Robust Factor Analysis Parameter Estimation*

Rui Zhou, Junyan Liu, Sandeep Kumar, and Daniel P. Palomar

Department of Electronic and Computer Engineering
The Hong Kong University of Science and Technology, Hong Kong
{rui.zhou, jliubl}@connect.ust.hk, {eesandeep, palomar}@ust.hk

Abstract. This paper considers the problem of robustly estimating the parameters of a heavy-tailed multivariate distribution when the covariance matrix is known to have the structure of a low-rank matrix plus a diagonal matrix as considered in factor analysis (FA). By assuming the observed data to follow the multivariate Student's t distribution, we can robustly estimate the parameters via maximum likelihood estimation (MLE). However, the MLE of parameters becomes an intractable problem when the multivariate Student's t distribution and the FA structure are both introduced. In this paper, we propose an algorithm based on the generalized expectation maximization (GEM) method to obtain estimators. The robustness of our proposed method is further enhanced to cope with missing values. Finally, we show the performance of our proposed algorithm using both synthetic data and real financial data.

Keywords: Robust parameter estimation · Factor Analysis · Student's t · Generalized expectation maximization · Missing values.

1 Introduction

Factor analysis (FA) is of great significance in various fields like finance, statistics, and cognitive ratio [11,14]. A basic FA model can be written as $\mathbf{x} = \boldsymbol{\mu} + \mathbf{B}\mathbf{f} + \boldsymbol{\varepsilon}$, where $\mathbf{x} \in \mathbb{R}^p$ is the observed vector, $\boldsymbol{\mu} \in \mathbb{R}^p$ is a constant vector, $\mathbf{B} \in \mathbb{R}^{p\times r}$ $(r \ll p)$ is the factors loading matrix, $\mathbf{f} \in \mathbb{R}^r$ is a vector of low-dimensional common factors, and $\boldsymbol{\varepsilon} \in \mathbb{R}^p$ is a vector of uncorrelated noise. For example, in a financial market, $\mathbf{x}$ can be the return of stocks, and $\mathbf{f}$ can be some macroeconomic factors like growth rate of the GDP, inflation rate, unemployment rate, etc. [2]. FA model typically assumes that $\mathbf{f}$ and $\boldsymbol{\varepsilon}$ are uncorrelated and both zero-mean, and the covariance matrix of $\mathbf{f}$ is an $r \times r$ identity matrix, denoted by $\mathbf{I}_r$. Following this, the covariance matrix of $\mathbf{x}$ can be expressed as $\boldsymbol{\Sigma} = \mathbf{B}\mathbf{B}^T + \boldsymbol{\Psi}$, where $\boldsymbol{\Psi}$ is a $p \times p$ diagonal matrix containing the variance of noise at its diagonal. Note that, with the FA structure, the number of parameters of the covariance matrix has been greatly reduced from $p\,(p+1)\,/2$ to $p\,(r+1)$. Therefore, the estimation of $\boldsymbol{\Sigma}$ could be improved due to the FA structure.

1.1 Related Works

Learn From $\boldsymbol{\Sigma}$: A large amount of literature has focused on estimating the covariance matrix with FA structure. One choice is decomposing the estimated $\boldsymbol{\Sigma}$ into the sum of a low-rank matrix and a diagonal matrix, e.g., constrained

* This work was supported by the Hong Kong RGC 16208917 research grant.

minimum trace factor analysis (MTFA) [13]. But a main drawback is imposing the hard equality constraint $\mathbf{\Sigma} = \mathbf{B}\mathbf{B}^T + \mathbf{\Psi}$, which does not allow any differences between $\mathbf{\Sigma}$ and $\mathbf{B}\mathbf{B}^T + \mathbf{\Psi}$. Its application is restricted as the exact $\mathbf{\Sigma}$ is usually not observed. Another choice is to approximate the target matrix by a FA structured one. A naive method is to obtain $\hat{\mathbf{B}}$ firstly via principal component analysis (PCA) and then $\hat{\mathbf{\Psi}}$ by taking directly the residual's sample variance. A joint estimation method over $\mathbf{B}$ and $\mathbf{\Psi}$ is usually chosen to minimize $\|\mathbf{\Sigma} - \left(\mathbf{B}\mathbf{B}^T + \mathbf{\Psi}\right)\|_F^2$, where $\|\cdot\|_F$ denotes the Frobenius norm of a matrix. This problem can be solved by applying PCA iteratively [11].

Learn From Data: Different from works mentioned above, the MLE for FA directly learns the parameters from raw data. It assumes that the data are generated from a certain statistical model, typically the multivariate Gaussian distribution, and then the parameters are estimated by maximizing the likelihood function. However, a disadvantage of the estimators under the Gaussian assumption is sensitiveness to outliers [2]. A popular way to obtain a more robust estimation result is to consider some robust heavy-tailed distribution, such as multivariate Student's t or multivariate Skew t [15,16] instead of Gaussian. The two aforementioned methods both assume that $\mathbf{f}$ and $\boldsymbol{\varepsilon}$ follow the same heavy tail distribution sharing the same degrees of freedom. As $\mathbf{f}$ and $\boldsymbol{\varepsilon}$ are not observed, such an assumption is very restrictive and difficult to verify in practice.

1.2 Contributions

This paper considers more general and practically verifiable assumptions on the FA model: we only assume that the observation $\mathbf{x}$ follows a multivariate Student's t distribution and the FA structure with no additional restrictions on $\mathbf{f}$ and $\boldsymbol{\varepsilon}$. For this more general model, we propose an efficient algorithm to estimate the parameters based on the generalized expectation maximization (GEM) [1] method. In addition, we use the PX-EM [7] method to accelerate the GEM. Our proposed algorithm can be easily extended to other situations, e.g., when observed data contains missing values [6] or when it follows the multivariate Skew t distribution. With synthetic data, our proposed algorithm shows great estimation accuracy and robustness to outliers and missing data, which is very meaningful in practical applications. We also consider real market data in the numerical results, where the global minimum variance portfolio is designed using our estimator and compared with those using other estimators.

2 Gaussian FA Problem

Given the sample covariance matrix of the observed data as $\mathbf{S}$, the MLE problem for FA under the Gaussian distribution assumption (GFA) is given as below:

$$\begin{aligned} \underset{\mathbf{\Sigma},\mathbf{B},\mathbf{\Psi}}{\text{maximize}} \quad & \log|\mathbf{\Sigma}^{-1}| - \operatorname{Tr}\left(\mathbf{\Sigma}^{-1}\mathbf{S}\right) \\ \text{subject to} \quad & \mathbf{\Sigma} = \mathbf{B}\mathbf{B}^T + \mathbf{\Psi} \\ & \mathbf{\Psi} = \operatorname{Diag}\left(\psi_1, \ldots, \psi_p\right) \succ \mathbf{0}. \end{aligned} \tag{1}$$

The solution to problem (1) would be $\mathbf{\Sigma}^\star = \mathbf{S}$ if the structure constraints were ignored, but becomes intractable when the FA structure is imposed. Here we introduce two algorithms for solving problem (1).

Alternate Algorithm: Problem (1) can be solved by an alternating optimization approach, which can be performed by alternately updating $\mathbf{B}$ and $\mathbf{\Psi}$ (but not equivalent to the alternating optimization over $\mathbf{B}$ and $\mathbf{\Psi}$), cf. [9,10]. For fixed $\mathbf{\Psi}$, the optimal update for $\mathbf{B}$ is given next.

Lemma 1. *([9]) Given a feasible $\mathbf{\Psi}$, the optimal $\mathbf{B}^\star$ maximizing problem (1) is $\mathbf{B}^\star = \mathbf{\Psi}^{\frac{1}{2}}\mathbf{U}\mathbf{D}^{\frac{1}{2}}$ where $\mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$ is the eigenvalue decomposition (EVD) of $\mathbf{\Psi}^{-\frac{1}{2}}\mathbf{S}\mathbf{\Psi}^{-\frac{1}{2}}$ and $\mathbf{D} = Diag\left(d_1, \ldots, d_r, 0, \ldots, 0\right)$ with $d_i = \max\left(\lambda_i - 1, 0\right)$.*
For fixed $\mathbf{B}$, the update for $\mathbf{\Psi}$ is set as $\text{Diag}\left(\mathbf{S} - \mathbf{B}\mathbf{B}^T\right)$.

MM Algorithm: Recently, a majorization-minimization (MM) [12] based method has been proposed in [4] to obtain the optimal $\mathbf{\Sigma}$. Plugging the optimal $\mathbf{B}^\star$ from Lemma 1 in (1), we can achieve a concentrated version of (1).

Lemma 2. *([4]) Denoting $\mathbf{\Phi} = \mathbf{\Psi}^{-1}$, the problem (1) is equivalent to minimizing $f\left(\boldsymbol{\phi}\right)$, where $f\left(\boldsymbol{\phi}\right) = f_1\left(\boldsymbol{\phi}\right) - f_2\left(\boldsymbol{\phi}\right)$ with $f_1\left(\boldsymbol{\phi}\right) = \sum_{i=1}^{p}\left(-\log\phi_i + S_{ii}\phi_i\right)$, $f_2\left(\boldsymbol{\phi}\right) = -\sum_{i=1}^{r}\left(\log\left(\max\left\{1, \lambda_i^*\right\}\right) - \max\left\{1, \lambda_i^*\right\} + 1\right)$ and $\left\{\lambda_i^*\right\}_{i=1}^{r}$ are the top r eigenvalues of $\mathbf{S}^* = \mathbf{\Phi}^{\frac{1}{2}}\mathbf{S}\mathbf{\Phi}^{\frac{1}{2}}$. Besides, $f_1\left(\boldsymbol{\phi}\right)$ and $f_2\left(\boldsymbol{\phi}\right)$ are both convex in $\boldsymbol{\phi}$.*
By linearizing the $f_2\left(\boldsymbol{\phi}\right)$ using its sub-gradient, we can majorize $f\left(\boldsymbol{\phi}\right)$ by $\bar{f}\left(\boldsymbol{\phi}\right) = \sum_{i=1}^{p}\left(-\log\phi_i + S_{ii}\phi_i - \bigtriangledown_i^{(k)}\phi_i\right)$, where $\bigtriangledown^{(k)}$ is a subgradient of $f_2\left(\boldsymbol{\phi}\right)$ at the kth iteration. The $\bigtriangledown^{(k)}$ can be calculated as $\bigtriangledown_i^{(k)} = \left(\mathbf{\Phi}^{-\frac{1}{2}}\mathbf{U}^*\mathbf{D}_1\mathbf{U}^{*T}\mathbf{\Phi}^{\frac{1}{2}}\mathbf{S}\right)_{ii}$ where $\mathbf{U}^*\mathbf{\Lambda}^*\mathbf{U}^{*T}$ is the EVD of $\mathbf{\Phi}^{\frac{1}{2}}\mathbf{S}^{(k)}\mathbf{\Phi}^{\frac{1}{2}}$ and $\mathbf{D}_1 = \text{Diag}\left(d_1, \ldots, d_r, 0, \ldots 0\right)$ with $d_i = \max\left\{0, 1 - 1/\lambda_i^*\right\}$ [4,5]. Then the update of $\mathbf{\Phi}$ can be easily obtained as $\phi_i^{(k+1)} = \left(S_{ii}^{(k)} - \bigtriangledown_i^{(k)}\right)^{-1}$ for $i = 1, \ldots, p$. By iteratively taking the above procedure, we can get a converged sequence of $\boldsymbol{\phi}^{(k)}$. Finally we can set $\mathbf{\Psi} = \mathbf{\Phi}^{-1}$ and compute $\mathbf{B}$ via Lemma 1.

It should be noted that the two algorithms are the same and can be verified by matrix algebra.

3 Problem Statement

The p-dimensional multivariate Student's t distribution, denoted as $\boldsymbol{t}_p\left(\boldsymbol{\mu}, \mathbf{\Sigma}, \nu\right)$, has the probability density function (pdf)

$$f\left(\mathbf{x}|\boldsymbol{\theta}\right) = \frac{\Gamma\left(\frac{\nu+p}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\nu^{\frac{p}{2}}\pi^{\frac{p}{2}}\left|\mathbf{\Sigma}\right|^{\frac{1}{2}}}\left[1 + \frac{1}{\nu}\left(\mathbf{x} - \boldsymbol{\mu}\right)^T\mathbf{\Sigma}^{-1}\left(\mathbf{x} - \boldsymbol{\mu}\right)\right]^{-\frac{\nu+p}{2}} \tag{2}$$

where $\boldsymbol{\theta} = \left(\boldsymbol{\mu}, \mathbf{\Sigma}, \nu\right)$, ν is the degrees of freedom, $\mathbf{\Sigma}$ is the scale $p \times p$ positive definite matrix, $\boldsymbol{\mu}$ is the p-dimensional mean vector, and $\Gamma\left(a\right) = \int_0^\infty t^{(a-1)}\exp\left(-t\right)dt$ is the gamma function. Note that the covariance matrix of $\mathbf{x}$ is $\frac{\nu}{\nu-2}\mathbf{\Sigma}$, and it is not defined for $\nu \leq 2$. Interestingly, the above multivariate Student's t distribution can be represented in a hierarchical structure as $\mathbf{x}|\tau \overset{i.i.d}{\sim} \mathcal{N}_p\left(\boldsymbol{\mu}, \frac{1}{\tau}\mathbf{\Sigma}\right)$ with $\tau \overset{i.i.d}{\sim} \text{Gamma}\left(\frac{\nu}{2}, \frac{\nu}{2}\right)$, where $\mathcal{N}_p\left(\boldsymbol{\mu}, \mathbf{\Sigma}\right)$ is the multivariate Gaussian distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\mathbf{\Sigma}$. $\text{Gamma}\left(a, b\right)$ means gamma distribution with shape a and rate b, whose pdf is $f\left(\tau\right) = b^a\tau^{(a-1)}\exp\left(-b\tau\right)/\Gamma\left(a\right)$.

We consider that the observed p-dimensional data $\mathbf{x}_t, t = 1, \ldots, T$ follows the independent and identical distributed (i.i.d.) multivariate Student's t distribution, i.e., $\mathbf{x}_t \sim \boldsymbol{t}_p\left(\boldsymbol{\mu}, \mathbf{\Sigma}, \nu\right)$. Besides, we assume that $\mathbf{x}_t$ follows the FA model,

which means $\mathbf{\Sigma} = \mathbf{B}\mathbf{B}^T + \mathbf{\Psi}$. Note that we omit a scaling factor in order to simplify the notation (recall that here $\mathbf{\Sigma}$ refers to the scale matrix). A natural approach is to obtain the parameter estimation through MLE method, i.e., maximizing $L(\boldsymbol{\theta}|\mathbf{X}) = \sum_t \log f(\mathbf{x}_t|\boldsymbol{\theta})$ w.r.t. $\boldsymbol{\theta}$, where $\mathbf{X} \in \mathbb{R}^{T\times p}$ with $\mathbf{x}_t$ along the t-th row.

4 Problem Solution

It is very difficult to directly solve the above MLE problem as the objective function and constraints are both non-convex. The expectation maximization (EM) algorithm is a powerful iterative method to handle such problem [8]. By incorporating the latent data $\mathbf{Z}$, EM can be employed to convert the maximization for $L(\mathbf{X}|\boldsymbol{\theta})$ to the maximization for a sequence of simpler and solvable problems. In each iteration, it requires $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})$, which is the expected log-likelihood function of $L(\mathbf{X}|\boldsymbol{\theta})$ with respect to the current conditional distribution of $\mathbf{Z}$ given the $\mathbf{X}$ and the current estimate of the parameter $\boldsymbol{\theta}^{(k)}$. Then it finds $\boldsymbol{\theta}^{(k+1)}$ by maximizing $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})$. However, the computational cost of solving the subproblem might still be rather heavy and make the whole EM algorithm impractical. That is when the GEM algorithm can help. The GEM is an iterative method based on the EM philosophy but requiring an improvement at each iteration instead of a full maximization as in EM.

4.1 The RFA Algorithm

In this section, we propose a robust factor analysis (RFA) algorithm to solve the above problem. By incorporating the latent variables τ from the multivariate Student's t hierarchical structure, the log-likelihood function of the complete data $(\mathbf{X}, \boldsymbol{\tau})$ is given in (3). Note that $\boldsymbol{\tau}$ corresponds to $\mathbf{Z}$ in our application.

$$\begin{aligned} L(\boldsymbol{\theta}|\mathbf{x}, \boldsymbol{\tau}) = \frac{T}{2}\log|\mathbf{\Sigma}^{-1}| - \frac{1}{2}\mathrm{Tr}\left(\mathbf{\Sigma}^{-1}\sum_{t=1}^{T}\tau_t(\mathbf{x}_t - \boldsymbol{\mu})(\mathbf{x}_t - \boldsymbol{\mu})^T\right) \\ + \frac{T\nu}{2}\log\frac{\nu}{2} + \frac{\nu}{2}\sum_{t=1}^{T}(\log\tau_t - \tau_t) - T\log\Gamma\left(\frac{\nu}{2}\right) + const. \end{aligned} \tag{3}$$

Expectation Step: The expectation step of the GEM algorithm is to find the conditional expectation of $L(\boldsymbol{\theta}|\mathbf{X}, \boldsymbol{\tau})$ over $\boldsymbol{\tau}$ given the observed $\mathbf{X}$ and the current estimation of $\boldsymbol{\theta}$, i.e., $\boldsymbol{\theta}^{(k)}$. Since the conditional expectation of τ_t and $\log\tau_t$ for $t = 1, \ldots, T$ can be directly calculated as

$$\begin{aligned} e_{1,t}^{(k)} &= \mathrm{E}\left(\tau_t|\mathbf{x}_t, \boldsymbol{\theta}^{(k)}\right) = \frac{\nu^{(k)} + p}{\nu^{(k)} + d\left(\mathbf{x}_t, \boldsymbol{\mu}^{(k)}, \mathbf{\Sigma}^{(k)}\right)} \\ e_{2,t}^{(k)} &= \mathrm{E}\left(\log\tau_t|\mathbf{x}_t, \boldsymbol{\theta}^{(k)}\right) = \psi\left(\frac{\nu^{(k)} + p}{2}\right) - \log\frac{\nu^{(k)} + d\left(\mathbf{x}_t, \boldsymbol{\mu}^{(k)}, \mathbf{\Sigma}^{(k)}\right)}{2} \end{aligned}$$

where $d(\mathbf{x}, \boldsymbol{\mu}, \mathbf{\Sigma}) = (\mathbf{x} - \boldsymbol{\mu})^T(\mathbf{\Sigma})^{-1}(\mathbf{x} - \boldsymbol{\mu})$ is the Mahalanobis distance between $\mathbf{x}_t$ and $\boldsymbol{\mu}$ [6], then the expectation of the complete data log-likelihood (3) is

$$Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}\right) = \frac{T}{2}\log|\boldsymbol{\Sigma}^{-1}| - \frac{1}{2}\mathrm{Tr}\left(\boldsymbol{\Sigma}^{-1}\left(\sum_{t=1}^{T} e_{1,t}^{(k)}\left(\mathbf{x}_t - \boldsymbol{\mu}\right)\left(\mathbf{x}_t - \boldsymbol{\mu}\right)^T\right)\right) + \frac{T\nu}{2}\log\frac{\nu}{2} + \frac{\nu}{2}\sum_{t=1}^{T}\left(e_{2,t}^{(k)} - e_{1,t}^{(k)}\right) - T\log\Gamma\left(\frac{\nu}{2}\right) + const. \tag{4}$$

Maximization Step: Here we devide the parameters update into two parts. *Update of* $\boldsymbol{\mu}$ *and* ν*:* It is easy to see that $\boldsymbol{\mu}$ and ν are actually decoupled in $Q\big(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}\big)$ and thus can be easily obtained by setting their derivative to zero. Then the update scheme for $(\boldsymbol{\mu}, \nu)$ is

$$\boldsymbol{\mu}^{(k+1)} = \sum_{t=1}^{T} e_{1,t}^{(k)}\mathbf{x}_t \Big/ \sum_{t=1}^{T} e_{1,t}^{(k)} \tag{5}$$

$$\nu^{(k+1)} = \underset{\nu>0}{\mathrm{argmax}}\left\{\frac{T\nu}{2}\log\frac{\nu}{2} + \frac{\nu}{2}\sum_{t=1}^{T}\left(e_{2,t}^{(k)} - e_{1,t}^{(k)}\right) - T\log\Gamma\left(\frac{\nu}{2}\right)\right\}. \tag{6}$$

According to Proposition 1 in [6], $\nu^{(k+1)}$ always exists and can be found by bisection search. Interestingly, the update for $\boldsymbol{\mu}$ and ν are independent from each other and irrelevant to $\boldsymbol{\Sigma}$.

Update of $\boldsymbol{\Sigma}$*:* Fixing $\boldsymbol{\mu}$ and ν, maximizing $Q\big(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}\big)$ w.r.t. $\boldsymbol{\Sigma}$ is reduced to a GFA problem with $\mathbf{S}^{(k)} = \frac{1}{T}\sum_{t=1}^{T} e_{1,t}^{(k)}\left(\mathbf{x}_t - \boldsymbol{\mu}^{(k+1)}\right)\left(\mathbf{x}_t - \boldsymbol{\mu}^{(k+1)}\right)^T$, which can be solved by any of the two iterative methods described in Section 2, which require several iterations until convergence. Considering that solving $\boldsymbol{\Sigma}$ exactly would require several iterations and could be time-consuming, we can instead only run the algorithms for one round, which would correspond to implementing the GEM instead of EM.

4.2 An Acceleration Scheme: PX-EM

A drawback of the EM algorithm is the slow convergence. The parameter expanded EM (PX-EM) [7] was proposed as an efficient method to accelerate the classical EM method and can be applied here. A well-known application of PX-EM on multivariate Student's t case is to assume that we have $\tau \overset{i.i.d}{\sim} \alpha\mathrm{Gamma}\left(\frac{\nu}{2}, \frac{\nu}{2}\right)$ in the Student's t hierarchical structure, where α is the expanded parameter.

PX-E step: The PX-E step needs only a few modifications on the original expectation step. Specifically, the Mahalanobis distance should be calculated as $d\left(\mathbf{x}_t, \boldsymbol{\mu}_*, \boldsymbol{\Sigma}_*/\alpha\right)$, where $(x)_*$ is the corresponding notation of x in the parameter expanded statistical model.

PX-M step: The update schemes for parameters is similar to those in EM with only few changes. The update of $\boldsymbol{\mu}$ and ν keeps unchanged but the update of $\boldsymbol{\Sigma}$ depends on $\left(\mathbf{S}^{(k)}\right)_* = \alpha^{(k)}\mathbf{S}^{(k)}$. The update of the new parameter α is $\alpha^{(k+1)} = \frac{\alpha^{(k)}}{T}\sum_{t=1}^{T} e_{1,t}^{(k)}$. After the algorithm achieves convergence, the real parameters should be recovered as $\boldsymbol{\mu} = \boldsymbol{\mu}_*$, $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_*/\alpha$ and $\nu = \nu_*$.

4.3 Robust Enhancement to Missing Data

Due to measurement problems or transmission/storage errors, the observed data $\mathbf{x}_t$ might contain some missing values, which has been well studied, cf. [6]. It turns out that the missing values can be regarded as latent data like $\boldsymbol{\tau}$. The new $Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}\right)$ at expectation step has the same expression w.r.t. $\boldsymbol{\theta}$ as in (4) [6]. Therefore the maximization step can also be achieved by first updating $\boldsymbol{\mu}$ and ν, and then $\boldsymbol{\Sigma}$ with FA structure imposed.

5 Numerical Experiments

5.1 Synthetic Data

We generate synthetic data following the multivariate Student's t distribution. The basic dimension setting is $p = 100$ and $r = 5$. The true distribution parameters are chosen as follows: ν_{true} is set to be 7, the elements of $\boldsymbol{\mu}_{\text{true}}$ are drawn i.i.d. from $\mathcal{N}(0, 1)$, $\mathbf{B}_{\text{true}}$ comes from BARRA industry model [3] with each sector has size 20, diagonal elements of $\boldsymbol{\Psi}_{\text{true}}$ are generated independently from an exponential distribution with mean 10. The initial $\boldsymbol{\mu}^{(0)}$, $\nu^{(0)}$ are set as the sample mean and 10, and $\boldsymbol{\Psi}^{(0)}$, $\mathbf{B}^{(0)}$ are given by the naive PCA method. The convergence condition is set as $|L^{(k+1)} - L^{(k)}| \leq 10^{-6}|L^{(k)}|$.

Convergence Illustration: In Fig. 1, we compare the convergence of our proposed methods. The log-likelihood function of the observed data increases monotonically with the iterations and can finally converge. The PX-EM can significantly accelerate the convergence in this case. Then, in Fig. 2, we show the parameter estimation convergence. The parameter estimation normalized errors are defined as $\text{NE}(\boldsymbol{\mu}) = \|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_{\text{true}}\|_2/\|\boldsymbol{\mu}_{\text{true}}\|_2$, $\text{NE}(\boldsymbol{\Sigma}) = \|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}_{\text{true}}\|_F/\|\boldsymbol{\Sigma}_{\text{true}}\|_F$ and $\text{NE}(\nu) = |s(\hat{\nu}) - s(\nu_{\text{true}})|/|s(\nu_{\text{true}})|$ for $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ and ν, where $s(\nu) = \frac{\nu}{\nu-2}$. In Figure 2, we can find all the errors are decreasing and finally converge.

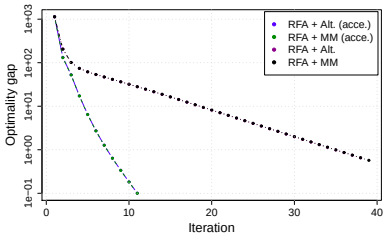

Fig. 1. Optimality gap vs iterations.

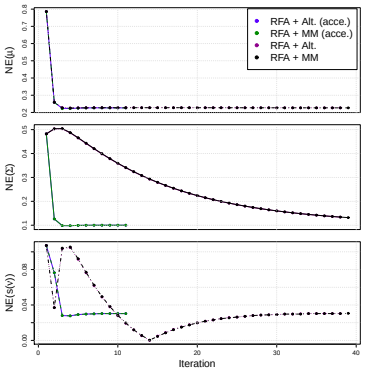

Fig. 2. Estimation error vs iterations.

Robustness Illustration: We compare our proposed RFA methods in covariance matrix estimation with sample covariance matrix (SCM), Student's t (Stu-t) estimation (without FA structure) [6], GFA, and iterative PCA (Iter-PCA) estimation. The results shown below are averaged over 100 different realizations of $\mathbf{X}$ following Gaussian distribution. In Fig. 3, we change the sample number n but fixing $p = 100$. All methods show better performance when n goes large while our proposed RFA method always gives the best result. In Fig. 4, we randomly pick some rows of $\mathbf{X}$ and element-wisely add outliers drawn from $\mathcal{N}(0, 50)$. It is significant that our proposed RFA method can still hold a good estimation while the results from non-robust estimation methods are totally destroyed. The results owe to the robustness of Student's t assumption in resisting the outliers. In Fig. 5, we randomly pick some rows of $\mathbf{X}$ and randomly set 10% values be missing for each row. Our proposed robust FA algorithm will be directly fed with incomplete data while for other methods we need to manually remove the rows containing the missing values. It is impressive that our proposed robust FA method can keep the lowest and almost unchanged performance.

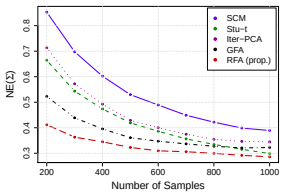

Fig. 3. Average NE $(\mathbf{\Sigma})$ when n changes.

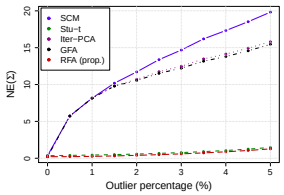

Fig. 4. Average NE $(\mathbf{\Sigma})$ with outliers.

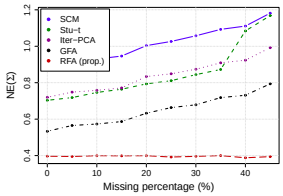

Fig. 5. Average NE $(\mathbf{\Sigma})$ with missing values.

5.2 Real Data

In this Section, we show the performance of our proposed algorithm based on the backtest with real financial data. We randomly choose 50 stocks for 10 times from Standard & Poor's 500 list and 2 years (2×252) continuous historical daily prices data between 01 Dec 2008 and 01 Dec 2018. Then for each resampling dataset, we perform the rolling window backtest with lookback window length set to 100 days and test window be 5 days. The rebalance is assumed to be done everyday without transaction cost. To fairly compare the estimation performance, we are particularly interested in the global minimum variance portfolio (GMVP): minimize $\mathbf{w}^T\mathbf{\Sigma}\mathbf{w}$ with constraint $\mathbf{1}^T\mathbf{w} = 1$, where $\mathbf{\Sigma}$ is the covariance matrix obtained from various methods. We respectively set $r = 2$ and $r = 4$ in Fig. 6 and 7. It turns out that our proposed RFA method can achieve smaller out-of-sample risk with less uncertainty.

6 Conclusion

In this paper, we have proposed the RFA algorithm to obtain the MLE of the multivariate Student's t distribution with FA structure imposed. The algorithm was based on the EM framework and had great estimation accuracy and robustness to outliers and missing values. The backtest over real financial data has shown advantages and practical usefulness of our proposed RFA algorithm.

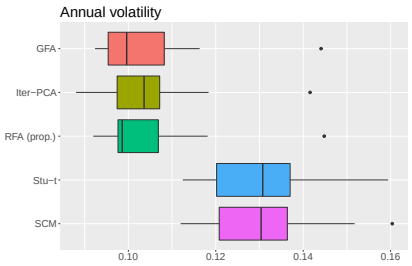

Fig. 6. Portfolio risk ($r = 2$).

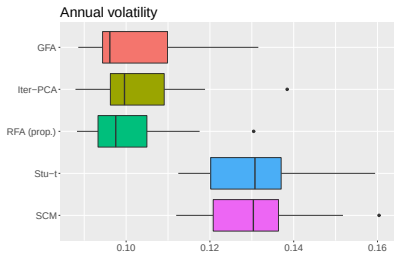

Fig. 7. Portfolio risk ($r = 4$).

References

1. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. Series B pp. 1–38 (1977)
2. Feng, Y., Palomar, D.P.: A signal processing perspective on financial engineering. Foundations and Trends® in Signal Processing (2016)
3. Heckman, L., Narayanan, S.R., Patel, S.A.: Country and industry importance in european returns. Available at SSRN 169455 (1999)
4. Khamaru, K., Mazumder, R.: Computation of the maximum likelihood estimator in low-rank factor analysis. arXiv preprint arXiv:1801.05935 (2018)
5. Lewis, A.S.: Derivatives of spectral functions. Mathematics of Operations Research **21**(3), 576–588 (1996)
6. Liu, C., Rubin, D.B.: ML estimation of the t distribution using EM and its extensions, ECM and ECME. Statistica Sinica pp. 19–39 (1995)
7. Liu, C., Rubin, D.B., Wu, Y.N.: Parameter expansion to accelerate EM: the PX-EM algorithm. Biometrika **85**(4), 755–770 (1998)
8. Moon, T.K.: The expectation-maximization algorithm. IEEE Signal Processing Magazine **13**(6), 47–60 (1996)
9. Ramírez, D., Santamaria, I., Van Vaerenbergh, S., Scharf, L.L.: An alternating optimization algorithm for two-channel factor analysis with common and uncommon factors. pp. 1743–1747 (2018)
10. Santamaria, I., Scharf, L.L., Via, J., Wang, H., Wang, Y.: Passive detection of correlated subspace signals in two mimo channels. IEEE Transactions on Signal Processing **65**(20), 5266–5280 (2017)
11. Sardarabadi, A.M., van der Veen, A.J.: Complex factor analysis and extensions. IEEE Transactions on Signal Processing **66**(4), 954–967 (2018)
12. Sun, Y., Babu, P., Palomar, D.P.: Majorization-minimization algorithms in signal processing, communications, and machine learning. IEEE Transactions on Signal Processing **65**(3), 794–816 (2017)
13. Ten Berge, J.M., Snijders, T.A., Zegers, F.E.: Computational aspects of the greatest lower bound to the reliability and constrained minimum trace factor analysis. Psychometrika **46**(2), 201–213 (1981)
14. Tsay, R.S.: Analysis of financial time series, vol. 543. John Wiley & Sons (2005)
15. Wang, W.L., Liu, M., Lin, T.I.: Robust skew-t factor analysis models for handling missing data. Statistical Methods & Applications **26**(4), 649–672 (2017)
16. Zhang, J., Li, J., Liu, C.: Robust factor analysis using the multivariate t-distribution. Statistica Sinica **24**(1), 291–312 (2014)