High Dimension Statistical Report

Covariance Matrix Estimation under Total Positivity for Portfolio Selection

Lyheng LAO, Guang YANG

January 18, 2026

Abstract

The paper [ARU20] introduces a new method for estimating the covariance matrix of asset returns, using the concept of total positivity of order 2 (MTP_2), which is given by the positive dependence among assets. This approach improves the performance of covariance matrix estimation compared to traditional methods, especially when the order of the assets N is similar to the order of the historical data periods T.

Contents

1 Introduction

The best way to select the portfolio with minimal risk or variance for a given level of expected return, commonly known as the Markowitz optimal portfolio, depends on two quantities: the vector of expected returns μ^* and the covariance matrix of returns Σ^*. These parameters are estimated from historical returns.

Based on the Markowitz portfolio optimization framework, the article focuses on estimating the covariance matrix of asset returns Σ^*. It highlights the problems with using traditional sample covariance matrices, particularly when N and T are of similar size, which leads to poor estimation performance due to insufficient effective samples per parameter. The article proposes a new method for estimating the covariance matrix by exploiting the concept of **multivariate total positivity of order** 2 **(**MTP_2**)**.

The article also introduces various other methods and compares them with the MTP_2 method through experiments.

Let us adopt the following notations and abbreviations throughout the report:

- N: number of assets.
- T: period (dates).
- $r_{i,t}$: observed return for asset i at date t for $1 \leq i \leq N$ and $1 \leq t \leq T$.
- $r_t := (r_{1,t}, ..., r_{N,t})^T$ consists of the returns of each asset on day t, and μ_t, Σ_t denote its expected and covariance matrix.
- $X \sim \mathcal{N}(\mu, \Sigma)$: Multivariate Gaussian distribution, where:
  - X is a random vector;
  - μ is a vector representing the mean of each variable;
  - Σ is a covariance matrix that describes the covariance between the variables.
- $K := \Sigma^{-1}$, the precision matrix.
- We use $\hat{*}$ to represent the estimator of $*$, such as $\hat{\Sigma}$ and $\hat{K}$.
- MTP_2: Multivariate Totally Positive of Order 2.
- **MLE**: Maximum Likelihood Estimation.
- **MSE**: Mean Squared Error.
- **HD**: High Dimension. This refers to situations where the data or problem involves a very high number of dimensions, especially when the number of parameters p is comparable to the number of samples n, or even when the number of parameters exceeds the number of samples.

2 Markowitz Portfolio

The *Markowitz portfolio* is a mathematical framework for constructing investment portfolios that seek to balance risk and return. Markowitz portfolio theory consists of assigning weights to a subset of N assets in order to minimize the portfolio variance for some threshold R (expected return).

Consider N assets. Let R_i be the return of i-th asset, and w_i the weights of this asset in the portfolio, where $1 \leq i \leq N$. The total return of our portfolio R is the weighted sum of the individual asset returns:

$$R = \sum_{i=1}^{N} w_i R_i. \tag{1}$$

Then, the variance of the portfolio return $\mathbf{var}(R)$ can be expressed as

$$\mathbf{var}(R) := \mathbb{E}((R - \mathbb{E}[R])^2).$$

Considering the equation (1), we have:

$$\begin{aligned} \mathbf{var}(R) &= \mathbb{E}\left(\sum_{i=1}^{N} w_i R_i - \mathbb{E}\left[\sum_{i=1}^{N} w_i R_i\right]\right)^2 \\ &= \mathbb{E}(\sum_{i=1}^{N} w_i (R - \mathbb{E}[R])^2 \\ &= \sum_{i=1}^{N} \sum_{j=1}^{N} w_i w_j \mathbf{cov}(R_i, R_j). \end{aligned}$$

Let Σ be the covaraince matrix of our return portfolio, and $\Sigma = (\sigma_{ij})$, where $\sigma_{ij} = \mathbf{cov}(R_i, R_j)$. Write $\mathbf{w} = (w_1, \ldots, w_N)^T$ Then,

$$\mathbf{var}(R) = \mathbf{w}\Sigma\mathbf{w}^T.$$

Given the expected returns R, ones need to minimize the portfolio variance. In other words, the original problem has been transformed into the following problem:

$$\begin{aligned} &\min_{\mathbf{w}\in\mathbb{R}^N} \mathbf{w}^T \Sigma^* \mathbf{w} \\ &\text{subject to } \mathbf{w}^T \mu^* = R \text{ and } \sum_{i=1}^{N} w_i = 1, \end{aligned}$$

where μ^* and Σ^* denote the true expected returns and covariance matrix of the returns.

Remark 2.1. Jagannathan and Ma explained in their paper that we can transform the search for a conditionally optimal solution here into the search for a globally optimal solution [JM03] [1].

Therefore, ignoring $\mathbf{w}^T \mu^* = R$, we only need to search the *global minimum variance* portfolio:

$$\min_{\mathbf{w}\in\mathbb{R}^N} \mathbf{w}^T \Sigma^* \mathbf{w} \quad \text{subject to} \quad \sum_{i=1}^{N} w_i = 1,$$

[1] For details, please see page 4 of the article, which is page 1653 of the journal.

To solve this equation we use the method of Lagrange multipliers. Write $\mathbf{1} = (1, \ldots, 1)^T$, then we have

$$\mathbf{1}^T \mathbf{w} = 1,$$

and the corresponding Lagrange function $\hat{\mathcal{L}}$ is:

$$\hat{\mathcal{L}}(\mathbf{w}, \lambda) = \mathbf{w}^T \Sigma^* \mathbf{w} - \lambda(\mathbf{1}^T \mathbf{w} - 1),$$

which leads to

$$\lambda = -\frac{2}{\mathbf{1}^T \Sigma^{-1} \mathbf{1}}$$

and

$$\mathbf{w} = \frac{\hat{\Sigma}_t^{-1} \mathbf{1}}{\mathbf{1}^T \hat{\Sigma}_t^{-1} \mathbf{1}}. \tag{2}$$

3 Multivariate Totally Positive of Order 2

The innovation in the paper [ARU20] lies in the discussion of multivariate totally positive of order 2. In this section, we will also focus on introducing the related concepts and details.

3.1 Basic Concepts

Definition 3.1. For vectors $\mathbf{p} = (p_1, p_2, \ldots, p_n)$ and $\mathbf{q} = (q_1, q_2, \ldots, q_n)$, the coordinate-wise minimum $\mathbf{p} \wedge \mathbf{q}$ is defined as

$$\mathbf{p} \wedge \mathbf{q} = (\min(p_1, q_1), \min(p_2, q_2), \ldots, \min(p_n, q_n)).$$

Similarly, the coordinate-wise maximum $\mathbf{p} \vee \mathbf{q}$ is defined as

$$\mathbf{p} \vee \mathbf{q} = (\max(p_1, q_1), \max(p_2, q_2), \ldots, \max(p_n, q_n)).$$

Definition 3.2 ([KR80])**.** A distribution on $\mathcal{X} \subseteq \mathbb{R}^M$ is multivariate totally positive of order 2 (MTP_2) if its density function p satisfies

$$p(x)p(y) \leq p(x \wedge y)p(x \vee y) \quad \text{for all} \quad x, y \in \mathcal{X},$$

where $\wedge, \vee$ denote the coordinate-wise minimum and maximum, respectively.

Remark 3.1. MTP_2 is a strong form of positive dependence that implies the positive association; see [CSS05].

Definition 3.3. Let A be a $n \times n$ real Z-matrix. That is,

$$A = \left(a_{ij}\right), \text{where } a_{ij} \leq 0 \text{ for all } i \neq j, 1 \leq i, j \leq n.$$

Then matrix A is also an M-matrix if it can be expressed in the form

$$A = sI - B,$$

where $B = \left(b_{ij}\right)$ with $b_{ij} \geq 0$, for all $1 \leq i, j \leq$ n, where s is at least as large as the maximum of the moduli of the eigenvalues of B, and I is an identity matrix.

Theorem 3.1. A multivariate Gaussian distribution is MTP$_2$ if and only if the precision matrix $K := \Sigma^{-1}$ is an M-matrix.

For the proof of the above theorem, see the appendix.

3.2 Why We discuss MTP_2?

The reasons stem from both theoretical and practical aspects.

Single-factor Analysis Model

Consider the following theorem.

Theorem 3.2 ([LUZ19]). Let $X \in \mathbb{R}^M$ follow a multivariate Gaussian distribution that factorizes according to a tree. If $\mathrm{Cov}(X) \geq 0$, then X is MTP_2 and any marginal of X is MTP_2.

Remark 3.2. The factor analysis model with a single factor is a particular example of a latent tree model consisting of an unobserved root variable that is connected to all the observed variables.

At the same time, we have the prominent capital asset pricing model (CAPM), which is a *single-factor analysis model*: the return of stock i is modeled as

$$r_i = r_f + \beta_i \left(r_m - r_f\right) + u_i, \quad \beta_i \in \mathbb{R}, \tag{3}$$

where r_f is known as the risk-free rate of return, r_m is the market return, and u_i is the uncorrelated, zero mean idiosyncratic error term.

Positive Correlation between Asset Returns

An interesting phenomenon is that asset returns are often *positively correlated* since assets typically move together with the market.

The following is the sample correlation matrix of global stock market indices based on monthly returns from 2013-2016[2].

Nasdaq	Canada	Europe	UK	Australia	
1.000	0.606	0.731	0.618	0.613	Nasdaq
0.606	1.000	0.550	0.661	0.598	Canada
0.731	0.550	1.000	0.644	0.569	Europe
0.618	0.661	0.644	1.000	0.615	UK

$$S = \begin{pmatrix} 1.000 & 0.606 & 0.731 & 0.618 & 0.613 \\ 0.606 & 1.000 & 0.550 & 0.661 & 0.598 \\ 0.731 & 0.550 & 1.000 & 0.644 & 0.569 \\ 0.618 & 0.661 & 0.644 & 1.000 & 0.615 \\ 0.613 & 0.598 & 0.569 & 0.615 & 1.000 \end{pmatrix}, \; S^{-1} = \begin{pmatrix} 2.629 & -0.480 & -1.249 & -0.202 & -0.490 \\ -0.480 & 2.109 & -0.039 & -0.790 & -0.459 \\ -1.249 & -0.039 & 2.491 & -0.675 & -0.213 \\ -0.202 & -0.790 & -0.675 & 2.378 & -0.482 \\ -0.490 & -0.459 & -0.213 & -0.482 & 1.992 \end{pmatrix}.$$

Remark 3.3. In the equation (3), the parameters β_i are positive, which explains why the covariance between stock returns is usually positive.

The shortcomings of traditional methods

Traditional methods are too crude and lack a real connection with the financial markets; if we are in high dimensional, we will face to the problem of *sparcity, overfitting, bad estimation, computing problem(distance between point is far hard to seek for max), etc.*

[2] Nasdaq: National Association of Securities Dealers Automated Quotations

Let us consider the multi-Gaussian case and pick a simple case to demonstrate that the Maximum Likelihood Estimator (MLE) performs poorly in **HD** (High Dimension). Let $X = (X_1, \dots, X_T)$ be N-dimensional random variables following a multivariate normal distribution. For $\mu \in \mathbb{R}^N$ as its expectation and Σ as its covariance matrix, our MLE for the covariance in this case is:

$$\hat{\mu}_{MLE} = \frac{1}{T}\sum_{i=1}^{T} X_i,$$

$$\hat{\Sigma}_{MLE} = \frac{1}{T}\sum_{i=1}^{T}(X_i - \bar{X})^T(X_i - \bar{X}).$$

We will concentrate in our expectation $\hat{\mu}$ and this estimator is unbiased estimator. Let calculate our Mean Squared Error (MSE):

$$MSE(\hat{\mu}) = \mathbb{E}(\|\hat{\mu} - \mu\|^2) = \mathbb{E}(\sum_{j=1}^{N}(\hat{\mu}_j - \mu_j)^2) = \sum_{j=1}^{N}\frac{1}{T}\Sigma_{jj} \leq \frac{N}{T}\max|\Sigma_{jj}|.$$

If we are in high dimension(when $N \gg T$), our ***MSE*** obviously very distorted, and the estimator works badly.

3.3 MTP_2 Estimator

Consider *multivariate Gaussian distribution*. In this case, given $D := \{r_t\}_{t=1}^{T}$, every r_t is a multi-Gaussian module,

$$r_t \sim \mathcal{N}(0, \Sigma),$$

and its density function is

$$f(r_t) = \frac{1}{(2\pi)^{N/2}\det(\Sigma)^{1/2}}\mathbf{e}^{-\frac{1}{2}r_t^T\Sigma^{-1}r_t}.$$

Therefore, ones have the maximum likelihood function:

$$L(D, \Sigma) = \prod_{t=1}^{T} f(r_t) = \prod_{t=1}^{T}\frac{1}{(2\pi)^{N/2}\det(\Sigma)^{1/2}}\mathbf{e}^{-\frac{1}{2}r_t^T\Sigma^{-1}r_t}.$$

It leads to the log-likelihood function:

$$\begin{aligned}\mathcal{L}(D, \Sigma) &= \sum_{t=1}^{T}[-\frac{N}{2}\log(2\pi) - \frac{1}{2}\log(\det(\Sigma)) - \frac{1}{2}r_t^T\Sigma^{-1}r_t] \\ &= \mathbf{C} + \sum_{t=1}^{T}(\frac{1}{2}\log(\det(\Sigma^{-1})) - \frac{1}{2}r_t^{\mathcal{T}}\Sigma^{-1}r_t),\end{aligned}$$

where **C** is constant.

Recall that $\det(\Sigma^{-1}) = (\det(\Sigma))^{-1}$ and $K := \Sigma^{-1}$. Ignored the constants, then:

$$\begin{aligned}
\mathcal{L}(D, K) &= \frac{T}{2}\log(\det(K)) - \sum_{t=1}^{T} \frac{1}{2} r_t^T K r_t \\
&= \frac{T}{2}\log(\det(K)) - \sum_{t=1}^{T} \frac{1}{2}\text{trace}(K r_t^T r_t) \\
&= \frac{T}{2}\log(\det(K)) - \frac{1}{2}\text{trace}(K \sum_{t=1}^{T} r_t^T r_t) \\
&= \frac{T}{2}\log(\det(K)) - \frac{T}{2}\text{trace}(KS), \quad \text{where } S = \frac{1}{T}\sum_{t=1}^{T} r_t^T r_t.
\end{aligned}$$

Remark 3.4. (a) When $T \leq N$, this the maximum of log-likelyhood function does not exist. *For the proof, please see the appendix.*

(b) However, if we solve the problem under the constraint of MTP_2, it works perfectly. [SH15] [LUZ19]

Note that Σ^{-1} is a M-matrix under the MTP_2 condition, so the problem become:

$$\arg\max_{K \geq 0} \mathcal{L}(D, K), \quad \text{subject to } K_{ij} \leq 0, \quad \text{for } i \neq j, \tag{4}$$

where $K \geq 0$ means that K is a positive semi-definite matrix, which is given by the definition of an M-matrix.

With this method, the estimated covariance matrix contains all partial correlations, which vary between -1 and 1.

Remark 3.5. The estimator $\hat{K}$ obtained by the above method is usually *sparse* [LUZ19] [3]. In this case, we can apply L_1-regularization to our *MLE*, ensuring that only significant, non-negative dependencies are modeled. This sparsity is achieved without any tuning parameter, which is an immediate advantage over methods that explicitly add sparsity-inducing L_1 penalties, such as the graphical lasso (Glasso).

4 Other Methods

This section will briefly introduce some methods that have appeared in other articles. The authors compared the MTP_2 estimator with these methods.

4.1 Glasso

Since the matrix $\hat{K}$ is sparse, we can naturally set a constraint: there exists $\kappa \in \mathbb{R}$ such that

$$\|K\|_0 := \sum_{i \neq j} I\left[K_{ij} \neq 0\right] \leq \kappa.$$

Under this constraint, estimating the matrix K remains difficult, so we replace the L_0 norm by the L_1 norm. This method is called *graphical lasso*, or *GLASSO*.

[3] There is an issue with the citation in the article: the authors state that the result comes from Corollary 2.9 of that article; in fact, this conclusion comes from Corollary 2.4, not 2.9; that article does not have a Corollary 2.9.

Therefore, we get the following convex optimization problem:

$$\arg\max_{K\geq 0} \mathcal{L}(D,K), \quad \text{subject to } \|K\|_1 \leq \lambda, \text{ where } \lambda \in \mathbb{R}^+. \tag{5}$$

The L_1 norm will bring the following advantages:

- **Promoting Sparsity**: Adding the L_1 norm as a regularization term to the objective function encourages the optimization process to produce sparse solutions. In other words, during the optimization process, many elements in the precision matrix K are shrunk to zero, resulting in a sparse matrix.
- **Feature Selection**: Since the L_1 norm pushes some elements towards zero, it effectively performs feature selection. In the estimation of the precision matrix, non-zero elements represent direct dependencies between variables, while zero elements indicate conditional independence.
- **Numerical Stability**: The L_1 regularization term also enhances the numerical stability of the estimation process. Especially when dealing with high-dimensional data, the covariance matrix can be singular or nearly singular, making the computation of its inverse unstable. Introducing L_1 regularization can prevent this instability, ensuring that the estimated precision matrix is positive definite.

4.2 CLIME

Now we have a similar estimator to *GLASSO*, which is *CLIME*. For this estimator, instead of maximizing the log likelihood, it finds a sparse estimate of the precision matrix by solving:

$$\hat{K} = \arg\min_{K\geq 0} \|K\|_1, \text{ subject to } \|SK - I\|_\infty \leq \lambda. \tag{6}$$

Indeed, if we reformulate equation (5) in an equivalent way, we have:

$$\arg\min_{K} \|K\|_1, \text{ subject to } \|trace(KS) - \log\det(K)\|_1 \leq \epsilon, \text{ where } \epsilon \in \mathbb{R}^+.$$

Thus, we can see that these two estimators have the same purpose of estimating the precision matrix of our return portfolio, but with different constraints.

CLIME method has following advantages:

- **Suitable for High-Dimensional Data**: The CLIME method can effectively handle high-dimensional datasets, especially when the number of variables is much greater than the number of samples (i.e., in high-dimensional, small-sample situations).
- **Column-Wise Solution**: The CLIME method solves the precision matrix column by column. By addressing multiple independent L_1-regularized optimization problems, it achieves higher computational efficiency.

Remark 4.1. In practical applications, the performance of the CLIME method may be inferior to that of the GLASSO method; this can be seen in our simulation experiments later in our paper. In fact, there may be a significant amount of information contained in λ of the equation (6), and selecting an appropriate λ can be challenging. This may also be related to our experimental setup: constrained by our equipment, we are unable to truly simulate high-dimensional scenarios.

4.3 Factors Models

Consider the model that give the return for day t by a linear combination of a collection of latent factor $f_{k,t}$ for $1 \leq k \leq K$. The returns are modeled as:

$$r_{i,t} = \alpha_i + \beta_i^T f_t + u_{i,t}, \text{where } f_t = (f_{1,t}, ..., f_{K,t}),$$

where $u_{i,t}$ is the idiosyncratic error term for asset i that is uncorrelated with f_t.

This model follow the fact of linear regression because for factors α and β can be determined by applying the linear regression. Let B be the matrix whose i-th column is β_i, $\Sigma_{u,t} = cov(u_{i,t})$ and $\Sigma_{f,t} = cov(f_t)$. Then our covariance matrix is:

$$\Sigma_t = B^T \Sigma_{f,t} B + \Sigma_{u,t}.$$

In this model, we have several types of factor models, such as the *dynamic factor model* shown in the above equation. The *static factor model* assumes that the covariance of $u_{i,t}$ and f_t is time-invariant. The *exact factor model* is characterized by a diagonal $\Sigma_{u,t}$. The *approximate factor model* assumes that Σ_u is bounded in L_1 and L_2 norms.

Remark 4.2. Compared to the models above, in some literature, you will see the famous Fama-French factor model [FF93]. There are two types, and their expressions are as follows:

- Fama-French 3-factor model,

$$E(R_i) = R_f + \beta_{i,M} \cdot (E(R_M) - R_f) + \beta_{i,SMB} \cdot SMB + \beta_{i,HML} \cdot HML,$$

- Fama-French 5-factor model,

$$\begin{aligned} E(R_i) =& R_f + \beta_{i,M} \cdot (E(R_M) - R_f) + \beta_{i,SMB} \cdot SMB + \beta_{i,HML} \cdot HML \\ &+ \beta_{i,RMW} \cdot RMW + \beta_{i,CMA} \cdot CMA, \end{aligned}$$

where $\beta_{i,X}$ represents the sensitivity of the stock i to factor X, where X can be the market (M) [4], size (SMB) [5], value (HML) [6], profitability (RMW) [7], or investment style (CMA) [8].

In the article only focus in *static factor model* and the following static factor-based covariance matrix estimators are popularly used in financial applications :

- **POET** : We apply the SVD by truncated rank K of sample covariance $\hat{\Sigma}$ to estimate $B^T \Sigma_f B$, and soft-thresholding the off-diagonal entries of the residual covariance matrix.
- **EFM** : $\hat{\Sigma}_f$ equals the sample covariance matrix of the factors f_t and $\hat{\Sigma}_u$ equal to diagonal of Σ_u

[4] Market Factor is the excess return of a broad market portfolio over the risk-free rate.

[5] SMB stands for "Small Minus Big" and represents the additional return investors can expect from investing in companies with a smaller market capitalization over those with a larger market capitalization.

[6] HML stands for "High Minus Low" and represents the additional return investors can expect from investing in companies with a high book-to-market ratio (value companies) over those with a low book-to-market ratio (growth companies).

[7] RMW for Robust Minus Weak: The excess return of stocks of companies with robust operating profitability over those with weak profitability, capturing the profitability effect.

[8] CMA for Conservative Minus Aggressive: The excess return of stocks of companies with conservative investment policies over those with aggressive investment policies, capturing the investment style effect.

- **AFM-POET** : $\hat{\Sigma}_f$ obtained by **EFM** ,whereas $\hat{\Sigma}_u$ by soft-thresholding from **POET**

Remark 4.3. There are some details about **POET**:

(a) When we approximate the given simple covariance by K-truncated SVD, all singular values are placed in order. Therefore, by choosing the first K-values, it can minimize the noisy data (Noise-reduction) which is for the low-rank approximation. By retaining the top K values and their corresponding singular vectors, it captures the most significant linear relationships in the data, effectively summarizing a large number of variables through a few factors.

(b) For the threshold method, it acts to control the number of false discoveries (non-zero entries) in the error matrix, directly affecting the model's complexity and overfitting tendencies. Statistically, this helps to reduce the expected number of falsely identified significant correlations among the residuals, effectively improving model robustness and interpretability.

4.4 Linear Shrinkage

Let S be our sample covariance matrix, and λ_i is the i-th eigenvalue of S and v_i the corresponding eigenvector. The eigen-decomposition of the sample covariance matrix S is

$$S = \mathbb{V}\Lambda\mathbb{V}^T,$$

where $\mathbb{V}$ is matrix of eigenvectors, and Λ is matrix of eigenvalues of S. Modify Λ to Λ' to adapt the our shrinkage:

$$\Lambda' = (1-\rho)\Lambda + \rho\bar{\lambda}\mathbf{I},$$

where λ is average of the eigenvalue, and $0 \leq \rho \leq 1$ is a tuning parameter that determines the amount of shrinkage. So our linear shrinkage estimator becomes

$$\hat{\Sigma}_{LS} = \mathbb{V}\Lambda'\mathbb{V}^T,$$

and

$$\begin{aligned}
\mathbf{MSE}(\Sigma_{LS}) &= \mathbb{E}(\|\Sigma_{LS} - \Sigma\|^2) \\
&= \mathbb{E}(\|\mathbb{V}(\Lambda' - \Lambda)\mathbb{V}^T\|^2) \\
&= \mathbb{E}(\|\mathbb{V}(\rho\bar{\lambda}I - \rho\Lambda)\mathbb{V}^T\|^2) \\
&= \mathbb{E}(\|\rho\bar{\lambda}I - \rho\mathcal{S}\|^2) \\
&= (\rho\bar{\lambda})^2 I - 2\rho^2\bar{\lambda}\mathbb{E}(\|S\|) + \rho^2\mathbb{E}(\|S\|^2) \\
&\leq \rho^2(\bar{\lambda}^2 I + E(\|\mathcal{S}\|^2)).
\end{aligned}$$

With a well chosen ρ, we can minimize the value of MSE. And all the same thing error of residue, and by applying the shinkage it can bring the extreme values to be around the means which means to reduce the Standard Deviation (std).

Remark 4.4. An extension of linear shrinkage is *non-linear shrinkage*, which is related to the Marcenko-Pastur distribution [LW12]. Ledoit and Wolf constructed a non-linear shrinkage estimator based on this. We provide a brief summary of their results in the appendix. Their paper is much more sophisticated than the one we discuss here, and we strongly recommend that interested readers refer to their original work.

5 Heavy-Tailed Distribution

This section requires an enormous amount of data, which our computer is completely incapable of handling for the relevant data analysis; we will only describe the related theory here.

Stock returns may be heavy tailed. The tails of the data distribution decay more slowly than those of a normal distribution, meaning that the probability of extreme values occurring in the tails is relatively higher. In fact, market news, political events, and the release of economic data can all cause significant fluctuations in stock prices. This is often accompanied by irrational behavior from investors, and financial leverage can amplify gains or losses, which can also lead to heavy-tailed distributions, such as t-distribution.

Remark 5.1. The Gaussian distribution does not belong to the heavy-tailed distributions.

The following concepts will be used for the extension of the MTP_2 method.

Definition 5.1. (a) A random vector X with density function $p(x)$, mean $\mu \in \mathbb{R}^M$ and covariance matrix $\Sigma \in \mathbb{R}^{M\times M}$ follows an *elliptical distribution* if its density function can be expressed as

$$g\left((x-\mu)^T\Sigma^{-1}(x-\mu)\right)$$

for some function g.

(b) X follows a *transelliptical distribution* if there exist monotonically increasing functions $f_i, i = 1, \ldots, M$, such that $(f_1(X_1), \cdots, f_M(X_M))$ follows an elliptical distribution.

We denote the covariance matrix of this elliptical distribution by Σ_f. The following result provides a necessary condition for a transelliptical distribution to be MTP$_2$.

Theorem 5.1 ([ARU20]). Suppose that the joint distribution of $(X_1, \cdots, X_M)$ is MTP_2 and transelliptical, i. e. , there exist increasing functions $f_i, i = 1, \ldots, M$, such that the density function of $(f_1(X_1), \cdots, f_M(X_M))$ can be written as $g\left((x-\mu)^T\Sigma_f^{-1}(x-\mu)\right)$. Then, Σ_f^{-1} is an M-matrix.

Given the above theorem, it is clear that for transelliptical distributions, the constraint that Σ^{-1} be an M-matrix is a relaxation of MTP_2.

Remark 5.2. To overcome the restrictive Gaussian assumption, recent work suggested replacing the sample covariance matrix S in Eq (5) and Eq (6) by Kendall's tau correlation matrix S_τ with $(S_\tau)_{ij} := \sin\left(\frac{\pi}{2}\hat{\tau}\right)$, where

$$\hat{\tau}_{ij} := \frac{1}{\binom{T}{2}} \sum_{1\le t\le t'\le T} \operatorname{sign}(X_{it} - X_{it'}) \operatorname{sign}\left(X_{jt} - X_{jt'}\right).$$

This method is also applicable to heavy-tailed distributions without a reduction in efficiency [LHZ12] [BK18].

6 Implementation

In this section, we will implement all of these estimators in Python.

We will use the data presented in the article, and we got :

- **ret** : 10344×3251 matrix; with daily stock returns;
- **rf** : 10344×1 vector with daily risk-free returns.

The following 2 data are in percentage.

- **investDate** : 360×1 vector with monthly investment dates (index set of mydate);
- **topMV95** : 360×1000 matrix that identifies the stocks in the investment universe.

And at investment data $h \in \{1, \ldots, 360\}$, you would do something like is for investing in N stocks and using a history of length T to estimate the covariance matrix:

- universe = topMV95(h,1:N);
- today = investDate(h);
- pastPeriod = (-T:-1)+today;
- pastRet = ret(pastPeriod,universe).

The corresponding out-of-sample returns of the stocks in the universe are given by:

- investPeriod = today(today+20),
- outRet = ret(investPeriod,universe).

For the definition of our data, we will use the code presented in the article, which provides the construction and organization of our datasets. We will compute only the mean, standard deviation (std), and mean/std for each estimator. Due to the performance limitations of our computer, we will choose only three cases: $T = 100, N = 25$; $T = 50, N = 50$; and $T = 25, N = 100$. Below are our results:

	POET(k=3)	**POET(k=5)**	**LS**	**NLS**	**CLIME**	**MTP2**	**GLASSO**
avg	10. 036	10. 724	9. 885	13. 627	nan	nan	12. 322
std	13. 029	12. 891	13. 807	19. 650	nan	12. 15	13. 807
avg/std	0. 770	0. 832	0. 716	0. 693	nan	nan	0. 716

Table 1: T = 25 , N = 100

	POET(k=3)	**POET(k=5)**	**LS**	**NLS**	**CLIME**	**MTP2**	**GLASSO**
avg	7. 728	9. 459	8. 108	11. 125	11. 84	nan	10. 687
std	14. 020	13. 881	14. 580	21. 278	27. 84	12. 31	14. 188
avg/std	0. 551	0. 681	0. 556	0. 523	0. 43	nan	0. 753

Table 2: T = 50 , N = 50

	POET(k=3)	**POET(k=5)**	**LS**	**NLS**	**CLIME**	**MTP2**	**GLASSO**
avg	8. 761	7. 633	9. 946	9. 301	8. 80	nan	10. 538
std	15. 888	17. 050	15. 766	16. 004	16. 89	12. 38	15. 162
avg/std	0. 551	0. 448	0. 631	0. 581	0. 52	nan	0. 695

Table 3: T = 100 , N = 25

We work in both R and Python because for some estimators, and there are direct functions available in R.

For the *CLIME* estimator, we can reduce its standard deviation by choosing a good λ. However, in **HD**, we encounter problems where we can't compute *CLIME* due to the small value of λ chosen. As mentioned before, this is due to the constraint computation, which exceeds our tuning parameter.

Our Python code can be found here.

7 Conclusion

All of these estimators are designed to estimate the best possible precision matrix, especially in **HD** (High Dimension), as our *std* is reduced by increasing our dimension. However, we can see that the new proposed method, MTP_2, leads to a better estimator by choosing only the positive covariance matrices, which minimizes the risk and maximizes the ideal return.

References

[ARU20] Raj Agrawal, Uma Roy, and Caroline Uhler. Covariance Matrix Estimation under Total Positivity for Portfolio Selection*. *Journal of Financial Econometrics*, 20(2):367–389, 09 2020.

[BK18] Rina Foygel Barber and Mladen Kolar. Rocket: Robust confidence intervals via kendall's tau for transelliptical graphical models. *The Annals of Statistics*, 46(6B):3422–3450, 2018.

[CSS05] Antonio Colangelo, Marco Scarsini, and Moshe Shaked. Some notions of multivariate positive dependence. *Insurance: Mathematics and Economics*, 37, 2005.

[FF93] Eugene F. Fama and Kenneth R. French. Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33(1):3–56, 1993.

[JM03] Ravi Jagannathan and Tongshu Ma. Risk reduction in large portfolios: Why imposing the wrong constraints helps. *Journal of Finance*, 58, 2003.

[KR80] Samuel Karlin and Yosef Rinott. Classes of orderings of measures and related correlation inequalities. i. multivariate totally positive distributions. *Journal of Multivariate Analysis*, 10(4):467–498, 1980.

[LHZ12] Han Liu, Fang Han, and Cun-hui Zhang. Transelliptical graphical models. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

[LUZ19] Steffen Lauritzen, Caroline Uhler, and Piotr Zwiernik. Maximum likelihood estimation in gaussian models under total positivity. *Annals of Statistics*, 47, 2019.

[LW12] Olivier Ledoit and Michael Wolf. Nonlinear shrinkage estimation of large-dimensional covariance matrices. *The Annals of Statistics*, 40(2):1024 – 1060, 2012.

[SH15] Martin Slawski and Matthias Hein. Estimation of positive definite m-matrices and structure learning for attractive gaussian markov random fields. *Linear Algebra and its Applications*, 473:145–179, 2015. Special issue on Statistics.

A Proof of the Remark 3.4

We will prove that the following equation has no upper bound:

$$\mathcal{L}(D, K) = \frac{T}{2}\log(\det(K)) - \frac{T}{2}\text{trace}(KS), \quad \text{where } S = \frac{1}{T}\sum_{t=1}^{T} r_t^T r_t.$$

In fact, when $N \geq T$, the number of samples T is less than or equal to the dimension N, and the rank of the sample covariance matrix S is at most T. This means that S is not a full-rank matrix and has at least $N - T$ eigenvalues equal to zero.

Assume the eigenvalue decomposition of S is

$$S = U\Lambda U^T,$$

where U is an orthogonal matrix whose columns are the eigenvectors of S, and Λ is a diagonal matrix whose diagonal elements are the eigenvalues of S, with at least $N - T$ of them being zero.

Similarly, assume the eigenvalue decomposition of K is

$$K = V\Sigma V^T,$$

where V is an orthogonal matrix whose columns are the eigenvectors of K, and Σ is a diagonal matrix whose diagonal elements are the eigenvalues of K.

Now consider trace(KS):

$$\text{trace}(KS) = \text{trace}(KU\Lambda U^T) = \text{trace}(V\Sigma V^T U\Lambda U^T).$$

Since the trace operation is invariant under cyclic permutations, we have:

$$\text{trace}(KS) = \text{trace}(\Sigma V^T U\Lambda U^T V).$$

Define the matrix $W = V^T U$. Since both V and U are orthogonal matrices, W is also an orthogonal matrix:

$$\text{trace}(KS) = \text{trace}(\Sigma W\Lambda W^T).$$

Since Λ is a diagonal matrix with at least $N - T$ zero eigenvalues, we can write Λ in a block diagonal form:

$$\Lambda = \begin{pmatrix} \Lambda_1 & 0 \\ 0 & 0 \end{pmatrix},$$

where Λ_1 is a $T \times T$ diagonal matrix containing the non-zero eigenvalues. Correspondingly, we can block partition W and Σ as well:

$$W = \begin{pmatrix} W_1 & W_2 \\ W_3 & W_4 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{pmatrix}.$$

Thus, we have

$$W\Lambda W^T = \begin{pmatrix} W_1\Lambda_1 W_1^T & 0 \\ 0 & 0 \end{pmatrix}.$$

Therefore,

$$\Sigma W \Lambda W^T = \begin{pmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{pmatrix} \begin{pmatrix} W_1 \Lambda_1 W_1^T & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} \Sigma_1 W_1 \Lambda_1 W_1^T & 0 \\ 0 & 0 \end{pmatrix}.$$

Finally, trace(KS) becomes:

$$\text{trace}(KS) = \text{trace}(\Sigma W \Lambda W^T) = \text{trace}(\Sigma_1 W_1 \Lambda_1 W_1^T) + \text{trace}(0) = \text{trace}(\Sigma_1 W_1 \Lambda_1 W_1^T).$$

Notice that the values of Σ_2 do not affect the value of trace(KS). This means we can arbitrarily increase the eigenvalues in Σ_2 without affecting trace(KS).

B Proof of the Theorem 3.1

The proof is divided into two steps.

($\Rightarrow$) We need to prove

$$(x^T \Sigma^{-1} x) + (y^T \Sigma^{-1} y) \geq ((x \wedge y)^T \Sigma^{-1} (x \wedge y)) + ((x \vee y)^T \Sigma^{-1} (x \vee y)).$$

Write

$$\Sigma^{-1} = sI - B, \text{ where } B = (b_{ij}) \text{ with } b_{ij} \geq 0, \text{ for all } 1 \leq i, j \leq n,$$

then the above inequality is equivalent to:

$$sx^T x + sy^T y - x^T Bx - y^T By \geq s(x \wedge y)^T (x \wedge y) + s(x \vee y)^T (x \vee y) - (x \wedge y)^T B (x \wedge y) - (x \vee y)^T B (x \vee y).$$

Since

$$sx^T x + sy^T y = s(x \wedge y)^T (x \wedge y) + s(x \vee y)^T (x \vee y),$$

the above inequality is equivalent to

$$(x \wedge y)^T B (x \wedge y) + (x \vee y)^T B (x \vee y) - x^T Bx - y^T By \geq 0.$$

If $B \in \mathbb{R}$, then the result is trivial.

Assume that the above inequality holds for all $B \in \mathbb{R}^{(n-1) \times (n-1)}$. Now, let $B \in \mathbb{R}^{n \times n}$. Rewrite

$$B = \begin{pmatrix} b & c^T \\ c & B_0 \end{pmatrix} \text{ where } b \in \mathbb{R}, c \in \mathbb{R}^{(n-1) \times 1} \text{ and } B_0 \in \mathbb{R}^{(n-1) \times (n-1)}$$

and

$$x = \begin{pmatrix} x' \\ x_0 \end{pmatrix}, \quad y = \begin{pmatrix} y' \\ y_0 \end{pmatrix}, \quad x \wedge y = \begin{pmatrix} \alpha \\ x_0 \wedge y_0 \end{pmatrix}, \quad x \vee y = \begin{pmatrix} \beta \\ x_0 \vee y_0 \end{pmatrix},$$

then

$$\begin{aligned} x^T Bx &= x' b x' + 2x' c^T x_0 + x_0^T B_0 x_0, \\ y^T By &= y' b y' + 2y' c^T y_0 + y_0^T B_0 y_0, \\ (x \wedge y)^T B (x \wedge y) &= \alpha b \alpha + 2\alpha c^T (x_0 \wedge y_0) + (x_0 \wedge y_0)^T B_0 (x_0 \wedge y_0), \\ (x \vee y)^T B (x \vee y) &= \beta b \beta + 2\beta c^T (x_0 \vee y_0) + (x_0 \vee y_0)^T B_0 (x_0 \vee y_0). \end{aligned}$$

Since the inequality holds for all matrices in $\mathbb{R}^{(n-1) \times (n-1)}$, we have

$$\begin{aligned} &(x_0 \wedge y_0)^T B (x_0 \wedge y_0) + (x_0 \vee y_0)^T B (x_0 \vee y_0) - x_0^T B x_0 - y_0^T B y_0 \geq 0, \\ &x' b x' + y' b y' - \alpha b \alpha - \beta b \beta = 0. \end{aligned}$$

Therefore, we only need to prove

$$\alpha c^T(x_0 \wedge y_0) + \beta c^T(x_0 \vee y_0) - x'c^T x_0 - y'c^T y_0 \geq 0.$$

Note that $\alpha \leq \beta$.

If $x' = \alpha$ and $y' = \beta$, then the above equation is obviously true, because

$$\alpha c^T[\mathbf{0}_{n-1} \wedge (y_0 - x_0)] + \beta c^T[(x_0 - y_0) \vee \mathbf{0}_{n-1}] \geq 0.$$

If $x' = \beta$ and $y' = \alpha$, then the equation still holds, because

$$\alpha c^T[(x_0 - y_0) \wedge \mathbf{0}_{n-1}] + \beta c^T[\mathbf{0}_{n-1} \vee (y_0 - x_0)] \geq 0.$$

($\Leftarrow$) Based on the previous proof, the statement is evident.

C Non Linear Shrinkage

We still strongly recommend that readers refer to the original text [LW12].

Let n denote the sample size and $p \equiv p(n)$ the number of variables, with $p/n \longrightarrow c \in (0,1)$ as $n \longrightarrow \infty$. Assume:

(a) The population covariance matrix $\Sigma_n \in \mathbb{R}^{p\times p}$ is a nonrandom positive definite matrix.

(b) Let $X_n \in \mathbb{R}^{n\times p}$ be a matrix of real independent and identically distributed (i.i.d.) random variables with zero mean and unit variance. One only observes $Y_n \equiv X_n \Sigma_n^{1/2}$, so neither X_n nor Σ_n are observed on their own.

(c) Let $(\tau_{n,1}, \ldots, \tau_{n,p})$ denotes a system of eigenvalues of Σ_n. The empirical distribution function (e.d.f.) of the population eigenvalues is defined as

$$\forall t \in \mathbb{R}, H_n(t) \equiv p^{-1} \sum_{i=1}^{p} \mathbf{1}_{[\tau_{n,i},+\infty)}(t),$$

where $\mathbf{1}$ denotes the indicator function of a set. We assume $H_n(t)$ converges to some limit $H(t)$ at all points of continuity of H.

(d) Supp(H), the support of H, is the union of a finite number of closed intervals, bounded away from zero and infinity. Furthermore, there exists a compact interval in $(0, +\infty)$ that contains Supp (H_n) for all n large enough.

Let $((\lambda_{n,1}, \ldots, \lambda_{n,p}); (u_{n,1}, \ldots, u_{n,p}))$ denote a system of eigenvalues and eigenvectors of the sample covariance matrix

$$S_n \equiv n^{-1} Y_n' Y_n = n^{-1} \Sigma_n^{1/2} X_n' X_n \times \Sigma_n^{1/2}.$$

The e.d.f. of the sample eigenvalues is defined as

$$F_n(\lambda) \equiv p^{-1} \sum_{i=1}^{p} \mathbf{1}_{[\lambda_i,+\infty)}(\lambda), \text{ for } \forall \lambda \in \mathbb{R}, .$$

The **Stieltjes transform** of a nondecreasing function G is defined by

$$m_G(z) \equiv \int_{-\infty}^{+\infty} \frac{1}{\lambda - z} dG(\lambda), \text{ for } \forall z \in \mathbb{C}^+,$$

where $\mathbb{C}^+$is the half-plane of complex numbers with strictly positive imaginary part. The Stieltjes transform of the e.d.f. of sample eigenvalues is

$$m_{F_n}(z) = \frac{1}{p}\sum_{i=1}^{p}\frac{1}{\lambda_i - z} = \frac{1}{p}\operatorname{Tr}\left[(S_n - zI)^{-1}\right], \text{ for } \forall z \in \mathbb{C}^+,$$

where I denotes a conformable identity matrix.

$F_n(\lambda)$ converges almost surely (a.s.) to some nonrandom limit $F(\lambda)$ at all points of continuity of F under certain sets of assumptions.

The most convenient expression of the Marčenko-Pastur equation is

$$m_F(z) = \int_{-\infty}^{+\infty}\frac{1}{\tau\left[1 - c - czm_F(z)\right] - z}dH(\tau), \text{ for } \forall z \in \mathbb{C}^+.$$

In addition, Silverstein and Choi (1995) showed that

$$\text{given } \forall\lambda \in \mathbb{R} - \{0\}, \quad \lim_{z\in\mathbb{C}^+\to\lambda} m_F(z) \equiv \breve{m}_F(\lambda)$$

exists, and that F has a continuous derivative $F' = \pi^{-1}\operatorname{Im}\left[\breve{m}_F\right]$ on all of $\mathbb{R}$ with $F' \equiv 0$ on $(-\infty, 0]$.

Let the linear operator L transform any c.d.f. G into $LG(x) \equiv \int_{-\infty}^{x}\tau dG(\tau)$. Combining L with the Stieltjes transform, we get

$$m_{LG}(z) = \int_{-\infty}^{+\infty}\frac{\tau}{\tau - z}dG(\tau) = 1 + zm_G(z).$$

In the absence of specific information about the true covariance matrix Σ_n, it appears reasonable to restrict attention to the class of estimators that are equivariant with respect to rotations of the observed data.

Definition C.1. Let W be an arbitrary p-dimensional orthogonal matrix. Let $\widehat{\Sigma}_n \equiv \widehat{\Sigma}_n\left(Y_n\right)$ be an estimator of Σ_n. Then the estimator is said to be **rotation-equivariant** if it satisfies $\widehat{\Sigma}_n\left(Y_n W\right) = W'\widehat{\Sigma}_n\left(Y_n\right)W$.

Every rotation-equivariant estimator is thus of the form

$$U_n A_n U_n', \quad \text{where } A_n \equiv \operatorname{Diag}\left(a_1, \ldots, a_p\right) \text{ is diagonal,}$$

and where U_n is the matrix whose i th column is the sample eigenvector $u_i \equiv u_{n,i}$.

Under the Frobenius norm ($\|A\| \equiv \sqrt{\operatorname{Tr}\left(AA'\right)/r}, A \in \mathbb{R}^{r\times m}$), we end up with the following minimization problem:

$$\min_{D_n}\left\|U_n A_n U_n' - \Sigma_n^{-1}\right\|.$$

Elementary matrix algebra shows that its solution is

$$A_n^* \equiv \operatorname{Diag}\left(a_1^*, \ldots, a_p^*\right), \quad \text{where } a_i^* \equiv u_i'\Sigma_n u_i \text{ for } i = 1, \ldots, p.$$

As a result, the finite-sample optimal estimator is given by

$$P_n^* \equiv U_n A_n^* U_n', \quad \text{where } A_n^* \text{ is defined as the above.}$$

Ledoit and Péché (2011) show that d_i^* can be approximated by the quantity

$$a_i^{or} \equiv \lambda_i^{-1} \left(1 - c - 2c\lambda_i \operatorname{Re}\left[\breve{m}_F\left(\lambda_i\right)\right]\right) \text{ for } i = 1, \ldots, p.$$

from which they deduce their oracle estimator

$$P_n^{or} \equiv U_n A_n^{or} U_n', \quad \text{where } A_n^{or} \equiv \operatorname{Diag}\left(d_1^{or}, \ldots, d_p^{or}\right).$$

One can see that the oracle estimator P_n^{or} involves the unknown quantities $\breve{m}_F\left(\lambda_i\right)$, for $i = 1, \ldots, p$. In fact, for every λ in the interior of Supp(F), there exists a unique $v \in \mathbb{C}^+$, denoted by v_λ, such that

$$v_\lambda - c v_\lambda m_{LH}\left(v_\lambda\right) = \lambda, \quad \text{and} \quad \breve{m}_F(\lambda) = \frac{1-c}{c\lambda} - \frac{1}{c}\frac{1}{v_\lambda}.$$