Statistics and Computing (2025) 35:39
https://doi.org/10.1007/s11222-025-10575-0

ORIGINAL PAPER

An EM algorithm for fitting matrix-variate normal distributions on interval-censored and missing data

Victor H. Lachos[1] · Salvatore D. Tomarchio[2] · Antonio Punzo[2] · Salvatore Ingrassia[2]

Received: 15 February 2024 / Accepted: 20 January 2025

Abstract
Matrix-variate distributions are powerful tools for modeling three-way datasets that often arise in longitudinal and multidimensional spatio-temporal studies. However, observations in these datasets can be missing or subject to some detection limits because of the restriction of the experimental apparatus. Here, we develop an efficient EM-type algorithm for maximum likelihood estimation of parameters, in the context of interval-censored and/or missing data, utilizing the matrix-variate normal distribution. This algorithm provides closed-form expressions that rely on truncated moments, offering a reliable approach to parameter estimation under these conditions. Results obtained from the analysis of both simulated data and real case studies concerning water quality monitoring are reported to demonstrate the effectiveness of the proposed method.

Keywords Censored data · ECM algorithm · Matrix-variate distribution · Missing data · Truncated moments

1 Introduction

In numerous practical applications and experimental studies, data collection often involves interval censoring and missing values. As concerns the first aspect, variables of interest can be prone to specific threshold values, beyond or below which measurements become unattainable; in other terms, the observed values are restricted to specific intervals. This scenario is notably prevalent in the examination of human immunodeficiency virus (HIV) behavior, wherein the quantification of HIV-1 RNA viral load is typically assessed according to certain upper and lower detection limits (see, e.g. De Gruttola and Lagakos,1989; Lachos et al., 2011). Another illustrative example is evident in environmental research, where concentration levels of chemical substances are influenced by multiple detection limit values (see, e.g. Galarza et al., 2022). Regarding the second aspect, missing values are often a commonplace occurrence, stemming from various factors such as non-disclosure of information, technical failures, data processing errors, or the researcher's inability to procure a specific observation. These circumstances have the potential to constrain or significantly diminish the comprehensive understanding and analysis of the phenomena of interest.

As well-known in the statistical literature (see, for example, Acock, 2005; Helsel, 2011), the least advisable approach in handling this kind of data is to omit or erase them. This introduces a significant bias into subsequent estimates and eliminates crucial information embedded within them. Likewise, imputing or substituting synthetic values, as if they were genuinely measured, often introduces a set of inaccuracies. Indeed, an intrusive signal to the data can be added which may either obscure the information present in the actual measurements or, conversely, might introduce a signal that is not genuinely inherent in the data. It is easy to understand that the analysis of data jointly containing interval-censored and missing values poses even more significant challenges. Our work focuses on data simultaneously presenting both characteristics.

In the univariate and multivariate literature, different works have been recently introduced to jointly handle censored (not necessarily of interval nature) and missing data. Some examples in this direction consists of the works of Lin et al. (2018); Valeriano et al. (2021); Faucheux et al. (2021); Galarza Morales et al. (2021); Bahari et al. (2021); de Alencar et al. (2022) and Valeriano et al. (2023). Nevertheless, our aim in this study is directed towards the examination of matrix-

✉ Salvatore D. Tomarchio
daniele.tomarchio@unict.it

[1] Department of Statistics, University of Connecticut, Storrs, CT 06269, USA

[2] Department of Economics and Business, University of Catania, Catania, Italy

Published online: 28 January 2025

variate data. This kind of data originates from examining p attributes in q occasions across a set of n units. Examples include spatial multivariate data, multivariate longitudinal data, and spatio-temporal data. In all these instances, a $p \times q$ matrix is observed for each statistical unit, signifying that a sample of n matrices can be organized into a $p \times q \times n$ three-way array.

The matrix-variate literature is gaining increasing attention, as evidenced by the growing number of contributions in this field (refer to, for instance, Sarkar et al., 2020; Gallaugher and McNicholas, 2020; Thompson et al., 2020; Tomarchio, 2024; Tomarchio et al., 2022; 2023). While there have been efforts to address the missing data problem within this context, as proposed by Triantafyllopoulos (2008); Allen and Tibshirani (2010); Glanz and Carvalho (2018); Zhang and Bandyopadhyay (2020), there is currently no work addressing the issue of censored data, to the best of the author's knowledge. Our paper introduces an approach designed to simultaneously handle both interval-censored and missing data within a matrix-variate framework, thus filling a gap in the existing literature. More in detail, we propose an approach based on the matrix-variate normal (MVN) distribution which, among the matrix-variate distributions, plays the same pivotal role as the multivariate normal distribution in the family of multivariate distributions. In particular, a fully likelihood-based approach is carried out, including the implementation of an expectation-conditional maximization (ECM) algorithm (Meng and Rubin 1993) for maximum likelihood (ML) parameter estimation. By using the properties of the MVN distribution, we compute the truncated moments required to obtain closed-form expressions for parameter estimates. The interval censoring mechanism therein illustrated allows us to handle missing (at random) and censored values simultaneously (de Alencar et al. 2022; Valeriano et al. 2023).

We assess the effectiveness of the suggested method through analyses conducted on both simulated and real data. In the simulated data analysis, we assess parameter recovery outcomes across various scenarios and contrast them with those obtained from an alternative method. In the real data analysis, we focus on examining concentration levels of chemical substances dissolved in water at the Chesapeake Bay (see, Murphy et al., 2019). Specifically, our investigation targets the data collected at two monitoring stations, each exhibiting a distinct combination of interval-censored and missing data.

The paper is organized as follows. In Sect. 2, we briefly discuss some preliminary concepts related to the MVN distribution, its properties, and the advantages of using the matrix-variate form factor over its vectorization. In Sect. 3, we present the MVN model for interval-censored and missing data, including the ECM algorithm for ML estimation. In Sect. 4, we illustrate the simulated and real data analyses. Finally, Sect. 5 concludes with some discussion and possible directions for future research.

2 Preliminaries

In this section, we begin by providing some notation that will be consistently employed throughout the manuscript. Specifically, a random matrix of dimensions $p \times q$ is labeled as $\boldsymbol{\mathcal{X}}$, and its observed realization is denoted as $\mathcal{X}$. Moreover, a random vector of dimension pq is identified as $\mathbf{X}$, while its observed realization is labeled as $\mathbf{x}$.

Then, we present a concise overview of the MVN distribution and its key features (Sect. 2.1). Lastly, we revisit the benefits associated with employing the matrix-variate form as opposed to its vectorization (Sect. 2.2).

2.1 Matrix-variate normal distribution

As introduced in Sect. 1, the MVN distribution is the most well-known in the matrix-variate setting, particularly for its properties and mathematical tractability (Gupta and Nagar 1999). An $p \times q$ random matrix $\boldsymbol{\mathcal{X}}$ follows an MVN distribution with mean parameter $\mathbf{M}$ and covariance matrices $\boldsymbol{\Sigma}$ and $\boldsymbol{\Psi}$ of dimensions $p \times p$ and $q \times q$, respectively, if its probability density function (pdf) is

$$f(\mathcal{X}|\mathbf{M}, \boldsymbol{\Sigma}, \boldsymbol{\Psi}) = \frac{1}{(2\pi)^{pq/2}|\boldsymbol{\Psi}|^{p/2}|\boldsymbol{\Sigma}|^{q/2}} \times \exp\left\{-\frac{1}{2}\text{tr}[\boldsymbol{\Psi}^{-1}(\mathcal{X}-\mathbf{M})^{\top}\boldsymbol{\Sigma}^{-1}(\mathcal{X}-\mathbf{M})]\right\}, \quad (1)$$

where $|\mathbf{A}|$ and $\text{tr}(\mathbf{A})$ denote the determinant and trace of the matrix $\mathbf{A}$, respectively. Notationally, we write $\boldsymbol{\mathcal{X}} \sim \mathcal{N}_{p\times q}(\mathbf{M}, \boldsymbol{\Sigma}, \boldsymbol{\Psi})$.

An equivalent definition specifies the MVN distribution as a special case of the multivariate normal (MN) distribution. Specifically,

$$\boldsymbol{\mathcal{X}} \sim \mathcal{N}_{p\times q}(\mathbf{M}, \boldsymbol{\Sigma}, \boldsymbol{\Psi}) \iff \text{vec}(\boldsymbol{\mathcal{X}}) = \mathbf{X} \sim \mathcal{N}_{pq}(\boldsymbol{\mu} = \text{vec}(\mathbf{M}), \boldsymbol{\Lambda} = \boldsymbol{\Psi} \otimes \boldsymbol{\Sigma}), \quad (2)$$

where $\mathcal{N}_{pq}(\cdot)$ denotes the MN distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Lambda}$, $\text{vec}(\cdot)$ is the vectorization operator, and $\otimes$ denotes the Kronecker product. Lastly, it is worth mentioning the existence of an identifiability issue involving the two covariance matrices. Indeed, $\boldsymbol{\Psi} \otimes \boldsymbol{\Sigma} = \boldsymbol{\Psi}^* \otimes \boldsymbol{\Sigma}^*$ if $\boldsymbol{\Sigma}^* = a\boldsymbol{\Sigma}$ and $\boldsymbol{\Psi}^* = a^{-1}\boldsymbol{\Psi}$. Therefore, the two covariance matrices are identifiable up to a multiplicative constant (Dutilleul 1999). In the matrix-variate literature, several approaches have been proposed to address this problem. Herein, we adopt the approach used in Melnykov and Zhu (2018); Sarkar et al.

(2020); Tomarchio et al. (2022), which consists of imposing the constraint $|\boldsymbol{\Psi}| = 1$.

2.2 Benefits over vectorization

From a theoretical point of view, the relationship in (2) is often convenient for establishing and calculating quantities of interest when working with MVN-based models (see, for example, Viroli, 2011; 2012). Excluding these cases, when conducting analyses, the matrix-variate formulation is preferred over its multivariate counterpart when data comes in a matrix-variate form factor. Indeed, a potential strategy involves the vectorization of the matrix-variate data, followed by the application of multivariate models. However, this data rearrangement presents several practical drawbacks, as extensively discussed and shown in the matrix-variate literature (see, for example, Allen and Tibshirani, 2010; Anderlucci et al., 2014; Gallaugher and McNicholas, 2018; Sarkar et al., 2020; Tomarchio et al., 2021). In the following, we provide a summary of the issues induced by such an approach.

1. Interpretability: the identification of the two sources of variability, governed by $\boldsymbol{\Sigma}$ and $\boldsymbol{\Psi}$, would not be possible if they were combined into a unique $\boldsymbol{\Psi} \otimes \boldsymbol{\Sigma}$ matrix. This leads to a loss of interpretation of the data.
2. Parsimony: the number of free covariance parameters in the matrix-variate setting is $p(p+1)/2-q(q+1)/2-1$. When vectorizing, this number rises to $pq(pq+1)/2$, potentially resulting in an overparameterization of the multivariate models.
3. Model selection: strictly connected to the previous problem, the increase in the number of parameters in the multivariate setting can introduce challenges in model selection. This is a consequence of the increased weight assigned to the penalty term of widely used information criteria.

3 Matrix-variate normal distribution for interval-censored and missing data

Let $\widetilde{\mathcal{X}} = \{\mathcal{X}_1, \ldots, \mathcal{X}_i, \ldots, \mathcal{X}_n\}$ be a set of n matrices from model (1). In this framework, we consider a similar approach to that proposed by Valeriano et al. (2023) to model the censored and missing multivariate responses. In detail, the observed data for the ith subject are given by the matrices $(\mathcal{V}_i, \mathcal{C}_i)$, where each element of the matrix $\mathcal{V}_i$ represents either the uncensored observations ($V_{ijk} = V_{0ijk}$) or the interval-censoring level ($V_{ijk} \in [V_{1ijk}, V_{2ijk}]$), and $\mathcal{C}_i$ is the matrix of censoring indicators, satisfying

$$C_{ijk} = \begin{cases} 1 & \text{if } V_{1ijk} \leq X_{ijk} \leq V_{2ijk} \\ 0 & \text{if } X_{ijk} = V_{0ijk} \end{cases}, \tag{3}$$

for all $i \in \{1, \ldots, n\}$, $j \in \{1, \ldots, p\}$ and $k \in \{1, \ldots, q\}$. Thus, the observed data for a set of n matrices are given by $\widetilde{\mathcal{V}} = \{\mathcal{V}_1, \ldots, \mathcal{V}_i, \ldots, \mathcal{V}_n\}$ and $\widetilde{\mathcal{C}} = \{\mathcal{C}_1, \ldots, \mathcal{C}_i, \ldots, \mathcal{C}_n\}$. In this case, (1) and (3) define the MVN interval-censored model (hereafter, the MVNC model). To provide further elucidation, the censoring structure results in truncation at both the upper and lower bounds of the distribution's support. This is due to our knowledge being limited to the fact that the true observation X_{ijk} is less than or equal to the observed quantity V_{2ijk} and also greater than or equal to the observed quantity V_{1ijk}. Missing observations can be conveniently handled by considering $V_{1ijk} = -\infty$ and $V_{2ijk} = +\infty$.

3.1 The likelihood function

For ease of calculation, we use the relationship in (2) in this section. Specifically, let us consider $\text{vec}(\widetilde{\mathcal{X}}) = \{\mathbf{x}_1, \ldots, \mathbf{x}_i, \ldots, \mathbf{x}_n\}$, $\text{vec}(\widetilde{\mathcal{V}}) = \{\mathbf{v}_1, \ldots, \mathbf{v}_i, \ldots, \mathbf{v}_n\}$, and $\text{vec}(\widetilde{\mathcal{C}}) = \{\mathbf{c}_1, \ldots, \mathbf{c}_i, \ldots, \mathbf{c}_n\}$. To obtain the likelihood function of the MVNC model, the first step is to partition each subject as $\mathbf{x}_i = (\mathbf{x}_i^{o\top}, \mathbf{x}_i^{c\top})^\top$, where the superscripts o and c refer to the observed and censored parts, respectively. Accordingly and after reordering, we have $\mathbf{c}_i = (\mathbf{c}_i^{o\top}, \mathbf{c}_i^{c\top})^\top$, $\mathbf{v}_i = (\mathbf{v}_i^{o\top}, \mathbf{v}_i^{c\top})^\top$, with $\mathbf{v}_i^c = (\mathbf{v}_{1i}^c, \mathbf{v}_{2i}^c)$,

$$\boldsymbol{\mu}_i = (\boldsymbol{\mu}_i^{o\top}, \boldsymbol{\mu}_i^{c\top})^\top, \qquad \boldsymbol{\Lambda}_i = \begin{pmatrix} \boldsymbol{\Lambda}_i^{oo} \boldsymbol{\Lambda}_i^{oc} \\ \boldsymbol{\Lambda}_i^{co} \boldsymbol{\Lambda}_i^{cc} \end{pmatrix}, \tag{4}$$

being the quantities in (4) the corresponding partitions of $\boldsymbol{\mu}_i$ and $\boldsymbol{\Lambda}_i$. The computation of the likelihood function involves marginalization and conditioning of an MN distribution. Thus, from the properties of MN distribution, we have that $\mathbf{X}_i^o \sim \mathcal{N}_{p_i^o}(\boldsymbol{\mu}_i^o, \boldsymbol{\Lambda}_i^{oo})$, and $\mathbf{X}_i^c \mid \mathbf{X}_i^o = \mathbf{x}_i^o \sim \mathcal{N}_{p_i^c}(\boldsymbol{\mu}_i^{co}, \boldsymbol{\Lambda}_i^{cc.o})$, where

$$\begin{aligned} \boldsymbol{\mu}_i^{co} &= \boldsymbol{\mu}_i^c + \boldsymbol{\Lambda}_i^{co}(\boldsymbol{\Lambda}^{oo})_i^{-1}(\mathbf{z}_i^o - \boldsymbol{\mu}_i^o), \\ \boldsymbol{\Lambda}_i^{cc.o} &= \boldsymbol{\Lambda}_i^{cc} - \boldsymbol{\Lambda}_i^{co}(\boldsymbol{\Lambda}_i^{oo})^{-1}\boldsymbol{\Lambda}_i^{oc}. \end{aligned} \tag{5}$$

Now, let $\Phi_p(\mathbf{u}_1, \mathbf{u}_2; \mathbf{a}, \mathbf{A})$ and $\phi_p(\mathbf{u}; \mathbf{a}, \mathbf{A})$ be the cdf (computed at interval $[\mathbf{u}_1, \mathbf{u}_2]$) and pdf (computed at vector $\mathbf{u}$), respectively, of a p-variate MN distribution; in symbols, $N_p(\mathbf{a}, \mathbf{A})$. From Valeriano et al. (2023) (see also, de Alencar et al. 2022), the likelihood function of $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\Lambda})$ given the observed data is

$$\ell(\boldsymbol{\theta} \mid \text{vec}(\widetilde{\mathcal{V}}), \text{vec}(\widetilde{\mathcal{C}})) = \sum_{i=1}^{n} \ln L_i, \tag{6}$$

where L_i represents the likelihood function of $\boldsymbol{\theta}$ for the ith sample observation, given by

$$L_i = f(\mathbf{v}_{1i}^c \le \mathbf{x}_i^c \le \mathbf{v}_{2i}^c \mid \mathbf{x}_i^o, \boldsymbol{\theta}) f(\mathbf{x}_i^o \mid \boldsymbol{\theta}) = \Phi_{p_i^c}(\mathbf{v}_{1i}^c, \mathbf{v}_{2i}^c; \boldsymbol{\mu}_i^{co}, \boldsymbol{\Lambda}_i^{cc.o}) \, \phi_{p_i^o}(\mathbf{x}_i^o; \boldsymbol{\mu}_i^o, \boldsymbol{\Lambda}_i^{oo}). \quad (7)$$

Despite the observed log-likelihood function (6) can be calculated without much computational burden, it involves complex expressions and it is not convenient for ML parameter estimation. However, the estimation process can be considerably simplified by the application of an EM-based algorithm, as discussed in the subsequent section.

3.2 The ECM algorithm

We describe here an ECM algorithm, originally proposed by Meng and Rubin (1993), for ML parameter estimation. The ECM replaces the M-step with a sequence of conditional maximization (CM) steps. This algorithm preserves the stability of the EM and has a typically faster convergence rate than the original EM (McLachlan and Krishnan 2008).

For the application of the ECM algorithm, the set $\widetilde{\mathcal{X}}$ is viewed as incomplete and it is augmented with the sets $\{\widetilde{\mathcal{V}}, \widetilde{\mathcal{C}}\}$ to obtain the complete data $\widetilde{S} = \{\widetilde{\mathcal{X}}, \widetilde{\mathcal{V}}, \widetilde{\mathcal{C}}\}$. Hence, the corresponding complete-data log-likelihood function is $\ell_c(\boldsymbol{\Theta}) = \sum_{i=1}^n \ell_{ic}(\boldsymbol{\Theta})$, where $\boldsymbol{\Theta} = \{\mathbf{M}, \boldsymbol{\Sigma}, \boldsymbol{\Psi}\}$ and the individual complete data log-likelihood is as follows

$$\begin{aligned}\ell_{ic}(\boldsymbol{\Theta}) &= -\frac{pq}{2}\log(2\pi) - \frac{p}{2}\log(|\boldsymbol{\Psi}|) - \frac{q}{2}\log(|\boldsymbol{\Sigma}|) \\ &\quad -\frac{1}{2}\mathrm{tr}[\boldsymbol{\Psi}^{-1}(\mathcal{X}_i - \mathbf{M})^\top \boldsymbol{\Sigma}^{-1}(\mathcal{X}_i - \mathbf{M})] \\ &= -\frac{pq}{2}\log(2\pi) - \frac{p}{2}\log(|\boldsymbol{\Psi}|) - \frac{q}{2}\log(|\boldsymbol{\Sigma}|) \\ &\quad -\frac{1}{2}\mathrm{tr}[(\boldsymbol{\Psi}\otimes\boldsymbol{\Sigma})^{-1}\mathrm{vec}(\mathcal{X}_i - \mathbf{M})\mathrm{vec}(\mathcal{X}_i - \mathbf{M})^\top].\end{aligned}$$

Subsequently, the ECM algorithm for the MVNC model can be summarized as follows:

E-step: Given the current estimate $\widehat{\boldsymbol{\Theta}}^{(k)} = \{\widehat{\mathbf{M}}^{(k)}, \widehat{\boldsymbol{\Sigma}}^{(k)}, \widehat{\boldsymbol{\Psi}}^{(k)}\}$ at the kth step of the algorithm, the E-step provides the conditional expectation of the complete data log-likelihood function, i.e.,

$$Q(\boldsymbol{\Theta} \mid \widehat{\boldsymbol{\Theta}}^{(k)}) = \mathbb{E}\Big[\ell_c(\boldsymbol{\Theta}) \mid \widetilde{\mathcal{V}}, \widetilde{\mathcal{C}}, \widehat{\boldsymbol{\Theta}}^{(k)}\Big] = \sum_{i=1}^n Q_i(\boldsymbol{\Theta} \mid \widehat{\boldsymbol{\Theta}}^{(k)}),$$

where

$$\begin{aligned}Q_i(\boldsymbol{\Theta} \mid \widehat{\boldsymbol{\Theta}}^{(k)}) &= -\frac{pq}{2}\log(2\pi) - \frac{p}{2}\log(|\boldsymbol{\Psi}|) - \frac{q}{2}\log(|\boldsymbol{\Sigma}|) \\ &\quad -\frac{1}{2}\mathrm{tr}\Big\{(\boldsymbol{\Psi}\otimes\boldsymbol{\Sigma})^{-1} \\ &\quad \times \Big[\big(\mathrm{vec}(\widehat{\mathcal{X}}_i^{(k)}) - \mathrm{vec}(\mathbf{M})\big)\big(\mathrm{vec}(\widehat{\mathcal{X}}_i^{(k)}) \\ &\quad -\mathrm{vec}(\mathbf{M})\big)^\top + \widehat{\boldsymbol{\Omega}}_i^{(k)}\Big]\Big\}, \quad (8)\end{aligned}$$

where $\mathrm{vec}(\widehat{\mathcal{X}}_i^{(k)}) = \mathbb{E}_{\boldsymbol{\mathcal{X}}_i}[\mathrm{vec}(\boldsymbol{\mathcal{X}}_i) \mid \widetilde{\mathcal{V}}_i, \widetilde{\mathcal{C}}_i, \widehat{\boldsymbol{\Theta}}^{(k)}]$ and $\widehat{\boldsymbol{\Omega}}_i^{(k)} = \mathbb{C}ov_{\boldsymbol{\mathcal{X}}_i}[\mathrm{vec}(\boldsymbol{\mathcal{X}}_i) \mid \widetilde{\mathcal{V}}_i, \widetilde{\mathcal{C}}_i, \widehat{\boldsymbol{\Theta}}^{(k)}]$.

CM-step 1: Conditionally maximizing $Q(\boldsymbol{\Theta} \mid \widehat{\boldsymbol{\Theta}}^{(k)})$ with respect to $\mathbf{M}$ and $\boldsymbol{\Sigma}$, we obtain the following updates

$$\widehat{\mathbf{M}}^{(k+1)} = \frac{1}{n}\sum_{i=1}^n \widehat{\mathcal{X}}_i^{(k)}, \quad (9)$$

$$\widehat{\boldsymbol{\Sigma}}^{(k+1)} = \frac{1}{nq}\sum_{i=1}^n\sum_{j=1}^{pq} \widehat{\mathbf{B}}_{ij}^{(k)} \widehat{\boldsymbol{\Psi}}^{(k)^{-1}} \widehat{\mathbf{B}}_{ij}^{(k)\top}, \quad (10)$$

where $\widehat{\mathbf{B}}_{ij}^{(k)}$ is a $p \times q$ matrix such that $\mathrm{vec}(\widehat{\mathbf{B}}_{ij}^{(k)}) = \widehat{\mathbf{L}}_{ij}^{(k)}$. Here $\widehat{\mathbf{L}}_{ij}^{(k)}$ is jth column of the $pq \times pq$ lower triangular matrix $\widehat{\mathbf{L}}_i^{(k)}$, obtained from the Cholesky decomposition of the matrix

$$\widehat{\boldsymbol{\Delta}}_i^{(k)} = \big(\mathrm{vec}(\widehat{\mathcal{X}}_i^{(k)}) - \mathrm{vec}(\widehat{\mathbf{M}}^{(k+1)})\big)\big(\mathrm{vec}(\widehat{\mathcal{X}}_i^{(k)}) - \mathrm{vec}(\widehat{\mathbf{M}}^{(k+1)})\big)^\top + \widehat{\boldsymbol{\Omega}}_i^{(k)},$$

such that $\boldsymbol{\Delta}_i^{(k)} = \widehat{\mathbf{L}}_i^{(k)}\widehat{\mathbf{L}}_i^{(k)\top}$. Note that $\widehat{\boldsymbol{\Delta}}_i^{(k)}$ is a symmetric positive-definite matrix since it is the sum of two symmetric positive-definite matrices.

CM-step 2: Conditionally maximizing $Q(\boldsymbol{\Theta} \mid \widehat{\boldsymbol{\Theta}}^{(k)})$ with respect to $\boldsymbol{\Psi}$, we get the last parameter update

$$\widehat{\boldsymbol{\Psi}}^{(k+1)} = \frac{1}{np}\sum_{i=1}^n\sum_{j=1}^{pq} \widehat{\mathbf{B}}_{ij}^{(k)\top} \widehat{\boldsymbol{\Sigma}}^{(k+1)^{-1}} \widehat{\mathbf{B}}_{ij}^{(k)}. \quad (11)$$

Note that, to satisfy the identifiability constraint $|\Psi| = 1$ in the estimation process, the estimator of $\boldsymbol{\Psi}$ in (11) is then replaced by:

$$\widehat{\boldsymbol{\Psi}}^{(k+1)} = \frac{\sum_{i=1}^n\sum_{j=1}^{pq} \widehat{\mathbf{B}}_{ij}^{(k)\top} \widehat{\boldsymbol{\Sigma}}^{(k+1)^{-1}} \widehat{\mathbf{B}}_{ij}^{(k)}}{|\sum_{i=1}^n\sum_{j=1}^{pq} \widehat{\mathbf{B}}_{ij}^{(k)\top} \widehat{\boldsymbol{\Sigma}}^{(k+1)^{-1}} \widehat{\mathbf{B}}_{ij}^{(k)}|^{1/q}}. \quad (12)$$

We stopped the algorithm when $|\ell(\widehat{\boldsymbol{\theta}}^{(k+1)}|\mathbf{V}, \mathbf{C})/\ell(\widehat{\boldsymbol{\theta}}^{(k)}|\mathbf{V}, \mathbf{C}) - 1| < \epsilon$, with $\epsilon = 10^{-6}$, i.e., the algorithm stops when the relative distance between two successive evaluations of the log-likelihood is less than the tolerance.

3.3 Details for the steps in the ECM algorithm

The E-step reduces only to the computation of $\mathrm{vec}(\widehat{\mathcal{X}}_i^{(k)}) = \mathbb{E}_{\boldsymbol{\mathcal{X}}_i}[\mathrm{vec}(\boldsymbol{\mathcal{X}}_i) \mid \widetilde{\mathcal{V}}_i, \widetilde{\mathcal{C}}_i, \widehat{\boldsymbol{\Theta}}^{(k)}]$ and $\widehat{\boldsymbol{\Omega}}_i^{(k)} = \mathbb{C}ov_{\boldsymbol{\mathcal{X}}_i}[\mathrm{vec}(\boldsymbol{\mathcal{X}}_i) \mid \widetilde{\mathcal{V}}_i, \widetilde{\mathcal{C}}_i, \widehat{\boldsymbol{\Theta}}^{(k)}]$, that is the first and second moments of a truncated MN distribution. These can be determined in closed

form, as a function of MN probabilities using a sequence of simple transformations, for which we use the `MomTrunc` package in `R` (Galarza et al. 2022). For more details on the computation of these moments, refer to Vaida and Liu (2009); Matos et al. (2013); Lachos et al. (2017) and Valeriano et al. (2023).

The update in (9) follows from both (8) and the result given, for instance, in Glanz and Carvalho (2018). To obtain the forms in (10) and (11), first note that

$$
\begin{aligned}
&Q(\{\mathbf{M}^{(k+1)}, \boldsymbol{\Sigma}, \boldsymbol{\Psi}\} \mid \widehat{\boldsymbol{\Theta}}^{(k)}) \\
&= T - \frac{np}{2}\log(|\boldsymbol{\Psi}|) - \frac{nq}{2}\log(|\boldsymbol{\Sigma}|) \\
&\quad - \frac{1}{2}\sum_{i=1}^{n} \mathrm{tr}\left[(\boldsymbol{\Psi}\otimes\boldsymbol{\Sigma})^{-1}\widehat{\mathbf{L}}_i^{(k)}\widehat{\mathbf{L}}_i^{(k)\top}\right] \\
&= T - \frac{np}{2}\log(|\boldsymbol{\Psi}|) - \frac{nq}{2}\log(|\boldsymbol{\Sigma}|) \\
&\quad - \frac{1}{2}\sum_{i=1}^{n} \mathrm{tr}\left[\widehat{\mathbf{L}}_i^{(k)\top}(\boldsymbol{\Psi}\otimes\boldsymbol{\Sigma})^{-1}\widehat{\mathbf{L}}_i^{(k)}\right] \\
&= T - \frac{np}{2}\log(|\boldsymbol{\Psi}|) - \frac{nq}{2}\log(|\boldsymbol{\Sigma}|) \\
&\quad - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{pq} \left[\widehat{\mathbf{L}}_{ij}^{(k)\top}(\boldsymbol{\Psi}\otimes\boldsymbol{\Sigma})^{-1}\widehat{\mathbf{L}}_{ij}^{(k)}\right] \\
&= T - \frac{np}{2}\log(|\boldsymbol{\Psi}|) - \frac{nq}{2}\log(|\boldsymbol{\Sigma}|) \\
&\quad - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{pq} \left[\mathrm{vec}(\widehat{\mathbf{B}}_{ij}^{(k)})^{\top}(\boldsymbol{\Psi}\otimes\boldsymbol{\Sigma})^{-1}\mathrm{vec}(\widehat{\mathbf{B}}_{ij}^{(k)})\right] \\
&= T - \frac{np}{2}\log(|\boldsymbol{\Psi}|) - \frac{nq}{2}\log(|\boldsymbol{\Sigma}|) \\
&\quad - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{pq} \mathrm{tr}\left[\boldsymbol{\Sigma}^{-1}\widehat{\mathbf{B}}_{ij}^{(k)}\boldsymbol{\Psi}^{-1}\widehat{\mathbf{B}}_{ij}^{\top(k)}\right],
\end{aligned}
$$

where we worked through known identities of the vec operator and Kronecker product (see, Glanz and Carvalho 2018, Subsection2.1), after replacing $\mathbf{M}$ by its estimate in (9), and T is a quantity that involves neither $\boldsymbol{\Sigma}$ nor $\boldsymbol{\Psi}$. This representation simplifies the matrix derivatives with respect to $\boldsymbol{\Sigma}$ and $\boldsymbol{\Psi}$, leading to the forms given in (10) and (11), respectively - see also Glanz and Carvalho (2018).

4 Numerical studies

In this section, we provide numerical studies based on both simulated data (Sect. 4.1) and real data concerning water quality monitoring (Sect. 4.2).

4.1 Simulated data

Firstly, we simulate data from an MVN distribution having $p = 3$, $q = 4$, and the following parameters

$$
\mathbf{M} = \begin{pmatrix} 3.00 & 6.00 & 3.00 & 6.00 \\ 6.00 & 6.00 & 3.00 & 3.00 \\ 9.00 & 9.00 & 6.00 & 9.00 \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} 1.50 & 0.60 & 0.24 \\ 0.60 & 1.50 & 0.60 \\ 0.24 & 0.60 & 1.50 \end{pmatrix},
$$

$$
\boldsymbol{\Psi} = \begin{pmatrix} 2.15 & 1.72 & 1.38 & 1.10 \\ 1.72 & 2.15 & 1.72 & 1.38 \\ 1.38 & 1.72 & 2.15 & 1.72 \\ 1.10 & 1.38 & 1.72 & 2.15 \end{pmatrix}.
$$

Then, for each dataset, we randomly take a certain percentage c of the pqn simulated values, where $c \in \{5\%, 15\%, 30\%\}$, and replace them in a way to encompass the following $s = 5$ scenarios:

1. *Only interval-censored values*: for each selected value, two numbers, namely a and b, are generated from a uniform distribution, such that $c \in (a, b)$, and considered as the lower and upper limit of the censoring interval.
2. *Both missing and interval-censored values*: 25% of the selected values are replaced by using the strategy described in point 5, and the remaining 75% are replaced via the strategy reported in point 1;
3. *Both missing and interval-censored values*: 50% of the selected values are replaced by using the strategy described in point 5, and the remaining 50% are replaced via the strategy reported in point 1;
4. *Both missing and interval-censored values*: 75% of the selected values are replaced by using the strategy described in point 5, and the remaining 25% are replaced via the strategy reported in point 1.
5. *Only missing values*: the selected values are substituted with $\pm\infty$ (refer to Sect. 3);

We also consider three levels for the sample size, i.e. $n \in \{100, 200, 400\}$. Therefore, by combining the experimental factors (c, s, and n), we obtain $3 \times 5 \times 3 = 45$ data configurations. Note that, for each data configuration, we generate 100 datasets, resulting in a total of 4500 simulated datasets.

On each simulated dataset, we fit our MVNC model and evaluate its capability to recover the true data-generating parameters. To this aim, on each simulated dataset, we compute the Frobenius norm for $\mathbf{M}$ as

$$
\widetilde{\mathbf{M}} = ||\widehat{\mathbf{M}}_m - \mathbf{M}||_F,
$$

and the following normalized loss functions for $\boldsymbol{\Sigma}$ and $\boldsymbol{\Psi}$,

$$
\widetilde{\boldsymbol{\Sigma}} = \frac{1}{\sqrt{p}}||\boldsymbol{\Sigma}^{-1/2}(\widehat{\boldsymbol{\Sigma}}_m - \boldsymbol{\Sigma})\boldsymbol{\Sigma}^{-1/2}||_F \quad \text{and}
$$

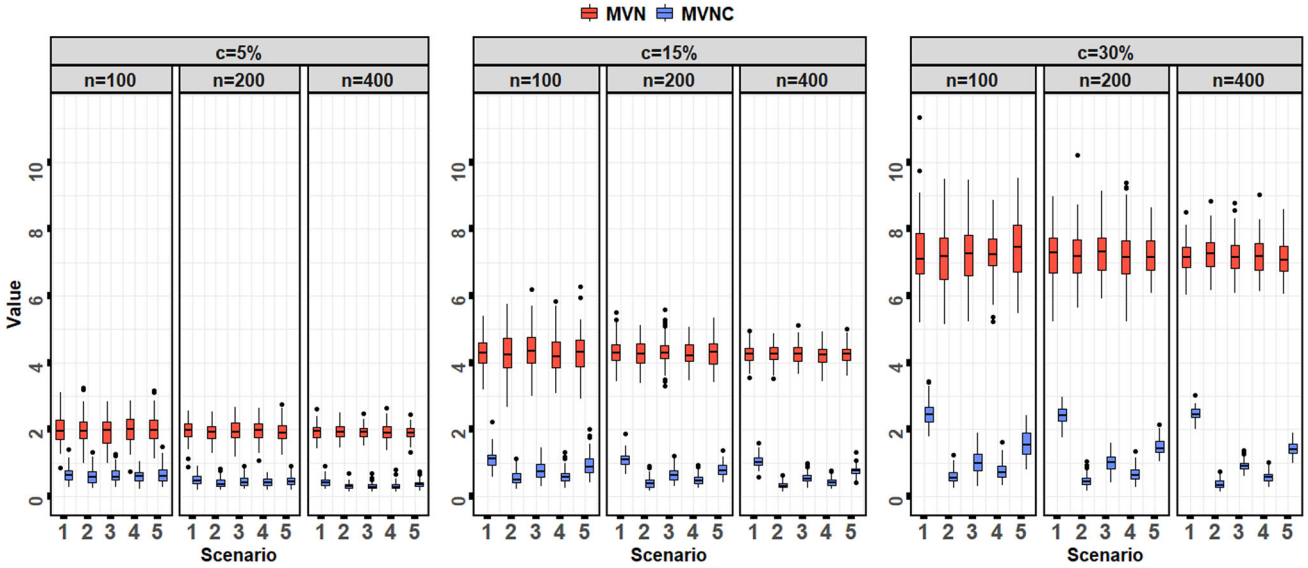

Fig. 1 Box-plots of $\widetilde{\mathbf{M}}$ over 100 simulated datasets for a given (c, n) pair and scenario

$$\widetilde{\Psi} = \frac{1}{\sqrt{q}} ||\Psi^{-1/2}(\widehat{\Psi}_m - \Psi)\Psi^{-1/2}||_F,$$

where $\widehat{\mathbf{M}}_m$, $\widehat{\Sigma}_m$, and $\widehat{\Psi}_m$ are the corresponding estimates from the mth dataset.

Given the absence of alternative models capable of managing data with the above-mentioned characteristics in the matrix-variate context (see Sect. 1), we implement the approach consisting of removing any observation presenting these values as an alternative procedure. Consequently, after cleaning each dataset, we proceed to fit the classical MVN distribution.

4.1.1 Results

Results are separately reported for each parameter using box-plots from Figs. 1, 2 and 3, respectively. Each figure illustrates the calculated differences for the MVNC and the MVN approaches for each (c, n) pair over the five scenarios. Specifically, for a given (c, n) pair and scenario, the box-plots summarize the results over 100 simulated datasets.

The results across the three figures reveal important insights. First, the MVNC model consistently outperforms the MVN model across all scenarios and parameters. Importantly, this discrepancy widens as c increases. As the proportion of missing or censored data grows, the MVN model's performance deteriorates more sharply due to its reliance on discarding incomplete observations, which reduces the effective sample size. In contrast, the MVNC model leverages all available data, leading to more accurate estimates.

A second facet to highlight is that the performance gap between the two models does not follow a uniform pattern across scenarios. For a given (c, n) pair, the performance of the MVN model remains relatively similar across the five scenarios. However, for the MVNC model, the estimation error is generally smaller in mixed scenarios (like Scenarios 2–4) than in scenarios with purely missing or censored data (Scenarios 1 and 5). This indicates that the MVNC model is particularly effective in dealing with a combination of missing and censored values, further boosting its advantage over MVN in such cases.

Finally, the quantities under evaluation improve as n increases, demonstrating the consistency of the estimators. The MVNC model's ability to incorporate incomplete data allows it to capture the underlying structure better as the sample size grows, whereas the MVN model suffers from the reduction in effective sample size, leading to more variability and estimation errors, particularly when n is small.

4.2 Real data case studies

Here, we consider the `dataCensored` dataset contained in `R` package `baytrends` (Murphy et al. 2023) and introduced earlier in Murphy et al. (2019). Specifically, it contains water quality variables for eight stations from the Chesapeake Bay monitoring program. More in detail, at each station, p variables are measured at q layers on n occasions, thus producing a $p \times q \times n$ dataset. In this study, our attention is directed toward two specific stations that exhibit interval-censored and missing data, albeit in varying proportions. The objective is to explore two distinct real-data scenarios concerning two stations labeled CB5.4 and EE2.1, respectively.

We begin our analysis with the CB5.4 station. For illustrative purposes, we consider the following $p = 3$ variables: orthophosphorus (`po4`), dissolved inorganic nitrogen (`din`),

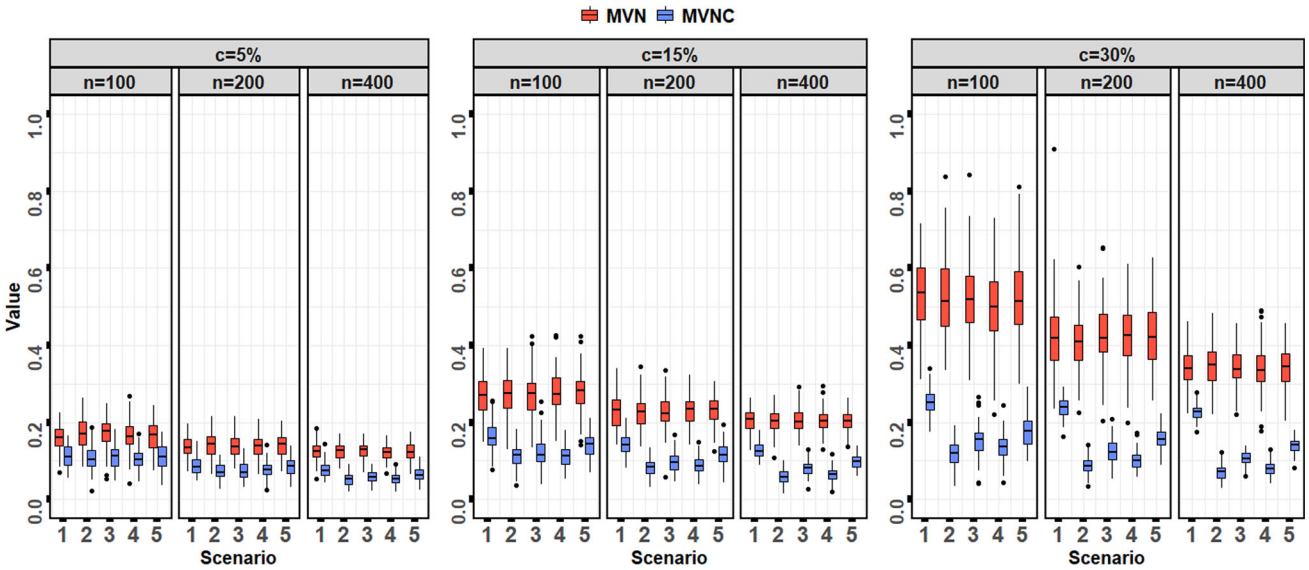

Fig. 2 Box-plots of $\widetilde{\boldsymbol{\Sigma}}$ over 100 simulated datasets for a given (c, n) pair and scenario

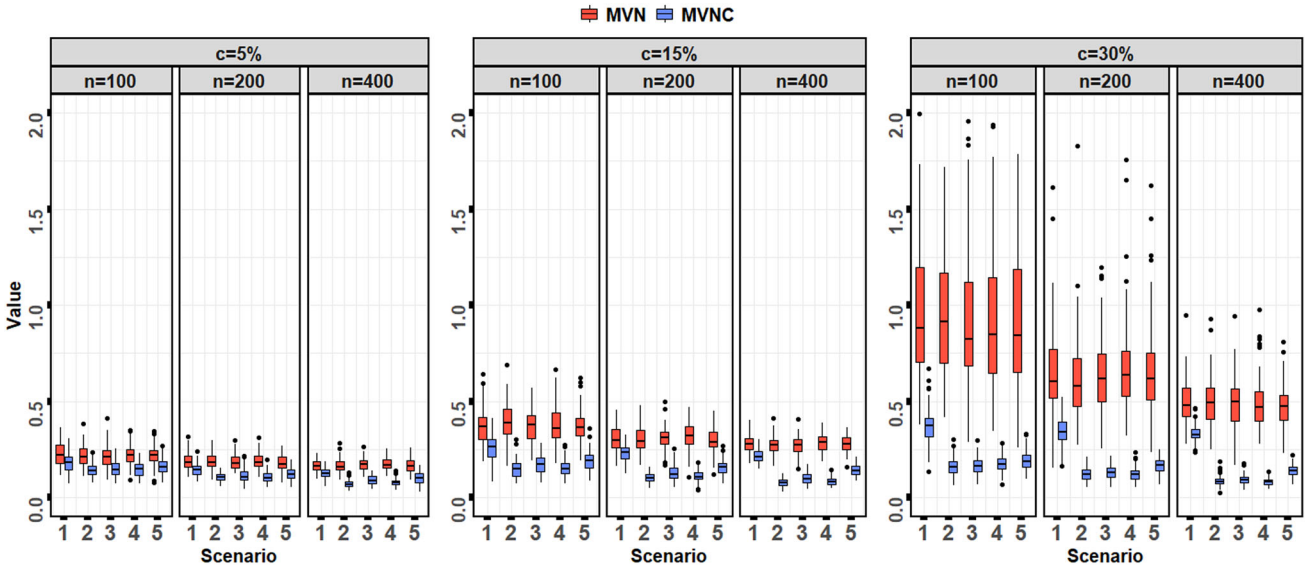

Fig. 3 Box-plots of $\widetilde{\boldsymbol{\Psi}}$ over 100 simulated datasets for a given (c, n) pair and scenario

and ammonium (nh4). These variables present the highest levels of interval-censored and missing values in the dataset. Specifically, po4 has 14.93% of missing values and 1.60% of interval-censored values, totaling 16.56%. For din, the proportions are 13.11% missing and 0.94% interval-censored, resulting in a total of 14.05%. Lastly, nh4 shows proportions of 7.80% missing and 0.72% interval-censored values, totaling 8.52%. Overall, we easily note that the proportion of missing values is far higher than that of interval-censored.

The variables are measured at $q = 4$ layers: surface (S), above-pycnocline (AP), below-pycnocline (BP), and bottom (B). The resulting 3×4 matrices are then evaluated on $n = 452$ occasions. Thus, the final structure of the first dataset is $3 \times 4 \times 452$.

Our second analysis concerns the EE2.1 station. Here, we take into account the following $p = 3$ variables: po4, total dissolved nitrogen (tdn), and total dissolved phosphorus (tdp). Similar to our earlier discussion, these variables exhibit the highest levels of interval-censored and missing values within this dataset. In detail, the variable po4 displays 1.81% missing values and 26.22% interval-censored values, resulting in a combined percentage of 28.03%. For din, the corresponding proportions are 5.21% missing and 7.39% interval-censored, resulting in a total of 12.60%. Finally, nh4 demonstrates proportions of 5.64% missing and 7.18%

interval-censored values, yielding a cumulative percentage of 12.82%. As it is evident, and in contrast to the preceding dataset, the presence of interval-censored values is notably higher than that of missing values.

These variables are measured across the same $q = 4$ layers previously outlined and are evaluated over $n = 470$ occasions. Thus, the second dataset is a $3 \times 4 \times 470$ array.

4.2.1 Results for the CB5.4 station

We start by presenting the results obtained by fitting the MVNC model to the first dataset. In particular, the estimated mean matrix $\mathbf{M}$ is

$$\mathbf{M} = \begin{array}{c} \\ \texttt{po4} \\ \texttt{din} \\ \texttt{nh4} \end{array} \begin{array}{c} \begin{array}{cccc} \texttt{S} & \texttt{AP} & \texttt{BP} & \texttt{B} \end{array} \\ \begin{bmatrix} 0.002 & 0.003 & 0.009 & 0.014 \\ 0.108 & 0.105 & 0.112 & 0.131 \\ 0.018 & 0.026 & 0.058 & 0.088 \end{bmatrix} \end{array}.$$

Notably, we observe a rise in the average values of both po4 and nh4 as they are assessed from the surface to the bottom. This trend could indicate that these substances are either being produced or accumulating at lower depths. Possible explanations might involve biological processes, such as the decomposition of organic matter or the release of these compounds from sediments. The accumulation of these nutrients at greater depths may influence biological productivity and cause eutrophication (Wang et al. 2022).

For din, the trend is similar, with the exception that its mean values decrease when moving from the surface to above the pycnocline. The pycnocline is a layer of rapidly changing density with depth in the ocean. The decrease in din above the pycnocline could be due to processes such as biological uptake by phytoplankton or other organisms in the upper water column.

To understand the relationships between the variables and among the different levels of depth, we now report the correlation matrices $\mathbf{R}(\cdot)$ related to $\boldsymbol{\Sigma}$ and $\boldsymbol{\Psi}$, that is,

$$\mathbf{R}(\boldsymbol{\Sigma}) = \begin{array}{c} \\ \texttt{po4} \\ \texttt{din} \\ \texttt{nh4} \end{array} \begin{array}{c} \begin{array}{ccc} \texttt{po4} & \texttt{din} & \texttt{nh4} \end{array} \\ \begin{bmatrix} 1.000 & 0.375 & 0.567 \\ 0.375 & 1.000 & 0.472 \\ 0.567 & 0.472 & 1.000 \end{bmatrix} \end{array} \quad \text{and}$$

$$\mathbf{R}(\boldsymbol{\Psi}) = \begin{array}{c} \\ \texttt{S} \\ \texttt{AP} \\ \texttt{BP} \\ \texttt{B} \end{array} \begin{array}{c} \begin{array}{cccc} \texttt{S} & \texttt{AP} & \texttt{BP} & \texttt{B} \end{array} \\ \begin{bmatrix} 1.000 & 0.840 & 0.442 & 0.316 \\ 0.840 & 1.000 & 0.555 & 0.419 \\ 0.442 & 0.555 & 1.000 & 0.821 \\ 0.316 & 0.419 & 0.821 & 1.000 \end{bmatrix} \end{array}.$$

Starting with $\mathbf{R}(\boldsymbol{\Sigma})$, we note there is a positive correlation between all the variables. To elaborate further, the positive correlation between po4 and din can be elucidated by a shared origin, such as agricultural runoff or the decomposition of organic matter. In the case of po4 and nh4, the positive correlation indicates that processes enhancing the availability of one nutrient may concurrently influence the other. Notably, both are commonly associated with the decomposition of organic matter and wastewater discharge. Turning to the correlation between din and nh4, the positive value implies that elevated ammonium levels may lead to increased microbial activity, consequently giving rise to the production of various forms of dissolved inorganic nitrogen. To sum up, the observed correlations can be rationalized by the intricate interplay of biological and chemical processes in aquatic ecosystems. Nutrient cycling, microbial activity, and chemical reactions collectively contribute to the discernible patterns in the correlation matrix. Understanding these relationships is paramount for the evaluation of water quality and the overall health of ecosystems.

Now, considering $\mathbf{R}(\boldsymbol{\Psi})$, we note that the correlations suggest varying degrees of influence and connection between measurements at different lake depths. For example, surface conditions (S) seem to strongly influence conditions just above the pycnocline (AP), and to a lesser extent, conditions below the pycnocline (BP) and at the bottom (B). These results also apply when the other pairs of layers are considered. Processes like vertical mixing, sedimentation, and biological activity likely contribute to the observed correlations.

To show the effectiveness of our proposal, we now consider the alternative approach discussed in Sect. 4.1, consisting of removing any observation presenting interval-censored and/or missing values. Specifically, once the dataset is cleaned, the new sample size is $n = 299$. Thus, we would lose one-third of the information in the data if our approach would not be available. On this data, we fit the classical MVN distribution, and we report the percentage change $\%(\cdot)$ between our and the parameters estimated using the classical MVN, to assess potential differences in the results. Specifically, we obtain

$$\%(\mathbf{M}) = \begin{array}{c} \\ \texttt{po4} \\ \texttt{din} \\ \texttt{nh4} \end{array} \begin{array}{c} \begin{array}{cccc} \texttt{S} & \texttt{AP} & \texttt{BP} & \texttt{B} \end{array} \\ \begin{bmatrix} -31.579 & -20.728 & -2.470 & -3.982 \\ -14.728 & -15.092 & -7.781 & -6.617 \\ -29.130 & -19.123 & -2.916 & -4.324 \end{bmatrix} \end{array},$$

$$\%(\mathbf{R}(\boldsymbol{\Sigma})) = \begin{array}{c} \\ \texttt{po4} \\ \texttt{din} \\ \texttt{nh4} \end{array} \begin{array}{c} \begin{array}{ccc} \texttt{po4} & \texttt{din} & \texttt{nh4} \end{array} \\ \begin{bmatrix} 0.000 & -9.282 & -6.861 \\ -9.282 & 0.000 & 1.612 \\ -6.861 & 1.612 & 0.000 \end{bmatrix} \end{array},$$

and $\%(\mathbf{R}(\boldsymbol{\Psi}))$

$$= \begin{array}{c} \\ \texttt{S} \\ \texttt{AP} \\ \texttt{BP} \\ \texttt{B} \end{array} \begin{array}{c} \begin{array}{cccc} \texttt{S} & \texttt{AP} & \texttt{BP} & \texttt{B} \end{array} \\ \begin{bmatrix} 0.000 & -0.667 & -10.454 & -8.290 \\ -0.667 & 0.000 & -2.832 & -4.189 \\ -10.454 & -2.832 & 0.000 & 3.399 \\ -8.290 & -4.189 & 3.399 & 0.000 \end{bmatrix} \end{array}.$$

As we note, there is a constant underestimation of the values in $\mathbf{M}$, which is particularly high for the first two layers. Regarding $\mathbf{R}(\boldsymbol{\Sigma})$, a negative and relatively higher underestimation is also noted in the correlations between po4 and din, whereas there is a slight overestimation of the correlation between din and nh4. Lastly, we note differences spanning from -10.454 to 3.399 in the correlations between the four layers in $\mathbf{R}(\boldsymbol{\Psi})$. Consequently, we can infer that if all available data information had not been utilized, we would have obtained relatively different estimates. This issue can be circumvented by employing our methodology, which comprehensively incorporates all available data.

4.2.2 Results for the EE2.1 station

As concerns the second dataset, we start by providing the estimated parameters by the MVNC model. In particular, the estimated mean matrix $\mathbf{M}$ is

$$\mathbf{M} = \begin{array}{c} \\ \texttt{po4} \\ \texttt{tdn} \\ \texttt{tdp} \end{array} \begin{array}{c} \begin{array}{cccc} \texttt{S} & \texttt{AP} & \texttt{BP} & \texttt{B} \end{array} \\ \begin{bmatrix} 0.005 & 0.004 & 0.005 & 0.005 \\ 0.545 & 0.533 & 0.532 & 0.537 \\ 0.017 & 0.017 & 0.018 & 0.018 \end{bmatrix} \end{array}.$$

We note that, differently from the previous station, the mean concentrations of po4 are relatively stable across all layers. Being this variable a key nutrient for aquatic organisms, including algae and aquatic plants, the uniformity in concentrations may indicate a relatively steady supply and demand for po4 in the studied ecosystem. The (overall) decreasing trend in mean concentrations of tdn from the surface (S) to the bottom (B), could be indicative of microbial nitrogen transformations as water descends through different layers. The variation in tdn concentrations could impact nutrient availability and affect the overall nitrogen cycle in the aquatic environment. Lastly, the slight increase in mean concentrations of tdp from the surface to the bottom may be associated with sediment interactions, where phosphorus is released from the sediments or adsorbed onto particles settling to the bottom. This could have implications for phosphorus availability and cycling in the water column. Additionally, biological processes such as phosphorus uptake by organisms could contribute to the observed variations.

To understand the relationships between the variables and among the different levels of depth, we now report the correlation matrices related to $\boldsymbol{\Sigma}$ and $\boldsymbol{\Psi}$, that is,

$$\mathbf{R}(\boldsymbol{\Sigma}) = \begin{array}{c} \\ \texttt{po4} \\ \texttt{tdn} \\ \texttt{tdp} \end{array} \begin{array}{c} \begin{array}{ccc} \texttt{po4} & \texttt{tdn} & \texttt{tdp} \end{array} \\ \begin{bmatrix} 1.000 & 0.073 & 0.204 \\ 0.073 & 1.000 & 0.095 \\ 0.204 & 0.095 & 1.000 \end{bmatrix} \end{array} \quad \text{and}$$

$$\mathbf{R}(\boldsymbol{\Psi}) = \begin{array}{c} \\ \texttt{S} \\ \texttt{AP} \\ \texttt{BP} \\ \texttt{B} \end{array} \begin{array}{c} \begin{array}{cccc} \texttt{S} & \texttt{AP} & \texttt{BP} & \texttt{B} \end{array} \\ \begin{bmatrix} 1.000 & 0.885 & 0.817 & 0.793 \\ 0.885 & 1.000 & 0.858 & 0.827 \\ 0.817 & 0.858 & 1.000 & 0.872 \\ 0.793 & 0.827 & 0.872 & 1.000 \end{bmatrix} \end{array}.$$

Starting with $\mathbf{R}(\boldsymbol{\Sigma})$, we note there is a positive correlation between all the variables. In detail, the weak positive correlation between po4 and tdn implies a limited biological association. Biological and chemical processes affecting one of these nutrients may not strongly influence the other. In the case of po4 and tdp, we note a moderate positive correlation. Common biological processes or sources may contribute to the observed correlation, such as sediment interactions or biological uptake. Regarding the correlation between tdn and tdp, the weak positive correlation implies that the factors influencing nitrogen and phosphorus concentrations may have limited overlap biologically.

Regarding $\mathbf{R}(\boldsymbol{\Psi})$, we note that, similarly to the previous station, the correlations suggest varying degrees of influence and connection between measurements at different lake depths. In particular, layers that are progressively more distant imply smaller correlations. However, the rate of decrease between one layer and another is significantly lower at this station than in the previous one. Thus, a stronger persistence and similarities between the four water layers are here observed.

Also at this station, we compare our parameter estimates with those obtained by discarding observations having interval-censored and/or missing values. The new sample has a size of $n = 270$. Therefore, almost half of the information in the data would be lost. On this data, the classical MVN distribution is fitted and, as for the previous station, its estimated parameters are compared (via the percentage change) with those of our proposal to explore potential differences. In detail, we obtain

$$\%(\mathbf{M}) = \begin{array}{c} \\ \texttt{po4} \\ \texttt{tdn} \\ \texttt{tdp} \end{array} \begin{array}{c} \begin{array}{cccc} \texttt{S} & \texttt{AP} & \texttt{BP} & \texttt{B} \end{array} \\ \begin{bmatrix} -5.601 & 1.531 & 3.446 & 2.287 \\ -3.346 & -3.723 & -3.544 & -3.320 \\ -17.771 & -19.176 & -18.508 & -21.177 \end{bmatrix} \end{array},$$

$$\%(\mathbf{R}(\boldsymbol{\Sigma})) = \begin{array}{c} \\ \texttt{po4} \\ \texttt{tdn} \\ \texttt{tdp} \end{array} \begin{array}{c} \begin{array}{ccc} \texttt{po4} & \texttt{tdn} & \texttt{tdp} \end{array} \\ \begin{bmatrix} 0.000 & 33.430 & 69.178 \\ 33.430 & 0.000 & -20.000 \\ 69.178 & -20.000 & 0.000 \end{bmatrix} \end{array},$$

and $\%(\mathbf{R}(\boldsymbol{\Psi}))$

$$= \begin{array}{c} \\ \texttt{S} \\ \texttt{AP} \\ \texttt{BP} \\ \texttt{B} \end{array} \begin{array}{c} \begin{array}{cccc} \texttt{S} & \texttt{AP} & \texttt{BP} & \texttt{B} \end{array} \\ \begin{bmatrix} 0.000 & -2.129 & -2.530 & -1.644 \\ -2.129 & 0.000 & -5.067 & -4.108 \\ -2.530 & -5.067 & 0.000 & -0.374 \\ -1.644 & -4.108 & -0.374 & 0.000 \end{bmatrix} \end{array}.$$

We immediately note that there is a regular underestimation of the average tdn and tdp values in $\mathbf{M}$. In

particular, for the tdp variable, this negative difference is constantly around 20%. Regarding $\mathbf{R}(\boldsymbol{\Sigma})$, we observed great divergences among the two approaches, regardless of the considered variable. Finally, we notice a constant underestimation of the correlations in $\mathbf{R}(\boldsymbol{\Psi})$. To sum up, even at this station, we appreciate the advantages of using our methodology, given that all the data information can be used. Much of this information would be lost without the introduced approach.

5 Conclusions

In this paper, we introduce an approach for modeling data with interval-censored and missing values within a matrix-variate framework. The current literature on matrix-variate models lacks statistical approaches that simultaneously address these two features. Thus, our proposed method introduces a new avenue for modeling such data. We rely on the matrix-variate normal distribution and its mathematical properties to conceptualize and derive an ECM algorithm for parameter estimation.

Simulated analyses demonstrate the reliability of our approach in providing accurate parameter estimates across various scenarios with different proportions of interval-censored and missing values, as well as varying sample sizes. Additionally, we compare our results with an alternative method that involves excluding observations with these characteristics. As it might be reasonable to expect, the alternative method leads to parameter estimates with greater variability and lower precision.

Furthermore, we apply our method to two real-data datasets focusing on the concentrations of various chemical components in water at two measurement stations. Both applications involve interval-censored and missing values, though in different proportions. The results offer valuable insights into the water quality features at the respective locations. When contrasted with the alternative approach used in the simulation study, our method demonstrates the importance of considering all data characteristics. Excluding such information leads to under/overestimation of parameters, yielding different results compared to our comprehensive approach.

The approach presented in this paper has the potential for extension in various directions. One possible avenue is exploring a model-based clustering method that accommodates the presence of latent clusters in data containing interval-censored and missing values. Furthermore, there is potential for extension in a regression setting, where consideration is given to a response variable exhibiting the aforementioned characteristics. Finally, to allow for time dependence, an extension to the hidden Markov setting can be considered.

Acknowledgements This paper was written while Víctor H. Lachos was visiting the Department of Economics and Business at the University of Catania, Italy. Victor H. Lachos acknowledges the partial financial support from the Office of the Vice President for Research (OVPR) Research Excellence Program (REP) and UConn - CLAS's Summer Research Funding Initiative 2023. Salvatore Ingrassia, Antonio Punzo, and Salvatore D. Tomarchio have been partially supported by MUR, grant no. 2022XRHT8R (CUP: E53D23005950006) - The SMILE project: Statistical Modelling and Inference to Live the Environment, funded by the European Union - Next Generation EU.

Funding Open access funding provided by Università degli Studi di Catania within the CRUI-CARE Agreement.

Declarations

Conflict of interest The authors have no conflict of interest to declare.

References

Acock, A.C.: Working with missing values. J. Marriage Fam. **67**(4), 1012–1028 (2005)

Alencar, F.H., Galarza, C.E., Matos, L.A., Lachos, V.H.: Finite mixture modeling of censored and missing data using the multivariate skew-normal distribution. Adv. Data Anal. Classif. **16**(3), 521–557 (2022)

Allen, G.I., Tibshirani, R.: Transposable regularized covariance models with an application to missing data imputation. Ann. Appl. Stat. **4**(2), 764 (2010)

Anderlucci, L., Montanari, A., Viroli, C.: A matrix-variate regression model with canonical states: An application to elderly Danish twins. Statistica (Bologna) **74**(4), 367–381 (2014)

Bahari, F., Parsi, S., Ganjali, M.: Reliability of a soccer player based on the bivariate Rayleigh distribution with right censored and ignorable missing data. J. Appl. Stat. **48**(2), 285–300 (2021)

De Gruttola, V., Lagakos, S.W.: Analysis of doubly-censored survival data, with application to AIDS. Biometrics **45**, 1–11 (1989)

Dutilleul, P.: The MLE algorithm for the matrix normal distribution. J. Stat. Comput. Simul. **64**(2), 105–123 (1999)

Faucheux, L., Resche-Rigon, M., Curis, E., Soumelis, V., Chevret, S.: Clustering with missing and left-censored data: a simulation study comparing multiple-imputation-based procedures. Biom. J. **63**(2), 372–393 (2021)

Galarza Morales, C.E., Lachos, V.H., Bourguignon, M.: A skew-t quantile regression for censored and missing data. Stat **10**(1), 379 (2021)

Galarza, C.E., Kan, R., Lachos, V.H.: MomTrunc: moments of folded and doubly truncated multivariate distributions. (2022). https://CRAN.R-project.org/package=MomTrunc

Galarza, C.E., Matos, L.A., Lachos, V.H.: An EM algorithm for estimating the parameters of the multivariate skew-normal distribution with censored responses. METRON **80**(2), 231–253 (2022)
Gallaugher, M.P., McNicholas, P.D.: Finite mixtures of skewed matrix variate distributions. Pattern Recogn. **80**, 83–93 (2018)
Gallaugher, M.P., McNicholas, P.D.: Mixtures of skewed matrix variate bilinear factor analyzers. Adv. Data Anal. Classif. **14**(2), 415–434 (2020)
Glanz, H., Carvalho, L.: An expectation-maximization algorithm for the matrix normal distribution with an application in remote sensing. J. Multivar. Anal. **167**, 31–48 (2018)
Gupta, A.K., Nagar, D.K.: Matrix variate distributions. Chapman and Hall/CRC, Boca Raton (1999)
Helsel, D.R.: Statistics for censored environmental data using Minitab and R, vol. 77. John Wiley & Sons, Hoboken (2011)
Lachos, V.H., Bandyopadhyay, D., Dey, D.K.: Linear and nonlinear mixed-effects models for censored HIV viral loads using normal/independent distributions. Biometrics **67**(4), 1594–1604 (2011)
Lachos, V.H., Moreno, E.J.L., Chen, K., Cabral, C.R.B.: Finite mixture modeling of censored data using the multivariate Student-t distribution. J. Multivar. Anal. **159**, 151–167 (2017)
Lin, T.-I., Lachos, V.H., Wang, W.-L.: Multivariate longitudinal data analysis with censored and intermittent missing responses. Stat. Med. **37**(19), 2822–2835 (2018)
Matos, L.A., Prates, M.O., Chen, M.H., Lachos, V.H.: Likelihood-based inference for mixed-effects models with censored response using the multivariate-t distribution. Stat. Sin. **23**, 1323–1342 (2013)
McLachlan, G.J., Krishnan, T.: The EM algorithm and extensions. John Wiley & Sons, Hoboken (2008)
Melnykov, V., Zhu, X.: On model-based clustering of skewed matrix data. J. Multivar. Anal. **167**, 181–194 (2018)
Meng, X.-L., Rubin, D.B.: Maximum likelihood estimation via the ECM algorithm: a general framework. Biometrika **80**(2), 267–278 (1993)
Murphy, R., Perry, E., Keisman, J., Harcum, J., Leppo, E.W.: Baytrends: long term water quality trend analysis. (2023). R package version 2.0.9. https://CRAN.R-project.org/package=baytrends
Murphy, R.R., Perry, E., Harcum, J., Jennifer, K.: A generalized additive model approach to evaluating water quality: Chesapeake bay case study. Environ. Modell. Softw. **118**, 1–13 (2019)
Sarkar, S., Zhu, X., Melnykov, V., Ingrassia, S.: On parsimonious models for modeling matrix data. Comput. Stat. Data Anal. **142**, 106822 (2020)
Thompson, G.Z., Maitra, R., Meeker, W.Q., Bastawros, A.F.: Classification with the matrix-variate-t distribution. J. Comput. Graph. Stat. **29**(3), 668–674 (2020)
Tomarchio, S.D.: Matrix-variate normal mean-variance Birnbaum-Saunders distributions and related mixture models. Comput. Stat. **39**(2), 405–432 (2024)
Tomarchio, S.D., McNicholas, P.D., Punzo, A.: Matrix normal cluster-weighted models. J. Classif. **38**(3), 556–575 (2021)
Tomarchio, S.D., Punzo, A., Bagnato, L.: On the use of the matrix-variate tail-inflated normal distribution for parsimonious mixture modeling. In: Salvati, N., Perna, C., Marchetti, S., Chambers, R. (eds.) Studies in theoretical and applied statistics, pp. 407–423. Springer, Cham (2022)
Tomarchio, S.D., Punzo, A., Maruotti, A.: Parsimonious hidden Markov models for matrix-variate longitudinal data. Stat. Comput. **32**(3), 53 (2022)
Tomarchio, S.D., Punzo, A., Maruotti, A.: Matrix-variate hidden Markov regression models: fixed and random covariates. J. Classif. **41**, 429–454 (2024)
Triantafyllopoulos, K.: Missing observation analysis for matrix-variate time series data. Stat. Probab. Lett. **78**(16), 2647–2653 (2008)
Vaida, F., Liu, L.: Fast implementation for normal mixed effects models with censored response. J. Comput. Graph. Stat. **18**, 797–817 (2009)
Valeriano, K.A., Lachos, V.H., Prates, M.O., Matos, L.A.: Likelihood-based inference for spatiotemporal data with censored and missing responses. Environmetrics **32**(3), 2663 (2021)
Valeriano, K.A., Galarza, C.E., Matos, L.A., Lachos, V.H.: Likelihood-based inference for the multivariate skew-t regression with censored or missing responses. J. Multivar. Anal. **196**, 105174 (2023)
Viroli, C.: Finite mixtures of matrix normal distributions for classifying three-way data. Stat. Comput. **21**, 511–522 (2011)
Viroli, C.: On matrix-variate regression analysis. J. Multivar. Anal. **111**, 296–309 (2012)
Wang, Z., Huang, Z., Zheng, B., Wu, D., Zheng, S.: Efficient removal of phosphate and ammonium from water by mesoporous tobermorite prepared from fly ash. J. Environ. Chem. Eng. **10**(3), 107400 (2022)
Zhang, L., Bandyopadhyay, D.: A graphical model for skewed matrix-variate non-randomly missing data. Biostatistics **21**(2), 80–97 (2020)