Three Skewed Matrix Variate Distributions

Michael P.B. Gallaugher and Paul D. McNicholas

Dept. of Mathematics & Statistics, McMaster University, Hamilton, Ontario, Canada.

Abstract

Three-way data can be conveniently modelled by using matrix variate distributions. Although there has been a lot of work for the matrix variate normal distribution, there is little work in the area of matrix skew distributions. Three new matrix variate distributions that incorporate skewness, as well as other flexible properties such as concentration, are presented. Equivalences to multivariate analogues are presented, and moment generating functions are derived. Maximum likelihood parameter estimation is discussed, and simulated data is used for illustration.

1 Introduction

Matrix variate distributions are very useful in modelling three way data, such as multivariate longitudinal data. Although the matrix normal distribution is widely used, there is relative paucity in the area of matrix skewed distributions. Herein, we present matrix variate extensions of three already well established multivariate distributions using matrix normal variance-mean mixtures. Specifically, we consider a matrix variate generalized hyperbolic distribution, a matrix variate variance-gamma distribution and a matrix variate normal inverse Gaussian (NIG) distribution. In addition, we discuss parameter estimation which is then performed on simulated data. The remainder of this paper is laid out as follows. Some
background is given (Section 2) before the three proposed matrix variate distributions are derived (Section 3). In Section 4, we discuss parameter estimation and Section 5 looks at some simulated data analyses. We end with a discussion in Section 6.

2 Background

2.1 The Matrix Variate Normal and Related Distributions

One of the most mathematically tractable examples of a matrix variate distribution is the matrix variate normal distribution. An \( n \times p \) random matrix \( \mathcal{X} \) follows a matrix variate normal distribution if its probability density function can be written as

\[
f(\mathcal{X}|M, \Sigma, \Psi) = \frac{1}{(2\pi)^{\frac{np}{2}}|\Sigma|^{\frac{p}{2}}|\Psi|^{\frac{n}{2}}} \exp\left\{-\frac{1}{2} \text{tr}\left(\Sigma^{-1}(\mathcal{X} - M)\Psi^{-1}(\mathcal{X} - M)^\prime\right)\right\},
\]

where \( M \) is an \( n \times p \) location matrix, \( \Sigma \) is an \( n \times n \) scale matrix for the rows of \( \mathcal{X} \) and \( \Psi \) is a \( p \times p \) scale matrix for the columns of \( \mathcal{X} \). We denote this distribution by \( \mathcal{N}_{n \times p}(M, \Sigma, \Psi) \) and, for notational clarity, we will denote the random matrix by \( \mathcal{X} \) and its realization by \( \mathcal{X} \). One useful property of the matrix variate normal distribution, as given in Harrar and Gupta (2008), is

\[
\mathcal{X} \sim \mathcal{N}_{n \times p}(M, \Sigma, \Psi) \iff \text{vec} \mathcal{X} \sim \mathcal{N}_{np}(M, \Psi \otimes \Sigma),
\]

(1)

where \( \mathcal{N}_{np}(\cdot) \) denotes the multivariate normal distribution with dimension \( np \), and \( \otimes \) denotes the Kronecker product.

Although the matrix variate normal is probably the most well known matrix variate distribution, there are other examples. For example, the Wishart distribution (Wishart, 1928) was shown to be the distribution of the sample covariance matrix for a random sample from a multivariate normal distribution. There are also a few examples of a matrix variate skew normal distribution such as Chen and Gupta (2005), Domínguez-Molina et al. (2007) and Harrar and Gupta (2008). Most recently, Gallaugher and McNicholas (2017a), considered a matrix variate skew t distribution using a matrix normal variance-mean mixture.
There are also a few examples of mixtures of matrix variate distributions. Anderlucci et al. (2015) considered a mixture of matrix variate normal distributions for clustering multivariate longitudinal data and Doğru et al. (2016) considered a mixture of matrix variate t distributions.

2.2 The Inverse and Generalized Inverse Gaussian Distributions

The derivation of the matrix distributions and parameter estimation discussed in Sections 3 and 4, will rely heavily on the generalized inverse Gaussian distribution, and to a lesser extent the inverse Gaussian distribution. A random variable, $Y$, follows an inverse Gaussian distribution if its probability density function is of the form

$$f(y|\delta, \gamma) = \frac{\delta}{\sqrt{2\pi}} \exp\{\delta \gamma\} y^{-3/2} \exp\left\{-\frac{1}{2} \left(\frac{\delta^2}{y} + \gamma^2 y\right)\right\}.$$ 

For notational purposes, we will denote this distribution by IG($\delta, \gamma$).

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

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

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 $\lambda$. Some expectations of functions of a GIG random variable with this parameterization have a mathematically tractable form and are given by

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

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

$$\mathbb{E}(Y^2) = \frac{b}{a} \frac{K_{\lambda+2}(\sqrt{ab})}{K_\lambda(\sqrt{ab})}.$$
\[ \mathbb{E}(\log Y) = \log \left( \sqrt{\frac{b}{a}} \right) + \frac{1}{K_\lambda(\sqrt{ab})} \frac{\partial}{\partial \lambda} K_\lambda(\sqrt{ab}). \] (4) 

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

\[ g(y|\omega, \eta, \lambda) = \left( \frac{w}{\eta} \right)^{\lambda-1} \frac{\exp \left\{ -\frac{\omega}{2} \left( \frac{w}{\eta} + \frac{\eta}{w} \right) \right\}}{2\eta K_\lambda(\omega)}, \] (5)

where \( \omega = \sqrt{ab} \) and \( \eta = \sqrt{a/b} \). For notational clarity, we will denote the parameterization given in (5) by \( I(\omega, \eta, \lambda) \).

### 2.3 Variance-Mean Mixtures

A \( p \)-variate random vector \( X \) defined in terms of a variance-mean mixture, has a probability density function of the form

\[ f(x) = \int_0^\infty \phi_p(x|\mu + w\alpha, w\Sigma)h(w|\theta)dw, \]

where the random variable \( W \) has density function \( h(w|\theta) \), and \( \phi_p(\cdot) \) represents the density function of the \( p \)-variate Gaussian distribution. This representation is equivalent to writing

\[ X = \mu + W\alpha + \sqrt{W}V, \] (6)

where \( \mu \) is a location parameter, \( \alpha \) is the skewness, \( V \sim \mathcal{N}_p(0, \Sigma) \) with \( \Sigma \) as the scale matrix, and \( W \) has density function \( h(w|\theta) \). Many multivariate distributions can be obtained through a variance mean mixture by changing the distribution of \( W \), (cf. McNicholas, 2016, Ch. 6). For example, the \( p \)-dimensional generalized hyperbolic distribution, \( \text{GH}_p(\mu, \alpha, \Sigma, \psi, \chi, \lambda) \), as given in McNeil et al. (2005), was shown to arise as a special case of (6) by taking \( W \sim GIG(\psi, \chi, \lambda) \). However, there was a restriction that \( |\Sigma| = 1 \). Simply relaxing this constraint results in an identifiability problem. In Browne and McNicholas (2015), this was discussed, and the authors proposed the reparameterization \( \omega = \sqrt{\psi\chi}, \eta = \sqrt{\chi/\psi} \). The representation of \( X \) is then as in (6), with \( W \sim I(\omega, 1, \lambda) \).
The $p$-dimensional variance-gamma distribution, $\text{VG}_p(\mu, \alpha, \Sigma, \lambda, \psi)$, results as a limiting case of the generalized hyperbolic by taking $\lambda > 0$, and $\chi \to 0$. The exact details can be found in McNicholas et al. (2017), however, essentially, the variance-gamma distribution also arises as a special case of (6), with $W \sim \text{gamma}(\lambda, \psi/2)$, where $\text{gamma}(a,b)$ denotes the gamma distribution with density

$$f(w|a,b) = \frac{b^a}{\Gamma(a)} w^{a-1} \exp\{-bw\}.$$ 

However, we again have an identifiability issue using this representation if we remove the constraint $|\Sigma| = 1$. In McNicholas et al. (2017), the authors propose setting $E(W) = 1$, resulting in the reparameterization $\gamma := \lambda = \psi/2$.

Finally, we have the $p$-dimensional Gaussian distribution, $\text{NIG}_p(\mu, \alpha, \Sigma, \delta, \gamma)$. In Karlis and Santourian (2009), the authors derived the $p$-dimensional NIG distribution using a variance-mean mixture with $W \sim \text{IG}(\delta, \gamma)$. However, there was once again a restriction on the determinant of $\Sigma$. To remove this restriction and maintain identifiability, the authors set $\delta = 1$, and set $\tilde{\gamma} = \gamma$. This formulation was also used in O’Hagan et al. (2016).

3 Three Matrix Variate Skew Distributions

3.1 Matrix Normal Variance-Mean Mixture

We now derive densities for matrix variate versions of the generalized hyperbolic, variance-gamma and NIG distributions. For all three of these distributions, we consider a matrix normal variance-mean mixture, where we can take the representation

$$\mathcal{X} = \mathbf{M} + W \mathbf{A} + \sqrt{W} \mathbf{Y},$$

where $\mathbf{Y} \sim \mathcal{N}_{n \times p}(\mathbf{0}_{n \times p}, \mathbf{\Sigma}, \mathbf{\Psi})$ with $\mathbf{0}_{n \times p}$ representing the $n \times p$ zero matrix, $\mathbf{M}$ is an $n \times p$ location matrix, $\mathbf{A}$ is an $n \times p$ skewness matrix, and $W$ is a random variable with density $h(\theta)$. We now derive three matrix variate distributions using this representation with different distributions for $W$. 

5
3.2 A Matrix Variate Generalized Hyperbolic Distribution

We now derive the density of a matrix variate generalized hyperbolic distribution. In this case, to avoid the identifiability issue discussed in Browne and McNicholas (2015), we take $W \sim I(\omega, 1, \lambda)$, where $\omega$ is a concentration parameter and $\lambda$ is the index parameter. It then follows that $X|w \sim N_{n \times p}(M + wA, w\Sigma, \Psi)$ and thus the joint density of $X$ and $W$ is

$$f(X, w|\vartheta) = f(X|w)f(w)$$

$$= w^{\lambda - \frac{np}{2} - 1} \frac{1}{(2\pi)^{\frac{np}{2}}|\Sigma|^\frac{p}{2}|\Psi|^\frac{p}{2}K_\lambda(\omega)}$$

$$\times \exp \left\{ -\frac{1}{2w} \left( \operatorname{tr}(\Sigma^{-1}(X - M - wA)\Psi^{-1}(X - M - wA)' + \omega) - \omega w/2 \right) \right\},$$

where $\vartheta = (M, A, \Sigma, \Psi, \omega, \lambda)$.

We note that the exponential term in (8) can be written as

$$\exp \left\{ \operatorname{tr}(\Sigma^{-1}(X - M)\Psi^{-1}A') \right\} \times \exp \left\{ -\frac{1}{2w} \left[ \delta(X; M, \Sigma, \Psi) + \omega \right] + w (\rho(A, \Sigma, \Psi) + \omega) \right\}$$

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

Therefore, the marginal density of $X$ is

$$f(X) = \int_0^\infty f(X, w)dw$$

$$= \frac{1}{(2\pi)^{\frac{np}{2}}|\Sigma|^\frac{p}{2}|\Psi|^\frac{p}{2}K_\lambda(\omega)} \exp \left\{ \operatorname{tr}(\Sigma^{-1}(X - M)\Psi^{-1}A') \right\}$$

$$\times \frac{1}{2} \int_0^\infty w^{\lambda - \frac{np}{2} - 1} \exp \left\{ -\frac{1}{2w} \left[ \delta(X; M, \Sigma, \Psi) + \omega \right] + w (\rho(A, \Sigma, \Psi) + \omega) \right\} dw.$$  (9)

Making the change of variables

$$y = \frac{\sqrt{\rho(A, \Sigma, \Psi) + \omega}}{\sqrt{\delta(X; M, \Sigma, \Psi) + \omega}} w,$$
(9) becomes,

\[
f_{\text{MVGH}}(X|\theta) = \exp \left\{ \frac{\text{tr}(\Sigma^{-1}(X - M)\Psi^{-1}A')}{(2\pi)^{np} |\Sigma|^{\frac{p}{2}} |\Psi|^{\frac{p}{2}} K_\lambda(\omega)} \right\}
\]

\[
\left( \frac{\delta(X; M, \Sigma, \Psi) + \omega}{\frac{\lambda - np}{2}} \right)^{\frac{(\lambda - np)}{2}}
\times K_{\lambda-np/2} \left( \sqrt{[\rho(A, \Sigma, \Psi) + \omega] [\delta(X; M, \Sigma, \Psi) + \omega]} \right),
\]

where \( \omega > 0 \) is a concentration parameter, and \( \lambda \in \mathbb{R} \) is an index parameter.

We note that the density of \( X \), as derived here, is similar to that in Browne and McNicholas (2015), and we denote this distribution by MVGH\(_{n \times p}(M, A, \Sigma, \Psi, \lambda, \omega)\). For the purposes of parameter estimation, note that the conditional density of \( W \) is

\[
f(w|X) = \frac{f(X|w)f(w)}{f(X)}
\]

\[
= \left( \frac{\rho(A, \Sigma, \Psi) + \omega}{\delta(X; M, \Sigma, \Psi) + \omega} \right)^{\frac{(\lambda - np)}{2}}
\times \frac{w^{\lambda-2np/2-1}}{2K_{\lambda-np/2} \sqrt{[\rho(A, \Sigma, \Psi) + \omega] [\delta(X; M, \Sigma, \Psi) + \omega]}}
\times \exp \left\{ -\frac{(\rho(A, \Sigma, \Psi) + \omega)w + [\delta(X; M, \Sigma, \Psi) + \omega]/w}{2} \right\}
\]

Therefore, \( W|X \sim \text{GIG} \left( \rho(A, \Sigma, \Psi) + \omega, \delta(X; M, \Sigma, \Psi) + \omega, \lambda - np/2 \right) \). We note that a matrix variate generalized hyperbolic distribution was derived in Thabane and Safiul Haq (2004); however, this formulation is different to the one considered herein.

### 3.3 A Matrix Variate Variance-Gamma Distribution

We now derive the density of a matrix variate variance-gamma distribution in much the same way as the generalized hyperbolic case. However, we now take \( W \sim \text{gamma}(\gamma, \gamma) \), resulting in the joint distribution

\[
f(X, W|\theta) = \frac{\gamma^\gamma}{(2\pi)^{\frac{np}{2}} |\Sigma|^{\frac{p}{2}} |\Psi|^{\frac{p}{2}} \Gamma(\gamma)} w^{\gamma-\frac{np}{2}-1}
\times \exp \left\{ -\frac{1}{2w} \text{tr} \left( \Sigma^{-1}(X - M - wA)\Psi^{-1}(X - M - wA)' \right) - \gamma w \right\}.
\]
Following the same procedure as before, the density of $\mathcal{X}$ is then

$$f_{\text{MVVG}}(X|\vartheta) = \frac{2^{2\gamma} \exp \left\{ \text{tr}(\Sigma^{-1}(X-M)\Psi^{-1}A') \right\} \left( \delta(X; M, \Sigma, \Psi) + 2\gamma \right)^{\frac{\gamma - np/2}{2}}}{(2\pi)^{\frac{d}{2}} |\Sigma|^\frac{s}{2} |\Psi|^\frac{s}{2} \Gamma(\gamma)} \times K_{\gamma - \frac{np}{2}} \left( \sqrt[\gamma - \frac{np}{2}] \left[ \rho(A, \Sigma, \Psi) + 2\gamma \right] \left[ \delta(X; M, \Sigma, \Psi) \right] \right),$$

where $\gamma > 0$. We will denote this distribution by $\text{MVVG}_{n\times p}(M, A, \Sigma, \Psi, \gamma)$. Note that $W|X \sim \text{GIG} (\rho(A, \Sigma, \Psi) + 2\gamma, \delta(X; M, \Sigma, \Psi), \gamma - np/2)$.

### 3.4 A Matrix Variate NIG Distribution

Finally, we consider a matrix variate NIG distribution. Derived in much the same way as the previous distributions, we take $W \sim \text{IG}(1, \tilde{\gamma})$. The joint density of $\mathcal{X}$ and $W$ is

$$f(X, W|\vartheta) = \frac{1}{(2\pi)^{\frac{d}{2} + 1} |\Sigma|^\frac{s}{2} |\Psi|^\frac{s}{2}} w^{-(\frac{3+np}{2})} w^{-\frac{1}{2} \text{tr} \left( \Sigma^{-1}(X-M-wA)\Psi^{-1}(X-M-wA)' \right) + 1} - \frac{w^{-\tilde{\gamma}^2}}{2} + \tilde{\gamma} \right)};$$

and the density of $\mathcal{X}$ is then

$$f_{\text{MVNIG}}(X|\vartheta) = \frac{2 \exp \left\{ \text{tr}(\Sigma^{-1}(X-M)\Psi^{-1}A') + \tilde{\gamma} \right\} \left( \delta(X; M, \Sigma, \Psi) + 1 \right)^{-(1+np)/4}}{(2\pi)^{\frac{d}{2} + 1} |\Sigma|^\frac{s}{2} |\Psi|^\frac{s}{2}} \times K_{-(1+np)/2} \left( \sqrt[\gamma - \frac{np}{2}] \left[ \rho(A, \Sigma, \Psi) + \tilde{\gamma}^2 \right] \left[ \delta(X; M, \Sigma, \Psi) + 1 \right] \right),$$

where $\tilde{\gamma} > 0$. We denote this distribution by $\text{MVNIG}_{n\times p}(M, A, \Sigma, \Psi, \tilde{\gamma})$, and note that $W|X \sim \text{GIG} (\rho(A, \Sigma, \Psi) + \tilde{\gamma}^2, \delta(X; M, \Sigma, \Psi) + 1, -(1 + np)/2)$.

### 3.5 Some Properties

One interesting element that we see for all three of these distributions is that there is an equivalence between each of these matrix variate distributions and their multivariate coun-
terparts. Specifically,

\[ X \sim MVGH_{n \times p}(M, A, \Sigma, \Psi, \omega, \lambda) \iff \text{vec}(X) \sim GH_{np}(\text{vec}(M), \text{vec}(A), \Psi \otimes \Sigma, \omega, \lambda), \]

\[ X \sim MVVG_{n \times p}(M, A, \Sigma, \Psi, \gamma) \iff \text{vec}(X) \sim VG_{np}(\text{vec}(M), \text{vec}(A), \Psi \otimes \Sigma, \gamma), \]

\[ X \sim MVNIG_{n \times p}(M, A, \Sigma, \Psi, \tilde{\gamma}) \iff \text{vec}(X) \sim NIG_{np}(\text{vec}(M), \text{vec}(A), \Psi \otimes \Sigma, \tilde{\gamma}). \]

These properties can be easily seen by using the representation of \( X \) given in (7) as well as the property of the matrix variate normal distribution given in (1).

We can also easily derive the moment generating functions for each of these three distributions. Using the representation for a random matrix \( X \) given in (7) and the moment generating function for the matrix variate normal distribution given in Dutilleul (1999), we have that the moment generating function in the general case of a matrix normal variance-mean mixture is

\[
M_X(T) = \mathbb{E}[\exp\{\text{tr}(T'X)\}]
\]

\[
= \mathbb{E}[\mathbb{E}[\exp\{\text{tr}(T'X)\} | W]]
\]

\[
= \exp\{\text{tr}(T'M)\} \mathbb{E}[\exp\{W \text{tr}(T'A + T\Sigma T'\Psi)\}]
\]

\[
= \exp\{\text{tr}(T'M)\} M_W(\text{tr}(T'A + T\Sigma T'\Psi)),
\]

where \( M_W(\cdot) \) is the moment generating function of \( W \).

Therefore, in the case of the generalized inverse Gaussian distribution, we have that the moment generating function is

\[
\exp\{\text{tr}(T'M)\} \left[1 - \frac{\text{tr}(T'A + T\Sigma T'\Psi)}{\omega}\right]^{-\frac{1}{2}} K_\lambda\left(\frac{\sqrt{\omega(\omega - 2 \text{tr}(T'A + T\Sigma T'\Psi))}}{K_\lambda(\omega)}\right)
\]

In the case of the variance gamma distribution, the moment generating function is

\[ M_X^{MVVG}(T) = \exp\{\text{tr}(T'M)\} \left(1 - \frac{\text{tr}(T'A + T\Sigma T'\Psi)}{\gamma}\right)^{-\gamma} \]

for \( \text{tr}(T'A + T\Sigma T'\Psi) < \gamma \).

Finally, in the case of the NIG distribution, the moment generating function is

\[ M_X^{MVVG}(T) = \exp\{\text{tr}(T'M)\} \exp\left\{\tilde{\gamma} \left(1 - \sqrt{1 - \frac{2 \text{tr}(T'A + T\Sigma T'\Psi)}{\tilde{\gamma}^2}}\right)\right\}.\]
4 Parameter Estimation

4.1 An ECM Algorithm Framework

Fortunately, parameter estimation is very similar for the three distributions derived previously. Suppose we observe $\mathcal{X} = (X_1, X_2, \ldots, X_N)$, where $\mathcal{X}$ is a random sample from one of the MVGH, MVVG or MVNIG distribution. It is convenient to consider this observed data to be incomplete, and introduce the latent variables $W_i$, where $W_i$ has density function $h(w|\theta)$. The complete-data log likelihood is then

$$L_c(\theta | \mathcal{X}, W) = (L_1 + C_1) + (L_2 + C_2), \quad (10)$$

where $C_1$ and $C_2$ are constants with respect to the parameters, $L_1 = \sum_{i=1}^N \log h(w_i|\theta) - C_1$,

$$L_2 = -\frac{Np}{2} \log |\Sigma| - \frac{Nn}{2} \log |\Psi| + \frac{1}{2} \sum_{i=1}^N \text{tr} (\Sigma^{-1}(X_i - M)\Psi^{-1}A') + \frac{1}{2} \sum_{i=1}^N \text{tr} (\Sigma^{-1}A\Psi^{-1}(X_i - M)')$$

$$- \frac{1}{2} \sum_{i=1}^N \frac{1}{w_i} \text{tr}(\Sigma^{-1}(X_i - M)\Psi^{-1}(X_i - M)') - \frac{1}{2} \sum_{i=1}^N w_i \text{tr}(\Sigma^{-1}A\Psi^{-1}A').$$

We take this opportunity to note that while it is true that

$$\frac{1}{2} \text{tr} (\Sigma^{-1}(X_i - M)\Psi^{-1}A') + \frac{1}{2} \text{tr} (\Sigma^{-1}A\Psi^{-1}(X_i - M)') = \text{tr} (\Sigma^{-1}(X_i - M)\Psi^{-1}A'),$$

it is necessary to separate the trace into two terms to ensure symmetry of the updates for $\Sigma$ and $\Psi$.

We now maximize (10) using an expectation conditional maximization algorithm (ECM; Meng and Rubin 1993). We first note that $L_2$ is the portion of the likelihood that depends on $M, A, \Sigma$ and $\Psi$ and not on the parameters related to the distribution of $W$. Thus the updates for these parameters would be of the same form for each of the distributions we consider. The ECM algorithm would proceed as follows.

1) Initialization: Initialize the parameters $M, A, \Sigma, \Psi$ and other parameters.

2) E Step: Update $a_i, b_i, c_i$, where

$$a_i^{(t+1)} = \mathbb{E}(W_i|X_i, \hat{\theta}^{(t)}), \quad b_i^{(t+1)} = \mathbb{E}(1/W_i|X_i, \hat{\theta}^{(t)}), \quad c_i^{(t+1)} = \mathbb{E}(\log(W_i)|X_i, \hat{\theta}^{(t)}).$$
The precise updates would depend on the distribution. However, recall that in each case, the conditional distribution of $W|X$ was a generalized inverse Gaussian distribution. Thus, we would just calculate these expectations using Equations (2)-(4) using an appropriate choice of $a, b$ and $\lambda$.

3) **First CM Step**: Update the parameters $M$ and $A$.

$$
\hat{M}^{(t+1)} = \frac{\sum_{i=1}^{N} X_i \left( \bar{a}^{(t+1)} b_i^{(t+1)} - 1 \right)}{\sum_{i=1}^{N} \bar{a}(t+1) b_i^{(t+1)} - N}, \quad \hat{A}^{(t+1)} = \frac{\sum_{i=1}^{N} X_i \left( \bar{b}^{(t+1)} - b_i^{(t+1)} \right)}{\sum_{i=1}^{N} \bar{a}(t+1) b_i^{(t+1)} - N},
$$

where $\bar{a}^{(t+1)} = (1/N) \sum_{i=1}^{N} a_i^{(t+1)}$ and $\bar{b}^{(t+1)} = (1/N) \sum_{i=1}^{N} b_i^{(t+1)}$.

4) **Second CM Step**: Update $\Sigma$

$$
\hat{\Sigma}^{(t+1)} = \frac{1}{Np} \left\{ \frac{1}{N} \sum_{i=1}^{N} \left[ b_i^{(t+1)} \left( X_i - \hat{M}^{(t+1)} \right) \hat{\Psi}^{(t)}^{-1} \left( X_i - \hat{M}^{(t+1)} \right)' \right. \right.
- \hat{A}^{(t+1)} \hat{\Psi}^{(t)}^{-1} \left( X_i - \hat{M}^{(t+1)} \right)' - \left( X_i - \hat{M}^{(t+1)} \right) \hat{\Psi}^{(t)}^{-1} \hat{A}^{(t+1)',}
\left. \left. + a_i^{(t+1)} \hat{A}^{(t+1)} \hat{\Psi}^{(t)}^{-1} \hat{A}^{(t+1)'} \right] \right\}. \tag{11}
$$

5) **Third CM Step**: Update $\Psi$

$$
\hat{\Psi}^{(t+1)} = \frac{1}{Nn} \left\{ \frac{1}{N} \sum_{i=1}^{N} \left[ b_i^{(t+1)} \left( X_i - \hat{M}^{(t+1)} \right)' \hat{\Sigma}^{(t+1)}^{-1} \left( X_i - \hat{M}^{(t+1)} \right) \right. \right.
- \hat{A}^{(t+1)'} \hat{\Sigma}^{(t+1)}^{-1} \left( X_i - \hat{M}^{(t+1)} \right) - \left( X_i - \hat{M}^{(t+1)} \right)' \hat{\Sigma}^{(t+1)}^{-1} \hat{A}^{(t+1)'}
\left. \left. + a_i^{(t+1)} \hat{A}^{(t+1)'} \hat{\Sigma}^{(t+1)}^{-1} \hat{A}^{(t+1)} \right] \right\}. \tag{12}
$$

We note that the updates given in the first, second and third CM steps are identical for all three distributions.

6) **Additional CM Steps**: Now update the additional parameters, $\theta$, by maximizing $L_1$. The particulars for each distribution are discussed in Sections 4.2-4.4.

7) **Check for convergence** If the convergence criterion is not met, iterate through steps 2-6 until convergence.
4.2 Generalized Hyperbolic Distribution

In the case of the generalized hyperbolic distribution, we would update \( \lambda \) and \( \omega \). In this case, \[ \mathcal{L}_{1}^{\text{MVGH}} = N \log K_{\lambda}(\omega) - \lambda \sum_{i=1}^{N} \log w_{i} - \frac{1}{2} \omega \sum_{i=1}^{N} \left( w_{i} + \frac{1}{w_{i}} \right). \] (13)

The updates for \( \lambda \) and \( \omega \) cannot be obtained in closed form. However, Browne and McNicholas (2015) discuss numerical methods for these updates, and therefore because the portion of the likelihood function that include these parameters is the same as in the multivariate case, the updates described in Browne and McNicholas (2015) can be used directly here.

The updates for \( \lambda \) and \( \omega \) rely on the log convexity of \( K_{\lambda}(\omega) \), Baricz (2010), in both \( \lambda \) and \( \omega \) and maximizing (13) via conditional maximization. The resulting updates are

\[
\hat{\lambda}^{(t+1)} = \bar{c}^{(t+1)} \hat{\lambda}^{(t)} \left[ \frac{\partial}{\partial s} \log(K_{s}(\hat{\omega}^{(t)})) \right]_{s=\hat{\lambda}^{(t)}}^{-1} \tag{14}
\]

\[
\hat{\omega}^{(t+1)} = \hat{\omega}^{(t)} - \left[ \frac{\partial}{\partial s} q(\hat{\lambda}^{(t+1)}, s) \right]_{s=\hat{\omega}^{(t)}}^{-1} \left[ \frac{\partial^{2}}{\partial s^{2}} q(\hat{\lambda}^{(t+1)}, s) \right]_{s=\hat{\omega}^{(t)}}^{-1} \tag{15}
\]

where the derivative in (14) is calculated numerically, \( q(\lambda, \omega) = \mathcal{L}_{1}^{\text{MVGH}} \) and \( \bar{c}^{(t+1)} = (1/N) \sum_{i=1}^{N} c_{i}^{(t+1)} \).

The partials in (15) are described in Browne and McNicholas (2015), and can be written as

\[
\frac{\partial}{\partial \omega} q(\lambda, \omega) = \frac{1}{2} \left[ R_{\lambda}(\omega) + R_{-\lambda}(\omega) - (\bar{a}^{(t+1)} + \bar{b}^{(t+1)}) \right],
\]

and

\[
\frac{\partial^{2}}{\partial \omega^{2}} q(\lambda, \omega) = \frac{1}{2} \left[ R_{\lambda}(\omega)^{2} - \frac{1 + 2\lambda}{\omega} R_{\lambda}(\omega) - 1 + R_{-\lambda}(\omega)^{2} - \frac{1 - 2\lambda}{\omega} R_{-\lambda}(\omega) - 1 \right],
\]

where \( R_{\lambda}(\omega) = K_{\lambda+1}(\omega)/K_{\lambda}(\omega) \).

4.3 Variance-Gamma Distribution

In the case of the matrix variate variance-gamma,

\[ \mathcal{L}_{1}^{\text{MVVG}} = N \gamma \log \gamma - N \log \Gamma(\gamma) + \gamma \sum_{i=1}^{N} (\log w_{i} - w_{i}). \]
The update for $\gamma$, like the generalized hyperbolic case, cannot be obtained in closed form. Instead, the update, $\gamma^{(t+1)}$, is obtained by solving (16) for $\gamma$:

$$\log \gamma + 1 - \varphi(\gamma) + \tilde{c}^{(t+1)} - \tilde{a}^{(t+1)} = 0,$$

(16)

where $\varphi(\cdot)$ denotes the digamma function.

### 4.4 NIG Distribution

Finally, in the NIG case,

$$L_{1}^{\text{MVNIG}} = N\tilde{\gamma} - \frac{\tilde{\gamma}^2}{2} \sum_{i=1}^{N} w_i.$$

(17)

In this case, the update can be obtained in closed form as

$$\hat{\tilde{\gamma}}^{(t+1)} = \frac{N}{\sum_{i=1}^{N} a_i^{(t+1)}}.$$

### 4.5 A Note on Identifiability

It is important to note that the estimates for $\Sigma$ and $\Psi$ are only unique up to a positive constant. Therefore, to remove this identifiability issue, we must impose a constraint on either $\Sigma$ or $\Psi$. Anderlucci et al. (2015) suggest taking the trace trace of $\Psi$ to be set to $p$. However, a simpler constraint would be to just take the first diagonal element of $\Sigma$ to be set to 1, which is the constraint we use for our analyses.

### 4.6 Convergence Criterion

There are different convergence criteria that could be used. For our analyses, we use a convergence criterion based on the Aitken acceleration, Aitken (1926). At iteration $t$, the Aitken acceleration is

$$a^{(t)} = \frac{l^{(t+1)} - l^{(t)}}{l^{(t)} - l^{(t-1)}},$$

(18)
where \( l^{(t)} \) is the (observed) log-likelihood at iteration \( t \). The quantity in (18) can be used to derive an asymptotic estimate (i.e., an estimate of the value after very many iterations) of the log-likelihood at iteration \( t + 1 \), i.e.,

\[
l_{\infty}^{(t+1)} = l^{(t)} + \frac{1}{1 - a^{(t)}} (l_{\infty}^{(t+1)} - l^{(t)})
\]

(cf. Böning et al. 1994; Lindsay 1995). As in McNicholas et al. (2010), we stop our EM algorithms when \( l_{\infty}^{(t+1)} - l^{(t)} < \epsilon \), provided this difference is positive.

5 Analyses

We performed two simulations for each of the three different distributions. Common elements between the distributions are as follows. For each simulation, we took 50 datasets each with 100 observations. For each distribution, in Simulation 1, we took

\[
M_1 = \begin{pmatrix} 0 & 1 & -1 & 0 \\ 1 & 0 & 0 & -1 \\ 0 & 1 & -1 & 0 \end{pmatrix} \quad A_1 = \begin{pmatrix} 1 & -1 & 0 & 1 \\ 1 & -1 & 0 & 1 \\ 1 & -1 & 0 & 1 \end{pmatrix},
\]

as the location and skewness matrices respectively. In Simulation 2, these parameters were

\[
M_2 = \begin{pmatrix} -5 & 0 & 0 & 1 \\ -2 & 1 & 3 & 0 \\ 0 & 0 & 6 & 1 \end{pmatrix} \quad A_2 = \begin{pmatrix} 1 & -1 & 0 & 1 \\ 0.5 & -1 & 0 & -0.5 \\ 0 & -1 & 0 & 0 \end{pmatrix}.
\]

In both simulations, for all three distributions, the scale matrices, \( \Sigma \) and \( \Psi \) were

\[
\Sigma = \begin{pmatrix} 1 & 0.5 & 0.1 \\ 0.5 & 1 & 0.5 \\ 0.1 & 0.5 & 1 \end{pmatrix} \quad \Psi = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0.5 & 0.5 \\ 0 & 0.5 & 1 & 0.1 \\ 0 & 0.5 & 0.1 & 1 \end{pmatrix}.
\]

In the first simulation, we took \( \lambda_1 = 2, \omega_1 = 2 \) for the generalized hyperbolic, \( \gamma_1 = 2 \) for the variance-gamma, and \( \tilde{\gamma}_1 = 4 \) for the NIG. For the second simulation, we took \( \lambda_2 = \)
−2, \omega = 2 for the generalized hyperbolic, \gamma_2 = 4 for the variance-gamma and \tilde{\gamma}_2 = 2 for the NIG. In Figure 1 (Appendix A), we show the marginal distributions of the columns for each distribution of a typical dataset for Simulation 1. We label the columns V1, V2, V3, and V4. The marginal location is shown by the red dashed line. We can clearly see the differences in concentration around the location parameter with the MVNIG distribution being the most concentrated, and the MVGH distribution being the least concentrated. We also note the obvious difference in the skewness parameter, with V3 being symmetric around the location, whereas for V1, V2, and V4, the distribution is clearly skewed. In Figure 2 (Appendix A), we show similar plots for Simulation 2. We can clearly see the difference in the concentration parameter for the MVGH distribution.

We now look at the parameter estimates. We show the component-wise means and standard deviations (in brackets), of the parameter estimates. In Table 1, we show the results for Simulation 1. We see that for all three of the distributions, we get good average estimates. However, we see that for the matrix variate generalized hyperbolic distribution, we have slightly more variation in the estimates of the location parameters. Also, the matrix variate NIG has a lot more variation in the estimates of the skewness.

In Table 2, we show similar results for Simulation 2. One obvious result that is unexpected is the estimate for \lambda for the matrix variate generalized hyperbolic distribution. The estimate is very different from the true value, and there is a very large amount of variation. We also notice a deflation in absolute value for the estimates of the skewness as well as a fair amount of variation. One possible explanation is that the generalized hyperbolic distribution is over-parameterized, and thus the deflation in the estimates for the skewness could be compensation for the increased value of \lambda.

6 Discussion

We derived the densities and described parameter estimation for three matrix variate skew distributions using a matrix normal variance-mean mixture. The three distributions were
the matrix variate generalized hyperbolic, variance-gamma, and NIG distributions. When looking at the estimates in the simulations, we obtained fairly good results. One exception was the average estimate of $\lambda$ and the skewness matrix for the matrix variate generalized hyperbolic distribution. However, this could be due to over-parameterization.

We mentioned previously that a matrix variate generalized hyperbolic distribution was
Table 2: Component wise averages and standard deviations for the estimated parameters for Simulation 2 for each of the three distributions

### Generalized Hyperbolic

| M(sd)         | A(sd)       | Σ(sd)       | Ψ(sd)       | λ(sd)       | ω(sdf) |
|---------------|-------------|-------------|-------------|-------------|--------|
| \(-4.97  0.05  -0.03  1.02\) | \(0.57  -0.69  0.02  0.64\) | \(1.00  0.50  0.10\) | \(0.63  0.00  0.01  0.00\) | \(0.00  0.64  0.33  0.32\) |
| \(-1.89  1.01  3.00  0.05\) | \(0.23  -0.68  -0.02  -0.34\) | \(0.50  0.99  0.50\) | \(0.01  0.33  0.63  0.07\) | \(0.00  0.32  0.07  0.64\) |
| \(0.10  -0.01  5.98  0.97\) | \(-0.02  -0.64  0.04  0.02\) | \(0.10  0.50  1.00\) | \(0.00  0.32  0.07  0.64\) | \(0.01  0.33  0.63  0.07\) |
| \(0.212  0.281  0.282  0.247\) | \(0.526  0.820  0.272  0.660\) | \(0.000  0.055  0.061\) | \(0.606  0.057  0.068  0.045\) | \(1.63  0.022  0.299  0.297\) |
| \(0.199  0.266  0.245  0.259\) | \(0.276  0.779  0.255  0.388\) | \(0.055  0.117  0.079\) | \(0.057  0.581  0.299  0.297\) | \(0.068  0.299  0.596  0.068\) |
| \(0.251  0.160  0.239  0.218\) | \(0.338  0.665  0.173  0.242\) | \(0.061  0.079  0.112\) | \(0.045  0.297  0.068  0.067\) | \(0.410  0.222  0.299  0.297\) |

### Variance Gamma

| M(sd)         | A(sd)       | Σ(sd)       | Ψ(sd)       | γ(sdf) |
|---------------|-------------|-------------|-------------|--------|
| \(-4.98  0.01  0.04  0.96\) | \(0.98  -0.99  -0.00  1.04\) | \(1.00  0.51  0.10\) | \(0.99  -0.01  -0.01  0.00\) | \(0.99  -0.01  -0.01  0.00\) |
| \(-1.98  1.00  3.02  0.02\) | \(0.49  -0.98  0.01  -0.52\) | \(0.51  1.01  0.51\) | \(-0.01  0.98  0.47  0.51\) | \(-0.01  0.98  0.47  0.51\) |
| \(0.02  0.05  6.07  1.03\) | \(0.00  -1.05  -0.06  -0.04\) | \(0.10  0.51  1.02\) | \(-0.01  0.98  0.47  0.51\) | \(-0.01  0.98  0.47  0.51\) |
| \(0.280  0.229  0.254  0.260\) | \(0.307  0.269  0.256  0.282\) | \(0.000  0.048  0.063\) | \(0.121  0.064  0.053  0.060\) | \(4.20  (1.04)\) |
| \(0.233  0.240  0.260  0.216\) | \(0.248  0.256  0.222  0.247\) | \(0.048  0.095  0.081\) | \(0.064  0.103  0.074  0.072\) | \(0.053  0.074  0.121  0.059\) |
| \(0.238  0.242  0.266  0.195\) | \(0.260  0.245  0.232  0.225\) | \(0.063  0.081  0.129\) | \(0.060  0.072  0.059  0.126\) | \(0.321  0.222  0.299  0.297\) |

### Normal Inverse Gaussian

| M(sd)         | A(sd)       | Σ(sd)       | Ψ(sd)       | γ(sdf) |
|---------------|-------------|-------------|-------------|--------|
| \(-5.02  0.04  0.01  1.03\) | \(1.16  -1.18  0.01  1.02\) | \(1.00  0.49  0.11\) | \(1.02  0.01  0.01  0.02\) | \(1.02  0.01  0.01  0.02\) |
| \(-1.99  1.04  2.99  0.05\) | \(0.55  -1.19  0.04  -0.64\) | \(0.49  1.01  0.51\) | \(0.01  1.06  0.54  0.53\) | \(0.01  1.06  0.54  0.53\) |
| \(0.02  0.05  5.98  1.01\) | \(0.01  -1.11  0.04  0.02\) | \(0.11  0.51  1.00\) | \(0.01  1.06  0.54  0.53\) | \(0.01  1.06  0.54  0.53\) |
| \(0.143  0.134  0.133  0.137\) | \(0.506  0.446  0.306  0.418\) | \(0.000  0.045  0.053\) | \(0.250  0.065  0.064  0.072\) | \(2.12  (0.50)\) |
| \(0.137  0.123  0.140  0.117\) | \(0.390  0.462  0.323  0.357\) | \(0.045  0.107  0.077\) | \(0.065  0.285  0.175  0.139\) | \(0.064  0.175  0.281  0.072\) |
| \(0.148  0.120  0.128  0.114\) | \(0.298  0.433  0.271  0.249\) | \(0.053  0.077  0.119\) | \(0.072  0.139  0.072  0.245\) | \(0.064  0.175  0.281  0.072\) |

derived in [Thabane and Safiul Haq (2004)](https://www.jstor.org/stable/2759398), and a detailed comparison will be a subject of future work. One possible extension of the current work is to consider a multiple-scaled version of the three distributions described here. Another extension would be to look at mixtures of these distributions for use in clustering three way data. Finally, it would be interesting to consider placing a constrained covariance structure on Σ for possible use in
multivariate longitudinal data analysis.

References

Aitken, A. C. (1926). A series formula for the roots of algebraic and transcendental equations. *Proceedings of the Royal Society of Edinburgh* 45, 14–22.

Anderlucci, L., C. Virolì, et al. (2015). Covariance pattern mixture models for the analysis of multivariate heterogeneous longitudinal data. *The Annals of Applied Statistics* 9(2), 777–800.

Baricz, A. (2010). Turn type inequalities for some probability density functions. *Studia Scientiarum Mathematicarum Hungarica* 47, 175–189.

Böhning, D., E. Dietz, R. Schaub, P. Schlattmann, and B. Lindsay (1994). The distribution of the likelihood ratio for mixtures of densities from the one-parameter exponential family. *Annals of the Institute of Statistical Mathematics* 46, 373–388.

Browne, R. P. and P. D. McNicholas (2015). A mixture of generalized hyperbolic distributions. *Canadian Journal of Statistics* 43(2), 176–198.

Chen, J. T. and A. K. Gupta (2005). Matrix variate skew normal distributions. *Statistics* 39(3), 247–253.

Doğru, F. Z., Y. M. Bulut, and O. Arslan (2016). Finite mixtures of matrix variate t distributions. *Gazi University Journal of Science* 29(2), 335–341.

Domínguez-Molina, J. A., G. González-Farías, R. Ramos-Quiroga, and A. K. Gupta (2007). A matrix variate closed skew-normal distribution with applications to stochastic frontier analysis. *Communications in Statistics—Theory and Methods* 36(9), 1691–1703.

Dutilleul, P. (1999). The mle algorithm for the matrix normal distribution. *Journal of statistical computation and simulation* 64(2), 105–123.
Gallaugher, M. P. B. and P. D. McNicholas (2017a). A matrix variate skew-t distribution. *Stat* 6(1).

Gallaugher, M. P. B. and P. D. McNicholas (2017b). A mixture of matrix variate skew-t distributions. arXiv preprint arXiv:1703.08882.

Harrar, S. W. and A. K. Gupta (2008). On matrix variate skew-normal distributions. *Statistics* 42(2), 179–194.

Karlis, D. and A. Santourian (2009). Model-based clustering with non-elliptically contoured distributions. *Statistics and Computing* 19(1), 73–83.

Lindsay, B. G. (1995). Mixture models: Theory, geometry and applications. In *NSF-CBMS Regional Conference Series in Probability and Statistics*, Volume 5. California: Institute of Mathematical Statistics: Hayward.

McNeil, A. J., R. Frey, and P. Embrechts (2005). *Quantitative Risk Management: Concepts, Techniques and Tools*. Princeton University Press.

McNicholas, P. D. (2016). *Mixture Model-Based Classification*. Boca Raton: Chapman & Hall/CRC Press.

McNicholas, P. D., T. B. Murphy, A. F. McDaid, and D. Frost (2010). Serial and parallel implementations of model-based clustering via parsimonious Gaussian mixture models. *Computational Statistics and Data Analysis* 54(3), 711–723.

McNicholas, S. M., P. D. McNicholas, and R. P. Browne (2017). A mixture of variance-gamma factor analyzers. In S. E. Ahmed (Ed.), *Big and Complex Data Analysis, Contributions to Statistics*, pp. 369–385. Cham: Springer International Publishing.

Meng, X.-L. and D. B. Rubin (1993). Maximum likelihood estimation via the ECM algorithm: a general framework. *Biometrika* 80, 267–278.
O’Hagan, A., T. B. Murphy, I. C. Gormley, P. D. McNicholas, and D. Karlis (2016). Clustering with the multivariate normal inverse Gaussian distribution. *Computational Statistics and Data Analysis 93*, 18–30.

Thabane, L. and M. Safiul Haq (2004). On the matrix-variate generalized hyperbolic distribution and its bayesian applications. *Statistics 38*(6), 511–526.

Wishart, J. (1928). The generalised product moment distribution in samples from a normal multivariate population. *Biometrika*, 32–52.
A Figures

Figure 1: Marginal distributions for the matrix variate GH, VG and NIG distributions for (a) V1, (b) V2, (c) V3 and (d) V4. The marginal location is by a red dashed line.
Figure 2: Marginal distributions for the matrix variate GH, VG and NIG distributions for (a) V1, (b) V2, (c) V3 and (d) V4. The marginal location is by a red dashed line.