimations, along with a principal theorem for the consistency of common component estimation. These form the theoretical basis for estimating portfolio weights in our multi-factor strategy.

Proposition 1. Within the framework of Assumptions 1–3, there exists a sequence of matrices $\widehat{\boldsymbol{H}}$ such that $\widehat{\boldsymbol{H}}^\top\boldsymbol{V}\widehat{\boldsymbol{H}} \xrightarrow{\text{p}} \boldsymbol{I}_m$ as $\min\{N,T\} \to \infty$. Furthermore, it has been established:

(ⅰ) The factor loading matrix estimation satisfies

$$\frac{1}{N}\left\|\widehat{\boldsymbol{L}} - \boldsymbol{L}\widehat{\boldsymbol{H}}\right\|_F^2 = O_p\left(\frac{1}{N}+\frac{1}{T}\right);$$

(ⅱ) for any $t \le T$, the factor estimation satisfies

$$\left\|\widehat{\boldsymbol{H}}\widehat{\boldsymbol{f}}_t - \boldsymbol{f}_t\right\|^2 = O_p\left(\frac{1}{N}+\frac{1}{T}\right).$$

Proposition 1 establishes the convergence rates for both factor and factor loading estimates without imposing moment constraints on the factors or idiosyncratic error terms, which is made possible by means of the GCM.

Theorem 1. Under the framework of Assumption 1–3 and $\min\{N,T\} \to \infty$, for any $t \le T$, we have

$$\frac{1}{N}\left\|\widehat{\boldsymbol{x}}_t - \boldsymbol{x}_t\right\|^2 = O_p\left(\frac{1}{N}+\frac{1}{T}\right).$$

Theorem 1 demonstrates that the estimated common component consistently approximates the actual component, thereby substantiating our ability to estimate the covariance matrix via the estimated common component.

4 Simulations

In this section, we conduct simulation studies to examine the empirical performance of the proposed method in finite samples. The foundation of these simulation exercises is rooted in the following model:

$$y_{it} = x_{it} + \sqrt{\theta}e_{it}, \quad x_{it} = \sum_{k=1}^{m} l_{ik}f_{kt},$$

$$e_{it} = \rho e_{i,t-1} + (1-\beta)v_{it} + \sum_{l=i-J}^{i+J}\beta v_{lt},\ i = 1,\cdots,N,\ t = 1,\cdots,T.$$

Notably, similar data-generating models have been employed in previous studies, including Refs. [17, 19, 20, 22], where they have exhibited favorable behavior in simulation exercises. In this study, we explore dimensions for (N,T), focusing on $(150,100)$ and $(250,150)$, with the number of factors set to $m = 3$. We also examine three parameter configurations and three distinct data-generating scenarios.

Setting 1. The vectors $(\boldsymbol{f}_t^\top, \boldsymbol{v}_t^\top)^\top$ are generated as independent and identically distributed (i.i.d.) samples from a multivariate Gaussian distribution $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{N+m})$, with the factor loadings $l_{ik} \sim \mathcal{N}(0,1)$.

Setting 2. The vectors $(\boldsymbol{f}_t^\top, \boldsymbol{v}_t^\top)^\top$ are generated as i.i.d. samples from a multivariate centralized t-distribution $t_2(\boldsymbol{0}, \boldsymbol{I}_{N+m})$, with factor loadings $l_{ik} \sim \mathcal{N}(0,1)$.

Setting 3. The factor $\boldsymbol{f}_t$ is generated as i.i.d. samples from a multivariate skewed-t_2 distribution, with the components of $\boldsymbol{v}_t$ independently sampled from $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_N)$. The factor loadings l_{ik} are drawn from a uniform distribution $\boldsymbol{U}(0,1)$.

Scenario A. The case of pure serial correlation is obtained when the cross-sectional correlation parameter $\beta = 0$. The following parameter values are used: $\beta = 0$, $\rho = 0.5$, $J = 0$, and $\theta = 1$.

Scenario B. The scenario of pure cross-sectional dependence is realized when $\rho = 0$. The following parameter values are employed: $\beta = 0.2$, $\rho = 0$, $J = \max\{N/20, 10\}$, and $\theta = 1$.

Scenario C. This scenario introduces serial and cross-sectional correlation settings. We set the parameters as follows: $\beta = 0.2$, $\rho = 0.5$, $J = \max\{N/20, 10\}$, and $\theta = 1$.

To evaluate the empirical performance of different methods, we employ the following performance indicators. MEAN_FL: Mean deviation in estimating the factor loading matrices; MEAN_FS: Mean deviation in estimating the factor matrices; MED_CC: Median deviation in normalized estimates of the common components. These indicators serve as quantitative measures to assess the effectiveness of the methods under consideration. Specifically,

$$\text{MEAN\_FL} = \frac{1}{R}\sum_{i=1}^{R}\mathcal{D}\left(\widehat{\boldsymbol{L}}_i, \boldsymbol{L}\right),$$

$$\text{MEAN\_FS} = \frac{1}{R}\sum_{i=1}^{R}\mathcal{D}\left(\widehat{\boldsymbol{F}}_i, \boldsymbol{F}\right),$$

$$\text{MED\_CC} = \text{median}\left\{\frac{\left\|\widehat{\boldsymbol{L}}_i\widehat{\boldsymbol{F}}_i^\top - \boldsymbol{L}\boldsymbol{F}^\top\right\|_F^2}{\|\boldsymbol{L}\boldsymbol{F}^\top\|_F^2},\ i = 1,\cdots,R\right\},$$

where R represents the number of replications, and $\widehat{\boldsymbol{L}}_i$ and $\widehat{\boldsymbol{F}}_i$ correspond to the estimated factor loading and factor matrices from the ith replication, respectively. Given two orthogonal matrices $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$ with dimensions $N \times p_1$ and $N \times p_2$, respectively, we define the distance metric $\mathcal{D}(\boldsymbol{M}_1, \boldsymbol{M}_2)$ as:

$$\mathcal{D}(\boldsymbol{M}_1, \boldsymbol{M}_2) = \left(1 - \frac{1}{\max(p_1,p_2)}\operatorname{tr}(\boldsymbol{M}_1\boldsymbol{M}_1^\top\boldsymbol{M}_2\boldsymbol{M}_2^\top)\right)^{1/2}.$$

If $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$ are not column orthogonal initially, we can use the Gram–Schmidt orthogonal transformation procedure. The metric $\mathcal{D}(\boldsymbol{M}_1, \boldsymbol{M}_2)$ quantifies the distance between the column spaces of $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$, with values ranging from 0 to 1. A value of 0 indicates that the column spaces of $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$ are identical, whereas a value of 1 signifies orthogonality between them.

Table 1 presents a summary of the simulation outcomes of the three scenarios. As described in this paper, "Gini" refers to the factor estimation method using the GCM, "PCA" denotes the method based on the sample covariance matrix, and "KEN" represents the factor estimation method employing

CSTR: 32290.14.JUSTC-2025-0003
DOI: 10.52396/JUSTC-2025-0003
JUSTC, **2025**, 55(8): 0806

Table 1. Simulation results for Scenarios A, B, and C.

		Method	$(N, T) = (150, 100)$			$(N, T) = (250, 150)$		
			MEAN_FL	MEAN_FS	MED_CC	MEAN_FL	MEAN_FS	MED_CC
Scenario A	Setting 1	Gini	0.12(0.01)	0.10(0.01)	0.02(0.00)	0.10(0.01)	0.07(0.01)	0.01(0.00)
		PCA	0.12(0.01)	0.10(0.01)	0.02(0.00)	0.10(0.01)	0.07(0.01)	0.01(0.00)
		KEN	0.12(0.01)	0.10(0.01)	0.02(0.00)	0.10(0.01)	0.07(0.01)	0.01(0.00)
	Setting 2	Gini	0.24(0.10)	0.23(0.15)	0.05(0.07)	0.19(0.12)	0.18(0.15)	0.03(0.03)
		PCA	0.33(0.21)	0.32(0.22)	0.07(0.36)	0.29(0.21)	0.27(0.22)	0.04(0.26)
		KEN	0.19(0.06)	0.18(0.10)	0.05(0.04)	0.15(0.04)	0.13(0.07)	0.03(0.02)
	Setting 3	Gini	0.12(0.02)	0.08(0.02)	0.01(0.01)	0.09(0.01)	0.06(0.01)	0.01(0.01)
		PCA	0.11(0.02)	0.08(0.02)	0.01(0.01)	0.08(0.01)	0.06(0.01)	0.01(0.01)
		KEN	0.15(0.02)	0.09(0.02)	0.03(0.01)	0.12(0.01)	0.06(0.01)	0.02(0.01)
Scenario B	Setting 1	Gini	0.15(0.02)	0.13(0.02)	0.04(0.01)	0.13(0.01)	0.10(0.01)	0.02(0.01)
		PCA	0.15(0.02)	0.13(0.02)	0.04(0.01)	0.12(0.01)	0.10(0.01)	0.02(0.01)
		KEN	0.15(0.02)	0.13(0.02)	0.04(0.01)	0.13(0.01)	0.10(0.01)	0.03(0.01)
	Setting 2	Gini	0.33(0.13)	0.33(0.16)	0.12(0.21)	0.25(0.12)	0.25(0.15)	0.06(0.08)
		PCA	0.45(0.20)	0.43(0.21)	0.26(0.63)	0.41(0.21)	0.39(0.22)	0.15(0.57)
		KEN	0.27(0.07)	0.27(0.11)	0.10(0.10)	0.21(0.05)	0.19(0.07)	0.06(0.04)
	Setting 3	Gini	0.16(0.05)	0.12(0.04)	0.03(0.02)	0.12(0.03)	0.08(0.02)	0.02(0.01)
		PCA	0.12(0.03)	0.10(0.03)	0.02(0.02)	0.10(0.02)	0.08(0.02)	0.01(0.01)
		KEN	0.23(0.07)	0.14(0.06)	0.05(0.03)	0.17(0.03)	0.09(0.03)	0.03(0.01)
Scenario C	Setting 1	Gini	0.18(0.02)	0.16(0.02)	0.06(0.02)	0.15(0.02)	0.12(0.02)	0.04(0.01)
		PCA	0.18(0.02)	0.15(0.02)	0.06(0.02)	0.15(0.02)	0.12(0.02)	0.04(0.01)
		KEN	0.19(0.02)	0.16(0.02)	0.06(0.02)	0.15(0.02)	0.12(0.02)	0.04(0.01)
	Setting 2	Gini	0.49(0.15)	0.48(0.17)	0.37(0.57)	0.40(0.16)	0.41(0.18)	0.25(0.48)
		PCA	0.54(0.17)	0.52(0.18)	0.48(0.75)	0.51(0.20)	0.50(0.20)	0.42(0.77)
		KEN	0.47(0.13)	0.46(0.15)	0.34(0.48)	0.37(0.12)	0.35(0.14)	0.20(0.21)
	Setting 3	Gini	0.21(0.08)	0.16(0.08)	0.04(0.04)	0.15(0.04)	0.10(0.03)	0.03(0.02)
		PCA	0.17(0.04)	0.13(0.04)	0.03(0.03)	0.13(0.03)	0.09(0.02)	0.02(0.02)
		KEN	0.31(0.11)	0.21(0.11)	0.09(0.08)	0.22(0.06)	0.13(0.05)	0.05(0.03)

The values in parentheses are the standard deviations for MEAN_FL, MEAN_FS, and the interquartile ranges for MED_CC.

the Kendall τ matrix. Several key observations emerge from the analysis of the tables. When both the factor and error terms follow normal distributions, the performances of the three methods are similar, suggesting that satisfactory factor estimates can be obtained via either moment or rank information. In Setting 2, where both the factor and error terms are heavy-tailed, the Gini and KEN methods significantly outperform PCA. This is because, in the heavy-tailed case, the Gini and KEN methods leverage rank information, which helps mitigate the absence of higher-order moments, a limitation of PCA. However, in Setting 3, where the factor and error terms follow t-distributions and normal distributions, respectively, the Gini and PCA methods perform better. This is because, compared with the KEN method, which considers only rank information, the Gini and PCA methods also utilize moment information, allowing them to capture the data structure more efficiently. The performance of all three methods improves as the sample sizes N and T increase, demonstrating their scalability and suitability for application to larger datasets. In conclusion, the Gini method proposed in this paper is highly applicable across a variety of distributional scenarios.

5 Portfolio analysis

This section provides an analysis of daily stock return data derived from the S&P 100 constituent companies, covering the period from January 1, 2020 to December 31, 2022. The S&P 100 index, comprising 100 large-cap blue-chip stocks traded on the New York Stock Exchange and Nasdaq, represents leading enterprises across diverse industries. These constituent companies not only serve as industry benchmarks but also reflect the developmental status and trends of crucial sectors within the U.S. economy. Given its composition, the performance of the S&P 100 index is strongly correlated with both the U.S. and global economic conditions, making it a focal point for investors, financial institutions, and analysts worldwide. In our methodology, we implement a sliding window technique to sequentially determine optimal investment

CSTR: 32290.14.JUSTC-2025-0003
DOI: 10.52396/JUSTC-2025-0003
JUSTC, **2025**, 55(8): 0806

weights, ensuring dynamic portfolio adjustments in response to market fluctuations. A standardized time window of approximately two months ($s = 44$ trading days) is used, with out-of-sample testing starting on March 6, 2020. For each investment day t, we calculate $\boldsymbol{G}_{m,t}$ via returns from the preceding 44 trading days, where m denotes the number of factors. This estimated covariance matrix is then employed to construct the optimal portfolio. To compare the performance of models with different numbers of factors, we set m equal to 3 and 5, mirroring the Fama–French 3-factor and 5-factor models.

Specifically, for each day following March 6, 2020, we recursively compile returns over the past $s = 44$ days, resulting in a data panel of dimensions 100×44. We then estimate $\widehat{\boldsymbol{G}}_{m,t}$ for day t via the methodology outlined in Algorithm 1. According to Section 2, we consider the following problem for portfolios that do not allow short selling:

$$\widehat{\boldsymbol{w}}_{\mathrm{opt},t} = \arg\min_{\boldsymbol{w}} \boldsymbol{w}^{\top}\widehat{\boldsymbol{G}}_{m,t}\boldsymbol{w}$$
$$\text{subject to} \quad \boldsymbol{w}^{\top}\boldsymbol{1} = 1,\ 0 \le w_i \le 1 \text{ for all } 1 \le i \le N,$$

where $\boldsymbol{w} = (w_1, \cdots, w_N)^{\top}$ denotes the optimal weights. This optimization problem incorporates a short-selling constraint into the framework of problem (3). Such a formulation constitutes a standard quadratic programming problem with boundary constraints, which can be solved via existing tools such as the R package **quadprog**. Hence, the portfolio return on day t can be expressed as $\widehat{\boldsymbol{w}}_{\mathrm{opt},t}^{\top}\boldsymbol{y}_t$, with $\boldsymbol{y}_t$ representing the original return for day t. Note that in a later analysis, we comparatively analyze the KEN method and the PCA method, denoting portfolios on the basis of the sample covariance estimates proposed by Fan et al.[12]

Fig. 1 shows the net value curves for the three methods at $m = 3$ and 5 from 2020 to 2022, excluding transaction costs and liquidity risk. Overall, the Gini method outperformed the PCA and KEN methods across most of the period, maintaining a small but stable advantage. This suggests that the Gini method has a certain superiority in portfolio optimization. However, the cumulative returns experienced several declines, reflecting the significant fluctuations in the U.S. stock market over the past few years. Early to mid-2020: global economic stagnation and market panic triggered by the COVID-19 outbreak. Mid-2022: the Federal Reserve's interest rate hike and rising inflation triggered market concerns.

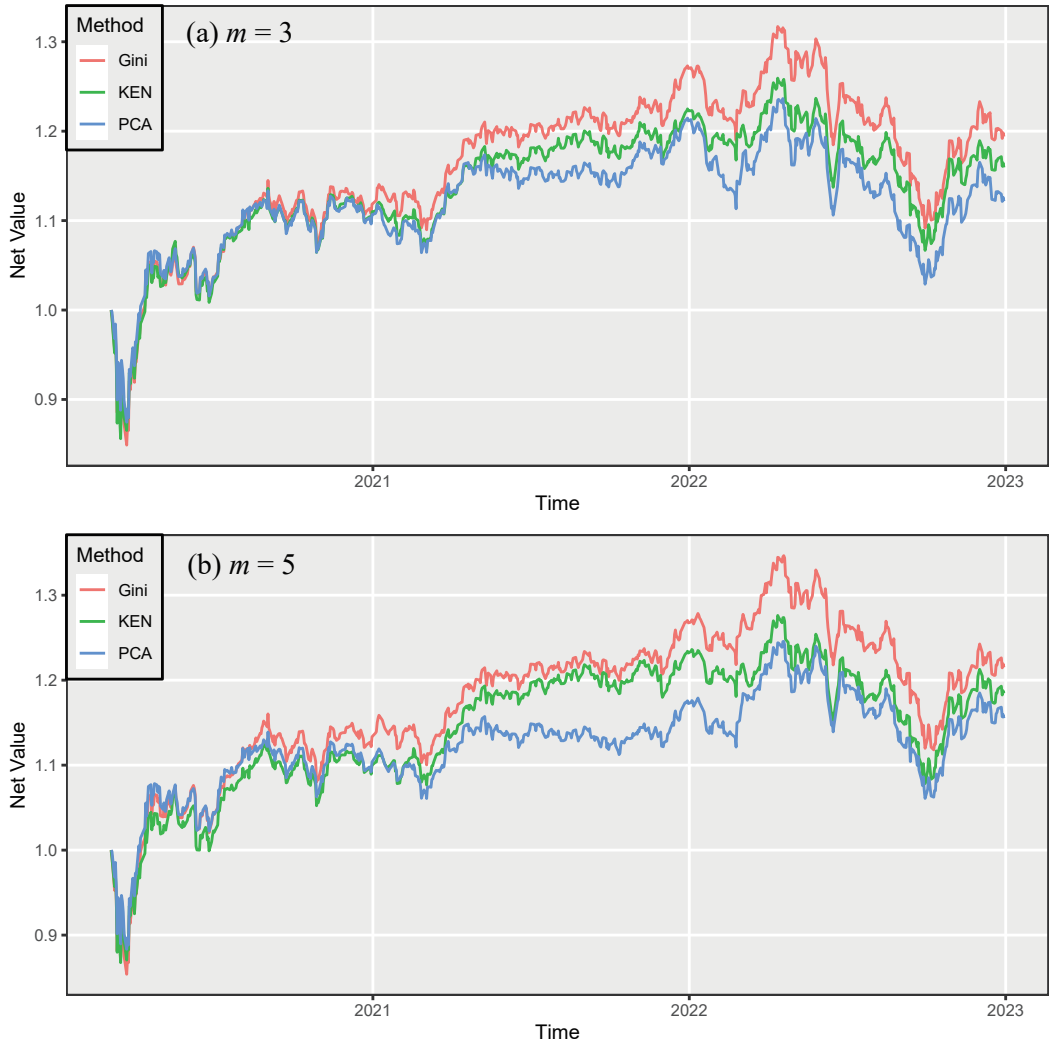


Fig. 1. Portfolio net value curves based on the Gini, PCA, and KEN methods with factors of (a) $m = 3$ and (b) $m = 5$.

CSTR: 32290.14.JUSTC-2025-0003
DOI: 10.52396/JUSTC-2025-0003
JUSTC, **2025**, 55(8): 0806

As mentioned earlier, the Gini method is more adaptable to various market conditions, which explains its superior performance in the constantly fluctuating financial market.

6 Conclusions

This paper introduces a novel approach to large-dimensional portfolio analysis: the Gini covariance strategy with multiple factors, which is designed to optimize portfolios in financial markets. Specifically, first, we estimate the factor loadings and factors. Next, we obtain the estimates of the GCM for the latent common components and the idiosyncratic errors with a sparse structure. Finally, we combine these estimates to derive the overall estimate. Theoretical demonstrations of the consistency of component estimation, along with simulations and real portfolio studies, reveal that the Gini method is highly applicable across a variety of distributional scenarios. Future research will delve into the theoretical properties of the method, focusing on the estimation error bounds of the return covariance matrix.

Supporting information

The supporting information for this article can be found online at http://doi.org/10.52396/JUSTC-2025-0003. It includes several lemmas and proofs of Proposition 1 and Theorem 1.

Acknowledgements

This work was supported by the Postdoctoral Fellowship Program of CPSF (GZC20241651), the National Natural Science Foundation of China (12501391), and the Natural Science Foundation of Anhui Province (2408085QA005).

Conflict of interest

The authors declare that they have no conflict of interest.

Biographies

Yongda Zhu is currently a graduate student at the University of Science and Technology of China. His research mainly focuses on the factor model.

Lei Shu is a postdoctoral fellow at the University of Science and Technology of China (USTC). He received his Ph.D. degree in Statistics from USTC in 2023. His research mainly focuses on high-dimensional statistical inference.

References

[1] Markowitz H. Portfolio selection. *The Journal of Finance*, **1952**, *7* (1): 77–91.
[2] DeMiguel V, Garlappi L, Uppal R. Optimal versus naive diversification: How inefficient is the 1/*N* portfolio strategy? *The Review of Financial Studies*, **2009**, *22* (5): 1915–1953.
[3] Choueifaty Y, Coignard Y. Toward maximum diversification. *The Journal of Portfolio Management*, **2008**, *35* (1): 40–51.
[4] Campbell J Y, Lo A W, MacKinlay A C. The Econometrics of Financial Markets. Princeton, NJ, USA: Princeton University Press, **1996**.
[5] McNeil A J, Frey R, Embrechts P. Quantitative Risk Management: Concepts, Techniques and Tools. Revised Edition. Princeton, NJ, USA: Princeton University Press, **2015**.
[6] Bodnar T, Schmid W. A test for the weights of the global minimum variance portfolio in an elliptical model. *Metrika*, **2008**, *67*: 127–143.
[7] Chamberlain G. A characterization of the distributions that imply mean–variance utility functions. *Journal of Economic Theory*, **1983**, *29* (1): 185–201.
[8] Owen J, Rabinovitch R. On the class of elliptical distributions and their applications to the theory of portfolio choice. *The Journal of Finance*, **1983**, *38* (3): 745–752.
[9] Sang Y, Dang X, Sang H. Symmetric Gini covariance and correlation. *The Canadian Journal of Statistics*, **2016**, *44* (3): 323–342.
[10] Dang X, Sang H, Weatherall L. Gini covariance matrix and its affine equivariant version. *Statistical Papers*, **2019**, *60* (3): 641–666.
[11] Bai J. Inferential theory for factor models of large dimensions. *Econometrica*, **2003**, *71* (1): 135–171.
[12] Fan J, Fan Y, Lv J. High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics*, **2008**, *147* (1): 186–197.
[13] Ledoit O, Wolf M. Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of Empirical Finance*, **2003**, *10* (5): 603–621.
[14] Chamberlain G. Funds, factors, and diversification in arbitrage pricing models. *Econometrica*, **1983**, *51* (5): 1305–1324.
[15] Fan J, Liu H, Wang W. Large covariance estimation through elliptical factor models. *The Annals of Statistics*, **2018**, *46* (4): 1383–1414.
[16] De Nard G, Ledoit O, Wolf M. Factor models for portfolio selection in large dimensions: The good, the better and the ugly. *Journal of Financial Econometrics*, **2021**, *19* (2): 236–257.
[17] He Y, Kong X, Yu L, et al. Large-dimensional factor analysis without moment constraints. *Journal of Business & Economic Statistics*, **2022**, *40* (1): 302–312.
[18] Adcock C, Azzalini A. A selective overview of skew-elliptical and related distributions and of their applications. *Symmetry*, **2020**, *12* (1): 118.
[19] Onatski A. Determining the number of factors from empirical distribution of eigenvalues. *The Review of Economics and Statistics*, **2010**, *92* (4): 1004–1016.
[20] Ahn S C, Horenstein A R. Eigenvalue ratio test for the number of factors. *Econometrica*, **2013**, *81* (3): 1203–1227.
[21] Bickel P J, Levina E. Covariance regularization by thresholding. *The Annals of Statistics*, **2008**, *36* (6): 2577–2604.
[22] Bai J, Ng S. Determining the number of factors in approximate factor models. *Econometrica*, **2002**, *70* (1): 191–221.

CSTR: 32290.14.JUSTC-2025-0003
DOI: 10.52396/JUSTC-2025-0003
JUSTC, **2025**, 55(8): 0806