Article

Statistical Inference for Finite Mixture of Matrix Variate t-distribution

Xinyu Yang [1,a,†,‡], Leigang Dong [1,b], Zhichuan Zhu [2,‡] and Weisan Wu [1,a*]

[1,a] School of Mathematics and Statistics, Baicheng Normal University, Baicheng, China; wuws009@outlook.com
[1,b] School of Computation Science, Baicheng Normal University, Baicheng, China; lgdong010@163.com
[2] School of Mathematics and Statistics, Liaoning University, Shenyang, China; zhuzcnh@126.com
* Correspondence: wuws009@outlook.com. School of Mathematics and Statistics, Baicheng Normal University, 57 Zhongxing West Road, Taobei District, Baicheng , Jilin,China.
† Current address: 57 Zhongxing West Road, Taobei District, Baicheng , Jilin,China.
‡ These authors contributed equally to this work.

Abstract: In the era of big data with increasingly complex data structures and ever-larger data scales, matrix-type data are becoming highly valued and their applications in the fields of medicine, industry, education, geography, and astronomy are growing in extent. In recent years, significant progress has been made in the practical use of matrix variable t-distribution finite mixture models for handling data in order to address the issues of multi-subgroup structures and long data tails. In this paper, the expectation-maximization (EM) algorithm with penalized maximum likelihood is proposed to resolve the problem of the unbounded nature of the likelihood function applied to the model by considering the degeneracy of the variance-covariance matrix of this model. Our data were analyzed through simulations and real data, and the results demonstrate that our model is effective in both preventing the likelihood function from being unbounded and in ensuring the accuracy of the estimated parameters of the EM algorithm.

Keywords: Matrix variate distribution; Mixture models; EM-algorithm; Penalized likelihood

1. Introduction

Finite mixture models are of great importance in the field of statistics, with the renowned statistician Pearson advocating the use of mixture models to study data with subgroup properties since the introduction of the mixed-normal model. For more than a century, it has been proven through persistent research that the application of mixture models to heterogeneous data and multiple subgroups of data can achieve remarkable results in areas including classification, clustering, and machine-learning (see McLachlan and Basford 1988[1] and Fruhwirth-Schnatter 2006 [21] for more information on this area of research).

The matrix variable t-distribution was proposed by Gupta A.K. in 2013[20]. It can be considered as a special case of the Pearson type VII distribution that can be regarded as a matrix variable normal distribution when the degrees of freedom tend to infinity. Doğru, Fatma Zehra[24] provided parameter estimates for the finite mixture matrix t-distribution (MMVT) in 2016 using the classic expectation-maximization (EM) algorithm to obtain asymptotic solutions of the parameters after iteration. They serve as a useful tool for describing the long-tail property. We observed that although the long-tail problem with multiple subgroups can be effectively described using MMVT, the MMVT model shares the same problems of unbounded likelihood function, infinite Fisher information, and unsatisfied strong identifiability as the normal mixed model. Therefore, the maximum likelihood estimate is difficult to define in the case of likelihood function degeneracy.

For this reason, we had drawn on the idea of a penalized likelihood in the 2009 finite normal mixed model [3] to penalize the variance-covariance matrices for the rows and columns of the components in the MMVT model to prevent degeneracy of its likelihood

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 January 2023 doi:10.20944/preprints202301.0524.v1

function [13][19],as well as designed an ECM-type algorithm that combines EM with channel matching (CM) to ensure the validity of the estimation through a hierarchical representation of the random variables and a Bayesian approach[2][3][4][17].

The remainder of this article is organized as follows: We show the foundation of matrix variate t distribution in the section 2. In Section 3, the MMVT model is introduced and its degenerate properties are discussed, along with the design of the penalized ECM type algorithm for the MMVT model. In Section 4, simulations are performed with the model and actual data is analyzed to verify the effectiveness of the algorithm. Finally, the ideas behind our model are discussed and future prospects are discussed in Section 5.

2. Preliminaries and Methods

In this section, the definition of the matrix variable t-distribution in Section 2.1, followed by the introduction of the concept of MMVT in Section 2.2.

2.1. Matrix variate t distribution

In this section, we introduce a novel mixture modeling using an original matrix variate t (MVT) distribution[19][22][20]. The density function of the normalized MVT distribution is given by

The $k \times n$ dimensional random matrix X called matrix variate t distribution with mean matrix M and row covariance Ψ and column covariance Σ, if the pdf of X is given by

$$f_{k,n}(X, M, \Psi \otimes \Sigma) = \frac{\Gamma(\frac{kn+v}{2})}{(\pi v)^{\frac{kn}{2}}\Gamma(\frac{v}{2})|\Sigma|^{\frac{k}{2}}|\Psi|^{\frac{n}{2}}}[1 + \frac{1}{v}\exp\{tr(\Psi^{-1}(X-M)\Sigma^{-1}(X-M)^T\}]^{-\frac{kn+v}{2}} \quad (1)$$

M is a $k \times n$ constant matrix, Σ and Ψ are $k \times k$ and $n \times n$ positive definite matrix.$E(X) = M, Var(vec(X)) = \Psi \otimes \Sigma$ are mean and covariance respectively.If variable X has the density (1), we denoted by $X \sim MVT_{k,n}(M, \Sigma, \Psi, v)$. The following lemma gives the relationship between the matrix variate t distribution and the t distribution, which will be used later in the parameter estimation. (Theorem 6.1 of [12]) $X \sim MVT_{k,n}(M, \Sigma, \Psi, \nu)$ if and only if $vec(X) \sim t_{kn}(vec(M), \Psi \otimes \Sigma, \nu)$.Further more,$E[X] = M$, for $\nu > 1$ and $Cov(X) = \frac{\nu}{\nu-2}(\Psi \otimes \Sigma)$, for $\nu > 2$.

2.2. The MMVT models

Given i.i.d. $k \times n$ observations $Y_1, Y_2, \cdots, Y_m$ from the mixture of matrix variate t (MMVT) distribution with G subgroups[24][14], the pdf of the MMVT is as follows:

$$f_{MMVT}(Y|\Theta) = \sum_{g=1}^{G} \pi_g f_{MVT}(Y|\theta_g), \quad (2)$$

where θ_g is the set of parameters for the g-th components, we write $\theta = (\theta_1, \cdots, \theta_G)$,where,$\theta_g = (\pi_g, M_g, \Sigma_g, \Psi_g, \nu_g), g = 1, \cdots, G$, or we can write the parameter space as

$$\Theta = \{\pi_1, \cdots, \pi_{g-1}, M_1, \cdots, M_G, \Sigma_1, \cdots, \Sigma_G, \Psi_1, \cdots, \Psi_G, \nu_1, \cdots, \nu_G\}. \quad (3)$$

where $0 \leq \pi_g \leq 1, \sum_{g=1}^{G} \pi_g = 1$ is mixing weight, $M_g \in \mathbb{R}^{k\times n}, \Sigma_g \in \mathbb{R}^{k\times k}, \Psi_g \in \mathbb{R}^{n\times n}$ are local parameter,row variance-covariance and column variance-covariance parameters of the g-th components respectively.

The matrix variate t distribution have additional representation as follows:

$$X = M + \Psi^{\frac{1}{2}} Z \Sigma^{\frac{1}{2}} U^{-\frac{1}{2}},$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 January 2023 doi:10.20944/preprints202301.0524.v1

where $Z \sim N_{k,n}(0, I_k, I_n), U \sim Gamma(\frac{\nu}{2}, \frac{\nu}{2})$. Furthermore, we can use hierarchical representation

$$\begin{aligned} &X|U = u \sim N_{k,n}(M, u^{-1}\Psi, \Sigma) \\ &U \sim Gamma(\frac{\nu}{2}, \frac{\nu}{2}) \\ &U|X \sim Gamma(\frac{kn+\nu}{2}, \frac{1}{2}(\nu + tr\{\Psi^{-1}(X-M)^T\Sigma^{-1}(X-M)\})). \end{aligned} \tag{4}$$

Take conditional expectation for the random variable $U|X$ and $\log U|X$

$$\begin{aligned} E[U|X] &= \frac{(kn+\nu)/2}{(\nu + tr\{\Psi^{-1}(X-M)^T\Sigma^{-1}(X-M)\})/2} \\ &= \frac{kn+\nu}{\nu + tr\{\Psi^{-1}(X-M)^T\Sigma^{-1}(X-M)\}} \\ E[\log U|X] &= DG(\frac{kn+\nu}{2}) - \log\frac{1}{2}(\nu + tr\{\Psi^{-1}(X-M)^T\Sigma^{-1}(X-M)\}) \end{aligned} \tag{5}$$

where $DG(x) = \frac{d}{dx}\log\Gamma(x)$.

3. PMLE

Firstly, we show illustration of how an unbounded likelihood function for MMVT is possible through a concrete example provided in Section 3.1. Secondly, a new method of a penalized likelihood function is given in Section 3.2, and finally, an EM algorithm is designed in Section 3.3 to estimate the parameters based on the penalized likelihood function.

3.1. Undesirable properties of MMVT models and the PMLE

We now present undesirable property of MMVT models, and then introduce the penalty functions to counter these problems.

Suppose $Y_1, Y_2, \cdots, Y_n$ from a random sample of size n from the MMVT model (2), the log-likelihood function is given by

$$\ell_n(\Theta) = \sum_{i=1}^{m}\log f(Y_i;\theta) = \sum_{i=1}^{m}\log\left\{\sum_{g=1}^{G}\pi_g f_{MVT}(Y_i;\theta_g)\right\}. \tag{6}$$

Given the sample $Y_1, Y_2, \cdots, Y_m$, let the location parameter of first component be $M_1 = Y_1$. Clearly, when $|\Sigma_1| \to 0, |\Psi_1| \to 0$, we have

$$\begin{aligned} f_{MVT}(Y_i;\theta_1) =& \frac{\Gamma(\frac{kn+v_1}{2})}{|\Sigma_1|^{\frac{k}{2}}|\Psi_1|^{\frac{n}{2}}(\pi v_1)^{\frac{kn}{2}}\Gamma(\frac{v_1}{2})}[1 + \frac{1}{v_1}\exp\{tr(\Psi_1^{-1}(Y_1-M_1)\Sigma_1^{-1}(Y_1-M_1)^T\}]^{-\frac{kn+v_1}{2}} \\ =& \frac{\Gamma(\frac{kn+v_1}{2})}{|\Sigma_1|^{\frac{k}{2}}|\Psi_1|^{\frac{n}{2}}(\pi v_1)^{\frac{kn}{2}}\Gamma(\frac{v_1}{2})} \to \infty. \end{aligned} \tag{7}$$

Let the other component parameters be $M_g = M_0, \Sigma_g = I_k, \Psi_g = I_n$ for $g = 2, 3, \cdots, G$,where $M_0 \neq Y_i$ for any $i = 2, 3, \cdots, m$. It is clear that the component density $f_{MVT}(Y_i;\theta_g)$ is bounded by

$$\begin{aligned} f_{MVT}(Y_i;\theta_g) =& \frac{\Gamma(\frac{kn+v_g}{2})}{|\Sigma_g|^{\frac{k}{2}}|\Psi_g|^{\frac{n}{2}}(\pi v_g)^{\frac{kn}{2}}\Gamma(\frac{v_g}{2})}[1 + \frac{1}{v_g}\exp\{tr(\Psi_g^{-1}(Y_i-M_g)\Sigma_k^{-1}(Y_i-M_g)^T\}]^{-\frac{kn+v_g}{2}} \\ \leq& \frac{\Gamma(\frac{kn+v_g}{2})}{|\Sigma_g|^{\frac{k}{2}}|\Psi_g|^{\frac{n}{2}}(\pi v_g)^{\frac{kn}{2}}\Gamma(\frac{v_g}{2})} \end{aligned} \tag{8}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 January 2023 doi:10.20944/preprints202301.0524.v1

for $i = 2, 3, \cdots, m, g = 2, 3, \cdots, G$.

Let $\theta_0 = (M_0, I_k, I_n, \nu_0)$, and the mixing proportion be $\pi_1 = 1/2$, and $\pi_g = \frac{1}{2(G-1)}$ for $g = 2, 3, \cdots, G$, we have

$$\begin{aligned}\ell_n(\Theta) &= \log\left\{\sum_{g=1}^{G}\pi_g f_{MVT}(Y_1;\theta_g)\right\} + \sum_{i=2}^{m}\log\left\{\sum_{g=1}^{G}\pi_g f_{MVT}(Y_i;\theta_g)\right\} \\ &\geq \log\{\pi_1 f_{MVT}(Y_1;\theta_1)\} + \sum_{i=2}^{m}\log\left\{\sum_{g=2}^{G}\pi_g f_{MVT}(Y_i;\theta_0)\right\} \\ &= -m\log 2 + \log f_{MVT}(Y_1;\theta_1) + \sum_{i=2}^{m}\log f_{MVT}(Y_i;\theta_0) \to \infty\end{aligned}$$

as $|\Sigma_1| \to 0, |\Psi_1| \to 0$. Therefore, the MLE of Θ is inconsistent.

3.2. The PMLE of parameter in MMVT models

In this part, to overcome both of the likelihood function's undesirable properties, we introduce the penalty function to the log-likelihood, which is defined as

$$\ell_m(\Theta) = p_m(\Psi) + p_m(\Sigma) = \sum_{g=1}^{G}\tilde{p}_m(\Psi_g) + \sum_{g=1}^{G}\tilde{p}_m(\Sigma_g),$$

where $p_m(\Psi), p_m(\Sigma)$ are the penalty terms for covariance matrix parameters to avoid the likelihood function are unbounded.

Then, we define the penalized log-likelihood function as follows:

$$p\ell_m(\Theta) = \ell_m(\Theta) + p_m(\Sigma) + p_m(\Psi), \tag{9}$$

and the PMLE of Θ would be obtained by $\tilde{\Theta} = \arg\max_{\Theta} p\ell_m(\Theta)$. The PMLE also can be computed using ECM algorithm. In order to obtain the consistent estimator, we select $\tilde{p}_m(\Sigma_g), \tilde{p}_m(\Psi_g)$ such that it goes to negative infinity when $|\Sigma_g|$ and $|\Psi_g|$ goes to 0 or to infinity. There are several options for the penalty function. For example, we can assume that the expression for the penalty function is $\tilde{p}_m(\Sigma_g) = -a_m(tr(S_g(\Sigma)\Sigma_g^{-1}) + \log(|\Sigma_g|))$, $\tilde{p}_m(\Psi_g) = -b_m(tr(S_g(\Psi)\Psi_g^{-1}) + \log(|\Psi_g|)$, where the $S_g(\Sigma)$ and $S_g(\Psi)$ are the sample row and column variance-covariance of the g-th components respectively.

3.3. The Penalized EM Algorithms

Like Geoffrey Z. Thompson(2020)[25],we known the ECM algorithm then proceeds as follows: Consider the complete data $(Y, Z) = \{Y_i, Z_i\}_{i=1}^{m}$, where the latent component-indicators vector $Z_i = (Z_{1i}, Z_{2i}, \cdots, Z_{Gi})$ follows a multinomial distribution with 1 trial and cell probabilities $\pi_1, \pi_2, \cdots, \pi_G$. Write it as $Z_i \sim Multi(1; \pi_1, \pi_2, \cdots, \pi_G)$. Note that $Z_1, Z_2, \cdots Z_m$ are mutually independent.

By the hierarchical representation

$$\begin{aligned}Y_i|U_i = u_i, Z_{gi} = 1 &\sim N_{k,n}(M_i, u_i^{-1}\Psi_g, \Sigma_g), \\ U_i|Z_{gi} = 1 &\sim Gamma(\frac{v_g}{2}, \frac{v_g}{2}) \\ Z_{gi} &\sim Multi(1:\pi_1, \pi_G).\end{aligned}$$

We denote the condition expectation on current iteration parameters as

$$\alpha_{gi}^{(t)} \triangleq E[U_i|Y_i, Z_{gi} = 1] = \frac{kn + v_g}{v_g + tr(\Psi_k^{-1}(Y_i - M_g)\Sigma_k^{-1}(Y_i - M_g)^T)}$$

$$\beta_{gi}^{(t)} \triangleq E[\log(U_i)|Y_i, Z_{gi} = 1] = \frac{d}{dv_g}\Gamma(\frac{kn + v_g}{2}) - \log(\frac{1}{2}(v_g + tr(\Psi_g^{-1}(Y_i - M_g)\Sigma_g^{-1}(Y_i - M_g)^T))).$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 January 2023 doi:10.20944/preprints202301.0524.v1

E-step: The Q function can be written as

$$\begin{aligned}
Q(\Theta|\Theta^{(t)}) =& E(p\ell_n^c(\Theta)|Y,\Theta^{(t)}) \\
=& \sum_{i=1}^{m}\sum_{g=1}^{g} E[\hat{z}_{gi}^{(t)}|Y_i]\Big\{ \log(\pi_g) - \frac{k}{2}\log|\Sigma_g| - \frac{n}{2}\log|\Psi_g| \\
&+ \frac{kn}{2}\beta_{gi}^{(t)} - \frac{v_g}{2}\log(\frac{v_g}{2}) - \log(\Gamma(\frac{v_g}{2})) + (\frac{v_g}{2}-1)\beta_{gi}^{(t)} - \frac{v_g}{2}\alpha_{gi}^{(t)} \\
&- \frac{1}{2}tr(\alpha_{gi}^{(t)}(\Psi_g)^{-1}(Y_i-M_g)^T\Sigma_g^{-1}(Y_i-M_g)) - a_n(tr(S_m^2\Sigma_g^{-1}) + \log(|\Sigma_g|(S_m^2)^{-1}) \\
&- b_n(tr(S_m^2\Psi_g^{-1}) + \log(|\Psi_g|(S_m^2)^{-1}).
\end{aligned}$$

M-Step: Let us maximize $Q(\Theta|\Theta^{(t)})$ with respect to Θ under the restriction with $\sum_{g=1}^{G}\pi_g = 1$.

1. Update $\pi_g^{(t)}$ by $\pi_g^{(t+1)} = m^{-1}\sum_{i=1}^{m}\hat{z}_{gi}^{(t)}$.

2. Update $M_g^{(t)}$ by

$$M^{(t+1)} = \frac{\sum_{i=1}^{m}\hat{z}_{gi}^{(t)}\alpha_{gi}^{(t)}Y_i}{\sum_{i=1}^{m}\hat{z}_{gi}^{(t)}}.$$

3. Fix $M_g^{(t)} = M_g^{(t+1)}$, then update $\Psi_g^{(t+1)}, \Sigma_g^{(t+1)}$ by

$$\begin{aligned}
(\Psi_g^{(t+1)}, \Sigma_g^{(t+1)}) =& \arg\max_{(\Psi_g^{(t)},\Sigma_g^{(t)})} Q(\Theta|\Theta^{(t)}) \\
=& \arg\max_{(\Psi_g^{(t)},\Sigma_g^{(t)})} \Bigg\{ \sum_{i=1}^{m}\sum_{g=1}^{G} E[\hat{z}_{gi}^{(t)}|Y_i]\Big\{ \log(\pi_g) - \frac{kn}{2}\log(2\pi) - \frac{k}{2}\log|\Sigma_g^{(t)}| - \frac{n}{2}\log|\Psi_g^{(t)}| \\
&+ \frac{kn}{2}\beta_{gi}^{(t)} - \frac{v_g}{2}\log(\frac{v_g}{2}) - \log(\Gamma(\frac{v_g}{2})) + (\frac{v_g}{2}-1)\beta_{gi}^{(t)} - \frac{v_g}{2}\alpha_{gi}^{(t)} \\
&- \frac{1}{2}tr(\alpha_{gi}^{(t)}(\Psi_g^{(t)})^{-1}(Y_i-M_g^{(t+1)})^T\Sigma_g^{-1}(Y_i-M_g^{(t+1)}))\Big\} \\
&- a_n(tr(S_m^2(\Sigma_g^{(t)})^{-1}) + \log(|\Sigma_g^{(t)}|(S_m^2)^{-1}) \\
&- b_n(tr(S_m^2(\Psi_g^{(t)})^{-1}) + \log(|\Psi_g^{(t)}|(S_m^2)^{-1})\Bigg\}.
\end{aligned}$$

4.The conditional maximization of ν_g given $(M_g^{(t+1)}, \Sigma_g^{(t+1)}, \Psi_g^{(t+1)})$ through solve the following equation

$$\frac{kn}{2}\beta_{gi}^{(t)} - \frac{v_g}{2}\log(\frac{v_g}{2}) - \log(\Gamma(\frac{v_g}{2})) + (\frac{v_g}{2}-1)\beta_{gi}^{(t)} - \frac{v_g}{2}\alpha_{gi}^{(t)} = 0.$$

For the given initial value $\Theta^{(0)}$, we can repeat the E-step and M-step and stop the iteration until the penalized log-likelihood $p\ell_m(\Theta)$ does not increase or remains in a satisfactory condition $|p\ell_m(\Theta^{(t+1)}) - p\ell_m(\Theta^{(t)})| \leq 10^{-3}$. Then, we can obtain the PMLE of the MSNM models. Generally, the EM algorithm requires the setting of more initial values to allow the model's estimator to arrive at the maximum.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 January 2023 doi:10.20944/preprints202301.0524.v1

4. Numerical studies

Our experimental implemented in R version 4.2.1 based on R (R Core Team 2019) package MatrixMixtures and MatrixLDA to get PMLE results, and all experiments are performed on a laptop with Intel Core i7-9750H and 2.60 GHz CPU and 6.00 RAM.

4.1. Simulation studies

For convenient, we denote elements of parameters $M = (m_{i,j}^{(g)}), \Sigma = (\sigma_{i,j}^{(g)}), \Psi = (\psi_{i,j}^{(g)}), \nu = \nu^{(g)}$ for the g-th component respectively. We given three types model of MMVT as follows,

Model I:

$$0.6MVT_{2,2}(M_1, \Sigma_1, \Psi_1, \nu_1) + 0.4MVT_{2,2}(M_2, \Sigma_2, \Psi_2, \nu_2),$$

where,

$$M_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, M_2 = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}, \Sigma_1 = \Psi_1 = \begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix},$$
$$\Sigma_2 = \Psi_2 = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}, \nu_1 = \nu_2 = 5.$$

Model II:

$$0.3MVT_{3,3}(M_1, \Sigma_1, \Psi_1, \nu_1) + 0.7MVT_{3,3}(M_2, \Sigma_2, \Psi_2, \nu_2),$$

where,

$$M_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{pmatrix}, M_2 = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \Sigma_1 = \Psi_1 = \begin{pmatrix} 1.4 & 1.5 & 1.2 \\ 1.5 & 1.7 & 1.5 \\ 1.2 & 1.5 & 1.9 \end{pmatrix},$$
$$\Sigma_2 = \Psi_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \nu_1 = \nu_2 = 2.$$

Model III:

$$0.3MVT_{3,3}(M_1, \Sigma_1, \Psi_1, \nu_1) + 0.2MVT_{3,3}(M_2, \Sigma_2, \Psi_2, \nu_2) + 0.5MVT_{3,3}(M_3, \Sigma_3, \Psi_3, \nu_3),$$

where,

$$M_1 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, M_2 = \begin{pmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{pmatrix}, M_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$
$$\Sigma_1 = \Psi_1 = \begin{pmatrix} 1.4 & 1.5 & 1.2 \\ 1.5 & 1.7 & 1.5 \\ 1.2 & 1.5 & 1.9 \end{pmatrix}, \Sigma_2 = \Psi_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$
$$\Sigma_3 = \Psi_3 = \begin{pmatrix} 2 & 1 & 3 \\ 1 & 4 & 2 \\ 3 & 2 & 5 \end{pmatrix}, \nu_1 = 5, \nu_2 = 10, \nu_3 = 20.$$

Consider the penalized function as follows:

$$\tilde{p}_m(\Sigma_g) = -a_m(tr(S_g(\Sigma)\Sigma_g^{-1}) + \log(|\Sigma_g|)), \tilde{p}_m(\Psi_g) = -b_m(tr(S_g(\Psi)\Psi_g^{-1}) + \log(|\Psi_g|).$$

The constant in penalized term as $a_m = 1/m$, $b_m = 1/\log(m)$,sample size as $g \times 100, g \times 200$, and $g \times 500$,we set $g = 2, k = n = 2$ in model I, set $g = 2, k = n = 3$ in model II and

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 January 2023 doi:10.20944/preprints202301.0524.v1

set $g = 3, k = n = 3$ in model III. The initial parameter size is $g \times (2kn + \frac{kn\times(kn+1)}{2} + 1)$, so we need the number of estimate parameter is $g \times (2kn + \frac{kn\times(kn+1)}{2} + 1) + 1$ because of the weight parameter's freedom is $g - 1$. We repeat 100 times for every models and compute (MSE).

From Table 1 we known the MSE of all parameters are small. From Table 2, we can see MSE of parameter $\nu^{(1)}$ in component 1 and $\nu^{(2)}$ are 1.235, 0.630, 0.507 and 1.354, 0.712, 0.143. These results all showed that MSE decreased with the increase of sample size. Similar results are presented in Table 3 of the 3 components mixtures model.

Table 1. MLE and PMLEs:model I

	n=100		n=200		n=500	
	m	mse	m	mse	m	mse
model I component 1						
$\pi = 0.6$	0.600	0.002	0.596	0.001	0.602	0.000
$m_{1,1}^{(1)} = 1$	1.166	0.306	1.151	0.331	1.030	0.020
$m_{2,2}^{(1)} = 1$	1.138	0.342	1.104	0.294	1.019	0.014
$\sigma_{1,11}^{(1)} = 3$	2.756	1.492	2.650	0.231	2.917	0.141
$\sigma_{1,12}^{(1)} = 1$	0.823	1.293	0.781	0.094	0.935	0.047
$\sigma_{1,22}^{(1)} = 3$	2.699	1.410	2.827	0.331	2.917	0.134
$\psi_{1,11}^{(1)} = 3$	2.675	1.692	2.832	0.161	2.944	0.101
$\psi_{1,12}^{(1)} = 1$	0.841	1.193	0.799	0.114	0.945	0.037
$\psi_{1,22}^{(1)} = 3$	2.704	1.349	2.866	0.312	2.922	0.106
$\nu^{(1)} = 5$	4.278	0.250	4.638	0.320	4.451	0.259
model I component 2						
$m_{1,1}^{(2)} = 2$	2.087	0.293	2.048	0.328	2.038	0.023
$m_{2,2}^{(2)} = 2$	2.114	0.348	2.039	0.295	2.024	0.017
$\sigma_{2,11}^{(2)} = 2$	1.717	0.331	1.797	0.195	1.947	0.106
$\sigma_{2,12}^{(2)} = 1$	0.711	0.232	0.754	0.114	0.938	0.051
$\sigma_{2,22}^{(2)} = 2$	1.721	0.329	1.813	0.217	1.955	0.106
$\psi_{2,11}^{(2)} = 2$	1.720	0.301	1.800	0.190	1.947	0.103
$\psi_{2,12}^{(2)} = 1$	0.717	0.202	0.791	0.110	0.950	0.047
$\psi_{2,22}^{(2)} = 2$	1.733	0.314	1.841	0.207	1.969	0.089
$\nu^{(2)} = 5$	2.561	0.441	2.704	0.421	2.981	0.340

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 January 2023 doi:10.20944/preprints202301.0524.v1

Table 2. MLE and PMLEs:model II

	n=100		n=200		n=500	
	m	mse	m	mse	m	mse
model II component 1						
$\pi = 0.3$	0.309	0.005	0.300	0.002	0.304	0.001
$m_{1,1}^{(1)} = 1$	1.166	0.160	1.147	0.097	1.056	0.037
$m_{2,2}^{(1)} = 2$	2.038	0.270	2.161	0.088	2.055	0.033
$m_{3,3}^{(1)} = 3$	2.977	0.308	3.095	0.068	3.020	0.033
$\sigma_{1,1}^{(1)} = 1.4$	1.182	0.253	1.177	0.117	1.314	0.063
$\sigma_{1,2}^{(1)} = 1.5$	1.200	0.344	1.239	0.131	1.402	0.069
$\sigma_{1,3}^{(1)} = 1.2$	0.924	0.314	0.944	0.148	1.106	0.062
$\sigma_{2,2}^{(1)} = 1.7$	1.537	0.691	1.402	0.153?	1.594	0.078
$\sigma_{2,3}^{(1)} = 1.5$	1.179	0.667	1.223	0.187	1.409	0.083
$\sigma_{3,3}^{(1)} = 1.9$	1.729	0.606	1.712	0.283	1.868	0.148
$\psi_{1,1}^{(1)} = 1.4$	1.191	0.251	1.185	0.101	1.360	0.058
$\psi_{1,2}^{(1)} = 1.5$	1.310	0.332	1.251	0.111	1.424	0.057
$\psi_{1,3}^{(1)} = 1.2$	0.985	0.301	0.947	0.119	1.134	0.058
$\psi_{2,2}^{(1)} = 1.7$	1.541	0.672	1.456	0.135	1.603	0.064
$\psi_{2,3}^{(1)} = 1.5$	1.188	0.641	1.231	0.164	1.433	0.074
$\psi_{3,3}^{(1)} = 1.9$	1.751	0.593	1.742	0.264	1.890	0.133
$\nu^{(1)} = 2$	2.214	1.235	2.040	0.630	2.003	0.507
model II component 2						
$m_{1,1}^{(2)} = 2$	1.801	0.182	1.917	0.109	1.991	0.026
$m_{2,2}^{(2)} = 1$	1.003	0.276	0.925	0.135	0.986	0.047
$m_{3,3}^{(2)} = 1$	1.186	0.222	1.122	0.116	1.004	0.034
$\sigma_{1,1}^{(2)} = 1$	1.051	0.075	1.031	0.046	1.009	0.014
$\sigma_{1,2}^{(2)} = 0$	0.184	0.121	0.036	0.070	-0.001	0.015
$\sigma_{1,3}^{(2)} = 0$	0.209	0.210	0.136	0.087	0.025	0.035
$\sigma_{2,2}^{(2)} = 2$	2.146	0.317	2.142	0.226	2.033	0.063
$\sigma_{2,3}^{(2)} = 0$	-0.133	0.527	-0.027	0.148	-0.026	0.055
$\sigma_{3,3}^{(2)} = 3$	2.832	0.664	2.783	0.414	2.981	0.171
$\psi_{1,1}^{(2)} = 1$	1.049	0.073	1.027	0.041	1.007	0.013
$\psi_{1,2}^{(2)} = 0$	0.181	0.119	0.030	0.051	0.002	0.011
$\psi_{1,3}^{(2)} = 0$	0.201	0.201	0.132	0.078	0.019	0.027
$\psi_{2,2}^{(2)} = 2$	2.016	0.287	2.001	0.201	2.011	0.033
$\psi_{2,3}^{(2)} = 0$	0.145	0.515	-0.016	0.128	0.002	0.003
$\psi_{3,3}^{(2)} = 3$	2.912	0.114	2.923	0.111	2.986	0.094
$\nu^{(2)} = 2$	1.817	1.354	1.899	0.712	2.060	0.143

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 January 2023 doi:10.20944/preprints202301.0524.v1

Table 3. MLE and PMLEs:model III

	n=100		n=200		n=500	
	m	mse	m	mse	m	mse
$\pi_1 = 0.3$	0.266	0.007	0.280	0.014	0.298	0.014
$m_{1,1}^{(1)} = 0$	-0.695	1.262	-0.971	0.651	-0.751	1.257
$m_{2,2}^{(1)} = 0$	0.824	0.443	0.871	0.892	0.853	0.855
$m_{3,3}^{(1)} = 0$	0.636	0.576	0.411	1.429	0.404	1.459
$\sigma_{1,1}^{(1)} = 1.4$	0.868	0.613	1.114	0.730	1.100	0.553
$\sigma_{1,2}^{(1)} = 1.5$	0.421	0.242	0.356	0.447	0.457	0.212
$\sigma_{1,3}^{(1)} = 1.2$	0.431	0.518	0.647	0.992	0.593	0.854
$\sigma_{2,2}^{(1)} = 1.7$	1.181	1.040	1.701	1.527	1.650	1.555
$\sigma_{2,3}^{(1)} = 1.5$	0.416	0.403	0.475	0.679	0.435	0.563
$\sigma_{3,3}^{(1)} = 1.9$	1.667	1.033	2.190	1.843	2.141	1.928
$\psi_{1,1}^{(1)} = 1.4$	0.802	0.521	1.014	0.638	1.055	0.513
$\psi_{1,2}^{(1)} = 1.5$	0.421	0.242	0.356	0.447	0.457	0.212
$\psi_{1,3}^{(1)} = 1.2$	0.431	0.518	0.647	0.992	0.593	0.854
$\psi_{2,2}^{(1)} = 1.7$	1.181	1.040	1.701	1.527	1.650	1.555
$\psi_{2,3}^{(1)} = 1.5$	0.416	0.403	0.475	0.679	0.435	0.563
$\psi_{3,3}^{(1)} = 1.9$	1.667	1.033	2.190	1.843	2.141	1.928
$\nu^{(1)} = -1$	0.245	2.566	-0.309	3.443	-0.468	3.453
$\pi_2 = 0.2$	0.389	0.012	0.346	0.015	0.371	0.017
$m_{1,1}^{(2)} = -1$	-1.023	0.693	-0.801	1.495	-0.798	1.139
$m_{2,2}^{(2)} = -1$	0.949	0.570	1.008	0.619	0.824	0.783
$m_{3,3}^{(2)} = -1$	0.867	0.642	1.239	0.959	1.142	0.982
$\sigma_{1,1}^{(2)} = 1$	1.253	0.702	1.198	0.557	1.078	0.361
$\sigma_{1,2}^{(2)} = 0$	-0.085	0.544	0.333	0.671	0.228	0.375
$\sigma_{1,3}^{(2)} = 0$	0.851	0.820	0.773	0.496	0.922	0.616
$\sigma_{2,2}^{(2)} = 2$	2.425	1.821	1.984	1.295	2.145	1.540
$\sigma_{2,3}^{(2)} = 0$	-0.239	0.711	0.276	0.629	0.274	0.525
$\sigma_{3,3}^{(2)} = 3$	2.052	1.171	1.975	0.837	2.141	1.082
$\psi_{1,1}^{(2)} = 1$	1.253	0.702	1.198	0.557	1.078	0.361
$\psi_{1,2}^{(2)} = 0$	-0.085	0.544	0.333	0.671	0.228	0.375
$\psi_{1,3}^{(2)} = 0$	0.851	0.820	0.773	0.496	0.922	0.616
$\psi_{2,2}^{(2)} = 2$	2.425	1.821	1.984	1.295	2.145	1.540
$\psi_{2,3}^{(2)} = 0$	-0.239	0.711	0.276	0.629	0.274	0.525
$\sigma_{3,3}^{(2)} = 3$	2.052	1.171	1.975	0.837	2.141	1.082
$\nu^{(2)} = -1$	-0.410	2.518	-0.215	2.598	-0.417	3.342
$\pi_3 = 0.5$	0.345	0.019	0.375	0.025	0.331	0.019
$m_{1,1}^{(3)} = 1$	-1.370	0.825	-1.445	0.992	-1.415	1.221
$m_{2,2}^{(3)} = 1$	0.773	0.784	0.388	1.031	0.625	1.253
$m_{3,3}^{(3)} = 1$	1.078	0.729	0.740	1.784	0.763	1.604
$\sigma_{1,1}^{(3)} = 2$	1.167	0.524	1.411	0.459	1.420	0.393
$\sigma_{1,2}^{(3)} = 1$	-0.033	0.501	0.191	0.665	0.479	0.500
$\sigma_{1,3}^{(3)} = 3$	0.838	0.595	1.226	1.099	1.059	1.247
$\sigma_{2,2}^{(3)} = 4$	2.142	1.186	2.726	1.454	2.665	1.250
$\sigma_{2,3}^{(3)} = 2$	-0.039	0.897	0.445	0.998	0.565	1.137
$\sigma_{3,3}^{(3)} = 5$	2.281	1.840	2.799	1.962	3.027	1.494
$\psi_{1,1}^{(3)} = 2$	1.167	0.524	1.411	0.459	1.420	0.393
$\psi_{1,2}^{(3)} = 1$	-0.033	0.501	0.191	0.665	0.479	0.500

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 January 2023 doi:10.20944/preprints202301.0524.v1

4.2. Application example

The actual data we used is a classic multivariate data set – the data of the Australian Institute of Sport (AIS). This data set contains 13 physical indicators of 102 men and 100 women exercising. We took variables BMI and Bfat to form a binary sample, where BMI represents the body quality index. Bfat represents the percentage of body fat. It can be seen that this data has an obvious mixed structure. We divided AIS data into two cases of two-component fitting and three-component fitting, and the results are shown in Table 4-Table 5. It can be found from Table 4-5 that the parameter estimates of the first component and the third component of the three-component model are similar in data fitting, and their MSE is slightly larger than the results of the two components. Therefore, the fitting effect of the two-component MMVT model for AIS data is better. The PMLE of MMVT model is -1070.055, which is larger than the MLE of -1097.79 in MMVT.

Table 4. The results of two components MMVT model for AIS dataset

	π_1	$m_{1,1}^{(1)}$	$m_{2,2}^{(1)}$	$\sigma_{1,1}^{(1)}$	$\sigma_{1,2}^{(1)}$	$\sigma_{2,2}^{(1)}$	$\nu^{(1)}$	$p\ell_m$
mle	0.36	21.88	8.37	4.43	1.01	2.30	2.76	-1070.037
pmle	0.38	21.53	5.82	6.01	5.77	7.90	0.24	-1070.055
	π_2	$m_{1,1}^{(2)}$	$m_{2,2}^{(2)}$	$\sigma_{1,1}^{(2)}$	$\sigma_{1,2}^{(2)}$	$\sigma_{2,2}^{(2)}$	$\nu^{(2)}$	
mle	0.64	19.41	18.34	23.13	-0.60	33.58	8.10	
pmle	0.62	19.42	18.40	22.23	0.13	33.17	7.69	

Table 5. The results of three components MMVT model for AIS dataset

	π_1	$m_{1,1}^{(1)}$	$m_{2,2}^{(1)}$	$\sigma_{1,1}^{(1)}$	$\sigma_{1,2}^{(1)}$	$\sigma_{2,2}^{(1)}$	$\nu^{(1)}$	$p\ell_m$
mle	0.20	27.12	17.55	15.84	6.07	24.59	-0.40	-1069.994
pmle	0.15	18.01	13.01	7.31	-3.04	7.28	8.82	-1070.01
	π_2	$m_{1,1}^{(2)}$	$m_{2,2}^{(2)}$	$\sigma_{1,1}^{(2)}$	$\sigma_{1,2}^{(2)}$	$\sigma_{2,2}^{(2)}$	$\nu^{(2)}$	
mle	0.39	24.05	8.54	3.66	1.99	2.56	-0.77	
pmle	0.37	23.92	8.35	3.54	1.90	2.39	0.12	
	π_3	$m_{1,1}^{(3)}$	$m_{2,2}^{(3)}$	$\sigma_{1,1}^{(3)}$	$\sigma_{1,2}^{(3)}$	$\sigma_{2,2}^{(3)}$	$\nu^{(3)}$	
mle	0.41	22.88	19.02	6.38	11.15	34.68	-1.67	
pmle	0.48	20.34	21.54	21.79	-9.78	34.50	7.72	

5. Conclusions

In this paper, penalizing the MMVT row and column covariances is proposed to overcome the parameter degeneracy of the model. Meanwhile, proof of the PMLE consistency is also provided. This approach ensures parameter consistency and estimation validity while overcoming the traditional drawback of having to impose additional restrictions on the parameter space.

There remain several unresolved issues regarding matrix-variable mixed models that must be solved in the future. First, the convergence rate of the EM algorithm is very slow, and especially for matrix variable models, which makes it a priority to reduce the computational complexity of the algorithm. Second, most of the mixed matrix variable distribution models face problems such as likelihood degeneracy, e.g., a matrix variable skewed normal distribution, matrix variable skewed t-distribution, and other skewed ellipsoidal distribution mixtures. The possibility of further generalization of our method to these models is promising.

Author Contributions: "Conceptualization, Xinyu Yang and Weisan Wu; methodology, Xinyu Yang; software, Leigang Dong; formal analysis, Zhichuan Zhu; writing—original draft preparation, Xinyu Yang. All authors have read and agreed to the published version of the manuscript."

Funding: This research was funded by natural science foundation of Jilin Province grant number YDZJ202201ZYTS666.The education science planning foundation of Jilin grant number GH21317. The education department of Jilin province industrialization cultivation project and Privacy Computing Joint Laboratory

Data Availability Statement: Not applicable.

Acknowledgments: The authors would like to thank the referees for their extraordinary comments, which helped to enrich the content of this paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. McLachlan, G.J., Basford, K.E. Mixture Models: Inference and Application to Clustering. . *Dekker* **New York**,*1988*.
2. Chen J.H and Li S.T and Tan X.M. Consistency of the penalized MLE for two-parameter gamma mixture models. *SCI CHINA MATH* **2016**, *12*, 2301–2318.
3. Chen J.H and Tan X.M. Inference for multivariate normal mixtures. *Journal of Multivariate Analysis* **2009**, *7*, 1367–1383.
4. Chen J.H and Tan X.M. and Zhang R.C. Inference for normal mixtures in mean and variance. *Statistica Sinica* **2008**, *2*, 443-465.
5. Azzalini A. and Capitanio A. Statistical Applications of the Multivariate Skew Normal Distribution. *Journal of the Royal Statistical Society B* **1999**, *3*, 579–602.
6. Azzalini A. and Dalla Valle A.. The multivariate skew-normal distribution. *Biometrika* **1996**, *4*, 715–726.
7. Azzalini A. A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics* **1985**, *1*, 171–178.
8. Dempster A. P and Laird N. M and RubinD. B. The title of the cited article. *Journal of the Royal Statistical Society* **1977**, *1*, 1–38.
9. Grigory Alexandrovich. A note on the article Inference for multivariate normal mixtures by J. Chen and X. Tan. *Journal of Multivariate Analysis* **2014**, , 245–248.
10. Hathaway, Richard J. A Constrained Formulation of Maximum-Likelihood Estimation for Normal Mixture Distributions. *The Annals of Statistics* **1985**, *2*, 795–800.
11. Jin L.B and Xu W. and Zhu L.X. Penalized Maximum Likelihood Estimator for Skew Normal Mixtures.2016, *phrase indicating stage of publication (submitted)*.
12. Yurchenko, Y.S.. Some matrix distributions.2021, *phrase indicating stage of publication (submitted)*.
13. Kiefer J. and Wolfowitz J. Consistency on the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Annals of Mathematical Statistics* **1956**, *4*, 887–906.
14. Meng X. L. and Rubin D. B. Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika* **1993**, *2*, 267–278.
15. Tan X.M and Chen J.H and Zhang R.C. Consistency of the constrained maximum likelihood estimator in finite normal mixture models. *American Statistical Association* **2007**, *4*, 2113–2119.
16. Wald A. Note on the consistency of the maximum likelihood estimate. *Annals of Mathematical Statistics* **1949**, *4*, 595–601.
17. Viroli, C.. Finite mixtures of matrix normal distributions for classifying three-way data. *Statistics and Computing* **2011**, *21*, 511–522.
18. Basford, K.E., McLachlan, G.J.. The mixture method of clustering applied to three-way data. *Journal of Classification* **1985**, *2*, 109–125.

19. Dawid, A.P.. Some matrix-variate distribution theory: notational considerations and a Bayesian application. *Biometrika* **1981**, *68*, 265–274.
20. Gupta, A.K., Nagar, D.K.. *Matrix Variate Distributions*.; Publisher: Chapman and Hall/CRC, London/Boca Raton, 2000.
21. Fruhwirth-Schnatter, S. *Finite Mixture and Markov Switching Models*.; Publisher: Springer, 2006.
22. Hunt, L.A., Basford, K.E.. Fitting a Mixture Model to three-mode three-way data with categorical and continuous variables. *Journal of Classification* **1999**, *68*, 283—296.
23. Dutilleul, Pierre. The mle algorithm for the matrix normal distribution. *Journal of Statistical Computation and Simulation* **1999**, *64*, 105–123.
24. Doğru, Fatma Zehra. FINITE MIXTURES OF MATRIX VARIATE T DISTRIBUTIONS. *gazi university journal of science* **2016**, *29*, 335–341.
25. Geoffrey Z. Thompson, Ranjan Maitra, William Q. Meeker and Ashraf F. Bastawros. Classification With the Matrix-Variate-t Distribution. *gJournal of Computational and Graphical Statistics* **2020**, *3*, 668–674.