Adv Data Anal Classif (2019) 13:325–341
https://doi.org/10.1007/s11634-018-0326-1

REGULAR ARTICLE

Studying crime trends in the USA over the years 2000–2012

Volodymyr Melnykov[1] · **Xuwen Zhu**[2]

Received: 29 April 2017 / Revised: 2 March 2018 / Accepted: 21 May 2018 /
Published online: 23 June 2018

Abstract Studying crime trends and tendencies is an important problem that helps to identify socioeconomic patterns and relationships of crucial significance. Finite mixture models are famous for their flexibility in modeling heterogeneity in data. A novel approach designed for accounting for skewness in the distributions of matrix observations is proposed and applied to the United States crime data collected between 2000 and 2012 years. Then, the model is further extended by incorporating explanatory variables. A step-by-step model development demonstrates differences and improvements associated with every stage of the process. Results obtained by the final model are illustrated and thoroughly discussed. Multiple interesting conclusions have been drawn based on the developed model and obtained model-based clustering partition.

Keywords Crime data · Finite mixture model · Matrix normal distribution · Manly transformation · EM algorithm

Mathematics Subject Classification 62P25

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11634-018-0326-1) contains supplementary material, which is available to authorized users.

✉ Xuwen Zhu
xuwen.zhu@louisville.edu

[1] Department of Information Systems, Statistics, and Management Science, University of Alabama, Tuscaloosa, AL 35487, USA

[2] Department of Mathematics, University of Louisville, Louisville, KY 40292, USA

1 Data description

Crime is a complex socioeconomic phenomenon relying on numerous factors and reflecting the well-being of the society. The analysis of crime data can provide necessary understanding and insight into the origin of the problem, allow coordinating preventive actions, and suggest possible solutions. Harries (1976) partitioned 729 American cities based on crime rates and other social indicators. Grubesic (2006) applied fuzzy clustering techniques to locate crime hot spots. Viroli (2011b) investigated partitioning Italian provinces by crime characteristics. Reich and Porter (2015) considered semisupervised clustering relying on spatiotemporal relationships in the series of burglary crimes.

The crime data analyzed in this paper are obtained from the United States Department of Justice, Federal Bureau of Investigation, where they are publicly available at the following Web address: http://www.ucrdatatool.gov/Search/Crime/Crime.cfm. The data are collected for agencies (we refer to them as cities from now on) with the population coverage of at least 100,000. The crime frequency and rate records are available up to the year of 2012. As the focus of this paper is on studying crime trends in the XXI century, we collected the data between 2000 and 2012. There are seven variables included in the data set. The first four represent the rates of violent crimes: *Murder*, *Rape*, *Robbery*, and *Aggravated assault*. The next three variables are the rates of property crimes: *Motor vehicle theft*, *Burglary*, and *Larceny-theft*. For a number of cities, there were incomplete records due to the failure to report crime statistics or other reasons. Such observations were excluded from the data set yielding the total of 236 cities with complete information and each of them taking the form of a 7×13 matrix, i.e., a seven-variate vector observed over 13 time points. The goal of the study is to apply cluster analysis techniques to reveal crime trends and tendencies common for the analyzed cities. Such research can help to address existing problems, identify arising crime patterns, and develop preventive actions.

A particularly flexible cluster analysis method is called model-based clustering (Fraley and Raftery 2002). It is based on the concept of finite mixture models (McLachlan and Peel 2000) and shows remarkable modeling flexibility. While there is literature on matrix-variate distributions including those that can accommodate for skewness (Chen and Gupta 2005; Akdemir and Gupta 2010; Gallaugher and McNicholas 2017), the majority of model-based clustering algorithms focus on partitioning data with observations provided in the form of scalars or vectors. Thus, clustering more complex structures such as matrices or tensors poses immediate challenges. Finite mixtures of matrix normal distributions were recently introduced by Viroli (2011a, b, 2012). Unfortunately, as we show in the paper, crime rates are characteristics with severely skewed distributions. Hence, a more general class of matrix distributions capable of modeling skeweness is proposed. This opens new horizons in the cluster analysis of matrix-valued observations.

The rest of the paper is organized as follows. Section 2 provides the details of the step-by-step model development. Section 3 illustrates and discusses the results obtained. Finally, Sect. 4 gives a brief summary and outlines some possible directions of the future research.

2 Model development

In this section, we outline some basic yet important concepts from finite mixture modeling and model-based clustering that are used in the development of our clustering approach. We proceed in a step-by-step fashion discussing the entire process of model development.

2.1 Preliminaries on mixture modeling and model-based clustering

Let an observed random sample $\boldsymbol{Y}_i \in I\!R^{p\times T}$, $i = 1, \ldots, n$ originate from a mixture model with a probability density function (pdf) given by $g(\boldsymbol{Y}; \boldsymbol{\theta}) = \sum_{k=1}^{K} \tau_k f_k(\boldsymbol{Y}; \boldsymbol{\vartheta}_k)$, where $\boldsymbol{Y}$ is a $p \times T$ real-valued matrix, K is the total number of components in the mixture, $f_k(\cdot; \boldsymbol{\vartheta}_k)$ is the pdf of the kth component with a corresponding parameter vector $\boldsymbol{\vartheta}_k$, and τ_k represents the kth mixing proportion, subject to the constraints $0 < \tau_k \leq 1$ and $\sum_{k=1}^{K} \tau_k = 1$. The traditional method of estimating $\boldsymbol{\theta} = \{\tau_1, \ldots, \tau_K, \boldsymbol{\vartheta}_1, \ldots, \boldsymbol{\vartheta}_K\}$ is the maximum likelihood estimation that can be implemented by means of the expectation-maximization (EM) algorithm (Dempster et al. 1977). At the E-step, the conditional expectation of the complete-data log-likelihood function (also known as the Q-function) needs to be found assuming that the class membership labels $Z_i, i = 1, \ldots, n$, are available. In the traditional mixture modeling framework, this leads to calculating posterior probabilities $\pi_{ik} = E\{I(Z_i = k|\boldsymbol{Y}_i)\}$. At the M-step the conditional expectation from the E-step is maximized with respect to the parameter vector $\boldsymbol{\theta}$. After iterating the E- and M-steps till convergence is reached, the MLE of $\boldsymbol{\theta}$ (i.e., $\hat{\boldsymbol{\theta}}$) as well as that of π_{ik} (i.e., $\hat{\pi}_{ik}$) are obtained. In those cases when the number of components K is not known in advance, it is usually estimated based on one of available information criteria among which Bayesian information criterion (BIC) (Schwarz 1978) is the most popular. The final assignment of observations to classes is based on the Bayes decision rule: $\hat{z}_i = \text{argmax}_k \hat{\pi}_{ik}$. This process is known as model-based clustering and assumes the existence of the one-to-one correspondence between mixture components and data groups.

2.2 Mixture of matrix normal distributions

In the context of the considered data, $\boldsymbol{Y}_i$ represents a matrix observation associated with the ith city. The number of observed crime variables denoted by p is equal to 7. Finally, $T = 13$ is the number of years over which the data have been collected. The choice of the functional form of the component $f_k(\cdot; \boldsymbol{\vartheta}_k)$ is rather limited as there are very few multivariate distributions with matrix arguments. An immediate candidate for this role and a starting point in our model development process is the matrix normal (or Gaussian) distribution with a pdf given by

$$\Phi(\boldsymbol{Y}; \boldsymbol{M}, \boldsymbol{\Sigma}, \boldsymbol{\Psi}) = (2\pi)^{-\frac{pT}{2}} |\boldsymbol{\Sigma}|^{-\frac{T}{2}} |\boldsymbol{\Psi}|^{-\frac{p}{2}} e^{-\frac{1}{2}\text{tr}\{\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y}-\boldsymbol{M})\boldsymbol{\Psi}^{-1}(\boldsymbol{Y}-\boldsymbol{M})^\top\}}, \quad (2.1)$$

where $\boldsymbol{M}$ is the $p \times T$ mean matrix and $\boldsymbol{\Sigma}$ and $\boldsymbol{\Psi}$ are $p \times p$ and $T \times T$ covariance matrices associated with p rows and T columns, respectively. In our context, $\boldsymbol{\Sigma}$ is

responsible for describing variability among crime variables and $\boldsymbol{\Psi}$ is the covariance matrix related to years.

It can be noticed that upon vectorization, the matrix normal distribution can be reformulated as a traditional pT-variate normal distribution with a covariance matrix given by $\boldsymbol{\Psi} \otimes \boldsymbol{\Sigma}$, where $\otimes$ represents the Kronecker product. In many problems, however, observations naturally appear in the matrix form and matrix normal specification is more convenient mathematically as well as from the interpretation standpoint. Clearly, our application represents one of such cases. It is worth mentioning that there is a lack of model identifiability related to the properties of the Kronecker product. $\boldsymbol{\Psi} \otimes \boldsymbol{\Sigma} = \boldsymbol{\Psi}^\star \otimes \boldsymbol{\Sigma}^\star$ if $\boldsymbol{\Sigma}^\star = a\boldsymbol{\Sigma}$ and $\boldsymbol{\Psi}^\star = a^{-1}\boldsymbol{\Psi}$. As a result, matrices $\boldsymbol{\Sigma}$ and $\boldsymbol{\Psi}$ are identifiable up to a positive multiplier a.

The matrix normal (or Gaussian) mixture model, from now on denoted as mGm, has the probability denisty function given by

$$g(\boldsymbol{Y}; \boldsymbol{\theta}) = \sum_{k=1}^{K} \tau_k \Phi(\boldsymbol{Y}; \boldsymbol{M}_k, \boldsymbol{\Sigma}_k, \boldsymbol{\Psi}_k). \tag{2.2}$$

It can be shown that in this setting the EM algorithm reduces to the following iterative expressions:

$$\begin{aligned}
\ddot{\pi}_{ik} &= \frac{\dot{\tau}_k \Phi(\boldsymbol{Y}_i; \dot{\boldsymbol{M}}_k, \dot{\boldsymbol{\Sigma}}_k, \dot{\boldsymbol{\Psi}}_k)}{\sum_{k'=1}^{K} \dot{\tau}_{k'} \Phi(\boldsymbol{Y}_i; \dot{\boldsymbol{M}}_{k'}, \dot{\boldsymbol{\Sigma}}_{k'}, \dot{\boldsymbol{\Psi}}_{k'})}, \quad \ddot{\tau}_k = \frac{\sum_{i=1}^{n} \ddot{\pi}_{ik}}{n}, \\
\ddot{\boldsymbol{M}}_k &= \frac{\sum_{i=1}^{n} \ddot{\pi}_{ik} \boldsymbol{Y}_i}{\sum_{i=1}^{n} \ddot{\pi}_{ik}}, \\
\ddot{\boldsymbol{\Sigma}}_k &= \frac{\sum_{i=1}^{n} \ddot{\pi}_{ik} (\boldsymbol{Y}_i - \ddot{\boldsymbol{M}}_k) \dot{\boldsymbol{\Psi}}_k^{-1} (\boldsymbol{Y}_i - \ddot{\boldsymbol{M}}_k)^\top}{T \sum_{i=1}^{n} \ddot{\pi}_{ik}}, \\
\ddot{\boldsymbol{\Psi}}_k &= \frac{\sum_{i=1}^{n} \ddot{\pi}_{ik} (\boldsymbol{Y}_i - \ddot{\boldsymbol{M}}_k)^\top \ddot{\boldsymbol{\Sigma}}_k^{-1} (\boldsymbol{Y}_i - \ddot{\boldsymbol{M}}_k)}{p \sum_{i=1}^{n} \ddot{\pi}_{ik}}.
\end{aligned} \tag{2.3}$$

Here, parameters marked with one dot correspond to the previous iteration and those marked with two dots represent the estimates at the current iteration.

Despite the convenience and mathematical appeal of this mixture model, there are many applications where the distribution of random variables is far from being normal or even symmetric. Figure 1 illustrates the sampling distributions of burglary and murder rates based on the full 13-year data. It is easy to see that both histograms are severely skewed to the right. In fact, distributions of all seven crime types demonstrate strong right skewness without the presence of obvious outliers. This observation suggests that the mGm model does not provide sufficient modeling power and motivates us to investigate alternative components capable of modeling skewed crime data.

2.3 Mixture of matrix transformation distributions

The problem of modeling skewed data is not new in the mixture modeling framework. Mixtures of skew-normal (Lin 2009; Cabral et al. 2012) and skew-t (Lee and McLach-

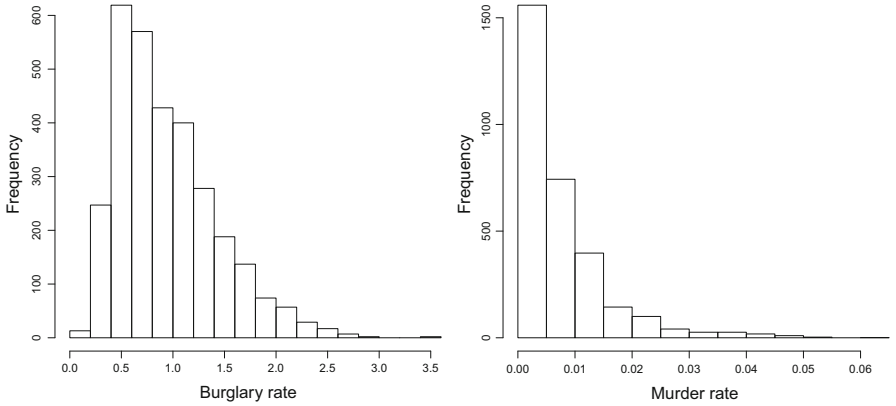

Fig. 1 Sampling distributions of burglary and murder rates showing deviations from normality due to severe skewness

lan 2013, 2014) distributions are the most popular choices in this context. Some other alternatives include shifted asymmetric Laplace (Franczak et al. 2014) and generalized hyperbolic (Browne and McNicholas 2015) components. All the above-mentioned models target vector-valued data. In our context, however, we should be able to model skewness in the distribution of matrix-valued observations.

Suppose there exists a transformation $\mathcal{T}(\cdot;\lambda)$ with a transformation parameter λ such that $\mathcal{T}(y;\lambda)$ is approximately normally distributed. Here, $\mathcal{T}$ represents the transformation operator. Multivariate transformation assumes that for an appropriate choice of the parameter $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p)^\top$, the vector $\boldsymbol{y} = (y_1, \ldots, y_p)^\top$ will be transformed to (approximate) normality, i.e.,

$$\mathcal{T}(\boldsymbol{y};\boldsymbol{\lambda}) = \big(\mathcal{T}(y_1;\lambda_1), \ldots, \mathcal{T}(y_p;\lambda_p)\big)^\top \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{2.4}$$

where $\mathcal{T}(\boldsymbol{y};\boldsymbol{\lambda})$ represents the transformed vector and $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the mean vector and covariance matrix of the transformed data, respectively. Starting from the normal distribution and applying back transformation, one can obtain what we call a $\mathcal{T}$-mixture component that is capable of modeling skeweness and given by $\phi(\mathcal{T}(\boldsymbol{y};\boldsymbol{\lambda}_k);\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k))|d\mathcal{T}(\boldsymbol{y};\boldsymbol{\lambda})/d\boldsymbol{y}^\top|$, where $\phi(\cdot;\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)$ represents the p-variate normal pdf and $|d\mathcal{T}(\boldsymbol{y};\boldsymbol{\lambda})/d\boldsymbol{y}^\top|$ is the Jacobian associated with the transformation. A good resource with ideas on multivariate transformations is the book by Atkinson et al. (2003).

The outlined idea can be generalized and applied to the matrix normal distribution. The skewness vector $\boldsymbol{\lambda}_k$ is applied to the p rows of the data yielding the matrix $\mathcal{T}$-mixture (m$\mathcal{T}$m) model with the density function given by

$$g(\boldsymbol{Y};\boldsymbol{\theta}) = \sum_{k=1}^{K} \tau_k \Phi(\mathcal{T}(\boldsymbol{Y};\boldsymbol{\lambda}_k);\boldsymbol{M}_k,\boldsymbol{\Sigma}_k,\boldsymbol{\Psi}_k)|d\mathcal{T}(\boldsymbol{Y};\boldsymbol{\lambda}_k)/d\boldsymbol{Y}^\top|, \tag{2.5}$$

where $\mathbf{1} = (1, \ldots, 1)^\top$ with cardinality $|\mathbf{1}| = T$. The corresponding Q-function is given by

$$Q(\boldsymbol{\theta};\dot{\boldsymbol{\theta}},\boldsymbol{Y}_1,\ldots,\boldsymbol{Y}_n)=\sum_{i=1}^{n}\sum_{k=1}^{K}\ddot{\pi}_{ik}\left\{-\frac{pT}{2}\log 2\pi-\frac{T}{2}\log|\boldsymbol{\Sigma}_k|-\frac{p}{2}\log|\boldsymbol{\Psi}_k|\right.$$
$$-\frac{1}{2}\mathrm{tr}\left\{\boldsymbol{\Sigma}_k^{-1}(\mathcal{T}(\boldsymbol{Y}_i;\boldsymbol{\lambda}_k)-\boldsymbol{M}_k)\boldsymbol{\Psi}_K^{-1}(\mathcal{T}(\boldsymbol{Y}_i;\boldsymbol{\lambda}_k)-\boldsymbol{M}_k)^\top\right\}$$
$$\left.+\log|d\mathcal{T}(\boldsymbol{Y}_i;\boldsymbol{\lambda}_k)/d\boldsymbol{Y}_i^\top|\right\}. \tag{2.6}$$

Expressions for updating parameters in the course of the EM algorithm are as follows below:

$$\begin{aligned}
\ddot{\pi}_{ik}&=\frac{\dot{\tau}_k\Phi(\mathcal{T}(\boldsymbol{Y}_i;\dot{\boldsymbol{\lambda}}_k);\dot{\boldsymbol{M}}_k,\dot{\boldsymbol{\Sigma}}_k,\dot{\boldsymbol{\Psi}}_k)|d\mathcal{T}(\boldsymbol{Y}_i;\dot{\boldsymbol{\lambda}}_k)/d\boldsymbol{Y}_i^\top|}{\sum_{k'=1}^{K}\dot{\tau}_{k'}\Phi(\mathcal{T}(\boldsymbol{Y}_i;\dot{\boldsymbol{\lambda}}_{k'});\dot{\boldsymbol{M}}_{k'},\dot{\boldsymbol{\Sigma}}_{k'},\dot{\boldsymbol{\Psi}}_{k'})|d\mathcal{T}(\boldsymbol{Y}_i;\dot{\boldsymbol{\lambda}}_{k'})/d\boldsymbol{Y}_i^\top|},\\
\ddot{\tau}_k&=\frac{\sum_{i=1}^{n}\ddot{\pi}_{ik}}{n},\quad \ddot{\boldsymbol{\lambda}}_k=\mathrm{argmax}\,Q_{\boldsymbol{\lambda}}(\boldsymbol{\lambda}_k),\quad \ddot{\boldsymbol{M}}_k=\frac{\sum_{i=1}^{n}\ddot{\pi}_{ik}\mathcal{T}(\boldsymbol{Y}_i;\ddot{\boldsymbol{\lambda}}_k)}{\sum_{i=1}^{n}\ddot{\pi}_{ik}},\\
\ddot{\boldsymbol{\Sigma}}_k&=\frac{\sum_{i=1}^{n}\ddot{\pi}_{ik}(\mathcal{T}(\boldsymbol{Y}_i;\ddot{\boldsymbol{\lambda}}_k)-\ddot{\boldsymbol{M}}_k)\dot{\boldsymbol{\Psi}}_k^{-1}(\mathcal{T}(\boldsymbol{Y}_i;\ddot{\boldsymbol{\lambda}}_k)-\ddot{\boldsymbol{M}}_k)^\top}{T\sum_{i=1}^{n}\ddot{\pi}_{ik}},\\
\ddot{\boldsymbol{\Psi}}_k&=\frac{\sum_{i=1}^{n}\ddot{\pi}_{ik}(\mathcal{T}(\boldsymbol{Y}_i;\ddot{\boldsymbol{\lambda}}_k)-\ddot{\boldsymbol{M}}_k)^\top\ddot{\boldsymbol{\Sigma}}_k^{-1}(\mathcal{T}(\boldsymbol{Y}_i;\ddot{\boldsymbol{\lambda}}_k)-\ddot{\boldsymbol{M}}_k)}{p\sum_{i=1}^{n}\ddot{\pi}_{ik}},
\end{aligned} \tag{2.7}$$

where

$$Q_{\boldsymbol{\lambda}}(\boldsymbol{\lambda}_k)=\sum_{i=1}^{n}\ddot{\pi}_{ik}\left\{-\frac{T}{2}\log|\boldsymbol{\Sigma}_k(\boldsymbol{\lambda}_k)|+\log|d\mathcal{T}(\boldsymbol{Y};\boldsymbol{\lambda}_k)/d\boldsymbol{Y}^\top|\right\} \tag{2.8}$$

is the part of the Q-function given in 2.6 that involves $\boldsymbol{\lambda}_k$ and $\boldsymbol{\Sigma}_k(\boldsymbol{\lambda}_k)$ is the matrix $\boldsymbol{\Sigma}_k$ expressed as a function of $\boldsymbol{\lambda}_k$. It is worth mentioning that skewness can be introduced for columns as well. In this paper, however, a reasonable assumption can be made that there is approximately same skewness over years. A lack of model identifiability regarding the skewness parameters naturally accommodates them into the expression of $\boldsymbol{\lambda}_k$'s.

2.4 Choice of transformation

After the general idea behind $\mathcal{T}$-mixtures is outlined, it is important to focus on some transformations that can be conveniently employed in the outlined setting. The first transformation we consider is the exponential transformation proposed by Manly (1976) and recently studied in the mixture modeling context by Zhu and Melnykov (2018). The transformation is given by $\mathcal{M}(y;\lambda)=I(\lambda\neq 0)(e^{\lambda y}-1)/\lambda+I(\lambda=0)y$ and is more flexible than the famous power transformation proposed by Box and Cox (1964) as it can model left- and right-skewed data and is not restricted to $\mathbb{R}^+$. Under this transformation, $f_Y(y;\boldsymbol{\vartheta})\approx\phi(\mathcal{M}(y;\lambda);\mu,\sigma^2)\exp(\lambda y)$. The equality is not strict as both, Manly and Box–Cox, transformations lead to a distribution with bounded support, while the support for normal distribution should be $\mathbb{R}$. As pointed out by Draper and Cox (1969) as well as Manly (1976), this issue is rather minor due

to the "regularization" that occurs in the course of the transformation. In other words, the parameter λ is chosen in such a way that the transformed variable is as close to normal as possible and the impact of the bound $-1/\lambda$ is minimal. While most of the researchers simply ignore the discussed issue in their likelihood-based inference due to its minor impact (Box and Cox 1964; Krzanowski and Marriott 1994; Atkinson et al. 2003; Lo and Gottardo 2012), we propose a slight modification. Upon the estimation of model parameters, it is easy to find a normalizing constant c such that the exact equality $f_Y(y;\boldsymbol{\vartheta}) = c^{-1}\phi(\mathcal{M}(y;\lambda);\mu,\sigma^2)\exp(\lambda y)$ holds. This correction relaxes the issue with improper pdf and results in slightly higher likelihood values as it can be shown that $c \leq 1$.

Another interesting transformation that we employ in this paper is proposed by Yeo and Johnson (2000). It is an improved version of a power transformation that is equally effective in modeling negative as well as positive data and can accommodate both left and right skewness. The transformation maps $\mathbb{R}$ to $\mathbb{R}$ and is given by

$$\begin{aligned}\mathcal{P}(y;\lambda) \;=\; & I(y \geq 0)\left\{ I(\lambda \neq 0)\frac{(1+y)^{\lambda}-1}{\lambda} + I(\lambda = 0)\log(1+y)\right\} \\ & - I(y < 0)\left\{ I(\lambda \neq 2)\frac{(1-y)^{2-\lambda}-1}{2-\lambda} + I(\lambda = 2)\log(1-y)\right\}.\end{aligned} \tag{2.9}$$

This implies that $f_Y(y;\boldsymbol{\vartheta}) = \phi(\mathcal{P}(y;\lambda);\mu,\sigma^2)(|y|+1)^{sgn(y)(\lambda-1)}$ and no correction is needed.

It is worth mentioning that none of the above-mentioned transformations is generally superior over the other one. However, when both transformations are equally effective in reaching normality, the produced log-likelihood values are equal to each other. As a final remark of this section, we would like to mention that analytical expressions for the moments of $f_Y(y;\boldsymbol{\vartheta})$ are not available in closed forms for all transformations discussed in this paper. Some expressions for univariate approximations can be found in Box and Cox (1964), Manly (1976), and Yeo and Johnson (2000). Zhu and Melnykov (2018) derived an expression for the univariate mode in the Manly transformation context: $m = \lambda^{-1}\log\left\{\frac{1}{2}\left(1+\lambda\mu+\sqrt{(1+\lambda\mu)^2+4\lambda^2\sigma^2}\right)\right\}$. In order to find a multivariate mode, a system of nonlinear equations must be solved. The authors also discussed an approach for estimating multivariate moments based on the ideas of propagation of errors. In our opinion, if the moments or other characteristics of the distribution $f_Y(y;\boldsymbol{\vartheta})$ are of interest, Monte Carlo methods such as importance sampling can present an effective and accurate alternative.

2.5 Covariance matrix parameterization

One serious shortcoming associated with the model given in (2.5) is the number of parameters involved in covariance matrices $\boldsymbol{\Psi}_k$ corresponding to data columns. Indeed, there are $KT(T+1)/2$ (i.e., $91K$ in our problem) parameters associated with $\boldsymbol{\Psi}_1,\ldots,\boldsymbol{\Psi}_K$. Employing one of traditional time series relationships can often fix

this drawback effectively. Mixture modeling of longitudinal data has been extensively studied in the recent literature. McNicholas and Murphy (2010) focused on covariance matrix parameterizations for autoregressive (AR) time series, Anderlucci and Viroli (2015) considered a related setting in the matrix-variate context, Melnykov (2012) employed a conditional likelihood function, and Michael and Melnykov (2016) made use of the famous Kalman filter to estimate parameters in the general autoregressive moving average (ARMA) time series framework. In the context of our problem, first order AR model, i.e., AR(1), can be effectively employed and dramatically reduce the number of parameters. Plots of sample autocorrelation functions that can be found in Supplement clearly suggest the AR(1) structure is reasonable.

Hence, each $\boldsymbol{\Psi}_k$ can be formulated in terms of two parameters σ_k^2 and ρ_k (i.e., $\boldsymbol{\Psi}_k \equiv \boldsymbol{\Psi}(\sigma_k^2, \rho_k)$), where σ_k^2 and ρ_k are the variance and correlation parameter, respectively. $\boldsymbol{\Psi}(\sigma_k^2, \rho_k)$ is given by

$$\boldsymbol{\Psi}(\sigma_k^2, \rho_k) = \frac{\sigma_k^2}{1-\rho_k^2} \begin{pmatrix} 1 & \rho_k & \rho_k^2 & \dots & \rho_k^{T-1} \\ \rho_k & 1 & \rho_k & \dots & \rho_k^{T-2} \\ \rho_k^2 & \rho_k & 1 & \dots & \rho_k^{T-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_k^{T-1} & \rho_k^{T-2} & \rho_k^{T-3} & \dots & 1 \end{pmatrix}. \tag{2.10}$$

Let the above-given correlation matrix be denoted as $\boldsymbol{R}(\rho_k)$, i.e., $\boldsymbol{\Psi}(\sigma_k^2, \rho_k) = \frac{\sigma_k^2}{1-\rho_k^2}\boldsymbol{R}(\rho_k)$. It can be shown that $|\boldsymbol{R}(\rho_k)| = (1-\rho_k^2)^{T-1}$ and $\boldsymbol{R}^{-1}(\rho_k) = \frac{1}{1-\rho_k^2}\boldsymbol{R}^{-}(\rho_k)$, where

$$\boldsymbol{R}^{-}(\rho_k) = \begin{pmatrix} 1 & -\rho_k & 0 & \dots & 0 \\ -\rho_k & 1+\rho_k^2 & -\rho_k & \dots & 0 \\ 0 & -\rho_k & 1+\rho_k^2 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{pmatrix}. \tag{2.11}$$

In our context, we obtain $\log|\boldsymbol{\Psi}(\sigma_k^2, \rho_k)| = \log\left|\frac{\sigma_k^2}{1-\rho_k^2}\boldsymbol{R}(\rho_k)\right| = T\log\sigma_k^2 - \log(1-\rho_k^2)$ and $\boldsymbol{\Psi}^{-1}(\sigma_k^2, \rho_k) = \frac{1-\rho_k^2}{\sigma_k^2}\boldsymbol{R}^{-1}(\rho_k) = \frac{1}{\sigma_k^2}\boldsymbol{R}^{-}(\rho_k) = \frac{1}{\sigma_k^2}\left(\rho_k^2\boldsymbol{A}_2 - \rho_k\boldsymbol{A}_1 + \boldsymbol{I}\right)$, where

$$\boldsymbol{A}_1 = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 & 0 \\ 1 & 0 & 1 & \dots & 0 & 0 \\ 0 & 1 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 0 & 1 \\ 0 & 0 & 0 & \dots & 1 & 0 \end{pmatrix} \quad \text{and} \quad \boldsymbol{A}_2 = \begin{pmatrix} 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 1 & 0 & \dots & 0 & 0 \\ 0 & 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 & 0 \\ 0 & 0 & 0 & \dots & 0 & 0 \end{pmatrix}. \tag{2.12}$$

For notational simplicity, let $c_{ik}(\boldsymbol{A})$ be defined as follows:

$$c_{ik}(\boldsymbol{A}) = \text{tr}\left\{\boldsymbol{\Sigma}_k^{-1}(\mathcal{T}(\boldsymbol{Y}_i;\boldsymbol{\lambda}_k) - \boldsymbol{M}_k)\boldsymbol{A}(\mathcal{T}(\boldsymbol{Y}_i;\boldsymbol{\lambda}_k) - \boldsymbol{M}_k)^\top\right\}, \tag{2.13}$$

where $\boldsymbol{A}$ is a $T \times T$ matrix. Also, define $C_k(\boldsymbol{A}) = \sum_{i=1}^n \pi_{ik} c_{ik}(\boldsymbol{A})$. Then, the EM algorithm can be written as the combination of the following expressions:

$$\begin{aligned}
\ddot{\pi}_{ik} &= \frac{\dot{\tau}_k \Phi(\mathcal{T}(\boldsymbol{Y}_i;\dot{\boldsymbol{\lambda}}_k);\dot{\boldsymbol{M}}_k,\dot{\boldsymbol{\Sigma}}_k,\boldsymbol{\Psi}(\dot{\sigma}_k^2,\dot{\rho}_k))|d\mathcal{T}(\boldsymbol{Y}_i;\dot{\boldsymbol{\lambda}}_k)/d\boldsymbol{Y}_i^\top|}{\sum_{k'=1}^K \dot{\tau}_{k'} \Phi(\mathcal{T}(\boldsymbol{Y}_i;\dot{\boldsymbol{\lambda}}_{k'});\dot{\boldsymbol{M}}_{k'},\dot{\boldsymbol{\Sigma}}_{k'},\boldsymbol{\Psi}(\dot{\sigma}_{k'}^2,\dot{\rho}_{k'}))|d\mathcal{T}(\boldsymbol{Y}_i;\dot{\boldsymbol{\lambda}}_{k'})/d\boldsymbol{Y}_i^\top|}, \\
\ddot{\tau}_k &= \frac{\sum_{i=1}^n \ddot{\pi}_{ik}}{n}, \quad \ddot{\boldsymbol{\lambda}}_k = \text{argmax} Q_{\boldsymbol{\lambda}}(\boldsymbol{\lambda}_k), \quad \ddot{\boldsymbol{M}}_k = \frac{\sum_{i=1}^n \ddot{\pi}_{ik}\mathcal{T}(\boldsymbol{Y}_i;\ddot{\boldsymbol{\lambda}}_k)}{\sum_{i=1}^n \ddot{\pi}_{ik}}, \\
\ddot{\boldsymbol{\Sigma}}_k &= \frac{\sum_{i=1}^n \ddot{\pi}_{ik}(\mathcal{T}(\boldsymbol{Y}_i;\ddot{\boldsymbol{\lambda}}_k) - \ddot{\boldsymbol{M}}_k)(\dot{\rho}_k^2\boldsymbol{A}_2 - \dot{\rho}_k\boldsymbol{A}_1 + \boldsymbol{I})(\mathcal{T}(\boldsymbol{Y}_i;\ddot{\boldsymbol{\lambda}}_k) - \ddot{\boldsymbol{M}}_k)^\top}{T\dot{\sigma}_K^2 \sum_{i=1}^n \ddot{\pi}_{ik}},
\end{aligned} \tag{2.14}$$

where $Q_{\boldsymbol{\lambda}}(\boldsymbol{\lambda}_k) = \sum_{i=1}^n \ddot{\pi}_{ik}\left\{-\frac{T}{2}\log|\boldsymbol{\Sigma}_k(\boldsymbol{\lambda}_k)| + \log|d\mathcal{T}(\boldsymbol{Y}_i;\boldsymbol{\lambda}_k)/d\boldsymbol{Y}_i^\top|\right\}$ is the part of the Q-function that depends on $\boldsymbol{\lambda}_k$. Estimates $\ddot{\sigma}_k^2$ and $\ddot{\rho}_k$ can be obtained by taking the derivatives of the function

$$\begin{aligned}
Q_{\sigma^2,\rho}(\sigma_k^2,\rho_k) = \sum_{i=1}^n \ddot{\pi}_{ik}\Bigg(&-\frac{pT}{2}\log\sigma_k^2 + \frac{p}{2}\log(1-\rho_k^2) \\
&-\frac{1}{2\sigma_k^2}\left(\ddot{c}_{ik}(\boldsymbol{A}_2)\rho_k^2 - \ddot{c}_{ik}(\boldsymbol{A}_1)\rho_k + \ddot{c}_{ik}(\boldsymbol{I})\right)\Bigg)
\end{aligned} \tag{2.15}$$

with respect to σ_k^2 and ρ_k, setting them to zero, and solving the obtained system of two equations. It can be shown that the parameter estimate $\ddot{\rho}_k$ can be found as a root of the following cubic equation (subject to $|\rho_k| < 1$ for stationarity as there are three roots):

$$2(T-1)\ddot{C}_k(\boldsymbol{A}_2)\rho_k^3 - (T-2)\ddot{C}_k(\boldsymbol{A}_1)\rho_k^2 - 2(\ddot{C}_k(\boldsymbol{I}) + T\ddot{C}_k(\boldsymbol{A}_2))\rho_k + T\ddot{C}_k(\boldsymbol{A}_1) = 0. \tag{2.16}$$

After that, the parameter estimate $\ddot{\sigma}_k^2$ can be found as

$$\ddot{\sigma}_k^2 = \frac{\ddot{C}_k(\boldsymbol{A}_2)\ddot{\rho}_k^2 - \ddot{C}_k(\boldsymbol{A}_1)\ddot{\rho}_k + \ddot{C}_k(\boldsymbol{I})}{pT\sum_{i=1}^n \ddot{\pi}_{ik}} \tag{2.17}$$

and this completes the steps of the EM algorithm.

2.6 Mixture of matrix transformation regression time series

It can be noticed that the number of parameters associated with each component mean matrix $\boldsymbol{M}_k$, $pT = 91$, is also rather high. A considerable reduction in the number of parameters can be achieved if crime rates are regressed on years. In this setting, the

$T \times (q+1)$ matrix $\boldsymbol{X}$ contains the information about the intercept and q explanatory variables and $\boldsymbol{\beta}_k$ is the $(q+1) \times p$ matrix of corresponding coefficients. Then, the model in (2.5) can be rewritten in the following way:

$$g(\boldsymbol{Y};\boldsymbol{\theta}) = \sum_{k=1}^{K} \tau_k \Phi(\mathcal{T}(\boldsymbol{Y};\boldsymbol{\lambda}_k); (\boldsymbol{X}\boldsymbol{\beta}_k)^\top, \boldsymbol{\Sigma}_k, \boldsymbol{\Psi}_k)|d\mathcal{T}(\boldsymbol{Y};\boldsymbol{\lambda}_k)/d\boldsymbol{Y}^\top|. \tag{2.18}$$

The calculation of model parameters $\tau_k, \boldsymbol{\lambda}_k, \boldsymbol{\Sigma}_k$, and $\boldsymbol{\Psi}_k$ requires rather minor and intuitive updates to the EM algorithm. The matrix of coefficients $\boldsymbol{\beta}_k$ should be estimated instead of the mean matrix $\boldsymbol{M}_k$. It can be accomplished by the expression

$$\ddot{\boldsymbol{\beta}}_k = \left(\sum_{i=1}^{n} \ddot{\pi}_{ik} \boldsymbol{X}_i^\top \dot{\boldsymbol{\Psi}}_k^{-1} \boldsymbol{X}_i\right)^{-1} \sum_{i=1}^{n} \ddot{\pi}_{ik} \boldsymbol{X}_i^\top \dot{\boldsymbol{\Psi}}_k^{-1} \mathcal{T}(\boldsymbol{Y}_i;\ddot{\boldsymbol{\lambda}}_k)^\top. \tag{2.19}$$

It can be noticed that the reduction in the number of parameters due to this model change is equal to $(T-q-1)pK$, which can be rather essential for high values of T.

Table 1 contains the results obtained for matrix Gaussian (mGm), Manly (mMm), and power (mPm) regression mixture models. As there are various polynomial regressions considered, the polynomial function degree ($q = 0, \ldots, 6$) is specified in the subscript of each model's name. The superscript $*$ indicates that the AR(1) structure has been applied. Corresponding results that do not assume AR(1) are generally worse and can be found in Supplement. To identify the best number of components, $K = 1, \ldots, 7$ have been studied for each model. To alleviate the risk of poor initialization, the EM algorithm was rerun 500 times from random starting points for each model and the best result in terms of BIC was recorded. The solution with the smallest BIC is provided in the bold font.

For higher numbers of q and K, we encountered a problem related to finding spurious solutions (McLachlan and Peel 2000). Such solutions reflect a random pattern in data rather than a systematic characteristic. They can be usually detected by the presence of data groups with very few observations or low eigenvalues compared to the other ones. As a result, overly high maximized log-likelihood values can be observed. Various approaches of alleviating the problem are available in the literature. We effectively fought with the issue by excluding from the consideration all solutions that involved clusters consisting of less than 5 points. The detailed analysis of obtained results is provided in the next section.

3 Analysis of results

As we can notice from Table 1, the majority of mGm models prefer solutions with 7 components, with the best one achieved at mGm_3^*, i.e., $\text{mGm}_3^*(7)$. The corresponding BIC value is $-76{,}627.3$. It can be noticed that BIC values associated with both mMm and mPm models are dramatically better. In particular, mMm models prefer three-, four-, or five-component solutions, with the best one reported for $\text{mMm}_5^*(4)$ (BIC is equal to $-80{,}404.4$). A close competitor that is just marginally worse is $\text{mMm}_5^*(4)$ with

Table 1 BIC values obtained for different numbers of components K based on matrix Gaussian (mGm), Manly (mMm), and power (mPm) regression mixtures with AR(1) sructure

Model	$K=1$	$K=2$	$K=3$	$K=4$	$K=5$	$K=6$	$K=7$
mGm_0^*	− 69932.6	− 74087.2	− 75160.5	− 75751.5	− 76049.2	− 76281.9	− **76477.8**
mGm_1^*	− 70169.2	− 74337.5	− 75404.4	− 75994.7	− 76279.1	− 76496.4	− **76682.7**
mGm_2^*	− 70198.9	− 74350.5	− 75427.3	− 75983.1	− 76245.7	− 76421.8	− **76566.1**
mGm_3^*	− 70393.6	− 74534.6	− 75595.7	− 76137.5	− 76358.2	− 76498.6	− **76627.3**
mGm_4^*	− 70463.4	− 74591.3	− 75634.7	− 76148.1	− 76345.2	− 76457.9	− **76583.6**
mGm_5^*	− 70552.8	− 74674.0	− 75695.9	− 76183.5	− 76343.2	− **76446.4**	− 76417.1
mGm_6^*	− 70584.9	− 74680.2	− 75682.1	− 76132.1	− 76264.1	− **76347.5**	− 76305.5
mMm_0^*	− 77733.2	− 79188.6	− 79742.3	− 79840.4	− **79905.4**	− 79855.3	− 79780.8
mMm_1^*	− 78019.1	− 79488.6	− 80041.7	− 80127.2	− **80181.6**	− 80120.7	− 80041.7
mMm_2^*	− 78107.2	− 79554.4	− 80080.3	− 80134.2	− **80157.7**	− 80073.3	− 79932.6
mMm_3^*	− 78317.9	− 79744.0	− 80248.4	− **80290.4**	− 80277.4	− 80148.4	− 79979.7
mMm_4^*	− 78421.2	− 79837.4	− 80313.0	− **80339.8**	− 80293.8	− 80132.5	− 79961.1
mMm_5^*	− 78553.0	− 79940.4	− 80402.0	− **80404.4**	− 80319.8	− 80144.5	− 79946.9
mMm_6^*	− 78578.7	− 79935.2	− **80381.4**	− 80353.2	− 80247.8	− 80045.5	− 79768.7
mPm_0^*	− 78150.2	− 79302.9	− 79857.7	− 79898.6	− **79922.2**	− 79869.0	− 79786.8
mPm_1^*	− 78441.1	− 79601.0	− 80156.3	− 80200.7	− **80217.3**	− 80151.0	− 80053.0
mPm_2^*	− 78527.0	− 79671.0	− 80193.2	− **80208.0**	− 80193.2	− 80086.6	− 79953.0
mPm_3^*	− 78734.8	− 79852.7	− **80365.1**	− 80340.3	− 80298.3	− 80171.2	− 80013.1
mPm_4^*	− 78839.3	− 79951.4	− **80429.6**	− 80398.1	− 80335.5	− 80170.3	− 80012.2
mPm_5^*	− 78969.8	− 80059.8	− **80520.9**	− 80466.5	− 80352.9	− 80167.6	− 79940.1
mPm_6^*	− 78994.2	− 80063.1	− **80498.8**	− 80411.8	− 80286.0	− 80068.2	− 79801.2

The subscript in model names specifies the polynomial function degree. Best models are given in the bold font

BIC − 80,402.0. Among mPm models, $\text{mMm}_5^*(3)$ is the best one with corresponding BIC − 80,520.9. The second and third performers are $\text{mMm}_6^*(3)$ and $\text{mMm}_5^*(4)$ with BICs − 80,498.8 and − 80,466.5, respectively. Thus, there is nearly a 4000 difference in BIC values associated with the best mGm and two transformation models.

It is worth mentioning that although the power transformation outperforms Manly transformation in the majority of cases, the difference is rather minor. For example, the relative difference between the best mPm and mMm models is 0.0014. For mMm and mPm models, we can notice that the mixture order decays from $K=5$ to $K=3$ along with the increase in the order of the polynomial regression function. Such an effect can be explained by the overparameterization issue (see, e.g., Melnykov 2016 for related discussions). An interesting observation can be also made with regard to the polynomial function order associated with the top solutions for transformation-based models. In all cases, $q=5$ is prefered. One might think that this order is rather high given just 13 time points. However, we have to recall that the parameter estimation is additionally supported by multiple cities and thus overparameterization is not an issue here.

Table 2 The agreement tables for the partitions discussed in Sect. 3

		$\text{mPm}_5^*(4)$			
		1	2	3	4
$\text{mPm}_5^*(3)$	1	**20**	**59**	0	0
	2	4	1	**84**	0
	3	1	0	0	**67**

		$\text{mcMm}_5^*(3)$			
		1	2	3	
1		**74**	1	4	
2		1	**86**	2	
3		0	0	**68**	

		$\text{mcMm}_5^*(4)$			
		1	2	3	4
1		**18**	**59**	2	0
2		10	0	**79**	0
3		1	0	1	**66**

		$\text{mPm}_5^*(4)$			
		1	2	3	4
1		**20**	4	1	
2		**54**	2	4	
3		1	**81**	2	
4		0	0	**67**	

		$\text{mcMm}_5^*(4)$			
		1	2	3	4
1		**21**	3	1	0
2		2	**56**	2	0
3		6	0	**78**	0
4		0	0	1	**66**

		$\text{mcMm}_5^*(4)$			
		1	2	3	4
$\text{mcMm}_5^*(3)$	1	**19**	**55**	1	0
	2	9	0	**78**	0
	3	1	4	3	**66**

The bold values represents consistent classifications between two partitions

Table 2 provides classification agreement results between the partition produced by the best performing model $\text{mPm}_5^*(3)$ and those yielded by the other models discussed above. First, we can notice that the agreement between $\text{mPm}_5^*(3)$ and $\text{mMm}_5^*(3)$ is rather high. There are just 8 disagreements out of 236. This corresponds to 96.6% match between partitions. The difference between $\text{mPm}_5^*(3)$ and $\text{mPm}_5^*(4)$ is primarily due to the first cluster being split into two. There are only 6 other disagreements that are not explained by this split. $\text{mPm}_5^*(4)$ and $\text{mMm}_5^*(4)$ solutions are also quite similar with 93.6% match in class assignments.

Figure 2 shows polynomial functions regressing crime rates on time that are obtained for two best-performing models. The first block represents the solution obtained by $\text{mPm}_5^*(3)$. The second block demonstrates the mean profiles of $\text{mPm}_5^*(4)$ model. Blue solid, red dashed, orange dotted, and magenta dotted-and-dashed curves illustrate polynomial functions associated with different components. For $\text{mPm}_5^*(3)$, the blue curve represents the cluster of safest cities. The red curve corresponds to the group of cities with the highest crime rates for the majority of variables (*Murder*, *Rape*, *Robbery*, *Aggravated assault*, and *Motor vehicle theft*). Only for the variable *Larceny-theft*, the orange curve reflects higher rates than the red one. The second block in Fig. 2 provides curves associated with the four components of the $\text{mPm}_5^*(4)$ model. It can be noticed that the cluster of safest cities remains nearly unchanged compared with the cluster of safest cities in the $\text{mPm}_5^*(3)$ solution. Based on the above-made observation, we can conclude that this cluster is relatively distinct and easy to detect. The red dashed curve represents the cities with consistently highest rates for all seven crime variables. Orange and magenta curves represent two groups of cities with intermediate rates. The magenta curve corresponds to higher rates for *Murder*, *Robbery*, *Aggravated assault*, and *Motor vehicle theft*. For the variables *Rape*, *Burglary*, and *Larceny-theft*, the orange curve is associated with higher rates. Thus, the magenta cluster mostly

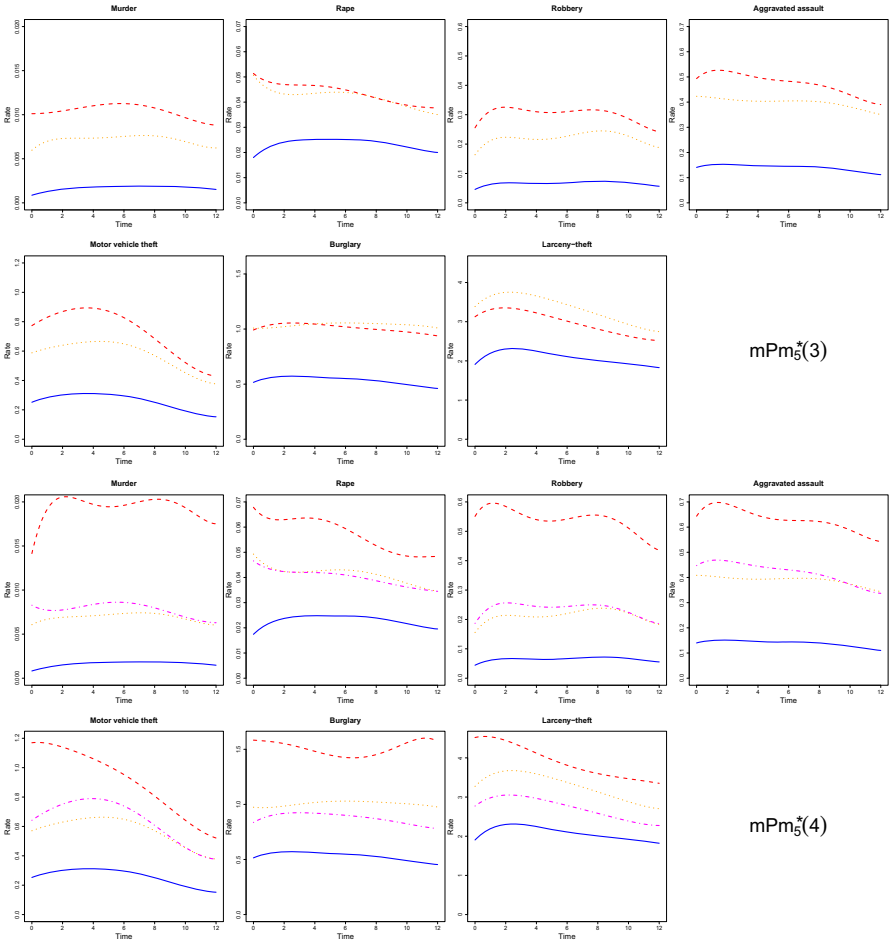

Fig. 2 Mean profiles for models mPm$_5^*$(3) (top) and mPm$_5^*$(4) (bottom). Colors and line types represent different components (color figure online)

shows higher rates for violent crimes and orange cluster has generally higher rates for property crimes. For the variables *Murder*, *Robbery*, *Aggravated assault*, and *Motor vehicle theft*, the difference between orange and magenta trajectories was more vivid before 2009 (magenta was always higher) and nearly disappeared in 2009–2012. This interesting observation suggests that the orange and magenta groups are primarily separated due to the differences in the crime rates of the above-mentioned variables before 2009 as well as discrepancies in variables *Burglary* and *Larceny-theft* (orange is always higher).

Another interesting observation can be made with regard to the periodic behavior of crime rates in the red component. This type of behavior is observed for the variables *Murder*, *Rape*, *Robbery*, *Aggravated assault*, and *Burglary*. This explains the rather high degree of polynomial functions prefered by mMm and mPm models. We can notice that around 2002 and 2009 there were spikes in crime rates of *Murder*, *Robbery*,

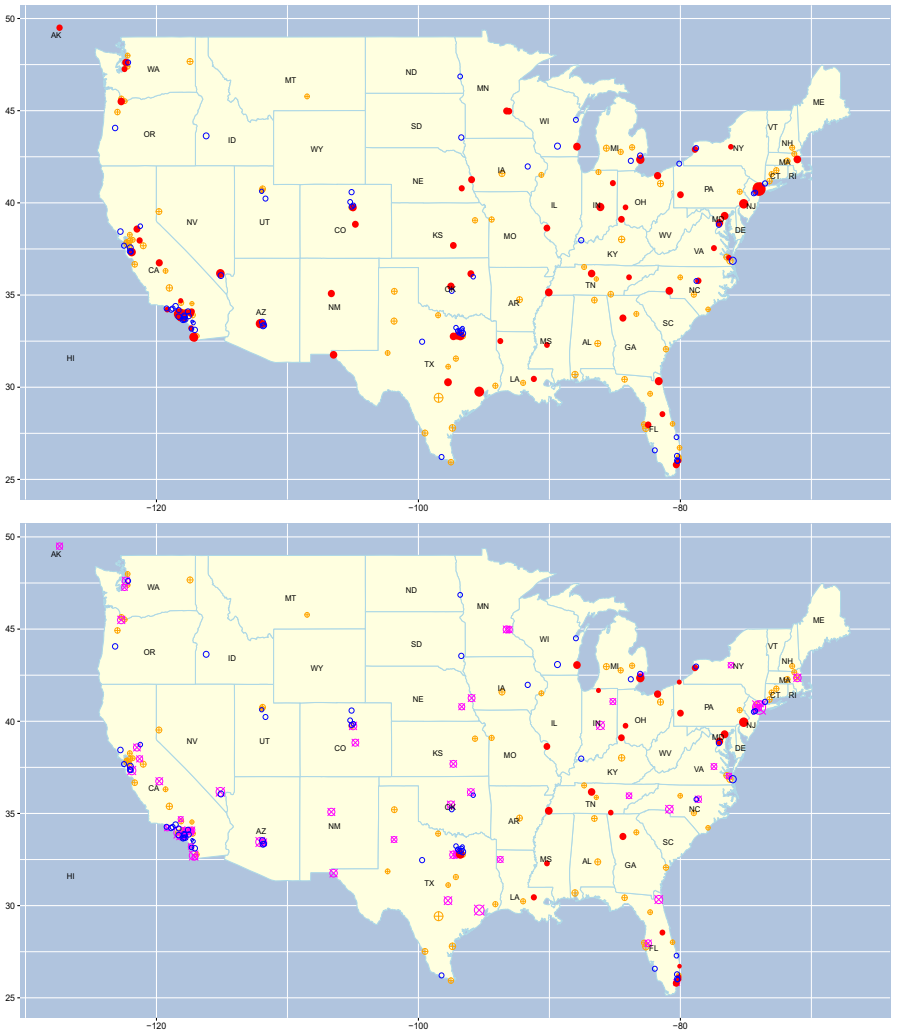

Fig. 3 Two maps representing partitions obtained by $\mathrm{mPm}_5^*(3)$ (top) and $\mathrm{mPm}_5^*(4)$ (bottom) models. The sizes of circles reflect city populations. Colors and symbols illustrate different clusters (color figure online)

and *Aggravated assault*. At the same time, the variable *Rape* showed its lowest values around these times. Some variables, especially *Motor vehicle theft* and *Larceny-theft*, are on decay in crime rates for all groups. At the same time, it can be noticed that the crime rates associated with the blue cluster representing the safest cities remain approximately at the same level for the majority of variables.

Two maps provided in Fig. 3 illustrate clustering solutions obtained by the two best models, i.e., $\mathrm{mPm}_5^*(3)$ (top) and $\mathrm{mPm}_5^*(4)$ (bottom), respectively. The plots for the models $\mathrm{mMm}_5^*(3)$ and $\mathrm{mMm}_5^*(4)$ are not included due to their similarity. Each point represents a specific city. The size of a point reflects the population size, i.e., the larger the city, the bigger the corresponding point. Different colors and symbols reflect

group assignments and match the colors of the polynomial functions constructed in Fig. 2. In the discussion to follow, we focus on the four-cluster solution [obtained from $\text{mPm}_5^*(4)$] presented in the bottom map and use the map in the top for comparison mainly. The entire partition corresponding to the model $\text{mPm}_5^*(4)$ is provided in Supplement.

Several interesting observations can be made based on the maps in Fig. 3. As we can see, the primary difference between the two pictures is that the large red cluster in the top one is split into red and magenta groups in the bottom plot. The red cluster after the separation is smaller than the magenta one. The size of circles in the red cluster (and partition table containing city names that is provided in Supplement) indicates that it includes many large cities. The magenta-colored cities are distributed almost uniformly over the map while the red cluster consists of the cities that are almost entirely located in the eastern part of the country. There is a distinct belt of red circles located along the Mississippi river that includes such cities as St. Louis, Memphis, Jackson, and Baton Rouge. The only red-colored city that falls to the west from this belt is Dallas in Texas. As the red cluster represents the most problematic areas, it can be concluded that the east of the country is considerably more affected by all investigated crime variables.

The blue cluster represents the safest cities in the country. They are all relatively small. We can notice several zones where we observe the majority of blue circles. First, there is a high number of such cities located in northern Texas close to Dallas. Second, there are many blue-colored circles in the Los Angeles—San Diego area. As we can clearly see from the map, the other large Californian metropolitan area around San Francisco is not as safe. Next, there is a high proportion of safe cities in the northern part of the country as there are many small cities that are declared safe according to our clustering solution. On the contrary, there is nearly complete absence of blue cluster representatives in the south-east corner of the country, with just few exceptions around the coast of Florida. Interestingly, we can notice the somewhat unusual location of the middle-sized blue point in the south of Indiana. It represents Evansville, the largest city in the cluster of safest cities. The analysis of the data associated with this city reveals that the assignment to the blue cluster is primarily driven by relatively low rates in *Motor vehicle theft*. The other variables do not support the observed classification so clearly. In fact, the variable *Rape* is rather high and atypical for the other cities in the blue cluster. Another interesting remark can be made regarding the presence of small blue points around large cities such as Dallas, Denver, Detroit, Los Angeles, Phoenix, Salt Lake City, San Diego, San Francisco, etc. This observation suggests that in some metropolitan areas there are well-populated suburbs that are considerably safer than the city they surround. On the other hand, this is not the case for metropolitan areas along what we called the Mississippi belt. Perhaps, the development of these areas used to be driven by the proximity to the Mississippi river. On the contrary, this factor does not contribute much to the area population growth nowadays.

4 Discussion

In this paper, a novel approach to clustering crime data was proposed. The developed procedure is based on the so-called matrix transformation mixture model and illus-

trated on exponential and power transformations. The importance of this model can be seen from its capability to model skewness in data with matrix-valued observations. The step-by-step model development was provided in the context of our problem. The obtained results show great interpretability. Numerous interesting conclusions are drawn based on the partition analysis.

Several ways to further develop the proposed approach can be investigated in the future work. We have discussed how the issue of having an excessive number of parameters associated with matrices $\boldsymbol{\Psi}_k$ and $\boldsymbol{M}_k$ can be addressed in the considered framework. Further model developments can be readily proposed. Specifically, constrained parameterizations of covariance matrices $\boldsymbol{\Sigma}_k$ (Banfield and Raftery 1993) can be employed. In our problem, the number of crime variables $p = 7$ is not high. However, in many problems with even moderate values of p, considering parsimonious models can lead to substantial improvements in BIC.

Acknowledgements The research is partially funded by the University of Louisville EVPRI internal research grant from the Office of the Executive Vice President for Research and Innovation.

References

Akdemir D, Gupta A (2010) A matrix variate skew distribution. Eur J Pure Appl Math 3:128–140
Anderlucci L, Viroli C (2015) Covariance pattern mixture models for the analysis of multivariate heterogeneous longitudinal data. Ann Appl Stat 9:777–800
Atkinson AC, Riani M, Cerioli A (2003) Exploring multivariate data with the forward search. Clarendon Press, Oxford
Banfield JD, Raftery AE (1993) Model-based Gaussian and non-Gaussian clustering. Biometrics 49(3):803–821
Box GE, Cox DR (1964) An analysis of transformations. J R Stat Soc Ser B 26(2):211–252
Browne RP, McNicholas PD (2015) A mixture of generalized hyperbolic distributions. Can J Stat 43(2):176–198
Cabral C, Lachos V, Prates M (2012) Multivariate mixture modeling using skew-normal independent distributions. Comput Stat Data Anal 56(1):126–142
Chen J, Gupta A (2005) Matrix variate skew normal distribution. Statistics 39:247–253
Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood for incomplete data via the EM algorithm (with discussion). J R Stat Soc Ser B 39(1):1–38
Draper NR, Cox DR (1969) On distributions and their transformations to normality. J R Stat Soc B 31:472–476
Fraley C, Raftery AE (2002) Model-based clustering, discriminant analysis, and density estimation. J Am Stat Assoc 97:611–631
Franczak BC, Browne RP, McNicholas PD (2014) Mixtures of shifted asymmetric Laplace distributions. IEEE Trans Pattern Anal Mach Intell 36(6):1149–1157
Gallaugher M, McNicholas P (2017) A matrix variate skew-t distribution. Stat 6:160–170
Grubesic T (2006) On the application of fuzzy clustering for crime hot spot detection. J Quant Criminol 22:77–105
Harries K (1976) A crime based analysis of 729 American cities. Soc Indic Res 2:467–487
Krzanowski WJ, Marriott FHC (1994) Multivariate analysis, part I: distributions, ordination and inference. Halstead Press/Edward Arnold, New York/London
Lee S, McLachlan GJ (2013) On mixtures of skew normal and skew t-distributions. Adv Data Anal Classif 7(3):241–266
Lee S, McLachlan GJ (2014) Finite mixtures of multivariate skew t-distributions: some recent and new results. Stat Comput 24(2):181–202
Lin TI (2009) Maximum likelihood estimation for multivariate skew normal mixture models. J Multivar Anal 100(2):257–265

Lo K, Gottardo R (2012) Flexible mixture modeling via the multivariate t distribution with the Box–Cox transformation: an alternative to the skew-t distribution. Stat Comput 22(1):35–52
Manly BFJ (1976) Exponential data transformations. J R Stat Soc Ser D 25(1):37–42
McLachlan GJ, Peel D (2000) Finite mixture models. Wiley, New York
McNicholas P, Murphy T (2010) Model-based clustering of longitudinal data. Can J Stat 38:153–168
Melnykov V (2012) Efficient estimation in model-based clustering of Gaussian regression time series. Stat Anal Data Min 5:95–99
Melnykov V (2016) Model-based biclustering of clickstream data. Comput Stat Data Anal 93C:31–45
Michael S, Melnykov V (2016) Finite mixture modeling of Gaussian regression time series with application to dendrochronology. J Classif 33:412–441
Reich B, Porter M (2015) Partially-supervised spatiotemporal clustering for burglary crime series identification. J R Stat Soc A 178:465–480
Schwarz G (1978) Estimating the dimensions of a model. Ann Stat 6(2):461–464
Viroli C (2011a) Finite mixtures of matrix normal distributions for classifying three-way data. Stat Comput 21:511–522
Viroli C (2011b) Model based clustering for three-way data structures. Bayesian Anal 6:573–602
Viroli C (2012) On matrix-variate regression analysis. J Multivar Anal 111:296–309
Yeo I-K, Johnson RA (2000) A new family of power transformations to improve normality or symmetry. Biometrika 87:954–959
Zhu X, Melnykov V (2018) Manly transformation in finite mixture modeling. Comput Stat Data Anal 121:190–208