Filomat 33:15 (2019), 4753–4767
https://doi.org/10.2298/FIL1915753M

Published by Faculty of Sciences and Mathematics,
University of Niš, Serbia
Available at: http://www.pmf.ni.ac.rs/filomat

An EM Algorithm for Singular Gaussian Mixture Models

Khalil Masmoudi[a], Afif Masmoudi[a]

[a] *Laboratory of Probability and Statistics - University of Sfax*

Abstract. In this paper, we introduce finite mixture models with singular multivariate normal components. These models are useful when the observed data involves collinearities, that is when the covariance matrices are singular. They are also useful when the covariance matrices are ill-conditioned. In the latter case, the classical approaches may lead to numerical instabilities and give inaccurate estimations. Hence, an extension of the Expectation Maximization algorithm, with complete proof, is proposed to derive the maximum likelihood estimators and cluster the data instances for mixtures of singular multivariate normal distributions. The accuracy of the proposed algorithm is then demonstrated on the grounds of several numerical experiments. Finally, we discuss the application of the proposed distribution to financial asset returns modeling and portfolio selection.

1. Introduction

The multivariate Normal distribution is of great importance in statistics. It stands out because it is relatively easy to use, at the heart of the central limit theorem and adapted to a large number of natural phenomena modeling. The dependence structure of a Gaussian vector is fully determined through its covariance matrix Σ. Several statistical methods, including the principal component analysis, linear discriminant analysis, clustering analysis, and regression models, require the knowledge of the covariance structure. When the observed data involves collinearities or the variable dimension d is larger than the sample size n, the covariance matrix is singular. In this case, its estimation is a challenging problem. The maximum likelihood estimator of Σ yields a nonunique estimate ([24], [18]). When the studied model includes many classes, a mixture of non-singular normal distributions is used. These models were studied by many authors such as [5], [28], [19], [6] and [12]. They are increasingly used to model the distributions of diverse phenomena. The Gaussian mixture can also be useful for approximating a multimodal density function [16]. The maximum Likelihood estimates of the mixture parameters are computed using the Expectation-Maximization (EM) algorithm (see [8]).
In this paper, we introduced the singular Gaussian mixture model which addresses two main shortcomings in the existing research. First, this model is useful to adapt several statistical techniques such as the discriminant analysis [15] and density approximation, in the presence of singular covariance matrices. Second, it is also effective with ill-conditioned covariance matrices and can be used in various applications such as returns modeling and risk management. Besides, an extended EM algorithm for the parameters estimation of a singular Gaussian mixture was provided. The proposed algorithm offers an interesting way to deal

2010 *Mathematics Subject Classification*. Primary 62H10 ; Secondary 62H12
Keywords. Finite mixture, Maximum likelihood, Singular multivariate normal distribution, EM algorithm, Portfolio selection
Received: 13 April 2019; Accepted: 15 August 2019
Communicated by Biljana Popović
Email addresses: khalil.masmoudi@centraliens.net (Khalil Masmoudi), afif.masmoudi@fss.rnu.tn (Afif Masmoudi)

with numerical problems in the presence of ill-conditioned covariance matrices.

The remaining of this paper is organized as follows: In Section 2, we remind the reader of the main properties of a singular multivariate normal distribution and the maximum likelihood estimators of its parameters. Section 3 discusses the suggested mixture model and its main properties. In Section 4, the maximum likelihood estimators of the mixture parameters are derived and a customized EM algorithm is proposed for their computation and the data clustering. The proposed approach is evaluated, in Section 5, on the basis of a simulation study. Finally, Section 6 illustrates the use of the mixture of singular multivariate Gaussian distributions to model financial assets returns in the context of portfolio selection.

2. Singular multivariate normal distributions

2.1. Density function

Consider a d-dimensional Gaussian random vector $\mathbf{X}$ with mean vector μ and covariance matrix Σ. When Σ is positive definite ($\Sigma > 0$), the density function of X with respect to the Lebesgue measure on $\mathbb{R}^d$ is given by

$$f(y) = (2\pi)^{-\frac{d}{2}}|\Sigma|^{-\frac{1}{2}} \exp(-\frac{1}{2}(y-\mu)^T\Sigma^{-1}(y-\mu)), \tag{1}$$

where $|\Sigma|$ denotes the determinant of Σ.
However, when Σ is singular, the density function is restricted to an affine subspace of $\mathbb{R}^d$ ([9]). Hence, Σ is diagonalized as follows

$$\Sigma = (P_1, P_2)\begin{pmatrix}\Lambda & 0\\ 0 & 0\end{pmatrix}\begin{pmatrix}{P_1}^T\\ {P_2}^T\end{pmatrix}$$

where P_1 is a $d \times r$ matrix whose orthonormal column vectors span the Σ's image and P_2 is a $d \times (d-r)$ matrix whose orthonormal column vectors span the Σ's null space. The rank of Σ is denoted by r and Λ is a $r \times r$ diagonal matrix containing positive Σ's eigenvalues $\Lambda = diag(\lambda_1, ..., \lambda_r)$. P_1 and P_2 are such that

$$P_1{P_1}^T + P_2{P_2}^T = I_d. \tag{2}$$

Where I_d is the d-dimensional identity matrix.
The density function is concentrated on the affine subspace

$$E = \{y \in \mathbb{R}^d; {P_2}^T y = {P_2}^T \mu\}. \tag{3}$$

A density function of X with respect to the restriction of the Lebesgue measure to E was first introduced by [14], for any generalized inverse Σ^- defined by $\Sigma\Sigma^-\Sigma = \Sigma$:

$$f(y|\mu, \Sigma) = (2\pi)^{-\frac{r}{2}}|\Lambda|^{-\frac{1}{2}} \exp(-\frac{1}{2}(y-\mu)^T\Sigma^-(y-\mu)), \tag{4}$$

A particular generalized inverse of Σ, denoted by Σ^+ is the Moore-Penrose pseudo-inverse ([22]):

$$\Sigma^+ = P_1\Lambda^{-1}{P_1}^T \tag{5}$$

2.2. Maximum likelihood estimation

This subsection provides the maximum likelihood estimators for singular multivariate normal distribution parameters ([25]). Consider a random sample of size n with observations matrix $\boldsymbol{y} = (y_1 : ... : y_n)$. The likelihood function is given by

$$\begin{aligned} l(\boldsymbol{y}, \mu, P_1, \Lambda) &= (2\pi)^{-\frac{nr}{2}}|\Lambda|^{-\frac{n}{2}} \exp(\sum_{i=1}^{n} -\frac{1}{2}(y_i-\mu)^T\Sigma^+(y_i-\mu)) \\ &= (2\pi)^{-\frac{nr}{2}}|\Lambda|^{-\frac{n}{2}} \exp(-\frac{1}{2}tr(\Lambda^{-1}{P_1}^T S P_1)). \end{aligned}$$

Where $S = (\boldsymbol{y} - \mu\mathbb{1}^T)(\boldsymbol{y} - \mu\mathbb{1}^T)^T$; $\mathbb{1}$ stands for $n \times 1$ vector of ones i.e., $\mathbb{1} = (1, ..., 1)^T$. The maximum likelihood estimators (MLE) are given by the following proposition.

Proposition 2.1. - *[25]*
Let $Y \sim N_{d,n}(\mu\mathbb{1}^T, \Sigma, I_n)$ a $d \times n$ random matrix whose columns are i.i.d and normally distributed with mean μ and covariance matrix Σ, where Σ is of rank r, and $\Sigma = P_1 \Lambda {P_1}^T$. Let $S = Y(I_n - \frac{1}{n}\mathbb{1}\mathbb{1}^T)Y^T$ and H be the $p \times r$ matrix of eigenvectors corresponding to the r largest eigenvalues of S, denoted by $L = diag(l_1, ..., l_r)$. Then the MLE of μ, Λ and P_1 are respectively given by

$$\begin{aligned}
&\hat{\mu} = \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \\
&\hat{\Lambda} = \frac{1}{n}L \\
&\hat{P_1} = H \\
&\hat{\Sigma} = \frac{1}{n}HLH^T
\end{aligned}$$

3. Mixture of singular multivariate normal distributions

Consider a mixture of K d-variate singular normal distributions with density

$$f(y|\Theta) = \sum_{k=1}^{K} \pi_k f_k(y|\mu_k, \Sigma_k) \tag{6}$$

where the π_k's are the mixing weights ($0 < \pi_k < 1$; $\sum_{k=1}^{K} \pi_k = 1$) and $f_k(y|\mu_k, \Sigma_k)$ is the density function of the multivariate singular Gaussian distribution with mean μ_k and covariance matrix Σ_k. The set of all the mixture parameters is denoted by $\Theta = \{\pi_k, \mu_k, \Sigma_k; k = 1, ..., K\}$. The mixture components are assumed to be concentrated on the same affine subspace E. This assumption is crucial since it allows having a dominated mixture model. In fact, under this assumption, the mixture model has the density function defined by (6) with respect to the restriction of the Lebesgue measure to an r-dimensional subspace.
Practically, this assumption seems natural when the data involves collinearties that are inherent in the studied population or when the empirical covariance matrix is ill-conditioned. In the latter case, the data is projected onto an r-dimensional affine subspace and consequently all the covariance matrices have the same rank r.

This mixture is useful when the considered sample $y = (y_1 : ... : y_n)$ is drawn from a population which consists of K groups G_1, G_2,...,G_K, in proportions $\pi_1, ..., \pi_K$. Given G_k, y_i is drawn from the Gaussian distribution with mean μ_k and covariance matrix Σ_k.
In the remaining text of this paper, the random variables are denoted by capital letters whereas their associated observations are denoted by the same lower-case letters. It is also convenient to associate a K-dimensional component-label vector $z_i = (z_{i1}, ..., z_{iK})$ to each observation y_i in order to manage the groups membership. So, the k^{th} element of z_i, $z_{ik} = (z_i)_k$, is defined to be one or zero, according to whether y_i is a member of the group G_k or not ($z_{ik} \in \{0, 1\}$; $\sum_{k=1}^{K} z_{ik} = 1$). Thus, Z_i is distributed according to a multinomial distribution with probabilities vector $\pi = (\pi_1, ..., \pi_K)$, we write

$$Z_i \sim Mult_K(1, \pi) \text{ and } P(Z_i = z_i) = \prod_{k=1}^{K} \pi_k^{z_{ik}}$$

The concept of associating a label z_i to each observation y_i is useful for computing the MLE (Maximum Likelihood Estimate) of the mixture distribution via a straightforward application of the EM algorithm.

4. Parameters estimation

In this section, we introduce a customized version of the EM algorithm to iteratively estimate the mixture parameters Θ. Consider a random sample with n random vectors $Y_1, ..., Y_n$ independent and identically distributed with mixture density function given by (6).

In the EM framework, the feature data $y_1, ..., y_n$ are viewed as incomplete data since their associated component-label vectors $z_1, ..., z_n$ are not available. The complete data is $(y^T, z^T)^T$ where $y = (y_1 : ... : y_n)$ is the incomplete data and $z = (z_1 : ... : z_n)$ is the hidden data.
The complete likelihood function is given by

$$l(y_1, ..., y_n, z_1, ..., z_n|\Theta) = \prod_{i=1}^{n}\prod_{k=1}^{K}[\pi_k^{z_{ik}} f_k^{z_{ik}}(y_i)] \tag{7}$$

The complete-data log-likelihood can be written as

$$L(y_1, ..., y_n, z_1, ..., z_n|\Theta) = \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik} log(\pi_k) + z_{ik} log(f_k(y_i)) \tag{8}$$

The EM algorithm proceeds in two steps, E-step (Expectation) and M-step (Maximization). The E-step takes the conditional expectation of the complete-data log-likelihood given the observed data y, using the current fit for Θ. The M-step gives the best update for Θ in the current iteration. The choice of the starting values $\Theta^{(0)}$ is very important especially for the singular normal distributions mixture.

4.1. Initial values choice

On the first iteration of the EM algorithm, we need to specify the initial values of Θ. Since each component of the mixture is concentrated on the affine subspace E, we should add some restrictions on the possible values for Σ_k and μ_k. The following values can be chosen:

- Any arbitrary value for $\pi_k^{(0)}$, for example $\frac{1}{K}$.
- Any arbitrary value verifying: $\mu_k{}^{(0)} \in E$. This can be obtained by taking a convex combination on elements of a subset of the data set. In fact, since all the observations are concentrated on the same affine subspace E, which is stable by convex combination, with probability one, the proposed initial value of $\mu_k{}^{(0)}$ belongs to E.
- Any arbitrary value Σ_k positive semi-definite verifying : $\varphi(\Sigma_k) = \bar{E}$ where $\varphi(A)$ denotes the column space of a matrix A and $\bar{E}$ is the vector subspace associated to E.

From a practical point of view, one can partition the data set into K subsets $\mathcal{D}_1, ..., \mathcal{D}_K$ (using prior knowledge or a deterministic clustering algorithm such as k-means). Then, the initial values of the parameters are constructed using the following formulas :

- $\pi_k^{(0)} = \frac{|\mathcal{D}_k|}{n}$,
- $\mu_k{}^{(0)} = \frac{1}{|\mathcal{D}_k|}\sum_{y_i \in \mathcal{D}_k} y_i$,
- $\Sigma_k{}^{(0)} = \frac{1}{|\mathcal{D}_k|}\sum_{y_i \in \mathcal{D}_k} (y_i - \mu_k{}^{(0)})(y_i - \mu_k^{(0)})^T$,

where $|\mathcal{D}_k|$ is the cardinality of $\mathcal{D}_k$.

4.2. E-Step

The addition of the hidden random variables $Z = (Z_1 : ... : Z_n)$ is handled by the E-step. After l iterations, the conditional expectation of the complete-data log likelihood given the random sample $Y = (Y_1 : ... : Y_n)$, using the current fit $\Theta^{(l)}$, can be written as follows

$$Q(\Theta||\Theta^{(l)}) = E_{\Theta^{(l)}}(L(Y, Z, \Theta)|Y) \tag{9}$$

Thanks to the linearity of L in the unobservable variable Z_{ik}, the E-step requires only the computation of the current expectation of Z_{ik} given $Y = y$. This expectation is

$$\tau_{ik}^{(l)} = E_{\Theta^{(l)}}(Z_{ik}|Y = y) = P_{\Theta^{(l)}}(Z_{ik} = 1|Y = y) = \frac{\pi_k^{(l)} f_k(y_i|\mu_k{}^{(l)}, \Sigma_k{}^{(l)})}{\sum_{k=1}^{K} \pi_k^{(l)} f_k(y_i|\mu_k{}^{(l)}, \Sigma_k{}^{(l)})} \tag{10}$$

$\tau_{ik}^{(l)}$ can be viewed as the probability that y_i belongs to the k^{th} component of the mixture. Using (10), the conditional expectation of the complete-data log-likelihood given $Y = y$ is

$$Q(\Theta||\Theta^{(l)}) = \sum_{i=1}^{n}\sum_{k=1}^{K} \tau_{ik}^{(l)}[log(\pi_k) + log(f_k(y_i|\mu_k, \Sigma_k))] \tag{11}$$

Using the density function defined in (4), we get

$$Q(\Theta||\Theta^{(l)}) = \sum_{i=1}^{n}\sum_{k=1}^{K} \tau_{ik}^{(l)}[log(\pi_k) - \frac{r}{2}log(2\pi) - \frac{1}{2}log|\Lambda_k| - \frac{1}{2}(y_i - \mu_k)^T\Sigma_k{}^{+}(y_i - \mu_k)] \tag{12}$$

4.3. M-Step

The M-step, at the $(l+1)^{th}$ iteration, requires the global maximization of $Q(\Theta||\Theta^{(l)})$ with respect to Θ over the parameters space to give the updated estimates:

$$\Theta^{(l+1)} = \arg\max_{\Theta} Q(\Theta||\Theta^{(l)}) \tag{13}$$

The mixing proportions are updated using the classical result

$$\pi_k^{(l+1)} = \frac{1}{n}\sum_{i=1}^{n} \tau_{ik}^{(l)} \tag{14}$$

We now state the main theorem which gives the updates of the component means and covariance matrices. This theorem relies on the same assumption as the model defined in (6): all the mixture components are concentrated on the same affine subspace with probability one. Hence, all the covariance matrices have the same rank r.

Theorem 4.1. *For $k \in \{1, ..., K\}$, let $r = rank(\Sigma_k)$,*

$$S_k = \sum_{i=1}^{n} \tau_{ik}^{(l)}(Y_i - \mu_k{}^{(l+1)})(Y_i - \mu_k{}^{(l+1)})^T \tag{15}$$

and H_k be the $d \times r$ matrix of eigenvectors corresponding to the r largest eigenvalues of S_k, denoted by $L_k = diag(l_{k1}, ..., l_{kr})$. Then, the maximum of Q with respect to Σ_k is achieved for

$$\Sigma_k{}^{(l+1)} = \frac{H_k L_k H_k{}^T}{n_k} \text{ where } n_k = \sum_{i=1}^{n} \tau_{ik}^{(l)}.$$

$\Lambda_k{}^{(l+1)} = \frac{1}{n_k} L_k.$

$P_{1k}{}^{(l+1)} = H_k.$

The updates of the components means μ_k are given by

$$\mu_k{}^{(l+1)} = \frac{\sum_{i=1}^{n} \tau_{ik}^{(l)} Y_i}{n_k} \tag{16}$$

In order to prove Theorem 4.1, we need the following two propositions. The first proposition is a well-known linear algebra result (see [7]):

Proposition 4.2. *Let A, B be $d \times d$ symmetric matrices. Then*

$$\min_{U \text{unitary}} tr(AU^T BU) = \sum_{k=1}^{d} \alpha_k \beta_k, \tag{17}$$

where $\alpha_1 \geq \alpha_2 \geq ... \geq \alpha_d$ and $\beta_1 \leq \beta_2 \leq ... \leq \beta_d$ are the eigenvalues of A and B respectively, inversely ordered. (Note that U is unitary if $U^{-1} = U^T$.)

The second proposition is the following:

Proposition 4.3. *For $k \in \{1, ..., K\}$, with the starting values chosen in Subsection 4.1, we have with probability one:*

(i) $\forall l \in \mathbb{N}, \mu_k{}^{(l)} \in E.$

(ii) $\forall l \in \mathbb{N}, \varphi(\Sigma_k{}^{(l)}) \subseteq \bar{E}$ *and if* $n \geq r$ *then* $\varphi(\Sigma_k{}^{(l)}) = \bar{E}$

Proof: Since $\mu_k{}^{(l)}$ and $\Sigma_k{}^{(l)}$ are defined iteratively and the initial values $\mu_k^{(0)}$ and $\Sigma_k{}^{(0)}$ meet the required conditions, we only need to prove that $\mu_k{}^{(l+1)} \in E$ and $\varphi(\Sigma_k{}^{(l+1)}) \subseteq \bar{E}$.
Since all the Y_i are elements of E with probability one, we get from (16) that $\mu_k{}^{(l+1)} \in E$ as a convex combination of elements of E.
For the second part of the proof, we see that $\Sigma_k{}^{(l+1)}$ is the sum of n elements having the form $V_i V_i{}^T$ where $V_i = (Y_i - \mu_k{}^{(l+1)}) \in \bar{E}$. The latter remark implies that $\varphi(\Sigma_k{}^{(l+1)}) \subseteq \bar{E}$ with probability one. And if $n \geq r$, the equality is achieved. □

Proof of Theorem 4.1
Note that Q is concave with respect to μ_k. In order to maximize it with respect to μ_k, we simply have to write the first-order condition

$$\frac{\partial Q}{\partial \mu_k} = 0 = \sum_{i=1}^{n} \tau_{ik}^{(l)} \Sigma_k{}^{+} (Y_i - \mu_k).$$

Using (5), the condition becomes

$$P_{1k} \Lambda_k{}^{-1} [\sum_{i=1}^{n} \tau_{ik}^{(l)} P_{1k}{}^T (Y_i - \mu_k)] = 0$$

Since $P_{1k} \Lambda_k{}^{-1}$ is full-column rank (one-to-one), with probability one, we necessarily have

$$\sum_{i=1}^{n} \tau_{ik}^{(l)} P_{1k}{}^T (Y_i - \mu_k) = 0$$

Multiplying the above equation by P_{1k} and using (2), we get

$$\sum_{i=1}^{n} \tau_{ik}^{(l)} P_{1k} {P_{1k}}^T (Y_i - \mu_k) = 0 \Leftrightarrow \sum_{i=1}^{n} \tau_{ik}^{(l)} (Y_i - \mu_k) = \sum_{i=1}^{n} \tau_{ik}^{(l)} P_{2k} {P_{2k}}^T (Y_i - \mu_k)$$

Since $Y_i \in E$, the right-hand term is equal to 0 (with probability one), we get

$$\sum_{i=1}^{n} \tau_{ik}^{(l)} Y_i = \mu_k \sum_{i=1}^{n} \tau_{ik}^{(l)}$$

which completes the second part of the proof (μ_k update).

As previously done, Σ_k is decomposed as $\Sigma_k = P_{1k} \Lambda_k {P_{1k}}^T$. Similarly, we write $S_k = H_k L_k {H_k}^T$ where H_k is semi-orthogonal i.e., ${H_k}^T H_k = I_r$. Then, from Proposition 4.3, it follows

$$\varphi(P_{1k}) = \varphi(\Sigma_k) = \bar{E} = \varphi(S_k) = \varphi(H_k) \tag{18}$$

In fact, the previous decompositions imply that $\varphi(\Sigma_k) \subseteq \varphi(P_{1k})$ and $\varphi(S_k) \subseteq \varphi(H_k)$. The equalities are achieved since Σ_k, P_{1k}, S_k and H_k have same rank r.
Thus, it exists a full-rank $r \times r$ matrix ψ_k such that $P_{1k} = H_k \psi_k$.

$$I_r = {P_{1k}}^T P_{1k} = {\psi_k}^T {H_k}^T H_k \psi_k = {\psi_k}^T \psi_k. \tag{19}$$

Since ψ_k is square and full-rank, it follows that ψ_k is unitary.
Moreover, using the fact that

$$(Y_i - \mu_k)^T {\Sigma_k}^+ (Y_i - \mu_k) = tr((Y_i - \mu_k)^T {\Sigma_k}^+ (Y_i - \mu_k)) = tr((Y_i - \mu_k)(Y_i - \mu_k)^T {\Sigma_k}^+),$$

the conditional expectation of the complete-data log likelihood given Y becomes

$$Q(\Theta || \Theta^{(l)}) = c - \frac{1}{2} \sum_{k=1}^{K} \sum_{i=1}^{n} \tau_{ik}^{(l)} [log|\Lambda_k| + tr((Y_i - \mu_k)(Y_i - \mu_k)^T {\Sigma_k}^+)]. \tag{20}$$

Where $c = \sum_{i=1}^{n} \sum_{k=1}^{K} \tau_{ik}^{(l)} [log(\pi_k) - \frac{r}{2} log(2\pi)]$ is constant with respect to μ_k and Σ_k. Then, thanks to trace function linearity, and using (15) we get

$$Q(\Theta || \Theta^{(l)}) = c - \frac{1}{2} \sum_{k=1}^{K} \left[tr(S_k {\Sigma_k}^+) + \sum_{i=1}^{n} \tau_{ik}^{(l)} log|\Lambda_k| \right]. \tag{21}$$

Replacing ${\Sigma_k}^+$ and S_k respectively by $P_{1k} {\Lambda_k}^{-1} {P_{1k}}^T$ and $H_k L_k {H_k}^T$ we get

$$Q(\Theta || \Theta^{(l)}) = c - \frac{1}{2} \sum_{k=1}^{K} \left[tr(H_k L_k {H_k}^T P_{1k} {\Lambda_k}^{-1} {P_{1k}}^T) + log|\Lambda_k| \sum_{i=1}^{n} \tau_{ik}^{(l)} \right] \tag{22}$$

$$= c - \frac{1}{2} \sum_{k=1}^{K} \left[tr({\Lambda_k}^{-1} {\psi_k}^T L_k \psi_k) + n_k log|\Lambda_k| \right]. \tag{23}$$

The latter equality (23) is obtained using (19), and the invariance of the trace function under cyclic permutations.
For each $k \in \{1, 2, ..., K\}$, we order the eigenvalues of S_k and Σ_k such that $l_{1k} \geq l_{2k} \geq ... \geq l_{rk}$ and $\lambda_{1k} \geq \lambda_{2k} \geq ... \geq \lambda_{rk}$ respectively. Then, Proposition 4.2 with ${\Lambda_k}^{-1}$, ψ_k, and L_k we get that for any ψ_k unitary,

$$tr({\Lambda_k}^{-1} {\psi_k}^T L_k \psi_k) \geq \sum_{m=1}^{r} \frac{l_{mk}}{\lambda_{mk}}$$

Therefore,

$$Q(\Theta||\Theta^{(l)}) = c - \frac{1}{2}\sum_{k=1}^{K}\left[tr({\Lambda_k}^{-1}{\psi_k}^T L_k\psi_k) + n_k log\left(\prod_{m=1}^{r}\lambda_m k\right)\right]$$
$$\leq c - \frac{1}{2}\sum_{k=1}^{K}\sum_{m=1}^{r}\left[\frac{l_{mk}}{\lambda_{mk}} + n_k log(\lambda_{mk})\right]. \quad (24)$$

The equality is obtained for $\psi_k = I_r$. Then, after replacing ψ_k by I_r, using the fact that $e^x \geq x+1$ with $x = log(\frac{l_{mk}}{n_k\lambda_{mk}})$, we get

$$\forall k \in \{1,...,K\}\,,\ \forall m \in \{1,...,r\}\,,\ \frac{l_{mk}}{n_k\lambda_{mk}} \geq 1 + log(\frac{l_{mk}}{n_k\lambda_{mk}})$$

By replacing and simplifying (24), we finally get

$$Q(\Theta||\Theta^{(l)}) \leq c - \frac{1}{2}\sum_{k=1}^{K}\sum_{m=1}^{r} n_k\left[1 + log(\frac{l_{mk}}{n_k})\right] \quad (25)$$

The equality is achieved for $x = 0$ i.e., $\lambda_{mk} = \frac{l_{mk}}{n_k}$. Thus, we obtain a global maximum of $Q(\Theta||\Theta^{(l)})$ with repect to Σ_k achieved for $\Sigma_k = H_k L_k {H_k}^T$. □

5. Simulation study

The performance of the proposed algorithm was evaluated on the grounds of a simulation study involving three five-variate singular normal components. We consider two different types of experiments using simulated data: the first analyzes the convergence properties of the algorithm; the second is designed to show how to use the algorithm with real data.

5.1. Example 1

5.1.1. Methodology

Each run of the proposed EM algorithm performs the following steps:

- We first generate a random sample (y, z) of size n from the singular Gaussian mixture distribution with density

$$f(y) = \pi_1 f_1(y|\mu_1, \Sigma_1) + \pi_2 f_2(y|\mu_2, \Sigma_2) + \pi_3 f_3(y|\mu_3, \Sigma_3) \quad (26)$$

for a parameters set $\Theta = (\mu_1, \Sigma_1, \pi_1, \mu_2, \Sigma_2, \pi_2, \mu_3, \Sigma_3, \pi_3)$.
Here, all the covariance matrices are singular and of rank $r = 4$. As described in Section 3, y is the observed data and z contains component-labels. We then split, arbitrarily, the generated data into a training set and a test set.

- The second step is to apply the described algorithm, using the observed data y only from the training set, in order to estimate the parameters. The algorithm is stopped when the distance between two successive estimations is small enough, i.e.,:

$$||\Theta^{(l+1)} - \Theta^{(l)}|| < \epsilon(1 + ||\Theta^{(l)}||), \quad (27)$$

where $||.||$ denotes a norm and ϵ is a relative tolerance (we choose $\epsilon = 10^{-4}$). Moreover, an upper bound on the number of iterations is added in order to avoid infinite loops.
The estimation errors are then computed as follows :

– Mean relative error of each component:

$$\frac{\|\mu_k - \widehat{\mu}_k\|_2}{\|\mu_k\|_2} = \frac{\sqrt{\sum_{i=1}^{d} \left[(\mu_k)_i - (\widehat{\mu}_k)_i\right]^2}}{\sqrt{\sum_{i=1}^{d} (\mu_k)_i^2}}$$

– Covariance matrix relative estimation error of each component :

$$\frac{\|\Sigma_k - \widehat{\Sigma}_k\|}{\|\Sigma_k\|} = \frac{\sqrt{trace((\Sigma_k - \widehat{\Sigma}_k)^2)}}{\sqrt{trace({\Sigma_k}^2)}}$$

– Proportion estimation error of each component:

$$|\pi_k - \widehat{\pi}_k|$$

- The test set is then used to compute the classification error rate (CER): This rate is given by the number of misclassified observed data divided by the size of the test set. In fact, the proposed algorithm provides a clustering criterion through the probability τ_{ij} defined in (10).

Using the above-described strategy, we performed several experiments. First, for a given parameters set Θ_0, the algorithm was applied using 100 different samples of size $n = 5000$. The goal here is to check the consistency of the estimation method with respect to the random sample. The obtained results reported in Table 1 show that the algorithm converges and gives almost the same parameters estimation for all the random samples. This fact appears more clearly in Figure 1 which reports the scatter diagrams of the relative errors.

Table 1: Average estimation errors and their associated confidence intervals for 100 EM runs with different random samples of size n=5000 drawn using a parameters set Θ_0.

	μ error	π error	Σ error	CER	Iterations Nb
Average	0.017	0.005	0.041	3.18%	32.47
Confidence interval	[0.016 ; 0.018]	[0.004 ; 0.006]	[0.038 ; 0.045]	[2.86% ; 3.06%]	[31.91 ; 33.03]

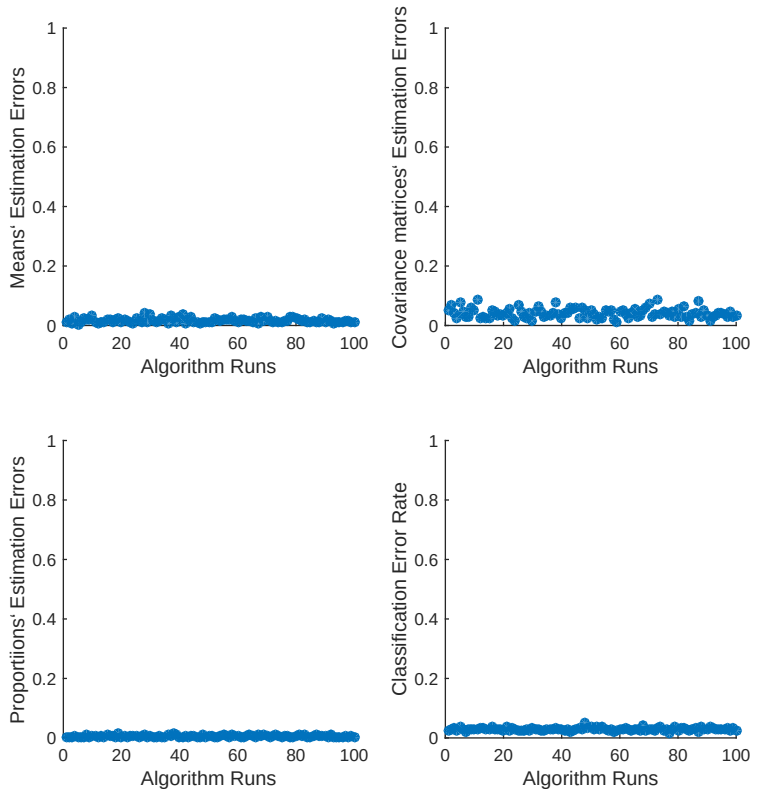

Figure 1: Estimation errors and misclassification rate in relation to the number of iterations with a sample of size n=5000

The second experiment aims at evaluating the algorithm consistency with respect to the model parameters. Hence, we randomly generated 15 parameter sets $(\Theta_1, ..., \Theta_{15})$ and for each parameters set Θ_i, we drew 10 different random samples of size $n = 5000$ and performed 10 runs using these samples. As previously, we computed the relative errors for each run. The average errors and their associated confidence intervals are reported in Table 2. The obtained results confirm the consistency of the algorithm. On average, the algorithm converges within almost 32 iterations. The obtained estimations are very close to the true parameters for all the random samples and for all the parameter sets.

Table 2: Average estimation errors and their associated confidence intervals for 150 EM runs with different random samples of size n=5000 drawn using 15 different parameter sets.

	μ error	π error	Σ error	CER	Iterations Nb
Average	0.028	0.008	0.052	6.00%	36.36
Confidence interval	[0.017 ; 0.038]	[0.004 ; 0.011]	[0.045 ; 0.059]	[5.36% ; 6.65%]	[33.14 ; 39.6]

In order to evaluate the algorithm convergence speed, we performed two additional experiments. First, we tested the effect of limiting the number of iterations on the estimation errors and classification error rates. Using a sample of size $n = 5000$ drawn with the parameters set Θ_0, we run the algorithm many times varying the iterations number from 5 to 100. For each run, we computed the the above-described errors. Figure 2 displays the obtained results. For all the error curves, the same kind of shape is observed : a rapid

improvement of the estimations followed by a slow decrease of the errors.

The last test aims at evaluating the algorithm convergence with respect to the sample size. Many simulations were performed using samples with different sizes n varying from 200 to 2000 (see Figure 3). We observe a rapid decrease until the sample size reaches 500. Thereafter, the errors are almost constant and close to zero.

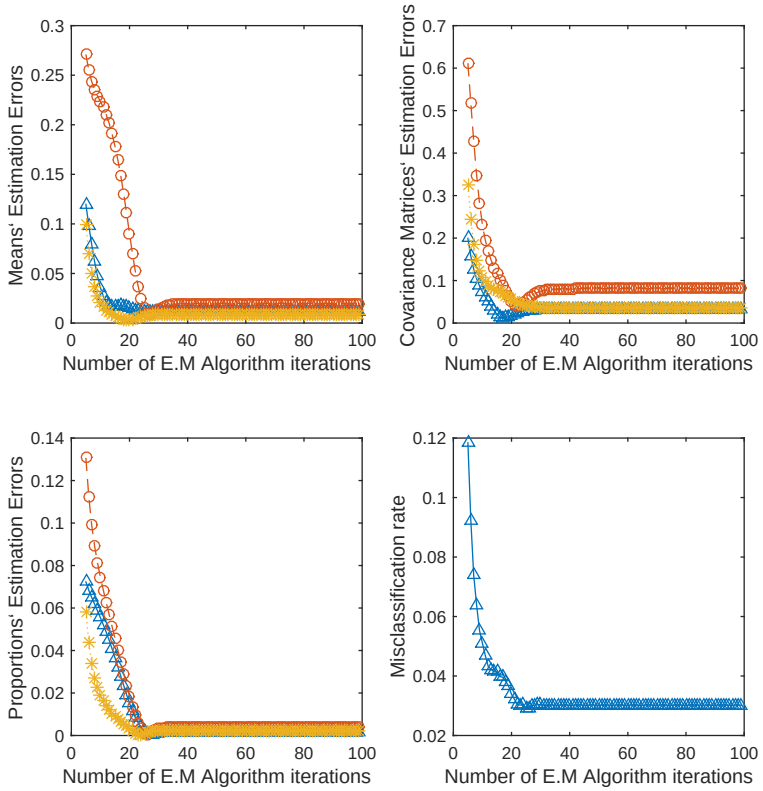

Figure 2: Estimation errors and misclassification rate in relation to the number of iterations with a sample of size n=2000

5.2. *Example 2*

In this second example, we consider the case of a mixture of normal distributions with ill-conditioned covariance matrices. In practice, when working with real data containing collinear features, the empirical covariance matrix S is ill-conditioned and not necessarily singular. However, the observed data are almost concentrated on an affine subspace. As illustration, we considered a mixture of two non-singular three-variate normal distributions with ill-conditioned covariance matrices. Hence, we generated a random sample of size $n = 500$ using the following mixture density function

$$f(y) = \pi_1 f_1(y|\mu_1, \Sigma_1) + \pi_2 f_2(y|\mu_2, \Sigma_2), \tag{28}$$

where the mixture parameters $\Theta = \{\pi_1, \mu_1, \Sigma_1, \pi_2, \mu_2, \Sigma_2\}$ are reported in Table 3. The 3D-scatter diagram of the random sample, reported in Figure 4, shows that the observations are almost concentrated on the 2-D plan plotted on the same figure. Consequently, projecting the data on this plan gives rise to a mixture of singular Gaussian distributions for which the proposed algorithm is suitable.

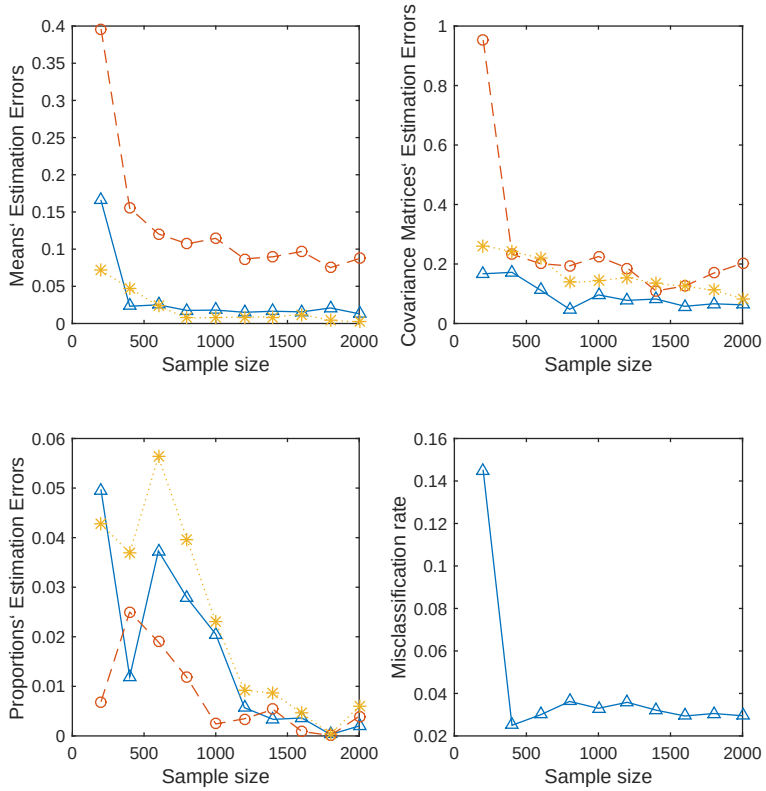

Figure 3: Estimation errors and misclassification rate in relation to the sample size

Table 3: Mixture parameters used for scatter plot in Figure 4

	π	μ	Σ		
Component (1)	0.5	$\begin{bmatrix} -0.12 \\ 0.69 \\ -1.15 \end{bmatrix}$	$\begin{bmatrix} 1.7456 & -0.3670 & 1.4447 \\ -0.3670 & 2.4747 & 0.7549 \\ 1.4447 & 0.7549 & 1.6641 \end{bmatrix}$		
Component (2)	0.5	$\begin{bmatrix} 4.96 \\ 3.45 \\ 4.75 \end{bmatrix}$	$\begin{bmatrix} 4.4157 & -0.9191 & 3.6591 \\ -0.9191 & 6.3672 & 1.9658 \\ 3.6591 & 1.9658 & 4.2378 \end{bmatrix}$		

When further generalized, for d-dimensional distributions, one can project data on the r-dimensional affine subspace E corresponding to the r largest eigenvalues of the empirical covariance matrix S. This projection procedure solves some numerical problems faced when working directly with the raw data. In fact, when applying the classical EM algorithm, the precision matrices are computed in each iteration to determine the probabilities τ_{ij}. The inverse computation fails after few iterations. In this context, the use of the projection combined with our proposed model improves the estimation results.

In order to illustrate this procedure, as in the previous example 5.1, we considered 15 parameter sets for the mixture density (28) and generated for each set of parameters, 10 different random samples of size

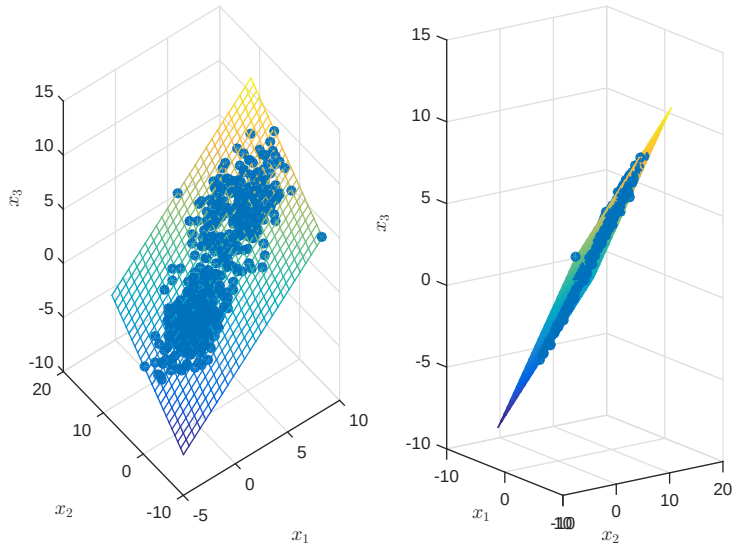

Figure 4: Visualization of sample drawn from mixture of two three-variate normal distributions with ill-conditioned covariance matrices.

$n = 3000$. We, then, compared the estimates obtained using our suggested algorithm with projected data and the classical EM algorithm applied to the raw data. The obtained results, summarized in Table 4, confirm the expected improvements due to the projection-based algorithm: on average, for parameter sets used in this test, the relative errors are reduced to the half. The error reduction factor may vary with mixture parameters and the used random samples. However, the most important gain is the robustness of the proposed method.

Table 4: Average estimation errors obtained using the classical EM algorithm and the updated EM with projected data for 150 runs with different random samples of size n=3000 drawn using 15 different parameter sets.

	μ error	π error	Σ error	CER
EM Raw Data	0.028	0.015	0.113	5.04%
EM Projected Data	0.015	0.012	0.052	5.74%

6. Asset returns modeling with singular Gaussian mixture

6.1. Historical context

The normal distribution is widely used to model the financial asset returns. This choice is motivated by the simplicity of this distribution and its coherence with the mean-variance paradigm [17]. However, two major problems exist within this model. First, as shown by many studies, the returns of a single asset are leptokurtic (fat tails) and asymmetric (see [10], [2]). Second, when several assets are considered, the covariance matrix of the returns distribution may be singular or ill-conditioned (see [4],[23]). In order to tackle the first problem, many authors proposed to use a finite mixture of Gaussian distributions ([11], [27], [1], [20]). Buser ([4]) proposed an extension of the mean-variance model of portfolio selection to solve the second issue. More recently, other authors ([23], [26], [13], [21]), have given some additional results concerning the portfolio selection when the covariance matrix is singular.
Moreover, the problem of portfolio optimization when the returns have a Gaussian mixture distribution was

addressed by [3]. In their work, they introduced an approach which improves the standard mean-variance procedure. The proposed methodology is reduced to three steps :

1. Model calibration from historical data (parameters estimation).
2. Solving a three-dimensional family of linear quadratic programs to obtain the efficient frontier.
3. Solving a three-dimensional non-linear program to determine the optimal portfolio.

6.2. Proposed model

In this subsection, we propose to use a mixture of singular Gaussian distributions to address simultaneously the two problems previously discussed : singularity of the covariance matrix and non-normality of the returns. The mixture components are associated to different market regimes. Here, we assume the existence of two regimes corresponding to bear and bull markets.
Consider d risky financial assets and let $R = (R_1, ..., R_d)^T$ be their random rates of return. The random vector R has a distribution with density function :

$$f(y) = \pi_D f_D(y|\mu_D, \Sigma_D) + \pi_T f_D(y|\mu_T, \Sigma_T), \tag{29}$$

This density function describes a market with two regimes :

- A "Distressed" market regime where the returns vector R has a singular Gaussian distribution with mean μ_D and covariance matrix Σ_D.

- A "Tranquil" market regime where the returns vector R has a singular Gaussian distribution with mean μ_T and covariance matrix Σ_T.

Note that π_D and π_T are the mixing weights ($\pi_T + \pi_D = 1$). Both covariance matrices Σ_D and Σ_T are singular with rank $r < d$.
Within this model, the methodology proposed in [3] remains applicable. Hence, the EM algorithm described in Section 4 can be used to achieve the first step (parameters estimation). The second and third steps remain unchanged. The optimal portfolio depends on the chosen objective function. A detailed analysis and further research are necessary to reveal more properties for the selected portfolios.

7. Conclusion

In this paper, we have introduced the finite singular multivariate Gaussian mixture model. The main objectives were to estimate the model parameters and give a clustering approach of the observed data. We proposed an extension of the EM algorithm in order to meet these objectives. More precisely, we gave, with a detailed proof, new formulas used in the maximization step of this algorithm. The performance of the proposed algorithm was evaluated through different numerical experiments. The obtained estimates are very close to the true parameters used in samples generation. Moreover, when covariance matrices are ill-conditioned, the use of the singular Gaussian mixture model with projected data addresses instability problems faced in the classical EM algorithm and improves the estimators quality and the model robustness. Finally, we illustrated a possible application of the proposed model to portfolio selection when the asset returns have a singular covariance matrix. We demonstrated the usefulness of the mixture of singular normal distributions in this context as it simultaneously addresses the non-normality of the returns and the singularity of the covariance matrix.

Our future perspective includes the theoretical investigation of mixture of singular multivariate normal distributions when the components are concentrated on different affine subspaces. Hence, several cases should be considered depending on the intersection of such subspaces.

References

[1] T. Ané and C. Labidi. Revisiting the finite mixture of gaussian distributions with application to futures markets. *Journal of Futures Markets*, 21(4):347–376, 2001.

[2] R. M. Bookstaber and J. B. McDonald. A general distribution for describing security price returns. *The Journal of Business*, 60(3):401–424, 1987.

[3] I. Buckley, D. Saunders, and L. Seco. Portfolio optimization when asset returns have the gaussian mixture distribution. *European Journal of Operational Research*, 185(3):1434–1461, 2008.

[4] S. A. Buser. Mean-variance portfolio selection with either a singular or nonsingular variance-covariance matrix. *The Journal of Financial and Quantitative Analysis*, 12(3):347–361, 1977.

[5] N. A. Campbell and R. J. Mahon. A multivariate study of variation in two species of rock crab of the genus leptograpsus. *Australian Journal of Zoology*, 22(3):417–425, 1974.

[6] G. Celeux and G. Govaert. Gaussian parsimonious clustering models. *Pattern Recognition*, 28(5):781–793, 1995.

[7] I. D. Coope and P. F. Renaud. Trace inequalities with application to orthogonal regression and matrix nearness problems. *Journal of inequalities in pure and applied mathematics*, 10(92), 2009.

[8] A. P. Dempster, N. M. Larid, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38, 1977.

[9] M. L. Eaton. *Multivariate statistics: a vector space approach*. Institute of Mathematical Statistics, 2007.

[10] E. F. Fama. The behavior of stock-market prices. *The Journal of Business*, 38(1):34–105, 1965.

[11] M. Haas and S. Mittnik. Portfolio selection with common correlation mixture models. In *Risk Assessment*, pages 47–76. Springer, 2009.

[12] R. J. Hathaway. A constrained formulation of maximum-likelihood estimation for normal mixture distributions. *The Annals of Statistics*, 13(2):795–800, 1985.

[13] C. Jiang. *Efficient Subset Selection in Large-Scale Portfolio with Singular Covariance Matrix*, pages 1443–1453. Springer Berlin Heidelberg, 2014.

[14] C. G. Khatri. Some results for the singular normal multivariate regression models. *Sankhyā: The Indian Journal of Statistics*, pages 267–280, 1968.

[15] W. J. Krzanowski, P. Jonathan, W. V. McCarthy, and M. R. Thomas. Discriminant analysis with singular covariance matrices: Methods and applications to spectroscopic data. *Journal of the Royal Statistical Society.*, 44(1):101–115, 1995.

[16] S. X. Lee, S.-K. Ng, and G. J. McLachlan. Chapter 4 - finite mixture models in biostatistics. In A. S. S. Rao, S. Pyne, and C. Rao, editors, *Disease Modelling and Public Health, Part A*, volume 36, pages 75 – 102. Elsevier, 2017.

[17] H. Markowitz. Portfolio selection. *The Journal of Finance*, 7(1):77–91, 1952.

[18] K. Masmoudi and A. Masmoudi. Singular gaussian graphical models: Structure learning. *Communications in Statistics-Simulation and Computation*, 47(10):3106–3117, 2018.

[19] G. Mclachlan and D. Peel. *Finite Mixture Models*. John Wiley & Sons, 2000.

[20] M. S. Paolella. Multivariate asset return prediction with mixture models. *The European Journal of Finance*, 21(13-14):1214–1252, 2015.

[21] D. Pappas and K. Kiriakopoulos. Optimal portfolio selection with singular covariance matrix. *International Mathematical Forum*, 5(47):2305–2318, 2010.

[22] R. Penrose. A generalized inverse for matrices. *Mathematical Proceedings of the Cambridge Philosophical Society*, 51(3):406–413, 1955.

[23] P. J. Ryan and J. Lefoll. A comment on mean-variance portfolio selection with either a singular or a non-singular variance-covariance matrix. *The Journal of Financial and Quantitative Analysis*, 16(3):389–395, 1981.

[24] M. S. Srivastava and C. Khatri. *An introduction to multivariate statistics*. North Holland, 1979.

[25] M. S. Srivastava and D. von Rosen. Regression models with unknown singular covariance matrix. *Linear Algebra and its Applications*, 354(1-3):255–273, 2002.

[26] J. Vörös. The explicit derivation of the efficient portfolio frontier in the case of degeneracy and general singularity. *European Journal of Operational Research*, 32(2):302–310, 1987.

[27] J. Wang. Generating daily changes in market variables using a multivariate mixture of normal distributions. *Proceeding of the 2001 Winter Simulation Conference*, 2001.

[28] M. Zitouni, M. Zribi, and A. Masmoudi. Asymptotic properties of the estimator for a finite mixture of exponential dispersion models. *Filomat*, 32(19):6575–6598, 2018.