Journal of Multivariate Analysis 167 (2018) 31–48

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

An expectation–maximization algorithm for the matrix normal distribution with an application in remote sensing

Hunter Glanz [a,*], Luis Carvalho [b]

[a] Department of Statistics, California Polytechnic State University, San Luis Obispo, CA 93407, USA
[b] Department of Mathematics and Statistics, Boston University, Boston, MA 02215, USA

ARTICLE INFO

Article history:
Received 23 January 2017
Available online 17 April 2018

AMS subject classifications:
62

Keywords:
Kronecker covariance structure
Missing data imputation

ABSTRACT

Dramatic increases in the size and dimensionality of many modern datasets make crucial the need for sophisticated methods that can exploit inherent structure and handle missing values. In this article we derive an expectation–maximization (EM) algorithm for the matrix normal distribution, a distribution well-suited for naturally structured data such as spatio-temporal data. We review previously established maximum likelihood matrix normal estimates, and then consider the situation involving missing data. We apply our EM method in a simulation study exploring errors across different dimensions and proportions of missing data. We compare these errors to those from three alternative methods and show that our proposed EM method outperforms them in all scenarios. Finally, we implement the proposed EM method in a novel way on a satellite image dataset to investigate land-cover classification separability.

1. Introduction

Technological advances in recent decades have ushered in a new era in data management and analysis. The dimension of datasets continues to grow alongside the number of observations. Consequently, the estimation of parameters or characteristics of these data remains a significant challenge. Specifically, the covariance matrix of such high dimensional data can be extremely difficult to estimate and handle. An increasingly common simplification is the assumption that this covariance has a Kronecker product structure.

A Kronecker structured covariance involves fewer parameters than a completely unstructured covariance, thus requiring less data to fit the model. This simpler structure also leads to more computationally efficient parameter estimation. Not only does this structure ease the estimation procedure, but it also naturally fits many situations in a more physical way. Multivariate repeated measurement data, for example, provide this type of framework [2,20,23,24]. That is, the response variables and time are two separate *dimensions* of the data that can be characterized independently. Thus, a straightforward way to model the full covariance is via the Kronecker product of a covariance for the responses and a covariance for time. Similarly, multivariate time series of other types such as longitudinal data [3,10] and spatio-temporal data [9,12,27] lend themselves to this kind of covariance decomposition.

The matrix normal distribution provides a natural synthesis of the ideas surrounding data with a Kronecker covariance. Specifically, a matrix normal random variate (X) is a matrix itself. The distribution is parameterized by a mean matrix and two covariance matrices representing the covariance of the rows and columns of X, respectively, denoted by

$$X \sim \mathrm{MN}(\mu, \Sigma_s, \Sigma_c), \tag{1}$$

* Corresponding author.
E-mail addresses: hglanz@calpoly.edu (H. Glanz), lecarval@math.bu.edu (L. Carvalho).

https://doi.org/10.1016/j.jmva.2018.03.010
0047-259X/

where X and μ are $p \times q$ matrices, Σ_s is a $p \times p$ matrix and Σ_c is a $q \times q$ matrix [5,7,28,29]. The matrix normal distribution is synonymous with the Kronecker covariance structure since if X has the matrix normal distribution in (1), then

$$\text{vec}(X) \sim \mathcal{N}[\text{vec}(\mu),\, \Sigma_c \otimes \Sigma_s],$$

where $\Sigma_c \otimes \Sigma_s$ denotes the Kronecker product of Σ_c and Σ_s and $\text{vec}(X_i)$ denotes the vectorization of X_i. When compared to an unconstrained normal model with a full covariance matrix, $\text{vec}(X) \sim \mathcal{N}[\text{vec}(\mu),\, \Sigma]$, the matrix normal model, when suitable, offers higher statistical and computational efficiency since its more parsimonious structure requires fewer data points to fit at the similar estimation error levels and lower computational complexity due to fewer parameters.

However, while maximum likelihood estimates have been derived along with tests of whether a covariance matrix has this structure [7], the issue of missing data in this context has not been fully addressed. Work by Allen and Tibshirani [1] treats a specific variation of this problem as an application of variance selection. Key differences between their work and what is presented here include constraints they impose on the mean, μ, and the number of observations used to estimate the parameters. Additionally, work in [17] attempts to estimate a Kronecker structure covariance in the presence of missing values, but focuses on estimation of the overall covariance instead of the two components.

We assume, for simplicity, that the data is missing completely at random [25] and derive an expectation–maximization (EM) algorithm [6] for estimating the parameters of the matrix normal distribution.

We discuss our proposed EM algorithm at Section 2.2. With expressions for the estimates in hand, we conduct a simulation study in Section 3.1 and apply the method to classify land cover using PCA compressed remotely sensed data in Section 3.2. Finally, we summarize the discussion in Section 4.

2. Model and methods

We begin with N independent observations from a matrix normal distribution,

$$X_1, \ldots, X_N \overset{\text{i.i.d.}}{\sim} \text{MN}(\mu,\, \Sigma_s,\, \Sigma_c),$$

as in (1). The use of this structure reduces the number of parameters by explicitly describing the covariance between the rows and the covariance between the columns as opposed to an individual covariance in each cell of the upper triangle of the full, $pq \times pq$, covariance matrix. Besides this simplification, the partitioning of the covariance follows naturally from a setup involving two *physical*, or *separable*, dimensions such as space and time.

2.1. Parameter estimation with complete data

It is straightforward to estimate the parameters of a matrix normal distribution using maximum likelihood when there is no missing data. If $X_i \overset{\text{i.i.d.}}{\sim} \text{MN}(\mu,\, \Sigma_s,\, \Sigma_c)$ then, equivalently,

$$\text{vec}(X_i) \overset{\text{i.i.d.}}{\sim} \mathcal{N}[\text{vec}(\mu),\, \Sigma_c \otimes \Sigma_s]$$

and so

$$\mathbb{P}(X_i;\, \mu,\, \Sigma_c,\, \Sigma_s) = (2\pi)^{-pq/2} |(\Sigma_c \otimes \Sigma_s)^{-1}|^{1/2} \exp\left\{ -\frac{1}{2} \text{vec}(X_i - \mu)^\top (\Sigma_c \otimes \Sigma_s)^{-1} \text{vec}(X_i - \mu) \right\}.$$

As usual in multivariate normal densities, in the exponential term we have the Mahalanobis distance $D_\Sigma(X_i, \mu)$, with $\Sigma = \Sigma_c \otimes \Sigma_s$, between X_i and μ,

$$D_\Sigma(X_i, \mu) = \text{vec}(X_i - \mu)^\top (\Sigma_c \otimes \Sigma_s)^{-1} \text{vec}(X_i - \mu).$$

This distance can be worked through known identities of the vec operator and Kronecker product to yield a simpler expression for the matrix normal density, viz.

$$\begin{aligned} D_\Sigma(X_i, \mu) &= \text{vec}(X_i - \mu)^\top (\Sigma_c \otimes \Sigma_s)^{-1} \text{vec}(X_i - \mu) \overset{\text{(i)}}{=} \text{vec}(X_i - \mu)^\top (\Sigma_c^{-1} \otimes \Sigma_s^{-1}) \text{vec}(X_i - \mu) \\ &\overset{\text{(ii)}}{=} \text{vec}(X_i - \mu)^\top \text{vec}\{\Sigma_s^{-1} (X_i - \mu) \Sigma_c^{-1}\} \overset{\text{(iii)}}{=} \text{tr}\{(X_i - \mu)^\top \Sigma_s^{-1} (X_i - \mu) \Sigma_c^{-1}\} \\ &= \text{tr}\{\Sigma_s^{-1} (X_i - \mu) \Sigma_c^{-1} (X_i - \mu)^\top\}, \end{aligned}$$

where (i), (ii), and (iii) are applications of identities (488), (496), and (497) in [21], respectively. Furthermore, since $|(\Sigma_c \otimes \Sigma_s)^{-1}| = |\Sigma_c^{-1}|^p |\Sigma_s^{-1}|^q$ by identity (492) in [21], we recover the characterization of Srivastava and Khatri [28]: $X_i \sim \text{MN}(\mu,\, \Sigma_s,\, \Sigma_c)$ if and only if the density of X_i is given by

$$\mathbb{P}(X_i;\, \mu,\, \Sigma_c,\, \Sigma_s) = (2\pi)^{-pq2} |\Sigma_s|^{-q/2} |\Sigma_c|^{-p/2} \exp\left[-\frac{1}{2} \text{tr}\left\{ \Sigma_s^{-1} (X_i - \mu) \Sigma_c^{-1} (X_i - \mu)^\top \right\} \right]. \tag{2}$$

The log likelihood of the parameters $\Theta = (\mu, \Sigma_c, \Sigma_s)$ is then

$$\ln \mathbb{P}(X; \Theta) = \sum_i \ln \mathbb{P}(X_i; \Theta) = \frac{pN}{2} \ln|\Sigma_c^{-1}| + \frac{qN}{2} \ln|\Sigma_s^{-1}| - \frac{1}{2} \sum_i D_\Sigma(X_i, \mu) \tag{3}$$

up to a normalizing constant. The form in (2) simplifies the matrix derivatives of (3) considerably leaving us with the following maximum likelihood estimates (MLEs):

$$\widehat{\mu} = \frac{1}{N} \sum_{i=1}^{N} X_i, \quad \widehat{\Sigma}_c = \frac{1}{pN} \sum_{i=1}^{N} (X_i - \widehat{\mu})^\top \widehat{\Sigma}_s^{-1} (X_i - \widehat{\mu}), \quad \widehat{\Sigma}_s = \frac{1}{qN} \sum_{i=1}^{N} (X_i - \widehat{\mu}) \widehat{\Sigma}_c^{-1} (X_i - \widehat{\mu})^\top. \tag{4}$$

The two covariance estimates depend on each other and thus their estimates must be computed in an iterative fashion until convergence. However, we note that since the negative log-likelihood is not convex in all its parameters this method might only converge to a local minimum; conditions for existence and uniqueness of these estimators are discussed in [22].

Handling non-identifiability

If for any $\kappa \neq 0$ we define $\tilde{\Sigma}_c = \kappa \Sigma_c$ and $\tilde{\Sigma}_s = \Sigma_s / \kappa$ then $\tilde{\Sigma}_c \otimes \tilde{\Sigma}_s = \Sigma_c \otimes \Sigma_s$, and so both estimates yield the same overall covariance matrix. To resolve this non-identifiability issue we propose the following amendment to the model:

$$\text{vec}(X_i) \sim \mathcal{N}[\text{vec}(\mu), \sigma^2 \Sigma_c \otimes \Sigma_s] \tag{5}$$

and require that $(\Sigma_c)_{11} = 1$ and $(\Sigma_s)_{11} = 1$ (the choice of the top-left entry is arbitrary). In this way we fix the scale of Σ_c and Σ_s, and estimate the scale of the overall covariance in σ^2.

The new log-likelihood of parameters $\Theta = (\mu, \sigma^2, \Sigma_c, \Sigma_s)$ is similar to (3),

$$\ln \mathbb{P}(X; \Theta) = \frac{pN}{2} \ln|\Sigma_c^{-1}| + \frac{qN}{2} \ln|\Sigma_s^{-1}| - \frac{pqN}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{N} D_\Sigma(X_i, \mu), \tag{6}$$

which yields the MLE of σ^2 as

$$\widehat{\sigma}^2 = \frac{1}{pqN} \sum_{i=1}^{N} (X_i - \widehat{\mu})^\top (\widehat{\Sigma}_c \otimes \widehat{\Sigma}_s)^{-1} (X_i - \widehat{\mu})$$

depending on the other estimates. The MLE for μ is clearly the same as in (4) since we only changed the variance of the model. However, since the variance scale is now captured by σ^2 we need to account for the scale constraints on Σ_c and Σ_s when deriving their MLEs. In this regard, we exploit the following result.

Theorem 1. *For $n > 0$, $\sigma^2 > 0$ and a symmetric positive definite matrix S, $S \succ 0$, procedure* ADJUST(n, σ^2, S) *obtains*

$$\underset{\Sigma \succ 0, \Sigma_{11}=1}{\arg\max} \frac{n}{2} \ln|\Sigma^{-1}| - \frac{1}{2\sigma^2} tr(\Sigma^{-1} S).$$

An outline of ADJUST and a proof of this result can be found in Appendices A–C. Now we just need to apply Theorem 1 to (6) as a function of Σ_c and Σ_s in turn to retrieve their respective MLEs:

$$\widehat{\Sigma}_c = \text{ADJUST}\left\{ pN, \widehat{\sigma}^2, \sum_{i=1}^{N} (X_i - \widehat{\mu})^\top \widehat{\Sigma}_s^{-1} (X_i - \widehat{\mu}) \right\}, \quad \widehat{\Sigma}_s = \text{ADJUST}\left\{ qN, \widehat{\sigma}^2, \sum_{i=1}^{N} (X_i - \widehat{\mu}) \widehat{\Sigma}_c^{-1} (X_i - \widehat{\mu})^\top \right\}.$$

Finally, we remark that, according to Theorem 3.1 in [29], we need $N > \max(p, q)$ for the maximum likelihood estimates to be unique.

2.2. Parameter estimation with missing data

Missing data presents a difficult, albeit well-studied challenge in parameter estimation. Traditional methods, such as the EM algorithm, can usually handle missing data in a straightforward way. However, as the dimensionality increases, as in our case, the method can become quite computationally expensive. Naturally, we aim to assess different ways of achieving accurate parameter estimates with an eye towards reducing computation time.

The first approach (which we label "MM") applies a maximization in two ways, being similar to cyclic gradient ascent: (1) "imputation" of missing values by maximum likelihood estimation conditional on parameters; and (2) maximum likelihood parameter estimation conditional on imputed missing values. In particular, the missing values get replaced by the most recent estimate of the mean. The next iteration of mean and covariance estimates come from the same maximum likelihood expressions in (4), with the addition of $\widehat{\sigma}^2$, based on the fully imputed data. The ease and simplicity of this method make it a natural first step in handling missing data, but also hinder its robustness and ability to capture all of the uncertainty associated with missing data.

The second approach ("GMM") applies the MM method to the most general version of the model. As opposed to estimating the parameters of (1), the GMM algorithm provides parameter estimates for the following model:

$$\mathrm{vec}(X_i) \sim \mathcal{N}[\mathrm{vec}(\mu),\ \Sigma], \tag{7}$$

where Σ is an unstructured covariance matrix of order pq. The third approach ("GEM") applies the EM algorithm to the model in (7). These multivariate normal EM estimates (from GEM) have the same form of those found in [19]. The GMM and GEM approaches do not simplify the original problem since they require more parameters to be estimated by not assuming the Kronecker structure. The simple form of (7) attracts much attention since its log-likelihood is convex and thus guarantees a global optimum for maximum likelihood estimation, but its complexity far exceeds that of (1). Where sources of variation in the data can be naturally partitioned, such as in space or time, the Kronecker structure surpasses (7) in both interpretability and statistical efficiency. The remainder of this article introduces an EM procedure for (1) and its superiority in situations involving these types of structured data.

EM algorithm for matrix normal distribution

In the situation where missing data exists the EM algorithm is a convenient way to estimate the parameters $\Theta = (\mu, \sigma^2, \Sigma_c, \Sigma_s)$ in (1). The rest of this section details the fourth approach ("EM") to parameter estimation with missing data. Let us denote $X = (Y, Z)$ where Z is the missing portion of X and Y is the observed portion of X. For the *E-step*, we need:

$$\begin{aligned} Q(\Theta;\Theta^{(t)}) &= \mathbb{E}_{Z|Y;\Theta^{(t)}}\{\ln \mathbb{P}(X_1,\ldots,X_n;\Theta)\} \\ &= \frac{pN}{2}\ln|\Sigma_c^{-1}| + \frac{qN}{2}\ln|\Sigma_s^{-1}| - \frac{pqN}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_i \mathbb{E}_{Z|Y;\Theta^{(t)}}\{D_\Sigma(X_i,\mu)\}, \end{aligned}$$

while the *M-step* updates Θ by maximizing Q, $\Theta^{(t+1)} = \arg\max_\Theta Q(\Theta;\Theta^{(t)})$, via matrix differentiation in our case.

The Mahalanobis distance obeys a Pythagorean relationship: if $\tilde{\mu}_i^{(t)} = \mathbb{E}_{Z|Y;\Theta^{(t)}}(X_i)$, then

$$\mathbb{E}_{Z|Y;\Theta^{(t)}}\{D_\Sigma(X_i,\mu)\} = \mathbb{E}_{Z|Y;\Theta^{(t)}}\{D_\Sigma(X_i,\mu_i^{(t)})\} + D_\Sigma(\mu_i^{(t)},\mu).$$

From here the update for μ follows from $\partial Q/\partial\mu = 0$, viz.

$$\widehat{\mu}^{(t+1)} = \frac{1}{N}\sum_{i=1}^{N}\tilde{\mu}_i^{(t)},$$

similarly to the plain MLE case in (4).

Updating Σ_c, Σ_s and σ^2 requires a bit more work. To this end we focus, first, on the following term:

$$\begin{aligned} R_i(\Theta;\Theta^{(t)}) &= \mathbb{E}_{Z|Y;\Theta^{(t)}}\{D_\Sigma(X_i,\mu_i^{(t)})\} \\ &= \mathbb{E}\Big[\mathrm{tr}\{(\Sigma_c\otimes\Sigma_s)^{-1}\mathrm{vec}(X_i-\tilde{\mu}_i^{(t)})\mathrm{vec}(X_i-\tilde{\mu}_i^{(t)})^\top\}\Big] \\ &= \mathrm{tr}\Big[(\Sigma_c^{-1}\otimes\Sigma_s^{-1})\underbrace{\mathbb{E}\{\mathrm{vec}(X_i-\tilde{\mu}_i^{(t)})\mathrm{vec}(X_i-\tilde{\mu}_i^{(t)})^\top\}}_{V_i^{(t)}}\Big], \end{aligned}$$

where we define the expected outer product

$$V_i^{(t)} = \mathbb{E}_{Z|Y;\Theta^{(t)}}\{\mathrm{vec}(X_i-\tilde{\mu}_i^{(t)})\mathrm{vec}(X_i-\tilde{\mu}_i^{(t)})^\top\}.$$

To get the partial derivatives of Q with respect to Σ_c we need

$$\frac{\partial R_i}{\partial(\Sigma_c^{-1})_{k\ell}} = \mathrm{tr}\left[\frac{\partial}{\partial(\Sigma_c^{-1})_{k\ell}}\{(\Sigma_c^{-1}\otimes\Sigma_s^{-1})V_i\}\right] = \mathrm{tr}\left[\left\{\underbrace{\frac{\partial\Sigma_c^{-1}}{\partial(\Sigma_c^{-1})_{k\ell}}}_{S_{k\ell}}\otimes\Sigma_s^{-1}\right\}V_i\right],$$

where $S_{k\ell}$ is the structure matrix [21] of a symmetric matrix, that is, $S_{k\ell} = e_k e_\ell^\top + e_\ell e_k^\top$. Thus, $(S_{k\ell}\otimes\Sigma_s^{-1})$ is a block matrix.

We can now look at V_i as a $q\times q$ block matrix where each element is a $p\times p$ matrix in the following way, e.g.,

$$V_{i,k\ell} = \begin{bmatrix} \circ & \circ & \circ & \circ & \circ & \circ & \circ \\ \circ & \bullet & \circ & \bullet & \circ & \bullet & \circ \\ \circ & \circ & \circ & \circ & \circ & \circ & \circ \\ \circ & \bullet & \circ & \bullet & \circ & \bullet & \circ \\ \circ & \circ & \circ & \circ & \circ & \circ & \circ \end{bmatrix} \begin{matrix} \\ \\ b \\ \\ \\ \\ \end{matrix} \quad b' \ . \tag{8}$$

So the matrix V_i being a block matrix leads to $V_{i,k\ell}$ being a symmetric $p \times p$ matrix with zeros at the empty circles in (8) and $\text{cov}_{Z_i|Y_i}(Z_{kb}, Z_{\ell b'})$ at the filled circles. Thus,

$$\frac{\partial R_i}{\partial(\Sigma_c^{-1})_{k\ell}} = \text{tr}[\Sigma_s^{-1}(V_{i,k\ell} + V_{i,k\ell}^\top)] = \sum_{b,b' \in \text{miss}(k,\ell)} (\Sigma_s^{-1})_{b,b'} \text{cov}_{Z_i|Y_i,\Theta^{(t)}}[Z_{kb}, Z_{\ell b'}].$$

Here $\text{miss}(k, \ell)$ are the row-column pairs for which there are missing entries in V_i, and $V_{i,k\ell}$ is the $p \times p$ block submatrix of V_i from rows $(k-1)p+1$ to kp and columns $(\ell-1)p+1$ to ℓp. Note that $\partial R_i/\partial(\Sigma_c^{-1})_{k,\ell}$ does not depend on Z_i. Moreover, the conditional covariances in $\text{var}_{Z_i|Y_i,\Theta^{(t)}}(Z_i)$ above can be obtained by applying the SWEEP operator [11] to the rows of $\Sigma_c^{-1} \otimes \Sigma_s^{-1}/\sigma^2$ that correspond to missing values to isolate the Schur complement.

Thus, from solving $\partial Q/\partial(\Sigma_c^{-1}) = 0$ with $(\Sigma_c)_{11} = 1$, we have

$$\widehat{\Sigma}_c^{(t+1)} = \text{ADJUST}\left[pN, \widehat{\sigma}^{2(t)}, \sum_i \left\{ \frac{\partial R_i}{\partial \Sigma_c^{-1}} + (\tilde{\mu}_i^{(t)} - \mu^{(t)})^\top \Sigma_s^{-1^{(t)}} (\tilde{\mu}_i^{(t)} - \mu^{(t)}) \right\}\right]. \tag{9}$$

Similarly, for $\widehat{\Sigma}_s$,

$$\widehat{\Sigma}_s^{(t+1)} = \text{ADJUST}\left[qN, \widehat{\sigma}^{2(t)}, \sum_i \left\{ \frac{\partial R_i}{\partial \Sigma_s^{-1}} + (\tilde{\mu}_i^{(t)} - \mu^{(t)}) \Sigma_c^{-1^{(t)}} (\tilde{\mu}_i^{(t)} - \mu^{(t)})^\top \right\}\right]. \tag{10}$$

Finally, for σ^2,

$$\widehat{\sigma}^{2(t+1)} = \frac{1}{pqN} \sum_i R_i + \text{vec}(\tilde{\mu}_i^{(t)} - \widehat{\mu}^{(t)})^\top (\widehat{\Sigma}_c^{(t)} \otimes \widehat{\Sigma}_s^{(t)})^{-1} \text{vec}(\tilde{\mu}_i^{(t)} - \widehat{\mu}^{(t)}). \tag{11}$$

Detailed information regarding the implementation of this EM method can be found in Appendix A, with R package code for all four methods, MM, GMM, EM, and GEM, included in the supplementary material. Interestingly, while the four methods have similar computational complexity per update, roughly $O(Np^2q^2)$, our experiments show that EM and GEM require more iterations to achieve convergence and so longer running times in practice.

3. Case studies

3.1. Simulation study

To empirically assess the model and algorithm we simulated data from a matrix normal distribution with randomly chosen parameters of dimensions (p, q): $(3, 5)$, $(3, 7)$, and $(10, 25)$. Since we need to estimate $pq(pq+1)/2$ parameters in the upper triangle of the covariance matrix under the GMM and GEM models – as opposed to only $p(p+1)/2+q(q+1)/2$ under the MM and EM models – we are constrained to modest dimension sizes in our simulation study due to the computational burden. To this end, the sample sizes used for $(3, 5)$ and $(3, 7)$ were 500 and 1000 and all four methods were applied. For the $(10, 25)$ setting, sample sizes of 1000 and 2000 were used and only the MM and EM methods were applied. Four different proportions of missing data were used: 10%, 25%, 50% and 75%. Data were simulated 100 times at each combination of sample size and proportion of missing data to evaluate how the accuracy of the estimates vary. The four different algorithms described in Section 2.2 were run in each of these combinations to provide a richer comparison. In order to compare these methods, the relative errors in the covariance estimates were always measured with respect to the full (Kronecker product) covariance matrix.

The relative errors of the mean estimates across the four methods and the four different proportions of missing data were consistently low. The methods differ very little when it comes to the estimate of the mean, and so we focus on the variance estimates. Indeed, the models and estimation procedure vary most when dealing with the covariance matrix, as expected.

Fig. 1 tells a rich story about how these four methods differ most. As the sample size increases the estimates appear to improve slightly. With a true underlying Kronecker structure, it is perhaps no surprise that the EM and MM methods tend to outperform the general methods. Most notably, our proposed EM method (denoted *EM*) outperforms the MM method in every situation and sometimes by large margins. Fig. 2 continues the story with results from the $(10, 25)$ setting. We see similar results here though, with our proposed EM method outperforming the MM method at every proportion of missing data.

If we instead simulate a general covariance matrix, with no particular structure, then the methods perform quite differently. Fig. 3 describes complex stories for this setting. The MM and EM methods tend to improve, relative to the other methods, as the proportion of missing data increases only clearly beating GMM and GEM in the 75% missing data situation. For most of the settings in Fig. 3 the MM and EM methods perform comparably. In the bottom graph of Fig. 3 we see a slightly opposite result from Fig. 2, with the MM method marginally outperforming the EM method. For additional metrics comparing these methods with the simulation study, including run times, see Appendix C.

Of course, this presumes the choice between the two models. In a situation where the physical dimensions of the data imply a Kronecker structure, we can take comfort in the above results. One such example is the following application to Remote Sensing.

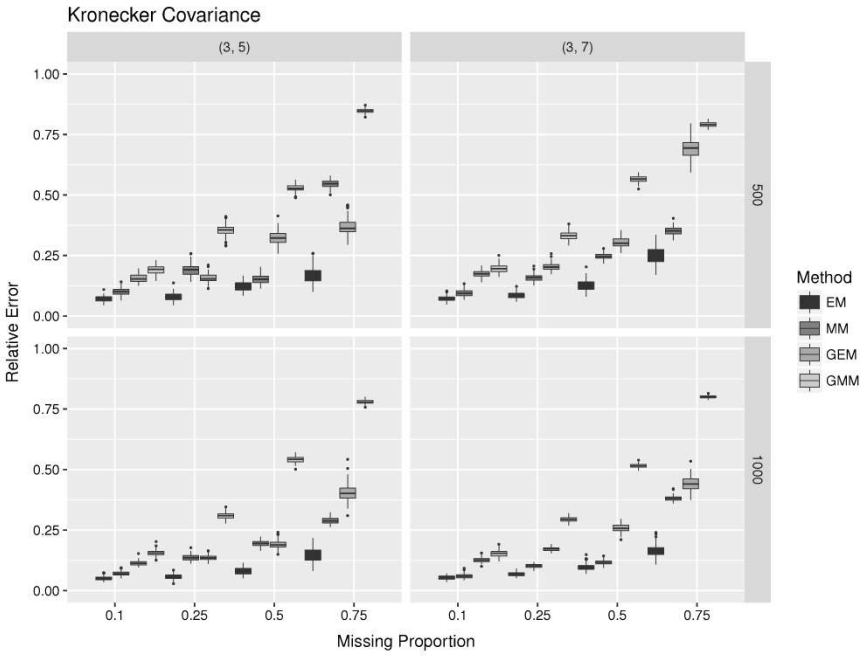

Fig. 1. Boxplots of the relative errors in the estimates of Kronecker Σ, in order of EM, MM, GEM and GMM for dimensions: $(3, 5)$ and $(3, 7)$.

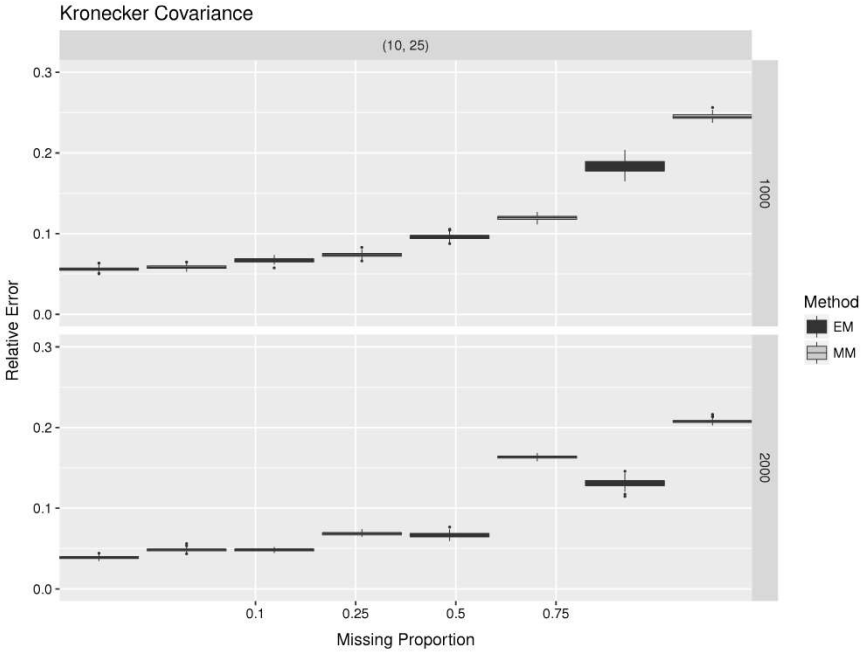

Fig. 2. Boxplots of the relative errors in the estimates of Kronecker Σ, in order of EM and MM for dimensions: $(10, 25)$.

3.2. Land cover classification using MODIS satellite image data

One of the biggest tasks in Remote Sensing is land-cover classification. In other words, taking remotely sensed images composed of millions of pixels and assigning to each pixel a particular land cover class. Many of the richest datasets are both multispectral and multitemporal. For satellite image data from the Moderate Resolution Imaging Spectroradiometer

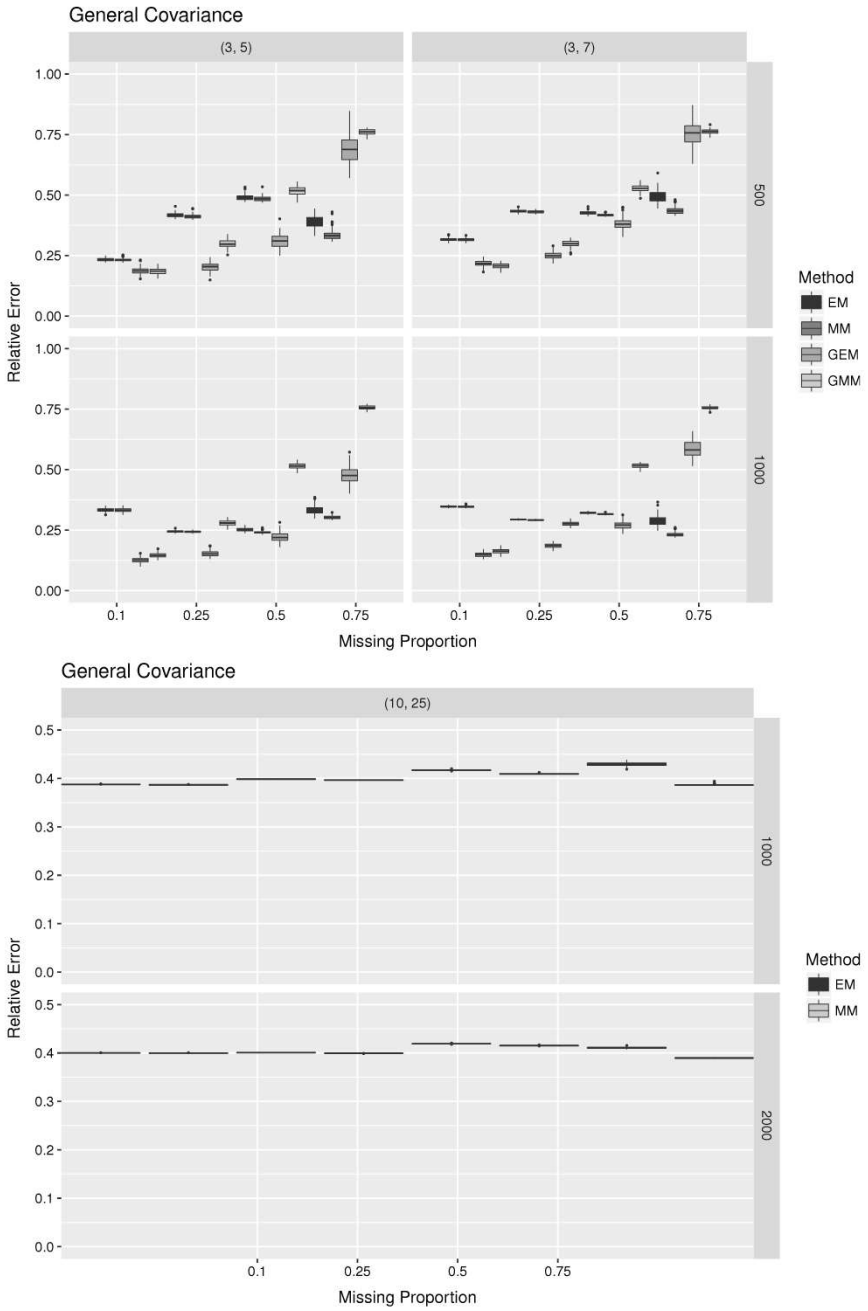

Fig. 3. Boxplots of the relative errors in the estimates of general Σ, in order of EM, MM, GEM and GMM for dimensions: (3, 5), (3, 7) and (10, 25).

(MODIS) sensor aboard the AQUA and TERRA satellite platforms [8] we observe data in 7 different spectral bands at each of 46 time points (comprising a multivariate annual profile). Fig. 4 gives an example of one such MODIS satellite image of part of North America plotted in true color with most of the water masked out (hence the black) for the purposes of land cover classification. Each pixel in this image is 1km x 1km in size and yields the seven spectral bands over 46 time points of data mentioned above. With this, we aim to assign a land cover classification (label) to it.

The analysis we present here is an application of the proposed EM algorithm to these multivariate time series data. We adopt the following model:

$$\text{vec}(X_v) \mid \theta_v = c \sim \mathcal{N}[\text{vec}(\mu_c), \sigma_c^2 \Sigma_c \otimes \Sigma_s], \tag{12}$$

Fig. 4. Example of satellite imagery of part of North America with water mostly masked out. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

where X_v and θ_v are the data and land cover class for pixel v, respectively. The data has two feature dimensions – spectral bands and time – and it is traditionally agreed in the field (see, e.g., the ubiquitous application of the tasseled cap transformation [15] to characterize "greenness", as in [31]) that spectral variation does not change with time and with land cover class since it is essentially due to device measurement error. In this case study we aim to formalize, using the Kronecker structure, this notion of covariance separability. Therefore, the Kronecker structure naturally models the covariance in the spectral bands (Σ_s) and the covariance in time for each class (Σ_c). That is, the full covariance is partitioned into spectral and temporal components. Note that in (12) we allow for a different mean, temporal covariance and scale (σ^2) for each land cover class. This is an adaptation of the EM algorithm we have proposed for the model in (5). The only changes to the estimating of our model parameters, mentioned in (9), (10), and (11), involve using just the data relevant to each particular parameter. That is, the mean for land cover class c gets estimated using only data from class c. Likewise for each Σ_c and each σ_c^2. The estimate of the common Σ_s makes use of all of the Σ_c estimates with their corresponding observations.

Winter months increase the occurrence of missing data due to snow and cloud occlusion, and so to keep a missing data mechanism closer to random we focus our attention on the middle of the year (the middle 28 time points). Thus X and μ_c are 7×28 matrices, Σ_c is a 28×28 matrix and Σ_s is a 7×7 matrix. For the analysis that follows we used a subset of the MODIS Land Cover Training site database that includes 204 sites located over the conterminous United States. These sites include 2,733 MODIS pixels and encompass most major biomes and land cover types in the lower 48 United States [26]. There are 12 land cover classes well represented in the training dataset, out of 17 original IGBP [18] classes.

Land cover classification has been approached in many different ways for quite some time. The high dimensionality of the data presents an increasingly significant challenge. Traditionally, principal components analysis (PCA) has been used to reduce the dimensionality of the data. However, usually all 196 (7×28) features are considered distinct features. Consequently, spectral and temporal variation cannot be isolated from an eigen-decomposition of the full (196×196) covariance matrix, as required from PCA.

To preserve the raw temporal information and still reduce the dimensionality of the data, we propose targeting Σ_s with the principal components analysis as opposed to $\Sigma_c \otimes \Sigma_s$. About 5% of the data is missing, making the Kronecker structure and our proposed EM algorithm even more ideal, from a purely computational standpoint, as evidenced by the simulation study above. Using the proposed EM algorithm we estimate the parameters of (12) for every land cover class and perform PCA on our estimate of Σ_s, the spectral covariance common to all land cover classes. To, again provide a comparison which confirms the quality of our *EM*, we compare the PCA results just described to those computed using the *MM* method.

Fig. 5 shows the data projected into the space of the first two principal components using the *MM* and *EM* methods. The added benefit of PCA applied to the spectral covariance term is that we recover well-established interpretations known as the tasseled-cap transformation [31]. This result identified the first three principal components as having physical meanings: brightness, greenness, wetness. The class separability seems reasonable and the different amounts of variation in each class are certainly identified. A closer inspection reveals that the *EM* method achieves better separability of the land-cover classes.

In particular, using the *EM*, the first two principal components capture 76.1% of the variation in the spectral bands. For the purposes of land-cover classification we chose to use the first *three* principal components since they capture 92.3% of the variation in the spectral bands.

In addition to visualizing the principal components, we performed a simple *10-fold* cross validation procedure to compare classification accuracies. For each fold, the parameters were estimated using the other 90% of the data and then used to classify the pixels in the withheld fold according to the MAP (*maximum a posteriori*) estimator.

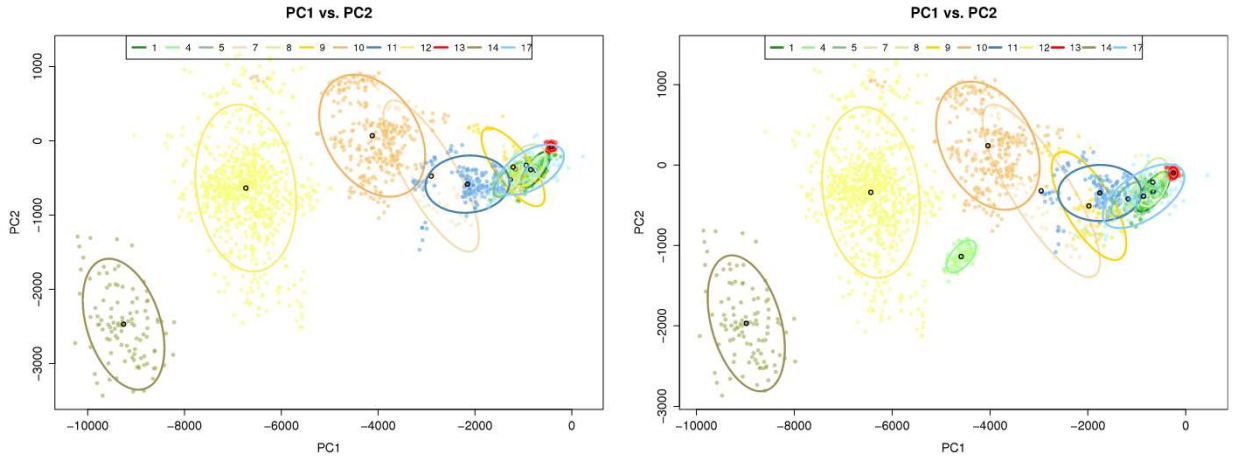

Fig. 5. The data projected into the space of the first two principal components, colored by their respective land color classes. Left: MM; Right: EM. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

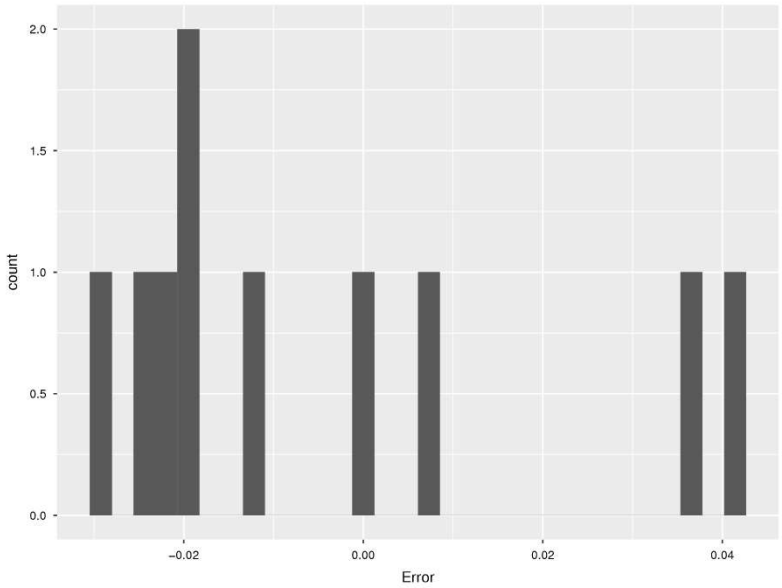

Fig. 6. Histogram of differences in overall classification errors between the EM and MM methods. Differences are *EM–MM*.

Fig. 6 displays the difference in overall miss-classification rate between the EM and MM methods, for each fold. There were only three folds where the MM method outperformed the EM method (indicated by the positive values in Fig. 6. Across all folds, the EM method's average miss-classification rate was 0.152 and the MM method's was 0.156. While this difference may seem small, it only adds support to the usefulness of our proposed EM method, whose advantages began during the simulation study.

4. Conclusion

The matrix normal distribution is a natural candidate for situations involving some sort of structure or separability in the dimensions of the data. In this article we derived an expectation–maximization algorithm for estimating the parameters of the matrix normal distribution in the presence of missing data.

A simulation study exploring different sample sizes and proportions of missing data showed the usefulness of this method when compared to a full, unconstrained multivariate normal distribution. An example of this type of scenario, in the field of Remote Sensing, produced physically useful and interpretable results. This case study also highlighted the flexibility of the proposed method in estimating potentially varying pieces of a Kronecker structured covariance. Many missing observations in this application show a seasonal pattern with nonignorable missing data concentrated on winter months. One direct

solution would be to follow the approach proposed by [13] and explicitly assume that missing data occurrences follow an AR(1) model and then use a Monte Carlo EM method [30] based on a Gibbs sampler to fit the matrix normal parameters for each pixel. Another approach is to model spatial associations between parameters of neighboring pixels, especially the means [4], to borrow information across pixels and improve robustness of the EM method. We also note that the strategy proposed by [13] (see also [14] for a review) is suitable for non-normal sampling distributions belonging to an exponential family and broader models with random effects.

As data becomes more abundant and higher in dimension the challenge of extracting information continues to grow in difficulty and importance. The Kronecker covariance structure can provide both a richer physical interpretation of the parameters as well as help the estimating procedure. Now, even with missing data, accurate estimates of these parameters are obtainable.

Appendix A. Estimation algorithms for matrix normal parameters

Here we give more detailed implementations of the methods discussed in the text. In what follows, $\mathsf{miss}(i)$ contains the indices for the missing entries (if any) in the ith observation, while $m_s(i)$ and $m_c(i)$ are the row and column indices of the entries in $\mathsf{miss}(i)$, respectively. Moreover, the operator $\mathbf{e}(\cdot)$ produces a *mask* matrix according to the corresponding row and column indices. Take the following example:

$$X_i = \begin{bmatrix} 1 & 4 & 7 & 10 & ? & 16 & ? \\ ? & 5 & 8 & 11 & 14 & 17 & 20 \\ ? & 6 & ? & 12 & 15 & ? & 21 \end{bmatrix}.$$

There are six missing values that correspond to $\mathsf{miss}(i) = [2,\ 3,\ 9,\ 13,\ 18,\ 19]$ (in column-major order), $m_s(i) = [2,\ 3,\ 3,\ 1,\ 3,\ 1]$, and $m_c(i) = [1,\ 1,\ 3,\ 5,\ 6,\ 7]$. In this case, $\mathbf{e}\{m_c(i)\}$ is 6×7 (the number of missing values by the number of columns of X) and $\mathbf{e}\{m_s(i)\}$ is 6×3 (the number of missing values by the number of rows of X):

$$\mathbf{e}\{m_c(i)\} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \quad \text{and} \quad \mathbf{e}\{m_s(i)\} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}.$$

In this way, there is an "indicator" row for each index in $m_c(i)$ and $m_s(i)$, respectively. We denote by $\circ$ the Hadamard (element-wise) product. We further denote by $A\{m_s(i)\}$ the rows of A indexed by $m_s(i)$, and by $A\{-m_s(i)\}$ the rows of A there are *not* in $m_s(i)$. A similar notation is used to subset columns based on index sets.

We start with unconstrained, general formulations in Algorithms 1 and 2 for ML and EM estimation updates in the case of a simple normal distribution. Algorithm 1 implements GMM and simply imputes missing data using the current mean, while Algorithm 2 implements GEM and takes into account the missing data uncertainty in the mean and covariance using EM conditional moments [16]. In terms of computational complexity, the dominating step in Algorithm 1 is the outer product for $\Sigma^{(t+1)}$ update in line 5 with overall complexity $O(Np^2q^2)$. The SWEEP operator requires $O(kn^2)$ when applied to k rows on a squared matrix of order n. Thus, with $m_i = |\mathsf{miss}(i)|$, Algorithm 2 costs $O(p^3q^3)$ to invert $\Sigma^{(t)}$ initially, and, for each observation i, $O(p^2q^2m_i)$ for the SWEEP operation in line 3 and $O(p^2q^2)$ for the $\Sigma^{(t+1)}$ update in line 7. Thus, the overall cost of $O(p^3q^3 + (N+M)p^2q^2)$, where $M = m_1 + \cdots + m_N$ is the total number of missing observations.

Algorithm 1: General MM update for matrix normal parameter (ML) estimation, currently at tth iteration.

1 $\mu^{(t+1)} \leftarrow 0_{pq}$; $\Sigma^{(t+1)} \leftarrow 0_{pq,pq}$;
2 **for** $i \leftarrow 1, \ldots, N$ **do** `// for each observation`
3 $\quad X_m \leftarrow Y_i$; $X_m[\mathsf{miss}(i)] \leftarrow \mu^{(t)}[\mathsf{miss}(i)]$; `// auxiliary variable`
$\quad$ `// Update estimates:`
4 $\quad \mu^{(t+1)} \leftarrow \mu^{(t+1)} + X_m$;
5 $\quad \Sigma^{(t+1)} \leftarrow \Sigma^{(t+1)} + (X_m - \mu^{(t)})(X_m - \mu^{(t)})^\top$;
6 **end**
7 $\mu^{(t+1)} \leftarrow \mu^{(t+1)}/N$; $\Sigma^{(t+1)} \leftarrow \Sigma^{(t+1)}/N$; `// scale`

Next, we provide an implementation for ML estimation ("MM" in our notation) for the matrix normal distribution in Algorithm 3 ; see also [7]. The computational complexity is $O(p^3 + q^3)$ for the initial inversions of $\Sigma_c^{(t)}$ and $\Sigma_s^{(t)}$, and, for each observation i, $O(pq^2)$ and $O(p^2q)$ for the updates of $\Sigma_c^{(t+1)}$ and $\Sigma_s^{(t+1)}$ at lines 5 and 6 respectively, and $O(p^2q^2)$ for the update of $\sigma^{2^{(t+1)}}$ at line 7. These dominating steps account for an overall time cost of $O(p^3 + q^3 + Np^2q^2)$ per MM update.

Finally, our proposed EM update is listed in Algorithm 4. Similarly to the MM update, the EM update requires $O(p^3 + q^3)$ for the initial covariance inversions but an extra $O(m_ip^2q^2)$ for each i-observation due to the SWEEP operation at line 3. The

Algorithm 2: General EM update for matrix normal parameter estimation, currently at tth iteration.

$\mu^{(t+1)} \leftarrow 0_{pq}$; $\Sigma^{(t+1)} \leftarrow 0_{pq,pq}$;
for $i \leftarrow 1, \ldots, N$ **do** // for each observation
  // Auxiliary variables: $X_m = \mathbb{E}_{Z_i|Y_i,\Theta^{(t)}}[Z_i]$, $R_m = \mathrm{var}_{Z_i|Y_i,\Theta^{(t)}}[Z_i]$.
  $R \leftarrow \Sigma^{-1^{(t)}}$; SWEEP the miss($i$) rows of R;
  $X_m \leftarrow Y_i$; $X_m[\mathrm{miss}(i)] \leftarrow \mu^{(t)}[\mathrm{miss}(i)] + R[-\mathrm{miss}(i), \mathrm{miss}(i)]^\top (Y_i - \mu^{(t)}[-\mathrm{miss}(i)])$;
  $R_m \leftarrow R[\mathrm{miss}(i), \mathrm{miss}(i)]$;
  // Update estimates:
  $\mu^{(t+1)} \leftarrow \mu^{(t+1)} + X_m$;
  $\Sigma^{(t+1)} \leftarrow \Sigma^{(t+1)} + (X_m - \mu^{(t)})(X_m - \mu^{(t)})^\top + R_m$;
end
$\mu^{(t+1)} \leftarrow \mu^{(t+1)}/N$; $\Sigma^{(t+1)} \leftarrow \Sigma^{(t+1)}/N$; // scale

Algorithm 3: MM update for matrix normal parameter (ML) estimation, currently at tth iteration.

$\mu^{(t+1)} \leftarrow 0_{p,q}$; $\Sigma_c^{(t+1)} \leftarrow 0_{q,q}$; $\Sigma_s^{(t+1)} \leftarrow 0_{p,p}$; $\sigma^{2^{(t+1)}} \leftarrow 0$;
for $i \leftarrow 1, \ldots, N$ **do** // for each observation
  $X_m \leftarrow Y_i$; $X_m[\mathrm{miss}(i)] \leftarrow \mu^{(t)}[\mathrm{miss}(i)]$; // auxiliary variable
  // Update estimates:
  $\mu^{(t+1)} \leftarrow \mu^{(t+1)} + X_m$;
  $\Sigma_c^{(t+1)} \leftarrow \Sigma_c^{(t+1)} + (X_m - \mu^{(t)})^\top \Sigma_s^{-1^{(t)}} (X_m - \mu^{(t)})$;
  $\Sigma_s^{(t+1)} \leftarrow \Sigma_s^{(t+1)} + (X_m - \mu^{(t)}) \Sigma_c^{-1^{(t)}} (X_m - \mu^{(t)})^\top$;
  $\sigma^{2^{(t+1)}} \leftarrow \sigma^{2^{(t+1)}} + \mathrm{tr}(\mathrm{vec}(X_m - \mu^{(t)})^\top (\Sigma_c^{-1^{(t)}} \otimes \Sigma_s^{-1^{(t)}}) \mathrm{vec}(X_m - \mu^{(t)}))$;
end
$\mu^{(t+1)} \leftarrow \mu^{(t+1)}/N$; $\sigma^{2^{(t+1)}} \leftarrow \sigma^{2^{(t+1)}}/(Npq)$; // scale
$\Sigma_c^{(t+1)} \leftarrow \mathrm{ADJUST}(pN, \sigma^{2^{(t)}}, \Sigma_c^{(t+1)})$;
$\Sigma_s^{(t+1)} \leftarrow \mathrm{ADJUST}(qN, \sigma^{2^{(t)}}, \Sigma_s^{(t+1)})$;

Algorithm 4: EM update for matrix normal parameter estimation, currently at tth iteration.

$\mu^{(t+1)} \leftarrow 0_{p,q}$; $\Sigma_c^{(t+1)} \leftarrow 0_{q,q}$; $\Sigma_s^{(t+1)} \leftarrow 0_{p,p}$; $\sigma^{2^{(t+1)}} \leftarrow 0$;
for $i \leftarrow 1, \ldots, N$ **do** // for each observation
  // Auxiliary variables: $X_m = \mathbb{E}_{Z_i|Y_i,\Theta^{(t)}}[Z_i]$, $R_m = \mathrm{var}_{Z_i|Y_i,\Theta^{(t)}}[Z_i]$.
  $R \leftarrow \Sigma_c^{-1^{(t)}} \otimes \Sigma_s^{-1^{(t)}}/\sigma^{2^{(t)}}$; SWEEP the miss($i$) rows of R;
  $X_m \leftarrow Y_i$; $X_m[\mathrm{miss}(i)] \leftarrow \mu^{(t)}[\mathrm{miss}(i)] + R[-\mathrm{miss}(i), \mathrm{miss}(i)]^\top (Y_i - \mu^{(t)}[-\mathrm{miss}(i)])$;
  $R_m \leftarrow R[\mathrm{miss}(i), \mathrm{miss}(i)]$;
  $S_{s,m} \leftarrow \Sigma_s^{-1^{(t)}}[m_s(i), m_s(i)]$; $S_{c,m} \leftarrow \Sigma_c^{-1^{(t)}}[m_c(i), m_c(i)]$;
  // Update estimates:
  $\mu^{(t+1)} \leftarrow \mu^{(t+1)} + X_m$;
  $\Sigma_c^{(t+1)} \leftarrow \Sigma_c^{(t+1)} + (X_m - \mu^{(t)})^\top \Sigma_s^{-1^{(t)}} (X_m - \mu^{(t)}) + \mathbf{e}\{m_c(i)\}^\top (R_m \circ S_{c,m}) \mathbf{e}\{m_c(i)\}$;
  $\Sigma_s^{(t+1)} \leftarrow \Sigma_s^{(t+1)} + (X_m - \mu^{(t)}) \Sigma_c^{-1^{(t)}} (X_m - \mu^{(t)})^\top + \mathbf{e}\{m_s(i)\}^\top (R_m \circ S_{s,m}) \mathbf{e}\{m_s(i)\}$;
  $\sigma^{2^{(t+1)}} \leftarrow \sigma^{2^{(t+1)}} + \mathrm{tr}(\mathrm{vec}(X_m - \mu^{(t)})^\top (\Sigma_c^{-1^{(t)}} \otimes \Sigma_s^{-1^{(t)}}) \mathrm{vec}(X_m - \mu^{(t)})) + \sum_j \mathrm{vec}(R_m \circ S_{c,m} \circ S_{s,m})$;
end
$\mu^{(t+1)} \leftarrow \mu^{(t+1)}/N$; $\sigma^{2^{(t+1)}} \leftarrow \sigma^{2^{(t+1)}}/(Npq)$; // scale
$\Sigma_c^{(t+1)} \leftarrow \mathrm{ADJUST}(pN, \sigma^{2^{(t)}}, \Sigma_c^{(t+1)})$;
$\Sigma_s^{(t+1)} \leftarrow \mathrm{ADJUST}(qN, \sigma^{2^{(t)}}, \Sigma_s^{(t+1)})$;

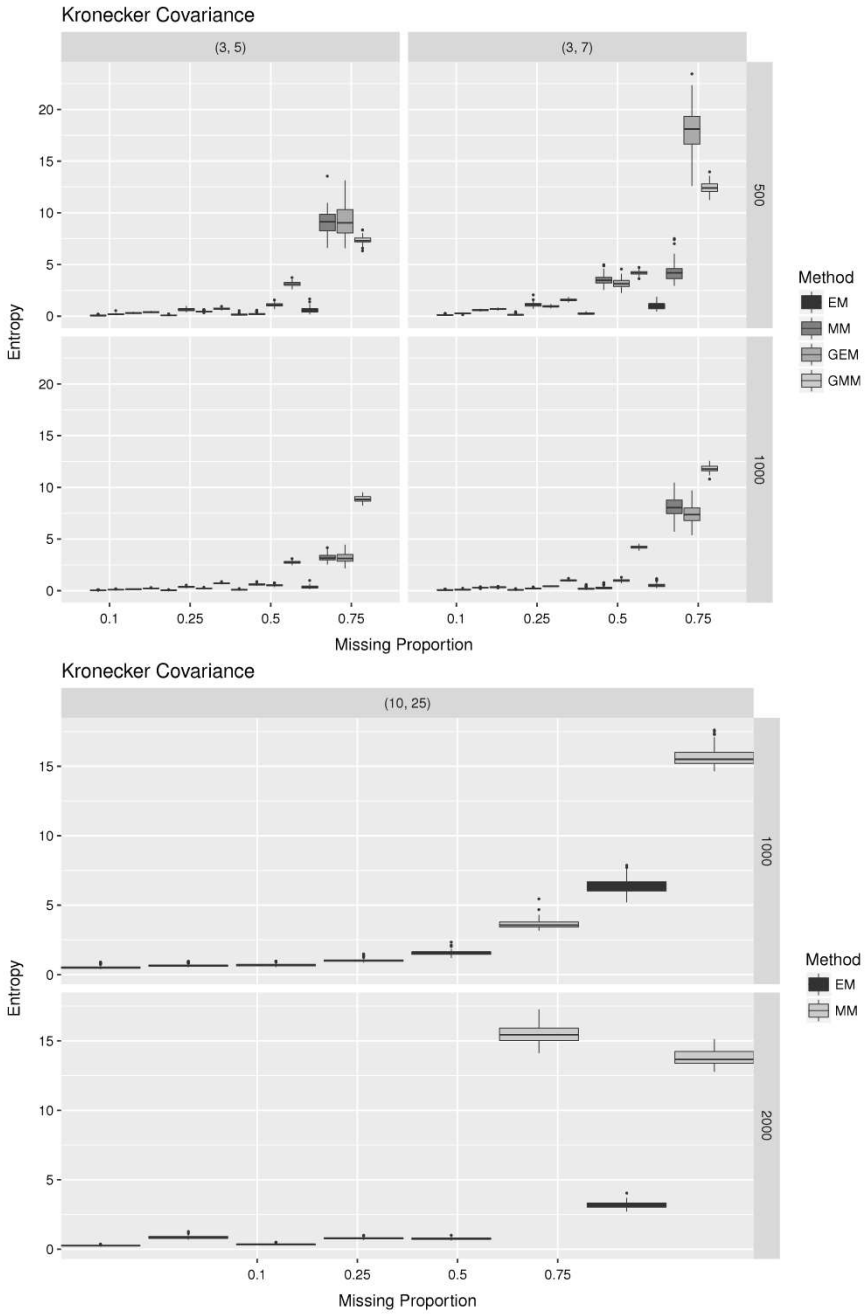

Fig. C.7. Boxplots of the entropy in the estimates of Kronecker Σ, in order of EM, MM, GEM, and GMM for dimensions: (3, 5), (3, 7), and (10, 25).

covariance and scale variance updates at lines 8, 9, and 10 cost $O(pq^2 + p^2q + p^2q^2)$ per observation, and so the overall complexity is $O(p^3 + q^3 + (N + M)p^2q^2)$ per EM update.

Each one of these update steps should be run until convergence. As a criterion to evaluate convergence, given a convergence tolerance ϵ, we stop at iteration t when

$$\frac{\|\widehat{\mu}^{(t+1)} - \widehat{\mu}^{(t)}\|_1}{\|\widehat{\mu}^{(t)}\|_1} + \frac{\|\widehat{\Sigma}_c^{(t+1)} - \widehat{\Sigma}_c^{(t)}\|_1}{\|\widehat{\Sigma}_c^{(t)}\|_1} + \frac{\|\widehat{\Sigma}_s^{(t+1)} - \widehat{\Sigma}_s^{(t)}\|_1}{\|\widehat{\Sigma}_s^{(t)}\|_1} + \frac{|\widehat{\sigma}^{2^{(t+1)}} - \widehat{\sigma}^{2^{(t)}}|}{|\widehat{\sigma}^{2^{(t)}}|} \leq \epsilon,$$

where we use relative (entrywise) L_1 norm, $\|A\|_1 = \|(\vec{A})\|_1 = \sum_{i,j}|A_{ij}|$, for better consistency when comparing across matrices with different dimensions.

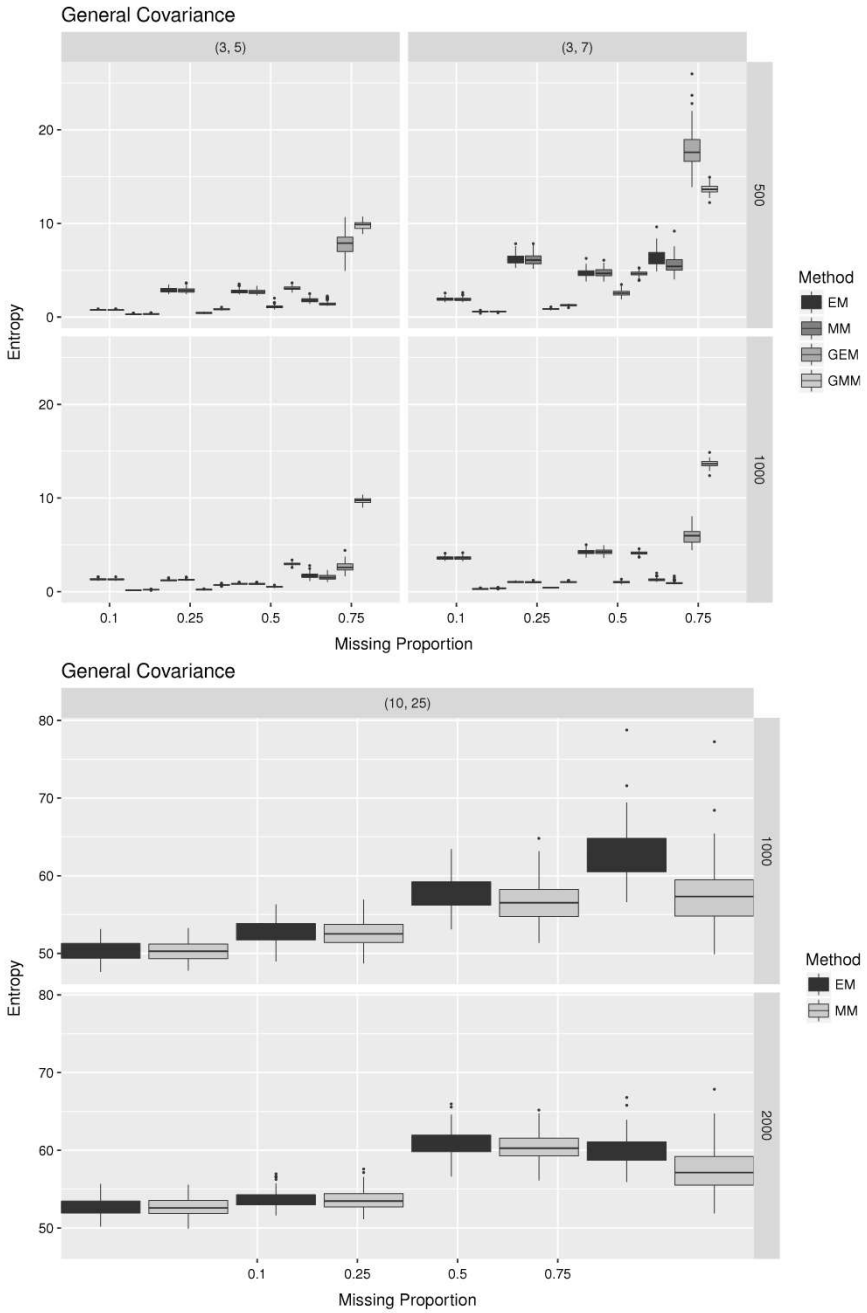

Fig. C.8. Boxplots of the entropy in the estimates of general Σ, in order of EM, MM, GEM, and GMM for dimensions: (3, 5), (3, 7), and (10, 25).

Appendix B. Proof of Theorem 1 and ADJUST procedure

Given $n > 0, \sigma^2 > 0$ and a symmetric positive definite matrix S let us define our objective function

$$f(\Sigma) = \frac{n}{2}\ln|\Sigma^{-1}| - \frac{1}{2\sigma^2}\mathrm{tr}(\Sigma^{-1}S)$$

for $\Sigma \succ 0$ and subject to $\Sigma_{11} = 1$. Moreover, let us define $g(\Sigma) = \Sigma_{11} - 1$. Then $\Sigma^* = \arg\max_{\Sigma\succ 0, \Sigma_{11}=1} f(\Sigma)$ is such that

$$\frac{\partial f}{\partial(\Sigma^{-1})}(\Sigma^*) - \frac{\lambda}{2}\frac{\partial g}{\partial(\Sigma^{-1})}(\Sigma^*) = 0$$

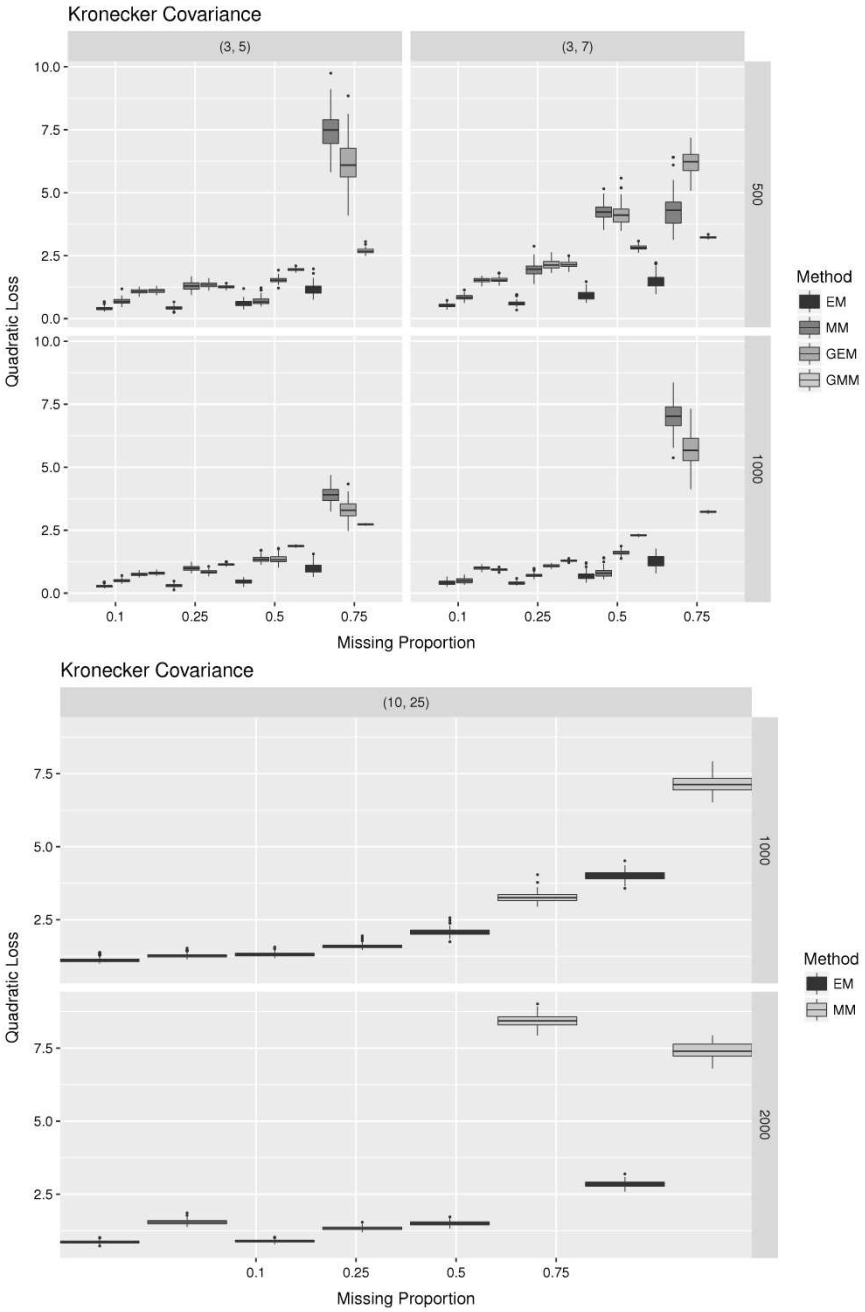

Fig. C.9. Boxplots of the quadratic loss in the estimates of Kronecker Σ, in order of EM, MM, GEM, and GMM for dimensions: $(3, 5)$, $(3, 7)$, and $(10, 25)$.

for some Lagrange multiplier λ. But since

$$\frac{\partial f}{\partial(\Sigma^{-1})} - \frac{\lambda}{2}\frac{\partial g}{\partial(\Sigma^{-1})} = \frac{n}{2}\{2\Sigma - (\Sigma \circ I)\} - \frac{1}{2\sigma^2}\{2S - (S \circ I)\} + \frac{\lambda}{2}\left[2\Sigma_{1,\cdot}\Sigma_{1,\cdot}^{\top} - \{(\Sigma_{1,\cdot}\Sigma_{1,\cdot}^{\top}) \circ I\}\right],$$

where $\circ$ is Hadamard product, I is the identity matrix, and $\Sigma_{1,\cdot}$ is the first row of Σ, we have

$$n\Sigma^* - \frac{1}{\sigma^2}S + \lambda\Sigma^*_{1,\cdot}\Sigma^{*\top}_{1,\cdot} = 0. \tag{B.1}$$

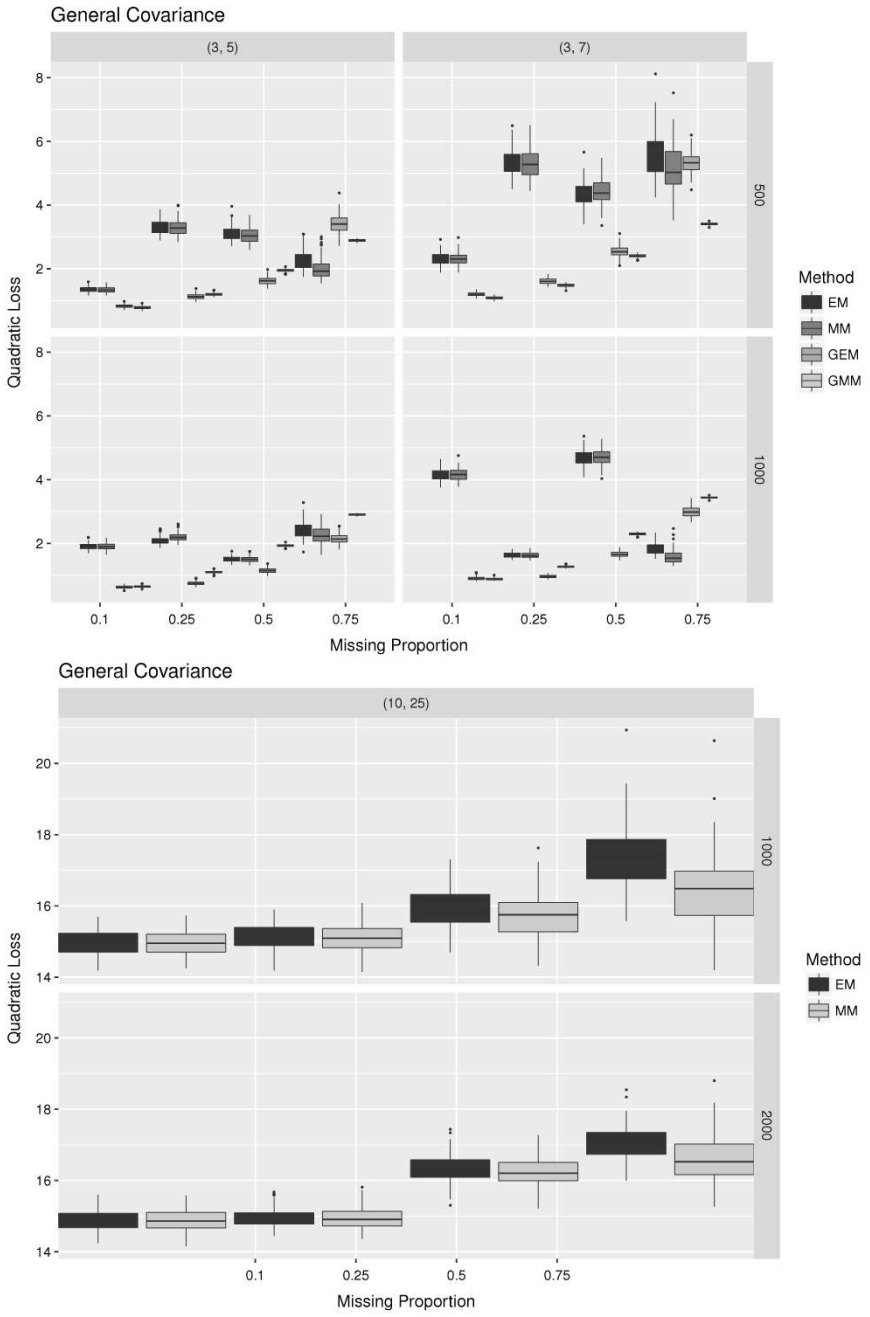

Fig. C.10. Boxplots of the quadratic loss in the estimates of general Σ, in order of EM, MM, GEM, and GMM for dimensions: $(3, 5)$, $(3, 7)$, and $(10, 25)$.

Thus, because $\Sigma^*_{11} = 1$, the $(1, 1)$ entry from (B.1) requires that $\lambda = S_{11}/\sigma^2 - n$. Next, for entries $(1, k)$ with $k > 1$ we have $n\Sigma^*_{1k} - S_{1k}/\sigma^2 + \lambda\Sigma^*_{11}\Sigma^*_{1k} = 0$ and so $\Sigma^*_{1k} = S_{1k}/S_{11}$. Finally, for all entries (j, k) with $j, k \neq 1$,

$$\Sigma^*_{jk} = \frac{1}{n}\left(\frac{S_{jk}}{\sigma^2} - \lambda\Sigma^*_{1j}\Sigma^*_{1k}\right) = \underbrace{\frac{S_{11}}{n\sigma^2}}_{=\eta}\frac{S_{jk}}{S_{11}} + \left(1 - \frac{S_{11}}{n\sigma^2}\right)\Sigma^*_{1j}\Sigma^*_{1k}.$$

This derivation can be summarized in the following simple procedure, where $S_{-1,-1}$ is submatrix of S excluding the first row and column and $S_{1,-1}$ is the first row excluding the first entry:

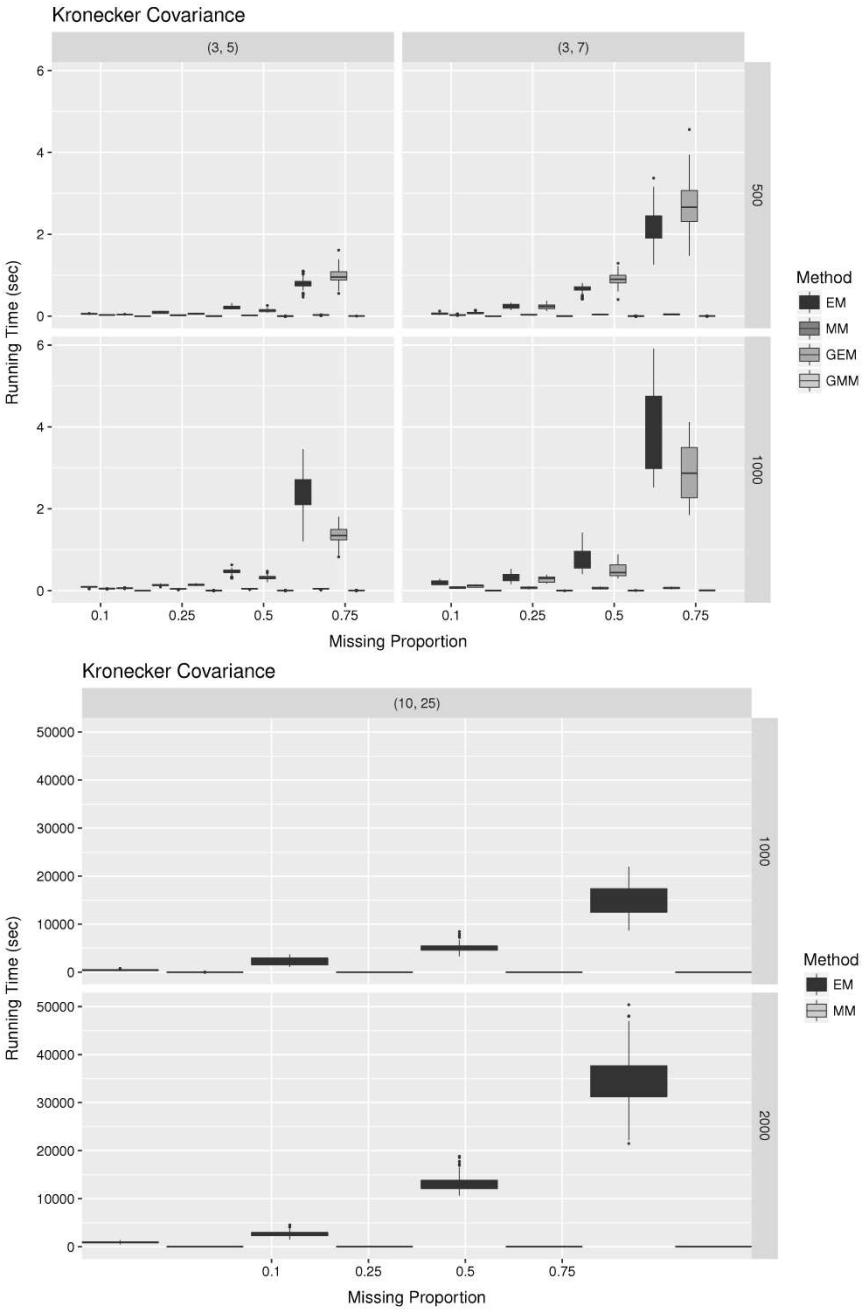

Fig. C.11. Boxplots of the run time in the estimation of Kronecker Σ, in order of EM, MM, GEM, and GMM for dimensions: (3, 5), (3, 7), and (10, 25). There was no constraint on the number of iterations.

Algorithm 5: ADJUST(n, σ^2, S) procedure

$\eta \leftarrow S_{11}/(n\sigma^2)$;
$S \leftarrow S/S_{11}$; // scale whole matrix
$S_{-1,-1} \leftarrow \eta S_{-1,-1} + (1-\eta) S_{1,-1} S_{1,-1}^{\top}$;

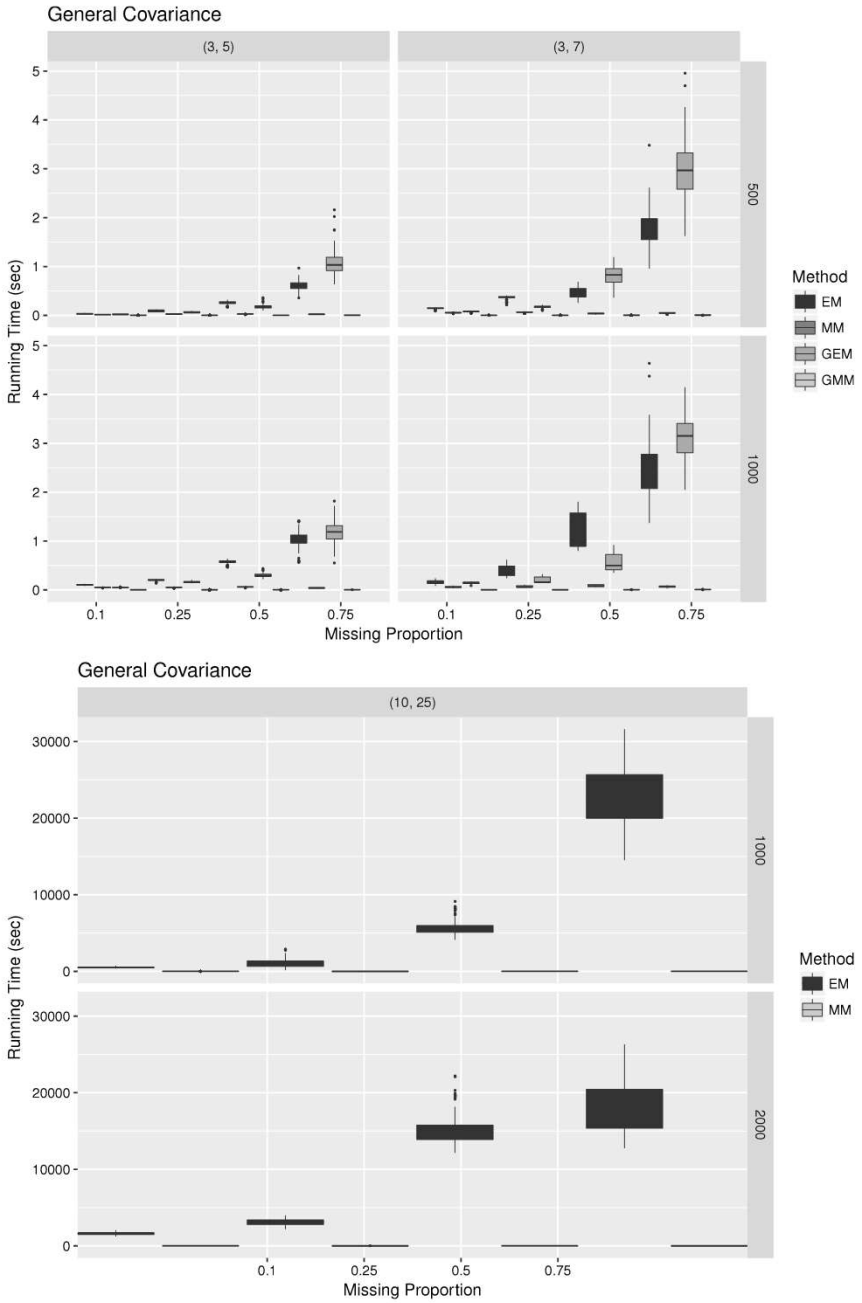

Fig. C.12. Boxplots of the run time in the estimation of general Σ, in order of EM, MM, GEM, and GMM for dimensions: (3, 5), (3, 7), and (10, 25). There was no constraint on the number of iterations.

Appendix C. Additional metrics for simulation study

See Figs. C.7–C.12.

Appendix D. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jmva.2018.03.010.

References

[1] G.I. Allen, R. Tibshirani, Transposable regularized covariance models with an application to missing data imputation, Ann. Appl. Stat. 4 (2010) 764–790.
[2] J.B. Boik, Scheffe's mixed model for multivariate repeated measures: A relative efficiency evaluation, Comm. Statist. Theory Methods 20 (1991) 1233–1255.
[3] N.R. Chaganty, D.N. Naik, Analysis of multivariate longitudinal data using quasi-least squares, J. Statist. Plann. Inference 103 (2002) 421–436.
[4] N. Cressie, The origins of kriging, Math. Geol. 22 (3) (1990) 239–252.
[5] A.P. Dawid, Some matrix-variate distribution theory: Notational considerations and a Bayesian application, Biometricka 68 (1) (1981) 265–274.
[6] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the EM Algorithm, J. Roy. Statist. Soc. 39 (1977) 1–38.
[7] P. Dutilleul, The MLE algorithm for the matrix normal distribution, J. Statist. Comput. Simul. 64 (2) (1999) 105–123.
[8] M.A. Friedl, D. Sulla-Menashe, B. Tan, A. Schneider, N. Ramankutty, A. Sibley, X. Huang, Modis collection 5 global land cover: Algorithm refinements characterization of new datasets, Remote Sens. Environ. 114 (2009) 168–182.
[9] M. Fuentes, Testing for separability of spatial-temporal covariance functions, J. Statist. Plann. Inference 136 (2004) 447–466.
[10] A.T. Galecki, General class of covariance structures for two or more repeated factors in longitudinal data analysis, Comm. Statist. Theory Methods 23 (11) (1994) 3105–3119.
[11] J.H. Goodnight, A tutorial on the SWEEP operator, Amer. Statist. 33 (3) (1979) 149–158.
[12] K. Greenewald, A.O. Hero, Robust kronecker product PCA for spatio-temporal covariance estimation, IEEE Trans. Signal Process. 63 (23) (2015) 6368–6378.
[13] J.G. Ibrahim, M.-H. Chen, S.R. Lipsitz, Missing responses in generalised linear mixed models when the missing data mechanism is nonignorable, Biometrika 88 (2) (2001) 551–564.
[14] J.G. Ibrahim, M.-H. Chen, S.R. Lipsitz, A.H. Herring, Missing-data methods for generalized linear models: A comparative review, J. Amer. Statist. Assoc. 100 (469) (2005) 332–346.
[15] R.J. Kauth, G.S. Thomas, The tasselled cap–a graphic description of the spectral-temporal development of agricultural crops as seen by Landsat, in: LARS Symposia, 1976, p. 159.
[16] R.J.A. Little, D.B. Rubin, Statistical Analysis with Missing Data, John Wiley & Sons, 2014.
[17] K. Lounici, High-dimensional covariance matrix estimation with missing observations, Bernoulli 20 (3) (2014) 1029–1058.
[18] T.R. Loveland, A.S. Belward, The IGBP-DIS global 1 km land cover data set. DISCover: First results, Int. J. Remote Sens. 18 (1997) 3291–3295.
[19] G.J. McLachlan, T. Krishnan, The EM Algorithm and Extensions, Wiley, New York, 1997.
[20] D.N. Naik, S. Rao, Analysis of multivariate repeated measures data with a Kronecker product structured covariance matrix, J. Appl. Stat. 28 (2001) 91–105.
[21] K.B. Petersen, M.S. Pedersen, The matrix cookbook, Technical University of Denmark, 2008.
[22] B. Roś, F. Bijma, J.C. de Munck, M.C.M. de Gunst, Existence and uniqueness of the maximum likelihood estimator for models with a Kronecker product covariance structure, J. Multivariate Anal. 143 (2016) 345–361.
[23] A. Roy, Testing the hypothesis of a kroneckar product covariance matrix in multivariate repeated measures data, Stat. Methodol. 2 (2005).
[24] A. Roy, R. Khattree, On discrimination and classification with multivariate repeated measures data, J. Statist. Plann. Inference 134 (2) (2005) 462–485.
[25] D.B. Rubin, Inference and missing data, Biometrika 63 (3) (1976) 581–592.
[26] C. Schaaf, F. Gao, A.H. Strahler, W. Lucht, X. Li, T. Tsang, N.C. Strugnell, X. Zhang, Y. Jin, J.-P. Muller, P. Lewis, M. Barnsley, P. Lewis, M. Barnsley, P. Hobson, M. Disney, G. Roberts, M. Dunderdale, C. Doll, R.P. dEntremont, B. Hug, S. Liang, J.L. Privette, D. Roy, First operational BRDF, albedo nadir reflectance products from modis, Remote Sens. Environ. 83 (1) (2002) 135–148.
[27] M. Shitan, P.J. Brockwell, An asymptotic test for separability of a spatial autoregressive model, Comm. Statist. Theory Methods 24 (8) (1995) 2027–2040.
[28] M.S. Srivastava, C.G. Khatri, An Introduction to Multivariate Statistics, North Holland, New York, USA, 1979, pp. 170–171.
[29] M.S. Srivastava, T. Nahtman, D. von Rosen, Models with a Kronecker product covariance structure: Estimation and testing, Math. Methods Statist. 17 (4) (2008) 357–370.
[30] G.C.G. Wei, M.A. Tanner, A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms, J. Amer. Statist. Assoc. 85 (411) (1990) 699–704.
[31] X. Zhang, C.B. Schaaf, M.A. Friedl, A.H. Strahler, F. Gao, J.C.F. Hodges, MODIS tasseled cap transformation and its utility, in: Geoscience and Remote Sensing Symposium, 2002. IGARSS'02. 2002 IEEE International, Vol. 2, IEEE, 2002, pp. 1063–1065.