Pattern Recognition 80 (2018) 83–93

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Finite mixtures of skewed matrix variate distributions☆

Michael P.B. Gallaugher, Paul D. McNicholas*

Department of Mathematics and Statistics, McMaster University, Ontario, Canada

ARTICLE INFO

Article history:
Received 9 July 2017
Revised 29 January 2018
Accepted 25 February 2018
Available online 27 February 2018

Keywords:
Clustering
Matrix variate
Mixture models
Skewed distributions

ABSTRACT

Clustering is the process of finding underlying group structures in data. Although mixture model-based clustering is firmly established in the multivariate case, there is a relative paucity of work on matrix variate distributions and none for clustering with mixtures of skewed matrix variate distributions. Four finite mixtures of skewed matrix variate distributions are considered. Parameter estimation is carried out using an expectation-conditional maximization algorithm, and both simulated and real data are used for illustration.

1. Introduction

Over the years, there has been increased interest in the applications involving three-way (matrix variate) data. Although there are countless examples of clustering for multivariate distributions using finite mixture models, as discussed in Section 2, there is very little work for matrix variate distributions. Moreover, the examples in the literature deal exclusively with symmetric (non-skewed) matrix variate distributions such as the matrix variate normal and the matrix variate *t* distributions.

There are many different areas of application for matrix variate distributions. One area is multivariate longitudinal data, where multiple variables are measured over time [e.g., [2]]. In this case, each row of a matrix would correspond to a time point and the columns would represent each of the variables. Furthermore, the two scale matrices, a defining characteristic of matrix variate distributions, allow for simultaneous modelling of the inter-variable covariances as well as the temporal covariances. A second application, considered herein, is image recognition. In this case, an image is analyzed as an $n \times p$ pixel intensity matrix. Herein, a finite mixture of four different skewed matrix distributions, the matrix variate skew-*t*, generalized hyperbolic, variance-gamma and normal inverse Gaussian (NIG) distributions are considered. These mixture models are illustrated for both clustering (unsupervised classification) and semi-supervised classification using both simulated and real data.

2. Background

2.1. Model-based clustering and mixture models

Clustering and classification look at finding and analyzing underlying group structures in data. One common method used for clustering is model-based, and generally makes use of a G-component finite mixture model. A multivariate random variable $\mathbf{X}$ from a finite mixture model has density

$$f(\mathbf{x} \mid \boldsymbol{\vartheta}) = \sum_{g=1}^{G} \pi_g f_g(\mathbf{x} \mid \boldsymbol{\theta}_g),$$

where $\boldsymbol{\vartheta} = \left(\pi_1, \pi_2, \ldots, \pi_G, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_G\right)$, $f_g(\cdot)$ is the gth component density, and $\pi_g > 0$ is the gth mixing proportion such that $\sum_{i=1}^{G} \pi_g = 1$. McNicholas [37] traces the association between clustering and mixture models back to Tiedeman [55], and the earliest use of a finite mixture model for clustering can be found in Wolfe [61], who uses a Gaussian mixture model. Other early work in this area can be found in [6,51], and a recent review of model-based clustering is given by McNicholas [38].

Although the Gaussian mixture model is well-established for clustering, largely due to its mathematical tractability, quite some work has been done in the area of non-Gaussian mixtures. For example, some work has been done using symmetric component densities that parameterize concentration (tail weight), e.g., the *t* distribution [3,4,33,47] and the power exponential distribution [14]. There has also been work on mixtures for discrete data [e.g.,

☆ The authors gratefully acknowledge the very helpful comments of two anonymous reviewers as well as the support of a Vanier Canada Graduate Scholarship (Gallaugher) and the Canada Research Chairs program (McNicholas).

* Corresponding author.
E-mail address: paul@math.mcmaster.ca (P.D. McNicholas).

https://doi.org/10.1016/j.patcog.2018.02.025
0031-3203/

[9,27]] as well as several examples of mixtures of skewed distributions such as the NIG distribution [28,54], the skew-*t* distribution [30–32,43–45,57,58], the shifted asymmetric Laplace distribution [18,42], the variance-gamma distribution [40], the generalized hyperbolic distribution [10], and others [e.g., [17,19]].

Recently, there has been an interest in the mixtures of matrix variate distributions, e.g., Anderlucci and Viroli [2] consider multivariate longitudinal data with the matrix variate normal distribution and Doğru et al. [15] consider a finite mixture of matrix variate *t* distributions.

2.2. Matrix variate distributions

Three-way data such as multivariate longitudinal data or greyscale image data can be easily modelled using a matrix variate distribution. There are many examples of such distributions presented in the literature, the most notable being the matrix variate normal distribution. In Section 2.1, $\mathbf{X}$ was used in the typical way to denote a multivariate random variable and $\mathbf{x}$ was used to denote its realization. Hereafter, $\mathbf{X}$ is used to denote a realization of a random matrix $\mathscr{X}$. An $n\times p$ random matrix $\mathscr{X}$ follows an $n\times p$ matrix variate normal distribution with location parameter $\mathbf{M}$ and scale matrices $\mathbf{\Sigma}$ and $\mathbf{\Psi}$ of dimensions $n\times n$ and $p\times p$, respectively, denoted by $\mathcal{N}_{n\times p}(\mathbf{M},\mathbf{\Sigma},\mathbf{\Psi})$ if the density of $\mathscr{X}$ can be written as

$$f(\mathbf{X}\mid\mathbf{M},\mathbf{\Sigma},\mathbf{\Psi}) = \frac{1}{(2\pi)^{\frac{np}{2}}|\mathbf{\Sigma}|^{\frac{p}{2}}|\mathbf{\Psi}|^{\frac{n}{2}}}\exp\left\{-\frac{1}{2}\operatorname{tr}\left(\mathbf{\Sigma}^{-1}(\mathbf{X}-\mathbf{M})\mathbf{\Psi}^{-1}(\mathbf{X}-\mathbf{M})'\right)\right\}. \tag{1}$$

A well known property of the matrix variate normal distribution, see [25], is

$$\mathscr{X}\sim\mathcal{N}_{n\times p}(\mathbf{M},\mathbf{\Sigma},\mathbf{\Psi}) \iff \operatorname{vec}(\mathscr{X})\sim\mathcal{N}_{np}(\operatorname{vec}(\mathbf{M}),\mathbf{\Psi}\otimes\mathbf{\Sigma}), \tag{2}$$

where $\mathcal{N}_{np}(\cdot)$ is the multivariate normal density with dimension np, $\operatorname{vec}(\cdot)$ is the vectorization operator, and $\otimes$ is the Kronecker product.

The matrix variate normal distribution has many elegant mathematical properties that have made it so popular, e.g., [56] uses a mixture of matrix variate normal distributions for clustering. However, there are non-normal examples such as the Wishart distribution [60] and the skew-normal distribution, e.g., [13,16,25]. More information on matrix variate distributions can be found in [24].

2.3. The generalized inverse Gaussian distribution

The generalized inverse Gaussian distribution has two different parameterizations, both of which will be useful. A random variable Y has a generalized inverse Gaussian (GIG) distribution parameterized by a, b and λ, denoted by $\text{GIG}(a,b,\lambda)$, if its probability density function can be written as

$$f(y|a,b,\lambda) = \frac{(a/b)^{\frac{\lambda}{2}}y^{\lambda-1}}{2K_{\lambda}(\sqrt{ab})}\exp\left\{-\frac{ay+b/y}{2}\right\},$$

where

$$K_{\lambda}(u) = \frac{1}{2}\int_0^{\infty}y^{\lambda-1}\exp\left\{-\frac{u}{2}\left(y+\frac{1}{y}\right)\right\}dy$$

is the modified Bessel function of the third kind with index λ. Expectations of some functions of a GIG random variable have a mathematically tractable form, e.g.:

$$\mathbb{E}(Y) = \sqrt{\frac{b}{a}}\frac{K_{\lambda+1}(\sqrt{ab})}{K_{\lambda}(\sqrt{ab})}, \tag{3}$$

$$\mathbb{E}(1/Y) = \sqrt{\frac{a}{b}}\frac{K_{\lambda+1}(\sqrt{ab})}{K_{\lambda}(\sqrt{ab})} - \frac{2\lambda}{b}, \tag{4}$$

$$\mathbb{E}(\log Y) = \log\left(\sqrt{\frac{b}{a}}\right) + \frac{1}{K_{\lambda}(\sqrt{ab})}\frac{\partial}{\partial\lambda}K_{\lambda}(\sqrt{ab}). \tag{5}$$

Although this parameterization of the GIG distribution will be useful for parameter estimation, for the purposes of deriving the density of the matrix variate generalized hyperbolic distribution, it is more useful to take the parameterization

$$g(y|\omega,\eta,\lambda) = \frac{(w/\eta)^{\lambda-1}}{2\eta K_{\lambda}(\omega)}\exp\left\{-\frac{\omega}{2}\left(\frac{w}{\eta}+\frac{\eta}{w}\right)\right\}, \tag{6}$$

where $\omega=\sqrt{ab}$ and $\eta=\sqrt{a/b}$. For notational clarity, we will denote the parameterization given in (6) by $\text{I}(\omega,\eta,\lambda)$.

2.4. Skewed matrix variate distributions

The work of Gallaugher and McNicholas [20,21] presents a total of four skewed matrix variate distributions, the matrix variate skew-*t*, generalized hyperbolic, variance-gamma and NIG distributions. Each of these distributions is derived from a matrix variate normal variance-mean mixture. In this representation, the random matrix $\mathscr{X}$ has the representation

$$\mathscr{X} = \mathbf{M} + W\mathbf{A} + \sqrt{W}\mathscr{V}, \tag{7}$$

where $\mathbf{M}$ and $\mathbf{A}$ are $n\times p$ matrices representing the location and skewness respectively, $\mathscr{V}\sim\mathcal{N}_{n\times p}(\mathbf{0},\mathbf{\Sigma},\mathbf{\Psi})$, and $W>0$ is a random variable with density $h(w|\boldsymbol{\theta})$.

In [20], the matrix variate skew-*t* distribution, with ν degrees of freedom, is shown to arise as a special case of (7) with $W^{\text{ST}}\sim\text{IGamma}(\nu/2,\nu/2)$, where $\text{IGamma}(\cdot)$ denotes the inverse Gamma distribution with density

$$f(y\mid a,b) = \frac{b^a}{\Gamma(a)}y^{-a-1}\exp\left\{-\frac{b}{y}\right\}.$$

The resulting density of $\mathscr{X}$ is

$$f_{\text{MVST}}(\mathbf{X}\mid\boldsymbol{\vartheta}) = \frac{2\left(\frac{\nu}{2}\right)^{\frac{\nu}{2}}\exp\left\{\operatorname{tr}(\mathbf{\Sigma}^{-1}(\mathbf{X}-\mathbf{M})\mathbf{\Psi}^{-1}\mathbf{A}')\right\}}{(2\pi)^{\frac{np}{2}}|\mathbf{\Sigma}|^{\frac{p}{2}}|\mathbf{\Psi}|^{\frac{n}{2}}\Gamma(\frac{\nu}{2})}\left(\frac{\delta(\mathbf{X};\mathbf{M},\mathbf{\Sigma},\mathbf{\Psi})+\nu}{\rho(\mathbf{A},\mathbf{\Sigma},\mathbf{\Psi})}\right)^{-\frac{\nu+np}{4}} \times K_{-\frac{\nu+np}{2}}\left(\sqrt{[\rho(\mathbf{A},\mathbf{\Sigma},\mathbf{\Psi})][\delta(\mathbf{X};\mathbf{M},\mathbf{\Sigma},\mathbf{\Psi})+\nu]}\right),$$

where

$$\delta(\mathbf{X};\mathbf{M},\mathbf{\Sigma},\mathbf{\Psi}) = \operatorname{tr}(\mathbf{\Sigma}^{-1}(\mathbf{X}-\mathbf{M})\mathbf{\Psi}^{-1}(\mathbf{X}-\mathbf{M})'),$$
$$\rho(\mathbf{A};\mathbf{\Sigma},\mathbf{\Psi}) = \operatorname{tr}(\mathbf{\Sigma}^{-1}\mathbf{A}\mathbf{\Psi}^{-1}\mathbf{A}')$$

and $\nu>0$. For notational clarity, this distribution will be denoted by $\text{MVST}(\mathbf{M},\mathbf{A},\mathbf{\Sigma},\mathbf{\Psi},\nu)$.

In [21], one of the distributions considered is a matrix variate generalized hyperbolic distribution. This again is the result of a special case of (7) with $W^{\text{GH}}\sim\text{I}(\omega,1,\lambda)$. This distribution will be denoted by $\text{MVGH}(\mathbf{M},\mathbf{A},\mathbf{\Sigma},\mathbf{\Psi},\lambda,\omega)$, and the density is

$$f_{\text{MVGH}}(\mathbf{X}|\boldsymbol{\vartheta}) = \frac{\exp\left\{\operatorname{tr}(\mathbf{\Sigma}^{-1}(\mathbf{X}-\mathbf{M})\mathbf{\Psi}^{-1}\mathbf{A}')\right\}}{(2\pi)^{\frac{np}{2}}|\mathbf{\Sigma}|^{\frac{p}{2}}|\mathbf{\Psi}|^{\frac{n}{2}}K_{\lambda}(\omega)}\left(\frac{\delta(\mathbf{X};\mathbf{M},\mathbf{\Sigma},\mathbf{\Psi})+\omega}{\rho(\mathbf{A},\mathbf{\Sigma},\mathbf{\Psi})+\omega}\right)^{\frac{(\lambda-\frac{np}{2})}{2}} \times K_{(\lambda-np/2)}\left(\sqrt{[\rho(\mathbf{A},\mathbf{\Sigma},\mathbf{\Psi})+\omega][\delta(\mathbf{X};\mathbf{M},\mathbf{\Sigma},\mathbf{\Psi})+\omega]}\right),$$

where $\lambda\in\mathbb{R}$ and $\omega>0$.