UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Matrix-Variate Skew-Normal and Censored Skew-Normal Models: Theory, Inference, and ECM Estimation

Átila Prates Correia

Tese de Doutorado do Programa Interinstitucional de Pós-Graduação em Estatística (PIPGEs)

ICMC USP SÃO CARLOS

Átila Prates Correia

Matrix-Variate Skew-Normal and Censored Skew-Normal Models: Theory, Inference, and ECM Estimation

Thesis submitted to the Institute of Mathematics and Computer Science – ICMC-USP and to the Department of Statistics – DEs-UFSCar – in accordance with the requirements of the Statistics Interagency Graduate Program, for the degree of Doctor in Statistics. *FINAL VERSION*

Concentration Area: Statistics

Advisor: Prof. Dr. Carlos Alberto Ribeiro Diniz

USP – São Carlos
February 2026

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

Prates Correia, Átila
P912m Matrix-Variate Skew-Normal and Censored Skew-Normal Models: Theory, Inference, and ECM Estimation / Átila Prates Correia; orientador Carlos Alberto Ribeiro Diniz. -- São Carlos, 2026.
135 p.

Tese (Doutorado - Programa Interinstitucional de Pós-graduação em Estatística) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2026.

1. Matrix Variate Skew-Normal Distribution. 2. Censored MVSN. 3. Asymmetric matrix models. 4. ECM algorithm. 5. Censored/Missing data. I. Ribeiro Diniz, Carlos Alberto, orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

UFSCar

UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Ciências Exatas

Programa Interinstitucional de Pós-Graduação em Estatística

Folha de Aprovação

Defesa de Tese de Doutorado do(a) candidato(a) Átila Prates Correia, realizada em 06/02/2026.

Comissão Julgadora:

Prof(a). Dr(a). Carlos Alberto Ribeiro Diniz (DEs-UFSCar)

Prof(a). Dr(a). Daiane Aparecida Zuanetti (DEs-UFSCar)

Prof(a). Dr(a). Jalmar Manuel Farfan Carrasco (UFBA)

Prof(a). Dr(a). Clecio da Silva Ferreira (UFJF)

Prof(a). Dr(a). Victor Hugo Lachos Dávila (UConn)

O Relatório de Defesa produzido pelos membros da Comissão Julgadora encontra-se arquivado junto ao Programa de Pós-Graduação.

Átila Prates Correia

Modelos Skew-Normal Matriciais e Skew-Normal Matriciais Censurados: Teoria, Inferência e Estimação via ECM

Tese apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP e ao Departamento de Estatística – DEs-UFSCar, como parte dos requisitos para obtenção do título de Doutor em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística. *VERSÃO REVISADA*

Área de Concentração: Estatística

Orientador: Prof. Dr. Carlos Alberto Ribeiro Diniz

USP – São Carlos
Fevereiro de 2026

Dedico este trabalho à minha família e amigos.

ACKNOWLEDGEMENTS

First and foremost, I express my profound gratitude to my family, whose unconditional support, patience, and understanding were present at every stage of my academic journey. Throughout this path, marked by challenges, uncertainties, and long periods of dedication to study and research, the constant presence of my family — despite the inevitable absences imposed by academic life — proved to be essential. The encouragement, affection, and confidence they demonstrated were fundamental in enabling me to persevere and ultimately complete this work.

I extend my sincere gratitude to Ana Maria Serra for her constant support, for her encouragement during the most challenging moments, and for the consideration she consistently showed throughout this journey. Her persistence in offering words of encouragement, combined with her kindness and tenderness, had a profound impact not only on the development of this work but also on my personal path during this period. Her presence and generosity were fundamental in enabling me to move forward with confidence and serenity.

Je tiens également à remercier chaleureusement mon ami Luis, qui a toujours fait preuve d'une grande gentillesse et d'une attention sincère à mon égard. Par sa présence constante, nos conversations chaque week-end, son écoute attentive et son soutien fidèle, il m'a aidé à traverser cette période exigeante avec davantage de sérénité. Son amitié, empreinte de bienveillance et de générosité, a compté plus que je ne saurais l'exprimer et a rendu ce chemin bien plus léger.

I express my sincere gratitude to my advisor, Carlos, for his academic rigor, always combined with availability and attentiveness throughout the entire development of this work. His scientific rigor, together with his openness to dialogue and constant guidance, was fundamental to the improvement of this study and to my academic development.

I would also like to thank my (informal) co-advisor Victor Hugo Lachos for the many productive academic discussions that helped me grow scientifically and develop my ideas. I am also grateful for his generosity and helpful suggestions, which were important in shaping this work and improving the overall quality of the thesis.

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

RESUMO

CORREIA, A. P. **Modelos Skew-Normal Matriciais e Skew-Normal Matriciais Censurados: Teoria, Inferência e Estimação via ECM**. 2026. 135 p. Tese (Doutorado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2026.

Esta tese desenvolve um arcabouço abrangente para modelar estruturas de dependência assimétricas em dados matriciais utilizando a distribuição Skew-Normal Matricial (MVSN) e sua extensão censurada. Estabelecemos propriedades teóricas fundamentais desses modelos, incluindo suas representações estocásticas, momentos e condições de identificabilidade. Com base nesses fundamentos, derivamos procedimentos de inferência baseados em verossimilhança e propomos algoritmos de Maximização por Etapas Condicionais (ECM) capazes de lidar com observações matriciais totalmente observadas, intervalarmente censuradas ou parcialmente faltantes. Estudos de simulação são conduzidos para avaliar a recuperação dos parâmetros, o comportamento de convergência e a robustez dos métodos de estimação propostos em diversos cenários. Por fim, demonstramos a utilidade prática dos modelos por meio de aplicações a conjuntos de dados reais com estruturas complexas de censura. Os resultados mostram que os modelos MVSN e MVSN censurado oferecem ferramentas flexíveis e interpretáveis para analisar dados multivariados e matriciais que apresentam assimetria, caudas leves e informações incompletas.

Palavras-chave: Distribuição Skew-Normal Matricial, MVSN censurado, Modelos matriciais assimétricos, Algoritmo ECM, Dados censurados.

ABSTRACT

CORREIA, A. P. **Matrix-Variate Skew-Normal and Censored Skew-Normal Models: Theory, Inference, and ECM Estimation**. 2026. 135 p. Tese (Doutorado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2026.

This thesis develops a comprehensive framework for modeling asymmetric dependence structures in matrix-valued data using the Matrix-Variate Skew-Normal (MVSN) distribution and its censored extension. We establish key theoretical properties of these models, including their stochastic representations, moments, and identifiability conditions. Building on these foundations, we derive likelihood-based inference procedures and propose Expectation–Conditional Maximization (ECM) algorithms capable of handling both fully observed and interval-censored or partially missing matrix-valued observations. Simulation studies are conducted to assess parameter recovery, convergence behavior, and robustness of the proposed estimation methods under diverse scenarios. Finally, we demonstrate the practical usefulness of the models through applications to real datasets with complex censoring structures. The results show that the MVSN and censored MVSN models offer flexible and interpretable tools for analyzing multivariate and matrix-structured data exhibiting asymmetry, moderate tails, and incomplete information.

Keywords: Matrix Variate Skew-Normal Distribution, Censored MVSN, Asymmetric matrix models, ECM algorithm, Censored/Missing data.

LIST OF FIGURES

CONTENTS

CHAPTER

1

INTRODUCTION

The normal distribution has long served as one of the foundational tools in statistics, largely due to its mathematical simplicity and strong theoretical underpinnings. Its analytical convenience makes it a natural choice for modeling data and performing estimation, which explains its central role in both applied and theoretical research. However, despite its importance, the normal distribution often fails to capture the complexity of real-world data. Many empirical datasets exhibit visible deviations from normality, such as skewness or light tails that allow for more extreme values than the Gaussian framework would predict. These limitations have motivated extensive research into more flexible families of distributions capable of modeling asymmetry and tail behavior while retaining the appealing properties that make the normal distribution so useful.

Responding to these limitations, the skew-normal family of distributions has emerged as a widely studied and powerful generalization of the normal model. The univariate skew-normal distribution was introduced by Azzalini (1985) and further developed in Azzalini (1986), providing a principled way to accommodate controlled skewness. This framework was later extended to the multivariate case by Azzalini and Valle (1996) and Azzalini and Capitanio (1999), yielding a richer class of models that preserve many of the desirable features of the multivariate normal distribution while enhancing flexibility in the presence of asymmetric data.

The multivariate skew-normal distribution is particularly appealing because it introduces a skewness mechanism that generalizes the symmetric multivariate normal, offering a more faithful representation of datasets where asymmetry is inherent—for example, in income measurements, financial returns, or biological processes. By maintaining mathematical tractability while allowing for asymmetric behavior, this distribution has found applications across statistics, econometrics, machine learning, and Bayesian inference. Numerous works, such as Sahu, Dey and Branco (2003), Lin, McLachlan and Lee (2016), and Queiroz, Loschi and Silva (2016), have further expanded its theoretical and methodological scope, illustrating the growing interest in models capable of handling non-Gaussian structure.

Motivated by these advancements, recent efforts have extended the skew-normal framework to the matrix-variate setting, which is essential for modern applications involving structured or high-dimensional data. Contributions by Gupta, González-Farıas and Domınguez-Molina (2004), Chen and Gupta (2005), and Ning and Gupta (2012) introduced several formulations of the matrix-variate skew-normal (MVSN) distribution, broadening its applicability to settings where preserving matrix structure is crucial. These extensions allow for more realistic modeling of dependencies across both rows and columns, expanding the utility of skew-normal models in increasingly complex statistical environments.

Despite this progress, key challenges persist regarding the theoretical properties and estimation procedures for MVSN models. The matrix structure introduces additional dependency layers that complicate inference, and classical techniques often require careful adaptation. Maximum likelihood estimation, in particular, becomes challenging due to latent variables and non-convex likelihood surfaces. EM-type algorithms provide a natural solution, with the Expectation–Conditional Maximization (ECM) variant offering computational advantages by decomposing complex optimization tasks into manageable components.

This work contributes to the ongoing development of matrix-variate skew-normal methodology by establishing essential distributional properties, clarifying identifiability conditions, and introducing a robust ECM algorithm specifically tailored to the MVSN model. A thorough examination of the model's theoretical structure is combined with extensive simulation studies, demonstrating the stability and accuracy of the proposed estimation technique. Real data applications further highlight the model's ability to capture intricate asymmetric patterns in multivariate and longitudinal datasets. Together, these developments form a rigorous and unified framework for inference under the MVSN distribution, positioning it as a powerful alternative for modeling non-Gaussian matrix-valued data.

By laying this theoretical and computational groundwork, the first part of this thesis establishes a solid foundation for the study and application of matrix-variate skew-normal models Azzalini and Capitanio (2014), Alencar, Gonçalves *et al.* (2022). The broader motivation for these tools becomes even more apparent when considering the growing complexity of real-world datasets. Modern applications frequently involve observations that deviate from Gaussian assumptions — not only through asymmetry or moderate tails but also through patterns of censoring and missingness. Such challenges arise in environmental monitoring, where measurements often fall below detection limits or quantification thresholds Helsel (2012), Huynh *et al.* (2014); in biomedical imaging, where MRI magnitude data exhibit intrinsically non-Gaussian Rician noise Gudbjartsson and Patz (1995), Coupé *et al.* (2010); in signal processing, where asymmetric and heavy-tailed noise distributions provide more realistic models than the classical Gaussian assumption Middleton (1986), Xu, Zhou and Li (2022); and in longitudinal studies, where incomplete observations and censoring require principled likelihood-based methods Little and Rubin (2002), Bandyopadhyay and Rao (2012). These data characteristics demand probabilistic

frameworks capable of jointly handling non-Gaussian behavior and incomplete-data mechanisms in a coherent and computationally tractable way (see Dempster, Laird and Rubin (1977), Dagne, Rousson *et al.* (2013), for instance).

Within this context, the matrix variate skew-normal censored (MVSNC) model provides a coherent extension of the MVSN framework by integrating three key elements: (i) matrix-variate dependence, (ii) a skewness mechanism, and (iii) a censoring structure. This unified formulation captures the essential features of matrix-valued data while accommodating asymmetry and partial observation. Unlike ad hoc approaches, the MVSNC model incorporates censored entries directly in the likelihood, ensuring that the information they contain is properly used during inference.

The model is built on a latent-variable representation in which skewness and censoring arise from hidden components that enrich the matrix-normal kernel. This hierarchical structure naturally yields truncated conditional distributions for censored entries, providing a principled treatment of incomplete data. By operating directly on matrices rather than vectorized forms, the MVSNC model preserves dependence patterns along both dimensions, resulting in more realistic representations of complex data arrays.

Inference under the MVSNC model is nontrivial due to interactions between latent skewness variables, censoring indicators, and nonlinear parameter relationships. As direct likelihood maximization is generally infeasible, EM-type procedures become essential. The ECM algorithm offers an efficient solution by decomposing the estimation problem into manageable conditional updates, relying on conditional expectations that arise naturally from the censoring mechanism.

The study of the MVSNC model advances both theory and methodology for censored matrix-valued data. It establishes core distributional properties, clarifies identifiability conditions, and develops a full likelihood-based estimation framework. Simulation studies and real-data applications illustrate substantial gains in flexibility and accuracy when modeling asymmetric and partially observed matrices. By jointly addressing skewness, structured dependence, and censoring, the MVSNC model emerges as a powerful tool for non-Gaussian incomplete data.

The second part of the thesis consolidates these developments by formally constructing the MVSNC model, deriving its latent-variable representation, and obtaining a tractable likelihood that incorporates censoring. A central contribution is the ECM algorithm, which leverages expectations from truncated distributions to deliver stable and efficient inference. Through theoretical results, simulations, and applications, this part demonstrates the model's ability to simultaneously capture skewness, preserve matrix structure, and handle censoring, offering a comprehensive framework for analyzing non-Gaussian censored matrix-valued data.

Organization of the Thesis

This thesis is organized to provide a coherent progression from foundational concepts to methodological developments, simulation studies, and real-data applications involving matrix-variate skew-normal models and their censored extensions.

Chapter 2 establishes the theoretical groundwork for the thesis. It reviews the matrix variate normal distribution and its main properties, introduces skewed matrix-variate extensions that motivate subsequent developments, and concludes with maximum likelihood estimation for the MVN model, providing a baseline for later methodological comparisons.

Chapter 3 presents the core theoretical contributions related to the matrix variate skew-normal (MVSN) distribution. It revisits the multivariate skew-normal model and extends the discussion to the matrix-variate setting, including the Matrix Variate Extended Skew-Normal (MVESN) distribution. This chapter develops key distributional properties and introduces a likelihood-based estimation strategy for the MVSN model, forming the basis for empirical assessments.

Chapter 4 evaluates the performance and practical utility of the MVSN model through a comprehensive set of simulation studies and a real-data application. It examines parameter recovery, compares the MVSN and MVN frameworks, and highlights the advantages of explicitly modeling skewness in matrix-valued contexts. The chapter concludes with an applied example that illustrates the interpretability and flexibility of the MVSN model in capturing asymmetric dependence structures.

Chapter 5 extends the modeling framework to handle incomplete observations through the Matrix Variate Skew-Normal Censored (MVSNC) model. It begins by formalizing the use of the MVSN distribution for interval-censored and missing data and develops an ECM algorithm specifically tailored for censored matrix-valued observations. This chapter focuses on the theoretical and methodological aspects of the MVSNC model, including its properties and the details of the ECM estimation procedure.

Chapter 6 presents simulation studies and real-data applications to evaluate the performance of the MVSNC model. The simulations assess estimation accuracy and compare the MVSNC and MVSN models, highlighting the advantages of explicitly modeling censoring. The chapter concludes with the application of the MVSNC model to a real dataset, demonstrating its ability to jointly accommodate skewness, structured dependence, and censoring while providing interpretable parameter estimates.

Chapter 7 introduces the computational framework of the `MVCens` package, presenting estimation and simulation tools for the MVN and MVSN models with complete or incomplete data. The chapter outlines the unified ECM-based inference approach, the treatment of censoring and missingness, and the generation of synthetic matrix-variate data, emphasizing the package's role in supporting simulation studies and applied matrix-variate analyses.

Overall, these chapters collectively provide a unified treatment of matrix-variate skew-normal models, encompassing theoretical foundations, methodological extensions, and practical applications. The thesis culminates in a comprehensive framework for modeling non-Gaussian matrix-valued data with or without censoring.

CHAPTER

2

THEORETICAL BACKGROUND

In the first place, we introduce the notation that will remain consistent throughout this chapter. To clarify, a random matrix with dimension $p \times q$ is referred to as $\boldsymbol{\mathcal{X}}$, such that its realization is denoted as $\mathcal{X}$. Furthermore, a random vector of size pq is identified as $\mathbf{X}$, while its realization is labeled as $\mathbf{x}$. Moreover, $|\cdot|$ represents the determinant of a square matrix, $\text{tr}(\cdot)$ is the trace of a square matrix, $\text{vec}(\cdot)$ vectorizes a matrix and $\otimes$ denotes the kronecker product.

2.1 Matrix Variate Normal Distribution

In Nguyen (1997), the authors extend earlier results from Ahsanullah (1985), which focused on characterizing the bivariate normal distribution, to a multidimensional setting. They then use these extended results to explore the Matrix Variate Normal Distribution, specifically in cases where the row vectors are identically distributed. Their approach builds on the same ideas used by Ahsanullah (1985) for the bivariate case, but adapts them using familiar concepts from matrix theory and linear transformations in real Euclidean space $\mathbb{R}^n$.

In the more general setting where the random matrices do not necessarily have identically distributed row vectors, the authors from the reference Gupta and Nagar (2018) describe the expression of the corresponding joint probability density function of the random matrices with Matrix Variate Normal Distributions (MVN distribution from now on).

A practical motivation for studying these distributions is outlined next. In many biomedical studies, each patient provides several correlated outcomes (e.g., blood pressure, cholesterol, weight), collected across multiple patients receiving the same treatment. Representing the data for each treatment as an $p \times q$ matrix reveals two dependence sources: correlations among the p measurements within each patient (rows) and among the q patients within the same treatment (columns). Standard multivariate models cannot properly accommodate this bidirectional structure and require estimating an impractically large covariance matrix. Matrix–variate models

address this by preserving the matrix form and separating the covariance into row and column components, yielding a more parsimonious and coherent framework for modeling treatment effects. More precisely, for each treatment group j, the data can be arranged as an $p \times q$ matrix.

$$\boldsymbol{\mathcal{X}}_j = \begin{bmatrix} x_{11}^{(j)} & x_{12}^{(j)} & \cdots & x_{1q}^{(j)} \\ x_{21}^{(j)} & x_{22}^{(j)} & \cdots & x_{2q}^{(j)} \\ \vdots & \vdots & \ddots & \vdots \\ x_{p1}^{(j)} & x_{p2}^{(j)} & \cdots & x_{pq}^{(j)} \end{bmatrix} \in \mathbb{R}^{p \times q}.$$

This matrix layout makes the structure clear: for each treatment group $j \in \{1, \ldots, n\}$, where each group corresponds to a distinct treatment, the rows represent the p measured variables and the columns correspond to the q patients within that treatment. This representation highlights the dependence structure in both directions, thereby motivating the use of matrix–variate rather than standard multivariate models.

The matrix–variate normal (MVN) distribution plays a central role in modeling multidimensional data that naturally arise in matrix form. Before developing more general models, it is essential to recall the basic structure of the MVN and the key properties that make it particularly suitable for matrix-structured observations. In this framework, the distribution is characterized by a mean matrix together with two covariance matrices that separately govern dependence across rows and across columns. This separable covariance formulation preserves the intrinsic layout of the data and provides a substantial reduction in the number of parameters compared to an unstructured multivariate normal model. These definitions and properties form the foundation upon which several extensions, including skewed, moderate-tailed, and mixture formulations, are constructed.

Definition 1. We say the random vector $\mathbf{X} \in \mathbb{R}^p$ has a Multivariate Normal distribution (MN distribution from now on) with mean vector $\mu \in \mathbb{R}^p$ and symmetric positive definite covariance matrix $|\Sigma| > 0$ if, and only if, its joint probability density function is given by:

$$f(\mathbf{X} \mid \mu, \Sigma) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{X}-\mu)^\top \Sigma^{-1}(\mathbf{X}-\mu)\right\},$$

Definition 2. A random vector $\mathbf{X} \in \mathbb{R}^n$ with mean vector μ and covariance matrix Σ is said to have a MN distribution if, and only if, the distribution of $\mathbf{c}^\top \mathbf{X} \sim \mathcal{N}(\mathbf{c}^\top \mu, \mathbf{c}^\top \Sigma \mathbf{c})$ for $\mathbf{c}^\top \in \mathbb{R}^n \backslash \{0\}$.

In both cases, this relationship is denoted by the notation $\mathbf{X} \sim \mathcal{N}_p(\mu, \Sigma)$. As demonstrated in Tong (1990), these two definitions are equivalent. In what follows, we present two important results, which are likewise provided in Tong (1990).

Proposition 1 (Affine transformation of the Multivariate Normal Distribution)**.** Let $\mathbf{Y} = \mathbf{C}\mathbf{X} + \mathbf{b}$ where $\mathbf{C}$ is an $p \times p$ matrix and $\mathbf{b}$ is a real vector (of size p). If $\mathbf{X} \sim \mathcal{N}_p(\mu, \Sigma)$, $|\Sigma| > 0$, and $\mathbf{C}$ satisfies $|\mathbf{C}| \neq 0$, then $\mathbf{Y} \sim \mathcal{N}_p(\mu_\mathbf{Y}, \Sigma_\mathbf{Y})$, $|\Sigma_\mathbf{Y}| > 0$, where

$$\mu_\mathbf{Y} = \mathbf{C}\mu + \mathbf{b}, \quad \Sigma_\mathbf{Y} = \mathbf{C}\Sigma\mathbf{C}^\top.$$

Remark. To present the next result, we shall need the following conventions. Let $1 \leq k < p$ be fixed, with $p \in \mathbb{N}$, $p \geq 2$, and define $p_1 := k$ and $p_2 := p - k$. On this basis, we adopt the convention that $\mathbf{X}_1 \in \mathbb{R}^{p_1}$, $\mathbf{X}_2 \in \mathbb{R}^{p_2}$, $\mu_1 \in \mathbb{R}^{p_1}$, $\mu_2 \in \mathbb{R}^{p_2}$ and $\Sigma_{ij} = \text{Cov}(\mathbf{X}_i, \mathbf{X}_j)$. More precisely,

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix}, \quad \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

Proposition 2 (Marginal density from the Multivariate Normal Distribution)**.** If $\mathbf{X} \sim \mathcal{N}_p(\mu, \Sigma)$, then for every fixed $1 \leq k < p$ the marginal distributions of $\mathbf{X}_1$ and $\mathbf{X}_2$ are $\mathcal{N}_k(\mu_1, \Sigma_{11})$ and $\mathcal{N}_{p-k}(\mu_2, \Sigma_{22})$, respectively.

In line with the results presented in Gupta and Nagar (2018), the next considerations apply:

Definition 3. We claim that a random matrix $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{p \times q}$ follows a Matrix Variate Normal Distribution (MVN distribution from now on) with location matrix $\mathbf{M} \in \mathbb{R}^{p \times q}$ and positive definite covariance matrix $\Psi \otimes \Sigma$, where $\Sigma \in \mathbb{R}^{p \times p}$ and $\Psi \in \mathbb{R}^{q \times q}$ are positive-definite symmetric matrices, iff $\text{vec}(\boldsymbol{\mathcal{X}}) \sim \mathcal{N}_{pq}(\text{vec}(\mathbf{M}), \Psi \otimes \Sigma)$, which we are going to denote by the notation $\boldsymbol{\mathcal{X}} \sim \mathcal{N}_{p \times q}(\mathbf{M}, \Sigma, \Psi)$.

Proposition 3. If $\boldsymbol{\mathcal{X}} \sim \mathcal{N}_{p \times q}(\mathbf{M}, \Sigma, \Psi)$, then the PDF of $\boldsymbol{\mathcal{X}}$ is given by:

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

Still in accordance with the reference Gupta and Nagar (2018), it can be proved that an analogous result to Proposition 1 also holds for matrix variate distributions. Precisely speaking, we may state it as follows.

Proposition 4 (Affine transformation of the Matrix Variate Normal Distribution)**.** If the random matrix $\boldsymbol{\mathcal{X}} \sim \mathcal{N}_{p \times q}(\mathbf{M}, \Sigma, \Psi)$, the matrix $\mathbf{C} \in \mathbb{R}^{r \times p}$ has rank $r \leq p$ and the matrix $\mathbf{D} \in \mathbb{R}^{q \times s}$ has rank $s \leq q$, then it can be concluded that:

$$\mathbf{C}\boldsymbol{\mathcal{X}}\mathbf{D} = \mathcal{N}_{r \times s}(\mathbf{CMD}, \mathbf{C}\Sigma\mathbf{C}^\top, \mathbf{D}^\top \Psi \mathbf{D}).$$

It is worth noting the presence of an identifiability issue concerning the two covariance matrices. Specifically, $\Psi \otimes \Sigma = \Psi^\star \otimes \Sigma^\star$ if $\Sigma^\star = a\Sigma$ and $\Psi^\star = a^{-1}\Psi$. Therefore the two covariance matrices are identifiable up to a multiplicative constant as discussed in Dutilleul (1999). In the matrix-variate literature, various methods have been suggested to tackle this issue. One possible solution consists in following the approach employed in Melnykov and Zhu (2018)), Sarkar *et al.* (2020) and Tomarchio, Punzo and Bagnato (2021), which involves imposing the constraint $|\Psi| = 1$.

Here is also worth to discuss the draws and backs of vectorization of random matrices. From a theoretical standpoint, consider the relationship described between the matrix variate distribution and its vectorial counterpart:

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

Such relation is often useful for establishing and computing quantities of interest when dealing with MVN-based models, as demonstrated in works by Viroli (2011), Viroli (2012). However, in scenarios other than those explicitly mentioned, the matrix-variate formulation is preferred over its multivariate counterpart when the data is presented in a matrix form. One potential approach involves transforming the matrix-variate data into a vector format and then applying multivariate models. Nonetheless, this rearrangement of data presents several practical challenges, extensively discussed and illustrated in the literature on matrix-variate analysis (e.g., Allen and Tibshirani (2010); Anderlucci *et al.* (2014); Gallaugher and McNicholas (2018); Sarkar *et al.* (2020); and Tomarchio, McNicholas and Punzo (2021)).

Hereafter, we offer a summary of the issues associated with such a method.

1. Interpretability would be compromised if the two sources of variability, governed by Σ and Ψ, were put together into a single $\Psi \otimes \Sigma$ matrix. Such extract would result in a loss of interpretability of the data.

2. Parsimony is compromised by the increase in the number of free covariance parameters in the matrix-variate setting, which is given by the formula $p(p+1)/2+q(q+1)/2-1$. When vectorized, this count escalates to $pq(pq+1)/2$, potentially leading to an overparameterization of the multivariate models.

3. Model selection becomes intricately linked to the preceding issue, as the proliferation of parameters in the multivariate context can pose challenges in selecting the appropriate model. This is a consequence of the increased weight assigned to the penalty term of widely used information criteria.

2.2 Matrix Differentiation Calculus

We shall start this section with some considerations involving the definitions of the Fréchet derivative and Gâteaux derivative in the context of random matrices.

Let $f : \mathbb{R}^{p\times q} \to \mathbb{R}$ be a scalar-valued function whose argument is a matrix. The space $\mathbb{R}^{p\times q}$ is a finite-dimensional vector space, and it is equipped with the Frobenius inner product $\langle \mathbf{A}, \mathbf{B} \rangle_{\mathrm{F}} = \mathrm{tr}(\mathbf{A}^\top \mathbf{B})$ and the associated norm $\|\mathbf{A}\|_{\mathrm{F}}^2 = \mathrm{tr}(\mathbf{A}^\top \mathbf{A})$. These structures allow us to define derivatives and gradients with respect to the Frobenius geometry.

Definition 4. Given $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{p\times q}$ and an arbitrary direction $\mathbf{H} \in \mathbb{R}^{p\times q}$, the Gâteaux derivative of f at $\boldsymbol{\mathcal{X}}$ in the direction $\mathbf{H}$ is defined by

$$Df(\boldsymbol{\mathcal{X}})[\mathbf{H}] = \lim_{\varepsilon \to 0} \frac{f(\boldsymbol{\mathcal{X}} + \varepsilon \mathbf{H}) - f(\boldsymbol{\mathcal{X}})}{\varepsilon}$$

where $\varepsilon \in \mathbb{R}$, provided the limit exists. This definition mirrors the usual directional derivative in $\mathbb{R}^n$: we restrict the function to the line $\varepsilon \mapsto \boldsymbol{\mathcal{X}} + \varepsilon \mathbf{H}$ and differentiate with respect to ε at $\varepsilon = 0$.

Thus, $Df(\boldsymbol{\mathcal{X}})[\mathbf{H}]$ represents the first-order variation of f when the matrix $\boldsymbol{\mathcal{X}}$ is perturbed in the direction $\mathbf{H}$. In general, the mapping $\mathbf{H} \mapsto Df(\boldsymbol{\mathcal{X}})\mathbf{H}$. In general, the Gâteaux derivative need not define a linear mapping in $\mathbf{H}$. When linearity holds, it defines a linear functional.

Definition 5. Let $f : \mathbb{R}^{p\times q} \to \mathbb{R}$ and let $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{p\times q}$. Then the function f is said to be Fréchet differentiable at $\boldsymbol{\mathcal{X}}$ if there exists a linear mapping $\mathcal{L} : \mathbb{R}^{p\times q} \to \mathbb{R}$ such that

$$\lim_{\|\mathbf{H}\|\to 0} \frac{\left|f(\boldsymbol{\mathcal{X}}+\mathbf{H}) - f(\boldsymbol{\mathcal{X}}) - \mathcal{L}(\mathbf{H})\right|_{\mathbb{R}}}{\|\mathbf{H}\|} = 0,$$

where $\|\cdot\|$ denotes any matrix norm on $\mathbb{R}^{p\times q}$, and the mapping $\mathcal{L}$ is called the Fréchet derivative of f at $\boldsymbol{\mathcal{X}}$ and is denoted by

$$Df(\boldsymbol{\mathcal{X}}) : \mathbb{R}^{p\times q} \to \mathbb{R}, \qquad \mathbf{H} \mapsto Df(\boldsymbol{\mathcal{X}})[\mathbf{H}].$$

In this case, $\mathcal{L}(\mathbf{H}) = Df(\boldsymbol{\mathcal{X}})[\mathbf{H}]$ provides the best linear approximation of f near $\boldsymbol{\mathcal{X}}$. Moreover, this linear mapping captures the first-order behavior of f in a neighborhood of $\boldsymbol{\mathcal{X}}$.

Theorem 1 (Riesz Representation Theorem)**.** Let V be a finite-dimensional inner product space over the field $\mathbb{F}$, and let $g : V \to \mathbb{F}$ be a linear functional. Then there exists a unique vector $y \in V$ such that $g(x) = \langle y, x\rangle$, for all $x \in V$.

Proof. A proof can be found in Friedberg, Insel and Spence (2013). □

Proposition 5. Let us suppose that $f : \mathbb{R}^{p\times q} \to \mathbb{R}$ is a function whose Gâteaux derivative exists in a neighborhood of $\boldsymbol{\mathcal{X}}$ and such that, for each $\boldsymbol{\mathcal{Y}}$ in this neighborhood, the mapping $\mathbf{H} \mapsto Df(\boldsymbol{\mathcal{Y}})[\mathbf{H}]$ is linear. If the gradient mapping $\boldsymbol{\mathcal{Y}} \mapsto \nabla_{\boldsymbol{\mathcal{Y}}} f$ (where $\nabla_{\boldsymbol{\mathcal{Y}}} f$ is the Riesz representative matrix of $Df(\boldsymbol{\mathcal{Y}})$) is continuous at $\boldsymbol{\mathcal{X}}$ (with respect to the Frobenius norm), then the mapping $\boldsymbol{\mathcal{Y}} \mapsto Df(\boldsymbol{\mathcal{Y}})$ is continuous at $\boldsymbol{\mathcal{X}}$ with respect to the operator norm.

Proof. Since the Gâteaux derivative of f exists in a neighborhood of $\boldsymbol{\mathcal{X}}$ and, for each $\boldsymbol{\mathcal{Y}}$ in this neighborhood, the mapping $\mathbf{H} \mapsto Df(\boldsymbol{\mathcal{Y}})[\mathbf{H}]$ is linear, it follows that $Df(\boldsymbol{\mathcal{Y}})$ defines a linear functional on $\mathbb{R}^{p\times q}$. As $\mathbb{R}^{p\times q}$ is a finite-dimensional inner-product space endowed with the Frobenius inner product, Theorem 1 ensures the existence of a unique matrix $\nabla_{\boldsymbol{\mathcal{Y}}} f \in \mathbb{R}^{p\times q}$ such that

$$Df(\boldsymbol{\mathcal{Y}})[\mathbf{H}] = \langle \nabla_{\boldsymbol{\mathcal{Y}}} f, \mathbf{H}\rangle_{\mathrm{F}} = \operatorname{tr}\big((\nabla_{\boldsymbol{\mathcal{Y}}} f)^{\top}\mathbf{H}\big),$$

for all directions $\mathbf{H} \in \mathbb{R}^{p\times q}$. Let $\boldsymbol{\mathcal{Y}}_n \to \boldsymbol{\mathcal{X}}$. For any $\mathbf{H} \in \mathbb{R}^{p\times q}$ with $\|\mathbf{H}\|_{\mathrm{F}} = 1$, the application of the Cauchy-Schwarz inequality yields the relation

$$\left|(Df(\boldsymbol{\mathcal{Y}}_n) - Df(\boldsymbol{\mathcal{X}}))[\mathbf{H}]\right| = \left|\langle \nabla_{\boldsymbol{\mathcal{Y}}_n} f - \nabla_{\boldsymbol{\mathcal{X}}} f, \mathbf{H}\rangle_{\mathrm{F}}\right| \leq \|\nabla_{\boldsymbol{\mathcal{Y}}_n} f - \nabla_{\boldsymbol{\mathcal{X}}} f\|_{\mathrm{F}}.$$

Taking the supremum over all $\mathbf{H}$ such that $\|\mathbf{H}\|_{\mathrm{F}} = 1$, it follows that

$$\|Df(\boldsymbol{\mathcal{Y}}_n) - Df(\boldsymbol{\mathcal{X}})\|_{\mathrm{op}} \leq \|\nabla_{\boldsymbol{\mathcal{Y}}_n} f - \nabla_{\boldsymbol{\mathcal{X}}} f\|_{\mathrm{F}}.$$

Since $\mathcal{Y} \mapsto \nabla_{\mathcal{Y}} f$ is continuous at $\mathcal{X}$, we have

$$\|\nabla_{\mathcal{Y}_n} f - \nabla_{\mathcal{X}} f\|_{\mathrm{F}} \xrightarrow{n\to+\infty} 0 \Rightarrow \|Df(\mathcal{Y}_n) - Df(\mathcal{X})\|_{\mathrm{op}} \xrightarrow{n\to+\infty} 0.$$

Therefore, the mapping $\mathcal{Y} \mapsto Df(\mathcal{Y})$ is continuous at $\mathcal{X}$ in accordance with the sequential characterization of continuity in metric spaces. □

Remark. In fact, in the present context, the operator norm of the derivative coincides with the Frobenius norm of its gradient. Indeed, for any $\mathcal{Y} \in \mathbb{R}^{p\times q}$,

$$\|Df(\mathcal{Y})\|_{\mathrm{op}} = \sup_{\|\mathbf{H}\|_{\mathrm{F}}=1} \left|\langle \nabla_{\mathcal{Y}} f, \mathbf{H}\rangle_{\mathrm{F}}\right| \geq \left|\langle \nabla_{\mathcal{Y}} f, \mathbf{H}^{\star}\rangle_{\mathrm{F}}\right| = \|\nabla_{\mathcal{Y}} f\|_{\mathrm{F}},$$

where the inequality is attained by taking $\mathbf{H}^{\star} = \nabla_{\mathcal{Y}} f / \|\nabla_{\mathcal{Y}} f\|_{\mathrm{F}}$ when $\nabla_{\mathcal{Y}} f \neq 0$, and is trivial otherwise. Since the converse inequality also holds by Cauchy–Schwarz, the correspondence given by the Riesz representation theorem is an isometry between linear functionals and their gradient representations.

Theorem 2. Let $U \subset \mathbb{R}^{p\times q}$ be an open set and let $f : U \to \mathbb{R}$ be a function. Suppose that f admits a Gâteaux derivative at every point of a neighborhood of $\mathcal{X} \in U$. Denote this derivative by $Df(\mathcal{Y})$. Assume that the mapping $\mathcal{Y} \mapsto Df(\mathcal{Y})$ is continuous at $\mathcal{X}$ with respect to the operator norm. Then f is Fréchet differentiable at $\mathcal{X}$, and its Fréchet derivative at $\mathcal{X}$ coincides with the Gâteaux derivative $Df(\mathcal{X})$.

Proof. The proof of such result can be found at Deimling (2013). □

Assume that the Gâteaux derivative of f exists in a neighborhood of $\mathcal{X}$ and that, for each $\mathcal{X}$, the mapping $\mathbf{H} \mapsto Df(\mathcal{X})[\mathbf{H}]$ is linear in $\mathbf{H}$. Since $\mathbb{R}^{p\times q}$ is a finite-dimensional inner-product space endowed with the Frobenius inner product, Theorem 1 applies in this setting. In particular, $Df(\mathcal{X})$ defines a linear functional on the Hilbert space $(\mathbb{R}^{p\times q}, \langle\cdot,\cdot\rangle_{\mathrm{F}})$. Accordingly, we define $\nabla_{\mathcal{X}} f$ as the Riesz representative matrix of $Df(\mathcal{X})$, the unique matrix $\nabla_{\mathcal{X}} f \in \mathbb{R}^{p\times q}$ such that

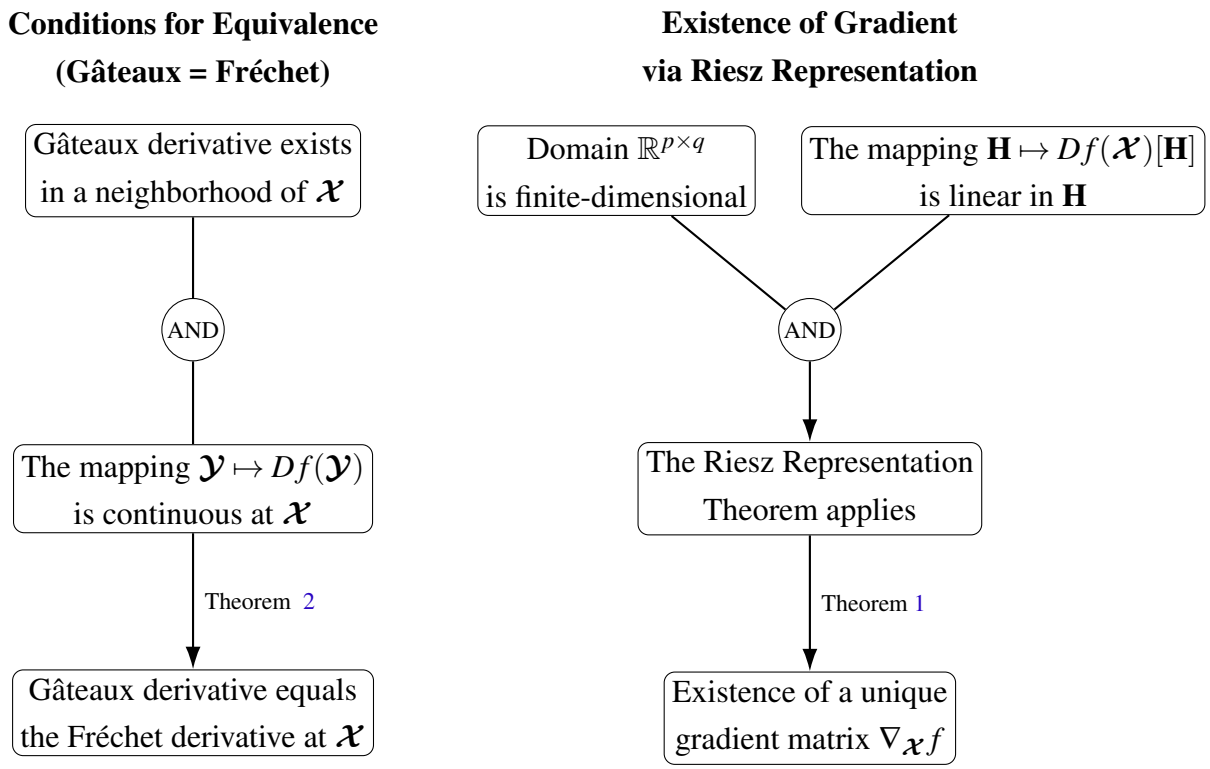
$$Df(\mathcal{X})[\mathbf{H}] = \langle \nabla_{\mathcal{X}} f, \mathbf{H}\rangle_{\mathrm{F}} = \mathrm{tr}((\nabla_{\mathcal{X}} f)^{\top}\mathbf{H})$$

for all $\mathbf{H} \in \mathbb{R}^{p\times q}$. The matrix $\nabla_{\mathcal{X}} f$, called the gradient of f at $\mathcal{X}$, provides the canonical matrix representation of the derivative with respect to the Frobenius inner product, thereby translating the linear functional into a concrete matrix expression.

Furthermore, if the mapping $\mathcal{Y} \mapsto Df(\mathcal{Y})$ is continuous at $\mathcal{X}$, then, in accordance with Theorem 2, f is Fréchet differentiable at $\mathcal{X}$ and the Gâteaux and Fréchet derivatives coincide. In this case, $Df(\mathcal{X})$ admits the representation $Df(\mathcal{X})[\mathbf{H}] = \langle \nabla_{\mathcal{X}} f, \mathbf{H}\rangle_{\mathrm{F}}$. Consequently, f possesses the first-order expansion

$$f(\mathcal{X} + \mathbf{H}) = f(\mathcal{X}) + Df(\mathcal{X})[\mathbf{H}] + o(\|\mathbf{H}\|_{\mathrm{F}}),$$

which characterizes the total derivative at $\mathcal{X}$ and yields a precise local linear approximation in the Frobenius norm. The main implications used in the argument are summarized in the flowchart.



where each box follows logically from the statements connected to it in the diagram. When f is expressed using trace operators, the derivative typically produces expressions involving both $\mathbf{H}$ and $\mathbf{H}^\top$. To rewrite these expressions in gradient form, standard trace identities are used, namely linearity of the trace, cyclic invariance $\mathrm{tr}(\mathbf{UVW}) = \mathrm{tr}(\mathbf{WUV})$, and the transpose identity $\mathrm{tr}(\mathbf{MH}^\top) = \mathrm{tr}(\mathbf{M}^\top\mathbf{H})$. These properties allow all first-order terms to be rearranged so that the derivative can be written as a Frobenius inner product with $\mathbf{H}$.

For quadratic trace functions of the form $f(\mathcal{X}) = \mathrm{tr}(\mathbf{A}\mathcal{X}\mathbf{B}\mathcal{X}^\top)$, a perturbation $\mathcal{X} \mapsto \mathcal{X} + \varepsilon\mathbf{H}$ generates terms of order ε^0, ε^1, and ε^2. Only the first-order terms — those in which exactly one occurrence of $\mathcal{X}$ is replaced by $\mathbf{H}$ — contribute to the derivative, while higher-order terms vanish after division by ε and taking the limit. This structure explains why the derivative of such functions consists of a sum of linear terms involving $\mathbf{H}$.

In summary, the Gâteaux derivative $\mathbf{H} \mapsto Df(\mathcal{X})\mathbf{H}$ extends the classical directional derivative to matrix spaces and characterizes the first-order variation of a scalar function with respect to matrix perturbations. Although it is defined along individual directions, when its domain is finite-dimensional and its variation is linear in the direction matrix, it admits a unique gradient representation through the Frobenius inner product. In addition, if the function $\mathcal{Y} \mapsto Df(\mathcal{Y})$ is continuous at $\mathcal{X}$, the Gâteaux derivative coincides with the Fréchet derivative.

The next result plays a fundamental role in rigorously establishing the connection between the two previously mentioned definitions of the derivative with respect to a particular function of interest that will be analyzed in greater detail later on, and some analogous results are presented in the sequence.

Lemma 1. Consider the function $f(\mathcal{X},\Sigma,\Psi) = \operatorname{tr}(\Sigma^{-1}\mathcal{X}\Psi^{-1}\mathcal{X}^\top)$ where both $\Sigma \succ 0$ and $\Psi \succ 0$ are fixed so that $\mathcal{X} \in \mathbb{R}^{p\times q}$ may vary, where the relation $\succ$ means that they are positive definite. Then its Gâteaux derivative with respect to $\mathcal{X}$ is represented, under the Frobenius inner product, by the gradient

$$\nabla_{\mathcal{X}} f(\mathcal{X},\Sigma,\Psi) = 2\Sigma^{-1}\mathcal{X}\Psi^{-1}$$

Notice this expression coincides with the Fréchet derivative in the present setting. Indeed, since the mapping $\mathcal{X} \mapsto \nabla_{\mathcal{X}} f$ is continuous, by Proposition 5, one has that $f \in C^1(\mathbb{R}^{p\times q})$. Based on Theorem 2, the desired claim holds.

Proof. Fix Σ and Ψ and let f vary as a function of $\mathcal{X}$. Let $\mathbf{H} \in \mathbb{R}^{p\times q}$ be an arbitrary direction. By definition of the Gâteaux derivative with respect to $\mathcal{X}$, we have

$$D_{\mathcal{X}} f(\mathcal{X},\Sigma,\Psi)[\mathbf{H}] = \lim_{\varepsilon\to 0} \frac{f(\mathcal{X}+\varepsilon\mathbf{H},\Sigma,\Psi) - f(\mathcal{X},\Sigma,\Psi)}{\varepsilon}.$$

To simplify the notation, set $\mathbf{A} = \Sigma^{-1}$ and $\mathbf{B} = \Psi^{-1}$. With this notation,

$$f(\mathcal{X},\Sigma,\Psi) = \operatorname{tr}(\mathbf{A}\mathcal{X}\mathbf{B}\mathcal{X}^\top).$$

We now evaluate the function at the perturbed argument $\mathcal{X}+\varepsilon\mathbf{H}$. Using linearity of matrix multiplication and the identity $(\mathcal{X}+\varepsilon\mathbf{H})^\top = \mathcal{X}^\top + \varepsilon\mathbf{H}^\top$, we obtain

$$f(\mathcal{X}+\varepsilon\mathbf{H},\Sigma,\Psi) = \operatorname{tr}(\mathbf{A}(\mathcal{X}+\varepsilon\mathbf{H})\mathbf{B}(\mathcal{X}^\top+\varepsilon\mathbf{H}^\top)).$$

Expanding the product inside the trace gives

$$(\mathcal{X}+\varepsilon\mathbf{H})\mathbf{B}(\mathcal{X}^\top+\varepsilon\mathbf{H}^\top) = \mathcal{X}\mathbf{B}\mathcal{X}^\top + \varepsilon\,\mathbf{H}\mathbf{B}\mathcal{X}^\top + \varepsilon\,\mathcal{X}\mathbf{B}\mathbf{H}^\top + \varepsilon^2\mathbf{H}\mathbf{B}\mathbf{H}^\top.$$

Substituting this expansion into the trace expression yields

$$f(\mathcal{X}+\varepsilon\mathbf{H},\Sigma,\Psi) = \operatorname{tr}(\mathbf{A}\mathcal{X}\mathbf{B}\mathcal{X}^\top) + \varepsilon\operatorname{tr}(\mathbf{A}\mathbf{H}\mathbf{B}\mathcal{X}^\top) + \varepsilon\operatorname{tr}(\mathbf{A}\mathcal{X}\mathbf{B}\mathbf{H}^\top) + \varepsilon^2\operatorname{tr}(\mathbf{A}\mathbf{H}\mathbf{B}\mathbf{H}^\top).$$

We now subtract $f(\mathcal{X},\Sigma,\Psi)$, divide by ε, and take the limit as $\varepsilon \to 0$. The terms of order ε^2 vanish in the limit, and we obtain

$$D_{\mathcal{X}} f(\mathcal{X},\Sigma,\Psi)[\mathbf{H}] = \operatorname{tr}(\mathbf{A}\mathbf{H}\mathbf{B}\mathcal{X}^\top) + \operatorname{tr}(\mathbf{A}\mathcal{X}\mathbf{B}\mathbf{H}^\top).$$

The next step is to express this directional derivative in the form of a Frobenius inner product with $\mathbf{H}$. Using cyclicity of the trace, the identity $\operatorname{tr}(\mathbf{M}\mathbf{H}^\top) = \operatorname{tr}(\mathbf{M}^\top\mathbf{H})$ as well as the fact that $\operatorname{tr}(\mathbf{X}) = \operatorname{tr}(\mathbf{X}^\top)$, we can rewrite both terms in the following way

$$\begin{aligned}
\operatorname{tr}(\mathbf{A}\mathbf{H}\mathbf{B}\mathcal{X}^\top) &= \operatorname{tr}((\mathbf{A}\mathbf{H}\mathbf{B}\mathcal{X}^\top)^\top) \\
&= \operatorname{tr}(\mathcal{X}\mathbf{B}^\top\mathbf{H}^\top\mathbf{A}^\top) \\
&= \operatorname{tr}(\mathbf{H}^\top\mathbf{A}^\top\mathcal{X}\mathbf{B}^\top) \\
&= \operatorname{tr}((\mathbf{A}^\top\mathcal{X}\mathbf{B}^\top)^\top\mathbf{H})
\end{aligned}$$

and $\operatorname{tr}(\mathbf{A}\boldsymbol{\mathcal{X}}\mathbf{B}\mathbf{H}^\top) = \operatorname{tr}((\mathbf{A}\boldsymbol{\mathcal{X}}\mathbf{B})^\top\mathbf{H})$. Combining these expressions gives

$$D_{\boldsymbol{\mathcal{X}}} f(\boldsymbol{\mathcal{X}},\Sigma,\Psi)[\mathbf{H}] = \operatorname{tr}(((\mathbf{A}^\top\boldsymbol{\mathcal{X}}\mathbf{B}^\top)^\top + (\mathbf{A}\boldsymbol{\mathcal{X}}\mathbf{B})^\top)\mathbf{H}).$$

Equivalently, due to the linearity of the transpose operator, one may conclude that

$$D_{\boldsymbol{\mathcal{X}}} f(\boldsymbol{\mathcal{X}},\Sigma,\Psi)[\mathbf{H}] = \operatorname{tr}((\mathbf{A}^\top\boldsymbol{\mathcal{X}}\mathbf{B}^\top + \mathbf{A}\boldsymbol{\mathcal{X}}\mathbf{B})^\top\mathbf{H}).$$

By definition of the gradient under the Frobenius inner product, the matrix $\nabla_{\boldsymbol{\mathcal{X}}} f(\boldsymbol{\mathcal{X}},\Sigma,\Psi)$ is characterized by the identity

$$D_{\boldsymbol{\mathcal{X}}} f(\boldsymbol{\mathcal{X}},\Sigma,\Psi)[\mathbf{H}] = \operatorname{tr}((\nabla_{\boldsymbol{\mathcal{X}}} f(\boldsymbol{\mathcal{X}},\Sigma,\Psi))^\top\mathbf{H}) \quad \text{for all } \mathbf{H}\in\mathbb{R}^{p\times q}.$$

Comparing both expressions, it results from the uniqueness of the gradient that

$$\nabla_{\boldsymbol{\mathcal{X}}} f(\boldsymbol{\mathcal{X}},\Sigma,\Psi) = \mathbf{A}^\top\boldsymbol{\mathcal{X}}\mathbf{B}^\top + \mathbf{A}\boldsymbol{\mathcal{X}}\mathbf{B} = \Sigma^{-\top}\boldsymbol{\mathcal{X}}\Psi^{-\top} + \Sigma^{-1}\boldsymbol{\mathcal{X}}\Psi^{-1}.$$

Finally, since Σ and Ψ are symmetric and positive definite, we have $\Sigma^{-\top} = \Sigma^{-1}$ and $\Psi^{-\top} = \Psi^{-1}$, which leads to the simplified expression

$$\nabla_{\boldsymbol{\mathcal{X}}} f(\boldsymbol{\mathcal{X}},\Sigma,\Psi) = 2\Sigma^{-1}\boldsymbol{\mathcal{X}}\Psi^{-1}.$$

□

Lemma 2. Consider the function $f(\boldsymbol{\mathcal{X}},\Sigma,\Psi) = \operatorname{tr}(\Sigma^{-1}\boldsymbol{\mathcal{X}}\Psi^{-1}\boldsymbol{\mathcal{X}}^\top)$, where both $\boldsymbol{\mathcal{X}}\in\mathbb{R}^{p\times q}$ and $\Psi \succ 0$ are held fixed and $\Sigma \succ 0$ is allowed to vary. Then the Gâteaux derivative of f with respect to Σ is represented, under the Frobenius inner product, by the gradient

$$\nabla_\Sigma f(\boldsymbol{\mathcal{X}},\Sigma,\Psi) = -\Sigma^{-1}\boldsymbol{\mathcal{X}}\Psi^{-1}\boldsymbol{\mathcal{X}}^\top\Sigma^{-1}.$$

Moreover, this expression coincides with the Fréchet derivative in the present finite-dimensional setting. Indeed, the mapping $\Sigma\mapsto\nabla_\Sigma f$ is continuous on the space of symmetric positive definite matrices $\mathbb{S}^p_{++}$, which implies that f is continuously differentiable by Proposition 5, that is to say, $f\in C^1(\mathbb{S}^p_{++})$. Consequently, due to Theorem 2, the Gâteaux and Fréchet derivatives agree.

Proof. Fix $\boldsymbol{\mathcal{X}}$ and Ψ and view f as a function of Σ only. Let $\mathbf{S}\in\mathbb{R}^{p\times p}$ be an arbitrary direction. By definition of the Gâteaux derivative with respect to Σ,

$$D_\Sigma f(\boldsymbol{\mathcal{X}},\Sigma,\Psi)[\mathbf{S}] = \lim_{\varepsilon\to 0}\frac{f(\boldsymbol{\mathcal{X}},\Sigma+\varepsilon\mathbf{S},\Psi) - f(\boldsymbol{\mathcal{X}},\Sigma,\Psi)}{\varepsilon}.$$

Set $\mathbf{B} = \Psi^{-1}$ and $\mathbf{C} = \boldsymbol{\mathcal{X}}\mathbf{B}\boldsymbol{\mathcal{X}}^\top$, so that

$$f(\boldsymbol{\mathcal{X}},\Sigma,\Psi) = \operatorname{tr}\big(\Sigma^{-1}\mathbf{C}\big).$$

To evaluate $f(\boldsymbol{\mathcal{X}},\Sigma+\varepsilon\mathbf{S},\Psi)$, we expand the inverse of the perturbed matrix. Using

$$(\Sigma+\varepsilon\mathbf{S})^{-1} = \big(\Sigma(\mathbf{I}_p+\varepsilon\Sigma^{-1}\mathbf{S})\big)^{-1} = (\mathbf{I}_p+\varepsilon\Sigma^{-1}\mathbf{S})^{-1}\Sigma^{-1},$$

and the first-order approximation $(\mathbf{I}_p + \varepsilon\mathbf{K})^{-1} = \mathbf{I}_p - \varepsilon\mathbf{K} + o(\varepsilon)$ as $\varepsilon \to 0$, we obtain

$$(\Sigma + \varepsilon\mathbf{S})^{-1} = \big(\mathbf{I}_p - \varepsilon\Sigma^{-1}\mathbf{S} + o(\varepsilon)\big)\Sigma^{-1} = \Sigma^{-1} - \varepsilon\Sigma^{-1}\mathbf{S}\Sigma^{-1} + o(\varepsilon).$$

Substituting into f yields

$$f(\boldsymbol{\mathcal{X}}, \Sigma + \varepsilon\mathbf{S}, \Psi) = \operatorname{tr}\big((\Sigma^{-1} - \varepsilon\Sigma^{-1}\mathbf{S}\Sigma^{-1} + o(\varepsilon))\mathbf{C}\big) = \operatorname{tr}(\Sigma^{-1}\mathbf{C}) - \varepsilon \operatorname{tr}(\Sigma^{-1}\mathbf{S}\Sigma^{-1}\mathbf{C}) + o(\varepsilon).$$

Therefore,

$$\frac{f(\boldsymbol{\mathcal{X}}, \Sigma + \varepsilon\mathbf{S}, \Psi) - f(\boldsymbol{\mathcal{X}}, \Sigma, \Psi)}{\varepsilon} = -\operatorname{tr}(\Sigma^{-1}\mathbf{S}\Sigma^{-1}\mathbf{C}) + \frac{o(\varepsilon)}{\varepsilon},$$

and letting $\varepsilon \to 0$ gives

$$D_\Sigma f(\boldsymbol{\mathcal{X}}, \Sigma, \Psi)[\mathbf{S}] = -\operatorname{tr}(\Sigma^{-1}\mathbf{S}\Sigma^{-1}\mathbf{C}).$$

Using cyclicity of the trace, we rewrite this as

$$D_\Sigma f(\boldsymbol{\mathcal{X}}, \Sigma, \Psi)[\mathbf{S}] = -\operatorname{tr}(\Sigma^{-1}\mathbf{C}\Sigma^{-1}\mathbf{S}) = \operatorname{tr}((-\Sigma^{-1}\mathbf{C}\Sigma^{-1})^\top\mathbf{S}).$$

By definition of the gradient under the Frobenius inner product,

$$D_\Sigma f(\boldsymbol{\mathcal{X}}, \Sigma, \Psi)[\mathbf{S}] = \operatorname{tr}((\nabla_\Sigma f)^\top\mathbf{S}) \quad \text{for all } \mathbf{S},$$

so we identify

$$\nabla_\Sigma f(\boldsymbol{\mathcal{X}}, \Sigma, \Psi) = -\Sigma^{-1}\mathbf{C}\Sigma^{-1} = -\Sigma^{-1}\boldsymbol{\mathcal{X}}\Psi^{-1}\boldsymbol{\mathcal{X}}^\top\Sigma^{-1}.$$

Finally, since Ψ is symmetric positive definite, $\mathbf{B} = \Psi^{-1}$ is symmetric, and therefore $\mathbf{C} = \boldsymbol{\mathcal{X}}\mathbf{B}\boldsymbol{\mathcal{X}}^\top$ is symmetric as well, implying that the gradient above is symmetric whenever Σ is symmetric. □

Lemma 3. Consider the function $f(\boldsymbol{\mathcal{X}}, \Sigma, \Psi) = \operatorname{tr}\big(\Sigma^{-1}\boldsymbol{\mathcal{X}}\Psi^{-1}\boldsymbol{\mathcal{X}}^\top\big)$, where $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{p\times q}$, $\Sigma \succ 0$ and $\Psi \succ 0$. Fix $\boldsymbol{\mathcal{X}}$ and Σ and let only the parameter Ψ vary. Then the Gâteaux derivative of f with respect to Ψ is represented by the gradient

$$\nabla_\Psi f(\boldsymbol{\mathcal{X}}, \Sigma, \Psi) = -\Psi^{-1}\boldsymbol{\mathcal{X}}^\top\Sigma^{-1}\boldsymbol{\mathcal{X}}\Psi^{-1}.$$

Moreover, in this finite-dimensional setting, this expression coincides with the Fréchet derivative. Indeed, since the mapping $\Psi \mapsto \nabla_\Psi f$ is continuous on the space of symmetric positive definite matrices $\mathbb{S}^q_{++}$, it follows that f is continuously differentiable by Proposition 5, that is, $f \in C^1(\mathbb{S}^q_{++})$. Therefore, by virtue of Theorem 2, the Gâteaux and Fréchet derivatives coincide.

Proof. Fix $\boldsymbol{\mathcal{X}}$ and Σ and view f as a function of Ψ only. Let $\mathbf{T} \in \mathbb{R}^{q\times q}$ be an arbitrary direction. By definition of the Gâteaux derivative with respect to Ψ,

$$D_\Psi f(\boldsymbol{\mathcal{X}}, \Sigma, \Psi)[\mathbf{T}] = \lim_{\varepsilon\to 0} \frac{f(\boldsymbol{\mathcal{X}}, \Sigma, \Psi + \varepsilon\mathbf{T}) - f(\boldsymbol{\mathcal{X}}, \Sigma, \Psi)}{\varepsilon}.$$

To simplify the notation, set $\mathbf{A} = \Sigma^{-1}$ and $\mathbf{C} = \boldsymbol{\mathcal{X}}^\top \mathbf{A} \boldsymbol{\mathcal{X}}$. Since

$$\operatorname{tr}\big(\mathbf{A}\boldsymbol{\mathcal{X}}\Psi^{-1}\boldsymbol{\mathcal{X}}^\top\big) = \operatorname{tr}\big(\Psi^{-1}\boldsymbol{\mathcal{X}}^\top\mathbf{A}\boldsymbol{\mathcal{X}}\big) = \operatorname{tr}\big(\Psi^{-1}\mathbf{C}\big),$$

we may write

$$f(\boldsymbol{\mathcal{X}}, \Sigma, \Psi) = \operatorname{tr}\big(\Psi^{-1}\mathbf{C}\big).$$

We now evaluate the function at the perturbed argument $\Psi + \varepsilon\mathbf{T}$. Using

$$(\Psi + \varepsilon\mathbf{T})^{-1} = \big(\Psi(\mathbf{I}_q + \varepsilon\Psi^{-1}\mathbf{T})\big)^{-1} = (\mathbf{I}_q + \varepsilon\Psi^{-1}\mathbf{T})^{-1}\Psi^{-1},$$

together with the first-order approximation

$$(\mathbf{I}_q + \varepsilon\mathbf{K})^{-1} = \mathbf{I}_q - \varepsilon\mathbf{K} + o(\varepsilon), \qquad \varepsilon \to 0,$$

we obtain

$$(\Psi + \varepsilon\mathbf{T})^{-1} = \Psi^{-1} - \varepsilon\,\Psi^{-1}\mathbf{T}\Psi^{-1} + o(\varepsilon).$$

Substituting into the trace expression yields

$$f(\boldsymbol{\mathcal{X}}, \Sigma, \Psi + \varepsilon\mathbf{T}) = \operatorname{tr}\big((\Psi^{-1} - \varepsilon\,\Psi^{-1}\mathbf{T}\Psi^{-1} + o(\varepsilon))\,\mathbf{C}\big) = \operatorname{tr}(\Psi^{-1}\mathbf{C}) - \varepsilon\operatorname{tr}(\Psi^{-1}\mathbf{T}\Psi^{-1}\mathbf{C}) + o(\varepsilon).$$

Therefore,

$$\frac{f(\boldsymbol{\mathcal{X}}, \Sigma, \Psi + \varepsilon\mathbf{T}) - f(\boldsymbol{\mathcal{X}}, \Sigma, \Psi)}{\varepsilon} = -\operatorname{tr}(\Psi^{-1}\mathbf{T}\Psi^{-1}\mathbf{C}) + \frac{o(\varepsilon)}{\varepsilon},$$

and letting $\varepsilon \to 0$ gives

$$D_\Psi f(\boldsymbol{\mathcal{X}}, \Sigma, \Psi)[\mathbf{T}] = -\operatorname{tr}(\Psi^{-1}\mathbf{T}\Psi^{-1}\mathbf{C}).$$

Using cyclicity of the trace, we rewrite this expression as

$$D_\Psi f(\boldsymbol{\mathcal{X}}, \Sigma, \Psi)[\mathbf{T}] = -\operatorname{tr}(\Psi^{-1}\mathbf{C}\Psi^{-1}\mathbf{T}) = \operatorname{tr}((-\Psi^{-1}\mathbf{C}\Psi^{-1})^\top\mathbf{T}).$$

By definition of the gradient under the Frobenius inner product,

$$D_\Psi f(\boldsymbol{\mathcal{X}}, \Sigma, \Psi)[\mathbf{T}] = \operatorname{tr}\big((\nabla_\Psi f)^\top\mathbf{T}\big) \quad \text{for all } \mathbf{T},$$

so we identify

$$\nabla_\Psi f(\boldsymbol{\mathcal{X}}, \Sigma, \Psi) = -\Psi^{-1}\mathbf{C}\Psi^{-1} = -\Psi^{-1}\boldsymbol{\mathcal{X}}^\top\Sigma^{-1}\boldsymbol{\mathcal{X}}\Psi^{-1}.$$

Finally, since Σ is symmetric positive definite, $\mathbf{A} = \Sigma^{-1}$ is symmetric, and therefore $\mathbf{C} = \boldsymbol{\mathcal{X}}^\top\mathbf{A}\boldsymbol{\mathcal{X}}$ is symmetric as well, implying that the gradient above is symmetric whenever Ψ is symmetric. □

Lemma 4. Consider the function $f(\Sigma) = \log|\Sigma^{-1}|$ where $\Sigma \succ 0$. Then the Gâteaux derivative of f with respect to Σ is represented by the gradient

$$\nabla_\Sigma f(\Sigma) = \Sigma^{-1}.$$

Moreover, this expression coincides with the Fréchet derivative. Indeed, since the mapping $\Sigma \mapsto \nabla_\Sigma f$ is continuous on the space of symmetric positive definite matrices $\mathbb{S}^p_{++}$, it follows that $f \in C^1(\mathbb{S}^p_{++})$ by Proposition 5. Therefore, by Theorem 2, the Gâteaux and Fréchet derivatives coincide.

Proof. Let $\mathbf{H} \in \mathbb{R}^{p\times p}$ be an arbitrary symmetric matrix ($\mathbf{H} = \mathbf{H}^\top$). Since the function f is defined on the cone of symmetric positive definite matrices, we consider perturbations of the form $\Sigma + \varepsilon\mathbf{H}$. Because this cone is an open subset of $\mathbb{R}^{p\times p}$, such perturbations remain positive definite for $|\varepsilon|$ sufficiently small, ensuring that the quantities below are well defined.

By definition, the Gâteaux derivative of f at Σ in the direction $\mathbf{H}$ is given by

$$Df(\Sigma)[\mathbf{H}] = \lim_{\varepsilon\to 0} \frac{f(\Sigma+\varepsilon\mathbf{H}) - f(\Sigma)}{\varepsilon}.$$

Our task is therefore to understand how f varies when Σ is subjected to a small symmetric perturbation. Using the determinant identity $|\mathbf{A}^{-1}| = |\mathbf{A}|^{-1}$ valid for every invertible matrix $\mathbf{A}$, we rewrite

$$f(\Sigma+\varepsilon\mathbf{H}) = \log\left|(\Sigma+\varepsilon\mathbf{H})^{-1}\right| = -\log\left|\Sigma+\varepsilon\mathbf{H}\right|.$$

Thus, differentiating f reduces to studying the variation of the log-determinant of the perturbed matrix. Substituting into the definition above, we obtain

$$Df(\Sigma)[\mathbf{H}] = -\lim_{\varepsilon\to 0} \frac{\log|\Sigma+\varepsilon\mathbf{H}| - \log|\Sigma|}{\varepsilon}.$$

To make the perturbation explicit, we factor Σ from the matrix $\Sigma+\varepsilon\mathbf{H}$:

$$\Sigma+\varepsilon\mathbf{H} = \Sigma\big(\mathbf{I}_p + \varepsilon\Sigma^{-1}\mathbf{H}\big).$$

This factorization is useful because the determinant of a product equals the product of determinants. Hence, $|\Sigma+\varepsilon\mathbf{H}| = |\Sigma|\left|\mathbf{I}_p+\varepsilon\Sigma^{-1}\mathbf{H}\right|$. Taking logarithms converts this multiplicative relation into an additive one:

$$\log|\Sigma+\varepsilon\mathbf{H}| = \log|\Sigma| + \log\left|\mathbf{I}_p+\varepsilon\Sigma^{-1}\mathbf{H}\right|.$$

When substituted into the difference quotient, the constant term $\log|\Sigma|$ cancels, leaving

$$Df(\Sigma)[\mathbf{H}] = -\lim_{\varepsilon\to 0} \frac{\log\left|\mathbf{I}_p+\varepsilon\Sigma^{-1}\mathbf{H}\right|}{\varepsilon}.$$

Let $\mathbf{A} := \Sigma^{-1}\mathbf{H}$. We must now determine the first-order behavior of $\log|\mathbf{I}_p+\varepsilon\mathbf{A}|$ as $\varepsilon\to 0$. A classical expansion of the determinant shows that

$$\left|\mathbf{I}_p+\varepsilon\mathbf{A}\right| = 1+\varepsilon\,\mathrm{tr}(\mathbf{A})+o(\varepsilon), \qquad \varepsilon\to 0.$$

This formula expresses the fact that the trace governs the first-order change of the determinant near the identity. Applying the scalar expansion $\log(1+t) = t+o(t)$ then yields

$$\log\left|\mathbf{I}_p+\varepsilon\mathbf{A}\right| = \varepsilon\,\mathrm{tr}(\mathbf{A})+o(\varepsilon).$$

Dividing by ε and passing to the limit gives

$$Df(\Sigma)[\mathbf{H}] = -\mathrm{tr}(\mathbf{A}) = -\mathrm{tr}\big(\Sigma^{-1}\mathbf{H}\big).$$

We have thus obtained the directional derivative as a linear function of the perturbation $\mathbf{H}$. To identify the gradient, we express this linear map using the Frobenius inner product. By definition,

$$Df(\Sigma)[\mathbf{H}] = \operatorname{tr}\big((\nabla_\Sigma f(\Sigma))^\top \mathbf{H}\big) \quad \text{for all symmetric } \mathbf{H}.$$

Comparing with the expression above, we conclude that $\nabla_\Sigma f(\Sigma) = -\Sigma^{-1}$. Finally, since Σ is symmetric positive definite, its inverse is also symmetric, so the gradient lies in the same space of symmetric matrices as the admissible perturbations. This completes the proof.

□

Remark. An entirely similar result can be deduced for Ψ.

Throughout the remainder of this work, derivatives with respect to matrix parameters are going to be written using the partial derivative notation, for example $\partial f/\partial \mathbf{M}$, instead of the previous gradient notation $\nabla_{\mathbf{M}} f$. This is purely a matter of notation. From now on, the functions under consideration are scalar-valued functions of matrix arguments and their Gâteaux derivatives are linear functionals acting on perturbation matrices. In the finite-dimensional space $\mathbb{R}^{p\times q}$ endowed with the Frobenius inner product, Theorem 1 ensures that each such functional can be uniquely represented as an inner product with a matrix of the same dimension.

Consequently, the symbols $\nabla_{\mathbf{M}} f$ and $\partial f/\partial \mathbf{M}$ denote the same matrix, characterized by

$$Df(\mathbf{M})[\mathbf{H}] = \operatorname{tr}((\nabla_{\mathbf{M}} f)^\top \mathbf{H}) = \operatorname{tr}\left\{ \left(\frac{\partial f}{\partial \mathbf{M}} \right)^\top \mathbf{H} \right\},$$

for all directions $\mathbf{H} \in \mathbb{R}^{p\times q}$.

Once we have discussed matrix differentiation, it remains to understand how to study the critical points in the matrix variate setting. Although the parameters of the matrix-variate normal model are matrices, the log-likelihood function is still a real-valued function,

$$\ell(\vartheta) := \log f_{\mathrm{MVN}}(\vartheta \mid \mathcal{X}) \in \mathbb{R}, \qquad \vartheta = (\mathbf{M}, \Sigma, \Psi).$$

This means that, regardless of the matrix structure of the parameters, the output of ℓ is a single real number. Therefore, from the viewpoint of optimization, we are still dealing with the problem of maximizing a scalar function.

The difference from the univariate case lies essentially in the nature of the argument of the function. Instead of depending on a scalar variable $x \in \mathbb{R}$, the function ℓ depends on matrices that belong to spaces such as $\mathbb{R}^{p\times q}$, $\mathbb{R}^{p\times p}$, and $\mathbb{R}^{q\times q}$, each of which forms a finite-dimensional vector space. For instance, a matrix $\mathbf{M} \in \mathbb{R}^{p\times q}$ contains pq real entries and can be identified with a vector in $\mathbb{R}^{pq}$ by stacking its columns via the vec operator. Similarly, when symmetry is taken into account, Σ and Ψ can be represented in terms of their independent entries, allowing them to be treated as elements of Euclidean spaces of appropriate dimension, thereby facilitating the extension of standard differential calculus arguments to this matrix setting.

Consequently, the parameter ϑ can be viewed as a point in a Euclidean space $\mathbb{R}^d$ for some finite d, since each matrix parameter may be identified with a vector of its entries (for instance, via the vec operator, once symmetry constraints are taken into account). In this sense, the matrix-variate log-likelihood may be regarded simply as a smooth function $\ell : \mathbb{R}^d \longrightarrow \mathbb{R}$,written in matrix notation for convenience and structural clarity. This identification is not merely notational: it guarantees that differentiating with respect to a matrix parameter is mathematically equivalent to differentiating with respect to all of its entries viewed as a vector. Matrix calculus therefore provides a structured way to compute the same derivatives that would arise from standard multivariate differentiation in $\mathbb{R}^d$.

Because the parameter space is finite-dimensional, all the familiar tools of multivariate calculus — directional derivatives, gradients, first-order optimality conditions, and convexity arguments — apply exactly as in the classical setting. The use of matrix notation does not alter the underlying optimization problem; it simply preserves the natural algebraic structure of the model while carrying out standard Euclidean analysis.

2.3 ML Estimation for the MVN Distribution

This section is partially based on the reference Glanz and Carvalho (2018). We begin by considering a random sample of size $n \in \mathbb{N}_{>0}$ drawn from a matrix-variate normal (MVN) distribution, $\boldsymbol{\mathcal{X}}_1, \boldsymbol{\mathcal{X}}_2, \ldots, \boldsymbol{\mathcal{X}}_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}_{p\times q}(\mathbf{M}, \Sigma, \Psi)$, where $\mathbf{M} \in \mathbb{R}^{p\times q}$ denotes the mean (or location) matrix, while $\Sigma \in \mathbb{R}^{p\times p}$ and $\Psi \in \mathbb{R}^{q\times q}$ represent, respectively, the row and column covariance matrices. These two matrices allow the model to capture dependence structures separately along rows and columns, which is one of the key features distinguishing the matrix-variate normal distribution from the classical multivariate normal model.

Suppose that $\mathcal{X} = (\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_n)$ is a realization of this sample. The likelihood function measures how plausible a given parameter value $\vartheta = (\mathbf{M}, \Sigma, \Psi)$ is in light of the observed data. Because the observations are independent and identically distributed, the likelihood factorizes into a product of individual densities, and therefore the log-likelihood becomes a sum:

$$\log f_{\text{MVN}}(\vartheta \mid \mathcal{X}) = \sum_{i=1}^{n} \log f_{\text{MVN}}(\vartheta \mid \mathcal{X}_i) = \frac{pn}{2} \log|\Psi^{-1}| + \frac{qn}{2} \log|\Sigma^{-1}| - \frac{1}{2} \sum_{i=1}^{n} \delta(\mathcal{X}_i \mid \mathbf{M}, \Sigma, \Psi),$$

up to an additive constant that does not depend on the parameters. Here, we adopt the notation

$$\delta(\mathcal{X}_i \mid \mathbf{M}, \Sigma, \Psi) := \operatorname{tr}(\Sigma^{-1}(\mathcal{X}_i - \mathbf{M})\Psi^{-1}(\mathcal{X}_i - \mathbf{M})^\top)$$

which plays the role of a matrix-valued Mahalanobis distance.

Throughout this section we assume that Σ and Ψ are symmetric and positive definite, that is to say, $\Sigma \succ 0$ and $\Psi \succ 0$. This assumption guarantees that the inverses and log-determinants appearing in the likelihood exist, and hence that the model is mathematically well defined.

It is useful to note that the log-likelihood naturally decomposes into two distinct parts. The first part, involving $\log|\Sigma|$ and $\log|\Psi|$, depends only on the covariance structure, while the second part measures how far each observation $\mathcal{X}_i$ is from the mean matrix $\mathbf{M}$. Indeed, δ can be rewritten as

$$\delta(\mathcal{X}_i \mid \mathbf{M}, \Sigma, \Psi) = \left\| \Sigma^{-1/2}(\mathcal{X}_i - \mathbf{M})\Psi^{-1/2} \right\|_{\mathrm{F}}^2,$$

showing that it is simply a squared Frobenius norm. Thus δ is always non-negative and quantifies the discrepancy between $\mathcal{X}_i$ and $\mathbf{M}$ after accounting for the covariance structure. Consequently, maximizing the log-likelihood with respect to the parameter $\mathbf{M}$ is equivalent to minimizing the sum of $\delta(\mathcal{X}_i \mid \mathbf{M}, \Sigma, \Psi)$, which is a least-squares type problem in matrix form.

Lemma 5 (Uniqueness of the MLE for the Mean Matrix)**.** For fixed $\Sigma \succ 0$ and $\Psi \succ 0$, the log-likelihood of the matrix-variate normal distribution is strictly concave in $\mathbf{M} \in \mathbb{R}^{p\times q}$. Hence the solution of the first-order condition is the unique global maximizer.

Proof. Fix $\Sigma \succ 0$ and $\Psi \succ 0$, as well as $\delta_i(\mathbf{M}) := \delta(\mathcal{X}_i \mid \mathbf{M}, \Sigma, \Psi)$ and consider the objective

$$Q(\mathbf{M}) := \sum_{i=1}^{n} \delta_i(\mathbf{M}) = \sum_{i=1}^{n} \operatorname{tr}(\Sigma^{-1}(\mathcal{X}_i - \mathbf{M})\Psi^{-1}(\mathcal{X}_i - \mathbf{M})^\top).$$

Using the Frobenius norm identity

$$\operatorname{tr}(\Sigma^{-1}(\mathcal{X}_i - \mathbf{M})\Psi^{-1}(\mathcal{X}_i - \mathbf{M})^\top) = \left\| \Sigma^{-1/2}(\mathcal{X}_i - \mathbf{M})\Psi^{-1/2} \right\|_{\mathrm{F}}^2,$$

we may rewrite

$$Q(\mathbf{M}) = \sum_{i=1}^{n} \left\| \Sigma^{-1/2}(\mathcal{X}_i - \mathbf{M})\Psi^{-1/2} \right\|_{\mathrm{F}}^2.$$

To prove that Q is strictly convex, let $\mathbf{M}_1, \mathbf{M}_2 \in \mathbb{R}^{p\times q}$ be such that $\mathbf{M}_1 \neq \mathbf{M}_2$, and let $t \in (0,1)$. We analyze the behavior of Q along the segment joining $\mathbf{M}_1$ and $\mathbf{M}_2$. To this end, consider the linear operator $\mathcal{L} : \mathbb{R}^{p\times q} \to \mathbb{R}^{p\times q}$ defined by

$$\mathcal{L}(\mathbf{A}) := \Sigma^{-1/2}\mathbf{A}\Psi^{-1/2}$$

Because $\Sigma^{-1/2}$ and $\Psi^{-1/2}$ are symmetric positive definite, they are invertible. Hence $\mathcal{L}$ is injective ($\ker(\mathcal{L}) = \{\mathbf{0}\}$). In particular, $\mathcal{L}(\mathbf{A}) = \mathbf{0} = \mathcal{L}(\mathbf{0})$ implies $\mathbf{A} = \mathbf{0}$. For each $i \in \{1, 2, \ldots, n\}$, define

$$\mathbf{Y}_i(\mathbf{M}) := \mathcal{L}(\mathcal{X}_i - \mathbf{M}).$$

The mapping $\mathbf{M} \mapsto \mathbf{Y}_i(\mathbf{M})$ is affine in $\mathbf{M}$, and we may write

$$\delta_i(\mathbf{M}) = \|\mathbf{Y}_i(\mathbf{M})\|_{\mathrm{F}}^2.$$

Thus, each δ_i is the squared Frobenius norm of an affine function of $\mathbf{M}$.

Since the squared norm is strictly convex on a Euclidean space,

$$\|t\mathbf{A}+(1-t)\mathbf{B}\|_{\mathrm{F}}^2 < t\|\mathbf{A}\|_{\mathrm{F}}^2+(1-t)\|\mathbf{B}\|_{\mathrm{F}}^2 \quad \text{whenever } \mathbf{A}\neq\mathbf{B}.$$

Based on the inequality above, we can establish that $\delta_i(\mathbf{M})$ is strictly convex. To this end, we first derive the following auxiliary identity, which will be instrumental in the proof of such claim.

$$\begin{aligned}
\mathbf{Y}_i\big(t\mathbf{M}_1+(1-t)\mathbf{M}_2\big) &= \mathcal{L}\big(\mathcal{X}_i-\big(t\mathbf{M}_1+(1-t)\mathbf{M}_2\big)\big)\\
&= \mathcal{L}(\mathcal{X}_i-t\mathbf{M}_1-(1-t)\mathbf{M}_2)\\
&= \mathcal{L}(t(\mathcal{X}_i-\mathbf{M}_1)+(1-t)(\mathcal{X}_i-\mathbf{M}_2))\\
&= t\,\mathcal{L}(\mathcal{X}_i-\mathbf{M}_1)+(1-t)\,\mathcal{L}(\mathcal{X}_i-\mathbf{M}_2)\\
&= t\,\mathbf{Y}_i(\mathbf{M}_1)+(1-t)\,\mathbf{Y}_i(\mathbf{M}_2).
\end{aligned}$$

Based on it, let us assume that $\mathbf{M}_1\neq\mathbf{M}_2$. Then one may deduce that

$$\begin{aligned}
\delta_i(t\mathbf{M}_1+(1-t)\mathbf{M}_2) &= \|\mathbf{Y}_i(t\mathbf{M}_1+(1-t)\mathbf{M}_2)\|_{\mathrm{F}}^2\\
&= \|t\mathbf{Y}_i(\mathbf{M}_1)+(1-t)\mathbf{Y}_i(\mathbf{M}_2)\|_{\mathrm{F}}^2\\
&< t\,\|\mathbf{Y}_i(\mathbf{M}_1)\|_{\mathrm{F}}^2+(1-t)\|\mathbf{Y}_i(\mathbf{M}_2)\|_{\mathrm{F}}^2\\
&= t\delta_i(\mathbf{M}_1)+(1-t)\delta_i(\mathbf{M}_2),
\end{aligned}$$

provided that $\mathbf{Y}_i(\mathbf{M}_1)\neq\mathbf{Y}_i(\mathbf{M}_2)$. But we do also know that

$$\mathbf{Y}_i(\mathbf{M}_1)=\mathbf{Y}_i(\mathbf{M}_2) \iff \mathcal{L}(\mathbf{M}_1-\mathbf{M}_2)=\mathbf{0} \iff \mathbf{M}_1=\mathbf{M}_2,$$

where the last equivalence follows from injectivity of $\mathcal{L}$. Therefore, for $\mathbf{M}_1\neq\mathbf{M}_2$, the strict inequality holds. Summing over $i\in\{1,2,\ldots,n\}$ gives

$$Q(t\mathbf{M}_1+(1-t)\mathbf{M}_2) < t\,Q(\mathbf{M}_1)+(1-t)\,Q(\mathbf{M}_2),$$

so Q is strictly convex on $\mathbb{R}^{p\times q}$. Finally, since for fixed Σ and Ψ the log-likelihood has the form

$$\ell(\mathbf{M}) = \text{constant}-\frac{1}{2}\,Q(\mathbf{M}),$$

it follows immediately that $\ell(\mathbf{M})$ is strictly concave in $\mathbf{M}$. □

Since $\mathbf{M}\in\mathbb{R}^{p\times q}$ is an unconstrained parameter, its maximum likelihood estimator is obtained by equating to zero the Fréchet derivative of the log-likelihood with respect to $\mathbf{M}$. Because the log-determinant terms do not depend on $\mathbf{M}$, this differentiation involves only the quadratic part of the likelihood. Carrying out the derivative yields

$$-\frac{1}{2}\frac{\partial}{\partial\mathbf{M}}\sum_{i=1}^{n}\delta(\mathcal{X}_i\mid\mathbf{M},\Sigma,\Psi)=0 \Longleftrightarrow \frac{\partial}{\partial\mathbf{M}}\sum_{i=1}^{n}\operatorname{tr}(\Sigma^{-1}(\mathcal{X}_i-\mathbf{M})\Psi^{-1}(\mathcal{X}_i-\mathbf{M})^{\top})=0.$$

Using the matrix differentiation rules earlier established as well as the chain rule, the derivative of each summand with respect to $\mathbf{M}$ is given by $\Sigma^{-1}(\mathcal{X}_i - \mathbf{M})\Psi^{-1}$. Summing these contributions over $i \in \{1, 2, \ldots, n\}$, the first-order necessary condition for maximizing the log-likelihood with respect to $\mathbf{M}$ becomes

$$\sum_{i=1}^{n} \Sigma^{-1}(\mathcal{X}_i - \mathbf{M})\Psi^{-1} = 0.$$

Because Σ^{-1} and Ψ^{-1} do not depend on the index i, they may be factored out of the summation, since they are constant with respect to the summation index and therefore commute with the finite sum. Consequently, the equation above can be rewritten as

$$\Sigma^{-1}\left(\sum_{i=1}^{n}(\mathcal{X}_i - \mathbf{M})\right)\Psi^{-1} = 0.$$

Since Σ and Ψ are assumed to be symmetric positive definite, they are invertible matrices. Therefore, the only way the preceding product can vanish is for the central term to be equal to zero, that is,

$$\sum_{i=1}^{n}(\mathcal{X}_i - \mathbf{M}) = 0.$$

Expanding the summation and rearranging terms yields

$$\widehat{\mathbf{M}} = \frac{1}{n}\sum_{i=1}^{n}\mathcal{X}_i.$$

Thus, the maximum likelihood estimator of the mean matrix is simply the sample average of the observed matrices, extending the classical result for the multivariate normal distribution to the matrix-variate case. Finally, by Lemma 5 (Uniqueness of the MLE for the Mean Matrix), the log-likelihood is strictly concave in $\mathbf{M}$ for fixed $\Sigma \succ 0$ and $\Psi \succ 0$. Therefore, this stationary point is unique and necessarily corresponds to the global maximizer of the log-likelihood with respect to $\mathbf{M}$.

We now turn to the estimation of Σ and Ψ. Unlike $\mathbf{M}$, these parameters are constrained to lie in the cone of symmetric positive definite matrices. Since the log-likelihood is strictly concave in Ψ (when Σ is fixed and vice-versa) over the cone of symmetric positive definite matrices, the stationary point obtained by solving the likelihood equation corresponds to the unique global maximizer with respect to that parameter. Indeed, this is what the following lemma proves.

For the sake of convenience, we still adopt the previous convention that $\mathbb{S}^p$ indicates the set of $p \times p$ symmetric matrices, $\mathbb{S}^p_+$ indicates the set of $p \times p$ symmetric positive semidefinite matrices and $\mathbb{S}^p_{++}$ denotes the set of $p \times p$ symmetric positive definite matrices. The proof relies on several standard identities from matrix differential calculus, in particular differentiation rules for the determinant and matrix inverse, together with fundamental trace properties that allow quadratic forms to be expressed in terms of Frobenius norms.

First, by Jacobi's formula, if $\mathbf{A}(t)$ is differentiable and invertible, then

$$\frac{\mathrm{d}}{\mathrm{d}t}\log\det\mathbf{A}(t) = \operatorname{tr}(\mathbf{A}(t)^{-1}\mathbf{A}'(t)).$$

Second, differentiating the identity $\mathbf{A}(t)\mathbf{A}(t)^{-1} = \mathbf{I}$ yields the derivative of the inverse,

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{A}(t)^{-1} = -\mathbf{A}(t)^{-1}\mathbf{A}'(t)\mathbf{A}(t)^{-1}.$$

We also use the cyclic property of the trace, $\operatorname{tr}(\mathbf{ABC}) = \operatorname{tr}(\mathbf{BCA})$, the identity $\operatorname{tr}(\mathbf{A}^\top\mathbf{A}) = \|\mathbf{A}\|_{\mathrm{F}}^2$ relating trace to the Frobenius norm, and the factorization $\mathbf{S} = \mathbf{S}^{1/2}\mathbf{S}^{1/2}$ for symmetric positive semidefinite matrices $\mathbf{S}$. These well-known results allow us to express second derivatives in terms of squared Frobenius norms, from which concavity properties follow directly.

Lemma 6 (Concavity of the matrix-normal log-likelihood in Ψ)**.** Fix $\mathbf{M} \in \mathbb{R}^{p\times q}$ and $\Sigma \in \mathbb{S}_{++}^p$. Let $\mathcal{X}_1, \ldots, \mathcal{X}_n \in \mathbb{R}^{p\times q}$ be observed matrices and define

$$\mathbf{S} \;=\; \sum_{i=1}^{n} (\mathcal{X}_i - \mathbf{M})^\top \Sigma^{-1} (\mathcal{X}_i - \mathbf{M}) \;\in\; \mathbb{S}_+^q.$$

Consider the (partial) log-likelihood of the matrix-variate normal model as a function of $\Psi \in \mathbb{S}_{++}^q$, up to an additive constant independent of Ψ,

$$\ell(\Psi) = -\frac{np}{2}\log\det(\Psi) - \frac{1}{2}\operatorname{tr}\big(\Psi^{-1}\mathbf{S}\big). \tag{2.1}$$

Then $\ell(\Psi)$ is concave on $\mathbb{S}_{++}^q$. Moreover, if $\mathbf{S} \succ \mathbf{0}$, then $\ell(\Psi)$ is strictly concave on $\mathbb{S}_{++}^q$.

Proof. Let $\Psi \in \mathbb{S}_{++}^q$ be fixed and let $\mathbf{H} \in \mathbb{S}^q$ be an arbitrary symmetric matrix. Recall that $\mathbb{S}_{++}^q$ is an open convex subset of the vector space $\mathbb{S}^q$. Consequently, for every symmetric direction $\mathbf{H}$, there exists an open interval around $t = 0$ such that

$$\Psi(t) := \Psi + t\mathbf{H}$$

remains in $\mathbb{S}_{++}^q$.

To establish concavity of ℓ on $\mathbb{S}_{++}^q$, it is sufficient to analyze its behavior along straight lines contained in the domain. Indeed, a twice differentiable function is concave if and only if its second directional derivative is nonpositive in every direction. Therefore, we consider the one–dimensional restriction

$$\phi(t) = \ell(\Psi(t)),$$

defined for t in a neighborhood of 0 such that $\Psi(t) \in \mathbb{S}_{++}^q$. Our goal is to show that the relation $\phi''(0) \le 0$ holds for every symmetric direction $\mathbf{H}$, thereby establishing concavity of ℓ on the entire domain. For notational convenience, define $\mathbf{P}(t) := (\Psi + t\mathbf{H})^{-1}$, which corresponds to the inverse along this path. Consequently, one has that $\mathbf{P} = \mathbf{P}(0) = \Psi^{-1}$.

Log-determinant term.

We first analyze the second derivative of the log-determinant term, as it determines the intrinsic curvature contribution of the determinant component of the likelihood. Using the well-known matrix calculus identities

$$\frac{\mathrm{d}}{\mathrm{d}t}\log\det(\Psi+t\mathbf{H})=\operatorname{tr}(\mathbf{P}(t)\mathbf{H}),\qquad \frac{\mathrm{d}}{\mathrm{d}t}\mathbf{P}(t)=-\mathbf{P}(t)\mathbf{H}\mathbf{P}(t),$$

Differentiating once more with respect to t, and applying the product rule together with the derivative of $\mathbf{P}(t)$, we obtain

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\log\det(\Psi+t\mathbf{H})=-\operatorname{tr}\big(\mathbf{P}(t)\mathbf{H}\mathbf{P}(t)\mathbf{H}\big).$$

Evaluating the second derivative at $t=0$, and recalling that $\mathbf{P}(0)=\Psi^{-1}=:\mathbf{P}$, we obtain

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\log\det(\Psi+t\mathbf{H})\bigg|_{t=0}=-\operatorname{tr}(\mathbf{PHPH}).$$

Since $\mathbf{P}\succ 0$, it admits an unique square root $\mathbf{P}^{1/2}\in\mathbb{S}^q_{++}$, and this representation allows us to express the quadratic form in terms of a Frobenius norm. In particular, we may rewrite it as next

$$\operatorname{tr}(\mathbf{PHPH})=\|\mathbf{P}^{1/2}\mathbf{H}\mathbf{P}^{1/2}\|_{\mathrm{F}}^2\geq 0,$$

which is nonnegative, since it is the squared Frobenius norm of a matrix. Consequently, one has

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\log\det(\Psi+t\mathbf{H})\bigg|_{t=0}=-\|\mathbf{P}^{1/2}\mathbf{H}\mathbf{P}^{1/2}\|_{\mathrm{F}}^2\leq 0.$$

Therefore, the log-determinant function is concave on $\mathbb{S}^q_{++}$.

Trace–inverse term.

We now turn our attention to the contribution of the trace–inverse term to the curvature of the log-likelihood. This component plays a complementary role to the log-determinant term and requires a separate analysis. To this end, consider the scalar-valued function

$$g(t)=\operatorname{tr}(\mathbf{P}(t)\mathbf{S}),$$

where $\mathbf{S}\succeq\mathbf{0}$ is fixed. This quantity measures how the inverse covariance matrix $\mathbf{P}(t)$ interacts with the empirical scatter matrix $\mathbf{S}$, and understanding its second-order behavior will allow us to determine its effect on the overall curvature.

We begin by computing the first derivative of $g(t)$. Differentiating with respect to t, and making use of the identity $\mathbf{P}'(t)=-\mathbf{P}(t)\mathbf{H}\mathbf{P}(t)$, we obtain

$$g'(t)=-\operatorname{tr}\big(\mathbf{P}(t)\mathbf{H}\mathbf{P}(t)\mathbf{S}\big).$$

This expression already reflects the interaction between the direction $\mathbf{H}$ and the inverse matrix $\mathbf{P}(t)$, but to assess convexity we must proceed further and examine the second derivative.

Differentiating once more and applying the product rule carefully, we arrive at

$$g''(t) = 2\,\mathrm{tr}\big(\mathbf{P}(t)\mathbf{H}\mathbf{P}(t)\mathbf{H}\mathbf{P}(t)\mathbf{S}\big).$$

At this stage, the structure of the expression is not yet fully transparent. In order to better understand its sign, it is convenient to rewrite it in a form that highlights its intrinsic nonnegativity.

To this end, we introduce the matrices

$$\mathbf{K}(t) := \mathbf{P}(t)^{1/2}\mathbf{H}\mathbf{P}(t)^{1/2}, \qquad \mathbf{Q}(t) := \mathbf{P}(t)^{1/2}\mathbf{S}\mathbf{P}(t)^{1/2}.$$

These definitions allow us to symmetrize the expression and isolate its essential components. Since $\mathbf{S} \succeq 0$ and $\mathbf{P}(t) \succ 0$, it follows immediately that $\mathbf{Q}(t) \succeq 0$ ($\mathbf{Q}(t)$ is positive semidefinite).

Using cyclicity of the trace, we can now rewrite the second derivative as

$$\mathrm{tr}\big(\mathbf{P}(t)\mathbf{H}\mathbf{P}(t)\mathbf{H}\mathbf{P}(t)\mathbf{S}\big) = \mathrm{tr}\big(\mathbf{K}(t)\mathbf{Q}(t)\mathbf{K}(t)\big).$$

This representation is particularly useful because it allows us to express the quantity as a squared Frobenius norm. Indeed, by standard matrix identities and the fact that $\mathbf{K}(t)$ is symmetric, one has that

$$\mathrm{tr}\big(\mathbf{K}(t)\mathbf{Q}(t)\mathbf{K}(t)\big) = \big\|\mathbf{Q}(t)^{1/2}\mathbf{K}(t)\big\|_{\mathrm{F}}^2 \geq 0.$$

This shows that the second derivative $g''(t)$ is nonnegative for all admissible values of t.

Therefore, we conclude that $g''(t) \geq 0$ in a neighborhood of 0, and in particular $g''(0) \geq 0$. This establishes that the mapping

$$\Psi \;\mapsto\; \mathrm{tr}(\Psi^{-1}\mathbf{S})$$

is convex on $\mathbb{S}^q_{++}$, as desired.

Conclusion.

We are now in a position to combine the contributions of both components of the log-likelihood. From the previous calculations, the second directional derivative of ℓ at Ψ in the direction $\mathbf{H}$ is given by

$$\phi''(0) = -\frac{np}{2}\|\mathbf{P}^{1/2}\mathbf{H}\mathbf{P}^{1/2}\|_{\mathrm{F}}^2 - \|\mathbf{Q}(0)^{1/2}\mathbf{K}(0)\|_{\mathrm{F}}^2.$$

Since both terms on the right-hand side are nonnegative, it follows immediately that

$$\phi''(0) \leq 0 \quad \text{for all } \mathbf{H} \in \mathbb{S}^q.$$

Therefore, ℓ is concave on $\mathbb{S}^q_{++}$.

Finally, suppose that $\mathbf{S} \succ 0$. In this case, $\mathbf{Q}(0) = \mathbf{P}^{1/2}\mathbf{S}\mathbf{P}^{1/2} \succ 0$, so that $\mathbf{Q}(0)^{1/2}$ is invertible. It follows that both quadratic forms above vanish if and only if $\mathbf{H} = 0$. Hence, for every nonzero direction $\mathbf{H}$, we have $\phi''(0) < 0$, which establishes that the concavity is strict. $\square$

From the score equation derived above, we obtain the relation

$$\mathbf{S} = pn\Psi, \qquad \text{where} \qquad \mathbf{S} = \sum_{i=1}^{n} (\mathcal{X}_i - \mathbf{M})^\top \Sigma^{-1} (\mathcal{X}_i - \mathbf{M}).$$

Solving for Ψ yields the (conditional) maximizer of the log-likelihood with respect to Ψ when $\mathbf{M}$ and Σ are held fixed, namely,

$$\widehat{\Psi} = \frac{1}{pn} \sum_{i=1}^{n} (\mathcal{X}_i - \widehat{\mathbf{M}})^\top \widehat{\Sigma}^{-1} (\mathcal{X}_i - \widehat{\mathbf{M}}). \tag{2.2}$$

By Lemma 6, the function $\ell(\Psi)$ is concave on $\mathbb{S}^q_{++}$ (strictly concave whenever $\mathbf{S} \succ \mathbf{0}$), so this stationary point is the unique global maximizer in $\mathbb{S}^q_{++}$ under the stated conditions.

The expression in (2.2) has a natural interpretation: $\widehat{\Psi}$ is a *weighted empirical column covariance* matrix of the centered observations. The weighting matrix $\widehat{\Sigma}^{-1}$ accounts for the dependence across rows, so that the remaining quadratic form captures variability attributable to the column structure.

By symmetry of the matrix-normal model, an entirely analogous calculation is obtained by interchanging the roles of rows and columns. In particular, if we fix $\mathbf{M}$ and Ψ and maximize the log-likelihood with respect to $\Sigma \in \mathbb{S}^p_{++}$, we obtain

$$\widehat{\Sigma} = \frac{1}{qn} \sum_{i=1}^{n} (\mathcal{X}_i - \widehat{\mathbf{M}}) \widehat{\Psi}^{-1} (\mathcal{X}_i - \widehat{\mathbf{M}})^\top. \tag{2.3}$$

The estimator in (2.3) may be viewed as the *weighted empirical row covariance* matrix, where the weighting $\widehat{\Psi}^{-1}$ adjusts for dependence across columns.

An important feature of these estimators is that the likelihood equations for Σ and Ψ are coupled, since each depends on the other. As a result, they cannot be solved simultaneously in closed form. In practice, numerical schemes such as the flip–flop algorithm of Dutilleul (1999) or ECM-type procedures are employed to obtain the maximum likelihood estimates.

2.4 Skewed Matrix Variate Distributions

Following the results presented in Chen and Gupta (2005), the authors are credited with introducing one of the first definitions of skew-normal distributions in the context of random matrices. Their contribution extends the classical skew-normal framework to matrix-valued random variables, thereby broadening its applicability to multivariate and structured data settings.

To introduce this concept, we begin by briefly recalling the definition of the skew-normal distribution in the univariate case, followed by its multivariate extension. These formulations provide the necessary foundation for understanding the matrix-variate generalization.

In what follows, we present one possible construction of skew-normal distributions in the univariate, vector, and matrix settings. In subsequent chapters of this thesis, however, we adopt an alternative definition of the matrix-variate skew-normal distribution. This alternative formulation is more convenient for theoretical developments and inference procedures, as it admits a tractable stochastic representation.

Definition 6. Given a univariate random variable $X \in \mathbb{R}$, we claim that it has Skew Normal (SN from now on) distribution iff its probability density function is given by:

$$f_{\text{SN}}(x \mid \lambda) := 2\phi(x)\Phi(\lambda x)$$

where ϕ and Φ are, respectively, the probability density function (PDF) and the cumulative distribution function (CDF) from a standard normal distribution.

Here, λ is the skewness parameter, in accordance to Azzalini (1985).

Definition 7. A random vector $\mathbf{X} \in \mathbb{R}^p$ is said to follow a Multivariate Skew Normal (MSN from now on) distribution if, and only if, its joint PDF is:

$$f_{\text{MSN}}(\mathbf{X}) := 2\phi_p(\mathbf{X}, \Omega)\Phi(\lambda^\top \mathbf{X}),$$

where the skewness vector $\lambda^\top \in \mathbb{R}^p$. Here, ϕ_p and Φ are the joint density of a p-dimensional normally distributed random vector and the CDF of an univariate standard normal variable, respectively. Its definition and properties can be checked at Azzalini and Capitanio (1999).

Definition 8. A random matrix $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{p\times q}$ is said to follow a Matrix Variate Skew Normal (MVSN from now on) distribution iff its PDF takes the following form:

$$f_{\text{MVSN}}(\boldsymbol{\mathcal{X}}) = c^\star \phi_{p\times q}(\boldsymbol{\mathcal{X}}, \Sigma, \Psi)\Phi_q(\boldsymbol{\mathcal{X}}^\top \mathbf{b}, \Psi),$$

where $\phi_{p\times q}(\boldsymbol{\mathcal{X}}, \Sigma, \Psi)$ is the PDF of the MVN distribution with location matrix given by $\mathbf{M} = 0_{p\times q}$ and (positive definite) covariance matrices $\Sigma \in \mathbb{R}^{p\times p}$ and $\Psi \in \mathbb{R}^{q\times q}$. Also, one has that $\mathbf{b} \in \mathbb{R}^p$ is the skewness vector, $\Phi_q(\boldsymbol{\mathcal{X}}^\top \mathbf{b}, \Psi)$ is the CDF of a q-dimensional MN vector with covariance Ψ, location matrix $\mathbf{M} = 0$ evaluated at $\boldsymbol{\mathcal{X}}^\top \mathbf{b}$, and $c^\star = (\Phi_q(0, (1 + \mathbf{b}^\top \Sigma \mathbf{b})\Psi))^{-1}$. Hence the MVSN random matrix is denoted as $\boldsymbol{\mathcal{X}} \sim \text{MVSN}_{p\times q}(\mathbf{b}, \Sigma, \Psi)$.

The verification that such mapping is a joint PDF indeed is documented in Chen and Gupta (2005). Additionally, the authors from the article Ning and Gupta (2012) discuss a generalization of the above-mentioned definition, which they name as the matrix variate extended skew normal (MVESN from now on) distribution. Furthermore, their investigations deal with the moment generating function, the distribution of the quadratic and the linear forms, as well as the analysis of marginal and conditional distributions within this family.

In addition to the definition and results presented in Chen and Gupta (2005) and Ning and Gupta (2012), several other articles explore variants of skewed matrix variate distributions.

Notably, in Gallaugher and McNicholas (2019), researchers introduce three distinct skewed matrix variate distributions based on the matrix normal variance-mean mixture. Specifically, they propose a matrix variate generalized hyperbolic distribution based on the generalized inverse Gaussian (GIG), a matrix variate variance-gamma distribution based on the gamma density and a matrix variate normal inverse Gaussian distribution based on the inverse Gaussian (IG). Still, as outlined in Gallaugher and McNicholas (2017), a matrix variate skew-t distribution is also derived through a matrix normal variance-mean mixture based on the inverse Gaussian (IG).

As previously mentioned, each of the three different skewed matrix variate distributions is defined as a particular choice of the matrix normal variance-mean mixture. For the sake of completeness, we provide here the stochastic representation for them:

$$\boldsymbol{\mathcal{X}} = \mathbf{M} + W\mathbf{A} + \sqrt{W}\boldsymbol{\mathcal{V}}$$

where $\boldsymbol{\mathcal{V}} \sim \mathcal{N}_{p\times q}(0_{p\times q}, \Sigma, \Psi)$. In such context, $0_{p\times q}$ represents the $p \times q$ zero matrix, $\mathbf{M}$ is an $p \times q$ location matrix, $\mathbf{A}$ is an $p \times q$ skewness matrix and $W > 0$ is a random variable with probability density function given by $h(w|\theta)$.

Starting with this chapter, we shift our focus to parameter estimation within the matrix-variate framework. We concentrate on the matrix-variate skew-normal distribution and its censored version, aiming to estimate their parameters via maximum likelihood. To accomplish this, we employ the ECM algorithm, applied to both fully observed data and data subject to interval censoring or missingness.

CHAPTER

3

MVSN PROPERTIES AND PARAMETER ESTIMATION

In this chapter, we introduce the matrix-variate skew-normal (MVSN) distribution through a stochastic representation involving a truncated latent variable, which provides a flexible and analytically tractable framework for modeling asymmetry in matrix-valued data. While the formulation is closely related to that of Naderi *et al.* (2020), the treatment developed here is self-contained and includes results that extend the existing literature. In addition, we propose an extended version (as in the reference Naderi *et al.* (2024)), termed the matrix-variate extended skew-normal (MVESN) distribution, and investigate its main structural properties. An EM-type algorithm is derived for maximum likelihood estimation under the MVSN model, allowing for efficient and stable inference. The proposed methodology is evaluated through simulation studies and further illustrated by an application to a historical Dow–Jones dividend dataset, highlighting its ability to capture asymmetric dependence structures in real data.

Before presenting the formal definition of the proposed distribution, we first establish the notation and basic conventions that will be used consistently throughout the chapter. The symbol $U \perp Z$ indicates that the random quantities U and Z are independent. Throughout, $\mathcal{N}_p(\mu,\Sigma)$ denotes the p-variate normal distribution with mean vector μ and covariance matrix Σ, with $\phi_p(\cdot \mid \mu,\Sigma)$ and $\Phi_p(\cdot \mid \mu,\Sigma)$ representing its probability density function (PDF) and cumulative distribution function (CDF), respectively. In the univariate case ($p=1$), we omit the index p, and when $\mu=0$ and $\sigma^2=1$, we simply write $\phi(\cdot)$ and $\Phi(\cdot)$ for the standard normal PDF and CDF. We also employ the notation $\mathrm{TN}(\mu,\sigma^2;[-\kappa,+\infty))$ for a truncated normal distribution with mean μ, variance σ^2, and support restricted to $[-\kappa,+\infty)$. In the special case $\mu=0$, $\sigma^2=1$ and $\kappa=0$, this reduces to the half-normal distribution $\mathrm{HN}(0,1)$, that is, $\mathrm{TN}(0,1;[0,+\infty))$. As is standard in probability theory and statistical modeling, we use upper-case letters to denote random variables, lower-case letters for their realizations, and boldface symbols to represent vectors and matrices, ensuring a clear and consistent notation throughout.

3.1 The Multivariate Skew Normal Distribution

We start this section by introducing the multivariate skew-normal (MSN) distribution, not only because of its inherent theoretical and practical significance, but also due to its foundational role in establishing a formal connection with the matrix-variate case. By presenting the multivariate form first, we make this relationship explicit, laying out a clear and logical pathway that naturally leads to the subsequent development of the matrix-valued formulation. Although there are different ways to describe the multivariate skew normal distribution, we shall prove that they are equivalent.

Definition 9. In accordance with the reference Lachos *et al.* (2007), we claim that the random vector $\mathbf{Z}$ has multivariate skew normal distribution if and only if it satisfies the stochastic representation:

$$\mathbf{Z} = \mu + \Omega^{1/2}(\delta|T_0| + (\mathbf{I}_p - \delta\delta^\top)^{1/2}\mathbf{T}_1),$$

where $\mu \in \mathbb{R}^p$ is the location vector, $\Omega \in \mathbb{R}^{p\times p}$ is a symmetric and positive definite matrix, $\lambda \in \mathbb{R}^p$ is the skewness vector, $\delta := \lambda(1+\lambda^\top\lambda)^{-1/2}$, $T_0 \sim \mathcal{N}(0,1)$, $\mathbf{T}_1 \sim \mathcal{N}_p(\mathbf{0},\mathbf{I}_p)$ and $|T_0| \perp \mathbf{T}_1$. We denote such relation by $\mathbf{Z} \sim \text{MSN}_p^\star(\mu,\Omega,\lambda)$

Definition 10. In accordance with the reference Azzalini and Valle (1996), we claim that the random vector $\mathbf{Z}$ has multivariate skew normal distribution iff it has the stochastic representation:

$$\mathbf{Z} = \mu + |X_0|\mathbf{b} + \mathbf{V},$$

where $\mu \in \mathbb{R}^p$ is the location vector, $X_0 \sim \mathcal{N}(0,1)$, $\mathbf{b} \in \mathbb{R}^p$ is the skewness vector, the distribution of $\mathbf{V}$ is given by $\mathbf{V} \sim \mathcal{N}_p(\mathbf{0},\Delta)$ so that Δ is a symmetric and positive definite matrix and $|X_0| \perp \mathbf{V}$. We denote such relation by $\mathbf{Z} \sim \text{MSN}_p(\mu,\Delta,\mathbf{b})$

Proposition 6. Suppose $\mathbf{Z}_1 \sim \text{MSN}_p^\star(\mu,\Omega,\lambda)$ and $\mathbf{Z}_2 \sim \text{MSN}_p(\mu,\Delta,\mathbf{b})$. Then $\mathbf{Z}_1 \overset{d}{=} \mathbf{Z}_2$.

Proof. Since stochastic representations fully characterize the distribution, to prove that $\mathbf{Z}_1 \overset{d}{=} \mathbf{Z}_2$ amounts to identifying $\mathbf{b}$ and Δ in the stochastic of $\mathbf{Z}_2$ in terms of Ω and δ in the stochastic representation of $\mathbf{Z}_1$. To do so, it suffices to make the (bijective) change of variables $\mathbf{b} := \Omega^{1/2}\delta$ and $\Delta := \Omega^{1/2}(\mathbf{I}_p - \delta\delta^\top)^{1/2}\Omega^{1/2}$. By the definitions involved, one has $|X_0| \overset{d}{=} |T_0| \sim \text{HN}(0,1)$. For the purpose of completeness sake, let us stress out the relation between the stochastic representation of $\mathbf{Z}_1$ and the proposed change of variables:

$$\mathbf{Z}_1 = \mu + \underbrace{\Omega^{1/2}\delta}_{\mathbf{b}}|T_0| + \underbrace{\Omega^{1/2}(\mathbf{I}_p - \delta\delta^\top)^{1/2}\mathbf{T}_1}_{\mathbf{V}}$$

The first part is exactly of the form $|T_0|\mathbf{b}$, with $\mathbf{b} := \Omega^{1/2}\delta$. The second part (call it $\mathbf{V}$) is Gaussian since it is a linear transformation of $\mathbf{T}_1 \sim \mathcal{N}_p(\mathbf{0},\mathbf{I}_p)$:

$$\mathbf{V} \sim \mathcal{N}_p(\mathbf{0}, \underbrace{\Omega^{1/2}(\mathbf{I}_p - \delta\delta^\top)\Omega^{1/2}}_{\Delta})$$

where $|T_0| \perp \mathbf{T}_1$ implies that $|T_0| \perp \mathbf{V}$ (once $\mathbf{V}$ is measurable function of $\mathbf{T}_1$). So we obtain

$$\mathbf{Z}_1 \overset{d}{=} \mu + |T_0|\mathbf{b} + \mathbf{V},$$

with $\mathbf{b}$ and $\mathbf{V}$ as defined above. On the other hand, the stochastic representation of $\mathbf{Z}_2$ is:

$$\mathbf{Z}_2 = \mu + |X_0|\mathbf{b} + \mathbf{V}.$$

This matches the stochastic representation of $\mathbf{Z}_1$ exactly, provided we set the above-mentioned change of variables. Since both $|T_0|$ and $|X_0|$ are identically distributed, the two stochastic representations are equivalent, from whence it can be concluded they are equally distributed, just as we wanted to demonstrate. □

Proposition 7. Let $\mathbf{X} \sim \text{MSN}_p^\star(\mu, \Omega, \lambda)$. Then its probability density function is given by:

$$f_{\text{MSN}}(\mathbf{x}) = 2\phi_p(\mathbf{x} \mid \mu, \Omega)\Phi(\lambda^\top \Omega^{-1/2}(\mathbf{x} - \mu)).$$

where $\phi_p(\cdot \mid \mu, \Omega)$ is the p-variate normal density and $\Phi(\cdot)$ is the standard normal cdf.

Proof. The reader is invited to check out the reference Lachos *et al.* (2007). □

Proposition 8. Let $\mathbf{X} \sim \text{MSN}_p(\mu, \Delta, \mathbf{b})$. Therefore, the probability density function of $\mathbf{X} \in \mathbb{R}^p$ can be written as

$$f_{\text{MSN}}(\mathbf{x}) = \frac{2\,\phi_p(\mathbf{x} \mid \mu, \Delta)}{\sqrt{1 + \mathbf{b}^\top \Delta^{-1}\mathbf{b}}} \exp\left\{\frac{1}{2}\frac{(\mathbf{b}^\top \Delta^{-1}(\mathbf{x} - \mu))^2}{(1 + \mathbf{b}^\top \Delta^{-1}\mathbf{b})}\right\} \Phi\left(\frac{\mathbf{b}^\top \Delta^{-1}(\mathbf{x} - \mu)}{\sqrt{1 + \mathbf{b}^\top \Delta^{-1}\mathbf{b}}}\right),$$

where $\phi_p(\cdot \mid \mu, \Delta)$ is the p-variate normal density and $\Phi(\cdot)$ is the standard normal cdf.

Proof. This result is a particular case of the Proposition 19, which we prove later on. □

Proposition 9. If the vector-valued random variable $\mathbf{X} \in \mathbb{R}^p$ satisfies $\mathbf{X} \sim \text{MSN}_p(\mu, \Delta, \mathbf{b})$ as mentioned in the Definition 10, then its first and second moments are given by:

$$\mathbb{E}(\mathbf{X}) = \mu + \sqrt{\frac{2}{\pi}}\mathbf{b} \quad \text{and} \quad \text{Var}(\mathbf{X}) = \Delta + \left(1 - \frac{2}{\pi}\right)\mathbf{b}\mathbf{b}^\top.$$

Proof. Given the corresponding stochastic representation, it suffices to apply the linearity of $\mathbb{E}$ to obtain $\mathbb{E}(\mathbf{X})$ and the independence between $|X_0|$ and $\mathbf{V}$ to calculate the covariance $\text{Var}(\mathbf{X})$. □

3.2 The MVSN Distribution

We begin this section by formally introducing the matrix variate skew-normal (MVSN) distribution through its stochastic representation as also proposed by Naderi *et al.* (2020). This formulation provides a clear foundation for the theoretical developments that follow and is particularly valuable in practical settings, as the stochastic representation greatly facilitates parameter estimation in the presence of censored data.

Definition 11. A random matrix $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{p\times q}$ is said to follow a MVSN distribution with parameters $\mathbf{M} \in \mathbb{R}^{p\times q}$, $\mathbf{A} \in \mathbb{R}^{p\times q}$, $\Sigma \in \mathbb{R}^{p\times p}$ and $\Psi \in \mathbb{R}^{q\times q}$ iff it admits the next stochastic representation:

$$\boldsymbol{\mathcal{X}} = \mathbf{M} + W\mathbf{A} + \boldsymbol{\mathcal{V}},$$

where $\boldsymbol{\mathcal{V}} \sim \mathcal{N}_{p\times q}(\mathbf{0}, \Sigma, \Psi)$, $W \sim \mathrm{HN}(0,1)$ is a positive scalar latent variable independent of $\boldsymbol{\mathcal{V}}$, $\mathbf{M}$ is a location matrix, $\mathbf{A}$ is a skewness matrix, Σ is the positive definite symmetric row scale matrix and Ψ is the positive definite symmetric column scale matrix, respectively. We denote this relation among $\boldsymbol{\mathcal{X}}$ and its parameters by $\boldsymbol{\mathcal{X}} \sim \mathrm{MVSN}_{p\times q}(\mathbf{M}, \mathbf{A}, \Sigma, \Psi)$.

Remark. The joint probability density function of the random pair $(\boldsymbol{\mathcal{X}}, W)$ is given by:

$$f_{\boldsymbol{\mathcal{X}},W}(\mathcal{X}, w) = \frac{1}{(2\pi)^{pq/2}|\Sigma|^{q/2}|\Psi|^{p/2}} \sqrt{\frac{2}{\pi}} \exp\left\{ -\frac{1}{2}\mathrm{tr}(\Sigma^{-1}(\mathcal{Z} - w\mathbf{A})\Psi^{-1}(\mathcal{Z} - w\mathbf{A})^\top) - \frac{w^2}{2} \right\},$$

where $w \in \mathbb{R}_{>0}$ and we set up that $\boldsymbol{\mathcal{Z}} := \boldsymbol{\mathcal{X}} - \mathbf{M}$. In order to conclude so, let us consider the Definition 11. If the random matrix $\boldsymbol{\mathcal{X}} \sim \mathrm{MVSN}(\mathbf{M}, \mathbf{A}, \Sigma, \Psi)$ is conditioned on the latent variable $W = w$, one obtains the following result:

$$\boldsymbol{\mathcal{X}} \mid W = w \sim \mathcal{N}_{p\times q}(\mathbf{M} + w\mathbf{A}, \Sigma, \Psi).$$

Building on this, to establish the desired relation it is enough to invoke the definition of the probability density function of W together with the product rule, from whence the desired relation holds.

Proposition 10. If $\boldsymbol{\mathcal{X}} \sim \mathrm{MVSN}_{p\times q}(\mathbf{M}, \mathbf{A}, \Sigma, \Psi)$, then the marginal density of $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{p\times q}$ is

$$f_{\mathrm{MVSN}}(\mathcal{X}) = \frac{2}{\tau} \exp\left(\frac{d_{\mathbf{A}}^2(\mathcal{X})}{2\tau^2} \right) \phi_{p\times q}(\mathbf{X} \mid \mathbf{M}, \Sigma, \Psi) \Phi\left(\frac{d_{\mathbf{A}}(\mathcal{X})}{\tau} \right),$$

where $\phi_{p\times q}(\cdot \mid \mathbf{M}, \Sigma, \Psi)$ is the PDF of the MVN distribution, $\tau^2 := 1 + \mathrm{tr}(\Sigma^{-1}\mathbf{A}\Psi^{-1}\mathbf{A}^\top)$ and $d_{\mathbf{A}}(\mathcal{X}) := \mathrm{tr}(\Sigma^{-1}\mathcal{Z}\Psi^{-1}\mathbf{A}^\top)$.

Proof. As previously mentioned in the remark, due to the product rule, the PDF of the random pair $(\boldsymbol{\mathcal{X}}, W)$ is expressed as the product $f_{\boldsymbol{\mathcal{X}}|W}(\mathcal{X} \mid w) f_W(w)$. Therefore, the marginal density of $\boldsymbol{\mathcal{X}}$ is obtained by integrating this joint density with respect to w, which results into the expression:

$$f_{\boldsymbol{\mathcal{X}}}(\mathcal{X}) = \int_0^\infty f_{\boldsymbol{\mathcal{X}}|W}(\mathcal{X} \mid w) f_W(w) \mathrm{d}w.$$

Let us keep the same notation $\mathcal{Z} := \mathcal{X} - \mathbf{M}$ as before. If we expand the trace term, due to the linearity and the cyclic properties of the trace function, it can be rewritten as follows:

$$\mathrm{tr}(\Sigma^{-1}(\mathcal{Z} - w\mathbf{A})\Psi^{-1}(\mathcal{Z} - w\mathbf{A})^\top) = \mathrm{tr}(\Sigma^{-1}\mathcal{Z}\Psi^{-1}\mathcal{Z}^\top) - 2w\,\mathrm{tr}(\Sigma^{-1}\mathcal{Z}\Psi^{-1}\mathbf{A}^\top) + w^2\,\mathrm{tr}(\Sigma^{-1}\mathbf{A}\Psi^{-1}\mathbf{A}^\top)$$

In favor of convenience, we shall adopt the notations (restricted to the context of this proposition):

$$\begin{aligned} a &:= \mathrm{tr}(\Sigma^{-1}\mathbf{A}\Psi^{-1}\mathbf{A}^\top) + 1, \\ b &:= \mathrm{tr}(\Sigma^{-1}\mathcal{Z}\Psi^{-1}\mathbf{A}^\top), \\ c &:= \mathrm{tr}(\Sigma^{-1}\mathcal{Z}\Psi^{-1}\mathcal{Z}^\top), \end{aligned}$$

Consequently, the probability density function $f_{\boldsymbol{\mathcal{X}}}(\mathcal{X})$ can also be reformulated in the following fashion in terms of the labels a, b and c:

$$f_{\boldsymbol{\mathcal{X}}}(\mathcal{X}) \propto \int_0^{\infty} \exp\left\{-\frac{1}{2}\left(aw^2 - 2bw + c\right)\right\} \mathrm{d}w.$$

Therefore, after completing the squares inside the exp function, one arrives at the expression:

$$-\frac{1}{2}(aw^2 - 2bw + c) = -\frac{a}{2}\left(w - \frac{b}{a}\right)^2 - \frac{1}{2}\left(c - \frac{b^2}{a}\right).$$

Substituting into the integral, the marginal density becomes

$$f_{\boldsymbol{\mathcal{X}}}(\mathcal{X}) = \frac{1}{(2\pi)^{pq/2}|\Sigma|^{q/2}|\Psi|^{p/2}}\sqrt{\frac{2}{\pi}}\exp\left\{-\frac{1}{2}\left(c - \frac{b^2}{a}\right)\right\}\int_0^{\infty}\exp\left\{-\frac{a}{2}\left(w - \frac{b}{a}\right)^2\right\}\mathrm{d}w.$$

If we evaluate the integral through convenient changes of variables, it results that:

$$\int_0^{\infty}\exp\left\{-\frac{a}{2}\left(w - \frac{b}{a}\right)^2\right\}\mathrm{d}w = \sqrt{\frac{2\pi}{a}}\Phi\left(\frac{b}{\sqrt{a}}\right),$$

Hence the final expression for the marginal PDF of $\boldsymbol{\mathcal{X}}$ corresponds to:

$$f_{\mathrm{MVSN}}(\mathcal{X}) = \frac{2}{(2\pi)^{pq/2}|\Sigma|^{q/2}|\Psi|^{p/2}\sqrt{a}}\exp\left\{-\frac{1}{2}\left(c - \frac{b^2}{a}\right)\right\}\Phi\left(\frac{b}{\sqrt{a}}\right),$$

Thence, if we replace a, b and c by their corresponding definitions, this procedure yields the explicit expression for the desired marginal probability density function:

$$\begin{aligned} f_{\mathrm{MVSN}}(\mathcal{X}) = {} & \frac{2}{(2\pi)^{pq/2}|\Sigma|^{q/2}|\Psi|^{p/2}\sqrt{1+\mathrm{tr}(\Sigma^{-1}\mathbf{A}\Psi^{-1}\mathbf{A}^{\top})}} \\ & \times \exp\left(-\frac{1}{2}\left[\mathrm{tr}(\Sigma^{-1}(\mathcal{X}-\mathbf{M})\Psi^{-1}(\mathcal{X}-\mathbf{M})^{\top}) - \frac{\left(\mathrm{tr}(\Sigma^{-1}(\mathcal{X}-\mathbf{M})\Psi^{-1}\mathbf{A}^{\top})\right)^2}{1+\mathrm{tr}(\Sigma^{-1}\mathbf{A}\Psi^{-1}\mathbf{A}^{\top})}\right]\right) \\ & \times \Phi\left(\frac{\mathrm{tr}(\Sigma^{-1}(\mathcal{X}-\mathbf{M})\Psi^{-1}\mathbf{A}^{\top})}{\sqrt{1+\mathrm{tr}(\Sigma^{-1}\mathbf{A}\Psi^{-1}\mathbf{A}^{\top})}}\right), \end{aligned}$$

By combining terms (that is to say, by indentifying the expression of the MVN PDF inside the obtained formula), the sought PDF can be expressed as a product involving the matrix variate normal distribution density and a skewness correction factor:

$$\begin{aligned} f_{\mathrm{MVSN}}(\mathcal{X}) = {} & \frac{2}{\sqrt{1+\mathrm{tr}\left(\Sigma^{-1}\mathbf{A}\Psi^{-1}\mathbf{A}^{\top}\right)}} \\ & \times \phi_{p\times q}(\mathcal{X};\mathbf{M},\Sigma,\Psi) \\ & \times \exp\left\{\frac{1}{2}\frac{\left(\mathrm{tr}\left(\Sigma^{-1}(\mathcal{X}-\mathbf{M})\Psi^{-1}\mathbf{A}^{\top}\right)\right)^2}{1+\mathrm{tr}\left(\Sigma^{-1}\mathbf{A}\Psi^{-1}\mathbf{A}^{\top}\right)}\right\} \\ & \times \Phi\left(\frac{\mathrm{tr}\left(\Sigma^{-1}(\mathcal{X}-\mathbf{M})\Psi^{-1}\mathbf{A}^{\top}\right)}{\sqrt{1+\mathrm{tr}\left(\Sigma^{-1}\mathbf{A}\Psi^{-1}\mathbf{A}^{\top}\right)}}\right), \end{aligned}$$

Finally, we are in conditions to arrive at the proposed formula. Indeed, in accordance with previous conventions $\tau := (\mathrm{tr}(\Sigma^{-1}\mathbf{A}\Psi^{-1}\mathbf{A}^\top)+1)^{1/2}$ and $d_{\mathbf{A}}(\boldsymbol{\mathcal{X}}) := \mathrm{tr}\left(\Sigma^{-1}(\boldsymbol{\mathcal{X}}-\mathbf{M})\Psi^{-1}\mathbf{A}^\top\right)$, the marginal PDF above coincides with the expression exhibited in the body of the proposition. □

Remark. Since this result was only briefly addressed in Naderi *et al.* (2020), we present a complete and self-contained proof here for clarity and completeness.

The MVSN distribution introduces asymmetry through the term $W\mathbf{A}$, while variability and dependence are governed by the Kronecker-structured covariance matrices Σ and Ψ. It is important to observe that the asymmetry is entirely captured by the factor

$$\frac{2}{\tau}\exp\left(\frac{d_{\mathbf{A}}^2(\mathcal{X})}{2\tau^2}\right)\Phi\left(\frac{d_{\mathbf{A}}(\mathcal{X})}{\tau}\right).$$

This term distorts the symmetric structure of the MVN distribution to accommodate the skewness induced by the matrix $\mathbf{A}$. Note that this factor reduces to 1 when $\mathbf{A}=\mathbf{0}$.

It is worth emphasizing that, unlike in the multivariate setting, the product of the first three terms in the MVSN density, namely, the exponential adjustment, the normalization constant, and the matrix normal kernel, does not correspond to a matrix-variate normal distribution with modified covariance matrices. This limitation arises from the presence of a squared trace term, which cannot be recast as a quadratic form under the matrix normal framework. In contrast, after vectorization, the full expression can be interpreted as a multivariate skew normal density with an appropriately adjusted covariance structure. This observation is particularly relevant for establishing equivalences with known multivariate models and is formalized in Proposition 13.

Lemma 7. Suppose that $Z\sim\mathcal{N}(0,1)$. Then the following identity is valid whenever $a\in\mathbb{R}$:

$$\int_{\mathbb{R}}\phi(z)\Phi(az)\mathrm{d}z=\frac{1}{2}$$

where ϕ and Φ are the PDF and CDF of the standard univariate normal distribution, respectively.

Proof. To begin with, let us denote such integral by $f(a)$. Hence, in accordance with the Leibniz' rule and the chain rule, it can be concluded through the Feynman's trick that:

$$\begin{aligned} f'(a) &= \frac{\mathrm{d}}{\mathrm{d}a}\int_{\mathbb{R}}\phi(z)\Phi(az)\mathrm{d}z \\ &= \int_{\mathbb{R}}\frac{\partial}{\partial a}\phi(z)\Phi(az)\mathrm{d}z \\ &= \int_{\mathbb{R}} z\phi(z)\phi(az)\mathrm{d}z \end{aligned}$$

Since the integrand in the previous expression is an odd function and the integration limits are symmetric about the origin, the integral evaluates to zero. This immediately shows that $f'(a)=0$, so the derivative of $f(a)$ vanishes identically, implying that $f(a)$ is constant. To determine this constant, we can evaluate the function at a convenient point, such as $a=0$, where the proposed relation holds, thus completing the argument. □

Theorem 3. The candidate for the probability density function of $\boldsymbol{\mathcal{X}} \sim \text{MVSN}(\mathbf{M}, \mathbf{A}, \Sigma, \Psi)$, as previously mentioned in Proposition 10, integrates to one indeed.

Proof. We start the proof by performing the change of variable $\boldsymbol{\mathcal{Z}} := \Sigma^{-1/2}(\boldsymbol{\mathcal{X}} - \mathbf{M})\Psi^{-1/2}$ and introducing the auxiliary definitions that shall be useful later on:

$$\mathbf{B} := \Sigma^{-1/2}\mathbf{A}\Psi^{-1/2} \quad \text{and} \quad K := \text{tr}(\Sigma^{-1}\mathbf{A}\Psi^{-1}\mathbf{A}^\top).$$

As a result from such conventions, the next identities hold:

$$\text{tr}(\Sigma^{-1}(\mathcal{X}-\mathbf{M})\Psi^{-1}(\mathcal{X}-\mathbf{M})^\top) = \text{tr}(\mathcal{Z}\mathcal{Z}^\top) \quad \text{and} \quad \text{tr}(\Sigma^{-1}(\mathcal{X}-\mathbf{M})\Psi^{-1}\mathbf{A}^\top) = \text{tr}(\mathcal{Z}\mathbf{B}^\top).$$

Due to the identity $\text{vec}(\mathbf{P}\mathcal{Z}\mathbf{Q}) = (\mathbf{Q}^\top \otimes \mathbf{P})\,\text{vec}(\mathcal{Z})$ of the vec operator as well as the kronecker product property $|\mathbf{Q}^\top \otimes \mathbf{P}| = |\mathbf{P}|^{\dim(\mathbf{Q})}|\mathbf{Q}|^{\dim(\mathbf{P})}$, the change of variables formula gives:

$$\begin{aligned}
\mathcal{Z} = \Sigma^{-1/2}(\mathcal{X}-\mathbf{M})\Psi^{-1/2} &\Rightarrow \mathcal{X} = \mathbf{M} + \Sigma^{1/2}\mathcal{Z}\Psi^{1/2} \\
&\Rightarrow \text{vec}(\mathcal{X}) = \text{vec}(\mathbf{M}) + \text{vec}(\Sigma^{1/2}\mathcal{Z}\Psi^{1/2}) \\
&\Rightarrow \text{vec}(\mathcal{X}) = \text{vec}(\mathbf{M}) + (\Psi^{1/2} \otimes \Sigma^{1/2})\,\text{vec}(\mathcal{Z}) \\
&\Rightarrow \text{d}\,\text{vec}(\mathcal{X}) = \text{d}((\Psi^{1/2} \otimes \Sigma^{1/2})\,\text{vec}(\mathcal{Z})) \\
&\Rightarrow \text{d}\,\text{vec}(\mathcal{X}) = |\Psi^{1/2} \otimes \Sigma^{1/2}|\,\text{d}\,\text{vec}(\mathcal{Z}) \\
&\Rightarrow \text{d}\mathcal{X} = |\Sigma|^{q/2}|\Psi|^{p/2}\text{d}\mathcal{Z}
\end{aligned}$$

Accordingly, the original integral reduces to:

$$\int_{\mathbb{R}^{p\times q}} f_{\text{MVSN}}(\mathcal{X})\text{d}\mathcal{X} \propto \int_{\mathbb{R}^{p\times q}} \exp\left(-\frac{1}{2}\text{tr}(\mathcal{Z}\mathcal{Z}^\top) + \frac{\text{tr}^2(\mathcal{Z}\mathbf{B}^\top)}{2(K+1)}\right)\Phi\left(\frac{\text{tr}(\mathcal{Z}\mathbf{B}^\top)}{\sqrt{K+1}}\right)|\Sigma|^{q/2}|\Psi|^{p/2}\text{d}\mathcal{Z}.$$

Now, define the scalar random variable

$$Y := \text{tr}(\mathcal{Z}\mathbf{B}^\top),$$

which can be interpreted as a linear combination of all entries of the matrix $\mathcal{Z}$, weighted by the corresponding entries of the matrix $\mathbf{B}$. Importantly, the integrand in our expression depends on $\mathcal{Z}$ only through the Frobenius norm squared $\|\mathcal{Z}\|_F^2$, the linear combination Y, and its square Y^2. In other words, once $\|\mathcal{Z}\|_F^2$ and Y are known, all remaining terms of the integrand are completely determined. Moreover, notice that the first term in the exponent can be directly expressed in terms of the Frobenius norm as $\text{tr}(\mathcal{Z}\mathcal{Z}^\top) = \|\mathcal{Z}\|_F^2 = \|\,\text{vec}(\mathcal{Z})\|^2$. This identity will be useful for separating the contribution of Y from the other components of $\mathcal{Z}$ that are orthogonal to Y when computing the integral.

Having said that, the integral of interest (denoted by I from now on) becomes:

$$I \propto (2\pi)^{pq/2}|\Sigma|^{q/2}|\Psi|^{p/2}\int_{\mathbb{R}^{pq}} \phi_{pq}(\text{vec}(\mathcal{Z}))\exp\left(\frac{Y^2}{2(K+1)}\right)\Phi\left(\frac{Y}{\sqrt{K+1}}\right)\text{d}(\text{vec}(\mathcal{Z})).$$

Due to the nature of the integrand, it sounds convenient to decompose $\text{vec}(\boldsymbol{\mathcal{Z}})$ onto directions parallel and orthogonal to $\text{vec}(\mathbf{B})$ so that we can split the MVN PDF into two components. With the purpose of doing so, let us reinforce the definition of Y and define the unit vector $\mathbf{u}$ as follows:

1. $\mathbf{u} := \operatorname{vec}(\mathbf{B})/\|\operatorname{vec}(\mathbf{B})\|$.

2. $Y := \operatorname{vec}(\boldsymbol{\mathcal{Z}})^\top \operatorname{vec}(\mathbf{B})$.

To proceed, let us consider the matrix $\mathbf{Q} \in \mathbb{R}^{pq\times(pq-1)}$ whose columns constitute an orthonormal basis for the vector space $\mathbf{u}^\perp$ satisfying the relation $\mathbf{Q}^\top\mathbf{Q} = \mathbf{I}_{pq-1}$ as well as $\mathbf{u}^\top\mathbf{Q} = \mathbf{0}_{pq-1}^\top$. Then the change of variables $\operatorname{vec}(\boldsymbol{\mathcal{Z}}) \mapsto (\widetilde{Y}, \operatorname{vec}(\boldsymbol{\mathcal{Z}})^\perp)^\top := (\operatorname{vec}(\boldsymbol{\mathcal{Z}})^\top\mathbf{u}, \operatorname{vec}(\boldsymbol{\mathcal{Z}})^\top\mathbf{Q})^\top$, which is a linear transformation whose matrix representation is called $\mathbf{J}$ from now on, is an orthogonal linear mapping (since $\mathbf{J}\mathbf{J}^\top = \mathbf{I}_{pq}$) with Jacobian determinant equal to one (in modulus). Indeed, one has:

$$\mathbf{J}\mathbf{J}^\top = \begin{bmatrix}\mathbf{u}^\top \\ \mathbf{Q}^\top\end{bmatrix}\begin{bmatrix}\mathbf{u} & \mathbf{Q}\end{bmatrix} = \begin{bmatrix}\mathbf{u}^\top\mathbf{u} & \mathbf{u}^\top\mathbf{Q} \\ \mathbf{Q}^\top\mathbf{u} & \mathbf{Q}^\top\mathbf{Q}\end{bmatrix} = \begin{bmatrix}1 & \mathbf{0}^\top \\ \mathbf{0} & \mathbf{I}_{pq-1}\end{bmatrix} = \mathbf{I}_{pq}.$$

Here, we introduce the notation $\operatorname{vec}(\boldsymbol{\mathcal{Z}})^\perp \in \mathbb{R}^{pq-1}$ to denote the component of $\operatorname{vec}(\boldsymbol{\mathcal{Z}})$ that lies in the orthogonal complement of the vector $\operatorname{vec}(\mathbf{B})$, that is, in the linear subspace $\mathbf{u}^\perp \subset \mathbb{R}^{pq}$. By construction, this orthogonal component captures all directions of variation of $\operatorname{vec}(\boldsymbol{\mathcal{Z}})$ that are perpendicular to $\operatorname{vec}(\mathbf{B})$. Notice that we can write

$$Y = \operatorname{vec}(\mathbf{B})^\top \operatorname{vec}(\boldsymbol{\mathcal{Z}}) = \|\operatorname{vec}(\mathbf{B})\|\,\mathbf{u}^\top \operatorname{vec}(\boldsymbol{\mathcal{Z}}) = \sqrt{K}\widetilde{Y},$$

Due to the orthogonality between $\widetilde{Y}$ and $\operatorname{vec}(\boldsymbol{\mathcal{Z}})^\perp$, we can assert that these two components are independent in the sense that they describe completely separate directions of variation within $\operatorname{vec}(\boldsymbol{\mathcal{Z}})$, which allows us to decompose certain functions of $\operatorname{vec}(\boldsymbol{\mathcal{Z}})$ into functions of $\widetilde{Y}$ and $\operatorname{vec}(\boldsymbol{\mathcal{Z}})^\perp$ separately. More precisely, it follows that:

$$\|\operatorname{vec}(\boldsymbol{\mathcal{Z}})\|^2 = \widetilde{Y}^2 + \|\operatorname{vec}(\boldsymbol{\mathcal{Z}})^\perp\|^2$$

from whence it results that:

$$\phi_{pq}(\operatorname{vec}(\boldsymbol{\mathcal{Z}})) = \frac{1}{\sqrt{2\pi}}\phi_{pq-1}(\operatorname{vec}(\boldsymbol{\mathcal{Z}})^\perp)\exp\left\{-\frac{1}{2}\widetilde{Y}^2\right\}.$$

Consequently, due to the Fubini's theorem, the obtained integral simplifies to

$$\begin{aligned}
I &= \frac{2}{\sqrt{K+1}}\int_{\mathbb{R}^{pq}} \phi_{pq}(\operatorname{vec}(\mathcal{Z}))\exp\left(\frac{Y^2}{2(K+1)}\right)\Phi\left(\frac{Y}{\sqrt{K+1}}\right)\mathrm{d}(\operatorname{vec}(\mathcal{Z})) \\
&\propto \frac{2}{\sqrt{K+1}}\int_{\mathbb{R}^{pq}} \phi_{pq-1}(\operatorname{vec}(\mathcal{Z})^\perp)\exp\left\{-\frac{\widetilde{Y}^2}{2}\right\}\exp\left(\frac{K\widetilde{Y}^2}{2(K+1)}\right)\Phi\left(\frac{\sqrt{K}\widetilde{Y}}{\sqrt{K+1}}\right)\mathrm{d}\widetilde{Y}\mathrm{d}(\operatorname{vec}(\mathcal{Z})^\perp) \\
&= \frac{2}{\sqrt{K+1}}\int_{\mathbb{R}}\exp\left(\frac{\widetilde{Y}^2}{2(K+1)}\right)\Phi\left(\frac{\sqrt{K}\widetilde{Y}}{\sqrt{K+1}}\right)\mathrm{d}\widetilde{Y}\int_{\mathbb{R}^{pq-1}}\phi_{pq-1}(\operatorname{vec}(\mathcal{Z})^\perp)\mathrm{d}(\operatorname{vec}(\mathcal{Z})^\perp) \\
&= \frac{2}{\sqrt{K+1}}\int_{\mathbb{R}}\exp\left(\frac{\widetilde{Y}^2}{2(K+1)}\right)\Phi\left(\frac{\sqrt{K}\widetilde{Y}}{\sqrt{K+1}}\right)\mathrm{d}\widetilde{Y},
\end{aligned}$$

Now, if we apply the change of variable $\widetilde{Y} = z\sqrt{K+1}$, so that $\mathrm{d}\widetilde{Y} = \sqrt{K+1}\mathrm{d}z$, we get that:

$$\begin{aligned} I &= \frac{2}{\sqrt{K+1}} \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) \Phi(\sqrt{K}z)\sqrt{K+1}\mathrm{d}z \\ &= 2\int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) \Phi(\sqrt{K}z)\mathrm{d}z. \end{aligned}$$

Finally, the desired result follows from the application of Lemma 7 as previously proved:

$$\int_{\mathbb{R}} \phi(z)\Phi(az)\mathrm{d}z = \frac{1}{2}$$

□

Remark. Although this verification is not strictly necessary — since normalization follows directly from the stochastic representation — we include it here for completeness and to enrich the presentation.

Proposition 11. Based on the assumption that $\boldsymbol{\mathcal{X}} \sim \mathrm{MVSN}_{p\times q}(\mathbf{M}, \mathbf{A}, \Sigma, \Psi)$, one may obtain the following formula for the PDF of the latent variable W conditioned on the random matrix $\boldsymbol{\mathcal{X}}$:

$$f_{\mathrm{TN}}(w \mid \boldsymbol{\mathcal{X}}) = \frac{1}{\sqrt{2\pi\sigma^2}} \left[\Phi\left(\frac{\mu}{\sigma}\right)\right]^{-1} \exp\left(-\frac{(w-\mu)^2}{2\sigma^2}\right),$$

where $w \geq 0$ and the parameters $\mu \in \mathbb{R}$ and $\sigma \in \mathbb{R}_{>0}$ are given by:

$$\mu = \frac{\mathrm{tr}(\Sigma^{-1}\boldsymbol{\mathcal{Z}}\Psi^{-1}\mathbf{A}^\top)}{\mathrm{tr}(\Sigma^{-1}\mathbf{A}\Psi^{-1}\mathbf{A}^\top)+1} = \frac{d_{\mathbf{A}}(\mathcal{X})}{\tau^2} \quad \text{and} \quad \sigma^2 = \frac{1}{\mathrm{tr}(\Sigma^{-1}\mathbf{A}\Psi^{-1}\mathbf{A}^\top)+1} = \frac{1}{\tau^2}.$$

as previously conventioned. The resulting probability density function $f(w \mid \boldsymbol{\mathcal{X}})$ corresponds to the distribution $W \mid \boldsymbol{\mathcal{X}} \sim \mathrm{TN}(\mu, \sigma^2; [0, +\infty))$.

Proof. To preserve notational clarity, we shall adopt the (redundant) conventions $b := d_{\mathbf{A}}(\mathcal{X})$ and $c := \tau^2$ throughout the context of this proposition. Consequently, we obtain that:

$$\begin{aligned} f(w \mid \boldsymbol{\mathcal{X}}) = \frac{f(\boldsymbol{\mathcal{X}}, w)}{f(\boldsymbol{\mathcal{X}})} &= \sqrt{\frac{c}{2\pi}} \left[\Phi\left(\frac{b}{\sqrt{c}}\right)\right]^{-1} \exp\left[-\frac{1}{2}\left(\frac{b^2}{c} - 2bw + cw^2\right)\right] \\ &= \sqrt{\frac{c}{2\pi}} \left[\Phi\left(\frac{b}{\sqrt{c}}\right)\right]^{-1} \exp\left[-\frac{c}{2}\left(\frac{b^2}{c^2} - \frac{2b}{c}w + w^2\right)\right] \\ &= \sqrt{\frac{c}{2\pi}} \left[\Phi\left(\frac{b}{\sqrt{c}}\right)\right]^{-1} \exp\left[-\frac{c}{2}\left(w - \frac{b}{c}\right)^2\right]. \end{aligned}$$

Since the truncated normal distribution defined over the interval $[0, +\infty)$ (with mean parameter $\mu \in \mathbb{R}$ and standard deviation parameter $\sigma \in \mathbb{R}_{>0}$) has PDF given by the following expression

$$f_{\mathrm{TN}}(w \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \left[\Phi\left(\frac{\mu}{\sigma}\right)\right]^{-1} \exp\left[-\frac{1}{2}\frac{(w-\mu)^2}{\sigma^2}\right],$$

it can be concluded (by comparison) that the conditional PDF $f(w \mid \boldsymbol{\mathcal{X}})$ belongs to the truncated normal family where:

$$\mu = \frac{b}{c} = \frac{d_{\mathbf{A}}(\mathcal{X})}{\tau^2} = \frac{\operatorname{tr}(\Sigma^{-1}\boldsymbol{\mathcal{Z}}\Psi^{-1}\mathbf{A}^\top)}{\operatorname{tr}(\Sigma^{-1}\mathbf{A}\Psi^{-1}\mathbf{A}^\top)+1} \quad \text{and} \quad \sigma^2 = \frac{1}{c} = \frac{1}{\tau^2} = \frac{1}{\operatorname{tr}(\Sigma^{-1}\mathbf{A}\Psi^{-1}\mathbf{A}^\top)+1}$$

and the proposed result holds, just as we wanted to demonstrate. □

We now establish key properties of the MVSN distribution. Proposition 12 gives explicit expressions for its first-order and second-order moments, while Proposition 13 confirms coherence between the matrix-variate and vectorized formulations. Proposition 14 shows that the class is closed under affine transformations, and Proposition 15 characterizes the distribution of block submatrices. Although these results already appear in the literature (including Proposition 11), we independently provide their proofs here for completeness.

Proposition 12. The mean and covariance matrices related to the the distribution of the random matrix $\boldsymbol{\mathcal{X}} \sim \text{MVSN}_{p\times q}(\mathbf{M},\mathbf{A},\Sigma,\Psi)$ as previously defined, where $W \perp \boldsymbol{\mathcal{V}}$, are given by:

$$\mathbb{E}(\boldsymbol{\mathcal{X}}) = \mathbf{M} + \sqrt{\frac{2}{\pi}}\mathbf{A} \quad \text{and} \quad \operatorname{Cov}(\operatorname{vec}(\boldsymbol{\mathcal{X}})) = \left(1-\frac{2}{\pi}\right)\operatorname{vec}(\mathbf{A})\operatorname{vec}(\mathbf{A})^\top + \Psi\otimes\Sigma.$$

Proof. The expression of $\mathbb{E}(\boldsymbol{\mathcal{X}})$ follows immediately from the stochastic representation of the random matrix $\boldsymbol{\mathcal{X}}$. In order to calculate the covariance matrix of $\boldsymbol{\mathcal{X}}$, we shall first vectorize it for the sake of convenience. Based on the independence of W and $\boldsymbol{\mathcal{V}}$, as well as on the fact that $\boldsymbol{\mathcal{V}} \sim \mathcal{N}_{p\times q}(0,\Sigma,\Psi)$, it can be deduced that:

$$\begin{aligned}
\operatorname{Cov}(\operatorname{vec}(\boldsymbol{\mathcal{X}})) &= \operatorname{Cov}(\operatorname{vec}(W\mathbf{A})) + \operatorname{Cov}(\operatorname{vec}(\mathcal{V})) \\
&= \operatorname{Cov}(W\operatorname{vec}(\mathbf{A})) + \Psi\otimes\Sigma \\
&= \operatorname{Var}(W)\operatorname{vec}(\mathbf{A})\operatorname{vec}(\mathbf{A})^\top + \Psi\otimes\Sigma \\
&= \left(1-\frac{2}{\pi}\right)\operatorname{vec}(\mathbf{A})\operatorname{vec}(\mathbf{A})^\top + \Psi\otimes\Sigma.
\end{aligned}$$

just as we wanted to demonstrate. □

Lemma 8. Let $\mathbf{P}\in\mathbb{R}^{m\times n}$, $\mathbf{Q}\in\mathbb{R}^{n\times p}$, $\mathbf{R}\in\mathbb{R}^{p\times q}$ and $\mathbf{S}\in\mathbb{R}^{q\times m}$. Then the identity holds:

$$\operatorname{tr}(\mathbf{PQRS}) = \operatorname{vec}(\mathbf{S}^\top)^\top(\mathbf{R}^\top\otimes\mathbf{P})\operatorname{vec}(\mathbf{Q}).$$

Proof. Here is the line by line proof of the proposed identity:

$$\begin{aligned}
\operatorname{tr}(\mathbf{PQRS}) &= \operatorname{tr}(\mathbf{SPQR}) && \text{(ciclicity of the trace function)} \\
&= \operatorname{vec}(\mathbf{S}^\top)^\top \operatorname{vec}(\mathbf{PQR}) && \text{(trace-vec identity)} \\
&= \operatorname{vec}(\mathbf{S}^\top)^\top(\mathbf{R}^\top\otimes\mathbf{P})\operatorname{vec}(\mathbf{Q}) && \text{(vec of triple product)}
\end{aligned}$$

and the desired relation holds, just as we wanted to demonstrate. □

Proposition 13. Let $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{p\times q}$ be a random matrix. Therefore the following relation holds:

$$\boldsymbol{\mathcal{X}} \sim \text{MVSN}_{p\times q}(\mathbf{M},\mathbf{A},\Sigma,\Psi) \iff \text{vec}(\boldsymbol{\mathcal{X}}) \sim \text{MSN}_{pq}(\text{vec}(\mathbf{M}),\text{vec}(\mathbf{A}),\Psi\otimes\Sigma),$$

where MSN_{pq} corresponds to the multivariate skew normal distribution as in Definition 10.

Proof. In probability theory, the distribution of a random element (variable, vector, or matrix) is fully characterized by its law, CDF, or PDF. Since both distributions are absolutely continuous, showing that their probability density functions coincide (after vectorization) is sufficient to establish equivalence in distribution. This is because the vec operator is a bijective linear mapping with determinant of modulus one, so it preserves probabilities under the change of variables. Having said that, we may now proceed.

By vectorizing the matrix $\boldsymbol{\mathcal{X}}$, we obtain $\mathbf{X} = \text{vec}(\boldsymbol{\mathcal{X}}) \in \mathbb{R}^{pq}$, which by its turn follows the MSN distribution $\mathbf{X} = \mu + W\mathbf{b} + \mathbf{V}$ so that the relations $\mu = \text{vec}(\mathbf{M})$, $\mathbf{b} = \text{vec}(\mathbf{A})$, $\mathbf{V} \sim \mathcal{N}_{pq}(\mathbf{0},\Delta)$ hold, where $\Delta = \Psi\otimes\Sigma$. The marginal PDF of $\mathbf{X} \in \mathbb{R}^{pq}$ is given by the formula:

$$f_{\text{MSN}}(\mathbf{x}) = \frac{2\phi_{pq}(\mathbf{x}\mid\mu,\Delta)}{\sqrt{1+\mathbf{b}^\top\Delta^{-1}\mathbf{b}}}\exp\left\{\frac{1}{2}\frac{(\mathbf{b}^\top\Delta^{-1}(\mathbf{x}-\mu))^2}{1+\mathbf{b}^\top\Delta^{-1}\mathbf{b}}\right\}\Phi\left(\frac{\mathbf{b}^\top\Delta^{-1}(\mathbf{x}-\mu)}{\sqrt{1+\mathbf{b}^\top\Delta^{-1}\mathbf{b}}}\right).$$

Based on Lemma 8, the kronecker product identity $\mathbf{P}^{-1}\otimes\mathbf{Q}^{-1} = (\mathbf{P}\otimes\mathbf{Q})^{-1}$ for invertible matrices $\mathbf{P}$ and $\mathbf{Q}$, one obtains the following result:

$$\begin{aligned}\text{tr}(\Sigma^{-1}(\boldsymbol{\mathcal{X}}-\mathbf{M})\Psi^{-1}\mathbf{A}^\top) &= \text{vec}(\mathbf{A})^\top(\Psi^{-1}\otimes\Sigma^{-1})\,\text{vec}(\boldsymbol{\mathcal{X}}-\mathbf{M})\\ &= \text{vec}(\mathbf{A})^\top(\Psi\otimes\Sigma)^{-1}\,\text{vec}(\boldsymbol{\mathcal{X}}-\mathbf{M})\\ &= \mathbf{b}^\top\Delta^{-1}(\mathbf{x}-\mu).\end{aligned}$$

Similarly, we do also have the next identity:

$$\begin{aligned}\text{tr}(\Sigma^{-1}\mathbf{A}\Psi^{-1}\mathbf{A}^\top) &= \text{vec}(\mathbf{A})^\top(\Psi^{-1}\otimes\Sigma^{-1})\,\text{vec}(\mathbf{A})\\ &= \text{vec}(\mathbf{A})^\top(\Psi\otimes\Sigma)^{-1}\,\text{vec}(\mathbf{A})\\ &= \mathbf{b}^\top\Delta^{-1}\mathbf{b}.\end{aligned}$$

At last, but not least, it results that:

$$\begin{aligned}\text{tr}(\Sigma^{-1}(\boldsymbol{\mathcal{X}}-\mathbf{M})\Psi^{-1}(\boldsymbol{\mathcal{X}}-\mathbf{M})^\top) &= \text{vec}(\boldsymbol{\mathcal{X}}-\mathbf{M})^\top(\Psi^{-1}\otimes\Sigma^{-1})\,\text{vec}(\boldsymbol{\mathcal{X}}-\mathbf{M})\\ &= \text{vec}(\boldsymbol{\mathcal{X}}-\mathbf{M})^\top(\Psi\otimes\Sigma)^{-1}\,\text{vec}(\boldsymbol{\mathcal{X}}-\mathbf{M})\\ &= (\mathbf{x}-\mu)^\top\Delta^{-1}(\mathbf{x}-\mu).\end{aligned}$$

These identities show that the matrix-variate formulation and the multivariate MSN model are fully compatible under vectorization, with the densities coinciding exactly when expressed in their respective parameterizations. □

Proposition 14. Let $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{p\times q}$ follow an stochastic matrix variate skew-normal distribution with parameters $\mathbf{M}$, $\mathbf{A}$, Σ, and Ψ, denoted by

$$\boldsymbol{\mathcal{X}} \sim \text{MVSN}_{p\times q}(\mathbf{M}, \mathbf{A}, \Sigma, \Psi).$$

Consider the affine transformation $\mathcal{U} = \mathbf{C} + \mathbf{D}\boldsymbol{\mathcal{X}}$, where $\mathbf{C} \in \mathbb{R}^{p\times q}$ and $\mathbf{D} \in \mathbb{R}^{p\times p}$ is a nonsingular matrix (that is to say, $|\mathbf{D}| > 0$). Then it can be claimed that:

$$\mathcal{U} \sim \text{MVSN}_{p\times q}\left(\mathbf{C} + \mathbf{DM},\ \mathbf{DA}, \mathbf{D}\Sigma\mathbf{D}^\top,\ \Psi\right),$$

which shows that the family of MVSN distribution is closed under affine row transformations.

Proof. Substituting the relation $\mathcal{U} = \mathbf{C} + \mathbf{D}\boldsymbol{\mathcal{X}}$ into the stochastic representation of $\boldsymbol{\mathcal{X}}$ leads to:

$$\mathcal{U} = (\mathbf{C} + \mathbf{DM}) + W(\mathbf{DA}) + \mathbf{D}\boldsymbol{\mathcal{X}}.$$

Since linear transformations preserve normality, we have $\mathbf{D}\boldsymbol{\mathcal{X}} \sim \mathcal{N}_{p\times q}(\mathbf{0}, \mathbf{D}\Sigma\mathbf{D}^\top, \Psi)$. Thus, $\mathcal{U}$ admits the same stochastic form as a MVSN distribution with transformed parameters. □

Proposition 15. Let $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{p\times q}$ follow a matrix variate skew-normal distribution, that is to say, $\boldsymbol{\mathcal{X}} \sim \text{MVSN}_{p\times q}(\mathbf{M}, \mathbf{A}, \Sigma, \Psi)$. Then let us partition it as $\boldsymbol{\mathcal{X}} = [\boldsymbol{\mathcal{X}}_1^\top, \boldsymbol{\mathcal{X}}_2^\top]^\top$. Likewise, let us split the location matrix $\mathbf{M}$, the skewness matrix $\mathbf{A}$ and the covariance matrix Σ. Assuming that $\mathbf{T}_1 = [\mathbf{I}_{p_1} \quad \mathbf{0}] \in \mathbb{R}^{p_1\times p}$ represents the selection matrix extracting the first p_1 rows of $\boldsymbol{\mathcal{X}}$. Thus:

$$\mathcal{U} = \mathbf{T}_1\boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{X}}_1 \sim \text{MVSN}_{p_1\times q}(\mathbf{M}_1, \mathbf{A}_1, \Sigma_{11}, \Psi).$$

Proof. The result is a direct consequence of the closure property of the MVSN distribution under affine row transformations. Specifically, let us define the linear transformation $\boldsymbol{\mathcal{U}}$ as $\mathcal{U} = \mathbf{T}_1\boldsymbol{\mathcal{X}}$, where $\mathbf{T}_1 = [\mathbf{I}_{p_1} \quad \mathbf{0}]$ is a selection matrix that extracts the first p_1 rows of $\boldsymbol{\mathcal{X}}$. This corresponds precisely to taking $\mathbf{C} = 0$ and $\mathbf{D} = \mathbf{T}_1$ in the general affine form $\mathcal{U} = \mathbf{C} + \mathbf{D}\boldsymbol{\mathcal{X}}$. Applying Proposition 14, we obtain

$$\boldsymbol{\mathcal{U}} \sim \text{MVSN}_{p_1\times q}\left(\mathbf{T}_1\mathbf{M},\ \mathbf{T}_1\mathbf{A},\ \mathbf{T}_1\Sigma\mathbf{T}_1^\top,\ \Psi\right).$$

Therefore the partitions of $\mathbf{M}$, $\mathbf{A}$ and the row covariance matrix Σ are defined as:

$$\mathbf{M} = \begin{bmatrix}\mathbf{M}_1\\ \mathbf{M}_2\end{bmatrix},\ \mathbf{A} = \begin{bmatrix}\mathbf{A}_1\\ \mathbf{A}_2\end{bmatrix},\ \Sigma = \begin{bmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{bmatrix},$$

From the partitioned structure, we immediately identify

$$\mathbf{T}_1\mathbf{M} = \mathbf{M}_1, \quad \mathbf{T}_1\mathbf{A} = \mathbf{A}_1, \quad \mathbf{T}_1\Sigma\mathbf{T}_1^\top = \Sigma_{11}.$$

Substituting these results into the general expression provided by Proposition 14, it results that:

$$\mathcal{U} = \boldsymbol{\mathcal{X}}_1 \sim \text{MVSN}_{p_1\times q}(\mathbf{M}_1, \mathbf{A}_1, \Sigma_{11}, \Psi).$$

This imples that the marginal distribution of any set of rows of $\boldsymbol{\mathcal{X}}$ remains a MVSN model. □

With the purpose of obtaining the moment generating function (MGF) of the MVSN distribution, we shall use its stochastic representation, the MGF of the matrix variate normal distribuiton and the law of iterated expectations (also known as the tower property). The first of them has already been proposed in the beginning of this chapter. The second one can be checked at the reference Mathai, Provost and Haubold (2022) and the third one is available at the reference Rolla and Lima (2025), for example. We present the MGF of the MVN distribution as a Lemma.

Lemma 9. Let $\boldsymbol{\mathcal{X}} \sim \mathcal{N}_{p\times q}(\mathbf{M}, \Sigma, \Psi)$ be a random matrix which follows a Gaussian distribution. If $\mathbf{T} \in \mathbb{R}^{p\times q}$, then its moment generating function has the expression:

$$M_{\boldsymbol{\mathcal{X}}}(\mathbf{T}) = \exp\left[\operatorname{tr}(\mathbf{T}'\mathbf{M}) + \frac{1}{2}\operatorname{tr}(\Psi\mathbf{T}'\Sigma\mathbf{T})\right].$$

Proof. The proof of such result can be checked at Mathai, Provost and Haubold (2022). □

Proposition 16. Suppose that $\boldsymbol{\mathcal{X}} \sim \text{MVSN}(\mathbf{M}, \mathbf{A}, \Sigma, \Psi)$, where the latent variable has distribution given by $W \sim \text{HalfNormal}(0,1)$ and is independent of $\boldsymbol{\mathcal{V}} \sim \mathcal{N}_{p\times q}(\mathbf{0}, \Sigma, \Psi)$, which comes from the stochastic representation of $\boldsymbol{\mathcal{X}}$. If we set up that $\mathbf{T} \in \mathbb{R}^{p\times q}$, it may be inferred that

$$M_{\boldsymbol{\mathcal{X}}}(\mathbf{T}) = 2\exp\left[\operatorname{tr}(\mathbf{T}'\mathbf{M}) + \frac{1}{2}\operatorname{tr}(\Psi\mathbf{T}'\Sigma\mathbf{T}) + \frac{1}{2}\operatorname{tr}^2(\mathbf{T}'\mathbf{A})\right]\Phi(\operatorname{tr}(\mathbf{T}'\mathbf{A})).$$

Proof. Since we know that $\mathbb{E}[\mathbb{E}[f(\boldsymbol{\mathcal{X}}) \mid W]] = \mathbb{E}[f(\boldsymbol{\mathcal{X}})]$ as previously mentioned (known as the tower property) as well as the distribution of $\boldsymbol{\mathcal{X}} \mid W$ (which is normally distributed), we claim:

$$M_{\boldsymbol{\mathcal{X}}|W}(\mathbf{T}) = \mathbb{E}\left[\exp\left(\operatorname{tr}\left(\mathbf{T}'\boldsymbol{\mathcal{X}}\right)\right) \mid W\right] = \exp\left[\operatorname{tr}(\mathbf{T}'(\mathbf{M} + W\mathbf{A})) + \frac{1}{2}\operatorname{tr}(\Psi\mathbf{T}'\Sigma\mathbf{T})\right].$$

Consequently, if we establish that $\kappa = \operatorname{tr}(\mathbf{T}'\mathbf{A})$, the conditional MGF is given by the formula:

$$M_{\boldsymbol{\mathcal{X}}|W}(\mathbf{T}) = \exp\left[\operatorname{tr}(\mathbf{T}'\mathbf{M}) + \frac{1}{2}\operatorname{tr}(\Psi\mathbf{T}'\Sigma\mathbf{T}) + \kappa W\right].$$

Therefore it remains to apply the tower property to the last expression, which by its turn consists in calculating the expected value $\mathbb{E}_W[\exp(\kappa W)]$. Based on such considerations, one results that:

$$\begin{aligned}
\mathbb{E}[\exp(\kappa W)] &= \sqrt{\frac{2}{\pi}}\int_0^\infty \exp\left(\kappa w - \frac{w^2}{2}\right) \mathrm{d}w \\
&= \sqrt{\frac{2}{\pi}}\int_0^\infty \exp\left(\frac{\kappa^2}{2} - \frac{1}{2}(w-\kappa)^2\right) \mathrm{d}w \\
&= \sqrt{\frac{2}{\pi}}\exp\left(\frac{\kappa^2}{2}\right)\int_{-\kappa}^\infty \exp\left(-\frac{z^2}{2}\right) \mathrm{d}z \\
&= 2\exp\left(\frac{\kappa^2}{2}\right)\Phi(\kappa).
\end{aligned}$$

Gathering all these steps together, one gets the desired result, just as we wanted to demonstrate. □

3.3 The MVESN Distribution

The multivariate extended skew-normal (MESN) distribution is a flexible family of models that builds on the normal distribution by adding a parameter vector that allows it to handle asymmetry in the data. This extra feature gives the MESN more adaptability, making it especially useful when dealing with datasets that do not follow the symmetric shape of the normal distribution. Because of this, the MESN has become more and more popular in statistical modeling, since it can capture patterns and structures that the standard normal model often overlooks. Although its matrix variate version has been introduced in Naderi *et al.* (2024), we provide our own discussion here (which was developed independently) and present additional results. For completeness sake, we now present the formal definition of the MESN model.

Definition 12. As it can be deduced from the reference Galarza, Matos and Lachos (2022), we say that $\mathbf{Z} \in \mathbb{R}^p$ is a multivariate extended skew normal random vector if and only if it satisfies the stochastic representation:

$$\mathbf{Z} = \mu + \Omega^{1/2}(\delta T_0 + (\mathbf{I}_p - \delta\delta^\top)^{1/2}\mathbf{T}_1),$$

where $\mu \in \mathbb{R}^p$ is the location vector, $\Omega \in \mathbb{R}^{p\times p}$ is a positive definite dispersion matrix, $\lambda \in \mathbb{R}^p$ is the skewness parameter vector, the latent variable T_0 satisfies $T_0 \sim \mathrm{TN}(0,1;[-\widetilde{\kappa},\infty))$, κ is the shift paramter so that $\widetilde{\kappa} := \kappa/(1+\lambda^\top\lambda)^{1/2}$, $\mathbf{T}_1 \sim \mathcal{N}_p(\mathbf{0},\mathbf{I}_p)$ and $\delta := \lambda/(1+\lambda^\top\lambda)^{1/2}$. We denote such a relation among $\mathbf{Z}$ and its parameters by $\mathbf{Z} \sim \mathrm{MESN}_p^\star(\mu,\Omega,\lambda,\kappa)$.

Definition 13. We claim that $\mathbf{Z} \in \mathbb{R}^p$ is a multivariate extended skew normal random vector if and only if it satisfies the stochastic representation:

$$\mathbf{Z} = \mu + W\mathbf{b} + \mathbf{V},$$

where $\mu \in \mathbb{R}^p$ is the location vector, $W \sim \mathrm{TN}(0,1,[-\widetilde{\kappa},+\infty))$ where κ is the shift parameter and $\widetilde{\kappa} := \kappa/(1+\mathbf{b}^\top\Delta^{-1}\mathbf{b})^{1/2}$, $\mathbf{b} \in \mathbb{R}^p$ is the skewness vector, the distribution of $\mathbf{V}$ is given by $\mathbf{V} \sim \mathcal{N}_p(\mathbf{0},\Delta)$ so that Δ is a symmetric and positive definite dispersion matrix and $|X_0| \perp \mathbf{V}$. We denote such relation among $\mathbf{Z}$ and its parameters by $\mathbf{Z} \sim \mathrm{MESN}_p(\mu,\Delta,\mathbf{b},\kappa)$.

Proposition 17. Suppose $\mathbf{Z}_1 \sim \mathrm{MESN}_p^\star(\mu,\Omega,\lambda,\kappa)$ and $\mathbf{Z}_2 \sim \mathrm{MESN}_p(\mu,\Delta,\mathbf{b},\kappa)$. Then it can be concluded that $\mathbf{Z}_1 \overset{d}{=} \mathbf{Z}_2$.

Proof. The proof of such result is exactly the same as the one provided to Proposition 6. □

Proposition 18. The probability density function corresponding to the vector-valued random variable $\mathbf{X} \sim \mathrm{MESN}_p^\star(\mu,\Omega,\lambda,\kappa)$ is given by the following formula:

$$f_{\mathrm{MESN}}(\mathbf{x}) = \frac{1}{\Phi\left(\kappa/\sqrt{1+\lambda^\top\lambda}\right)}\phi_p(\mathbf{x}\mid\mu,\Omega)\Phi\left(\kappa+\lambda^\top\Omega^{-1/2}(\mathbf{x}-\mu)\right),$$

where the parameters involved are considered as in the Definition 13.

Proof. The proof of such result is mentioned in Galarza, Matos and Lachos (2022). □

Lemma 10. The eigenvalues of the linear operator $\mathbf{I}_p - \mathbf{u}\mathbf{u}^\top$, where $\|\mathbf{u}\| < 1$ are $1 - \|\mathbf{u}\|^2$ and 1.

Proof. To start with, observe that the linear operator $\mathbf{I}_p - \mathbf{u}\mathbf{u}^\top$ is symmetric. Consequently, the spectral theorem applies, that is to say, its eigenvalues are real and there exists an orthonormal basis of eigenvectors. We claim that $\mathbf{u}$ is an eigenvalue. Indeed, one has that

$$(\mathbf{I}_p - \mathbf{u}\mathbf{u}^\top)\mathbf{u} = \mathbf{u} - \mathbf{u}(\mathbf{u}^\top\mathbf{u}) = (1 - \|\mathbf{u}\|^2)\mathbf{u}.$$

Therefore $\mathbf{u}$ is an eigenvector with eigenvalue $1 - \|\mathbf{u}\|^2$. Now we shall prove that every vector orthogonal to $\mathbf{u}$ is an eigenvector with eigenvalue 1. In fact, let $\mathbf{v} \in \mathbb{R}^p$ so that $\langle \mathbf{v}, \mathbf{u} \rangle = 0$. Then:

$$(\mathbf{I}_p - \mathbf{u}\mathbf{u}^\top)\mathbf{v} = \mathbf{v} - \mathbf{u}(\mathbf{u}^\top\mathbf{v}) = \mathbf{v}.$$

and the desired relation holds. Hence the subspace $\{\mathbf{v} \in \mathbb{R}^p \mid \mathbf{v}^\top\mathbf{u} = \mathbf{0}\} = \mathbf{u}^\perp$ has dimension $p-1$ (if $\mathbf{u} \neq \mathbf{0}$). Choose an orthonormal basis $\{\mathbf{v}_1, \ldots, \mathbf{v}_{p-1}\}$ of $\mathbf{u}^\perp$. Each $\mathbf{v}_i$ satisfies $(\mathbf{I}_p - \mathbf{u}\mathbf{u}^\top)\mathbf{v}_i = \mathbf{v}_i$. Consequently, these $p-1$ vectors are linearly independent eigenvectors whose associated eigenvalues equal one.

Now take it into account the set $\{\mathbf{u}, \mathbf{v}_1, \ldots, \mathbf{v}_{p-1}\}$. These vectors are linearly independent. Hence they form a basis of $\mathbb{R}^p$. Because $\mathbf{I}_p - \mathbf{u}\mathbf{u}^\perp$ is symmetric and we exhibited p linearly independent eigenvectors, $\mathbf{I}_p - \mathbf{u}\mathbf{u}^\top$ is diagonalizable with these eigenvectors as an orthogonal eigenbasis. Thus the spectrum of $\mathbf{I}_p - \mathbf{u}\mathbf{u}^\top$ consists exactly of:

- eigenvalue $1 - \|\mathbf{u}\|^2$ with eigenvector $\mathbf{u}$ (geometric multiciplicity one),
- eigenvalue 1 with eigenvectors spanning $\mathbf{u}^\perp$ (geometric multiplicity $p-1$).

For symmetric matrices, the algebraic multiplicity equals the geometric multiplicity for every eigenvalue. Therefore the algebraic multiplicities are the same: 1 occurs with multiplicity $p-1$, and $1 - \|\mathbf{u}\|^2$ occurs with multiplicity 1. If $\mathbf{u} = \mathbf{0}$, the result is trivial. □

Theorem 4. Suppose $\lambda \in \mathbb{R}^p$ and $\Omega \in \mathbb{R}^{p\times p}$, where the latter is symmetric and positive definite. If the relations $\mathbf{b} := \Omega^{1/2}\lambda/(1+\lambda^\top\lambda)^{1/2}$ and $\Delta := \Omega - \mathbf{b}\mathbf{b}^\top$ are established, then we may express that par (λ, Ω) in terms of the pair $(\mathbf{b}, \Delta)$ precisely as

$$\Omega = \Delta + \mathbf{b}\mathbf{b}^\top, \quad \lambda = (1+\mathbf{b}^\top\Delta^{-1}\mathbf{b})^{1/2}(\Delta+\mathbf{b}\mathbf{b}^\top)^{-1/2}\mathbf{b} \quad \text{and} \quad \lambda^\top\Omega^{-1/2} = \frac{\mathbf{b}^\top\Delta^{-1}}{(1+\mathbf{b}^\top\Delta^{-1}\mathbf{b})^{1/2}}$$

This shall be particularly useful later on in Chapter 5 when we deal with reparametrizations.

Proof. The first identity is clearly true. To prove that Δ is invertible, we demonstrate that it is positive definite. Indeed, let $\mathbf{x} \in \mathbb{R}^p \backslash \{0\}$. Due to the definitions involved, we may conclude that:

$$\mathbf{x}^\top\Delta\mathbf{x} = \mathbf{x}^\top\Omega\mathbf{x} - \mathbf{x}^\top\mathbf{b}\mathbf{b}^\top\mathbf{x} = (\mathbf{x}^\top\Omega^{1/2})(\Omega^{1/2}\mathbf{x}) - (\mathbf{x}^\top\mathbf{b})(\mathbf{x}^\top\mathbf{b})^\top = (\Omega^{1/2}\mathbf{x})^\top(\Omega^{1/2}\mathbf{x}) - (\mathbf{x}^\top\mathbf{b})^2.$$

From then on, we apply the definition $\mathbf{b}$ to arrive at the desired conclusion. As a matter of fact, if we conveniently define $\mathbf{y} := \Omega^{1/2}\mathbf{x}$, then $\mathbf{x} = \Omega^{-1/2}\mathbf{y}$ and it is possible to conclude that:

$$\mathbf{x}^\top \mathbf{b} = \frac{\mathbf{x}^\top \Omega^{1/2}\lambda}{(1+\lambda^\top\lambda)^{1/2}} \Rightarrow (\mathbf{x}^\top \mathbf{b})^2 = \frac{(\mathbf{x}^\top \Omega^{1/2}\lambda)^2}{1+\lambda^\top\lambda} = \frac{(\mathbf{y}^\top\lambda)^2}{1+\lambda^\top\lambda}$$

Therefore, in accordance with the Cauchy-Schwarz inequality, it can be deduced that:

$$\mathbf{x}^\top \Delta \mathbf{x} = \mathbf{y}^\top\mathbf{y} - \frac{(\mathbf{y}^\top\lambda)^2}{1+\|\lambda\|^2} \geq \|\mathbf{y}\|^2 - \frac{\|\mathbf{y}\|^2\|\lambda\|^2}{1+\|\lambda\|^2} = \|\mathbf{y}\|^2\left(1 - \frac{\|\lambda\|^2}{1+\|\lambda\|^2}\right) = \frac{\|\mathbf{y}\|^2}{1+\|\lambda\|^2} > 0.$$

The second point to observe is that we can rearrange the definition of $\mathbf{b}$ to express λ as follows:

$$\mathbf{b} = \frac{\Omega^{1/2}\lambda}{(1+\lambda^\top\lambda)^{1/2}} \Rightarrow \lambda = (1+\lambda^\top\lambda)^{1/2}\Omega^{-1/2}\mathbf{b}.$$

The third point to observe consists in noticing that we can express the matrix Δ as we do next:

$$\Delta = \Omega - \mathbf{b}\mathbf{b}^\top = \Omega - \frac{\Omega^{1/2}\lambda\lambda^\top\Omega^{1/2}}{1+\lambda^\top\lambda} = \Omega^{1/2}\left[\mathbf{I}_p - \frac{\lambda\lambda^\top}{1+\lambda^\top\lambda}\right]\Omega^{1/2},$$

from whence it can be concluded that its inverse corresponds to:

$$\Delta^{-1} = \Omega^{-1/2}\left[\mathbf{I}_p - \frac{\lambda\lambda^\top}{1+\lambda^\top\lambda}\right]^{-1}\Omega^{-1/2}.$$

Let us convention that $\mathbf{u} := \lambda/(1+\lambda^\top\lambda)^{1/2}$. Given that $\|u\|^2 < 1$ due its definition, it results (via Lemma 10) that $|\mathbf{I}_p - \mathbf{u}\mathbf{u}^\top| = 1 - \|\mathbf{u}\|^2 > 0$, thence invertible. Having said that, we may now apply the Sherman-Morrison identity to obtain:

$$(\mathbf{I}_p - \mathbf{u}\mathbf{u}^\top)^{-1} = \mathbf{I}_p + \frac{\mathbf{u}\mathbf{u}^\top}{1-\|\mathbf{u}\|^2} = \mathbf{I}_p + \frac{\lambda\lambda^\top/(1+\lambda^\top\lambda)}{1-\lambda^\top\lambda/(1+\lambda^\top\lambda)} = \mathbf{I}_p + \lambda\lambda^\top.$$

We may now proceed and prove that $\lambda\lambda^\top = \mathbf{b}^\top\Delta^{-1}\mathbf{b}$. Indeed, gathering the previous results,

$$\begin{aligned}
\mathbf{b}^\top\Delta^{-1}\mathbf{b} &= \mathbf{b}^\top(\Omega^{-1/2}(\mathbf{I}_p + \lambda\lambda^\top)\Omega^{-1/2})\mathbf{b} \\
&= (\mathbf{b}^\top\Omega^{-1/2})(\mathbf{I}_p + \lambda\lambda^\top)(\Omega^{-1/2}\mathbf{b}) \\
&= (\Omega^{-1/2}\mathbf{b})^\top(\mathbf{I}_p + \lambda\lambda^\top)(\Omega^{-1/2}\mathbf{b}) \\
&= \left[\frac{\lambda}{(1+\lambda^\top\lambda)^{1/2}}\right]^\top(\mathbf{I}_p + \lambda\lambda^\top)\left[\frac{\lambda}{(1+\lambda^\top\lambda)^{1/2}}\right] \\
&= \frac{\lambda^\top\lambda}{1+\lambda^\top\lambda} + \frac{(\lambda^\top\lambda)(\lambda^\top\lambda)}{1+\lambda^\top\lambda} \\
&= \lambda^\top\lambda\left(\frac{1+\lambda^\top\lambda}{1+\lambda^\top\lambda}\right) = \lambda^\top\lambda.
\end{aligned}$$

Hence the formula relating λ in terms of $\mathbf{b}$ and Δ is valid. To complete the proof, we may apply the Sherman-Morrison identity once again to get (where clearly $\mathbf{b}^\top\Omega^{-1}\mathbf{b} = (\Omega^{-1/2}\mathbf{b})^\top(\Omega^{-1/2}\mathbf{b}) < 1$ due the definitions involved):

$$\Delta^{-1} = (\Omega - \mathbf{b}\mathbf{b}^\top)^{-1} = \Omega^{-1} + \frac{\Omega^{-1}\mathbf{b}\mathbf{b}^\top\Omega^{-1}}{1-\mathbf{b}^\top\Omega^{-1}\mathbf{b}}$$

from where it is possible to deduce that

$$\mathbf{b}^\top\Delta^{-1} = \mathbf{b}^\top\Omega^{-1} - \frac{\mathbf{b}^\top\Omega^{-1}\mathbf{b}\mathbf{b}^\top\Omega^{-1}}{1-\mathbf{b}^\top\Omega^{-1}\mathbf{b}} = \mathbf{b}^\top\Omega^{-1}\left(1 + \frac{\mathbf{b}^\top\Omega^{-1}\mathbf{b}}{1-\mathbf{b}^\top\Omega^{-1}\mathbf{b}}\right) = \frac{\mathbf{b}^\top\Omega^{-1}}{1-\mathbf{b}^\top\Omega^{-1}\mathbf{b}}$$

Consequently, due to the definition of $\mathbf{b}$, it is possible to conclude that:

$$\mathbf{b}^\top\Omega^{-1}\mathbf{b} = (\Omega^{-1/2}\mathbf{b})^\top(\Omega^{-1/2}\mathbf{b}) = \left[\frac{\lambda}{(1+\lambda^\top\lambda)^{1/2}}\right]^\top\left[\frac{\lambda}{(1+\lambda^\top\lambda)^{1/2}}\right] = \frac{\lambda^\top\lambda}{1+\lambda^\top\lambda}.$$

At last, we obtain the sought-after relation:

$$\begin{aligned}\mathbf{b}^\top\Delta^{-1} &= \frac{\mathbf{b}^\top\Omega^{-1}}{1-\mathbf{b}^\top\Omega^{-1}\mathbf{b}} = \frac{(\mathbf{b}^\top\Omega^{-1/2})\Omega^{-1/2}}{1-\mathbf{b}^\top\Omega^{-1}\mathbf{b}} = \frac{(\Omega^{-1/2}\mathbf{b})^\top\Omega^{-1/2}}{1/(1+\lambda^\top\lambda)} \\ &= \frac{\lambda^\top\Omega^{-1/2}}{(1+\lambda^\top\lambda)^{1/2}}\frac{1}{1/(1+\lambda^\top\lambda)} = \lambda^\top\Omega^{-1/2}(1+\lambda^\top\lambda)^{1/2},\end{aligned}$$

as we just wanted to demonstrate. □

Proposition 19. Suppose that $\mathbf{X} \sim \text{MESN}_p^\star(\mu, \Omega, \lambda, \kappa)$. If we define

$$\mathbf{b} := \frac{\Omega^{1/2}\lambda}{\sqrt{1+\lambda^\top\lambda}} \quad \text{and} \quad \Delta := \Omega - \mathbf{b}\mathbf{b}^\top,$$

then $\mathbf{X}$ can equivalently be expressed as $\mathbf{X} \sim \text{MESN}_p(\mu, \Delta, \mathbf{b}, \kappa)$, that is to say, the MESN* distribution admits a reparametrization in terms of $(\mu, \Delta, \mathbf{b}, \kappa)$. In such case, the probability density function of $\mathbf{X}$ can be rewritten as:

$$f_{\text{MESN}}(\mathbf{x}) = \frac{1}{\Phi(\kappa/\sqrt{1+\delta^2})}\frac{\phi_p(\mathbf{x} \mid \mu, \Delta)}{\sqrt{1+\delta^2}}\exp\left\{\frac{1}{2}\frac{(\mathbf{b}^\top\Delta^{-1}(\mathbf{x}-\mu))^2}{1+\delta^2}\right\}\Phi\left(\kappa + \frac{\mathbf{b}^\top\Delta^{-1}(\mathbf{x}-\mu)}{\sqrt{1+\delta^2}}\right),$$

where $\delta^2 = \mathbf{b}^\top\Delta^{-1}\mathbf{b}$.

Proof. This result is a direct application of Proposition 18, the Theorem 4, the Sherman-Morrison identity and the Matrix Determinant Lemma. In fact, by applying the above change of variables to the probability density function corresponding to the parametrization $\mathbf{X} \sim \text{MESN}_p^\star(\mu, \Omega, \lambda, \kappa)$, we obtain

$$f_{\text{MESN}}(\mathbf{x}) = \frac{1}{\Phi(\kappa/\sqrt{1+\mathbf{b}^\top\Delta^{-1}\mathbf{b}})}\phi_p(\mathbf{x} \mid \mu, \Delta + \mathbf{b}\mathbf{b}^\top)\Phi\left(\kappa + \frac{\mathbf{b}^\top\Delta^{-1}(\mathbf{x}-\mu)}{\sqrt{1+\mathbf{b}^\top\Delta^{-1}\mathbf{b}}}\right).$$

Since Δ is positive definite, it may be concluded that $\Delta + \mathbf{b}\mathbf{b}^\top$ is positive definite too (hence invertible). Thus we are in conditions to apply the Sherman-Morrison identity, which states that:

$$(\Delta + \mathbf{b}\mathbf{b}^\top)^{-1} = \Delta^{-1} - \frac{\Delta^{-1}\mathbf{b}\mathbf{b}^\top\Delta^{-1}}{1 + \mathbf{b}^\top\Delta^{-1}\mathbf{b}}.$$

The Matrix Determinant Lemma, by its turn, states that $|\Delta + \mathbf{b}\mathbf{b}^\top| = |\Delta|(1+\delta^2)$. Therefore, in accordance with the definition of the multivariate normal probability density function, it can be deduced that:

$$\phi_p(\mathbf{x} \mid \mu, \Delta + \mathbf{b}\mathbf{b}^\top) = \frac{\phi_p(\mathbf{x} \mid \mu, \Delta)}{\sqrt{1+\delta^2}} \exp\left\{\frac{1}{2}\frac{(\mathbf{b}^\top\Delta^{-1}(\mathbf{x}-\mu))^2}{1+\delta^2}\right\}$$

and the desired result holds, just as we wanted to demonstrate. □

Corollary 1. To prove Proposition 8, it suffices to let $\kappa = 0$ in Proposition 19.

We now extend the multivariate formulation to the matrix-variate case by introducing the MVESN distribution. In this setting, the matrix-valued random variable $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{p\times q}$ is described through a stochastic representation analogous to the one used in the convolution form of the multivariate MESN distribution. The derivation of its marginal density relies on the vectorization of the matrix $\boldsymbol{\mathcal{X}}$, combined with the MESN density presented above. The result below formally defines the MVESN distribution.

Definition 14. Consider the random matrix $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{p\times q}$ defined through the following stochastic representation:

$$\boldsymbol{\mathcal{X}} = \mathbf{M} + W\mathbf{A} + \boldsymbol{\mathcal{V}},$$

where the location matrix is given by $\mathbf{M} \in \mathbb{R}^{p\times q}$, the skewness matrix is given by $\mathbf{A} \in \mathbb{R}^{p\times q}$, the random matrix $\boldsymbol{\mathcal{V}} \sim \mathcal{N}_{p\times q}(\mathbf{0}, \Sigma, \Psi)$ is independent of the random variable $W \sim \mathrm{TN}(0,1;[-\widetilde{\kappa},\infty))$, the shape parameter is denoted by $\widetilde{\kappa} = \kappa/\tau$, with $\tau^2 = 1 + \mathrm{vec}(\mathbf{A})^\top\Delta^{-1}\,\mathrm{vec}(\mathbf{A})$ and $\Delta = \Psi \otimes \Sigma$. We describe it by $\boldsymbol{\mathcal{X}} \sim \mathrm{MVESN}_{p\times q}(\mathbf{M}, \mathbf{A}, \Sigma, \Psi, \kappa)$.

Proposition 20. Given the random matrix $\boldsymbol{\mathcal{X}} \sim \mathrm{MVESN}_{p\times q}(\mathbf{M}, \mathbf{A}, \Sigma, \Psi, \kappa)$ as previously mentioned and defined, its probability density function is given by the formula:

$$f_{\mathrm{MVESN}}(\mathcal{X}) = \frac{1}{\Phi(\kappa/\tau)\,\tau}\phi_{p\times q}(\mathcal{X} \mid \mathbf{M}, \Sigma, \Psi)\exp\left(\frac{d_{\mathbf{A}}(\mathcal{X})^2}{2\tau^2}\right)\Phi\left(\kappa + \frac{d_{\mathbf{A}}(\mathcal{X})}{\tau}\right),$$

where $\phi_{p\times q}(\cdot \mid \mathbf{M}, \Sigma, \Psi)$ denotes the MVN density.

Proof. Let $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{p\times q}$ be a random matrix defined in accordance with the previous stochastic representation. Hence its vectorization has the following representation:

$$\mathrm{vec}(\boldsymbol{\mathcal{X}}) = \mathrm{vec}(\mathbf{M}) + W\,\mathrm{vec}(\mathbf{A}) + \mathbf{V},$$

Thus we can write the marginal PDF of $\mathrm{vec}(\boldsymbol{\mathcal{X}})$ as:

$$f_{\mathrm{MESN}}(\mathbf{x}) = \frac{\phi_p(\mathbf{x} \mid \mu, \Delta)}{\Phi(\kappa/\tau)\,\tau} \exp\left\{ \frac{\left(\mathbf{b}^\top \Delta^{-1}(\mathbf{x}-\mu)\right)^2}{2(1+\delta^2)} \right\} \Phi\left(\kappa + \frac{\mathbf{b}^\top \Delta^{-1}(\mathbf{x}-\mu)}{\sqrt{1+\delta^2}} \right),$$

where $\mathbf{x} = \mathrm{vec}(\boldsymbol{\mathcal{X}})$, $\mu = \mathrm{vec}(\mathbf{M})$, $\mathbf{b} = \mathrm{vec}(\mathbf{A})$, $\Delta = \Psi \otimes \Sigma$ and $\delta^2 = \mathbf{b}^\top \Delta^{-1} \mathbf{b}$. With the purpose of obtaining the corresponding matrix-variate PDF, we now use the following useful identities:

1. $\phi_{pq}(\mathbf{x} \mid \mu, \Delta) = \phi_{p\times q}(\mathcal{X} \mid \mathbf{M}, \Sigma, \Psi)$,
2. $\mathbf{b}^\top \Delta^{-1}(\mathbf{x}-\mu) = \mathrm{tr}\left(\Sigma^{-1}(\mathcal{X}-\mathbf{M})\Psi^{-1}\mathbf{A}^\top\right)$,
3. $\mathbf{b}^\top \Delta^{-1} \mathbf{b} = \mathrm{tr}(\Sigma^{-1}\mathbf{A}\Psi^{-1}\mathbf{A}^\top)$,

which follows from the well known identity:

$$\mathrm{vec}(\mathbf{P})^\top (\Psi \otimes \Sigma)^{-1} \mathrm{vec}(\mathbf{Q}) = \mathrm{tr}(\Sigma^{-1}\mathbf{Q}\Psi^{-1}\mathbf{P}^\top).$$

Specifically, each case under consideration corresponds to a distinct pairing of $\mathbf{P}$ and $\mathbf{Q}$, as outlined in the following descriptions.

1. In this case, $\mathbf{P} = \mathbf{Q} = \mathcal{X} - \mathbf{M}$.
2. In this case, $\mathbf{P} = \mathbf{A}$ and $\mathbf{Q} = \mathcal{X} - \mathbf{M}$.
3. In this case, $\mathbf{P} = \mathbf{Q} = \mathbf{A}$.

Thus, the final form of the MVESN density is:

$$f_{\mathrm{MVESN}}(\mathcal{X}) = \frac{1}{\Phi(\kappa/\tau)\,\tau} \phi_{p\times q}(\mathcal{X} \mid \mathbf{M}, \Sigma, \Psi) \exp\left(\frac{d_{\mathbf{A}}^2(\mathcal{X})}{2\tau^2} \right) \Phi\left(\kappa + \frac{d_{\mathbf{A}}(\mathcal{X})}{\tau} \right),$$

which concludes the proof. □

To facilitate comparison, we now recall the expression for the marginal PDF of the MVSN distribution, which arises under the same stochastic representation when the latent variable W follows a half-normal distribution, $\mathrm{HN}(0,1)$. This corresponds to the particular case where $\kappa = 0$ in the MVESN framework. In this case, the density takes the form

$$f_{\mathrm{MVSN}}(\mathcal{X}) = \frac{2}{\tau} \exp\left(\frac{d_{\mathbf{A}}^2(\mathcal{X})}{2\tau^2} \right) \phi_{p\times q}(\mathcal{X} \mid \mathbf{M}, \Sigma, \Psi) \Phi\left(\frac{d_{\mathbf{A}}(\mathcal{X})}{\tau} \right),$$

Proposition 21. Let $\boldsymbol{\mathcal{X}} \sim \mathrm{MVESN}_{p\times q}(\mathbf{M}, \mathbf{A}, \Sigma, \Psi, \kappa)$ and $W \sim \mathrm{TN}(0, 1; [-\widetilde{\kappa}, +\infty))$. Therefore, conditionally on $W = w$, the random matrix $\boldsymbol{\mathcal{X}}$ is normally distributed according to the relation:

$$\boldsymbol{\mathcal{X}} \mid W = w \sim \mathcal{N}_{p\times q}(\mathbf{M} + w\mathbf{A}, \Sigma, \Psi)$$

On the other hand, conditionally on $\boldsymbol{\mathcal{X}}$, the univariate random variable W is normally truncated distributed in accordance to the relation:

$$W \mid \boldsymbol{\mathcal{X}} \sim \mathrm{TN}\left(\frac{d_{\mathbf{A}}(\mathcal{X})}{\tau^2}, \frac{1}{\tau^2}; [-\widetilde{\kappa}, \infty) \right),$$

Proof. The proof of such a result is essentially the same as the proof of the Proposition 11. The main difference lies in the domain of definition of the random variable W, from which the factor $\Phi(\tilde{\kappa})$ appears in the denominator of the PDF of W instead of $\Phi(0)$. Thus, the resulting distribution also follows a truncated normal distribution whose parameters coincide with the parameters of the HN distribution as in Proposition 11. □

Proposition 22 (CDF of the MESN distribution)**.** Let $\mathbf{X} \sim \text{MESN}_p(\mu, \mathbf{b}, \Delta, \kappa)$. Its stochastic representation is given by $\mathbf{X} = \mu + W\mathbf{b} + \mathbf{V}$, where $W \sim \text{TN}(0, 1, [-\widetilde{\kappa}, +\infty))$. Then the CDF of the random vector $\mathbf{X}$ can be expressed as follows:

$$F_{\mathbf{X}}(\mathbf{x}) = \frac{\Phi_{p+1}(\mathbf{x}_u \mid \mu^\star, \Omega) - \Phi_{p+1}(\mathbf{x}_l \mid \mu^\star, \Omega)}{\Phi(\widetilde{\kappa})},$$

where we conveniently define $\mathbf{x}_u := (\mathbf{x}^\top, +\infty)^\top$ and $\mathbf{x}_l := (\mathbf{x}^\top, -\widetilde{\kappa})^\top$, so that

$$\mu^\star = \begin{bmatrix} \mu \\ 0 \end{bmatrix}, \quad \Omega = \begin{bmatrix} \Delta + \mathbf{b}\mathbf{b}^\top & \mathbf{b} \\ \mathbf{b}^\top & 1 \end{bmatrix}.$$

Proof. Let $\mathbf{Y} = \mu + U\mathbf{b} + \mathbf{V}$, where $U \sim \mathcal{N}(0, 1)$. The central idea underlying the proof is that the joint vector $(\mathbf{Y}, U)$ can be written as an affine transformation of the pair $(\mathbf{V}, U)$, whose joint distribution is multivariate normal. This observation enables us to describe $(\mathbf{Y}, U)$ as having a Gaussian structure.

$$\begin{bmatrix} \mathbf{V} \\ U \end{bmatrix} \sim \mathcal{N}_{p+1}\left(\begin{bmatrix} \mathbf{0} \\ 0 \end{bmatrix}, \begin{bmatrix} \Delta & \mathbf{0} \\ \mathbf{0}^\top & 1 \end{bmatrix} \right).$$

This holds because any linear combination $\mathbf{a}^\top\mathbf{V} + bU$ (with $\mathbf{a}^\top \in \mathbb{R}^p$) is normally distributed, since $\mathbf{a}^\top\mathbf{V} \perp bU$ and the sum of independent Gaussian variables remains Gaussian. Moreover, a vector-valued random variable is multivariate normal iff every linear combination of its components is Gaussian, and therefore the claim follows. With this in place, the relationship between $\mathbf{Y}$ and U can be written as:

$$\begin{bmatrix} \mathbf{Y} \\ U \end{bmatrix} = \begin{bmatrix} \mathbf{I}_p & \mathbf{b} \\ \mathbf{0}^\top & 1 \end{bmatrix} \begin{bmatrix} \mathbf{V} \\ U \end{bmatrix} + \begin{bmatrix} \mu \\ 0 \end{bmatrix},$$

which implies that $(\mathbf{Y}, U)$ follows a $(p+1)$-dimensional normal distribution with mean vector $\mu^\star$ and covariance matrix Ω, as defined in the statement of the proposition. Indeed, one has that:

- $\text{Var}(\mathbf{Y}) = \text{Var}(U\mathbf{b} + \mathbf{V}) = \mathbf{b}\,\text{Var}(U)\mathbf{b}^\top + \text{Var}(\mathbf{V}) = \mathbf{b}\mathbf{b}^\top + \Delta$,
- $\text{Cov}(\mathbf{Y}, U) = \mathbb{E}[(\mathbf{Y} - \mu)U] = \mathbb{E}[(U\mathbf{b} + \mathbf{V})U] = \mathbb{E}[U^2]\mathbf{b} + \mathbb{E}[\mathbf{V}U] = \mathbf{b} = \text{Cov}(U, \mathbf{Y})^\top$.

where $\text{Var}(U) = 1$. We now compute the CDF of $\mathbf{X}$, which is given by $\mathbb{P}(\mathbf{X} \leq \mathbf{x})$. A convenient way to introduce the CDF of $\mathbf{X}$ is to note that its distribution can be expressed by conditioning the joint normal vector $(\mathbf{Y}, U)$ on the event $U \geq -\widetilde{\kappa}$. This leads to a representation entirely in

terms of $(p+1)$-variate normal probabilities: the numerator gives the probability that $\mathbf{Y} \leq \mathbf{x}$ while U remains in $[-\widetilde{\kappa}, \infty)$, and the denominator accounts for the corresponding normalization.

$$F_{\mathbf{X}}(\mathbf{x}) = \mathbb{P}(\mathbf{Y} \leq \mathbf{x} \mid U \geq -\widetilde{\kappa}) = \frac{\mathbb{P}(\mathbf{Y} \leq \mathbf{x}, U \geq -\widetilde{\kappa})}{\mathbb{P}(U \geq -\widetilde{\kappa})} = \frac{\Phi_{p+1}(\mathbf{x}, +\infty) - \Phi_{p+1}(\mathbf{x}, -\widetilde{\kappa})}{\Phi(\widetilde{\kappa})}.$$

This completes the proof. □

Remark. A closely related formulation appears in Morales *et al.* (2022), in which the expressions, though presented differently and derived through an alternative route, remain fully consistent and mathematically equivalent to ours.

Proposition 23 (CDF of the MVESN distribution)**.** If $\boldsymbol{\mathcal{X}} \sim \text{MVESN}_{p\times q}(\mathbf{M}, \mathbf{A}, \Sigma, \Psi, \kappa)$, then the CDF of $\boldsymbol{\mathcal{X}}$ is given by:

$$F_{\text{MVESN}}(\text{vec}(\boldsymbol{\mathcal{X}})) = \frac{\Phi_{pq+1}(\mathbf{x}_u \mid \mu^\star, \Omega) - \Phi_{pq+1}(\mathbf{x}_l \mid \mu^\star, \Omega)}{\Phi(\widetilde{\kappa})},$$

where $\mathbf{x}_u = (\text{vec}(\boldsymbol{\mathcal{X}})^\top, +\infty)^\top$, $\mathbf{x}_l = (\text{vec}(\boldsymbol{\mathcal{X}})^\top, -\widetilde{\kappa})^\top$, with

$$\mu^\star = \begin{bmatrix} \text{vec}(\mathbf{M}) \\ 0 \end{bmatrix} \in \mathbb{R}^{pq+1}, \quad \Omega = \begin{bmatrix} \Psi \otimes \Sigma + \text{vec}(\mathbf{A})\,\text{vec}(\mathbf{A})^\top & \text{vec}(\mathbf{A}) \\ \text{vec}(\mathbf{A})^\top & 1 \end{bmatrix} \in \mathbb{R}^{(pq+1)\times(pq+1)}.$$

Proof. The random matrix $\boldsymbol{\mathcal{X}}$ follows the MVESN distribution, which corresponds to claim that its stochastic representation is given by $\boldsymbol{\mathcal{X}} = \mathbf{M} + W\mathbf{A} + \mathcal{V}$, with the latent variable $W \sim \mathcal{N}(0,1)$ truncated to $[-\widetilde{\kappa}, +\infty)$ and $\mathcal{V} \sim \mathcal{N}_{p\times q}(\mathbf{0}, \Sigma, \Psi)$.

By applying the vectorization operator, we obtain:

$$\text{vec}(\boldsymbol{\mathcal{X}}) = \text{vec}(\mathbf{M}) + W\,\text{vec}(\mathbf{A}) + \mathbf{V},$$

where $\mathbf{V} \sim \mathcal{N}_{pq}(\mathbf{0}, \Psi \otimes \Sigma)$. The resulting stochastic representation aligns exactly with the MESN form previously considered, with p replaced by pq. Therefore, the CDF of the matrix-valued random variable $\boldsymbol{\mathcal{X}}$ is equivalent to the CDF of a MESN vector with augmented dimension pq, as given above. □

Proposition 24. Let $\boldsymbol{\mathcal{X}} \sim \text{MVESN}_{p\times q}(\mathbf{M}, \mathbf{A}, \Sigma, \Psi, \kappa)$, that is, $\boldsymbol{\mathcal{X}} = \mathbf{M} + W\mathbf{A} + \boldsymbol{\mathcal{V}}$. Then the row covariance matrix C_{row} and the column covariance matrix C_{column} are respectively given by:

$$C_{\text{row}} = \mathbb{E}((\boldsymbol{\mathcal{X}} - \mathbb{E}(\boldsymbol{\mathcal{X}}))(\boldsymbol{\mathcal{X}} - \mathbb{E}(\boldsymbol{\mathcal{X}}))^\top) = \sigma_W^2 \mathbf{A}\mathbf{A}^\top + \text{tr}(\Psi)\Sigma$$
$$C_{\text{column}} = \mathbb{E}((\boldsymbol{\mathcal{X}} - \mathbb{E}(\boldsymbol{\mathcal{X}}))^\top(\boldsymbol{\mathcal{X}} - \mathbb{E}(\boldsymbol{\mathcal{X}}))) = \sigma_W^2 \mathbf{A}^\top\mathbf{A} + \text{tr}(\Sigma)\Psi.$$

where σ_W^2 is the variance of the random variable W.

Proof. To begin with, let us rearrange the inner expression of the row covariance matrix definition as follows:

$$\begin{aligned}(\boldsymbol{\mathcal{X}} - \mathbb{E}(\boldsymbol{\mathcal{X}}))(\boldsymbol{\mathcal{X}} - \mathbb{E}(\boldsymbol{\mathcal{X}}))^\top &= ((W - \mu_W)\mathbf{A} + \boldsymbol{\mathcal{V}})((W - \mu_W)\mathbf{A} + \boldsymbol{\mathcal{V}})^\top \\ &= (W - \mu_W)^2\mathbf{A}\mathbf{A}^\top + (W - \mu_W)(\mathbf{A}\boldsymbol{\mathcal{V}}^\top + \boldsymbol{\mathcal{V}}\mathbf{A}^\top) + \boldsymbol{\mathcal{V}}\boldsymbol{\mathcal{V}}^\top\end{aligned}$$

Moreover, due to the independence between W and $\mathcal{V}$, it also can be claimed that:

$$\begin{aligned}\mathbb{E}((W-\mu_W)(\mathbf{A}\boldsymbol{\mathcal{V}}^\top+\boldsymbol{\mathcal{V}}\mathbf{A}^\top)) &= \mathbf{A}\mathbb{E}((W-\mu_W)\boldsymbol{\mathcal{V}}^\top)+\mathbb{E}((W-\mu_W)\boldsymbol{\mathcal{V}})\mathbf{A}^\top \\ &= \mathbf{A}\mathbb{E}(W-\mu_W)\mathbb{E}(\boldsymbol{\mathcal{V}})^\top+\mathbb{E}(W-\mu_W)\mathbb{E}(\boldsymbol{\mathcal{V}})\mathbf{A}^\top \\ &= \mathbf{0}.\end{aligned}$$

With respect to the fisrt term, based on the definition of variance, it can be claimed that:

$$\mathbb{E}((W-\mu_W)^2\mathbf{A}\mathbf{A}^\top)=\mathbb{E}((W-\mu_W)^2)\mathbf{A}\mathbf{A}^\top=\sigma_W^2\mathbf{A}\mathbf{A}^\top.$$

At last, but not least, the expected value $\mathbb{E}(\boldsymbol{\mathcal{V}}\boldsymbol{\mathcal{V}}^\top)$ can be handled in the following manner. To start with, note that we can write $\boldsymbol{\mathcal{V}}=\Sigma^{1/2}\mathcal{Z}\Psi^{1/2}$, where $\boldsymbol{\mathcal{Z}}\sim\mathcal{N}_{p\times q}(\mathbf{0},\mathbf{I}_p,\mathbf{I}_q)$ so that $\boldsymbol{\mathcal{Z}}_{ij}\overset{\text{iid}}{\sim}\mathcal{N}(0,1)$. Consequently, it can be claimed that:

$$\boldsymbol{\mathcal{V}}\boldsymbol{\mathcal{V}}^\top=\Sigma^{1/2}\boldsymbol{\mathcal{Z}}\Psi\boldsymbol{\mathcal{Z}}^\top\Sigma^{1/2}\Rightarrow\mathbb{E}(\boldsymbol{\mathcal{V}}\boldsymbol{\mathcal{V}}^\top)=\Sigma^{1/2}\mathbb{E}(\boldsymbol{\mathcal{Z}}\Psi\boldsymbol{\mathcal{Z}}^\top)\Sigma^{1/2}.$$

Thus the original problem is reduced to calculate the expression $\mathbb{E}(\boldsymbol{\mathcal{Z}}\Psi\boldsymbol{\mathcal{Z}}^\top)$. To handle it, we may approach it by taking into account each entry separately:

$$\begin{aligned}(\boldsymbol{\mathcal{Z}}\Psi\boldsymbol{\mathcal{Z}}^\top)_{ij}=\sum_{k=1}^{q}\sum_{l=1}^{q}\boldsymbol{\mathcal{Z}}_{ik}\Psi_{kl}\boldsymbol{\mathcal{Z}}_{jl}\Rightarrow\mathbb{E}((\boldsymbol{\mathcal{Z}}\Psi\boldsymbol{\mathcal{Z}}^\top)_{ij})&=\sum_{k=1}^{q}\sum_{l=1}^{q}\Psi_{kl}\mathbb{E}(\boldsymbol{\mathcal{Z}}_{ik}\boldsymbol{\mathcal{Z}}_{jl}) \\ \Rightarrow\mathbb{E}((\boldsymbol{\mathcal{Z}}\Psi\boldsymbol{\mathcal{Z}}^\top)_{ij})&=\sum_{k=1}^{q}\sum_{l=1}^{q}\Psi_{kl}\delta_{ij}\delta_{kl}\end{aligned}$$

where δ represents the Kronecker delta. That is because the target expression $\mathbb{E}(\boldsymbol{\mathcal{Z}}_{ik}\boldsymbol{\mathcal{Z}}_{jl})$ equals $\mathbb{E}(\boldsymbol{\mathcal{Z}}_{ik}\boldsymbol{\mathcal{Z}}_{jl})=\mathbb{E}(\boldsymbol{\mathcal{Z}}_{ik}^2)=1$ if $(i,k)=(j,l)$ and $\mathbb{E}(\boldsymbol{\mathcal{Z}}_{ik}\boldsymbol{\mathcal{Z}}_{jl})=0$ if $(i,k)\neq(j,l)$ since each entry of $\boldsymbol{\mathcal{Z}}$ is normally distributed and independent of each other as previously mentioned. Therefore we may claim that the following holds:

$$\mathbb{E}((\boldsymbol{\mathcal{Z}}\Psi\boldsymbol{\mathcal{Z}}^\top)_{ij})=\sum_{k=1}^{q}\sum_{l=1}^{q}\Psi_{kl}\delta_{ij}\delta_{kl}=\delta_{ij}\sum_{k=1}^{q}\sum_{l=1}^{q}\Psi_{kl}\delta_{kl}=\delta_{ij}\sum_{k=1}^{q}\Psi_{kk}=\delta_{ij}\operatorname{tr}(\Psi),$$

from where it can be deduced that $\mathbb{E}(\boldsymbol{\mathcal{Z}}\Psi\boldsymbol{\mathcal{Z}}^\top)=\operatorname{tr}(\Psi)\mathbf{I}_p$. Finally, one obtains the result:

$$\begin{aligned}\mathbb{E}(\boldsymbol{\mathcal{V}}\boldsymbol{\mathcal{V}}^\top) &= \Sigma^{1/2}\mathbb{E}(\boldsymbol{\mathcal{Z}}\Psi\boldsymbol{\mathcal{Z}}^\top)\Sigma^{1/2} \\ &= \Sigma^{1/2}(\operatorname{tr}(\Psi)\mathbf{I}_p)\Sigma^{1/2} \\ &= \operatorname{tr}(\Psi)\Sigma^{1/2}\mathbf{I}_p\Sigma^{1/2} \\ &= \operatorname{tr}(\Psi)\Sigma.\end{aligned}$$

just as we wanted to demonstrate. A similar approach proves the other identity. □

To recover the standard matrix-variate skew-normal (MVSN) model, it suffices to set $\kappa=0$, so that $W\sim\text{TN}(0,1,[0,\infty))$, that is, a half-normal latent variable. In this case, the specification reduces to the usual MVSN formulation, and its distribution function can be expressed in terms of Gaussian probabilities (equivalently, evaluations of multivariate normal cumulative distribution functions).

Definition 15. Following the general framework of higher-order moments for Banach space–valued random elements Janson and Kaijser (2015), a random matrix $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{p\times q}$ may be regarded as a random element of the finite-dimensional vector space $\mathbb{R}^{p\times q}$. In this setting, its k-th central moment is naturally defined through the k-fold tensor product of the centered random matrix. Formally, the k-th central moment of any random matrix $\boldsymbol{\mathcal{X}}$ (provided that it exists) is given by:

$$\mathcal{M}_k(\boldsymbol{\mathcal{X}}) := \mathbb{E}\Big[(\boldsymbol{\mathcal{X}} - \mathbb{E}[\boldsymbol{\mathcal{X}}])^{\otimes k}\Big],$$

where the operator $\otimes$ denotes the tensor product taken in $(\mathbb{R}^{p\times q})^{\otimes k}$.

Lemma 11. Let $\boldsymbol{\mathcal{V}} \sim \mathcal{N}_{p\times q}(\mathbf{0}, \Sigma, \Psi)$ be a matrix-variate normal random variable with zero mean. Consider arbitrary indices $i, j, k \in \{1, 2, \ldots, p\}$ and $\alpha, \beta, \gamma \in \{1, 2, \ldots, q\}$. Then, owing to the joint Gaussianity and symmetry of the underlying distribution, the expectation of any triple product of its entries necessarily vanishes. In particular, one has

$$\mathbb{E}\big(\boldsymbol{\mathcal{V}}_{i\alpha}\, \boldsymbol{\mathcal{V}}_{j\beta}\, \boldsymbol{\mathcal{V}}_{k\gamma}\big) = 0.$$

Proof. By the definition of the matrix variate normal distribution, $\mathrm{vec}(\boldsymbol{\mathcal{V}}) \sim \mathcal{N}_{pq}(0, \Psi \otimes \Sigma)$. Hence any finite sub-collection of components of $\mathrm{vec}(\boldsymbol{\mathcal{V}})$ is jointly Gaussian with mean zero. Having said that, consider the 3-dimensional vector $\mathbf{Z} := (Z_1, Z_2, Z_3)^\top := (\boldsymbol{\mathcal{V}}_{i\alpha}, \boldsymbol{\mathcal{V}}_{j\beta}, \boldsymbol{\mathcal{V}}_{k\gamma})^\top$. Then $\mathbf{Z} \sim \mathcal{N}_3(\mathbf{0}, \Lambda)$ where Λ is some covariance matrix extractable from $\Psi \otimes \Sigma$. Therefore, its joint probability density function exists and can be written as:

$$f_{\mathbf{Z}}(z) = \frac{1}{(2\pi)^{3/2}|\Lambda|^{1/2}} \exp\left\{-\frac{1}{2} z^\top \Lambda^{-1} z\right\}$$

where $z \in \mathbb{R}^3$. Observe that the joint density $f_{\mathbf{Z}}(z)$ is an even function, in the sense that it satisfies the symmetry property $f_{\mathbf{Z}}(z) = f_{\mathbf{Z}}(-z)$ for every $z \in \mathbb{R}^3$. Consequently, by performing the change of variables $u = -z$, we obtain

$$\begin{aligned}
\mathbb{E}(Z_1 Z_2 Z_3) &= \int_{\mathbb{R}^3} z_1 z_2 z_3 \, f_{\mathbf{Z}}(z)\, \mathrm{d}z \\
&= \int_{\mathbb{R}^3} (-u_1)(-u_2)(-u_3)\, f_{\mathbf{Z}}(-u)\, \mathrm{d}u \\
&= -\int_{\mathbb{R}^3} u_1 u_2 u_3 \, f_{\mathbf{Z}}(u)\, \mathrm{d}u \\
&= -\mathbb{E}(Z_1 Z_2 Z_3),
\end{aligned}$$

which implies $\mathbb{E}(Z_1 Z_2 Z_3) = 0$. and the desired relation holds, just as we wanted to demonstrate. □

Proposition 25. Let $\boldsymbol{\mathcal{X}} \sim \mathrm{MVESN}_{p\times q}(\mathbf{M}, \mathbf{A}, \Sigma, \Psi, \kappa)$ be a random matrix distributed according to the matrix-variate extended skew-normal model. In this setting, the third central moment of $\boldsymbol{\mathcal{X}}$, denoted by $\mathcal{M}_3(\boldsymbol{\mathcal{X}})$, admits the following explicit representation:

$$\mathcal{M}_3(\boldsymbol{\mathcal{X}}) = \mu_3(\mathbf{A} \otimes \mathbf{A} \otimes \mathbf{A}).$$

where μ_3 corresponds to third central moment of W given by $\mathbb{E}(W - \mathbb{E}(W))^3$

Proof. Let $\mu_W := \mathbb{E}(W)$. Then it may be concluded that:

$$\mathbb{E}(\boldsymbol{\mathcal{X}}) = \mathbb{E}(\mathbf{M}) + \mathbb{E}(W)\mathbf{A} + \mathbb{E}(\boldsymbol{\mathcal{V}}) = \mathbf{M} + \mu_W\mathbf{A}$$

For the sake of convenience, we shall adopt the next notation within this proof:

$$\widetilde{\boldsymbol{\mathcal{X}}} = \boldsymbol{\mathcal{X}} - \mathbb{E}(\boldsymbol{\mathcal{X}}) = (W - \mu_W)\mathbf{A} + \boldsymbol{\mathcal{V}}.$$

Consequently, it follows that $\mathcal{M}_3^{i\alpha,j\beta,k\gamma}(\boldsymbol{\mathcal{X}}) = \mathbb{E}\left(\widetilde{\boldsymbol{\mathcal{X}}}_{i\alpha}\widetilde{\boldsymbol{\mathcal{X}}}_{j\beta}\widetilde{\boldsymbol{\mathcal{X}}}_{k\gamma}\right)$, where the indices $i,j,k \in \{1,2,\ldots,p\}$ and $\alpha,\beta,\gamma \in \{1,2,\ldots,q\}$ identify the corresponding row and column positions within the matrix. With this notation and indexing structure in place, we are now in a position to carefully expand the product and examine each term in detail, thereby clarifying the contributions that enter the resulting expression.

$$\begin{aligned}
\widetilde{\boldsymbol{\mathcal{X}}}_{i\alpha}\widetilde{\boldsymbol{\mathcal{X}}}_{j\beta}\widetilde{\boldsymbol{\mathcal{X}}}_{k\gamma} &= ((W-\mu_W)\mathbf{A}_{i\alpha} + \boldsymbol{\mathcal{V}}_{i\alpha})((W-\mu_W)\mathbf{A}_{j\beta} + \boldsymbol{\mathcal{V}}_{j\beta})((W-\mu_W)\mathbf{A}_{k\gamma} + \boldsymbol{\mathcal{V}}_{k\gamma}) \\
&= (W-\mu_W)^3\mathbf{A}_{i\alpha}\mathbf{A}_{j\beta}\mathbf{A}_{k\gamma} + (W-\mu_W)^2(\mathbf{A}_{i\alpha}\mathbf{A}_{j\beta}\boldsymbol{\mathcal{V}}_{k\gamma} + \mathbf{A}_{i\alpha}\boldsymbol{\mathcal{V}}_{j\beta}\mathbf{A}_{k\gamma} + \boldsymbol{\mathcal{V}}_{i\alpha}\mathbf{A}_{j\beta}\mathbf{A}_{k\gamma}) \\
&+ (W-\mu_W)(\mathbf{A}_{i\alpha}\boldsymbol{\mathcal{V}}_{j\beta}\boldsymbol{\mathcal{V}}_{k\gamma} + \boldsymbol{\mathcal{V}}_{i\alpha}\mathbf{A}_{j\beta}\boldsymbol{\mathcal{V}}_{k\gamma} + \boldsymbol{\mathcal{V}}_{i\alpha}\boldsymbol{\mathcal{V}}_{j\beta}\mathbf{A}_{k\gamma}) + \boldsymbol{\mathcal{V}}_{i\alpha}\boldsymbol{\mathcal{V}}_{j\beta}\boldsymbol{\mathcal{V}}_{k\gamma}.
\end{aligned}$$

We may now take expectations on both sides of the equality and, by appealing to the linearity of the expectation operator $\mathbb{E}$, decompose the expression and analyze each resulting term separately. In doing so, this step allows us to systematically isolate the contribution of each component of the expansion. To initiate this analysis, observe that:

$$\mathbb{E}((W-\mathbb{E}(W))^3\mathbf{A}_{i\alpha}\mathbf{A}_{j\beta}\mathbf{A}_{k\gamma}) = \mathbb{E}(W-\mathbb{E}(W))^3\mathbf{A}_{i\alpha}\mathbf{A}_{j\beta}\mathbf{A}_{k\gamma} = \mu_3\mathbf{A}_{i\alpha}\mathbf{A}_{j\beta}\mathbf{A}_{k\gamma}.$$

Moreover, due to the independence between W and $\boldsymbol{\mathcal{V}}$, it results that:

$$\mathbb{E}((W-\mathbb{E}(W))^2\mathbf{A}_{i\alpha}\mathbf{A}_{j\beta}\boldsymbol{\mathcal{V}}_{k\gamma}) = \mathbf{A}_{i\alpha}\mathbf{A}_{j\beta}\mathbb{E}(W-\mathbb{E}(W))^2\mathbb{E}(\boldsymbol{\mathcal{V}}_{k\gamma}) = 0.$$

Once again, because of the independence between W and $\boldsymbol{\mathcal{V}}$, one may conclude that:

$$\mathbb{E}((W-\mathbb{E}(W))\mathbf{A}_{i\alpha}\boldsymbol{\mathcal{V}}_{j\beta}\boldsymbol{\mathcal{V}}_{k\gamma}) = \mathbf{A}_{i\alpha}\mathbb{E}(W-\mu_W)\mathbb{E}(\boldsymbol{\mathcal{V}}_{j\beta}\boldsymbol{\mathcal{V}}_{k\gamma}) = 0.$$

Finally, and no less importantly, we turn our attention to the remaining term: because the entries of $\boldsymbol{\mathcal{V}}$ form a jointly Gaussian vector with zero mean, we have $\mathbb{E}(\boldsymbol{\mathcal{V}}_{i\alpha}\boldsymbol{\mathcal{V}}_{j\beta}\boldsymbol{\mathcal{V}}_{k\gamma}) = 0$, since all third-order central moments of a zero-mean Gaussian distribution are identically zero in accordance with Lemma 11. Consequently, after keeping only the surviving nonvanishing terms, we arrive at the desired relation.

$$\mathcal{M}_3^{i\alpha,j\beta,k\gamma}(\boldsymbol{\mathcal{X}}) = \mu_3\mathbf{A}_{i\alpha}\mathbf{A}_{j\beta}\mathbf{A}_{k\gamma},$$

that is to say, $\mathcal{M}_3(\boldsymbol{\mathcal{X}}) = \mu_3\mathbf{A}^{\otimes 3}$, just as we wanted to demonstrate. □

Proposition 26. Let $\boldsymbol{\mathcal{X}} \sim \text{MVESN}_{p\times q}(\mathbf{M},\mathbf{A},\Sigma,\Psi,\kappa)$ denote a random matrix following the matrix-variate extended skew normal distribution as in Definition 14. In such context, higher-order dependence and tail behavior can be characterized through central moments. In particular, the fourth central moment of $\boldsymbol{\mathcal{X}}$, denoted by $\mathcal{M}_4(\boldsymbol{\mathcal{X}})$, admits the following explicit analytical representation.

$$\begin{aligned}\mathcal{M}_4^{ij,k\ell,mn,rs}(\boldsymbol{\mathcal{X}}) &= \mu_4\mathbf{A}_{ij}\mathbf{A}_{k\ell}\mathbf{A}_{mn}\mathbf{A}_{rs}\\ &\quad+\sigma_W^2(\mathbf{A}_{ij}\mathbf{A}_{k\ell}\Sigma_{mn}\Psi_{rs}+\mathbf{A}_{ij}\mathbf{A}_{mn}\Sigma_{kr}\Psi_{\ell s}+\mathbf{A}_{ij}\mathbf{A}_{rs}\Sigma_{km}\Psi_{\ell n})\\ &\quad+\sigma_W^2(\mathbf{A}_{k\ell}\mathbf{A}_{mn}\Sigma_{ir}\Psi_{js}+\mathbf{A}_{k\ell}\mathbf{A}_{rs}\Sigma_{im}\Psi_{jn}+\mathbf{A}_{mn}\mathbf{A}_{rs}\Sigma_{km}\Psi_{\ell n})\\ &\quad+(\Sigma_{ik}\Psi_{j\ell})(\Sigma_{mr}\Psi_{ns})+(\Sigma_{im}\Psi_{jn})(\Sigma_{kr}\Psi_{\ell s})+(\Sigma_{ir}\Psi_{js})(\Sigma_{km}\Psi_{\ell n}),\end{aligned}$$

where $i,k,m,r\in\{1,2,\ldots,p\}$ and $j,\ell,n,s\in\{1,2,\ldots,q\}$ so that $\mu_4=\mathbb{E}[(W-\mathbb{E}(W))^3]$ is the fourth central moment of the random variable W.

Proof. Given the model $\boldsymbol{\mathcal{X}}\sim\text{MVESN}_{p\times q}(\mathbf{M},\mathbf{A},\Sigma,\Psi,\kappa)$, we are going to convention (within the context of this proof) that $\mu:=\mathbb{E}(W)$, $\widetilde{\boldsymbol{\mathcal{X}}}=\boldsymbol{\mathcal{X}}-\mathbb{E}(\boldsymbol{\mathcal{X}})=\widetilde{W}\mathbf{A}+\boldsymbol{\mathcal{V}}$, where $\widetilde{W}:=W-\mathbb{E}(W)$. Entry-wise, for any $i\in\{1,2,\ldots,p\}$ and $j\in\{1,2,\ldots,q\}$, one has that $\widetilde{\boldsymbol{\mathcal{X}}}_{ij}=\widetilde{W}\mathbf{A}_{ij}+\boldsymbol{\mathcal{V}}_{ij}$. On the basis of the preceding considerations, the expression of interest can be written as

$$\begin{aligned}\mathcal{M}_4^{ij,k\ell,mn,rs}&:=\mathbb{E}(\widetilde{\boldsymbol{\mathcal{X}}}_{ij}\widetilde{\boldsymbol{\mathcal{X}}}_{k\ell}\widetilde{\boldsymbol{\mathcal{X}}}_{mn}\widetilde{\boldsymbol{\mathcal{X}}}_{rs})\\ &=\mathbb{E}((\widetilde{W}\mathbf{A}_{ij}+\boldsymbol{\mathcal{V}}_{ij})(\widetilde{W}\mathbf{A}_{k\ell}+\boldsymbol{\mathcal{V}}_{k\ell})(\widetilde{W}\mathbf{A}_{mn}+\boldsymbol{\mathcal{V}}_{mn})(\widetilde{W}\mathbf{A}_{rs}+\boldsymbol{\mathcal{V}}_{rs}))\end{aligned}$$

If we establish the convention that

$$\boldsymbol{\mathcal{Z}}^{ij,k\ell,mn,rs}:=(\widetilde{W}\mathbf{A}_{ij}+\boldsymbol{\mathcal{V}}_{ij})(\widetilde{W}\mathbf{A}_{k\ell}+\boldsymbol{\mathcal{V}}_{k\ell})(\widetilde{W}\mathbf{A}_{mn}+\boldsymbol{\mathcal{V}}_{mn})(\widetilde{W}\mathbf{A}_{rs}+\boldsymbol{\mathcal{V}}_{rs}),$$

then fourth central moment entries satisfy the relation

$$\mathcal{M}_4(\boldsymbol{\mathcal{X}})=\mathbb{E}(\mathbb{E}(\boldsymbol{\mathcal{Z}}\mid\widetilde{W})).$$

Inside the conditional expectation, we shall split its arguments into the shift and the Gaussian part. For convenience, define the deterministic (given $\widetilde{W}$) shifts

$$S_{ij}:=\widetilde{W}\mathbf{A}_{ij},\quad S_{k\ell}:=\widetilde{W}\mathbf{A}_{k\ell},\quad S_{mn}:=\widetilde{W}\mathbf{A}_{mn},\quad S_{rs}:=\widetilde{W}\mathbf{A}_{rs}.$$

Conditionally on $\widetilde{W}$, we assume

$$\widetilde{\boldsymbol{\mathcal{X}}}_{ij}=S_{ij}+\boldsymbol{\mathcal{V}}_{ij},\quad \widetilde{\boldsymbol{\mathcal{X}}}_{k\ell}=S_{k\ell}+\boldsymbol{\mathcal{V}}_{k\ell},\quad \widetilde{\boldsymbol{\mathcal{X}}}_{mn}=S_{mn}+\boldsymbol{\mathcal{V}}_{mn},\quad \widetilde{\boldsymbol{\mathcal{X}}}_{rs}=S_{rs}+\boldsymbol{\mathcal{V}}_{rs}.$$

from where the next expression follows

$$\mathbb{E}(\boldsymbol{\mathcal{Z}}^{ij,k\ell,mn,rs}\mid\widetilde{W})=\mathbb{E}((S_{ij}+\boldsymbol{\mathcal{V}}_{ij})(S_{k\ell}+\boldsymbol{\mathcal{V}}_{k\ell})(S_{mn}+\boldsymbol{\mathcal{V}}_{mn})(S_{rs}+\boldsymbol{\mathcal{V}}_{rs})\mid\widetilde{W}).$$

Now expand conceptually by grouping terms according to the number of $\boldsymbol{\mathcal{V}}$'s appearing:

- 0 Gaussian terms: $S_{ij}S_{k\ell}S_{mn}S_{rs}$.
- 1 Gaussian term: $S_{ij}S_{k\ell}S_{mn}\boldsymbol{\mathcal{V}}_{rs}$ (and other three possibilities).
- 2 Gaussian terms: $S_{ij}S_{k\ell}\boldsymbol{\mathcal{V}}_{mn}\boldsymbol{\mathcal{V}}_{rs}$ (and five other pairings).
- 3 Gaussian terms: $S_{ij}\boldsymbol{\mathcal{V}}_{k\ell}\boldsymbol{\mathcal{V}}_{mn}\boldsymbol{\mathcal{V}}_{rs}$ (and three others).
- 4 Gaussian terms: $\boldsymbol{\mathcal{V}}_{ij}\boldsymbol{\mathcal{V}}_{k\ell}\boldsymbol{\mathcal{V}}_{mn}\boldsymbol{\mathcal{V}}_{rs}$.

Because $\boldsymbol{\mathcal{V}}$ is jointly centered Gaussian, the odd terms vanish. Thus, it remains to deal with the surviving terms with 0, 2 and 4 Gaussian terms. In other words, one has the following expression:

$$\mathbb{E}(\boldsymbol{\mathcal{Z}}^{ij,k\ell,mn,rs} \mid \widetilde{W}) = S_{ij}S_{k\ell}S_{mn}S_{rs} + \sum_{\text{6choices}} S_\alpha S_\beta \mathbb{E}(\boldsymbol{\mathcal{V}}_\gamma \boldsymbol{\mathcal{V}}_\delta) + \mathbb{E}(\boldsymbol{\mathcal{V}}_{ij}\boldsymbol{\mathcal{V}}_{k\ell}\boldsymbol{\mathcal{V}}_{mn}\boldsymbol{\mathcal{V}}_{rs}).$$

where $(\alpha,\beta,\gamma,\delta)$ runs over the six ways to choose which two slots are Gaussian. Consequently, based on Isserlis's Theorem, it may be deduced that

$$\begin{aligned}\mathbb{E}(\boldsymbol{\mathcal{V}}_{ij}\boldsymbol{\mathcal{V}}_{k\ell}\boldsymbol{\mathcal{V}}_{mn}\boldsymbol{\mathcal{V}}_{rs}) &= (\Sigma_{ik}\Psi_{j\ell})(\Sigma_{mr}\Psi_{ns}) \\ &\quad + (\Sigma_{im}\Psi_{jn})(\Sigma_{kr}\Psi_{\ell s}) \\ &\quad + (\Sigma_{ir}\Psi_{js})(\Sigma_{km}\Psi_{\ell n}),\end{aligned}$$

since $\mathbb{E}(\boldsymbol{\mathcal{V}}_{ab}\boldsymbol{\mathcal{V}}_{cd}) = \Sigma_{ac}\Psi_{bd}$. We may now write out the 0 and 2 blocks explicitly. On the one hand, the 0 block corresponds to

$$\begin{aligned}S_{ij}S_{k\ell}S_{mn}S_{rs} &= (\widetilde{W}\mathbf{A}_{ij})(\widetilde{W}\mathbf{A}_{k\ell})(\widetilde{W}\mathbf{A}_{mn})(\widetilde{W}\mathbf{A}_{rs}) \\ &= \widetilde{W}^4\mathbf{A}_{ij}\mathbf{A}_{k\ell}\mathbf{A}_{mn}\mathbf{A}_{rs}.\end{aligned}$$

On the other hand, the 2 blocks can be expressed as follows:

$$\begin{aligned}S_\alpha S_\beta \mathbb{E}(\boldsymbol{\mathcal{V}}_\gamma \boldsymbol{\mathcal{V}}_\delta) &= (\widetilde{W}\mathbf{A}_\alpha)(\widetilde{W}\mathbf{A}_\beta)\mathbb{E}(\boldsymbol{\mathcal{V}}_\gamma \boldsymbol{\mathcal{V}}_\delta) \\ &= \widetilde{W}^2\mathbf{A}_\alpha\mathbf{A}_\beta\mathbb{E}(\boldsymbol{\mathcal{V}}_\gamma \boldsymbol{\mathcal{V}}_\delta).\end{aligned}$$

Gathering everything discussed so far and taking the expected value with respect to $\widetilde{W}$ in the final expression, one arrives at the following formula for the fourth central moment of the random matrix $\boldsymbol{\mathcal{X}}$

$$\begin{aligned}\mathbb{E}(\widetilde{\boldsymbol{\mathcal{X}}}_{ij}\widetilde{\boldsymbol{\mathcal{X}}}_{k\ell}\widetilde{\boldsymbol{\mathcal{X}}}_{mn}\widetilde{\boldsymbol{\mathcal{X}}}_{rs}) &= \mu_4\mathbf{A}_{ij}\mathbf{A}_{k\ell}\mathbf{A}_{mn}\mathbf{A}_{rs} \\ &\quad + \sigma_W^2(\mathbf{A}_{ij}\mathbf{A}_{k\ell}\Sigma_{mn}\Psi_{rs} + \mathbf{A}_{ij}\mathbf{A}_{mn}\Sigma_{kr}\Psi_{\ell s} + \mathbf{A}_{ij}\mathbf{A}_{rs}\Sigma_{km}\Psi_{\ell n}) \\ &\quad + \sigma_W^2(\mathbf{A}_{k\ell}\mathbf{A}_{mn}\Sigma_{ir}\Psi_{js} + \mathbf{A}_{k\ell}\mathbf{A}_{rs}\Sigma_{im}\Psi_{jn} + \mathbf{A}_{mn}\mathbf{A}_{rs}\Sigma_{km}\Psi_{\ell n}) \\ &\quad + (\Sigma_{ik}\Psi_{j\ell})(\Sigma_{mr}\Psi_{ns}) + (\Sigma_{im}\Psi_{jn})(\Sigma_{kr}\Psi_{\ell s}) + (\Sigma_{ir}\Psi_{js})(\Sigma_{km}\Psi_{\ell n}),\end{aligned}$$

where μ_4 is the fourth central moment of the random variable W. □

3.4 ML Estimation for the MVSN Distribution

To implement the ECM algorithm for estimating the parameters of the MVSN distribution — building upon the seminal EM framework of Dempster, Laird and Rubin (1977) and its conditional maximization extension developed in Meng and Rubin (1993) — we proceed by exploiting the latent-variable representation of the model. This representation expresses the skew-normal structure through an augmented formulation involving an unobserved mixing variable, which allows the likelihood function to be handled within a more tractable complete-data framework.

The central idea of the EM-type methodology is to treat the unobserved variable W as missing data, thereby embedding the original model into a complete-data framework $(\mathcal{X}_i, W_i)$. Under this augmented representation, the likelihood assumes a simpler structure than the marginal (observed-data) likelihood, which generally involves integrating out the latent variable W and is therefore analytically intractable. Instead of attempting to maximize the observed-data log-likelihood directly, the algorithm proceeds iteratively by maximizing the expected value of the complete-data log-likelihood.

More precisely, at iteration k, the E-step consists of computing the conditional expectation

$$Q(\vartheta \mid \vartheta^{(k)}) = \mathbb{E}\left[\log f(\mathcal{X}_i, W_i \mid \vartheta) \,\middle|\, \mathcal{X}_i, \vartheta^{(k)}\right],$$

where the expectation is taken with respect to the conditional density

$$f(W_i \mid \mathcal{X}_i, \vartheta^{(k)}).$$

In practical terms, this step requires the evaluation of specific conditional moments of the latent variable W_i given the observed matrix $\mathcal{X}_i$ and the current parameter estimates $\vartheta^{(k)}$. These quantities summarize all the necessary information from the missing component and allow the incomplete-data problem to be replaced by a sequence of tractable optimization steps.

In the subsequent CM-steps (Conditional Maximization steps), rather than maximizing $Q(\vartheta \mid \vartheta^{(k)})$ jointly with respect to all components of ϑ, the ECM algorithm partitions the parameter vector into blocks and performs a sequence of conditional maximizations. At each CM-step, one block of parameters is updated by maximizing $Q(\vartheta \mid \vartheta^{(k)})$ while the remaining blocks are held fixed at their most recently updated values. This blockwise strategy typically simplifies the optimization problem and may lead to closed-form updating equations, or at least to lower-dimensional optimization tasks, while still preserving the monotone ascent property of the EM algorithm.

The E- and CM-steps are alternated until convergence. Convergence is commonly assessed through the relative increment of the observed-data log-likelihood, the norm of successive parameter differences, or both. Under standard regularity conditions, the ECM algorithm generates a sequence of parameter estimates with non-decreasing likelihood values and converges to a stationary point of the observed-data log-likelihood.

For clarity, we explicitly provide the full expression of the E-step function $Q_i(\vartheta \mid \vartheta^{(k)})$ used for inference. This formulation allows us to examine each parameter individually, which is the key distinction between the EM and ECM algorithms. To do so, we shall convention that:

$$\widehat{w}_i^{(k+1)} = \mathbb{E}_W(W \mid \mathcal{X}_i, \vartheta^{(k)}),$$
$$\widehat{w}_i^{2(k+1)} = \mathbb{E}_W(W^2 \mid \mathcal{X}_i, \vartheta^{(k)}).$$

Since $W \mid \boldsymbol{\mathcal{X}}$ follows a truncated normal, such expressions are well known. Suppose that we have $n \in \mathbb{N}_{>0}$ samples at our disposal, that is to say, our random sample is constituted by the set of observations $\{\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_n\}$. To apply the expectation step, we need to determine the individual log-likelihood of the joint distribution of the pair $(\boldsymbol{\mathcal{X}}, W)$ for each $\mathcal{X}_i$ where $i \in \{1, 2, \ldots, n\}$. Precisely, if we convention that $\boldsymbol{\mathcal{Z}} = \boldsymbol{\mathcal{X}} - \mathbf{M}$, it can be expressed as:

$$\ell_{ic}(\vartheta) \propto -\frac{q}{2}\log|\Sigma| - \frac{p}{2}\log|\Psi| - \frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}(\mathcal{Z}_i - w\mathbf{A})\Psi^{-1}(\mathcal{Z}_i - w\mathbf{A})^\top\right)$$

With the purpose of applying the ECM algorithm, the following result is crucial.

Proposition 27. Given $\mathbf{u} \in \mathbb{R}^{pq}\backslash\{0\}$ and $\mathbf{v} \in \mathbb{R}^{pq}\backslash\{0\}$, the symmetric matrix $\mathbf{M}$ defined by the expression $\mathbf{M} := \mathbf{u}\mathbf{u}^\top - a(\mathbf{u}\mathbf{v}^\top + \mathbf{v}\mathbf{u}^\top) + b\mathbf{v}\mathbf{v}^\top$ is positive definite if, and only if, $b > a^2$.

Proof. Indeed, let $\mathbf{x} \in \mathbb{R}^{pq}$. Consequently, the quadratic form $\mathbf{x}'\mathbf{M}\mathbf{x}$ corresponds to:

$$\begin{aligned}
\mathbf{x}^\top\mathbf{M}\mathbf{x} &= \mathbf{x}^\top(\mathbf{u}\mathbf{u}^\top - a(\mathbf{u}\mathbf{v}^\top + \mathbf{v}\mathbf{u}^\top) + b\mathbf{v}\mathbf{v}^\top)\mathbf{x} \\
&= \mathbf{x}^\top\mathbf{u}\mathbf{u}^\top\mathbf{x} - a(\mathbf{x}^\top\mathbf{u}\mathbf{v}^\top\mathbf{x} + \mathbf{x}^\top\mathbf{v}\mathbf{u}^\top\mathbf{x}) + b\mathbf{x}^\top\mathbf{v}\mathbf{v}^\top\mathbf{x} \\
&= \mathbf{x}^\top\mathbf{u}(\mathbf{x}^\top\mathbf{u})^\top - a(\mathbf{x}^\top\mathbf{u}(\mathbf{x}^\top\mathbf{v})^\top + \mathbf{x}^\top\mathbf{v}(\mathbf{x}^\top\mathbf{u})^\top) + b\mathbf{x}^\top\mathbf{v}(\mathbf{x}^\top\mathbf{v}) \\
&= (\mathbf{x}^\top\mathbf{u})^2 - 2a(\mathbf{x}^\top\mathbf{u})(\mathbf{x}^\top\mathbf{v}) + b(\mathbf{x}^\top\mathbf{v})^2
\end{aligned}$$

Therefore, if we set up that $u := \mathbf{x}^\top\mathbf{u}$ and $v := \mathbf{x}^\top\mathbf{v}$, then it results that:

$$\mathbf{x}^\top\mathbf{M}\mathbf{x} = u^2 - 2auv + bv^2.$$

As it can be seen, this is a quadratic form in u and v, whence we may express it as:

$$Q(u,v) = \begin{bmatrix} u & v \end{bmatrix} \begin{bmatrix} 1 & -a \\ -a & b \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix}$$

Thus the positive definiteness of $\mathbf{M}$ reduces to whether the matrix:

$$\mathbf{P} = \begin{bmatrix} 1 & -a \\ -a & b \end{bmatrix}$$

is positive definite. Recall that a 2×2 symmetric matrix is positive definite if and only if all of its leading principal minors are strictly positive. In the present case, the first leading principal minor corresponds to the top-left entry, which equals 1 and is clearly positive. The second leading principal minor is the determinant of $\mathbf{P}$, given by $\det(\mathbf{P}) = b - a^2$. Therefore, $\mathbf{P}$ is positive definite precisely when $b - a^2 > 0$, that is, when $b > a^2$. Consequently, the quadratic form $\mathbf{x}^\top\mathbf{M}\mathbf{x}$ is strictly positive for every nonzero $\mathbf{x}$, which establishes the positive definiteness of $\mathbf{M}$ and completes the proof. □

E-step: Based on the previous definitions of $\widehat{w}_i^{(k+1)}$ and $\widehat{w}_i^{2(k+1)}$, it may be concluded that:

$$Q_i(\vartheta \mid \vartheta^{(k)}) \propto -\frac{q}{2}\log|\Sigma| - \frac{p}{2}\log|\Psi| - \frac{1}{2}\left[\operatorname{tr}(\Sigma^{-1}\mathcal{Z}_i\Psi^{-1}\mathcal{Z}_i^\top) - \widehat{w}_i^{(k+1)}\operatorname{tr}(\Sigma^{-1}\mathcal{Z}_i\Psi^{-1}\mathbf{A}^\top)\right] - \frac{1}{2}\left[\widehat{w}_i^{2(k+1)}\operatorname{tr}(\Sigma^{-1}\mathbf{A}\Psi^{-1}\mathbf{A}^\top) - \widehat{w}_i^{(k+1)}\operatorname{tr}(\Sigma^{-1}\mathbf{A}\Psi^{-1}\mathcal{Z}_i^\top)\right]$$

Lemma 12 (Concavity of Q_i with respect to $\mathbf{M}$)**.** Let $\Sigma \in \mathbb{S}_{++}^p$ and $\Psi \in \mathbb{S}_{++}^q$ be fixed, and let $\mathbf{A} \in \mathbb{R}^{p\times q}$ be fixed. For a given observation $\mathcal{X}_i \in \mathbb{R}^{p\times q}$ define

$$\mathcal{Z}_i(\mathbf{M}) := \mathcal{X}_i - \mathbf{M}, \qquad \mathbf{M} \in \mathbb{R}^{p\times q}.$$

Fix also scalars $c_1 \in \mathbb{R}$ and $c_2 \in \mathbb{R}$ (in the ECM context, $c_1 = \widehat{w}_i^{(k+1)}$ and $c_2 = \widehat{w}_i^{2(k+1)}$), which are treated as known constants at the current iteration of the algorithm. Consider the function

$$Q_i(\mathbf{M}) \propto -\frac{1}{2}\Big[\operatorname{tr}\big(\Sigma^{-1}\mathcal{Z}_i(\mathbf{M})\,\Psi^{-1}\mathcal{Z}_i(\mathbf{M})^\top\big) - c_1\operatorname{tr}\big(\Sigma^{-1}\mathcal{Z}_i(\mathbf{M})\,\Psi^{-1}\mathbf{A}^\top\big)\Big] - \frac{1}{2}\Big[c_2\operatorname{tr}\big(\Sigma^{-1}\mathbf{A}\Psi^{-1}\mathbf{A}^\top\big) - c_1\operatorname{tr}\big(\Sigma^{-1}\mathbf{A}\Psi^{-1}\mathcal{Z}_i(\mathbf{M})^\top\big)\Big],$$

where the proportionality sign means that terms not depending on $\mathbf{M}$ are omitted. Then $Q_i(\mathbf{M})$ is a concave function of $\mathbf{M}$ on $\mathbb{R}^{p\times q}$ (indeed, it is a concave quadratic function).

Proof. Since Σ, Ψ and $\mathbf{A}$ are fixed, the dependence on $\mathbf{M}$ occurs only through $\mathcal{Z}_i(\mathbf{M}) = \mathcal{X}_i - \mathbf{M}$. Using the cyclic property of the trace and the identity $\operatorname{tr}(\mathbf{B}) = \operatorname{tr}(\mathbf{B}^\top)$, we note that

$$\operatorname{tr}\big(\Sigma^{-1}\mathbf{A}\Psi^{-1}\mathcal{Z}_i(\mathbf{M})^\top\big) = \operatorname{tr}\big(\big(\Sigma^{-1}\mathbf{A}\Psi^{-1}\mathcal{Z}_i(\mathbf{M})^\top\big)^\top\big) = \operatorname{tr}\big(\Sigma^{-1}\mathcal{Z}_i(\mathbf{M})\Psi^{-1}\mathbf{A}^\top\big).$$

Hence, the two linear trace terms involving $\mathcal{Z}_i(\mathbf{M})$ coincide by virtue of the cyclic property of the trace and the symmetry of the scalar expression, which allows us to combine them. Consequently, the part of Q_i that depends on $\mathbf{M}$ can be rewritten as

$$Q_i(\mathbf{M}) = \text{constant} - \frac{1}{2}\operatorname{tr}\big(\Sigma^{-1}\mathcal{Z}_i(\mathbf{M})\Psi^{-1}\mathcal{Z}_i(\mathbf{M})^\top\big) + c_1\operatorname{tr}\big(\Sigma^{-1}\mathcal{Z}_i(\mathbf{M})\Psi^{-1}\mathbf{A}^\top\big),$$

where "constant" collects all terms not depending on $\mathbf{M}$ (in particular, the term involving c_2).

To investigate concavity, consider the real-valued function of a scalar variable t defined by $\phi(t) := Q_i(\mathbf{M} + t\mathbf{H})$, where $\mathbf{H} \in \mathbb{R}^{p\times q}$ is an arbitrary direction matrix. This one-dimensional restriction allows us to analyze the curvature of Q_i along straight lines in the parameter space. Since

$$\mathcal{Z}_i(\mathbf{M} + t\mathbf{H}) = \mathcal{X}_i - (\mathbf{M} + t\mathbf{H}) = \mathcal{Z}_i(\mathbf{M}) - t\mathbf{H},$$

we obtain

$$\phi(t) \propto -\frac{1}{2}\operatorname{tr}\Big(\Sigma^{-1}\big(\mathcal{Z}_i(\mathbf{M}) - t\mathbf{H}\big)\Psi^{-1}\big(\mathcal{Z}_i(\mathbf{M}) - t\mathbf{H}\big)^\top\Big) + c_1\operatorname{tr}\Big(\Sigma^{-1}\big(\mathcal{Z}_i(\mathbf{M}) - t\mathbf{H}\big)\Psi^{-1}\mathbf{A}^\top\Big).$$

Differentiating twice with respect to t (using linearity of the trace) yields

$$\phi''(t) = -\operatorname{tr}\big(\Sigma^{-1}\mathbf{H}\Psi^{-1}\mathbf{H}^\top\big),$$

because the second derivative of the linear trace term is zero, while the quadratic term contributes the constant negative quantity above.

Finally, since $\Sigma^{-1} \succ \mathbf{0}$ and $\Psi^{-1} \succ \mathbf{0}$, we have

$$\operatorname{tr}\big(\Sigma^{-1}\mathbf{H}\Psi^{-1}\mathbf{H}^\top\big) = \big\|\Sigma^{-1/2}\mathbf{H}\Psi^{-1/2}\big\|_F^2 \geq 0,$$

with equality if and only if $\mathbf{H} = \mathbf{0}$. Therefore, $\phi''(t) \leq 0$ for all t, which shows that ϕ is concave in t along every line $\mathbf{M} + t\mathbf{H}$. Hence $Q_i(\mathbf{M})$ is concave in $\mathbf{M}$ on $\mathbb{R}^{p\times q}$, concluding the proof. □

CM-step 1: Based on the expression obtained for $Q_i(\vartheta \mid \vartheta^{(k)})$, we now derive the updating formula for the location parameter $\mathbf{M}$. Recall from Lemma 12 that Q_i is concave with respect to $\mathbf{M}$ when the remaining parameters are held fixed. Therefore, any stationary point obtained by solving the first-order condition corresponds to the global maximizer in $\mathbf{M}$.

To compute the derivative of $Q_i(\vartheta \mid \vartheta^{(k)})$ with respect to $\mathbf{M}$, we make use of standard matrix differentiation rules, which can be found in Petersen, Pedersen *et al.* (2008), for instance. Since $\mathcal{Z}_i = \mathcal{X}_i - \mathbf{M}$, differentiation with respect to $\mathbf{M}$ requires the application of the chain rule, taking into account that $\partial \mathcal{Z}_i / \partial \mathbf{M} = -\mathbf{I}$. Restricting attention to the $\mathbf{M}$-dependent part of Q_i, we obtain

$$\frac{\partial Q_i}{\partial \mathbf{M}} = \Sigma^{-1}\mathcal{Z}_i\Psi^{-1} - \widehat{w}_i^{(k+1)}\Sigma^{-1}\mathbf{A}\Psi^{-1}.$$

Setting this gradient equal to zero and multiplying on the left by Σ and on the right by Ψ yields the first-order condition $\mathcal{Z}_i = \widehat{w}_i^{(k+1)}\mathbf{A}$. Recalling that $\mathcal{Z}_i = \mathcal{X}_i - \mathbf{M}$, we obtain $\mathbf{M} = \mathcal{X}_i - \widehat{w}_i^{(k+1)}\mathbf{A}$. Averaging over $i \in \{1, 2, \ldots, n\}$, gives the updating formula

$$\widehat{\mathbf{M}}^{(k+1)} = \frac{1}{n}\sum_{i=1}^{n}\left(\mathcal{X}_i - \widehat{w}_i^{(k+1)}\widehat{\mathbf{A}}^{(k)}\right),$$

which completes the first conditional maximization step.

Lemma 13 (Concavity of Q_i with respect to $\mathbf{A}$)**.** Let $\Sigma \in \mathbb{S}_{++}^p$ and $\Psi \in \mathbb{S}_{++}^q$ be fixed, and let $\mathcal{Z}_i \in \mathbb{R}^{p\times q}$ be fixed. Fix also scalars $c_1 \in \mathbb{R}$ and $c_2 \in \mathbb{R}$ with $c_2 \geq 0$ (in the ECM context, $c_1 = \widehat{w}_i^{(k+1)}$ and $c_2 = \widehat{w}_i^{2(k+1)}$), which are treated as known constants at the current iteration of the algorithm. Consider the function of $\mathbf{A} \in \mathbb{R}^{p\times q}$ defined (up to additive constants not depending on $\mathbf{A}$) by

$$Q_i(\mathbf{A}) \propto -\frac{1}{2}\Big[c_2 \operatorname{tr}\big(\Sigma^{-1}\mathbf{A}\Psi^{-1}\mathbf{A}^\top\big) - c_1 \operatorname{tr}\big(\Sigma^{-1}\mathcal{Z}_i\Psi^{-1}\mathbf{A}^\top\big)\Big] - \frac{1}{2}\Big[-c_1 \operatorname{tr}\big(\Sigma^{-1}\mathbf{A}\Psi^{-1}\mathcal{Z}_i^\top\big)\Big],$$

where the proportionality sign means that terms not depending on $\mathbf{A}$ are omitted. Then $Q_i(\mathbf{A})$ is a concave function of $\mathbf{A}$ on $\mathbb{R}^{p\times q}$ (indeed, it is a concave quadratic function). Moreover, if $c_2 > 0$, then $Q_i(\mathbf{A})$ is strictly concave in $\mathbf{A}$.

Proof. The proof of concavity with respect to $\mathbf{A}$ follows exactly the same line of reasoning as the previous proof for $\mathbf{M}$. In both cases, the argument begins by isolating the part of Q_i that effectively depends on the parameter of interest and combining the symmetric linear trace terms using the cyclic property of the trace. Concavity is then established by restricting the function to a one–dimensional path of the form $\mathbf{A}+t\mathbf{H}$ and computing the second derivative with respect to t. In each case, the linear trace terms vanish after the second differentiation, while the quadratic trace term yields a constant curvature term of the form $-\operatorname{tr}(\Sigma^{-1}\mathbf{H}\Psi^{-1}\mathbf{H}^{\top})$ (up to a nonnegative multiplicative constant). Positive definiteness of Σ^{-1} and Ψ^{-1} ensures that this quantity is nonnegative, implying that the second derivative is nonpositive along every direction. Therefore, the concavity argument for $\mathbf{A}$ is structurally identical to that established for $\mathbf{M}$. □

CM-step 2: We now derive the updating formula for the skewness parameter $\mathbf{A}$. From Lemma 13, the function Q_i is concave with respect to $\mathbf{A}$ when the remaining parameters are kept fixed (and strictly concave whenever $\widehat{w}_i^{2(k+1)} > 0$). Consequently, any stationary point obtained by solving the first-order condition corresponds to the global maximizer in $\mathbf{A}$.

To compute the derivative of $Q_i(\vartheta \mid \vartheta^{(k)})$ with respect to $\mathbf{A}$, we again apply standard matrix differentiation rules (see, e.g., Petersen, Pedersen *et al.* (2008)). Restricting attention to the $\mathbf{A}$-dependent part of Q_i, we obtain

$$\frac{\partial Q_i}{\partial \mathbf{A}} = \widehat{w}_i^{(k+1)}\Sigma^{-1}\mathcal{Z}_i\Psi^{-1} - \widehat{w}_i^{2(k+1)}\Sigma^{-1}\mathbf{A}\Psi^{-1}.$$

Summing the gradient contributions over $i \in \{1,2,\ldots,n\}$ and recalling that the CM-step consists of maximizing the aggregated objective function Q with respect to $\mathbf{A}$ while keeping the remaining parameters fixed, we impose the corresponding first-order optimality condition

$$\sum_{i=1}^{n} \frac{\partial Q_i}{\partial \mathbf{A}} = \mathbf{0}.$$

Substituting the expression for the gradient and grouping the terms involving $\mathbf{A}$, we obtain

$$\Sigma^{-1}\left(\sum_{i=1}^{n} \widehat{w}_i^{(k+1)}\mathcal{Z}_i - \sum_{i=1}^{n} \widehat{w}_i^{2(k+1)}\mathbf{A}\right)\Psi^{-1} = \mathbf{0}.$$

Since Σ^{-1} and Ψ^{-1} are symmetric positive definite matrices (and therefore necessarily invertible), we may multiply on the left by Σ and on the right by Ψ without altering the set of solutions:

$$\sum_{i=1}^{n} \widehat{w}_i^{(k+1)}\mathcal{Z}_i = \left(\sum_{i=1}^{n} \widehat{w}_i^{2(k+1)}\right)\mathbf{A}.$$

Solving the previous matrix equation explicitly for $\mathbf{A}$, and recalling that $\mathcal{Z}_i = \mathcal{X}_i - \widehat{\mathbf{M}}^{(k)}$, so that the residuals are expressed in terms of the current iterate of the location parameter, we obtain

$$\widehat{\mathbf{A}}^{(k+1)} = \left(\sum_{i=1}^{n} \widehat{w}_i^{2(k+1)}\right)^{-1} \sum_{i=1}^{n} \left(\mathcal{X}_i - \widehat{\mathbf{M}}^{(k)}\right)\widehat{w}_i^{(k+1)}.$$

Whenever the matrices $\mathbf{P}$, $\mathbf{Q}$, $\mathbf{R}$ and $\mathbf{S}$ have conformable dimensions so that all products are well defined, it holds that $\mathrm{tr}(\mathbf{PQRS}) = \mathrm{vec}(\mathbf{S}^\top)^\top(\mathbf{R}^\top \otimes \mathbf{P})\mathrm{vec}(\mathbf{Q})$, see, for instance, Petersen, Pedersen *et al.* (2008). In the case of interest, we take $\mathbf{P} := \Sigma^{-1}$, $\mathbf{Q} := \boldsymbol{\mathcal{Z}} - w\mathbf{A}$, $\mathbf{R} := \Psi^{-1}$ and $\mathbf{S} := (\boldsymbol{\mathcal{Z}} - w\mathbf{A})^\top$.Using also the cyclic invariance property of the trace operator, namely $\mathrm{tr}(\mathbf{PQR}) = \mathrm{tr}(\mathbf{RPQ}) = \mathrm{tr}(\mathbf{QRP})$ whenever the products are well defined, we obtain:

$$
\begin{aligned}
\ell_{ic}(\vartheta) &\propto -\frac{q}{2}\log|\Sigma| - \frac{p}{2}\log|\Psi| - \frac{1}{2}\mathrm{tr}\left[\mathrm{vec}(\mathcal{Z}_i - w\mathbf{A})^\top(\Psi^{-1}\otimes\Sigma^{-1})\mathrm{vec}(\mathcal{Z}_i - w\mathbf{A})\right] \\
&= -\frac{q}{2}\log|\Sigma| - \frac{p}{2}\log|\Psi| - \frac{1}{2}\mathrm{tr}\left[(\Psi\otimes\Sigma)^{-1}\mathrm{vec}(\mathcal{Z}_i - w\mathbf{A})\mathrm{vec}(\mathcal{Z}_i - w\mathbf{A})^\top\right] \\
&= -\frac{q}{2}\log|\Sigma| - \frac{p}{2}\log|\Psi| - \frac{1}{2}\mathrm{tr}\left[(\Psi\otimes\Sigma)^{-1}(\mathrm{vec}(\mathcal{Z}_i)\mathrm{vec}(\mathcal{Z}_i)^\top - w\mathrm{vec}(\mathcal{Z}_i)\mathrm{vec}(\mathbf{A})^\top)\right] \\
&\quad - \frac{1}{2}\mathrm{tr}\left[(\Psi\otimes\Sigma)^{-1}(-w\mathrm{vec}(\mathbf{A})\mathrm{vec}(\mathcal{Z}_i)^\top + w^2\mathrm{vec}(\mathbf{A})\mathrm{vec}(\mathbf{A})^\top)\right].
\end{aligned}
$$

In order to better estimate the parameters Σ and Ψ, we are going to handle it through the use of the last formula for $\ell_{ic}(\vartheta)$. To do so, we shall use Proposition 27. Based on such result, we stress that the matrix presented below is symmetric and positive definite, as $\mathrm{Var}(W) = \mathbb{E}(W^2) - \mathbb{E}(W)^2 \geq 0$.

$$
\widehat{\Delta}_i^{(k)} := \mathrm{vec}(\mathcal{Z}_i)\mathrm{vec}(\mathcal{Z}_i)^\top - \widehat{w}_i^{(k+1)}\left(\mathrm{vec}(\mathcal{Z}_i)\mathrm{vec}(\mathbf{A})^\top + \mathrm{vec}(\mathbf{A})\mathrm{vec}(\mathcal{Z}_i)^\top\right) + \widehat{w}_i^{2(k+1)}\mathrm{vec}(\mathbf{A})\mathrm{vec}(\mathbf{A})^\top
$$

Hence, the estimator $\widehat{\Delta}_i^{(k)}$ admits a Cholesky decomposition, since it is symmetric and positive definite under the stated conditions. This guarantees the existence and uniqueness of a lower triangular factor, yielding a numerically stable and computationally efficient representation that preserves the structural properties of the covariance components throughout the ECM iterations. In particular, this decomposition substantially simplifies subsequent matrix operations within the update steps for each observation index $i \in \{1, 2, \ldots, n\}$. Consequently, we may write

$$
\widehat{\Delta}_i^{(k)} = \widehat{\mathbf{L}}_i^{(k)}(\widehat{\mathbf{L}}_i^{(k)})^\top,
$$

where $\widehat{\mathbf{L}}_i^{(k)}$ denotes the corresponding lower triangular Cholesky factor.

If we denote by $\widehat{\mathbf{B}}_{ij}^{(k)}$ the $p \times q$ matrix such that $\mathrm{vec}(\widehat{\mathbf{B}}_{ij}^{(k)}) := \widehat{\mathbf{L}}_{ij}^{(k)}$, where $\widehat{\mathbf{L}}_{ij}^{(k)}$ is the j-th column of the $pq \times pq$ lower triangular matrix $\widehat{\mathbf{L}}_i^{(k)}$, it results that:

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

CM-step 3: Taking into account equivalent formula for the E-step function as well as the results presented in Petersen, Pedersen *et al.* (2008), it may be deduced that

$$\begin{aligned}\frac{\partial}{\partial\Psi}Q(\vartheta|\vartheta^{(k)}) &= -\frac{np}{2}\frac{\partial}{\partial\Psi}\log|\Psi| - \frac{1}{2}\frac{\partial}{\partial\Psi}\sum_{i=1}^{n}\sum_{j=1}^{pq}\operatorname{tr}\left[\Sigma^{-1}\widehat{\mathbf{B}}_{ij}^{(k)}\Psi^{-1}(\widehat{\mathbf{B}}_{ij}^{(k)})^{\top}\right]\\ &= -\frac{np}{2}\frac{1}{|\Psi|}\frac{\partial}{\partial\Psi}|\Psi| - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{pq}\frac{\partial}{\partial\Psi}\operatorname{tr}\left[\Sigma^{-1}\widehat{\mathbf{B}}_{ij}^{(k)}\Psi^{-1}(\widehat{\mathbf{B}}_{ij}^{(k)})^{\top}\right]\\ &= -\frac{np}{2}\Psi^{-1} + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{pq}(\Psi^{-1}(\widehat{\mathbf{B}}_{ij}^{(k)})^{\top}\Sigma^{-1}\widehat{\mathbf{B}}_{ij}^{(k)}\Psi^{-1})\\ &= -\frac{np}{2}\Psi^{-1} + \Psi^{-1}\left(\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{pq}(\widehat{\mathbf{B}}_{ij}^{(k)})^{\top}\Sigma^{-1}\widehat{\mathbf{B}}_{ij}^{(k)}\right)\Psi^{-1}\end{aligned}$$

From whence it results that:

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

Notice that such update corresponds to the global maximum of the E-step function Q due to the result presented in Lemma 6. The analogous version of it related to the parameter Σ also applies to the next CM-step, which implies that the updated parameter corresponds to a global maximum as well.

CM-step 4: Similarly to the procedure applied to estimate the parameter Ψ, and following the same differentiation and first-order optimality arguments presented above, the following relation holds:

$$\begin{aligned}\frac{\partial}{\partial\Sigma}Q(\vartheta|\vartheta^{(k)}) &= -\frac{nq}{2}\frac{\partial}{\partial\Sigma}\log|\Sigma| - \frac{1}{2}\frac{\partial}{\partial\Sigma}\sum_{i=1}^{n}\sum_{j=1}^{pq}\operatorname{tr}\left[\Sigma^{-1}\widehat{\mathbf{B}}_{ij}^{(k)}\Psi^{-1}(\widehat{\mathbf{B}}_{ij}^{(k)})^{\top}\right]\\ &= -\frac{nq}{2}\frac{1}{|\Sigma|}\frac{\partial}{\partial\Sigma}|\Sigma| - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{pq}\frac{\partial}{\partial\Sigma}\operatorname{tr}\left[\Sigma^{-1}\widehat{\mathbf{B}}_{ij}^{(k)}\Psi^{-1}(\widehat{\mathbf{B}}_{ij}^{(k)})^{\top}\right]\\ &= -\frac{nq}{2}\Sigma^{-1} + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{pq}(\Sigma^{-1}\widehat{\mathbf{B}}_{ij}^{(k)}\Psi^{-1}(\widehat{\mathbf{B}}_{ij}^{(k)})^{\top}\Sigma^{-1})\\ &= -\frac{nq}{2}\Sigma^{-1} + \Sigma^{-1}\left(\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{pq}\widehat{\mathbf{B}}_{ij}^{(k)}\Psi^{-1}(\widehat{\mathbf{B}}_{ij}^{(k)})^{\top}\right)\Sigma^{-1}\end{aligned}$$

From whence we may claim that:

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

Remark. To monitor convergence, we employ the Aitken acceleration criterion. Let

$$\ell^{(k)} = \ell\left(\widehat{\mathbf{M}}^{(k)}, \widehat{\mathbf{A}}^{(k)}, \widehat{\Sigma}^{(k)}, \widehat{\Psi}^{(k)}\right)$$

denote the log-likelihood evaluated at the parameter estimates obtained at iteration k. The Aitken acceleration at iteration k is defined as

$$a^{(k)} = \frac{\ell^{(k)} - \ell^{(k-1)}}{\ell^{(k-1)} - \ell^{(k-2)}}.$$

An extrapolated estimate of the asymptotic log-likelihood is then given by

$$\ell_{\infty}^{(k)} = \ell^{(k-1)} + \frac{\ell^{(k)} - \ell^{(k-1)}}{1 - a^{(k)}}.$$

Accordingly, the convergence measure is computed as

$$\text{crit}^{(k)} = \frac{\ell^{(k)} - \ell^{(k-1)}}{1 - a^{(k)}},$$

and the algorithm is terminated when $\text{crit}^{(k)}$ is smaller than a prescribed tolerance.

Summary of Computational Results

In summary, this section establishes all the conditional expectations, auxiliary matrices, and updating equations necessary to implement the ECM algorithm effectively within the MVSN setting. A key component of the procedure is the matrix $\widehat{\Delta}_i^{(k)}$, which, under the stated assumptions, is symmetric and positive definite and therefore admits a Cholesky factorization. Adopting this decomposition is not merely theoretical; it ensures a unique lower triangular representation and enhances numerical stability while reducing computational cost. In addition, the Cholesky structure streamlines the matrix calculations required at each iteration, preserving the covariance structure and facilitating efficient updates for every observation index $i \in \{1, 2, \ldots, n\}$.

Altogether, the expressions derived in this chapter constitute the essential computational machinery driving the algorithm, ensuring that each expectation and conditional maximization cycle is executed in a coherent, stable, and analytically tractable manner.

CHAPTER

4

MVSN: SIMULATIONS AND APPLICATION

In this chapter, we examine the matrix-variate skew-normal distribution in depth, assessing its practical behavior through both controlled simulation experiments and an application to real data. The simulation study is specifically structured to investigate the performance of the ECM algorithm when estimating the parameters of the model. Our analysis emphasizes several key aspects of the algorithm's behavior, including the accuracy of the estimates, their consistency as the sample size increases, and the overall numerical stability of the procedure. By systematically altering the sample size while keeping the number of replications fixed, we are able to observe how reliably the algorithm recovers the true underlying parameters under different data conditions. This approach also allows us to detect potential weaknesses or limitations of the estimator in finite-sample scenarios, providing insight into when the method performs well and when caution may be needed. In addition to the simulation study, we also apply the model to a real dataset to illustrate how it performs with practical matrix-valued observations. This empirical analysis showcases the model's capacity to represent intricate dependence structures and asymmetric patterns, while also emphasizing the interpretability of its parameter estimates. Combined, the simulated and real-data investigations offer a well-rounded evaluation of the estimation strategy and demonstrate its suitability for high-dimensional settings.

4.1 Simulation studies

To evaluate the proposed MVSN distribution, we conducted two complementary simulation studies. The first examined how effectively the ECM algorithm recovers the true parameter values under controlled settings, allowing us to assess the accuracy and reliability of the estimation procedure. The second compared the MVSN model with the classical MVN by fitting both to data generated from a matrix normal variance–mean mixture with latent variable $W \sim \text{Exp}(1)$, following Gallaugher and McNicholas (2019), providing a benchmark for assessing whether the added flexibility of the MVSN framework improves performance.

Model performance across different sample sizes was evaluated using the Bayesian Information Criterion (BIC). In all simulations, we used the same values for the location matrix $\mathbf{M}$, the skewness matrix $\mathbf{A}$, and the covariance matrices Σ and Ψ to ensure consistency.

This setup was designed to isolate the effects of sample size and skewness from those of varying parameter configurations. The specific values for the parameter matrices used to simulate the data are detailed below. More precisely, the location and skewness matrices are correspondingly given by:

$$\mathbf{M} = \begin{bmatrix} 0.5 & 1.0 & 0.5 & 1.0 \\ 1.0 & 1.0 & 0.5 & 0.5 \\ 1.5 & 1.5 & 1.0 & 1.5 \end{bmatrix}, \quad \mathbf{A} = \begin{bmatrix} 0.18 & 1.62 & 1.16 & 0.02 \\ 0.25 & 0.73 & 0.83 & 0.13 \\ 0.71 & 1.42 & 0.22 & 0.81 \end{bmatrix}.$$

The scale matrices Σ and Ψ are defined as:

$$\Sigma = \begin{bmatrix} 0.100 & 0.040 & 0.016 \\ 0.040 & 0.100 & 0.040 \\ 0.016 & 0.040 & 0.100 \end{bmatrix}, \quad \Psi = \begin{bmatrix} 2.151657 & 1.721326 & 1.377061 & 1.101649 \\ 1.721326 & 2.151657 & 1.721326 & 1.377061 \\ 1.377061 & 1.721326 & 2.151657 & 1.721326 \\ 1.101649 & 1.377061 & 1.721326 & 2.151657 \end{bmatrix}.$$

It is important to note that $|\Psi| = 1$, which agrees with the previously mentioned constraint used to avoid identifiability problems. The remainder of this chapter is devoted to the study of the MVSN distribution and its comparison with the MVN model.

4.1.1 Parameter Recovery in the MVSN Model

The first study was structured into four groups, each consisting of $s = 200$ independent replications, with every group corresponding to a different sample size: $n = 50$, $n = 100$, $n = 200$, and $n = 400$ data matrices per replication. This design was chosen to systematically evaluate the performance of the ECM algorithm in recovering the true model parameters under varying sample sizes, allowing us to observe how estimation accuracy and stability change as the amount of available data increases. To ensure that the results are fully reproducible and comparable across different runs, a fixed random seed was employed throughout the simulations.

In each replication, data matrices were generated from the stochastic representation of the MVSN model using the fixed parameter values introduced earlier. Parameter estimation was carried out until convergence of the ECM algorithm was attained or until numerical precision was reached, after which the difference between the estimated and true parameter matrices ($\mathbf{M}$, $\mathbf{A}$, Σ, and Ψ) was computed. The initial parameter values are defined as follows. Let $\{\mathcal{X}_1, \ldots, \mathcal{X}_n\}$ be a sample of matrices in $\mathbb{R}^{3\times 4}$. The initial estimate of the location matrix is taken as the sample mean, $\mathbf{M}^{(0)} = \overline{\mathcal{X}_i}$. The initial row and column scale matrices are set equal to the identity matrices, $\Sigma^{(0)} = \mathbf{I}_3$ and $\Psi^{(0)} = \mathbf{I}_4$. Finally, the initial value of the skewness matrix is also taken as the sample mean, $\mathbf{A}^{(0)} = \overline{\mathcal{X}}_i$. The estimation error was quantified using the Frobenius norm, $\|\mathbf{H}\|_{\mathrm{F}}^2 = \mathrm{tr}(\mathbf{H}\mathbf{H}^\top)$, where $\mathbf{H}$ denotes the error matrix.

To summarize the results, we present box plots of the Frobenius norms across all replications, grouped according to sample size, as displayed in Figure 1. These plots offer a clear visual overview of the distribution of estimation errors and illustrate how the accuracy and stability of the parameter estimates generally improve as the sample size increases, providing insight into the effect of sample size on estimation performance.

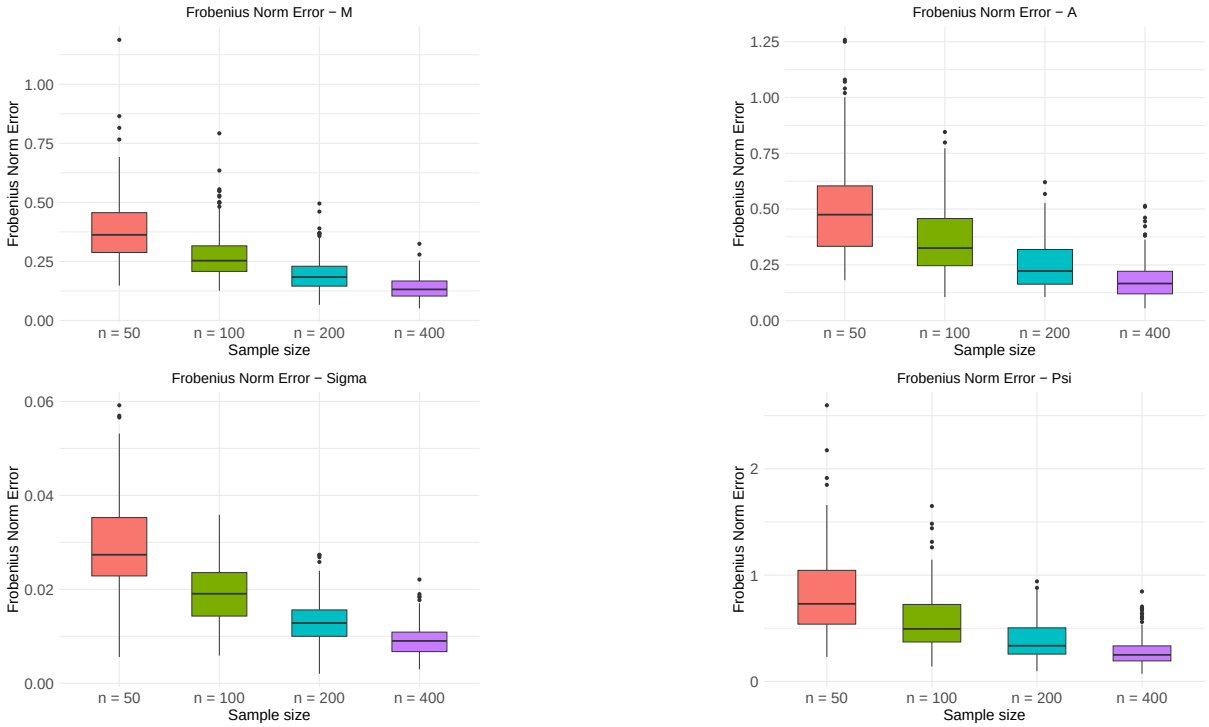


Figure 1 – Boxplots of Frobenius norm errors for the estimated parameter matrices $\mathbf{M}$, $\mathbf{A}$, Σ, and Ψ.

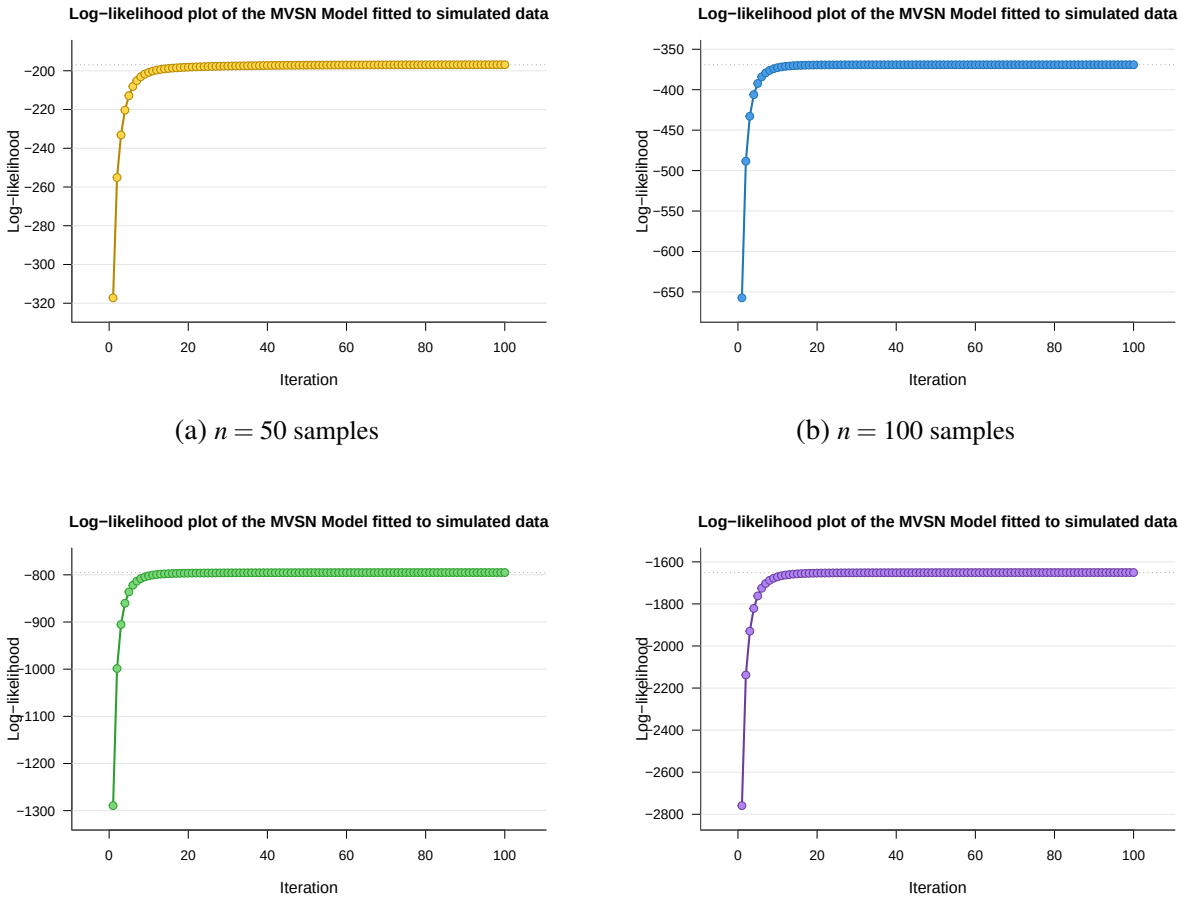
The box-plots show a clear decline in the Frobenius norm error of the estimator of $\mathbf{M}$ as the sample size increases from $n = 50$ to $n = 400$. Both the center and dispersion of the error distributions decrease with n, indicating greater accuracy and stability in larger samples. The steady reduction in the median and the narrowing interquartile range suggest consistency under the proposed model. Larger errors for small n reflect higher variability in low-information settings, which diminishes as the sample grows. Overall, the results provide empirical evidence of convergence and confirm the reliability of the estimation procedure for moderate to large samples, supporting its practical applicability in finite-sample contexts.

A similar pattern is observed for the estimator of $\mathbf{A}$, whose Frobenius norm error decreases steadily as the sample size increases. Both the median error and the dispersion diminish with n, reflecting enhanced estimation accuracy and reduced variability in larger samples. The wider spread seen at smaller sample sizes highlights the greater instability of the skewness matrix estimator when information is limited, an effect that is progressively alleviated as the sample size increases. These findings provide empirical support for the consistency and robustness of the proposed estimation procedure for $\mathbf{A}$.

The Frobenius norm error of the estimator of Σ shows a clear downward trend as the sample size increases. The reduction in both the median error and its variability reflects improved precision and stability in larger samples. Moreover, the consistently small magnitude of the error indicates that Σ is estimated more accurately than the location and skewness parameters, and the observed pattern further supports the consistency of the proposed method.

Finally, the estimator of Ψ exhibits a marked decrease in Frobenius norm error as the sample size increases. Both the median and dispersion of the errors shrink with n, indicating greater accuracy and stability in larger samples. The relatively larger errors for small sample sizes highlight the difficulty of estimating Ψ with limited information, an effect that diminishes as n grows. Overall, the results provide empirical support for the consistency and reliability of the proposed estimator.

In Figure 2, the log-likelihood trajectories exhibit the typical ECM pattern: a sharp initial rise followed by stabilization, showing that most improvement occurs in early iterations. For larger samples, the curves are smoother and more stable, reflecting lower variability and more regular likelihood surfaces. The algorithm quickly stabilizes, generally within 10–15 iterations, using the Aitken acceleration criterion ($\varepsilon = 10^{-15}$) with a maximum of 100 iterations.



(a) $n = 50$ samples

(b) $n = 100$ samples

(c) $n = 200$ samples

(d) $n = 400$ samples

Figure 2 – Log-likelihood trajectories across ECM iterations for different sample sizes.

4.1.2 Comparing the MVSN and MVN Models

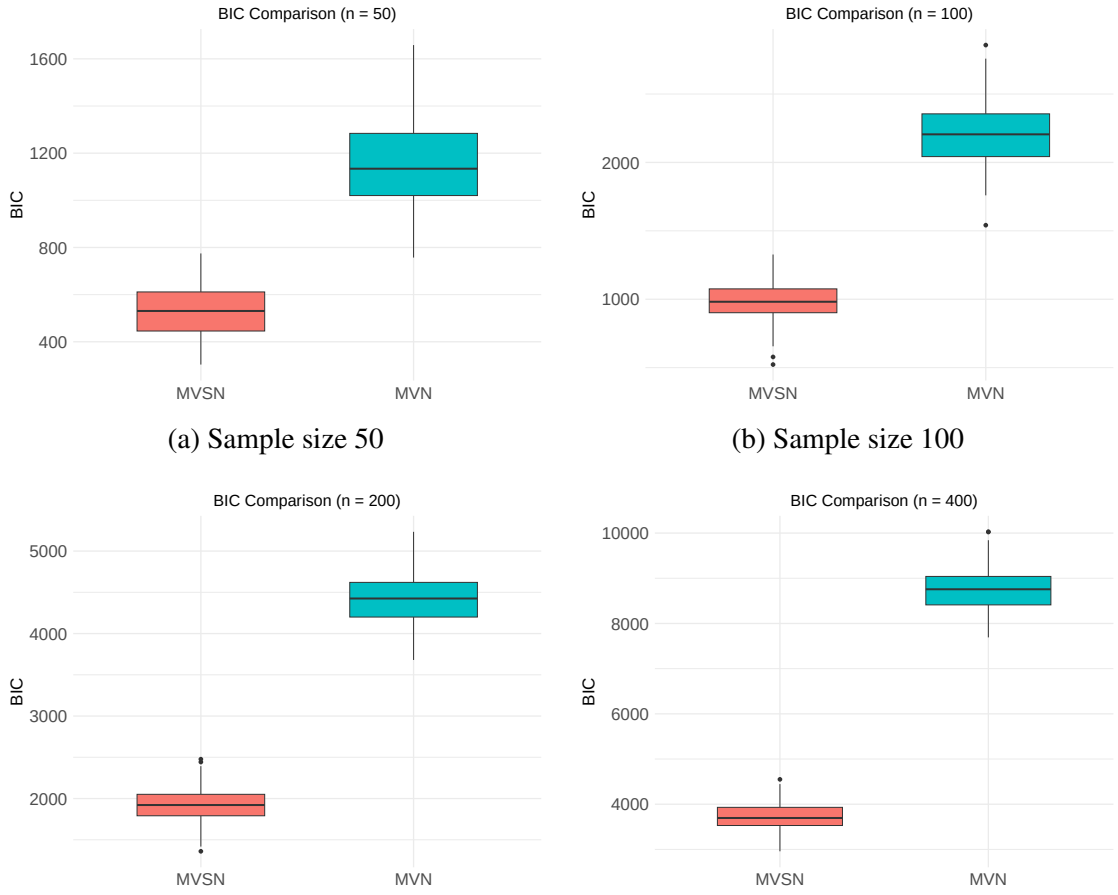
The second simulation study was conducted with the aim of evaluating the benefits of explicitly modeling asymmetry when it occurs in the data-generating process. To achieve this, we generated synthetic datasets following the stochastic representation $\boldsymbol{\mathcal{X}} = \mathbf{M} + W\mathbf{A} + \sqrt{W}\boldsymbol{\mathcal{V}}$, where $\mathbf{M} \in \mathbb{R}^{3\times4}$ is the location matrix, $\mathbf{A} \in \mathbb{R}^{3\times4}$ is the skewness matrix, and $\boldsymbol{\mathcal{V}} \sim \mathcal{N}_{3\times4}(\mathbf{0}, \Sigma, \Psi)$ represents the matrix-normal random component. The parameters $\mathbf{M}$, $\mathbf{A}$, Σ, and Ψ were kept the same as in the previous study to ensure comparability. In this setup, the latent variable was assumed to follow an exponential distribution, $W \sim \text{Exp}(1)$, introducing both scale variability and skewness in $\boldsymbol{\mathcal{X}}$. As mentioned in the reference Gallaugher and McNicholas (2019), this distribution is a particular case of the Matrix Variate Variance-Gamma model. The MVSN and MVN models were then fitted to these simulated datasets to determine whether explicitly incorporating skewness results in a superior fit compared to the symmetric alternative.

Model comparison was carried out using the Bayesian Information Criterion (BIC), computed for each sample size. The purpose of this analysis was to evaluate whether the added flexibility of the MVSN model, particularly in capturing skewness and heteroskedasticity, translates into improved performance relative to the classical MVN distribution. For each of the $s = 200$ replications and every sample size, both models were fitted to the simulated data using their respective estimation procedures — ECM for the MVSN model and ML for the MVN distribution. The BIC was defined as $\text{BIC} := -2\ell(\widehat{\Theta}) + k\log(n)$, where $\ell(\widehat{\Theta})$ is the maximized log-likelihood, k the number of estimated parameters, and n the number of observed matrices per replication. Lower BIC values correspond to a better balance between model fit and complexity.

As briefly noted earlier, in this second experiment we used the same parameter values as in the first simulation study for the location matrix $\mathbf{M}$, the skewness matrix $\mathbf{A}$, the row covariance matrix Σ, and the column covariance matrix Ψ, ensuring a consistent comparison. The results show a clear preference for the MVSN model, which consistently achieves lower BIC values across all sample sizes and conditions, highlighting its improved ability to capture departures from Gaussianity. For clarity, Figure 3 displays the distribution of BIC differences between the MVSN and MVN models for each sample size.

4.2 Application

This dataset consists of quarterly records (Q1–Q4) of Dow-Jones Industrial Common Stocks dividends and the Dow-Jones divisor, from 1920 to 1934 Rappoport and White (1993). The DJ divisor accounts for structural adjustments such as stock splits, ensuring consistency in index computation over time. This historical series has been explored in statistical modeling literature and was also considered by Rezaei, Yousefzadeh and Arellano-Valle (2020) to illustrate the use of matrix variate extended skew-normal distributions in modeling economic data with complex dependence structures.



(a) Sample size 50

(b) Sample size 100

(c) Sample size 200

(d) Sample size 400

Figure 3 – Box plots of the BIC Values for each sample size scenario.

The data are arranged in 4×2 matrices, where each matrix corresponds to one year. Rows represent quarters, and columns contain values for DJ dividends and DJ divisor, respectively. These matrices provide a compact and interpretable format suitable for models that incorporate temporal and cross-variable dependencies.

1920:

	DJ dividends	DJ divisor
Q1	31.97	20.00
Q2	30.00	20.00
Q3	30.00	20.00
Q4	28.75	20.00

1921:

	DJ dividends	DJ divisor
Q1	26.81	20.00
Q2	23.81	20.00
Q3	22.06	20.00
Q4	22.06	20.00

1922:

	DJ dividends	DJ divisor
Q1	21.13	20.00
Q2	19.38	20.00
Q3	19.38	20.00
Q4	20.88	20.00

1923:

	DJ dividends	DJ divisor
Q1	24.19	20.00
Q2	26.94	20.00
Q3	25.94	20.00
Q4	25.94	20.00

1924:

	DJ dividends	DJ divisor
Q1	30.00	20.00
Q2	26.75	19.40
Q3	27.20	19.40
Q4	27.20	18.90

1925:

	DJ dividends	DJ divisor
Q1	32.39	18.40
Q2	30.70	18.40
Q3	30.00	19.00
Q4	27.75	19.00

1926:

	DJ dividends	DJ divisor
Q1	37.13	17.42
Q2	27.13	16.67
Q3	29.88	16.67
Q4	26.88	16.67

1929:

	DJ dividends	DJ divisor
Q1	31.58	12.11
Q2	27.08	10.77
Q3	28.05	10.47
Q4	28.75	10.47

1932:

	DJ dividends	DJ divisor
Q1	15.54	15.46
Q2	18.23	15.46
Q3	16.51	15.46
Q4	16.26	15.46

1927:

	DJ dividends	DJ divisor
Q1	33.31	16.67
Q2	29.31	16.67
Q3	31.56	16.67
Q4	28.81	16.67

1930:

	DJ dividends	DJ divisor
Q1	30.05	10.47
Q2	27.90	9.85
Q3	27.50	10.38
Q4	25.95	10.38

1933:

	DJ dividends	DJ divisor
Q1	13.34	15.46
Q2	12.94	15.46
Q3	12.74	15.71
Q4	12.49	15.71

1928:

	DJ dividends	DJ divisor
Q1	31.38	16.67
Q2	27.63	16.67
Q3	30.63	16.17
Q4	27.45	13.92

1931:

	DJ dividends	DJ divisor
Q1	25.56	10.38
Q2	22.43	10.38
Q3	19.56	10.38
Q4	19.93	10.38

1934:

	DJ dividends	DJ divisor
Q1	13.90	15.71
Q2	13.95	15.71
Q3	14.10	15.74
Q4	15.30	15.74

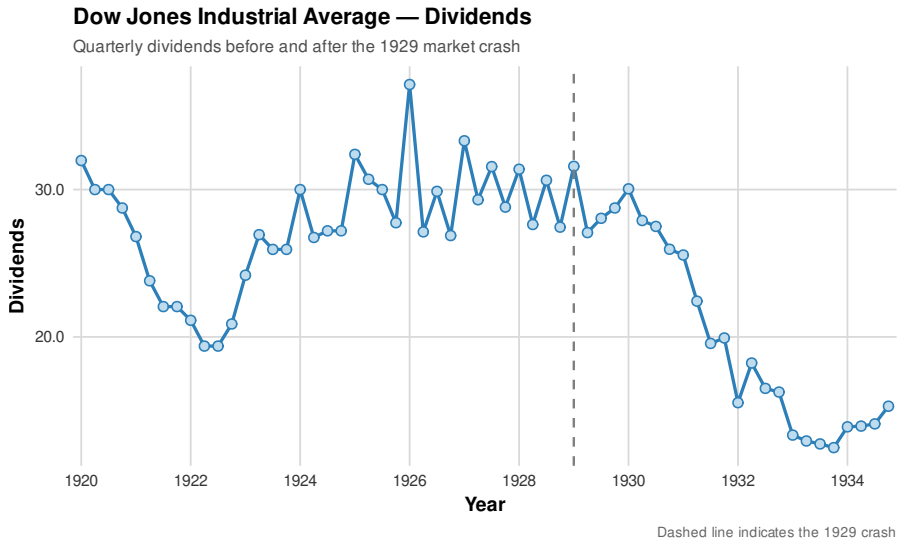
The time-series plots in Figure 4 illustrate the quarterly evolution of Dow-Jones Dividends and the Divisor from 1920 to 1934, providing a clear view of their temporal dynamics across different economic phases. Arranged chronologically, the plots reveal long-term trends, short-term fluctuations, and notable structural changes, particularly during the Great Depression.

To capture the complex dependence structure present in this dataset, across both temporal (quarterly) and cross-sectional (dividends and divisors) dimensions, we propose modeling the annual matrices in accordance with the distribution whose stochastic representation corresponds to the MVSN model. Specifically, we consider the formulation

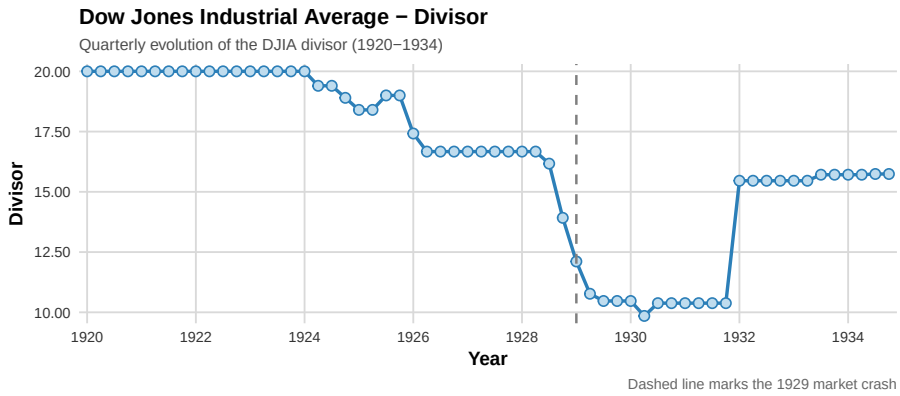
$$\boldsymbol{\mathcal{X}} = \mathbf{M} + W\mathbf{A} + \boldsymbol{\mathcal{V}},$$

where $\mathbf{M} \in \mathbb{R}^{p\times q}$ is a location matrix, $\mathbf{A} \in \mathbb{R}^{p\times q}$ encodes the directional skewness, $W \sim \mathrm{HN}(0,1)$ is a half normal scalar latent variable, and $\boldsymbol{\mathcal{X}} \sim \mathcal{N}_{p\times q}(\mathbf{0},\Sigma,\Psi)$ captures Gaussian noise with a separable covariance structure, as previously discussed. This makes it particularly well-suited for modeling economic and financial data that exhibit asymmetry. Moreover, the representation facilitates tractable likelihood-based inference and enables efficient parameter estimation even in high-dimensional settings.

To assess the goodness of fit of the proposed MVSN model, we employ a diagnostic based on Mahalanobis distances combined with a model-based parametric bootstrap. Mahalanobis distances are computed from vectorized observations using the Kronecker covariance structure implied by the fitted model. A parametric bootstrap is then used to generate a reference distribution under the estimated MVSN parameters, accounting for skewness and matrix-variate



(a) Quarterly evolution of Dow Jones dividends, 1920–1934.



(b) Quarterly evolution of the Dow Jones Industrial Average divisor, 1920–1934.

Figure 4 – Quarterly evolution of Dow Jones dividends and divisor before and after the 1929 market crash. The vertical dashed line marks the 1929 stock market crash. Dividends exhibit a sharp contraction followed by a slow recovery, while the divisor shows discrete structural adjustments reflecting changes in index composition and methodology during and after the crisis.

dependence. Figure 5 presents the resulting model-based Q–Q plot, together with the bootstrap mean curve and a pointwise 90% envelope. The empirical distances closely follow the identity line and remain within the envelope across the full range of quantiles, indicating an adequate fit of the MVSN model.

Figure 6 displays the evolution of the observed-data log-likelihood across iterations of the EM algorithm used to fit the MVSN model to the DJ dataset. The log-likelihood increases monotonically, with a steep improvement during the early iterations, followed by progressively smaller gains as the algorithm approaches a stable region. After approximately 30–40 iterations, the curve essentially plateaus, indicating that convergence has been reached and that further iterations yield negligible improvement. Overall, this behavior is consistent with stable numerical performance of the estimation procedure and supports the adequacy of the fitted model.

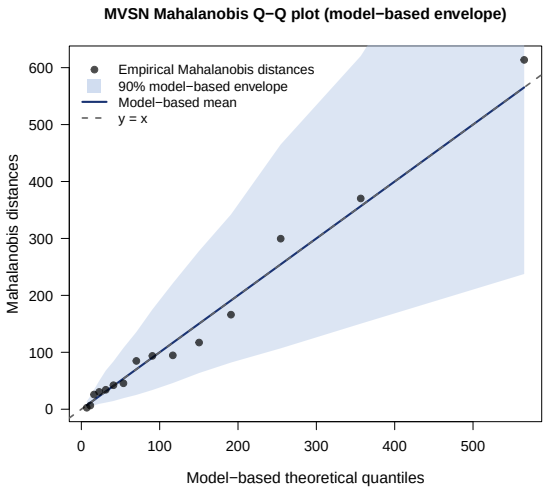


Figure 5 – Model-based Q–Q plot of Mahalanobis distances under the fitted MVSN model.

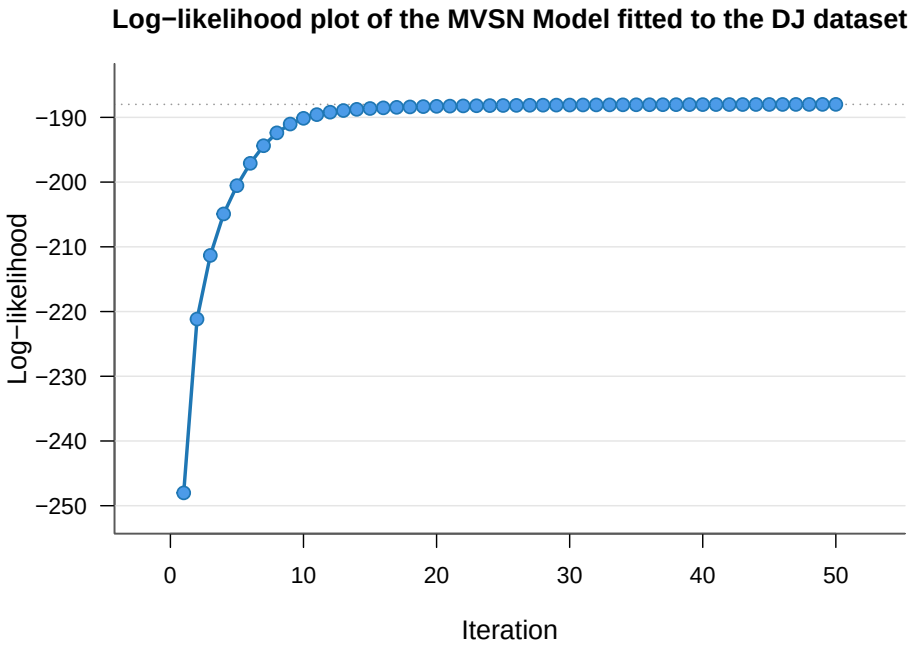


Figure 6 – Evolution of the observed-data log-likelihood over the EM iterations for the fitted MVSN model.

4.2.1 Interpreting the MVSN model estimates

This subsection examines the parameter estimates of the matrix-variate skew-normal (MVSN) model fitted to quarterly Dow Jones dividend and divisor data, emphasizing key aspects of the dependence, skewness, and variability structures. Particular attention is given to how the estimates reflect asymmetry and cross-variable interactions present in the data. Sampling variability is evaluated via a nonparametric bootstrap scheme, in which the model is repeatedly refitted to datasets resampled with replacement from the original observations, preserving the empirical structure while providing a flexible, distribution-free measure of estimation uncertainty.

Based on these bootstrap replicates, quantile-based confidence intervals are constructed for each entry of the estimated parameter matrices $\mathbf{M}$, $\mathbf{A}$, Σ, and Ψ. The resulting bootstrap standard errors account for both the matrix-variate dependence and finite-sample uncertainty, providing a practical basis for inference on the location, skewness, and covariance parameters. For ease of interpretation, bootstrap standard errors and corresponding quantile confidence intervals are reported alongside the parameter estimates.

Estimated matrix $\widehat{\mathbf{M}}_{MVSN}$

The estimated $\mathbf{M}$ matrix reveals moderate variation in dividend levels across quarters, with greater dispersion than that observed for the divisor. Although the quarterly means do not follow a strictly monotonic pattern, dividends display more pronounced quarter-to-quarter variability, as reflected in their larger nonparametric bootstrap standard errors and wider quantile-based confidence intervals for each matrix entry. In contrast, the divisor exhibits more stable estimates, with smaller associated uncertainties and narrower confidence intervals, consistent with its rule-based adjustment mechanism. These patterns are further supported by the bootstrap inference, which highlights the comparatively higher variability of dividends relative to the divisor. Such behavior is in line with economic downturn periods, during which dividends tend to adjust more rapidly than index construction components such as the divisor.

$$\widehat{\mathbf{M}}_{\text{MVSN}} = \begin{bmatrix} 19.32\,(4.36) & 16.41\,(2.57) \\ 20.48\,(2.92) & 16.42\,(2.79) \\ 18.59\,(3.42) & 16.56\,(2.82) \\ 19.34\,(3.10) & 16.70\,(2.91) \end{bmatrix} = \begin{bmatrix} 19.32\,[14.01, 31.32] & 16.41\,[9.93, 19.04] \\ 20.48\,[15.27, 26.78] & 16.42\,[9.36, 18.98] \\ 18.59\,[13.96, 28.59] & 16.56\,[9.39, 19.09] \\ 19.34\,[14.87, 27.50] & 16.70\,[9.39, 19.18] \end{bmatrix}$$

Estimated skewness matrix $\widehat{\boldsymbol{A}}_{MVSN}$

The estimated skewness matrix reveals a clearer asymmetry in the dividend component than in the divisor. The first-column entries are predominantly positive, indicating moderate right-skewness in dividends. However, the relatively large nonparametric bootstrap standard errors and wide quantile-based confidence intervals, many including zero, suggest caution in interpretation. In contrast, the second-column estimates are close to zero relative to their uncertainties, with confidence intervals spanning both positive and negative values, indicating weak or negligible skewness in the divisor. Overall, bootstrap inference suggests that asymmetry is mainly driven by dividends, with no strong evidence of systematic skewness accumulation once estimation uncertainty is taken into account.

$$\widehat{\mathbf{A}}_{\text{MVSN}} = \begin{bmatrix} 8.71\,(5.88) & 0.20\,(3.25) \\ 4.58\,(3.47) & -0.07\,(3.53) \\ 6.92\,(4.45) & -0.19\,(3.60) \\ 5.15\,(3.70) & -0.58\,(3.85) \end{bmatrix} = \begin{bmatrix} 8.71\,[-3.47, 17.81] & 0.20\,[-1.94, 9.63] \\ 4.58\,[-1.53, 12.32] & -0.07\,[-2.20, 10.30] \\ 6.92\,[-2.85, 14.97] & -0.19\,[-2.36, 10.34] \\ 5.15\,[-2.36, 11.64] & -0.58\,[-3.15, 10.21] \end{bmatrix}$$

Estimated Row Covariance Matrix $\widehat{\Sigma}_{MVSN}$

The estimated row covariance matrix indicates temporal dependence across quarters, with relatively larger covariances for adjacent periods, reflecting persistence in economic conditions. A mild decline in variances and covariances is observed in later quarters, suggesting a gradual easing of volatility over the year. However, the nonparametric bootstrap standard errors are of comparable magnitude and the quantile-based confidence intervals are relatively wide, indicating that this pattern should be interpreted with caution. Overall, the bootstrap inference supports a smooth evolution in quarterly dependence rather than abrupt structural changes.

$$\widehat{\Sigma}_{\text{MVSN}} = \begin{bmatrix} 17.51\,(8.54) & 17.52\,(8.91) & 16.98\,(8.80) & 16.55\,(8.97) \\ 17.52\,(8.91) & 18.21\,(9.29) & 17.84\,(9.21) & 17.22\,(9.36) \\ 16.98\,(8.80) & 17.84\,(9.21) & 17.66\,(9.11) & 17.16\,(9.27) \\ 16.55\,(8.97) & 17.22\,(9.36) & 17.16\,(9.27) & 17.18\,(9.40) \end{bmatrix}$$

$$= \begin{bmatrix} 17.51\,[4.10, 38.32] & 17.52\,[2.53, 39.64] & 16.98\,[2.49, 39.26] & 16.55\,[2.11, 38.85] \\ 17.52\,[2.53, 39.64] & 18.21\,[2.22, 41.53] & 17.84\,[2.00, 40.82] & 17.22\,[1.69, 40.38] \\ 16.98\,[2.49, 39.26] & 17.84\,[2.00, 40.82] & 17.66\,[2.28, 40.52] & 17.16\,[1.88, 40.41] \\ 16.55\,[2.11, 38.85] & 17.22\,[1.69, 40.38] & 17.16\,[1.88, 40.41] & 17.18\,[1.74, 40.23] \end{bmatrix}$$

Estimated column covariance matrix $\widehat{\Psi}_{MVSN}$

The estimated column covariance matrix indicates a weak positive association between dividends and the divisor, suggesting limited co-movement. The variances differ substantially, with dividends exhibiting greater variability than the divisor, consistent with their higher sensitivity to economic conditions. However, the relatively large nonparametric bootstrap standard errors and the width of the quantile-based confidence intervals—particularly for the covariance term, whose interval includes zero—indicate considerable uncertainty in these estimates. The determinant of the estimated matrix, $|\widehat{\Psi}| \approx 1$, confirms that the identifiability constraint is satisfied. Overall, the bootstrap inference suggests that, while variability differences are evident, the strength of dependence between the two components is not precisely estimated.

$$\widehat{\Psi}_{\text{MVSN}} = \begin{bmatrix} 2.26\,(3.47) & 0.11\,(0.15) \\ 0.11\,(0.15) & 0.45\,(0.15) \end{bmatrix} = \begin{bmatrix} 2.26\,[1.79, 12.22] & 0.11\,[-0.17, 0.46] \\ 0.11\,[-0.17, 0.46] & 0.45\,[0.08, 0.57] \end{bmatrix}$$

The MVSN model offers several advantages for the analysis of matrix-valued financial data. It captures complex dependence structures by jointly modeling temporal (row-wise) and cross-variable (column-wise) covariances, while the inclusion of skewness through the matrix $\mathbf{A}$ allows for departures from Gaussian symmetry, particularly evident in the dividend component. By preserving the matrix structure, the model facilitates interpretation and supports efficient likelihood-based estimation. Moreover, nonparametric bootstrap standard errors and quantile-based confidence intervals provide a flexible, distribution-free assessment of estimation uncertainty across all parameter matrices.

Model adequacy is assessed by comparing the Bayesian Information Criterion (BIC) of the MVSN model with that of the classical matrix-variate normal (MVN) model. The BIC values are 467.2621 for the MVN model and 451.8138 for the MVSN model, indicating a clear preference for the MVSN specification. This improvement reflects the benefits of accounting for asymmetry and richer dependence structures in the data.

Overall, the MVSN model provides a flexible and interpretable framework for analyzing matrix-valued financial series. The estimated parameters shed light on temporal dynamics, asymmetric behavior, and interactions between dividends and the divisor, while respecting the inherent structure of the data. Potential extensions include heavier-tailed matrix-variate models, such as skew-t or generalized hyperbolic formulations, as well as time-varying specifications to further enhance modeling flexibility.

CHAPTER

5

MVSNC PROPERTIES AND PARAMETER ESTIMATION

Matrix-variate normal and skew-normal models provide a natural and flexible framework for the analysis of data with an inherent two-way structure, such as longitudinal panels, spatio-temporal observations, image data, and repeated measurements. By preserving the matrix form of the observations, these models allow row-wise and column-wise dependence to be modeled separately through parsimonious covariance structures, while maintaining interpretability and computational tractability. Extensions to matrix-variate skew-normal models further enhance this framework by allowing for asymmetric behavior, which is frequently observed in empirical applications and cannot be adequately captured by symmetric Gaussian assumptions. In many applied contexts, matrix-valued observations are subject to censoring, either because of measurement limitations, detection thresholds, or institutional and design constraints. Ignoring censoring can lead to biased parameter estimates and distorted inference, particularly when censoring interacts with skewness and dependence structures. Although censored normal and skew-normal models are well developed for vector-valued data, corresponding methodological developments for matrix-variate distributions remain comparatively limited. The joint presence of matrix structure, skewness, and censoring substantially increases both theoretical and computational complexity, as it requires integrating over censored regions of high-dimensional sample spaces while simultaneously accounting for latent variables that induce asymmetry.

This chapter addresses these challenges by focusing on matrix-variate skew-normal models with censoring. The first section introduces the matrix-variate skew-normal censored (MVSNC) model, which extends existing matrix-variate skew-normal formulations by explicitly incorporating observation-level censoring. The model preserves separable covariance structures and interpretable skewness parameters, while remaining flexible enough to accommodate a broad range of censoring patterns encountered in practice. Its formulation relies on latent-variable representations that clarify the probabilistic structure and facilitate likelihood-based inference.

The second section of the chapter develops an estimation procedure for the MVSNC model based on an Expectation–Conditional Maximization (ECM) algorithm. The ECM framework is particularly well suited to this setting, as it exploits the hierarchical structure induced by both skewness and censoring, leading to computationally efficient and stable parameter updates. The chapter presents the construction of the complete-data likelihood, the derivation of the required conditional expectations, and the resulting conditional maximization steps. Practical aspects such as identifiability, convergence behavior, and implementation considerations are also discussed. As far as we know, the sole article discussing matrix variate distributions in the presence of censoring is Lachos *et al.* (2025a).

Together, these two sections provide a coherent methodological framework for modeling and inference with censored matrix-variate skew-normal data. The developments presented in this chapter lay the foundation for the empirical analyses and extensions explored in the subsequent chapters.

5.1 MVSN 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 observed matrices from the MVSN model. 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 i-th subject are given by the pair of matrices $(\mathcal{Y}_i, \mathcal{C}_i)$, where each element of the matrix $\mathcal{Y}_i$ represents either the uncensored observations $(\mathcal{Y}_{ijk} = \mathcal{Y}_{0ijk})$ or the interval-censoring level $(\mathcal{Y}_{ijk} \in [\mathcal{Y}_{1ijk}, \mathcal{Y}_{2ijk}])$, and $\mathcal{C}_i$ is the matrix of censoring indicators, satisfying

$$\mathcal{C}_{ijk} = \begin{cases} 1 & \text{if } \mathcal{Y}_{1ijk} \leq \mathcal{X}_{ijk} \leq \mathcal{Y}_{2ijk}, \\ 0 & \text{if } \mathcal{X}_{ijk} = \mathcal{Y}_{0ijk}, \end{cases} \tag{5.1}$$

for all $i \in \{1, \ldots, n\}$, $j \in \{1, \ldots, p\}$ and $k \in \{1, \ldots, q\}$. Accordingly, the observed information for the sample of n matrices is represented by the collections $\widetilde{\mathcal{Y}} = \{\mathcal{Y}_1, \ldots, \mathcal{Y}_i, \ldots, \mathcal{Y}_n\}$ and $\widetilde{\mathcal{C}} = \{\mathcal{C}_1, \ldots, \mathcal{C}_i, \ldots, \mathcal{C}_n\}$, where $\mathcal{Y}_i$ contains the observed interval bounds and $\mathcal{C}_i$ indicates the corresponding censoring status.

Under this formulation, the MVSN model combined with the relation defined in (5.1) gives rise to the MVSN interval-censored model (hereafter referred to as the MVSNC model). In this setting, the censoring mechanism induces truncation on both sides of the support of the underlying distribution. Specifically, for each entry, we only know that the latent observation $\mathcal{X}_{ijk}$ lies within the interval determined by the lower and upper bounds, that is, $\mathcal{Y}_{1ijk} \leq \mathcal{X}_{ijk} \leq \mathcal{Y}_{2ijk}$. Missing observations can be conveniently accommodated within this framework by setting $\mathcal{Y}_{1ijk} = -\infty$ and $\mathcal{Y}_{2ijk} = +\infty$, thereby preserving a unified representation for fully observed, censored, and missing data.

5.2 ML Estimation for the MVSNC Distribution

Since the censored log-likelihood of the MVSNC model is analytically intractable (at least to the best of our current knowledge), we overcome this difficulty by transferring the problem to the multivariate framework. This is possible due to Proposition 13. In particular, we exploit the results presented in the references Galarza, Matos and Lachos (2022) and Morales *et al.* (2022), where the authors discuss the multivariate skew normal censored model and the `MomTrunc` R package. To be more precise, here is the vector-valued version of the MVSNC model (MSNC* model from now on).

Let $\widetilde{\mathbf{x}} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ be n observed vectors from the model $\mathbf{X} \sim \text{MSN}^\star(\mu, \Omega, \lambda)$. However, the response vector $\mathbf{x}_i$ may not be fully observed due to censoring. We therefore define $(\widetilde{\mathbf{y}}, \widetilde{\mathbf{c}})$ as the observed data for the i-th sample, where $\widetilde{\mathbf{y}} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n\}$ contains either an uncensored observation ($\mathbf{y}_{ik} = \mathbf{y}_{0ik}$) or an interval-censored value ($\mathbf{y}_{ik} \in [\mathbf{y}_{1ik}, \mathbf{y}_{2ik}]$). Each vector in the set $\widetilde{\mathbf{c}} = \{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_n\}$ denotes the censoring indicators, defined by

$$\mathbf{c}_{ik} = \begin{cases} 1, & \text{if } \mathbf{y}_{1ik} \leq \mathbf{y}_{ik} \leq \mathbf{y}_{2ik}, \\ 0, & \text{if } \mathbf{y}_{ik} = \mathbf{y}_{0ik}, \end{cases}$$

for all $i \in \{1, \ldots, n\}$ and $k \in \{1, \ldots, p\}$. In other words, $\mathbf{c}_{ik} = 1$ whenever the observation $\mathbf{y}_{ik}$ falls within a specified censoring interval, indicating that its exact value is not directly observed but only known to lie between predetermined bounds.

Under this formulation, the model $\mathbf{X} \sim \text{MSN}^\star(\mu, \Omega, \lambda)$ defines the multivariate skew-normal interval-censored model (hereafter referred to as the MSNC* model). Similarly, one can define the distribution $\mathbf{X} \sim \text{MSNC}(\mu, \Delta, \mathbf{b})$ which comes from the distribution MSN with parametrization $(\mu, \Delta, \mathbf{b})$ in the presence of censoring. Missing observations can be incorporated into this framework by setting $\mathbf{y}_{1ik} = -\infty$ and $\mathbf{y}_{2ik} = +\infty$.

Proposition 28. Let $\mathbf{X} \sim \text{MESN}_p^\star(\mu, \Omega, \lambda, \kappa)$, and suppose that $\mathbf{X}$ is partitioned into two parts as $\mathbf{X} = (\mathbf{X}_1^\top, \mathbf{X}_2^\top)^\top$, where the subvectors $\mathbf{X}_1$ and $\mathbf{X}_2$ have dimensions $p_1 \geq 1$ as well as $p_2 \geq 1$, respectively, satisfying $p_1 + p_2 = p$. Let us also set up that

$$\Omega = \begin{pmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{pmatrix}, \quad \mu = (\mu_1^\top, \mu_2^\top)^\top, \quad \lambda = (\lambda_1^\top, \lambda_2^\top)^\top \quad \text{and} \quad \varphi = (\varphi_1^\top, \varphi_2^\top)^\top,$$

which represent the corresponding partitions of Ω, μ, λ and $\varphi := \Omega^{-1/2}\lambda$, where each is defined consistently with the same block structure adopted for the random vector and its associated covariance decomposition. Consequently, it can be concluded firstly that:

$$\mathbf{X}_1 \sim \text{MESN}_{p_1}^\star(\mu_1, \Omega_{11}, c_{12}\Omega_{11}^{1/2}\widetilde{\varphi}_1, c_{12}\kappa)$$

Secondly, one may also claim that

$$\mathbf{X}_2 \mid \mathbf{X}_1 = \mathbf{x}_1 \sim \text{MESN}_{p_2}^\star(\mu_{2.1}, \Omega_{22.1}, \Omega_{22.1}^{1/2}\varphi_2, \kappa_{2.1}),$$

where the following relations hold

$$
\begin{aligned}
c_{12} &= \left(1+\varphi_2^\top \Omega_{22.1}\varphi_2\right)^{-1/2} && \text{(normalizing constant)}\\
\widetilde{\varphi}_1 &= \varphi_1 + \Omega_{11}^{-1}\Omega_{12}\varphi_2 && \text{(adjusted shape vector)}\\
\Omega_{22.1} &= \Omega_{22} - \Omega_{21}\Omega_{11}^{-1}\Omega_{12} && \text{(Schur complement)}\\
\mu_{2.1} &= \mu_2 + \Omega_{21}\Omega_{11}^{-1}(\mathbf{x}_1-\mu_1) && \text{(conditional mean)}\\
\kappa_{2.1} &= \kappa + \widetilde{\varphi}_1^\top(\mathbf{x}_1-\mu_1) && \text{(updated skewness)}
\end{aligned}
$$

Proof. This result is mentioned in the reference Morales *et al.* (2022), for instance. □

Proposition 29. Let $\mathbf{X} \sim \text{MESN}_p(\mu,\Delta,\mathbf{b},\kappa)$ in accordance with Definition 13, and let $\mathbf{X}$ be partitioned as $\mathbf{X} = (\mathbf{X}_1^\top, \mathbf{X}_2^\top)^\top$ where $p_1 + p_2 = p$. Let us also consider the partition of Δ:

$$
\Delta = \begin{pmatrix} \Delta_{11} & \Delta_{12} \\ \Delta_{21} & \Delta_{22} \end{pmatrix}
$$

where $\mu = (\mu_1^\top, \mu_2^\top)^\top$ and $\mathbf{b} = (\mathbf{b}_1^\top, \mathbf{b}_2^\top)^\top$ be the corresponding partitions of μ and $\mathbf{b}$. Then $\mathbf{X}_1 \sim \text{MESN}_{p_1}(\mu_1,\Delta_{11},\mathbf{b}_1,\kappa_1)$ as well as $\mathbf{X}_2 \mid \mathbf{X}_1 = \mathbf{x}_1 \sim \text{MESN}_{p_2}(\mu_{2.1} + \mathbf{b}_{2.1}\frac{\mathrm{d}_{\mathbf{b}_1}(\mathbf{x}_1)}{\delta_1}, \Delta_{22.1}, \mathbf{b}_{2.1}, \kappa_{2.1})$, where the following relations hold

$$
\begin{aligned}
\kappa_1 &= \kappa\delta_1/\delta && \text{(marginal skewness parameter)},\\
\mathbf{b}_{2.1} &= \frac{1}{\delta_1}\left(\mathbf{b}_2 - \Delta_{21}\Delta_{11}^{-1}\mathbf{b}_1\right) && \text{(conditional shape vector)},\\
\Delta_{22.1} &= \Delta_{22} - \Delta_{21}\Delta_{11}^{-1}\Delta_{12} && \text{(Schur complement / conditional scale matrix)},\\
\mu_{2.1} &= \mu_2 + \Delta_{21}\Delta_{11}^{-1}(\mathbf{x}_1-\mu_1) && \text{(conditional mean)},\\
\kappa_{2.1} &= \delta_{2.1}\left(\kappa_1 + \frac{\mathrm{d}_{\mathbf{b}_1}(\mathbf{x}_1)}{\delta_1}\right) && \text{(updated conditional skewness)}.
\end{aligned}
$$

with $\delta_1 = \sqrt{1+\mathbf{b}_1^\top\Delta_{11}^{-1}\mathbf{b}_1}$, $\mathrm{d}_{\mathbf{b}_1}(\mathbf{x}_1) = \mathbf{b}_1^\top\Delta_{11}^{-1}(\mathbf{x}_1-\mu_1)$ and $\delta_{2.1} = \sqrt{1+\mathbf{b}_{2.1}^\top\Delta_{22.1}^{-1}\mathbf{b}_{2.1}}$. Assume throughout that Δ is symmetric positive definite. In particular, this implies that Δ_{11} is positive definite and therefore invertible, and that the Schur complement $\Delta_{22.1}$ is also positive definite.

Proof. The stochastic representation of the target distribution is given by $\mathbf{X} = \mu + \mathbf{b}W + \mathbf{V}$ with $W \sim \text{TN}(0,1,[-\kappa/\delta,\infty))$, $\mathbf{V} \sim \mathcal{N}_p(\mathbf{0},\Delta)$, $W \perp \mathbf{V}$, and $\delta = \sqrt{1+\mathbf{b}^\top\Delta^{-1}\mathbf{b}}$. Partition conformably

$$
\mathbf{X} = \begin{pmatrix}\mathbf{X}_1\\ \mathbf{X}_2\end{pmatrix},\quad \mu = \begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix},\quad \mathbf{b} = \begin{pmatrix}\mathbf{b}_1\\ \mathbf{b}_2\end{pmatrix},\quad \Delta = \begin{pmatrix}\Delta_{11} & \Delta_{12}\\ \Delta_{21} & \Delta_{22}\end{pmatrix},\quad \mathbf{V} = \begin{pmatrix}\mathbf{V}_1\\ \mathbf{V}_2\end{pmatrix}.
$$

We start with the derivation of the marginal distribution of $\mathbf{X}_1$. From the stochastic representation we immediately obtain $\mathbf{X}_1 = \mu_1 + \mathbf{b}_1 W + \mathbf{V}_1$. By the marginalization property of the multivariate normal distribution, $\mathbf{V}_1 \sim \mathcal{N}_{p_1}(\mathbf{0},\Delta_{11})$, and since $W \perp \mathbf{V}$, we also have $W \perp \mathbf{V}_1$. Hence $\mathbf{X}_1$ admits the same latent–truncation representation as the MESN model in dimension p_1. Define

$$
\delta_1 = \sqrt{1+\mathbf{b}_1^\top\Delta_{11}^{-1}\mathbf{b}_1}.
$$

Given that the same latent variable W appears in the marginal stochastic representation of $\mathbf{X}_1$, its truncation bound remains $-\kappa/\delta$. In the reduced MESN parametrization, however, the truncation is written as $-\kappa_1/\delta_1$. Therefore consistency of the parametrization requires

$$-\frac{\kappa_1}{\delta_1} = -\frac{\kappa}{\delta},$$

which implies $\kappa_1 = \kappa\delta_1/\delta$.

Before determining the distribution of the $\mathbf{X}_2$ given $\mathbf{X}_1 = \mathbf{x}_1$, it is convenient to define the transformed random vector $\boldsymbol{\varepsilon} := \mathbf{V}_2 - \Delta_{21}\Delta_{11}^{-1}\mathbf{V}_1$. From the theory of partitioned multivariate normal distributions, it follows that $\boldsymbol{\varepsilon} \sim \mathcal{N}_{p_2}(\mathbf{0}, \Delta_{22.1})$, where the conditional covariance matrix is given by $\Delta_{22.1} = \Delta_{22} - \Delta_{21}\Delta_{11}^{-1}\Delta_{12}$.

Since $W \perp \mathbf{V}$ and $\boldsymbol{\varepsilon}$ is a linear transformation of $\mathbf{V}$, it follows that $W \perp \boldsymbol{\varepsilon}$. Substituting $\mathbf{V}_1 = \mathbf{X}_1 - \mu_1 - \mathbf{b}_1 W$ into $\mathbf{V}_2 = \Delta_{21}\Delta_{11}^{-1}\mathbf{V}_1 + \boldsymbol{\varepsilon}$, and the result into $\mathbf{X}_2 = \mu_2 + \mathbf{b}_2 W + \mathbf{V}_2$, it yields that

$$\mathbf{X}_2 = \mu_2 + \Delta_{21}\Delta_{11}^{-1}(\mathbf{X}_1 - \mu_1) + (\mathbf{b}_2 - \Delta_{21}\Delta_{11}^{-1}\mathbf{b}_1)W + \boldsymbol{\varepsilon}.$$

Therefore, upon conditioning on $\mathbf{X}_1 = \mathbf{x}_1$, the conditional random vector $\mathbf{X}_2 \mid \mathbf{X}_1 = \mathbf{x}_1$ can be written explicitly in the following stochastic representation form:

$$\mathbf{X}_2 \mid \mathbf{X}_1 = \mathbf{x}_1 = \mu_{2.1} + \mathbf{r}W + \boldsymbol{\varepsilon},$$

where the conditional location term is given by

$$\mu_{2.1} = \mu_2 + \Delta_{21}\Delta_{11}^{-1}(\mathbf{x}_1 - \mu_1),$$

and the adjusted shape vector is $\mathbf{r} = \mathbf{b}_2 - \Delta_{21}\Delta_{11}^{-1}\mathbf{b}_1$.

At last, it remains to derive explicitly the conditional distribution of $W \mid \mathbf{X}_1 = \mathbf{x}_1$, which is the key step in completing the hierarchical representation of the model. Recall that, from the stochastic representation $\mathbf{X}_1 = \mu_1 + \mathbf{b}_1 W + \mathbf{V}_1$ with $\mathbf{V}_1 \sim \mathcal{N}_{p_1}(\mathbf{0}, \Delta_{11})$ and $W \perp \mathbf{V}_1$, it follows that, for any fixed $w \in \mathbb{R}$,

$$\mathbf{X}_1 \mid W = w \sim \mathcal{N}_{p_1}(\mu_1 + \mathbf{b}_1 w, \Delta_{11}).$$

Consequently, the conditional likelihood of $\mathbf{X}_1$ given the latent variable W is proportional to the Gaussian kernel

$$f_{\mathbf{X}_1|W}(\mathbf{x}_1 \mid w) \propto \exp\left\{-\frac{1}{2}(\mathbf{x}_1 - \mu_1 - \mathbf{b}_1 w)^\top \Delta_{11}^{-1}(\mathbf{x}_1 - \mu_1 - \mathbf{b}_1 w)\right\}.$$

On the other hand, by assumption, $W \sim \mathrm{TN}(0, 1, [-\kappa/\delta, +\infty))$ so its density is proportional to the standard normal kernel $\exp(-w^2/2)$ restricted to the interval $[-\kappa/\delta, +\infty)$. Since

$$f_{W|\mathbf{X}_1}(w \mid \mathbf{x}_1) \propto f_W(w) f_{\mathbf{X}_1|W}(\mathbf{x}_1 \mid w),$$

and the prior density of W already incorporates the truncation indicator $\mathbf{1}\{w \geq -\kappa/\delta\}$, the posterior density is proportional to the product of the Gaussian likelihood kernel and the truncated normal prior kernel. The posterior density of W given $\mathbf{X}_1 = \mathbf{x}_1$ is, up to a normalizing constant,

$$f_{W|\mathbf{X}_1}(w \mid \mathbf{x}_1) \propto \exp\left\{-\frac{1}{2}\left[w^2 + (\mathbf{x}_1 - \mu_1 - \mathbf{b}_1 w)^\top \Delta_{11}^{-1}(\mathbf{x}_1 - \mu_1 - \mathbf{b}_1 w)\right]\right\} \mathbf{1}\left\{w \geq -\frac{\kappa}{\delta}\right\}.$$

We now expand the quadratic form in w appearing in the exponent. A straightforward algebraic calculation yields the result

$$(\mathbf{x}_1 - \mu_1 - \mathbf{b}_1 w)^\top \Delta_{11}^{-1}(\mathbf{x}_1 - \mu_1 - \mathbf{b}_1 w) = (\mathbf{x}_1 - \mu_1)^\top \Delta_{11}^{-1}(\mathbf{x}_1 - \mu_1) - 2w\mathbf{b}_1^\top \Delta_{11}^{-1}(\mathbf{x}_1 - \mu_1) + w^2 \mathbf{b}_1^\top \Delta_{11}^{-1}\mathbf{b}_1$$

Collecting the terms that depend on w, the exponent becomes

$$\text{exponent} = -\frac{1}{2}\left[\delta_1^2 w^2 - 2\mathrm{d}_{\mathbf{b}_1}(\mathbf{x}_1) w\right],$$

where we have introduced the quantities $\mathrm{d}_{\mathbf{b}_1}(\mathbf{x}_1) = \mathbf{b}_1^\top \Delta_{11}^{-1}(\mathbf{x}_1 - \mu_1)$ and $\delta_1^2 = 1 + \mathbf{b}_1^\top \Delta_{11}^{-1}\mathbf{b}_1$.

The posterior kernel is therefore quadratic in w, and the distribution can be identified by completing the square. Indeed,

$$\delta_1^2 w^2 - 2\mathrm{d}_{\mathbf{b}_1}(\mathbf{x}_1) w = \delta_1^2 \left(w - \frac{\mathrm{d}_{\mathbf{b}_1}(\mathbf{x}_1)}{\delta_1^2}\right)^2 - \frac{\mathrm{d}_{\mathbf{b}_1}^2(\mathbf{x}_1)}{\delta_1^2},$$

Since the last term does not depend on w, it is absorbed into the normalizing constant. We conclude that the conditional density corresponds to a normal kernel with mean $\mathrm{d}_{\mathbf{b}_1}(\mathbf{x}_1)/\delta_1^2$ and variance $1/\delta_1^2$, truncated to the original support of W. Therefore,

$$W \mid \mathbf{X}_1 = \mathbf{x}_1 \sim \mathrm{TN}\left(\frac{\mathrm{d}_{\mathbf{b}_1}(\mathbf{x}_1)}{\delta_1^2}, \frac{1}{\delta_1^2}, \left[-\frac{\kappa}{\delta}, \infty\right)\right),$$

which establishes the claimed conditional Gaussian structure and completes the derivation.

We now introduce the location-scale transformation

$$U = \delta_1 \left(W - \frac{\mathrm{d}_{\mathbf{b}_1}(\mathbf{x}_1)}{\delta_1^2}\right),$$

so that the conditional latent variable is expressed in standardized form. This reparametrization allows the distribution to be written in the canonical MESN structure. Then it follows that $U \mid \mathbf{X}_1 = \mathbf{x}_1$ is standard normal, and the truncation transforms linearly to

$$U \geq -\left(\kappa_1 + \frac{\mathrm{d}_{\mathbf{b}_1}(\mathbf{x}_1)}{\delta_1}\right).$$

Consequently, we obtain that the latent variable U given $\mathbf{X}_1 = \mathbf{x}_1$ follows a standard normal distribution truncated at the updated lower threshold. That is to say,

$$U \mid \mathbf{X}_1 = \mathbf{x}_1 \sim \mathrm{TN}\left(0, 1, \left[-\left(\kappa_1 + \frac{\mathrm{d}_{\mathbf{b}_1}(\mathbf{x}_1)}{\delta_1}\right), +\infty\right)\right),$$

which restores the latent variable to the standard $\mathrm{TN}(0,1,\cdot)$ form required by the target MESN stochastic representation. At last, since

$$W = \frac{\mathrm{d}_{\mathbf{b}_1}(\mathbf{x}_1)}{\delta_1^2} + \frac{U}{\delta_1},$$

we substitute this expression into the conditional representation of $\mathbf{X}_2$. A direct rearrangement then yields the following result

$$\mathbf{X}_2 \mid \mathbf{X}_1 = \mathbf{x}_1 = \mu_{2.1} + \mathbf{b}_{2.1}\frac{\mathrm{d}_{\mathbf{b}_1}(\mathbf{x}_1)}{\delta_1} + \mathbf{b}_{2.1}U + \varepsilon, \quad \text{where} \quad \mathbf{b}_{2.1} = \frac{1}{\delta_1}\left(\mathbf{b}_2 - \Delta_{21}\Delta_{11}^{-1}\mathbf{b}_1\right).$$

The first two terms define the updated conditional location, while the term $\mathbf{b}_{2.1}U$ represents the new skewing mechanism induced by the standardized truncated normal variable U. The noise term ε remains Gaussian with covariance matrix $\Delta_{22.1}$ and is independent of U.

This is the MESN stochastic representation in dimension p_2, as the conditional model preserves the latent–truncation structure with parameters updated according to $\mathbf{x}_1$. In particular, the Gaussian component is governed by the Schur complement $\Delta_{22.1}$, and the truncation level is adjusted as

$$\frac{\kappa_{2.1}}{\delta_{2.1}} = \kappa_1 + \frac{\mathrm{d}_{\mathbf{b}_1}(\mathbf{x}_1)}{\delta_1} \quad \text{and} \quad \delta_{2.1} = \sqrt{1 + \mathbf{b}_{2.1}^\top \Delta_{22.1}^{-1}\mathbf{b}_{2.1}}.$$

Hence,

$$\kappa_{2.1} = \delta_{2.1}\left(\kappa_1 + \frac{\mathrm{d}_{\mathbf{b}_1}(\mathbf{x}_1)}{\delta_1}\right),$$

which ensures that the truncation is written in the canonical MESN parametrization. Therefore,

$$\mathbf{X}_2 \mid \mathbf{X}_1 = \mathbf{x}_1 \sim \mathrm{MESN}_{p_2}\left(\mu_{2.1} + \mathbf{b}_{2.1}\frac{\mathrm{d}_{\mathbf{b}_1}(\mathbf{x}_1)}{\delta_1}, \Delta_{22.1}, \mathbf{b}_{2.1}, \kappa_{2.1}\right),$$

showing that the MESN family is closed under conditioning with respect to subvectors. □

To calculate the censored likelihood, let us start by considering the vectorization of the observed data $\widetilde{\mathcal{X}} = \{\mathcal{X}_1, \ldots, \mathcal{X}_n\}$. Its vectorized version is denoted by $\mathrm{vec}(\widetilde{\mathcal{X}}) = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$. Similarly, the vectorization of the censored entries $\widetilde{\mathcal{Y}} = \{\mathcal{Y}_1, \ldots, \mathcal{Y}_n\}$ is $\mathrm{vec}(\widetilde{\mathcal{Y}}) = \{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$. Lastly, the vectorization of the censoring indicator matrices $\widetilde{\mathcal{C}} = \{\mathcal{C}_1, \ldots, \mathcal{C}_n\}$ is denoted by $\mathrm{vec}(\widetilde{\mathcal{C}}) = \{\mathbf{c}_1, \ldots, \mathbf{c}_n\}$.

In order to construct the observed-data likelihood under censoring, we decompose each observed realization into its observed and censored components, thereby separating the fully observed entries from those subject to interval restrictions. More precisely, for each $i \in \{1, 2, \ldots, n\}$, we write $\mathbf{x}_i = (\mathbf{x}_i^{o\top}, \mathbf{x}_i^{c\top})^\top$, where the superscripts o and c denote the observed and censored parts of the vector, respectively. This decomposition is performed componentwise according to the censoring pattern of the i-th observation.

Accordingly, after a suitable reordering of the entries to align the observed and censored components, the associated censoring indicators and latent variables may be written in a compatible block form as $\mathbf{c}_i = (\mathbf{c}_i^{o\top}, \mathbf{c}_i^{c\top})^\top$ and $\mathbf{y}_i = (\mathbf{y}_i^{o\top}, \mathbf{y}_i^{c\top})^\top$, where, in particular, the censored component of the latent vector is further partitioned as $\mathbf{y}_i^c = (\mathbf{y}_{1i}^c, \mathbf{y}_{2i}^c)$.

This block representation is fundamental for deriving the likelihood contribution of each observation, since it allows us to express the joint density in terms of marginal and conditional distributions. In particular, depending on the parametrization adopted for the multivariate extended skew normal model, one may invoke either of the Propositions 28 or 29 to compute the likelihood function in closed form.

We assume that our base model is described by $\boldsymbol{\mathcal{X}} \sim \text{MVSNC}_{p\times q}(\mathbf{M}, \mathbf{A}, \Sigma, \Psi)$. In order to reformulate the log-likelihood from the matrix-variate context into an equivalent vector-valued representation, we invoke Proposition 13. As a result, the following relation holds $\text{vec}(\boldsymbol{\mathcal{X}}) \sim \text{MSNC}(\text{vec}(\mathbf{M}), \text{vec}(\mathbf{A}), \Psi \otimes \Sigma)$. If one chooses to approach the problem through Proposition 28, we need to make a parametrization convertion. More precisely, in light of Propositions 6 and 4, $\text{vec}(\boldsymbol{\mathcal{X}}) \sim \text{MSNC}^\star_{pq}(\text{vec}(\mathbf{M}), \Omega, \mu)$, where $\Omega = \Psi \otimes \Sigma + \text{vec}(\mathbf{A})\,\text{vec}(\mathbf{A})^\top$ and the expression of λ is given by

$$\lambda = \frac{(\Psi \otimes \Sigma + \text{vec}(\mathbf{A})\,\text{vec}(\mathbf{A})^\top)^{-1/2}\,\text{vec}(\mathbf{A})}{(1 + \text{vec}(\mathbf{A})^\top(\Psi \otimes \Sigma)^{-1}\,\text{vec}(\mathbf{A}))^{-1/2}} = \frac{\Omega^{-1/2}\,\text{vec}(\mathbf{A})}{(1 - \text{vec}(\mathbf{A})^\top \Omega^{-1}\,\text{vec}(\mathbf{A}))^{1/2}}$$

Taking into account the notation in Proposition 28, the corresponding relations are

$$\mu_i = (\mu_i^{o\top}, \mu_i^{c\top})^\top, \quad \Omega_i = \begin{pmatrix} \Omega_i^{oo} & \Omega_i^{oc} \\ \Omega_i^{co} & \Omega_i^{cc} \end{pmatrix}, \quad \lambda_i = (\lambda_i^{o\top}, \lambda_i^{c\top})^\top \quad \text{and} \quad \varphi_i = (\varphi_i^{o\top}, \varphi_i^{c\top})^\top.$$

Then, still in accordance with Proposition 28, we have the following.

$$\mathbf{X}_i^o \sim \text{MSN}_{p_i^o}(\mu_i^o, \Omega_i^{oo}, c_i^{oc}\Omega_i^{oo\,1/2}\widetilde{\varphi}_i^o)$$

and $\mathbf{X}_i^c \mid \mathbf{X}_i^o = \mathbf{x}_i^o \sim \text{MESN}_{p_i^c}(\mu_i^{co}, \Omega_i^{cc.o}, \Omega_i^{cc.o\,1/2}\varphi_i^c, \kappa_i^{co})$, where

1. $\mu_i^{co} = \mu_i^c + \Omega_i^{co}(\Omega_i^{oo})^{-1}(\mathbf{x}_i^o - \mu_i^o)$,
2. $\Omega_i^{cc.o} = \Omega_i^{cc} - \Omega_i^{co}(\Omega_i^{oo})^{-1}\Omega_i^{oc}$,
3. $\widetilde{\varphi}_i^o = \varphi_i^o + (\Omega^{oo})_i^{-1}\Omega_i^{oc}\varphi_i^c$,
4. $c_i^{oc} = (1 + \varphi_i^{c\top}\Omega_i^{cc.o}\varphi_i^c)^{-1/2}$,
5. $\kappa_i^{co} = \widetilde{\varphi}_i^{o\top}(\mathbf{x}_i^o - \mu_i^o)$.

The likelihood function corresponding to $\vartheta := (\mathbf{M}, \mathbf{A}, \Sigma, \Psi)$ given the observed data $(\widetilde{\mathcal{Y}}, \widetilde{\mathcal{C}})$ is

$$\ell(\vartheta \mid \text{vec}(\widetilde{\mathcal{Y}}), \text{vec}(\widetilde{\mathcal{C}})) = \sum_{i=1}^{n} \log \ell_i(\vartheta \mid \text{vec}(\mathcal{X}_i)), \tag{5.2}$$

where ℓ_i represents the likelihood function of ϑ for the i-th sample observation $\text{vec}(\mathcal{X}_i)$:

$$\ell_i(\text{vec}(\mathcal{X}_i) \mid \vartheta) = f(\mathbf{y}_{1i}^c \leq \mathbf{x}_i^c \leq \mathbf{y}_{2i}^c \mid \mathbf{x}_i^o, \vartheta) f(\mathbf{x}_i^o \mid \vartheta)$$

in accordance with the previous distributions of the random vectors $\mathbf{X}_i^o$ and $\mathbf{X}_i^c \mid \mathbf{X}_i^o = \mathbf{x}_i^o$.

An analogous approach can be developed by invoking Proposition 29. By partitioning the vector-valued random variable $\mathbf{X}$ into observed and censored components and adopting the corresponding block decomposition of $(\mu, \Delta, \mathbf{b})$, the marginal distribution of $\mathbf{X}_1$ and the conditional distribution of $\mathbf{X}_2 \mid \mathbf{X}_1 = \mathbf{x}_1$ follow from the closure of the MESN_p family under partitioning. The updated parameters arise from the same Schur complement structure and affine adjustments described in Proposition 29, preserving the MESN stochastic representation.

Although the observed log-likelihood function (5.2) can be directly evaluated through the `MomTrunc` R package without imposing a heavy computational cost, its analytical form involves intricate expressions that make it impractical for straightforward maximum likelihood (ML) parameter estimation. To address this issue and improve numerical stability as well as interpretability, the estimation procedure can be greatly facilitated through the use of an EM-based algorithm, which decomposes the optimization into simpler expectation and maximization steps, providing an efficient iterative framework for convergence to the ML estimates.

5.2.1 The ECM algorithm

To implement the ECM algorithm for estimating the parameters of the MVSN distribution under censoring — following the EM framework of Dempster, Laird and Rubin (1977) and its conditional maximization extension of Meng and Rubin (1993) — we adopt the latent-variable representation of the model. In this formulation, both the unobserved skewing variable and the censored entries are incorporated into an augmented complete-data structure, which makes the likelihood substantially more tractable than its observed-data counterpart and facilitates the subsequent optimization steps.

The core idea of the EM-type approach is to treat the latent variable W together with the censored components of $\mathcal{X}_i$ as missing data, thereby embedding the problem into a complete-data framework $(\mathcal{X}_i, W_i)$. Under this augmented representation, the complete-data log-likelihood admits a simpler analytical form, avoiding the intractable integrations required in the observed-data likelihood. Consequently, the algorithm proceeds iteratively by maximizing the conditional expectation of the complete-data log-likelihood given the observed (censored) data and the current parameter estimates.

For clarity, we present the full expression of the E-step function $Q_i(\vartheta \mid \vartheta^{(k)})$ in the censored-data setting. This formulation makes explicit how the conditional expectations incorporate both the latent skewing variable and the unobserved censored components. It also allows each parameter to be analyzed and updated separately, which distinguishes the classical EM algorithm from its ECM extension.

To proceed, we adopt the following conventions:

$$\begin{aligned}
\widehat{w}_i^{(k+1)} &:= \mathbb{E}_{W,\mathrm{vec}(\mathcal{X}_i)}(W_i \mid \mathrm{vec}(\widetilde{\mathcal{Y}}_i), \mathrm{vec}(\widetilde{\mathcal{C}}_i), \widehat{\vartheta}^{(k)}),\\
\widehat{w}_i^{2(k+1)} &:= \mathbb{E}_{W,\mathrm{vec}(\mathcal{X}_i)}(W_i^2 \mid \mathrm{vec}(\widetilde{\mathcal{Y}}_i), \mathrm{vec}(\widetilde{\mathcal{C}}_i), \widehat{\vartheta}^{(k)}),\\
\mathrm{vec}(\widehat{w\mathcal{X}}_i)^{(k+1)} &:= \mathbb{E}_{W,\mathrm{vec}(\mathcal{X}_i)}(W_i \mathrm{vec}(\mathcal{X}_i) \mid \mathrm{vec}(\widetilde{\mathcal{Y}}_i), \mathrm{vec}(\widetilde{\mathcal{C}}_i), \widehat{\vartheta}^{(k)})\\
\mathrm{vec}(\widehat{\mathcal{X}}_i^{(k+1)}) &:= \mathbb{E}_{W,\mathrm{vec}(\mathcal{X}_i)}(\mathrm{vec}(\mathcal{X}_i) \mid \mathrm{vec}(\widetilde{\mathcal{Y}}_i), \mathrm{vec}(\widetilde{\mathcal{C}}_i), \widehat{\vartheta}^{(k)})\\
\mathrm{vec}(\widehat{\mathcal{X}}_i^2)^{(k+1)} &:= \mathbb{E}_{W,\mathrm{vec}(\mathcal{X}_i)}(\mathrm{vec}(\mathcal{X}_i)\mathrm{vec}(\mathcal{X}_i)^\top \mid \mathrm{vec}(\widetilde{\mathcal{Y}}_i), \mathrm{vec}(\widetilde{\mathcal{C}}_i), \widehat{\vartheta}^{(k)})
\end{aligned}$$

which can be derived following the results presented in the reference Galarza, Matos and Lachos (2022). Indeed, the expressions for them can be found in such reference and, whenever truncated moments are needed, we do also use Morales *et al.* (2022) to overcome such difficulty. This conversion is possible because the conditional expectations required in the matrix-variate case can be equivalently expressed in terms of the corresponding conditional expectations obtained after vectorizing the data. To apply the expectation step, we need to determine the individual log-likelihood of the joint distribution of the pair $(\mathcal{X}, W)$ for each $(\mathcal{Y}_i, \mathcal{C}_i)$ where $i \in \{1,2,\ldots,n\}$. Specifically, if we stablish the following convention $\mathcal{Z} = \mathcal{X} - \mathbf{M}$, it can be expressed as:

$$\begin{aligned}
\ell_{ic}(\vartheta) \propto & -\frac{q}{2}\log|\Sigma| - \frac{p}{2}\log|\Psi| - \frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}(\mathcal{Z}_i - w\mathbf{A})\Psi^{-1}(\mathcal{Z}_i - w\mathbf{A})^\top\right)\\
= & -\frac{q}{2}\log|\Sigma| - \frac{p}{2}\log|\Psi| - \frac{1}{2}\mathrm{tr}((\Psi\otimes\Sigma)^{-1}\mathrm{vec}(\mathcal{Z}_i - w\mathbf{A})\mathrm{vec}(\mathcal{Z}_i - w\mathbf{A})^\top)\\
= & -\frac{q}{2}\log|\Sigma| - \frac{p}{2}\log|\Psi|\\
& -\frac{1}{2}\mathrm{tr}((\Psi\otimes\Sigma)^{-1}(\mathrm{vec}(\mathcal{Z}_i)\mathrm{vec}(\mathcal{Z}_i)^\top - w(\mathrm{vec}(\mathcal{Z}_i)\mathrm{vec}(\mathbf{A})^\top + \mathrm{vec}(\mathbf{A})\mathrm{vec}(\mathcal{Z}_i)^\top))\\
& -\frac{1}{2}\mathrm{tr}((\Psi\otimes\Sigma)^{-1}(w^2\mathrm{vec}(\mathbf{A})\mathrm{vec}(\mathbf{A})^\top)).
\end{aligned}$$

Remark. By virtue of the definition of $\mathcal{Z} := \mathcal{X} - \mathbf{M}$ and the conventions defined above, both expressions $\mathrm{vec}(\widehat{w\mathcal{Z}}_i)^{(k+1)} := \mathbb{E}_{W,\mathrm{vec}(\mathcal{X}_i)}(W_i\mathrm{vec}(\mathcal{Z}_i) \mid \mathrm{vec}(\widetilde{\mathcal{Y}}_i), \mathrm{vec}(\widetilde{\mathcal{C}}_i), \widehat{\vartheta}^{(k)})$ and $\mathrm{vec}(\widehat{\mathcal{Z}}_i^2)^{(k+1)} := \mathbb{E}_{W,\mathrm{vec}(\mathcal{X}_i)}(\mathrm{vec}(\mathcal{Z}_i)\mathrm{vec}(\mathcal{Z}_i)^\top \mid \mathrm{vec}(\widetilde{\mathcal{Y}}_i), \mathrm{vec}(\widetilde{\mathcal{C}}_i), \widehat{\vartheta}^{(k)})$ have well known formulas. To be more precise, it can be concluded that:

$$\mathrm{vec}(\widehat{w\mathcal{Z}}_i)^{(k)} = \mathrm{vec}(\widehat{w\mathcal{X}}_i)^{(k)} - \widehat{w}_i^{(k)}\mathrm{vec}(\mathbf{M}^{(k)})$$

Likewise, it can also be claimed that:

$$\begin{aligned}
\mathrm{vec}(\widehat{\mathcal{Z}}_i^2)^{(k)} = & \mathrm{vec}(\widehat{\mathcal{X}}_i^2)^{(k)} - \mathrm{vec}(\widehat{\mathcal{X}}_i^{(k)})\mathrm{vec}(\mathbf{M}^{(k)})^\top\\
& - \mathrm{vec}(\mathbf{M}^{(k)})\mathrm{vec}(\widehat{\mathcal{X}}_i^{(k)})^\top + \mathrm{vec}(\mathbf{M}^{(k)})\mathrm{vec}(\mathbf{M}^{(k)})^\top.
\end{aligned}$$

E-step: Given the current estimate $\widehat{\vartheta}^{(k)} = \{\widehat{\mathbf{M}}^{(k)}, \widehat{\mathbf{A}}^{(k)}, \widehat{\Sigma}^{(k)}, \widehat{\Psi}^{(k)}\}$ in the k-th step of the ECM algorithm, the E-step provides the conditional expectation of the complete data log-likelihood function (Q-function), whose expression is given by:

$$Q(\vartheta \mid \widehat{\vartheta}^{(k)}) = \mathbb{E}(\ell_c(\vartheta) \mid \widetilde{\mathcal{Y}}, \widetilde{\mathcal{C}}, \widehat{\vartheta}^{(k)}) = \sum_{i=1}^{n} Q_i(\vartheta \mid \widehat{\vartheta}^{(k)}), \tag{5.3}$$

where each Q_i corresponds to:

$$Q_i(\vartheta \mid \widehat{\vartheta}^{(k)}) = -\frac{q}{2}\log|\Sigma| - \frac{p}{2}\log|\Psi| - \frac{1}{2}\mathrm{tr}\{(\Psi\otimes\Sigma)^{-1}\widehat{\Delta}_i^{(k)}\},$$

and the expression of $\widehat{\Delta}_i^{(k)}$ is given by:

$$\widehat{\Delta}_i^{(k)} = \mathrm{vec}(\widehat{\mathcal{Z}}_i^2)^{(k)} - \mathrm{vec}(\widehat{w\mathcal{Z}}_i)^{(k)}\mathrm{vec}^\top(\mathbf{A}) - \mathrm{vec}(\mathbf{A})\,\mathrm{vec}^\top(\widehat{w\mathcal{Z}}_i)^{(k)} + \widehat{w}_i^{2(k)}\mathrm{vec}(\mathbf{A})\mathrm{vec}^\top(\mathbf{A}),$$

Lemma 14. To start with, let us convention that $\mathbf{m} := \mathbb{E}(\mathrm{vec}(\mathcal{Z}_i) - w\,\mathrm{vec}(\mathbf{A}) \mid \widetilde{\mathcal{Y}}_i, \widetilde{\mathcal{C}}_i, \vartheta^{(k)})$ and $\mathbf{S} := \mathrm{Cov}(\mathrm{vec}(\mathcal{Z}_i) - w\,\mathrm{vec}(\mathbf{A}) \mid \widetilde{\mathcal{Y}}_i, \widetilde{\mathcal{C}}_i, \vartheta^{(k)}))$. Under the MVSNC model assumption, the matrix $\widehat{\Delta}_i^{(k)}$ of dimension $pq \times pq$ is symmetric and positive definite iff $\mathbf{S}$ is positive definite or $\mathbf{v}^\top\mathbf{m} \neq 0$ whenever $\mathbf{v} \neq \mathbf{0}$.

Proof. To begin with, note that we can rewrite the matrix $\widehat{\Delta}_i^{(k)}$ as follows:

$$\widehat{\Delta}_i^{(k)} = \mathbb{E}((\mathrm{vec}(\mathcal{Z}_i) - w\,\mathrm{vec}(\mathbf{A}))(\mathrm{vec}(\mathcal{Z}_i) - w\,\mathrm{vec}(\mathbf{A}))^\top \mid \widetilde{\mathcal{Y}}_i, \widetilde{\mathcal{C}}_i, \vartheta^{(k)}) = \mathbf{S} + \mathbf{m}\mathbf{m}^\top.$$

from which it can be concluded that $\widehat{\Delta}_i^{(k)}$ is always symmetric and positive semi-definite. Consequently, the matrix $\widehat{\Delta}_i^{(k)}$ is positive definite iff for every $\mathbf{v} \in \mathbb{R}^{pq}\backslash\{0\}$, one has that:

$$\mathbf{v}^\top\widehat{\Delta}_i^{(k)}\mathbf{v} = \mathbf{v}^\top\mathbf{S}\mathbf{v} + \mathbf{v}^\top\mathbf{m}\mathbf{m}^\top\mathbf{v} = \mathbf{v}^\top\mathbf{S}\mathbf{v} + (\mathbf{v}^\top\mathbf{m})^2 > 0$$

This relation, by its turn, occurs iff $\mathbf{v}^\top\mathbf{S}\mathbf{v} > 0$ or $\mathbf{v}^\top\mathbf{m} \neq 0$. □

Remark. Because $\widehat{\Delta}_i^{(k)}$ is not always strictly positive definite, at each ECM iteration the intermediate matrix $\widehat{\Delta}_i^{(k)}$ is projected onto the cone of symmetric positive semidefinite matrices using the `nearPD` function from the `Matrix` package in R. This function, which implements the algorithm of Higham (1988), computes the nearest positive semidefinite matrix to a given symmetric input in the Frobenius-norm sense. Although the `nearPD` algorithm formally yields the nearest symmetric positive semidefinite matrix, in practice this projection can be safely regarded as positive definite when small negative eigenvalues (typically of order 10^{-16}) arise from floating-point errors. This adjustment is numerically justified and widely adopted in EM-type algorithms, as it ensures the feasibility of the Cholesky decomposition required for updating the scale matrices Σ and Ψ, without compromising the statistical validity of the parameter estimates. In the highly exceptional case where $\widehat{\Delta}_i^{(k)}$ is exactly singular, the Cholesky factorization may fail. To handle such occurrences, a `tryCatch` mechanism was implemented to prevent premature termination of the algorithm: whenever singularity was detected, a new data sample was generated for that replicate. This safeguard enhances the numerical robustness of the estimation procedure without introducing bias into the overall Monte Carlo results. Such corrective projections are widely employed in EM-type algorithms and Gaussian process models to preserve numerical stability when covariance structures are theoretically positive semidefinite but may lose definiteness due to finite-precision arithmetic (see, for example, Higham (1988), Seeger (2004), Boyd and Vandenberghe (2004)).

Theorem 5. At the k-th step of the EM algorithm, the Q-function can be expressed as:

$$Q(\vartheta \mid \widehat{\vartheta}^{(k)}) = -\frac{np}{2}\log|\Psi| - \frac{nq}{2}\log|\Sigma| - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{pq}\operatorname{tr}\left[\Sigma^{-1}\widehat{\mathbf{B}}_{ij}^{(k)}\Psi^{-1}\widehat{\mathbf{B}}_{ij}^{\top(k)}\right], \tag{5.4}$$

where $\widehat{\mathbf{B}}_{ij}^{(k)}$ is a $p \times q$ matrix such that $\operatorname{vec}(\widehat{\mathbf{B}}_{ij}^{(k)}) = \widehat{\mathbf{L}}_{ij}^{(k)}$. 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 $\widehat{\Delta}_{i}^{(k)}$ as previously mentioned and defined.

Proof. First, recall that $\widehat{\Delta}_{i}^{(k)} = \widehat{\mathbf{L}}_{i}^{(k)}\widehat{\mathbf{L}}_{i}^{(k)\top}$ since $\widehat{\Delta}_{i}^{(k)}$ is a positive definite matrix. Thus:

$$\begin{aligned}
Q(\vartheta \mid \widehat{\vartheta}^{(k)}) &= -\frac{np}{2}\log|\Psi| - \frac{nq}{2}\log|\Sigma| - \frac{1}{2}\sum_{i=1}^{n}\operatorname{tr}\left[(\Psi\otimes\Sigma)^{-1}\widehat{\Delta}_{i}^{(k)}\right] \\
&= -\frac{np}{2}\log|\Psi| - \frac{nq}{2}\log|\Sigma| - \frac{1}{2}\sum_{i=1}^{n}\operatorname{tr}\left[(\Psi\otimes\Sigma)^{-1}\widehat{\mathbf{L}}_{i}^{(k)}\widehat{\mathbf{L}}_{i}^{(k)\top}\right] \\
&= -\frac{np}{2}\log|\Psi| - \frac{nq}{2}\log|\Sigma| - \frac{1}{2}\sum_{i=1}^{n}\operatorname{tr}\left[\widehat{\mathbf{L}}_{i}^{(k)\top}(\Psi\otimes\Sigma)^{-1}\widehat{\mathbf{L}}_{i}^{(k)}\right] \\
&= -\frac{np}{2}\log|\Psi| - \frac{nq}{2}\log|\Sigma| - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{pq}\left[\widehat{\mathbf{L}}_{ij}^{(k)\top}(\Psi\otimes\Sigma)^{-1}\widehat{\mathbf{L}}_{ij}^{(k)}\right],
\end{aligned}$$

where the last equality holds because $\widehat{\mathbf{L}}_{ij}^{(k)}$ is the jth column of the $pq \times pq$ lower triangular matrix $\widehat{\mathbf{L}}_{i}^{(k)}$. Using standard identities involving the vec operator, the trace, and the Kronecker product, such as the key relation

$$\operatorname{tr}(\mathbf{PQRS}) = \operatorname{vec}(\mathbf{P}^{\top})^{\top}(\mathbf{S}^{\top}\otimes\mathbf{Q})\operatorname{vec}(\mathbf{R}).$$

It follows that

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

which concludes the proof. □

The updated Q-function in (5.4), when reformulated in terms of the Cholesky decomposition of the matrix $\widehat{\Delta}_{i}^{(k)}$, closely mirrors the structure of the corresponding function in the multivariate normal (MVN) framework discussed by Glanz and Carvalho (2018). In that setting, analytical solutions for the model parameters are explicitly derived in Glanz and Carvalho (2018). Building on this analogy, Theorem 5 not only facilitates the simplification of the required matrix derivatives with respect to Σ and Ψ but also ensures that their closed-form solutions can be obtained in a straightforward and computationally efficient manner, as detailed in the subsequent discussion.

CM-step 1: The conditional maximization of the function $Q(\vartheta \mid \widehat{\vartheta}^{(k)})$, as defined in Theorem 5, with respect to the covariance matrices Σ and Ψ, leads — following the same reasoning as in Glanz and Carvalho (2018) — to the update equations presented below.

$$\widehat{\Sigma}^{(k+1)} = \frac{1}{nq}\sum_{i=1}^{n}\sum_{j=1}^{pq}\widehat{\mathbf{B}}_{ij}^{(k)}\widehat{\Psi}^{(k)^{-1}}\widehat{\mathbf{B}}_{ij}^{(k)^{\top}}, \tag{5.5}$$

$$\widehat{\Psi}^{(k+1)} = \frac{1}{np}\sum_{i=1}^{n}\sum_{j=1}^{pq}\widehat{\mathbf{B}}_{ij}^{(k)^{\top}}\widehat{\Sigma}^{(k)^{-1}}\widehat{\mathbf{B}}_{ij}^{(k)}. \tag{5.6}$$

Note that, to satisfy the identifiability constraint $|\Psi| = 1$ in the estimation process, the estimator of Ψ in (5.6) is then replaced by:

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

Remark. Before deriving the update equation for the location matrix $\mathbf{M}$ and the skewness matrix $\mathbf{A}$, we first need to unwrap the argument of the trace function appearing in the expression of $\ell_{ic}(\vartheta)$. More precisely, we have:

$$\begin{aligned}
\Sigma^{-1}(\mathcal{Z}_i - w\mathbf{A})\Psi^{-1}(\mathcal{Z}_i - w\mathbf{A})^{\top} &= \Sigma^{-1}(\mathcal{X}_i - \mathbf{M} - w\mathbf{A})\Psi^{-1}(\mathcal{X}_i - \mathbf{M} - w\mathbf{A})^{\top} \\
&= \Sigma^{-1}\mathcal{X}_i\Psi^{-1}\mathcal{X}_i^{\top} - \Sigma^{-1}\mathcal{X}_i\Psi^{-1}\mathbf{M}^{\top} - \Sigma^{-1}(w\mathcal{X}_i)\Psi^{-1}\mathbf{A}^{\top} \\
&- \Sigma^{-1}\mathbf{M}\Psi^{-1}\mathcal{X}_i^{\top} + \Sigma^{-1}\mathbf{M}\Psi^{-1}\mathbf{M}^{\top} + w\Sigma^{-1}\mathbf{M}\Psi^{-1}\mathbf{A}^{\top} \\
&- \Sigma^{-1}\mathbf{A}\Psi^{-1}(w\mathcal{X}_i)^{\top} + w\Sigma^{-1}\mathbf{A}\Psi^{-1}\mathbf{M} + w^2\Sigma^{-1}\mathbf{A}\Psi^{-1}\mathbf{A}^{\top}.
\end{aligned}$$

Based on this relation, we are now in a position to compute the conditional expectation of the complete-data log-likelihood function ℓ_c and subsequently maximize it with respect to the location matrix $\mathbf{M}$ and the skewness matrix $\mathbf{A}$, respectively.

CM-step 2: To obtain the update equation of the location matrix $\mathbf{M}$, notice that

$$\begin{aligned}
\ell_{ic}(\vartheta) \propto\ & \operatorname{tr}(\Sigma^{-1}\mathcal{X}_i\Psi^{-1}\mathbf{M}^{\top}) \\
&+ \operatorname{tr}(\Sigma^{-1}\mathbf{M}\Psi^{-1}\mathcal{X}_i^{\top}) \\
&- \operatorname{tr}(\Sigma^{-1}\mathbf{M}\Psi^{-1}\mathbf{M}^{\top}) \\
&- w\operatorname{tr}(\Sigma^{-1}\mathbf{M}\Psi^{-1}\mathbf{A}^{\top}) \\
&- w\operatorname{tr}(\Sigma^{-1}\mathbf{A}\Psi^{-1}\mathbf{M})
\end{aligned}$$

Therefore, since the linear operator vec is invertible, the knowledge about the conditional expected value of $\mathcal{X}_i$ and $w\mathcal{X}_i$ are known. Having said that, due to the linearity of the trace

function as well as the linearity of the operator $\mathbb{E}$, it results that:

$$\begin{aligned} Q_i(\vartheta \mid \widehat{\vartheta}^{(k)}) \propto & \operatorname{tr}(\Sigma^{-1}\widehat{\mathcal{X}}_i^{(k+1)}\Psi^{-1}\mathbf{M}^\top) \\ & + \operatorname{tr}(\Sigma^{-1}\mathbf{M}\Psi^{-1}(\widehat{\mathcal{X}}_i^{(k+1)})^\top) \\ & - \operatorname{tr}(\Sigma^{-1}\mathbf{M}\Psi^{-1}\mathbf{M}^\top) \\ & - \widehat{w}_i^{(k+1)} \operatorname{tr}(\Sigma^{-1}\mathbf{M}\Psi^{-1}\mathbf{A}^\top) \\ & - \widehat{w}_i^{(k+1)} \operatorname{tr}(\Sigma^{-1}\mathbf{A}\Psi^{-1}\mathbf{M}) \end{aligned}$$

Thus, after differentiating the Q-function with respect to $\mathbf{M}$ and solving the corresponding equation in terms of the updated location matrix, one may conclude that:

$$\widehat{\mathbf{M}}^{(k+1)} = \frac{1}{n}\sum_{i=1}^{n}(\widehat{\mathcal{X}}_i^{(k+1)} - \widehat{w}_i^{(k+1)}\mathbf{A}^{(k)})$$

CM-step 3: Analogously, in order to obtain the update equation of the skewness matrix $\mathbf{A}$, we first need to notice that:

$$\begin{aligned} Q_i(\vartheta \mid \widehat{\vartheta}^{(k)}) \propto & \operatorname{tr}(\Sigma^{-1}(\widehat{w\mathcal{X}_i})^{(k+1)}\Psi^{-1}\mathbf{A}^\top) \\ & - \widehat{w}_i^{(k+1)} \operatorname{tr}(\Sigma^{-1}\mathbf{M}\Psi^{-1}\mathbf{A}^\top) \\ & + \operatorname{tr}(\Sigma^{-1}\mathbf{A}\Psi^{-1}(\widehat{w\mathcal{X}_i})^{(k+1)})^\top \\ & - \widehat{w}_i^{(k+1)} \operatorname{tr}(\Sigma^{-1}\mathbf{A}\Psi^{-1}\mathbf{M}) \\ & - \widehat{w}_i^{2(k+1)} \operatorname{tr}(\Sigma^{-1}\mathbf{A}\Psi^{-1}\mathbf{A}^\top). \end{aligned}$$

Hence, after differentiating the Q-function with respect to $\mathbf{A}$ and solving the corresponding equation in terms of the updated skewness matrix, it results that:

$$\widehat{\mathbf{A}}^{(k+1)} = \left(\sum_{i=1}^{n}\widehat{w}_i^{2(k+1)}\right)^{-1}\sum_{i=1}^{n}\left\{(\widehat{w\mathcal{X}_i})^{(k+1)} - \widehat{w}_i^{(k+1)}\mathbf{M}^{(k)}\right\}.$$

Remark. Finally, we stopped the algorithm when

$$\left|\frac{\ell(\widehat{\vartheta}^{(k+1)} \mid \operatorname{vec}(\widetilde{\mathcal{V}}), \operatorname{vec}(\widetilde{\mathcal{C}}))}{\ell(\widehat{\vartheta}^{(k)} \mid \operatorname{vec}(\widetilde{\mathcal{V}}), \operatorname{vec}(\widetilde{\mathcal{C}}))} - 1\right| < \varepsilon$$

with $\varepsilon = 10^{-6}$, that is to say, the algorithm stops when the relative distance between two successive evaluations of the log-likelihood is less than the tolerance.

Although we have not explicitly mentioned in the body of the text, the E-step function $Q(\vartheta \mid \vartheta^{(k)})$ is concave with respect to any of its parameters when the others are held fixed, which ensures the existance of a unique global maximum at each step. This is justified by the same arguments as proven for the uncensored model MVSN.

CHAPTER

6

MVSNC: SIMULATIONS AND APPLICATION

This chapter investigates the Matrix-Variate Skew-Normal Censored (MVSNC) model through simulation studies and real-data analysis. The first section focuses on fitting the MVSNC model to simulated data under censoring, examining the model's ability to recover skewness and covariance structures in controlled settings. The second section applies the MVSNC model to real data with missing or censored entries, illustrating its practical utility in uncovering latent patterns and handling incomplete observations. Together, these sections highlight the MVSNC model's flexibility and effectiveness in modeling matrix-variate data subject to censoring.

6.1 Simulation Studies

To assess the finite-sample performance of the proposed estimators, we have carried out an extensive simulation study using data generated from a matrix-variate skew-normal (MVSN) distribution with dimensions $p = 3$ and $q = 4$. The true location and skewness matrices are:

$$\mathbf{M} = \begin{bmatrix} 1.00 & 2.00 & 1.00 & 2.00 \\ 2.00 & 2.00 & 1.00 & 1.00 \\ 3.00 & 3.00 & 2.00 & 3.00 \end{bmatrix}, \quad \mathbf{A} = \begin{bmatrix} -1.00 & -2.00 & -1.00 & -2.00 \\ 2.00 & -2.00 & 1.00 & -1.00 \\ 3.00 & -3.00 & 2.00 & -3.00 \end{bmatrix},$$

whereas the true row covariance and column covariance matrices were defined as

$$\Sigma = \begin{bmatrix} 1.50 & 0.60 & 0.24 \\ 0.60 & 1.50 & 0.60 \\ 0.24 & 0.60 & 1.50 \end{bmatrix} \quad \Psi = \begin{bmatrix} 2.151 & 1.721 & 1.377 & 1.101 \\ 1.721 & 2.151 & 1.721 & 1.377 \\ 1.377 & 1.721 & 2.151 & 1.721 \\ 1.101 & 1.377 & 1.721 & 2.151 \end{bmatrix}.$$

For each simulated dataset, we randomly selected a percentage $c \in \{5\%, 15\%, 30\%\}$ of the total pqn elements and replaced them following three distinct censoring or missing-data scenarios ($s \in \{1,2,3\}$), in order to assess the impact of different levels and patterns of incomplete data on model performance:

1. Only interval-censored values: A proportion `cens` of the smallest observations was selected using the empirical quantile of the simulated sample. For each selected value x, the exact observation was replaced by a symmetric interval $[L,U] = [x-2s, x+2s]$ where s denotes the empirical standard deviation of the censored subset.

2. Only missing values: All selected entries were substituted with $\pm\infty$.

3. Mixed (50% missing): Half of the selected entries were replaced using Scenario 2, and the remaining half followed Scenario 1.

We examined three sample sizes, $n \in \{100, 200, 400\}$, leading to $3 \times 3 \times 3 = 27$ simulation settings from all combinations of (c, s, n). For each setting, 200 independent datasets were generated. To evaluate estimation accuracy, we employed the Frobenius norm as a metric to measure the discrepancy between the true parameter values and their corresponding estimates $(\widehat{\mathbf{M}}, \widehat{\mathbf{A}}, \widehat{\Sigma}, \widehat{\Psi})$ derived from the MVSNC model applied to each simulated dataset.

The initial model comparison employed a baseline approach in which all incomplete observations were removed from the dataset prior to analysis. Following this deletion, the standard MVSN model was fitted to the remaining complete cases, providing a benchmark to evaluate the relative efficiency and performance of the proposed MVSNC method against this traditional deletion-based strategy.

The second model comparison followed the methodological framework proposed by the reference Lachos *et al.* (2025b), where the parameters of the MVN distribution are estimated in the presence of censored data. Specifically, we generated censored samples from the matrix variate normal inverse gaussian (MVNIG) distribution according to the formulation of Gallaugher and McNicholas (2019), and then computed the BIC values for both the MVSNC and MVNC models. The results consistently showed that, across all simulated scenarios, the MVSNC model outperformed the MVNC model, demonstrating its robustness and flexibility in modeling censored matrix-variate data with skewness and moderately heavy-tailed behavior.

6.1.1 Comparing the MVSNC and MVSN Models

We begin by noting that the parameter initialization adopted in the simulation study follows the procedure described in Correia, Davila and Diniz (2025). The plots below present the estimation errors for the MVSNC and MVSN models across all combinations of censoring levels (c) and sample sizes (n) considered in the three simulation scenarios. For each configuration, the boxplots summarize the empirical bias and variability of the parameter estimates based on 200 independent replications, capturing both the central tendency and the dispersion across runs. These visualizations provide a comprehensive assessment of model stability and accuracy, clearly illustrating the impact of censoring intensity and sample size on estimation reliability.

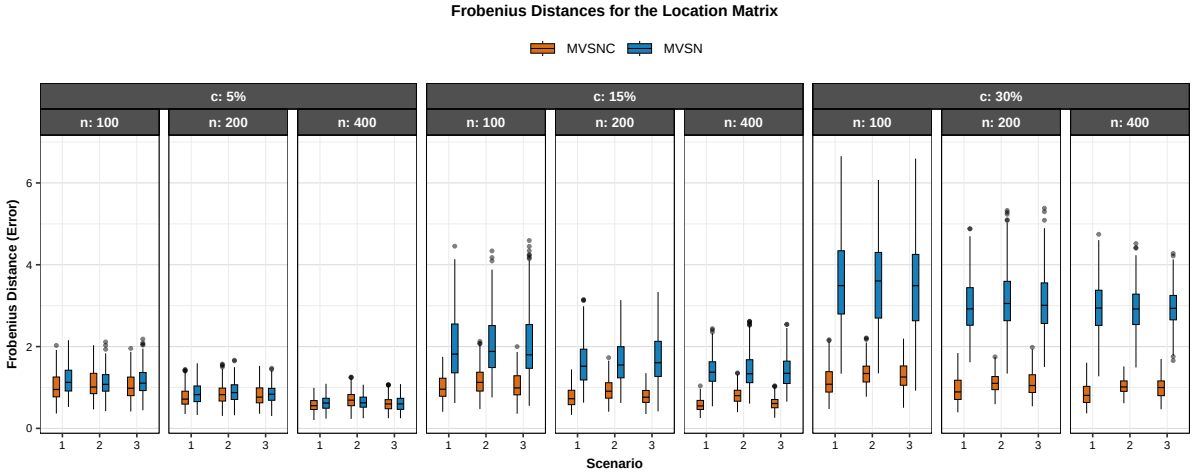


Figure 7 – Boxplots of the Frobenius distances for the location matrix **M**

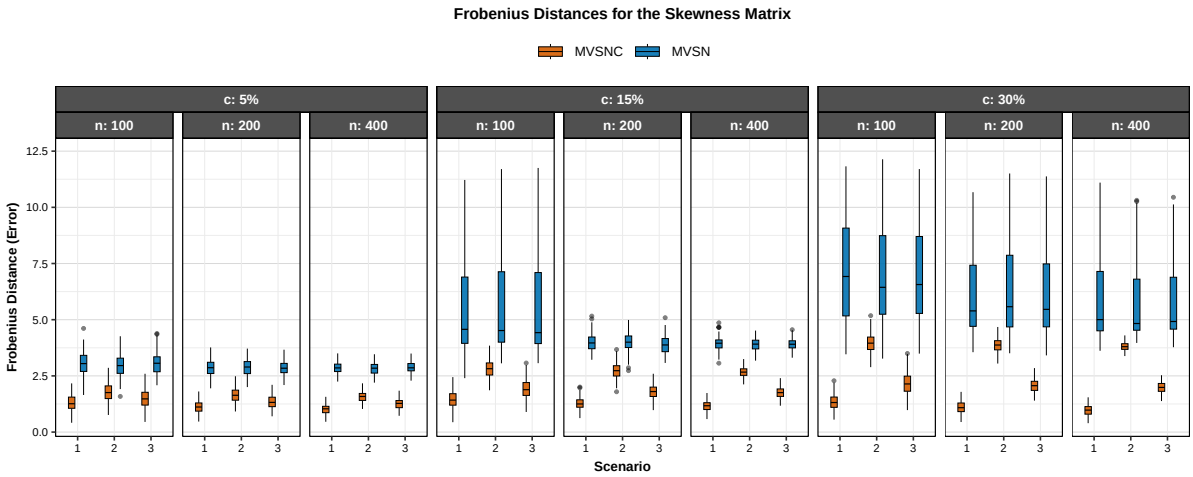


Figure 8 – Boxplots of the Frobenius distances for the skewness matrix **A**

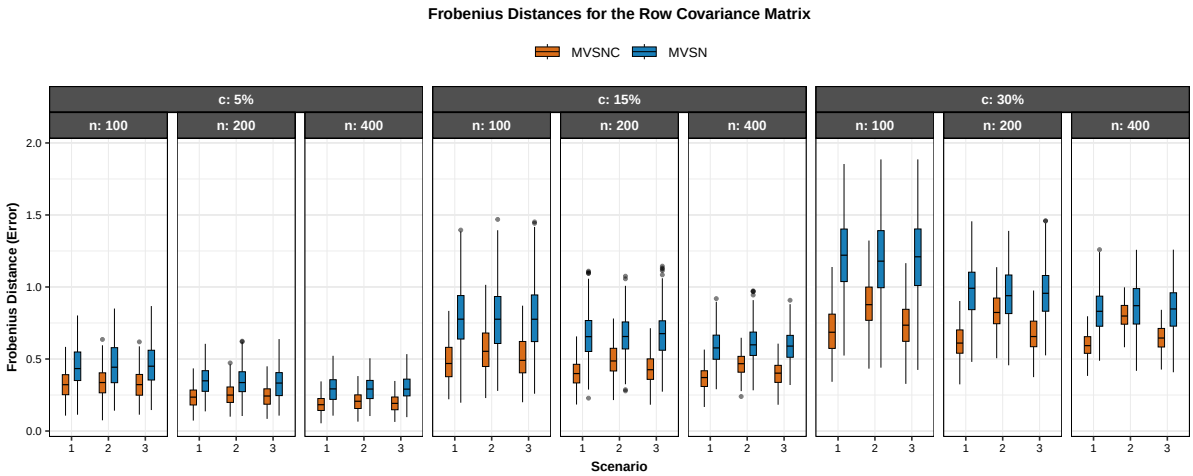


Figure 9 – Boxplots of the Frobenius distances for the row covariance matrix Σ

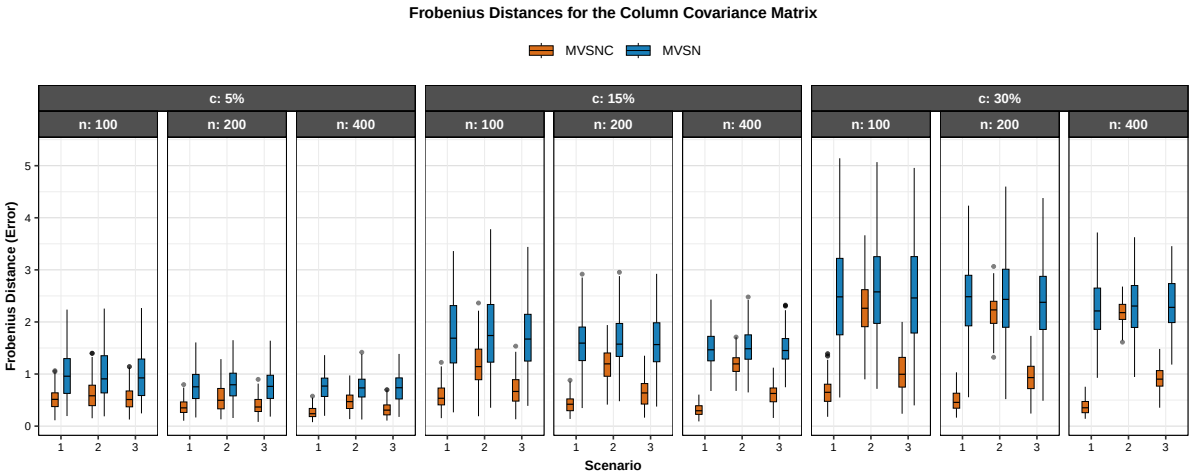


Figure 10 – Boxplots of the Frobenius distances for the column covariance matrix Ψ

Several important trends can be observed from these results. First, the MVSNC model consistently outperforms the MVSN model across all estimated parameters and simulation settings. This advantage stems from the fact that the MVSN model ignores incomplete observations, which effectively reduces the usable sample size and weakens the accuracy and efficiency of its estimates. By contrast, the MVSNC model leverages the information contained in both censored and observed data, yielding parameter estimates that are more precise, stable, and reliable, even when a substantial portion of the data is incomplete.

Another important observation relates to how the performance gap between the two models varies across different scenarios. For a given (c,n) combination, the MVSN model generally delivers similar results across all scenarios. On the other hand, the MVSNC model shows notably smaller estimation errors, especially in scenarios involving only censoring. This highlights the MVSNC model's robustness in handling datasets affected solely by censoring.

The accuracy of the estimates improves steadily as the sample size n grows, confirming the asymptotic consistency and dependability of the proposed estimators. With larger n, the MVSNC model becomes more capable of utilizing the information from incomplete observations to recover the underlying data structure. In contrast, the MVSN model remains affected by the loss of information from discarded cases, leading to increased variability, higher estimation errors, and lower efficiency, especially when the sample size is small or the degree of censoring is high.

Figure 11 presents the boxplots illustrating the distribution of the Bayesian Information Criterion (BIC) values obtained for both the MVSNC and MVNC models under a variety of simulation settings, where the datasets were generated according to the MVNIG distribution. The figure clearly demonstrates a consistent pattern across all examined configurations. Specifically, for every combination of the censoring proportion (c) and sample size (n), and across all considered scenarios, the MVSNC model systematically yields lower BIC values than the

MVNC model, which reflects a consistently superior overall fit. Moreover, as the proportion of censored observations (c) increases, the disparity in performance between the two models becomes even more evident, emphasizing the robustness of the MVSNC approach under heavier censoring conditions. This observed behavior aligns with theoretical expectations: while both models incorporate the effects of censoring in their likelihood formulations, the MVSNC model further accommodates distributional asymmetry through its explicit skewness component. As a result, it provides a more flexible and realistic approximation of the underlying data-generating process, ultimately leading to lower BIC values, enhanced model adequacy, and a more accurate reflection of the true structural characteristics of the data.

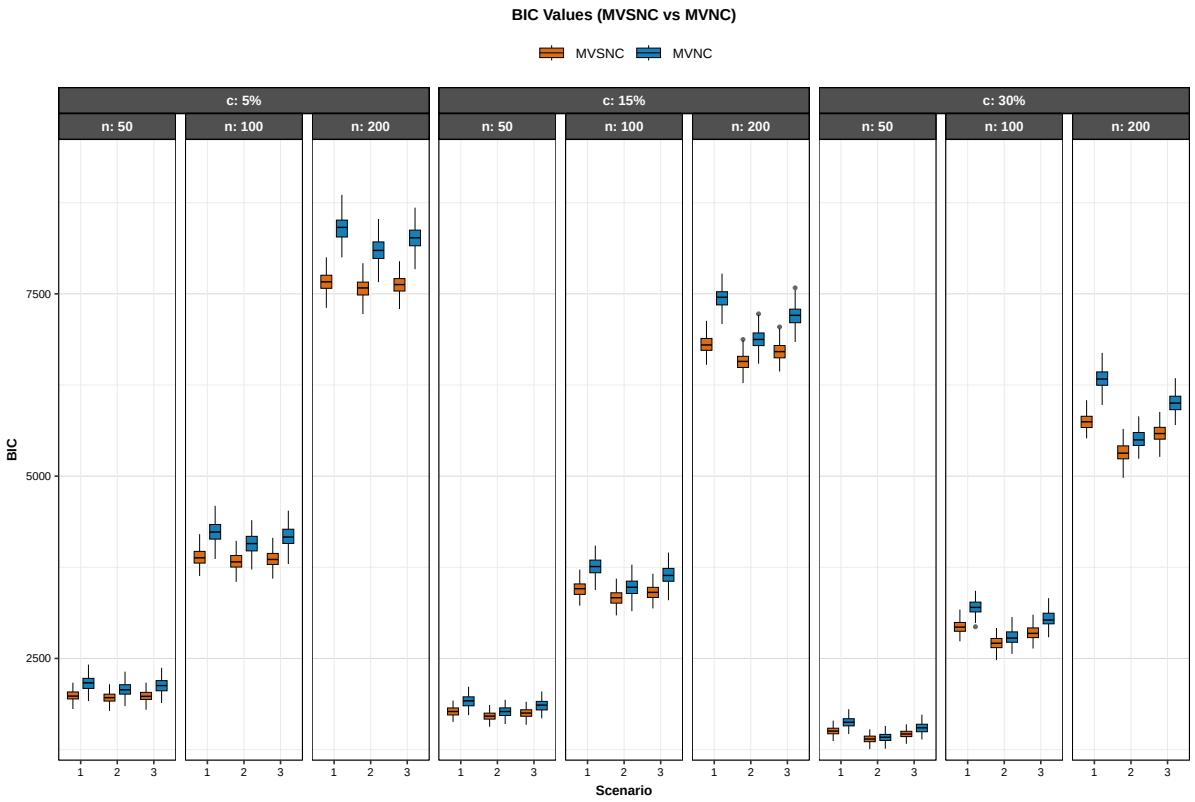


Figure 11 – Boxplots of the BICS used to compare the models MVSNC and MVNC

It is noteworthy that the MVSNC model achieves lower BIC values despite having more parameters, suggesting that the gains in model fit outweigh the BIC's complexity penalty. This pattern is consistent across all simulation settings, highlighting the robustness and stability of the proposed approach under varying levels of censoring and sample sizes. From a theoretical standpoint, these results emphasize the benefits of incorporating skewness into matrix-variate models, as it improves the ability to capture asymmetric patterns, directional effects, and latent heterogeneity commonly observed in real data. In practical terms, this added flexibility leads to more reliable inference and better model adequacy, particularly when censored observations occur alongside asymmetric or moderately heavy-tailed distributions. Overall, these findings reinforce the value of skew-extended models for providing a more accurate, expressive, and informative representation of matrix-valued data.

6.2 Application

To evaluate the empirical performance and practical applicability of the proposed MVSNC model, we conduct an analysis using a real-world environmental dataset focused on water quality measurements in the Chesapeake Bay. Specifically, we employ the `dataCensored` dataset provided in the `baytrends` package for R, which was originally introduced by Murphy *et al.* (2019) and later made publicly available through Murphy *et al.* (2023). This dataset consists of long-term monitoring data that track multiple chemical and nutrient indicators at a series of sampling stations within the Chesapeake Bay Program.

For each monitoring station, a collection of p chemical variables is recorded across q distinct vertical water layers, repeated over n separate sampling occasions. Consequently, the dataset can be naturally represented as a three-way array with dimensions $p \times q \times n$, which aligns well with the matrix-variate modeling framework. This structure enables the explicit modeling of row-wise dependencies that capture spatial or inter-variable correlations among the chemical indicators, as well as column-wise dependencies that reflect vertical stratification within the water column. By leveraging separable covariance matrices for these two dimensions, the model can account for complex correlation patterns in the data while maintaining interpretability and computational tractability.

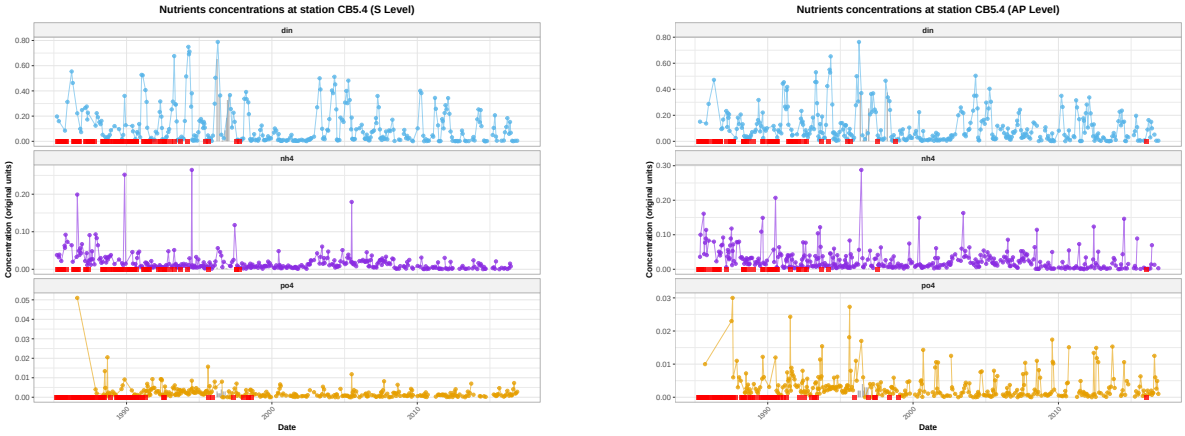
From the eight monitoring stations, we selected **CB5.4** and **EE2.1** due to their contrasting patterns of incomplete data. Station CB5.4 mainly exhibits sporadic missing values, whereas station EE2.1 is largely characterized by interval-censored measurements. This contrast enables a focused evaluation of the MVSNC model's performance and robustness under different types of data loss while preserving the key covariance and skewness structures of the dataset.

Station CB5.4. At this station, we examine $p = 3$ variables: PO_4 (orthophosphorus), DIN (dissolved inorganic nitrogen), and NH_4 (ammonium). The proportions of missing and censored observations are: 14.93% missing and 1.60% censored (PH_4), 13.11% missing and 0.94% censored (DIN), and 7.80% missing and 0.72% censored (NH_4), giving total incompleteness levels of 16.56%, 14.05%, and 8.52%, respectively. Each variable is recorded at $q = 4$ layers — surface (S), above-pycnocline (AP), below-pycnocline (BP), and bottom (B) — over $n = 452$ sampling occasions, forming a $3 \times 4 \times 452$ data array. At this site, the majority of the incomplete data is due to missing observations, whereas censored measurements account for only a small fraction of the overall incompleteness.

Exploratory visualization. To illustrate the temporal patterns, variability, and potential trends of nutrient concentrations over time at station CB5.4, Figures 12a–12d present the time series of phosphate (PO_4), dissolved inorganic nitrogen (DIN), and ammonium (NH_4) across the four water layers: surface (S), above-pycnocline (AP), below-pycnocline (BP), and bottom (B). Each plot clearly highlights the occurrence of censored observations, represented by gray vertical segments, as well as missing observations, indicated by red squares.

(a) Surface (S) layer

(b) Above-pycnocline (AP) layer



(c) Below-pycnocline (BP) layer

(d) Bottom (B) layer

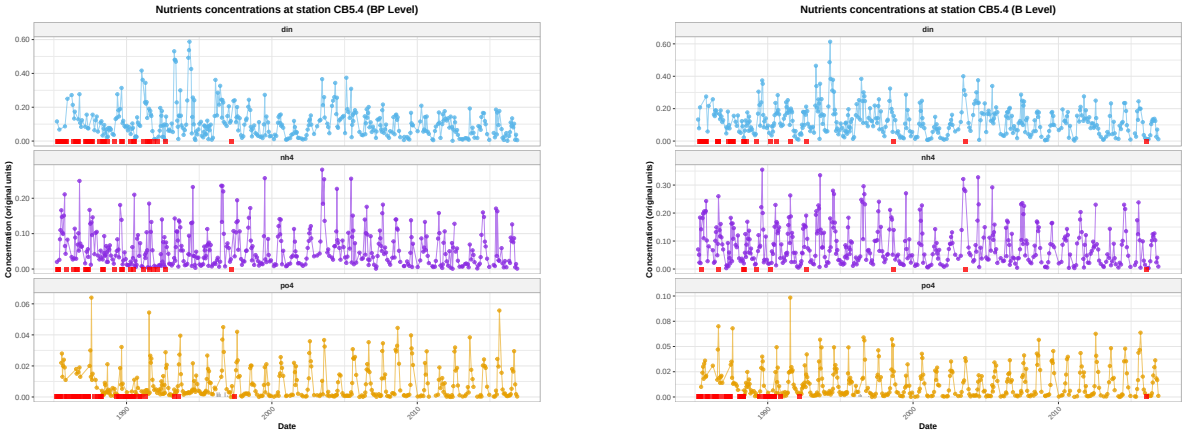


Figure 12 – Temporal evolution of censored nutrient concentrations (PO_4, DIN, and NH_4) at station CB5.4 across the four depth layers, which are represented by the orange, blue and purple colors, respectively. Each panel displays the corresponding time series for a specific layer, with censored and missing values

To illustrate the distribution of nutrient concentrations at station CB5.4, Figure 13 presents the histograms of phosphate (PO_4), dissolved inorganic nitrogen (DIN), and ammonium (NH_4) across the four water layers — surface (S), above-pycnocline (AP), below-pycnocline (BP), and bottom (B). The figure distinguishes observed measurements, interval-censored observations, and missing values using the color scheme defined below.

Station EE2.1. At this station, censoring is more prevalent. We consider $p = 3$ variables: PO_4 (orthophosphorus), TDN (total dissolved nitrogen), and TDP (total dissolved phosphorus). The proportions of missing and censored observations are 1.81% missing and 26.22% censored (PO_4), 5.21% missing and 7.39% censored (TDN), and 5.64% missing and 7.18% censored (TDP), giving total incompleteness levels of 28.03%, 12.60%, and 12.82%, respectively. Each variable is measured at $q = 4$ layers — surface (S), above-pycnocline (AP), below-pycnocline (BP), and bottom (B) — across $n = 470$ sampling occasions, forming a $3 \times 4 \times 470$ dataset. The predominance of censored observations here contrasts with the pattern observed at station CB5.4.

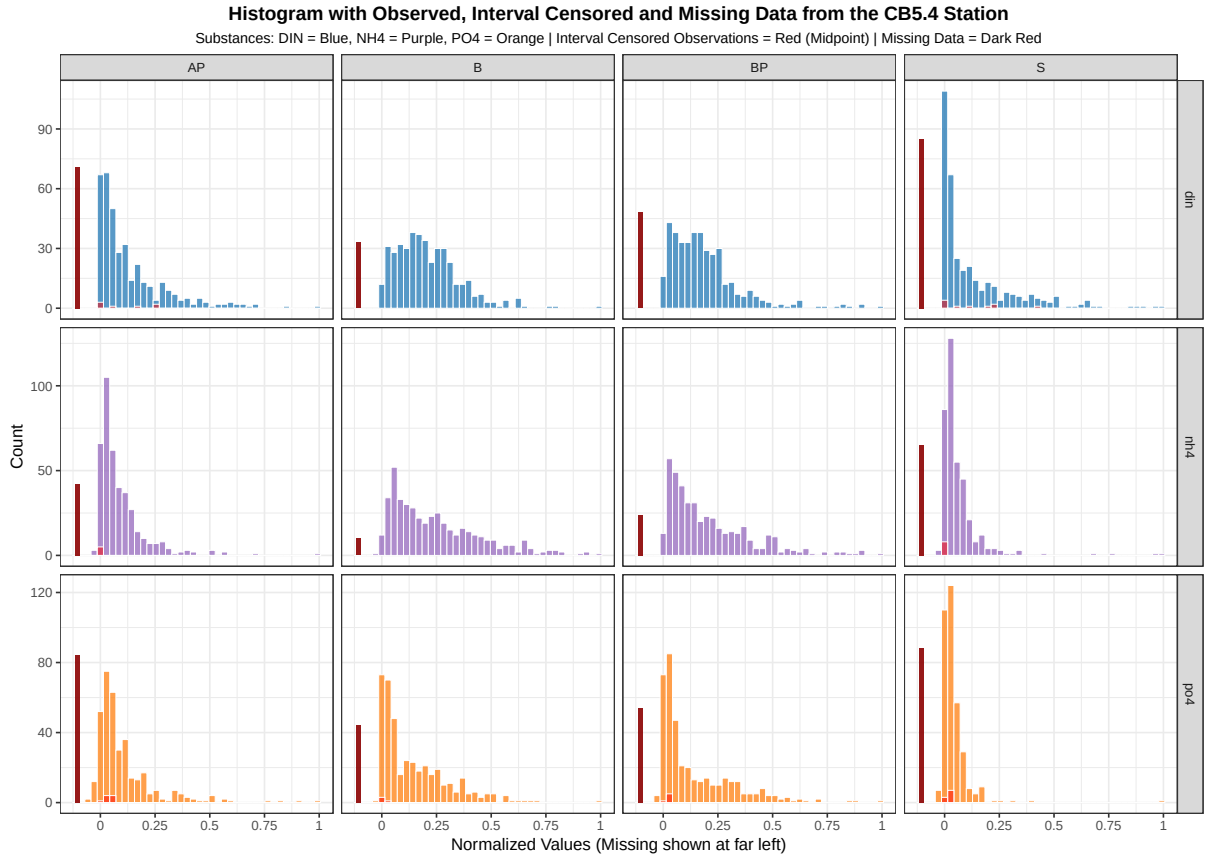
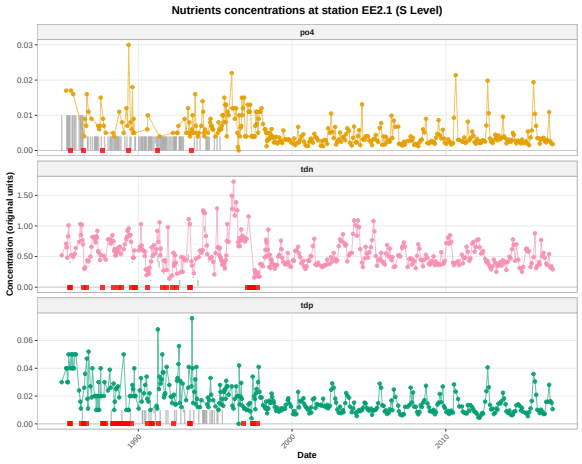


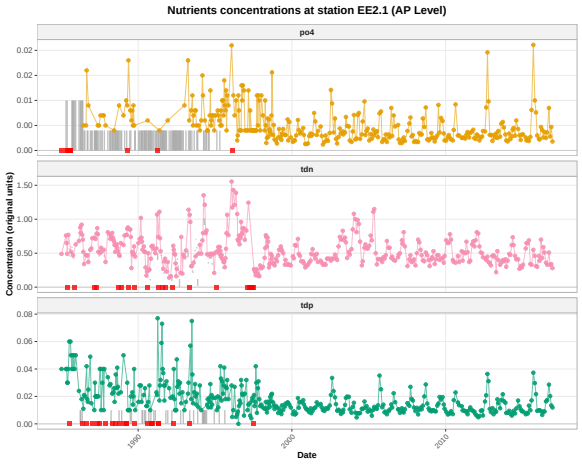
Figure 13 – Histogram showing the distribution of observed nutrient measurements, interval-censored observations, and missing data collected at station CB5.4

Exploratory visualization. Figures 14a–14d illustrate the censored nutrient time series for station EE2.1, focusing on phosphate (PO_4), total dissolved nitrogen (TDN), and total dissolved phosphorus (TDP) across the four depth layers. Each plot indicates the presence of censored observations (shown as gray vertical segments) and missing values (shown as red squares). The figures reveal a considerable proportion of censored measurements, particularly for PO_4, where values often fall below the analytical detection limit. This high prevalence of interval censoring underscores the need for models that explicitly incorporate censoring mechanisms in the likelihood function. Additionally, the co-movement observed between TDN and TDP across layers points to a strong vertical dependence structure, which is effectively captured by the matrix-variate modeling approach.

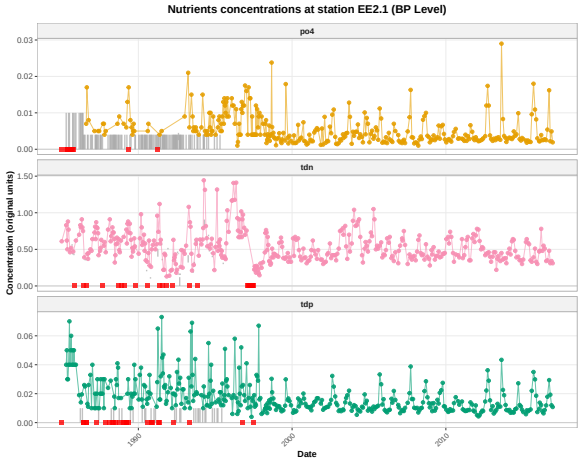
In order to examine the distributional patterns of nutrient concentrations at station EE2.1, Figure 15 presents histograms of phosphate (PO_4), total dissolved nitrogen (TDN), and total dissolved phosphorus (TDP) across the four depth layers. The figure distinguishes observed measurements from interval-censored values and missing data using the established color scheme, enabling a clear assessment of data completeness. A substantial portion of PO_4 measurements falls below the analytical detection limit, resulting in interval censoring. Furthermore, the similarities in the distributions of TDN and TDP across depth layers reflect a vertical dependence structure that is naturally accommodated by the matrix-variate modeling framework.



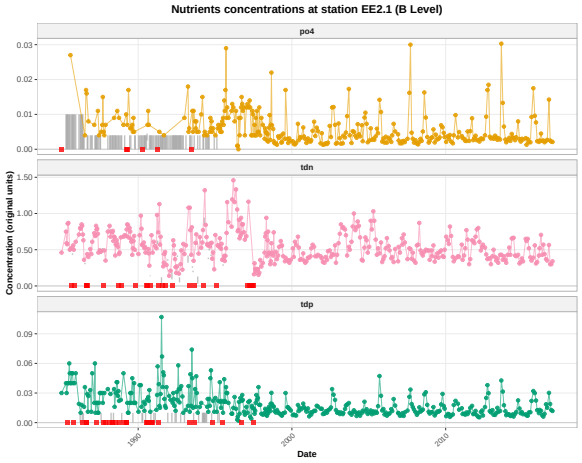
(a) Surface (S) layer



(b) Above-pycnocline (AP) layer



(c) Below-pycnocline (BP) layer



(d) Bottom (B) layer

Figure 14 – Temporal evolution of censored nutrient concentrations (PO_4, TDN, and TDP) at station EE2.1 across the four depth layers, which are represented by the orange, pink and green colors, respectively. Each panel displays the corresponding time series for a specific layer, with censored and missing values

For both stations, the data are organized as random matrices $\mathcal{X}_1, \ldots, \mathcal{X}_n$, with rows representing variables and columns corresponding to sampling layers, where left- and right-censoring result from laboratory detection limits. Each station exhibits a distinct pattern of censored measurements, reflecting variations in depth, chemical properties, and environmental conditions. We analyze the parameter estimates from the Matrix Variate Skew-Normal Censored (MVSNC) model applied to the `dataCensored` dataset, which extends the matrix-variate skew-normal distribution by incorporating censoring indicators within its hierarchical structure. This framework simultaneously models skewness and incomplete observations, yielding more accurate estimates of location and scale parameters and effectively capturing the asymmetric, moderately heavy-tailed behavior typical of environmental data under detection-limit constraints. Moreover, the model coherently represents the dependence structure across variables and layers, enhancing interpretability and robustness in practical applications.

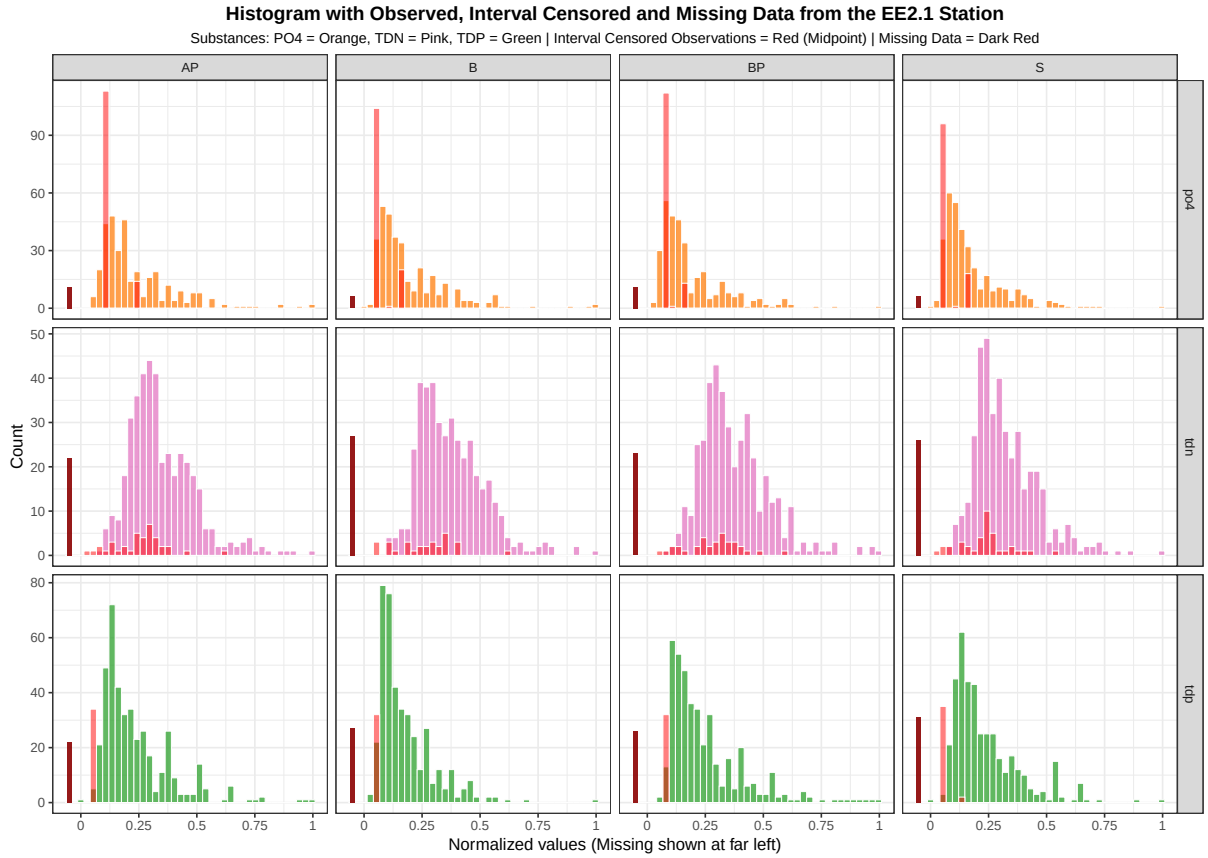


Figure 15 – Histogram illustrating the distribution of observed nutrient measurements, interval-censored observations, and missing data recorded over time at station EE2.1.

Readers interested in a comprehensive discussion of the estimation and interpretation of the parameters $\mathbf{M}_{\text{MVNC}}$, Σ_{MVNC}, and Ψ_{MVNC} for stations EE2.1 and CB5.4 are referred to Lachos *et al.* (2025b). Their work offers both theoretical and empirical insights into how the estimated structures capture spatial and vertical variability under censoring. Building on this framework, we concentrate on interpreting the parameter estimates obtained under the MVSNC model for each station, analyzing their magnitude, spatial configuration, and depth-specific patterns, and discussing how these characteristics reflect the underlying variability present in the observed data.

For the estimation of the location matrix ($\widehat{\mathbf{LM}}$) and the row and column covariance matrices, we employed the expressions introduced in Chapter 3. In this framework, the location matrix is obtained from $\mathbf{M}$ and $\mathbf{A}$ as described in Proposition 12, while the row and column covariance matrices correspond to C_{row} and C_{col}, respectively (see Proposition 24). The location matrix is interpreted following the standard convention, whereas C_{row} and C_{col} provide complementary insights into the dependence structure of the matrix-valued data. Specifically, C_{row} summarizes joint variability across rows, combining a skewness-driven term ($\sigma_W^2\mathbf{A}\mathbf{A}^\top$) with Gaussian row variability ($\text{tr}(\Psi)\Sigma$), while C_{col} reflects co-movements among columns, incorporating both structured dependence induced by $\mathbf{A}$ and underlying Gaussian fluctuations. Together, these matrices capture the principal patterns of dependence observed in practice.

6.2.1 Interpreting the MVSNC Model Estimates

In this section, we provide a detailed presentation and discussion of the results obtained from the application of the Matrix-Variate Skew-Normal with Censoring (MVSNC) model to the interval-censored and missing observations contained in the `dataCensored` dataset. Following the estimation of the model parameters, we thoroughly examine several aspects of the fitted model, including the inferred dependence structure across variables and sampling layers, the skewness patterns identified within the data, and the manner in which the model accommodates and interprets censored observations. The results presented in the following subsections illustrate that the MVSNC framework is capable of effectively capturing not only the main characteristics and inherent variability of the observed data but also the more subtle structural relationships that may exist among variables. Furthermore, the model provides a coherent and comprehensive representation of the underlying multivariate process, demonstrating clear practical advantages when dealing with incomplete datasets, asymmetric distributions, and the complex patterns often encountered in environmental measurements. This comprehensive modeling approach therefore offers both robust parameter estimation and enhanced interpretability, making it particularly suitable for real-world applications where data incompleteness and skewness are prevalent.

Results for the CB5.4 station

Based on the BIC results for both fitted models (MVSNC and MVNC), we observed that the MVSNC model (BIC $= -20765.6$) achieved a lower value than the MVNC model (BIC $= -20602.84$). Because smaller BIC values indicate a better trade-off between model fit and complexity, this finding suggests that the MVSNC formulation offers a more accurate representation of the censored observations in the `baytrends` dataset. The observed improvement in model adequacy motivates a detailed examination of the parameter estimates under the MVSNC model, allowing us to investigate how the explicit incorporation of skewness and censoring mechanisms contributes to more reliable and informative inference in the matrix-variate framework, particularly in the presence of incomplete data.

The location matrix corresponding to the CB5.4 station, estimated as $\widehat{\mathbf{LM}}_{\text{CB5.4}}$ and fully consistent with the results established in Proposition 12, can be expressed as follows:

$$\widehat{\mathbf{LM}}_{\text{CB5.4}} = \left[\begin{array}{c|cccc} & \texttt{S} & \texttt{AP} & \texttt{BP} & \texttt{B} \\ \hline \texttt{po4} & 0.002 & 0.003 & 0.008 & 0.013 \\ \texttt{din} & 0.122 & 0.116 & 0.115 & 0.129 \\ \texttt{nh4} & 0.017 & 0.025 & 0.054 & 0.082 \end{array}\right].$$

The matrix $\widehat{\mathbf{LM}}_{\text{CB5.4}}$ presents the estimated average nutrient concentrations for each of the four depth layers, offering an overview of the vertical variation of key chemical variables within the `dataCensored` dataset. These estimates can be interpreted as follows:

- **Phosphate (PO_4):** The estimated mean values exhibit a clear increasing trend with depth, ranging from approximately 0.002 at the surface (S) to 0.013 at the bottom (B) layer. This pattern suggests that phosphate concentrations rise in deeper waters, a behavior typically associated with remineralization processes and sediment nutrient release.
- **Dissolved Inorganic Nitrogen (DIN):** The DIN estimates remain relatively stable across the water column, varying slightly between 0.115 and 0.129. This homogeneity indicates that nitrogen compounds are more uniformly distributed, possibly due to vertical mixing or balanced biological uptake and regeneration processes.
- **Ammonium (NH_4):** The concentration of ammonium increases markedly with depth, from 0.017 at the surface to 0.082 at the bottom. This gradient reflects the accumulation of NH_4 in deeper layers, consistent with oxygen depletion and organic matter decomposition that enhance ammonium release near the sediment.

Overall, the estimated location matrix $\widehat{\mathbf{LM}}_{\text{CB5.4}}$ reveals coherent ecological patterns consistent with the expected nutrient cycling dynamics of stratified aquatic systems. The results further indicate that the MVSNC model captures realistic spatial variation while maintaining interpretability, even in the presence of censored or missing observations. These findings reinforce the suitability of the MVSNC approach for modeling asymmetric and incomplete environmental data.

The estimated skewness matrix $\widehat{\mathbf{A}}_{\text{CB5.4}}$ is given by:

$$\widehat{\mathbf{A}}_{\text{CB5.4}} = \left[\begin{array}{c|cccc} & \texttt{AP} & \texttt{B} & \texttt{BP} & \texttt{S} \\ \hline \texttt{din} & 0.0731 & 0.0554 & 0.0617 & 0.0832 \\ \texttt{nh4} & 0.0080 & 0.0275 & 0.0148 & 0.0042 \\ \texttt{po4} & 0.0002 & 0.0008 & -0.0001 & 0.0002 \end{array}\right].$$

This matrix represents the estimated skewness parameters associated with each nutrient (rows) across the four depth layers (columns) obtained from the fitted MVSNC model.

- **DIN:** The positive skewness coefficients across all layers (ranging from 0.055 to 0.083) indicate a mild right-skewed behavior in the distribution of dissolved inorganic nitrogen concentrations. This suggests occasional high DIN values relative to the mean, consistent with transient nutrient pulses or episodic enrichment events.
- **NH_4:** The skewness parameters are positive but smaller in magnitude, varying between 0.004 and 0.027. This pattern implies a weak asymmetry, with slightly heavier right tails especially near the bottom (B) layer, possibly reflecting localized ammonium accumulation under low-oxygen conditions.
- **PO_4:** The skewness values are close to zero and include a small negative entry at the BP layer (-0.0001), indicating that phosphate concentrations are nearly symmetric and stable across depths, without pronounced deviations from normality.

The estimated skewness pattern shows that asymmetry is more evident for nitrogen compounds (DIN and NH_4), while phosphate exhibits an almost symmetric behavior. These observations align with the underlying chemical and biological processes in the system: nitrogen-related variables tend to display episodic fluctuations driven by biological uptake and remineralization, whereas phosphate remains relatively stable. The MVSNC model successfully captures both the direction and extent of skewness across nutrients and depth layers, demonstrating its ability to model non-Gaussian characteristics in censored environmental data.

The correlation matrix corresponding to the estimated row covariance matrix $\widehat{C}^{\text{row}}_{\text{CB5.4}}$ is presented below, providing a clear summary of the pairwise relationships among the rows in the dataset:

$$\widehat{R}(\widehat{C}^{\text{row}}_{\text{CB5.4}}) = \left[\begin{array}{c|ccc} & \texttt{po4} & \texttt{din} & \texttt{nh4} \\ \hline \texttt{po4} & 1.000 & 0.394 & 0.577 \\ \texttt{din} & 0.394 & 1.000 & 0.522 \\ \texttt{nh4} & 0.577 & 0.522 & 1.000 \end{array}\right].$$

The estimated correlation structure of $\widehat{R}(\widehat{C}^{\text{row}}_{\text{CB5.4}})$ provides insights into the linear dependence among the nutrient variables (PO_4, DIN, and NH_4) at station CB5.4. All correlation coefficients are positive, indicating that increases in one nutrient tend to be associated with increases in the others.

- The strongest association is observed between PO_4 and NH_4 ($r = 0.577$), suggesting that phosphate and ammonium concentrations vary together across the samples. This relationship likely reflects their common origin from organic matter remineralization processes in deeper or less-oxygenated waters.
- The correlation between DIN and NH_4 ($r = 0.522$) is also moderate, indicating that variations in total dissolved inorganic nitrogen are influenced by fluctuations in ammonium levels, as NH_4 constitutes one of its key components.
- The weakest correlation occurs between PO_4 and DIN ($r = 0.394$), although still positive. This moderate relationship suggests that, while both nutrients are linked through biogeochemical cycling, they may respond to distinct processes or time scales within the aquatic system.

The correlation structure displays a consistent nutrient pattern, with strong connections observed among the different nitrogen forms and moderate links between nitrogen and phosphorus. The positive and moderately high correlations suggest that the estimated $\widehat{R}(\widehat{C}^{\text{row}}_{\text{CB5.4}})$ accurately reflects realistic co-variation among nutrient concentrations, in accordance with ecological expectations for estuarine systems influenced by nutrient recycling and sediment interactions.

The correlation matrix corresponding to the estimated column covariance matrix $\widehat{\mathrm{C}}^{\mathrm{column}}_{\mathrm{CB5.4}}$ is presented below, providing a comprehensive overview of the pairwise relationships among the columns in the dataset:

$$\widehat{R}(\widehat{\mathrm{C}}^{\mathrm{column}}_{\mathrm{CB5.4}}) = \left[\begin{array}{c|cccc} & \mathtt{S} & \mathtt{AP} & \mathtt{BP} & \mathtt{B} \\ \hline \mathtt{S} & 1.000 & 0.854 & 0.474 & 0.357 \\ \mathtt{AP} & 0.854 & 1.000 & 0.571 & 0.442 \\ \mathtt{BP} & 0.474 & 0.571 & 1.000 & 0.824 \\ \mathtt{B} & 0.357 & 0.442 & 0.824 & 1.000 \end{array}\right].$$

The estimated correlation matrix $\widehat{\mathrm{R}}(\widehat{\mathrm{C}}^{\mathrm{column}}_{\mathrm{CB5.4}})$ describes the dependence structure among the four water-depth layers (Surface – `S`, Above-Pycnocline – `AP`, Below-Pycnocline – `BP`, and Bottom – `B`) at station CB5.4. The generally positive and relatively high correlations indicate that nutrient concentrations at different depths are strongly interrelated, reflecting the vertical coherence of the water column.

- The strongest correlation is observed between the `S` and `AP` layers ($r = 0.854$), suggesting that surface and above-pycnocline waters exhibit highly similar nutrient dynamics. This is expected given their close proximity and exposure to similar physical and biological processes such as mixing, primary production, and atmospheric exchange.

- The `BP` and `B` layers are also strongly correlated ($r = 0.824$), indicating a tight coupling between the deeper water masses. This likely arises from sediment–water interactions and vertical diffusion that homogenize nutrient conditions near the bottom.

- Moderate correlations between intermediate and surface layers (e.g., `AP`–`BP`: $r = 0.571$ and `S`–`BP`: $r = 0.474$) reveal some stratification effects, suggesting that while the water column remains partially mixed, vertical gradients still influence nutrient exchange between upper and deeper layers.

- The weakest correlations involve the surface and bottom layers (`S`–`B`: $r = 0.357$), reflecting distinct physicochemical conditions between these depths, possibly driven by differences in oxygen concentration, light availability, and sediment nutrient fluxes.

The correlation pattern across depth layers indicates a system that is both vertically structured and interconnected, where the upper layers are dynamically influenced by mixing and biological processes, while the deeper layers exhibit strong internal consistency governed by benthic dynamics. The estimated $\widehat{\mathrm{R}}(\widehat{\mathrm{C}}^{\mathrm{column}}_{\mathrm{CB5.4}})$ effectively captures the combined features of stratification and inter-layer coupling characteristic of estuarine environments.

Results for the EE2.1 station

Based on the BIC values computed for the two fitted models, the MVSNC model (BIC $= -24180.83$) clearly outperforms the MVNC model (BIC $= -24021.84$). Since smaller BIC values indicate a better compromise between model fit and complexity, this result suggests that the MVSNC formulation provides a more flexible and precise representation of the censored data in the `baytrends` dataset. The improved performance of the MVSNC model further justifies a thorough examination of its parameter estimates to evaluate how the integration of skewness and censoring components enhances inference within the matrix-variate modeling framework.

The estimated location matrix $\widehat{\mathbf{LM}}_{\text{EE2.1}}$ is given by:

$$\widehat{\mathbf{LM}}_{\text{EE2.1}} = \left[\begin{array}{c|cccc} & \texttt{S} & \texttt{AP} & \texttt{BP} & \texttt{B} \\ \hline \texttt{po4} & 0.004 & 0.003 & 0.004 & 0.004 \\ \texttt{tdn} & 0.558 & 0.547 & 0.544 & 0.549 \\ \texttt{tdp} & 0.017 & 0.017 & 0.017 & 0.018 \end{array}\right].$$

This matrix presents the estimated average nutrient concentrations across the four depth layers (Surface – `S`, Above-Pycnocline – `AP`, Below-Pycnocline – `BP`, and Bottom – `B`) at station EE2.1, as obtained from the fitted MVSNC model, providing a detailed summary of the vertical distribution of key chemical variables in the water column.

- **Phosphate (PO_4):** The mean phosphate concentrations remain nearly constant across all depths, ranging from 0.003 to 0.004. This stability suggests a relatively homogeneous vertical distribution, possibly due to efficient vertical mixing or a steady-state balance between biological uptake and regeneration.

- **Total Dissolved Nitrogen (TDN):** The estimated means show minimal variation across layers (0.544–0.558), indicating that nitrogen levels are well distributed throughout the water column. Such consistency may reflect strong hydrodynamic connectivity and sustained nutrient availability across depths.

- **Total Dissolved Phosphorus (TDP):** The TDP values are almost constant (0.017–0.018), showing a very stable phosphorus profile with a slight increase near the bottom layer, potentially associated with weak sedimentary release processes.

The estimated location matrix $\widehat{\mathbf{LM}}_{\text{EE2.1}}$ exhibits a relatively uniform nutrient pattern along the vertical profile, indicating minimal stratification effects at station EE2.1. This consistency implies that nutrient concentrations are largely controlled by mixing processes rather than pronounced vertical gradients, and demonstrates that the MVSNC model effectively represents the balanced distribution of nutrients, even in the presence of censored observations.

The estimated skewness matrix $\widehat{\mathbf{A}}_{\text{EE2.1}}$ is given by:

$$\widehat{\mathbf{A}}_{\text{EE2.1}} = \left[\begin{array}{c|cccc} & \texttt{AP} & \texttt{B} & \texttt{BP} & \texttt{S} \\ \hline \texttt{po4} & 3.051063 \times 10^{-5} & -0.0005145 & -0.0003387 & 0.0002372 \\ \texttt{tdn} & 0.2138 & 0.1847 & 0.1979 & 0.2459 \\ \texttt{tdp} & -0.002142 & -0.002307 & -0.0002156 & -0.001321 \end{array}\right].$$

This matrix presents the estimated skewness parameters for each nutrient (rows) across the four depth layers (columns), as derived from the MVSNC model fitted to station EE2.1, providing a comprehensive summary of how asymmetry varies both among nutrients and throughout the vertical profile.

- **Phosphate (PO_4):** The skewness coefficients for PO_4 are very close to zero in all layers, with small positive values at `AP` and `S` and small negative values at `B` and `BP`. This pattern indicates an essentially symmetric distribution of phosphate concentrations at EE2.1, with no pronounced tail behavior in any particular direction.

- **Total Dissolved Nitrogen (TDN):** The skewness parameters for TDN are clearly positive in all layers (ranging approximately from 0.18 to 0.25), with the largest value at the surface layer `S` (≈ 0.246). These positive coefficients signal a right-skewed distribution, suggesting the presence of occasional high TDN values relative to the mean. The slightly stronger skewness near the surface may reflect episodic inputs or biological events (e.g., primary production or external nutrient pulses) that intermittently increase nitrogen concentrations.

- **Total Dissolved Phosphorus (TDP):** For TDP, the skewness estimates are predominantly negative (around -0.002 at `AP` and `B`, and slightly less negative at `S` and `BP`). This indicates a mild left-skewed behavior, with somewhat heavier lower tails. In practical terms, this suggests that relatively low TDP values occur more frequently than would be expected under symmetry, possibly due to strong biological uptake or limitation episodes that drive concentrations downward.

The estimated matrix $\widehat{\mathbf{A}}_{\text{EE2.1}}$ indicates that asymmetry is most pronounced for TDN, which shows a consistently right-skewed distribution across all depths, while PO_4 remains nearly symmetric and TDP exhibits a slight left skew. These patterns underscore the non-Gaussian characteristics of the nutrient profiles at station EE2.1 and demonstrate the capacity of the MVSNC model to capture distinct skewness behaviors across both variables and depth layers, even in the presence of censored observations.

The correlation matrix corresponding to the estimated row covariance matrix $\widehat{\mathbf{C}}^{\text{row}}_{\text{EE2.1}}$ is presented below, providing a detailed overview of the pairwise relationships among the different rows in the dataset and their relative associations across the measured variables:

$$\widehat{\mathrm{R}}(\widehat{\mathrm{C}}^{\mathrm{row}}_{\mathrm{EE2.1}}) = \left[\begin{array}{c|ccc} & \texttt{po4} & \texttt{tdn} & \texttt{tdp} \\ \hline \texttt{po4} & 1.000 & 0.021 & 0.218 \\ \texttt{tdn} & 0.021 & 1.000 & 0.046 \\ \texttt{tdp} & 0.218 & 0.046 & 1.000 \end{array}\right].$$

The estimated correlation matrix $\widehat{\mathrm{R}}(\widehat{\mathrm{C}}^{\mathrm{row}}_{\mathrm{EE2.1}})$ provides a summary of the dependence structure among the nutrient variables (PO_4, TDN, and TDP) at station EE2.1. The overall pattern indicates generally weak linear associations, suggesting that the nutrient variables largely vary independently within the water column.

- The correlation between PO_4 and TDP ($r = 0.218$) is the largest in the matrix, though still relatively low. This positive association is ecologically coherent, as both nutrients are phosphorus-related and may respond similarly to remineralization or sediment release.
- The correlation between TDN and TDP ($r = 0.046$) is very weak, indicating that nitrogen and phosphorus concentrations vary mostly independently, possibly reflecting distinct biogeochemical cycling or source processes.
- The correlation between PO_4 and TDN ($r = 0.021$) is negligible, suggesting that phosphate and total dissolved nitrogen follow unrelated temporal or spatial dynamics at this station.

The estimated $\widehat{\mathrm{R}}(\widehat{\mathrm{C}}^{\mathrm{row}}_{\mathrm{EE2.1}})$ reveals that nutrient co-variation at station EE2.1 is generally weak, with only a moderate correlation observed between the phosphorus forms (PO_4 and TDP). This correlation structure indicates that the processes controlling nitrogen and phosphorus distributions are largely independent and demonstrates that the MVSNC model effectively captures the limited interdependence among nutrients in the censored observations at this site.

The correlation matrix related to the estimated column covariance matrix $\widehat{\mathrm{R}}(\widehat{\mathrm{C}}^{\mathrm{column}}_{\mathrm{EE2.1}})$ is presented below, providing a summary of the pairwise relationships among the different depth layers.

$$\widehat{\mathrm{R}}(\widehat{\mathrm{C}}^{\mathrm{column}}_{\mathrm{EE2.1}}) = \left[\begin{array}{c|cccc} & \texttt{S} & \texttt{AP} & \texttt{BP} & \texttt{B} \\ \hline \texttt{S} & 1.000 & 0.919 & 0.873 & 0.846 \\ \texttt{AP} & 0.919 & 1.000 & 0.902 & 0.873 \\ \texttt{BP} & 0.873 & 0.902 & 1.000 & 0.901 \\ \texttt{B} & 0.846 & 0.873 & 0.901 & 1.000 \end{array}\right].$$

The estimated correlation matrix $\widehat{\mathrm{R}}(\widehat{\mathrm{C}}^{\mathrm{column}}_{\mathrm{EE2.1}})$ captures the dependence structure among the four depth layers (Surface – `S`, Above-Pycnocline – `AP`, Below-Pycnocline – `BP`, and Bottom – `B`) at station EE2.1. The correlations are uniformly high and positive, indicating a strong vertical coherence in nutrient concentrations across the water column.

- The strongest correlations are observed between adjacent layers, such as S-AP ($r = 0.919$) and BP–B ($r = 0.901$), reflecting smooth vertical transitions and the absence of sharp discontinuities in the nutrient profiles.
- The correlations between nonadjacent layers, while slightly smaller, remain strong (e.g., S–B: $r = 0.846$), suggesting that the entire water column behaves as an interconnected system with limited stratification.
- The overall pattern, with all $r > 0.84$, implies that temporal and spatial nutrient variations at one depth are closely mirrored at others, consistent with an environment where mixing processes dominate over stratifying forces.

The estimated $\widehat{\mathrm{R}}(\widehat{\mathrm{C}}^{\mathrm{column}}_{\mathrm{EE2.1}})$ highlights a strongly correlated vertical structure, suggesting that nutrient dynamics at station EE2.1 are vertically interconnected. This pattern is typical of well-mixed or moderately stratified estuarine systems and demonstrates that the MVSNC model effectively captures the pronounced interlayer dependencies present in the censored nutrient observations.

Discussion and Concluding Remarks

The analysis of the `dataCensored` dataset demonstrates that the MVSNC model successfully captures both the central tendencies and the non-Gaussian characteristics of nutrient distributions across the four depth layers, even in the presence of censored observations. The estimated location and skewness matrices reveal that nitrogen compounds exhibit pronounced right-skewed behavior, whereas phosphorus remains comparatively symmetric, reflecting the underlying chemical and biological processes such as episodic nitrogen uptake and remineralization. The correlation analyses indicate weak co-variation among the nutrient variables but strong vertical coherence across depth layers, suggesting that nutrient dynamics at station EE2.1 are largely vertically integrated, consistent with well-mixed or moderately stratified estuarine systems.

These findings highlight the ability of the MVSNC model to capture distinct patterns of variability and skewness, providing an interpretable framework for understanding nutrient interactions in complex environmental settings. While the approach proves highly effective, it relies on latent-variable assumptions and can be computationally intensive for larger datasets. Future work could consider Bayesian or hierarchical extensions, high-dimensional regularization, or spatio-temporal modeling to further enhance inference and applicability.

In summary, the application of the MVSNC model to the `dataCensored` dataset underscores its flexibility and robustness in modeling asymmetric and censored environmental data, offering valuable insights into nutrient distributions and interdependencies in estuarine systems.

CHAPTER

7

THE MVCENS PACKAGE

The `MVCens` package implements a comprehensive and versatile framework for the simulation and statistical inference of matrix-variate Normal (MVN) and matrix-variate Skew-Normal (MVSN) models, addressing a broad spectrum of data structures commonly encountered in practice. In particular, the framework is designed to accommodate both fully observed datasets and situations involving incomplete information. Such incompleteness may arise through interval censoring, left or right censoring, or the presence of missing values, all of which are explicitly incorporated into the modeling assumptions and estimation procedures supported by the package. This flexibility allows matrix-valued data to be analyzed under realistic observation schemes without relying on ad hoc treatments of incomplete entries.

In addition, the package provides a comprehensive set of functions for generating matrix-valued samples from MVN and MVSN distributions under a variety of censoring and missing-data mechanisms, supporting both methodological research and simulation studies. Beyond data generation, it includes dedicated routines for likelihood evaluation and parameter estimation tailored to matrix-variate models. Statistical inference relies on maximum likelihood estimation for complete MVN data and on ECM-type algorithms for MVN and MVSN models in the presence of censoring or missingness. These procedures are implemented within a unified computational framework that ensures methodological coherence and enables different model specifications—censored or uncensored, symmetric or asymmetric—to be treated in a consistent manner. The entire implementation is based on the stochastic representation $\boldsymbol{\mathcal{X}} = \mathbf{M} + \mathbf{A}W + \boldsymbol{\mathcal{V}}$, which underpins simulation, likelihood construction, and inference for skew-normal models while facilitating the generation of asymmetric matrix-variate data and EM-based estimation under incomplete observation schemes.

ECM Estimation for Matrix-Variate Models

The function `mv_ecm` implements a unified ECM-type estimation algorithm for a broad class of matrix-variate distributions, including the matrix-variate Normal (MVN), the matrix-variate Skew-Normal (MVSN), and their censored counterparts, namely the MVNC and MVSNC models. This unified implementation allows different distributional assumptions to be handled within a common estimation framework. The algorithm is specifically designed to accommodate datasets that include interval-censored observations, as well as datasets affected by missing values, which frequently arise in practical applications involving matrix-valued data.

Model specification

The target distribution is specified through a character argument, enabling the user to select among several model configurations. In particular, the function allows for the fitting of a matrix-variate Normal model under complete observation, a censored matrix-variate Normal model, a matrix-variate Skew-Normal model with complete data, or a censored matrix-variate Skew-Normal model. This flexibility makes it possible to seamlessly transition between symmetric and asymmetric models, as well as between complete and incomplete data scenarios, without modifying the overall estimation procedure.

Data structure

Observed data are supplied in the form of a three-dimensional array of dimension $p \times q \times n$, where each slice corresponds to a single matrix-valued observation. For uncensored entries, the array contains the observed values directly, whereas for censored entries it stores the corresponding lower bounds. In particular, missing values or left-censored observations are encoded as $-\infty$. Interval censoring is represented through an associated indicator array that identifies censored elements, while the corresponding upper bounds are provided in a separate array. For missing data or right-censored observations, these upper bounds are set to $+\infty$. This data representation allows the ECM algorithm to correctly account for the censoring structure during the estimation process.

Estimation output

The ECM algorithm returns estimates of the location matrix, the row covariance matrix, and the column covariance matrix. When skew-normal models are considered, estimates of the skewness matrix are also provided. In addition to parameter estimates, the output includes the final value of the observed-data log-likelihood, the Bayesian Information Criterion (BIC) for model comparison purposes, and the total number of ECM iterations performed until convergence. Convergence is assessed based on a user-specified tolerance level and a maximum number of iterations, ensuring numerical stability and computational control.

Generation of Censored Matrix-Variate Data

The function `rmatrix_censored` is designed to generate synthetic matrix-valued observations from either an MVN or an MVSN distribution, with the option to introduce censoring or missingness according to a user-defined mechanism. The degree of censoring is controlled through a quantile-based threshold, which determines the proportion of entries subject to censoring within each generated matrix.

Several censoring schemes are supported, including symmetric interval censoring, purely missing data, mixed censoring schemes combining missing and censored values, as well as one-sided censoring mechanisms such as left or right censoring. The function returns the censored data array, an indicator array identifying which entries are censored, and an array containing the corresponding upper censoring limits. These outputs are fully compatible with the input requirements of the `mv_ecm` function, facilitating simulation studies and method validation.

Matrix-variate observations are generated using standard matrix-normal simulation routines. In the case of skew-normal models, asymmetry is introduced by adding a skewness matrix multiplied by the absolute value of a standard normal variate, in accordance with the underlying stochastic representation adopted throughout the package.

Simulation of Matrix-Variate Skew-Normal Data

The function `rmvsn` generates samples from the matrix-variate Skew-Normal distribution under complete observation. Given user-specified dimensions and parameter matrices, the function produces a collection of matrix-valued observations according to the stochastic representation

$$\mathcal{X}_i = \mathbf{M} + \mathbf{A}|W_i| + \mathcal{V}_i,$$

where W_i denotes a standard normal random variable and $\mathcal{V}_i$ follows a matrix-variate Normal distribution with specified row and column covariance structures. The resulting output is returned as a three-dimensional array containing the simulated matrices, making it suitable for simulation experiments and illustrative examples.

Summary

The `MVCens` package provides a flexible and extensible computational environment for matrix-variate modeling in the presence of asymmetry and incomplete data. By combining stochastic representations with ECM-based estimation procedures, the package supports simulation studies, methodological development, and applied analyses involving censored or missing matrix-valued observations. The unified design of its simulation and estimation tools facilitates coherent modeling workflows and promotes reproducibility in matrix-variate statistical analyses.

CHAPTER

8

CONCLUSIONS

This thesis advances the theory and methodology of matrix-variate skewed distributions by developing a comprehensive framework for the matrix variate skew-normal (MVSN) and matrix variate ekew-normal censored (MVSNC) models. The first part establishes fundamental distributional properties, identifiability conditions, and computational tools for likelihood-based inference under the MVSN framework. Through explicit latent-variable representations and a tailored ECM algorithm, the proposed methodology provides stable, efficient, and interpretable parameter estimation. Simulation studies and real-data analyses demonstrate the practical advantages of incorporating skewness and structured matrix dependence, showing clear improvements over the classical matrix-variate normal model in settings where asymmetry is important.

Building on this foundation, the second part extends the framework to censored and partially observed matrix-valued data. The MVSNC model integrates skewness, matrix dependence, and censoring within a unified hierarchical structure, allowing censored entries to contribute directly to the likelihood via truncated conditional distributions. The resulting ECM algorithm, implemented in the R package `MVCens` Correia, Davila and Diniz (2025) and freely available at MVCens R package, offers a tractable and principled estimation strategy. Empirical results confirm that the MVSNC model effectively preserves information that would otherwise be lost through ad hoc preprocessing, outperforming standard approaches when data exhibit both asymmetry and censoring.

Overall, this thesis provides a rigorous and versatile framework for modeling non-Gaussian, asymmetric, and incomplete matrix-valued data. The combination of latent-variable formulations, stochastic representations, and tailored estimation strategies lays a solid foundation for future work, including Bayesian inference, high-dimensional regularization, hierarchical structures, and spatiotemporal or longitudinal applications. These contributions establish a strong platform for continued advances in modern matrix-variate statistical modeling.

BIBLIOGRAPHY

AHSANULLAH, M. Some characterizations of the bivariate normal distribution. **Metrika**, Springer, v. 32, n. 1, p. 215–218, 1985. Citation on page 23.

ALENCAR, F. H. C. de; GONÇALVES, L. W. N. *et al.* Finite mixture modeling of multivariate skew-normal data with censoring and missingness. **Communications in Statistics—Simulation and Computation**, Taylor & Francis, 2022. Citation on page 18.

ALLEN, G. I.; TIBSHIRANI, R. Transposable regularized covariance models with an application to missing data imputation. **The Annals of Applied Statistics**, NIH Public Access, v. 4, n. 2, p. 764, 2010. Citation on page 26.

ANDERLUCCI, L.; MONTANARI, A.; VIROLI, C. *et al.* A matrix-variate regression model with canonical states: An application to elderly danish twins. **Statistica**, v. 74, n. 4, p. 367–381, 2014. Citation on page 26.

AZZALINI, A. A class of distributions which includes the normal ones. **Scandinavian journal of statistics**, JSTOR, p. 171–178, 1985. Citations on pages 17 and 44.

____. Further results on a class of distributions which includes the normal ones. **Statistica**, v. 46, n. 2, p. 199–208, 1986. Citation on page 17.

AZZALINI, A.; CAPITANIO, A. Statistical applications of the multivariate skew normal distribution. **Journal of the Royal Statistical Society: Series B (Statistical Methodology)**, Wiley Online Library, v. 61, n. 3, p. 579–602, 1999. Citations on pages 17 and 44.

____. **The Skew-Normal and Related Families**. [S.l.]: Cambridge University Press, 2014. Citation on page 18.

AZZALINI, A.; VALLE, A. D. The multivariate skew-normal distribution. **Biometrika**, Oxford University Press, v. 83, n. 4, p. 715–726, 1996. Citations on pages 17 and 48.

BANDYOPADHYAY, D.; RAO, P. S. R. S. Linear mixed models for longitudinal data with skew-normal/independent distributions and censoring. **Biostatistics**, Oxford University Press, v. 13, n. 3, p. 409–424, 2012. Citation on page 18.

BOYD, S.; VANDENBERGHE, L. **Convex Optimization**. [S.l.]: Cambridge University Press, 2004. Citation on page 103.

CHEN, J. T.; GUPTA, A. K. Matrix variate skew normal distributions. **Statistics**, Taylor & Francis, v. 39, n. 3, p. 247–253, 2005. Citations on pages 18, 43, and 44.

CORREIA, A.; DAVILA, V. H. L.; DINIZ, C. A. R. **MVCens: An R Package for Estimation of Matrix-Variate Skew-Normal Censored Models**. 2025. <https://github.com/atilapcorreia/MVCens>. R package. Citations on pages 108 and 129.

COUPÉ, P.; MANJÓN, J. V.; COLLINS, D. L.; ROBLES, M. Robust rician noise estimation for mr images. **Medical Image Analysis**, Elsevier, v. 14, n. 4, p. 483–493, 2010. Citation on page 18.

DAGNE, G. A.; ROUSSON, V. *et al.* Bayesian analysis of multivariate skew-normal models with left-censored data. **Journal of Statistical Planning and Inference**, Elsevier, v. 143, n. 8, p. 1295–1306, 2013. Citation on page 19.

DEIMLING, K. **Nonlinear functional analysis**. [S.l.]: Springer Science & Business Media, 2013. Citation on page 28.

DEMPSTER, A. P.; LAIRD, N. M.; RUBIN, D. B. Maximum likelihood from incomplete data via the em algorithm. **Journal of the royal statistical society: series B (methodological)**, Wiley Online Library, v. 39, n. 1, p. 1–22, 1977. Citations on pages 19, 73, and 101.

DUTILLEUL, P. The mle algorithm for the matrix normal distribution. **Journal of statistical computation and simulation**, Taylor & Francis, v. 64, n. 2, p. 105–123, 1999. Citations on pages 25 and 43.

FRIEDBERG, S. H.; INSEL, A. J.; SPENCE, L. E. **Linear algebra: Pearson new international edition**. [S.l.]: Pearson Higher Ed, 2013. Citation on page 27.

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**, Springer, v. 80, n. 2, p. 231–253, 2022. Citations on pages 60, 61, 95, and 102.

GALLAUGHER, M. P.; MCNICHOLAS, P. D. A matrix variate skew-t distribution. **Stat**, Wiley Online Library, v. 6, n. 1, p. 160–170, 2017. Citation on page 45.

____. Finite mixtures of skewed matrix variate distributions. **Pattern Recognition**, Elsevier, v. 80, p. 83–93, 2018. Citation on page 26.

____. Three skewed matrix variate distributions. **Statistics & Probability Letters**, Elsevier, v. 145, p. 103–109, 2019. Citations on pages 45, 81, 85, and 108.

GLANZ, H.; CARVALHO, L. An expectation–maximization algorithm for the matrix normal distribution with an application in remote sensing. **Journal of Multivariate Analysis**, Elsevier, v. 167, p. 31–48, 2018. Citations on pages 36, 104, and 105.

GUDBJARTSSON, H.; PATZ, S. The rician distribution of noisy mri data. **Magnetic Resonance in Medicine**, Wiley, v. 34, n. 6, p. 910–914, 1995. Citation on page 18.

GUPTA, A. K.; GONZÁLEZ-FARIAS, G.; DOMINGUEZ-MOLINA, J. A. A multivariate skew normal distribution. **Journal of multivariate analysis**, Elsevier, v. 89, n. 1, p. 181–190, 2004. Citation on page 18.

GUPTA, A. K.; NAGAR, D. K. **Matrix variate distributions**. [S.l.]: Chapman and Hall/CRC, 2018. Citations on pages 23 and 25.

HELSEL, D. R. **Statistics for Censored Environmental Data Using Minitab and R**. 2nd. ed. [S.l.]: Wiley, 2012. Citation on page 18.

HIGHAM, N. J. Computing a nearest symmetric positive semidefinite matrix. **Linear Algebra and its Applications**, v. 103, p. 103–118, 1988. Citation on page 103.

HUYNH, T.; XUE, J.; DINH, H.; AL. et. Comparison of methods for analyzing left-censored exposure data. **Environmental Health Insights**, SAGE, v. 8, p. 1–10, 2014. Citation on page 18.

JANSON, S.; KAIJSER, S. **Higher moments of Banach space valued random variables**. [S.l.]: American Mathematical Society, 2015. Citation on page 69.

LACHOS, V. H.; BOLFARINE, H.; ARELLANO-VALLE, R. B.; MONTENEGRO, L. C. Likelihood-based inference for multivariate skew-normal regression models. **Communications in Statistics—Theory and Methods**, Taylor & Francis, v. 36, n. 9, p. 1769–1786, 2007. Citations on pages 48 and 49.

LACHOS, V. H.; TOMARCHIO, S. D.; PUNZO, A.; INGRASSIA, S. An em algorithm for fitting matrix-variate normal distributions on interval-censored and missing data. **Statistics and Computing**, v. 35, p. 39, 2025. Citation on page 94.

_____. An em algorithm for fitting matrix-variate normal distributions on interval-censored and missing data. **Statistics and Computing**, Springer, v. 35, n. 2, p. 1–11, 2025. Citations on pages 108 and 116.

LIN, T.-I.; MCLACHLAN, G. J.; LEE, S. X. Extending mixtures of factor models using the restricted multivariate skew-normal distribution. **Journal of Multivariate Analysis**, Elsevier, v. 143, p. 398–413, 2016. Citation on page 17.

LITTLE, R. J. A.; RUBIN, D. B. **Statistical Analysis with Missing Data**. 2nd. ed. [S.l.]: Wiley, 2002. Citation on page 18.

MATHAI, A. M.; PROVOST, S. B.; HAUBOLD, H. J. **Multivariate statistical analysis in the real and complex domains**. [S.l.]: Springer Nature, 2022. Citation on page 59.

MELNYKOV, V.; ZHU, X. On model-based clustering of skewed matrix data. **Journal of Multivariate Analysis**, Elsevier, v. 167, p. 181–194, 2018. Citation on page 25.

MENG, X.-L.; RUBIN, D. B. Maximum likelihood estimation via the ecm algorithm: A general framework. **Biometrika**, Oxford University Press, v. 80, n. 2, p. 267–278, 1993. Citations on pages 73 and 101.

MIDDLETON, D. **An Introduction to Statistical Communication Theory**. [S.l.]: IEEE Press, 1986. Classic reference on non-Gaussian noise in signal processing. Citation on page 18.

MORALES, C. E. G.; MATOS, L. A.; DEY, D. K.; LACHOS, V. H. On moments of folded and doubly truncated multivariate extended skew-normal distributions. **Journal of Computational and Graphical Statistics**, Taylor & Francis, v. 31, n. 2, p. 455–465, 2022. Citations on pages 67, 95, 96, and 102.

MURPHY, R.; PERRY, E.; KEISMAN, J.; HARCUM, J.; LEPPO, E. W. **baytrends: Long Term Water Quality Trend Analysis**. [S.l.], 2023. R package version 2.0.12. Available: <https://cran.r-project.org/package=baytrends>. Citation on page 112.

MURPHY, R. R.; PERRY, E.; HARCUM, J.; KEISMAN, J. A generalized additive model approach to evaluating water quality: Chesapeake bay case study. **Environmental Modelling & Software**, v. 118, p. 1–13, 2019. Citation on page 112.

NADERI, M.; BEKKER, A.; ARASHI, M.; JAMALIZADEH, A. A theoretical framework for landsat data modeling based on the matrix variate mean-mixture of normal model. **Plos one**, Public Library of Science San Francisco, CA USA, v. 15, n. 4, p. e0230773, 2020. Citations on pages 47, 49, and 52.

NADERI, M.; TAMANDI, M.; MIRFARAH, E.; WANG, W.-L.; LIN, T.-I. Three-way data clustering based on the mean-mixture of matrix-variate normal distributions. **Computational Statistics & Data Analysis**, Elsevier, v. 199, p. 108016, 2024. Citations on pages 47 and 60.

NGUYEN, T. T. A note on matrix variate normal distribution. **journal of multivariate analysis**, New York [etc.] Academic Press., v. 60, n. 1, p. 148–153, 1997. Citation on page 23.

NING, W.; GUPTA, A. K. Matrix variate extended skew normal distributions. **Random Operators and Stochastic Equations**, Walter de Gruyter GmbH, v. 20, n. 4, p. 299–310, 2012. Citations on pages 18 and 44.

PETERSEN, K. B.; PEDERSEN, M. S. *et al.* The matrix cookbook. **Technical University of Denmark**, v. 7, n. 15, p. 510, 2008. Citations on pages 76, 77, 78, and 79.

QUEIROZ, M. M. de; LOSCHI, R. H.; SILVA, R. W. Multivariate log-skewed distributions with normal kernel and their applications. **Statistics**, Taylor & Francis, v. 50, n. 1, p. 157–175, 2016. Citation on page 17.

RAPPOPORT, P.; WHITE, E. N. Was there a bubble in the 1929 stock market? **The Journal of Economic History**, Cambridge University Press, v. 53, n. 3, p. 549–574, 1993. Citation on page 85.

REZAEI, A.; YOUSEFZADEH, F.; ARELLANO-VALLE, R. B. Scale and shape mixtures of matrix variate extended skew normal distributions. **Journal of Multivariate Analysis**, Elsevier, v. 179, p. 104649, 2020. Citation on page 85.

ROLLA, L. T.; LIMA, B. N. B. de. **Probabilidade**. [S.l.: s.n.], 2025. Version dated March 18, 2025. Licensed under Creative Commons Attribution-NoDerivatives 4.0 International. Citation on page 59.

SAHU, S. K.; DEY, D. K.; BRANCO, M. D. A new class of multivariate skew distributions with applications to bayesian regression models. **Canadian Journal of Statistics**, Wiley Online Library, v. 31, n. 2, p. 129–150, 2003. Citation on page 17.

SARKAR, S.; ZHU, X.; MELNYKOV, V.; INGRASSIA, S. On parsimonious models for modeling matrix data. **Computational Statistics & Data Analysis**, Elsevier, v. 142, p. 106822, 2020. Citations on pages 25 and 26.

SEEGER, M. Gaussian processes for machine learning. **International Journal of Neural Systems**, v. 14, n. 2, p. 69–106, 2004. Citation on page 103.

TOMARCHIO, S. D.; MCNICHOLAS, P. D.; PUNZO, A. Matrix normal cluster-weighted models. **Journal of Classification**, Springer, v. 38, n. 3, p. 556–575, 2021. Citation on page 26.

TOMARCHIO, S. D.; PUNZO, A.; BAGNATO, L. On the use of the matrix-variate tail-inflated normal distribution for parsimonious mixture modeling. In: **Convegno della Società Italiana di Statistica**. [S.l.]: Springer, 2021. p. 407–423. Citation on page 25.

TONG, Y. L. **The Multivariate Normal Distribution**. [S.l.]: Springer New York, NY, 1990. Citation on page 24.

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. **Journal of Multivariate Analysis**, Elsevier, v. 196, p. 105174, 2023. Citation on page 94.

VIROLI, C. Finite mixtures of matrix normal distributions for classifying three-way data. **Statistics and Computing**, Springer, v. 21, p. 511–522, 2011. Citation on page 26.

____. On matrix-variate regression analysis. **Journal of Multivariate Analysis**, Elsevier, v. 111, p. 296–309, 2012. Citation on page 26.

XU, C.; ZHOU, J.; LI, Y. Nonlinear filtering with asymmetric measurement noise modeled by the skew-*t* distribution. **Signal Processing**, Elsevier, v. 197, p. 108532, 2022. Citation on page 18.


ICMC
USP
SÃO CARLOS