Technical University Dortmund

B.Sc. Data Science

Modul BD 14: Seminar

Estimation Methods of Matrix-valued AR Model
-
Reproduction and Analysis

Author:
Henry Huick

Supervisors:
Prof. Dr. Carsten Jentsch
Prof. Dr. Matei Demetrescu

January 11, 2026

Abstract

Matrix-valued time series are increasingly prominent in high-dimensional applications such as finance and neuroscience, where preserving the inherent row-column structure is essential for efficient modeling. Classical vector autoregressive (VAR) models often face a "parameter explosion" in these settings, leading to estimation challenges. This paper reproduces and critically analyzes the Matrix Autoregressive (MAR(1)) estimation framework proposed by (Kołodziejski, 2025), specifically focusing on Yule-Walker, Burg-type, and Least Squares Estimation (LSE) methods.

We extend the existing literature by evaluating these estimators under both correctly specified MAR(1) processes and misspecified VAR(1) environments, utilizing a "Kronecker Energy" metric to quantify structural sensitivity. Our simulation results indicate that while MAR models provide significant benefits in parameter parsimony, their predictive advantage is highly dependent on the proximity of the true data-generating process to the Kronecker subspace.

Furthermore, our replication of prior results suggests that certain counterintuitive findings regarding MAR superiority may be attributed to specific experimental configurations. Among the evaluated techniques, the LSE method demonstrated the most robust performance and the highest numerical stability across all tested setups.

Keywords: Matrix-valued time series, MAR(1), Yule-Walker, Burg estimation, Least Squares Estimation, Kronecker structure.

Contents

1 Introduction

Matrix-valued time series arise naturally in high-dimensional applications such as finance (multi-asset portfolios), neuroscience (EEG/fMRI functional connectivity), and sensor networks. Classical vector autoregressive (VAR) models treat each matrix observation as a long vector. However, this approach leads to a quadratic explosion in the number of parameters and fails to exploit the inherent row-column structure of the data. To address these limitations, Matrix Autoregressive (MAR) models were introduced to preserve matrix structure while modeling temporal dependencies efficiently (Chen et al., 2021, Li & Xiao, 2021, Samadi & Billard, 2025, Kołodziejski, 2025).

A first-order matrix autoregressive model, MAR(1), for a time series $\{\mathbf{X}_t; t \in \mathbb{Z}\}$ is defined as

$$\mathbf{X}_t = \mathbf{A}\mathbf{X}_{t-1}\mathbf{B}^\top + \mathbf{Z}_t, \tag{1}$$

where $\mathbf{X}_t : \Omega \to \mathbb{R}^{m\times n}$ is a random matrix process, $\mathbf{A} \in \mathbb{R}^{m\times m}$ and $\mathbf{B} \in \mathbb{R}^{n\times n}$ are coefficient matrices, and $\mathbf{Z}_t : \Omega \to \mathbb{R}^{m\times n}$ is a zero-mean matrix-valued white noise process. The noise process satisfies $\mathbb{E}[\mathrm{vec}(\mathbf{Z}_t)\mathrm{vec}(\mathbf{Z}_{t+h})^\top] = \mathbf{\Sigma}$ if $h = 0$ and $\mathbf{0}$ otherwise. Here, $\mathrm{vec}(\cdot)$ denotes the vectorization operator which stacks the columns of a matrix into a single vector.

The MAR(1) model can be related to the classical VAR(1) model via the vectorization identity:

$$\mathrm{vec}(\mathbf{X}_t) = (\mathbf{B} \otimes \mathbf{A})\mathrm{vec}(\mathbf{X}_{t-1}) + \mathrm{vec}(\mathbf{Z}_t), \tag{2}$$

where $\otimes$ denotes the Kronecker product and $\mathrm{vec}(\cdot)$ refers to the column-major vectorization (Kołodziejski, 2025). This representation highlights a significant advantage of the MAR specification: parsimony. By imposing a Kronecker structure on the transition matrix, the number of parameters is reduced from m^2n^2 (in a full VAR) to $m^2 + n^2$.

The vectorization representation is also used to establish stationarity and causality conditions. Specifically, the MAR(1) model is stationary and causal if

$$\rho(\mathbf{B} \otimes \mathbf{A}) = \rho(\mathbf{A}) \cdot \rho(\mathbf{B}) < 1,$$

where $\rho(\cdot)$ denotes the spectral radius (i.e., the largest absolute eigenvalue) (Chen et al., 2021).

A unique definition of the MAR model requires normalization, as the bilinearity in the coefficient matrices leads to an identification issue. The model is invariant to rescaling operations of the form $(c\mathbf{A})\mathbf{X}_{t-1}(\frac{1}{c}\mathbf{B})^\top = \mathbf{A}\mathbf{X}_{t-1}\mathbf{B}^\top$ for any non-zero scalar $c \in \mathbb{R}$ (Kołodziejski, 2025). Consequently, constraints such as $\|\mathbf{A}\|_F = 1$ are typically imposed to ensure identifiability.

While Chen et al. (2021) introduced maximum likelihood estimation and least squares estimation methods for the MAR model, these approaches do not guarantee that the estimated parameters satisfy the causality condition. To overcome this, Kołodziejski (2025) proposed method-of-moments approaches—specifically Yule-Walker and Burg-type estimators—that ensure causality by construction and preserve the autocovariance structure of the observed time series.

We reproduce the estimation framework of Kołodziejski (2025) and provide a critical extension to their experimental setup by evaluating the estimators across a broader range of configurations. Unlike the original study, which utilized a VAR(1) data-generating process, we analyze these estimators under both misspecified and correctly specified MAR(1) environments with symmetric and non-symmetric coefficient matrices. Furthermore, we utilize a 'Kronecker Energy' metric to quantify structural sensitivity and assess how predictive accuracy degrades as the process deviates from the Kronecker subspace. This approach allows for a robust assessment of the estimators' performance, numerical stability, and limitations across diverse matrix-valued time series settings.

The code to reproduce the experiments in this paper is available at https://github.com/lepremiere/mar-reproduction-analysis.

2 Estimators

In this section, we detail the estimation methods for the MAR(1) model used in our simulation study, alongside the classical VAR(1) approach for comparison. We focus on three primary estimation techniques: Yule-Walker (YW), Burg's method, and Least Squares (LSE). For each method, we distinguish between a direct matrix-valued estimation and a two-step *vectorized* approach that first estimates a VAR(1) model on the vectorized form of the data and subsequently approximates the MAR(1) structure via the Nearest Kronecker Product (NKP) method (Van Loan & Pitsianis, 1993).

We assume throughout that our time series have zero mean, i.e., $\mathbb{E}[\mathbf{X}_t] = \mathbf{0}$ and $\mathbb{E}[\mathbf{y}_t] = \mathbf{0}$, such that the computation of covariances reduces to $\mathbb{E}[\mathbf{X}_t\mathbf{X}_{t-h}^\top]$ and $\mathbb{E}[\mathbf{y}_t\mathbf{y}_{t-h}^\top]$, respectively. Additionally, all inverse matrices refer to the Moore-Penrose pseudoinverse to accommodate potential singularities. Higher-order models are omitted for brevity.

2.1 Yule-Walker Estimation

The Yule-Walker equations are particularly useful for estimating parameters in a stationary time series model by relating the model parameters to the autocovariance structure of the data. This method does not impose any distributional assumptions on the noise process beyond stationarity and zero-mean and can therefore be applied in a wide range of settings Kołodziejski (2025).

2.1.1 VAR(1) Yule-Walker

We start with the classical VAR(1) model defined as

$$\mathbf{y}_t = \mathbf{\Phi}\mathbf{y}_{t-1} + \mathbf{e}_t, \tag{3}$$

where $\mathbf{y}_t \in \mathbb{R}^d$ is the observation vector, $\mathbf{\Phi} \in \mathbb{R}^{d\times d}$ is the coefficient matrix, and $\mathbf{e}_t \in \mathbb{R}^d$ is a zero-mean white noise process.

To derive the Yule-Walker equations (Lütkepohl, 2005), we post-multiply the

model equation by $\mathbf{y}_{t-1}^\top$ and take expectations:

$$\mathbb{E}[\mathbf{y}_t \mathbf{y}_{t-1}^\top] = \mathbf{\Phi}\mathbb{E}[\mathbf{y}_{t-1}\mathbf{y}_{t-1}^\top] + \mathbb{E}[\mathbf{e}_t \mathbf{y}_{t-1}^\top].$$

Using the definition of the autocovariance matrix $\mathbf{\Gamma}_k = \mathbb{E}[\mathbf{y}_t \mathbf{y}_{t-k}^\top]$ and observing that $\mathbb{E}[\mathbf{e}_t \mathbf{y}_{t-1}^\top] = \mathbf{0}$ due to the white noise assumption, this simplifies to:

$$\mathbf{\Gamma}_1 = \mathbf{\Phi}\mathbf{\Gamma}_0.$$

Solving for $\mathbf{\Phi}$ yields the standard VAR(1) Yule-Walker estimator:

$$\mathbf{\Phi}_{\text{YW}} = \mathbf{\Gamma}_1 \mathbf{\Gamma}_0^{-1}.$$

Approximating the autocovariance matrices $\mathbf{\Gamma}_k$ with their sample counterparts provides the final estimator:

$$\hat{\mathbf{\Phi}}_{\text{YW}} = \left[\sum_{t=2}^{T} \mathbf{y}_t \mathbf{y}_{t-1}^\top\right] \left[\sum_{t=1}^{T} \mathbf{y}_t \mathbf{y}_t^\top\right]^{-1}.$$

2.1.2 MAR(1) Yule-Walker

Similarly, for the MAR(1) model, the Yule-Walker equations can be derived with a little more complexity due to the bilinear structure. We begin by applying the Kronecker product of $\mathbf{X}_{t-1}^\top$ to both sides of Eq. (1):

$$\mathbf{X}_t \otimes \mathbf{X}_{t-1}^\top = \mathbf{A}\mathbf{X}_{t-1}\mathbf{B}^\top \otimes \mathbf{X}_{t-1}^\top + \mathbf{Z}_t \otimes \mathbf{X}_{t-1}^\top.$$

Kołodziejski (2025) show that this term can be rearranged using the properties of the Kronecker product to yield

$$\mathbf{X}_t \otimes \mathbf{X}_{t-1}^\top = (\mathbf{A} \otimes \mathbf{I}_n)(\mathbf{X}_{t-1} \otimes \mathbf{X}_{t-1}^\top)(\mathbf{B}^\top \otimes \mathbf{I}_m) + \mathbf{Z}_t \otimes \mathbf{X}_{t-1}^\top,$$

where $\mathbf{I}_m \in \mathbb{R}^{m \times m}$ and $\mathbf{I}_n \in \mathbb{R}^{n \times n}$ are identity matrices. Applying the expectation to both sides yields

$$\mathbb{E}[\mathbf{X}_t \otimes \mathbf{X}_{t-1}^\top] = (\mathbf{A} \otimes \mathbf{I}_n)\mathbb{E}[\mathbf{X}_{t-1} \otimes \mathbf{X}_{t-1}^\top](\mathbf{B}^\top \otimes \mathbf{I}_m) + \mathbb{E}[\mathbf{Z}_t \otimes \mathbf{X}_{t-1}^\top].$$

Noting that $\mathbb{E}[\mathbf{Z}_t \otimes \mathbf{X}_{t-1}^\top] = \mathbf{0}$ due to the white noise assumption, we obtain the MAR(1) Yule-Walker equations:

$$\Gamma_1^{\otimes} = (\mathbf{A} \otimes \mathbf{I}_n)\Gamma_0^{\otimes}(\mathbf{B}^\top \otimes \mathbf{I}_m),$$

where $\Gamma_h^{\otimes} = \mathbb{E}[\mathbf{X}_t \otimes \mathbf{X}_{t-h}^\top]$. Finding an exact solution for the estimated matrices $\Gamma_i^{\otimes}$ may not be possible in general. Kołodziejski (2025) therefore propose to solve the following optimization problem:

$$\min_{\mathbf{A},\mathbf{B}} \|\Gamma_1^{\otimes} - (\mathbf{A} \otimes \mathbf{I}_n)\Gamma_0^{\otimes}(\mathbf{B}^\top \otimes \mathbf{I}_m)\|_F,$$

subject to the normalization constraint $\|\mathbf{A}\|_F = 1$ to ensure uniqueness of the solution. This method simultaneously solves both the Nearest Kronecker Product (NKP) problem and the Yule-Walker equations. As in the original paper, we deploy the L-BFGS-B algorithm (Liu & Nocedal, 1989) to solve this optimization problem. We allow a maximum of 1000 iterations and set the convergence tolerance to 10^{-9}.

To obtain the matrices $\Gamma_k^{\otimes} = \mathbb{E}[\mathbf{X}_t \otimes \mathbf{X}_{t-k}^\top]$ the authors exploit the relationship of the MAR(1) model to the VAR(1) model in its vectorized form (Eq. 2) (Kołodziejski, 2025). Specifically, $\Gamma_k^{\otimes}$ contains the same elements as the autocovariance matrix of the vectorized process $\Gamma_k = \mathbb{E}[\mathrm{vec}(\mathbf{X}_t)\mathrm{vec}(\mathbf{X}_{t-k})^\top]$. However, we were unable to verify the correctness of the permutation mapping proposed by the Kołodziejski (2025); instead, we provide an alternative mapping

$$\Gamma_k^{\otimes} = [\gamma_{ij}^{\otimes}]_{i,j=0,\ldots,mn-1} = [\gamma_{(i//m)\cdot m+(j//n),(i\%n)\cdot m+(j\%m)}]_{i,j=0,\ldots,mn-1}$$

where $\Gamma_k = [\gamma_{ij}]_{i,j=0,\ldots,mn-1}$, % denotes the modulo operation and // denotes integer division. We verified this mapping numerically and show a detailed example in Appendix A.

2.2 Burg's Method

The second approach is a method-of-moments estimator originally developed for scalar time series by Burg (1975). Unlike the LSE method, which minimizes only the forward prediction error, Burg's method estimates the parameters by minimizing the sum of both the forward and backward prediction errors. This approach is particularly effective for short time series as it improves the stability of the estimated model (Kołodziejski, 2025).

2.2.1 VAR(1) Burg

For the VAR(1) model, we define the forward prediction error $\mathbf{f}_t$ and backward prediction error $\mathbf{b}_t$ as residuals of the predictive process:

$$\mathbf{f}_t = \mathbf{y}_t - \mathbf{\Phi}\mathbf{y}_{t-1}, \quad \mathbf{b}_t = \mathbf{y}_{t-1} - \mathbf{\Phi}^\top \mathbf{y}_t.$$

Burg's method seeks the transition matrix $\mathbf{\Phi}$ that minimizes the sum of the squared Euclidean norms of these errors, $\mathcal{J} = \sum_{t=2}^{T}(\|\mathbf{f}_t\|_2^2 + \|\mathbf{b}_t\|_2^2)$.

In the specific case of a first-order model, the algorithm initializes the residuals using the raw data itself. That is, for $m = 1$, the forward and backward prediction terms are simply the observations $\mathbf{y}_t$ and $\mathbf{y}_{t-1}$. While strict minimization of $\mathcal{J}$ leads to a Sylvester equation, we employ the closed-form *Harmonic Mean* estimator typically utilized in the multivariate Burg (Nuttall-Strand) algorithm (Kay, 1988). The estimate $\hat{\mathbf{\Phi}}$ is given by:

$$\hat{\mathbf{\Phi}}_{\text{Burg}} = \left[2\sum_{t=2}^{T}\mathbf{y}_t\mathbf{y}_{t-1}^\top\right]\left[\sum_{t=2}^{T}\left(\mathbf{y}_t\mathbf{y}_t^\top + \mathbf{y}_{t-1}\mathbf{y}_{t-1}^\top\right)\right]^{-1}.$$

2.2.2 MAR(1) Burg

The extension of Burg's method to the matrix-valued case is derived by Kołodziejski (2025). We define the forward and backward prediction residuals for the MAR(1) model as:

$$\mathbf{f}_t = \mathbf{X}_t - \mathbf{A}\mathbf{X}_{t-1}\mathbf{B}^\top, \quad \mathbf{b}_t = \mathbf{X}_{t-1} - \mathbf{A}\mathbf{X}_t\mathbf{B}^\top.$$

Ideally, we aim to find the matrices $\mathbf{A}$ and $\mathbf{B}$ that minimize the total error energy:

$$\min_{\mathbf{A},\mathbf{B}} \sum_{t=2}^{T} \left(\|\mathbf{f}_t\|_F^2 + \|\mathbf{b}_t\|_F^2 \right).$$

As shown by Kołodziejski (2025), no closed-form solution exists for both $\mathbf{A}$ and $\mathbf{B}$ simultaneously. Instead, an iterative approach is employed. If we fix $\mathbf{B}$, the optimal $\mathbf{A}$ is obtained by setting the derivative of the cost function to zero:

$$\mathbf{A} = \left(\sum_{t=2}^{T} \left[\mathbf{f}_t \mathbf{B} \mathbf{b}_{t-1}^\top + \mathbf{b}_{t-1} \mathbf{B} \mathbf{f}_t^\top \right] \right) \left(\sum_{t=2}^{T} \left[(\mathbf{B}\mathbf{b}_{t-1}^\top)^2 + (\mathbf{B}\mathbf{f}_t^\top)^2 \right] \right)^{-1}.$$

Conversely, fixing $\mathbf{A}$ yields the update rule for $\mathbf{B}$:

$$\mathbf{B}^\top = \left(\sum_{t=2}^{T} \left[(\mathbf{A}\mathbf{b}_{t-1})^2 + (\mathbf{A}\mathbf{f}_t)^2 \right] \right)^{-1} \sum_{t=2}^{T} \left[\mathbf{b}_{t-1}^\top \mathbf{A}^\top \mathbf{f}_t + \mathbf{f}_t^\top \mathbf{A}^\top \mathbf{b}_{t-1} \right].$$

Here, we use the notation $\mathbf{X}^2 := \mathbf{X}^\top \mathbf{X}$ and assume $\mathbf{f}_t, \mathbf{b}_t$ are the zero-th order errors (raw data). These update rules extend the harmonic mean logic by balancing the forward and backward projection energies in the matrix space. The algorithm alternates between these updates until convergence to get the final estimates $\hat{\mathbf{A}}_{\text{Burg}}$ and $\hat{\mathbf{B}}_{\text{Burg}}$, with $\mathbf{A}$ normalized at each step ($\|\mathbf{A}\|_F = 1$) to ensure uniqueness. We, again, allow a maximum of 1000 iterations and set the convergence tolerance to 10^{-9}.

Symmetry Requirements. The Burg method assumes the process is time-reversible, such that $\mathbf{X}_t \approx \mathbf{A}\mathbf{X}_{t-1}\mathbf{B}^\top$ and $\mathbf{X}_{t-1} \approx \mathbf{A}\mathbf{X}_t\mathbf{B}^\top$ share the same coefficient matrices. Substituting the forward model into the backward model yields $\mathbf{X}_{t-1} \approx \mathbf{A}^2\mathbf{X}_{t-1}(\mathbf{B}^\top)^2$. It can be shown that this equality holds for general $\mathbf{X}_{t-1}$ if and only if $\mathbf{A}$ and $\mathbf{B}$ are symmetric matrices (Appendix B).

2.3 Least Squares Estimation

The LSE approach minimizes only the forward prediction error. While computationally simple, it does not guarantee model stability in the same manner as Burg's method.

2.3.1 VAR(1) OLS

The LSE estimator for the VAR(1) model minimizes $\sum_{t=2}^{T} \|\mathbf{y}_t - \mathbf{\Phi}\mathbf{y}_{t-1}\|_2^2$. Finding the derivative with respect to $\mathbf{\Phi}$ and setting it to zero yields the closed-form solution (Lütkepohl, 2005):

$$\hat{\mathbf{\Phi}}_{LSE} = \left[\sum_{t=2}^{T} \mathbf{y}_t \mathbf{y}_{t-1}^\top \right] \left[\sum_{t=2}^{T} \mathbf{y}_{t-1} \mathbf{y}_{t-1}^\top \right]^{-1}.$$

This estimator is equivalent to the ordinary least squares (OLS) solution for linear regression and almost identical to the Yule-Walker estimator, differing only in the summation limits.

2.3.2 MAR(1) LSE

For the MAR(1) model, we minimize $\sum_{t=2}^{T} \|\mathbf{X}_t - \mathbf{A}\mathbf{X}_{t-1}\mathbf{B}^\top\|_F^2$. As with Burg's method, we employ an alternating minimization (Chen et al., 2021). Setting the derivative of the cost function to zero and fixing one of the matrices yields closed-form update rules. Fixing $\mathbf{B}$ yields:

$$\mathbf{A} = \left[\sum_{t=2}^{T} \mathbf{X}_t \mathbf{B} \mathbf{X}_{t-1}^\top\right] \left[\sum_{t=2}^{T} \mathbf{X}_{t-1} \mathbf{B}^\top \mathbf{B} \mathbf{X}_{t-1}^\top\right]^{-1}.$$

Fixing $\mathbf{A}$ yields:

$$\mathbf{B} = \left[\sum_{t=2}^{T} \mathbf{X}_t^\top \mathbf{A} \mathbf{X}_{t-1}\right] \left[\sum_{t=2}^{T} \mathbf{X}_t^\top \mathbf{A}^\top \mathbf{A} \mathbf{X}_{t-1}\right]^{-1}.$$

This iterative process is identical in structure to the Burg update but omits the backward error terms in the summations. Again, we normalize $\mathbf{A}$ at each step to ensure uniqueness and obtain the final estimates $\hat{\mathbf{A}}_{LSE}$ and $\hat{\mathbf{B}}_{LSE}$ upon convergence. We, again, allow a maximum of 1000 iterations and set the convergence tolerance to 10^{-9}.

2.4 Vectorized Estimation

The two-step estimation approach relies on approximating the estimated VAR(1) transition matrix $\mathbf{\Phi} \in \mathbb{R}^{mn \times mn}$ with a MAR-compliant structure $\mathbf{B} \otimes \mathbf{A}$. We implement the algorithm of Van Loan & Pitsianis (1993), which treats the NKP problem as a rank-one approximation of a rearranged matrix. Specifically, $\mathbf{\Phi}$ is mapped to a rearrangement matrix $\mathcal{R}(\mathbf{\Phi}) \in \mathbb{R}^{m^2 \times n^2}$. The estimates $\hat{\mathbf{A}}$ and $\hat{\mathbf{B}}$ are recovered from the primary singular components:

$$\text{vec}(\hat{\mathbf{A}}) = \sqrt{\sigma_1}\mathbf{u}_1, \quad \text{vec}(\hat{\mathbf{B}}) = \sqrt{\sigma_1}\mathbf{v}_1,$$

where σ_i are the singular values and $\mathbf{u}_i, \mathbf{v}_i$ are the corresponding left and right singular vectors of $\mathcal{R}(\mathbf{\Phi})$. Consistent with Kołodziejski (2025), we normalize the result such that $\|\hat{\mathbf{A}}\|_F = 1$ to ensure uniqueness.

3 Methods

In this section, we outline the experimental procedures employed to evaluate the performance of the MAR(1) and VAR(1) estimators. We conduct a comprehensive simulation study to assess estimation accuracy under various data-generating processes. Additionally, we analyze the structural sensitivity of the MAR(1) model using Kronecker Energy metrics.

3.1 Simulation Study

To evaluate the performance of the MAR(1) and VAR(1) estimators, we conduct a series of controlled experiments using synthetic data like Kołodziejski (2025). The

authors use a VAR(1) process to generate the data and compare the estimation performance. Since the MAR(1) model is misspecified for data generated by a VAR(1) process, we extend their experiments by also generating data from a MAR(1) process. Furthermore, we consider both symmetric and non-symmetric coefficient matrices for the MAR(1) process to investigate the impact of symmetry on the estimation performance, especially for the Burg method.

Interestingly, the results reported by Kołodziejski (2025) are counterintuitive: the MAR(1) estimators outperform the VAR(1) estimator even though the data is generated from a VAR(1) process. Furthermore, the p-values from Mardia's test reported by Kołodziejski (2025) were suspiciously low, indicating possible issues with the implementation. While we were not able to reproduce their results, we were able to match their findings by computing the metrics between the predictions on the training data and the actual test data. Thereby, comparing unrelated samples. We include this "flawed" setup to replicate the results of (Kołodziejski, 2025).

Experimental Setup: For each experiment, we generate 100 datasets of varying lengths $T \in \{100, 200, \ldots, 500\}$ and square matrix sizes $m \in \{2, 3, \ldots, 10\}$. Each simulation produces a time series of $T + 100$ samples, where the first T are used for training and the final 100 observations serve as a held-out test set.

Performance is quantified via the average Root Mean Square Error (RMSE):

$$\text{RMSE} = \sqrt{\frac{1}{Tm^2} \sum_{t=1}^{T} \|\mathbf{X}_t - \hat{\mathbf{X}}_t\|_F^2}$$

Following Kołodziejski (2025), we also evaluate the joint multivariate normality of the residuals using Mardia's test (Mardia, 1970) for $m < 10$.

Data Generating Processes: We consider three distinct data-generating processes (DGP) configurations to test the estimators under correct and incorrect specifications:

1. **VAR(1):** A general vector process (Eq. 3) with a randomly initialized transition matrix $\mathbf{\Phi}$ rescaled for stationarity.
2. **MAR(1) Non-Symmetric:** A structured process where $\mathbf{\Phi} = \mathbf{B} \otimes \mathbf{A}$ with standard normal factor matrices.
3. **MAR(1) Symmetric:** A structured process where factor matrices are symmetrized, specifically testing the assumptions of the Burg method.

VAR(1) Process: The parameters of the VAR(1) model are initialized randomly and adjusted to ensure stationarity. A matrix $\mathbf{\Phi}^{(r)} \in \mathbb{R}^{m^2 \times m^2}$ is drawn from a standard normal distribution and rescaled to satisfy the spectral radius condition $\rho(\mathbf{\Phi}) < 1$ as follows:

$$\mathbf{\Phi} := \frac{1}{\rho(\mathbf{\Phi}^{(r)}) + 1} \mathbf{\Phi}^{(r)} \tag{4}$$

The white noise covariance matrix $\mathbf{\Sigma} \in \mathbb{R}^{m^2 \times m^2}$ is constructed to be positive semi-definite. We sample $\mathbf{S}$ from a standard normal distribution, symmetrize

it via $\mathbf{S}^s = \frac{1}{2}(\mathbf{S} + \mathbf{S}^\top)$, and apply an eigendecomposition $\mathbf{S}^s = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^\top$. Defining $\mathbf{\Lambda}^+ = \text{diag}(|\lambda_i|)$, the covariance is set to $\mathbf{\Sigma} := \mathbf{Q}\mathbf{\Lambda}^+\mathbf{Q}^\top$.

We then generate the vector time series $\mathbf{y}_t$ according to the VAR(1) equation (3), initializing $\mathbf{y}_0$ as a zero vector and using white noise $\mathbf{e}_t \sim \mathcal{N}(0, \mathbf{\Sigma})$. To use MAR(1) models on this data, we apply the reverse vectorization to reshape $\mathbf{y}_t$ into matrix form $\mathbf{X}_t \in \mathbb{R}^{m \times m}$.

MAR(1) Process: For data generated directly from a MAR(1) process (1), we use the same approach as for the VAR(1) process but generate the coefficient matrix $\mathbf{\Phi}$ as a Kronecker product of two smaller matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{m \times m}$.

In the non-symmetric case, we draw $\mathbf{A}$ and $\mathbf{B}$ from a standard normal distribution and compute their Kronecker product $\mathbf{\Phi} = \mathbf{B} \otimes \mathbf{A}$.

For the symmetric case, we also draw $\mathbf{A}$ and $\mathbf{B}$ from a standard normal distribution, but symmetrize them via $\mathbf{A}^s = \frac{1}{2}(\mathbf{A} + \mathbf{A}^\top)$ and $\mathbf{B}^s = \frac{1}{2}(\mathbf{B} + \mathbf{B}^\top)$. The Kronecker product is then computed as $\mathbf{\Phi} = \mathbf{B}^s \otimes \mathbf{A}^s$.

3.2 Structural Sensitivity Analysis

While the MAR(1) model offers a parsimonious alternative to the VAR(1), its validity depends on the transition matrix residing in the Kronecker subspace. To investigate the practical limitations of this specification, we analyze the Kronecker Energy (KE) (Van Loan & Pitsianis, 1993).

We define KE as the energy ratio of the first singular value of the rearrangement matrix $\mathcal{R}(\mathbf{\Phi})$:

$$\text{KE} = \frac{\sigma_1^2}{\sum_i \sigma_i^2}.$$

This metric quantifies how "close" a general VAR(1) process is to a MAR(1) structure. To observe the degradation of the MAR(1) estimator, we construct a mixture transition matrix:

$$\mathbf{\Phi}_{mix} = (1 - \alpha)\mathbf{\Phi}_{VAR} + \alpha\mathbf{\Phi}_{MAR},$$

By varying $\alpha \in [0, 1]$, we move the DGP along a spectrum from a general high-dimensional system to a strictly structured MAR(1) model. To maintain stability, we rescale $\mathbf{\Phi}_{mix}$ to satisfy the spectral radius condition $\rho(\mathbf{\Phi}_{mix}) < 1$ like in Eq. (4). For each level of α, we record the KE and the resulting RMSE to evaluate how predictive accuracy degrades as the DGP moves out of the specialized Kronecker subspace. We conduct this analysis for matrix sizes $m \in \{2, 3, \ldots, 10\}$ and time series length $T = 500$ for the MAR(1) (LSE) and VAR(1) (OLS) estimators.

4 Results

In this section, we present the results of our simulation study evaluating the performance of the MAR(1) and VAR(1) estimators under various data-generating processes and configurations. We also explore the relationship between estimation

accuracy and the structural properties of the transition matrices using Kronecker Energy.

Across all setups, Mardia's test consistently confirmed the multivariate normality of the residuals in over 95% of the simulations for both MAR(1) and VAR(1) models, indicating that the estimators produced statistically sound residuals (see Appendix C).

Additionally, we recorded the spectral radii of the estimated transition to investigate the stability of the fitted models. As Appendix D shows, the majority of estimated models were stable ($\rho < 1$), with only a small fraction exhibiting instability ($\leq 1\%$). Unstable estimates were more prevalent for the Yule-Walker method, particularly in the VAR setup. Conversely, Burg's method had slightly more unstable estimates in the non-symmetric MAR compared to the symmetric setup. Finally, there were no unstable estimates in the symmetric MAR setup for any method.

4.1 Yule-Walker

The performance of the Yule-Walker approach across the four experimental setups is illustrated in Figure (1). In the flawed VAR setup, the MAR-based estimators appear to outperform the VAR(1) estimator like in (Kołodziejski, 2025), though this stems from a methodological mismatch rather than structural superiority. Under a general VAR(1) process, the VAR(1) Yule-Walker estimator significantly outperforms both MAR(1) approaches, as the RMSE for the MAR models increases sharply with matrix size m. Conversely, when the data follows a true MAR(1) process, the parsimonious MAR-based estimators show superior performance compared to the VAR(1) model. There seems to be no significant difference between the symmetric and non-symmetric MAR(1) setups for the Yule-Walker method.

4.2 Burg's Method

Figure (2) presents the RMSE results for Burg's method. In the flawed VAR setup, the MAR(1) Burg estimator again appears to outperform the VAR(1) estimator. In the general VAR(1) setup, the VAR(1) Burg estimator outperforms the MAR(1) Burg estimators very similarly to the Yule-Walker results. However, in the non-symmetric MAR setup, the VAR(1) and vectorized MAR(1) Burg estimators outperform the structured MAR(1) Burg estimator significantly, while the vectorized MAR(1) Burg estimator performs best. This indicates that the Burg method is sensitive to the symmetry assumption inherent in the structured MAR(1) model. In the symmetric MAR setup, the structured MAR(1) Burg estimator and its vectorized counterpart perform best and similarly, while the VAR(1) Burg estimator performs worst.

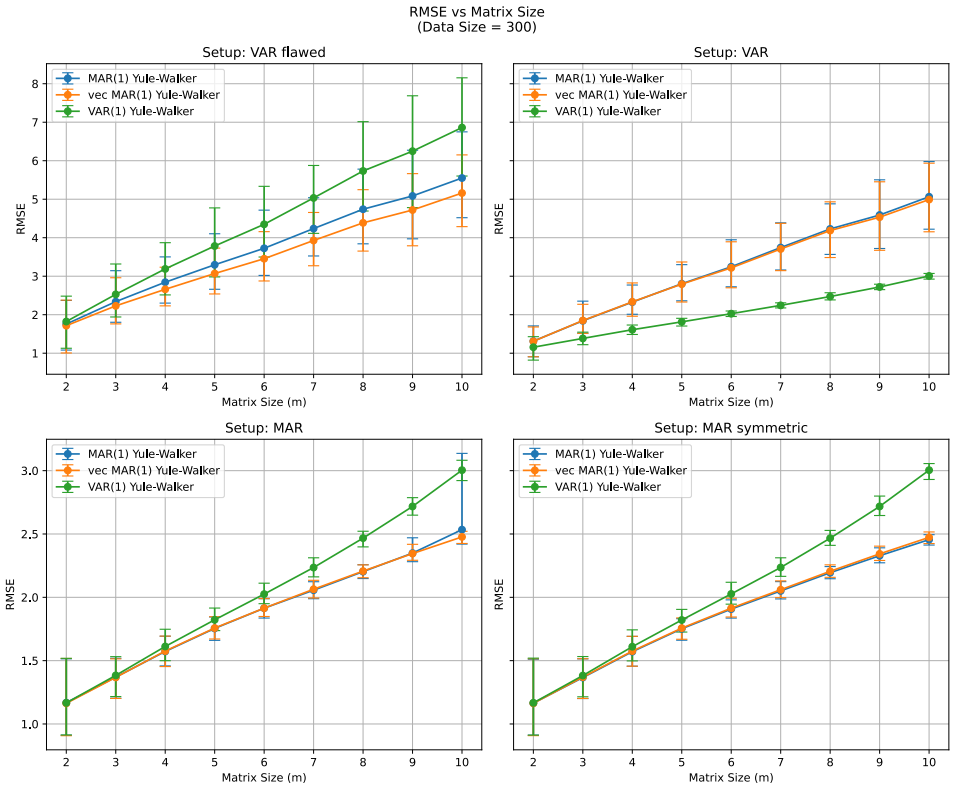

Figure 1. RMSE of the Yule-Walker approach for the MAR(1) and VAR(1) models for a time series of length 300 across the four setups. For details, see Section 3.1.

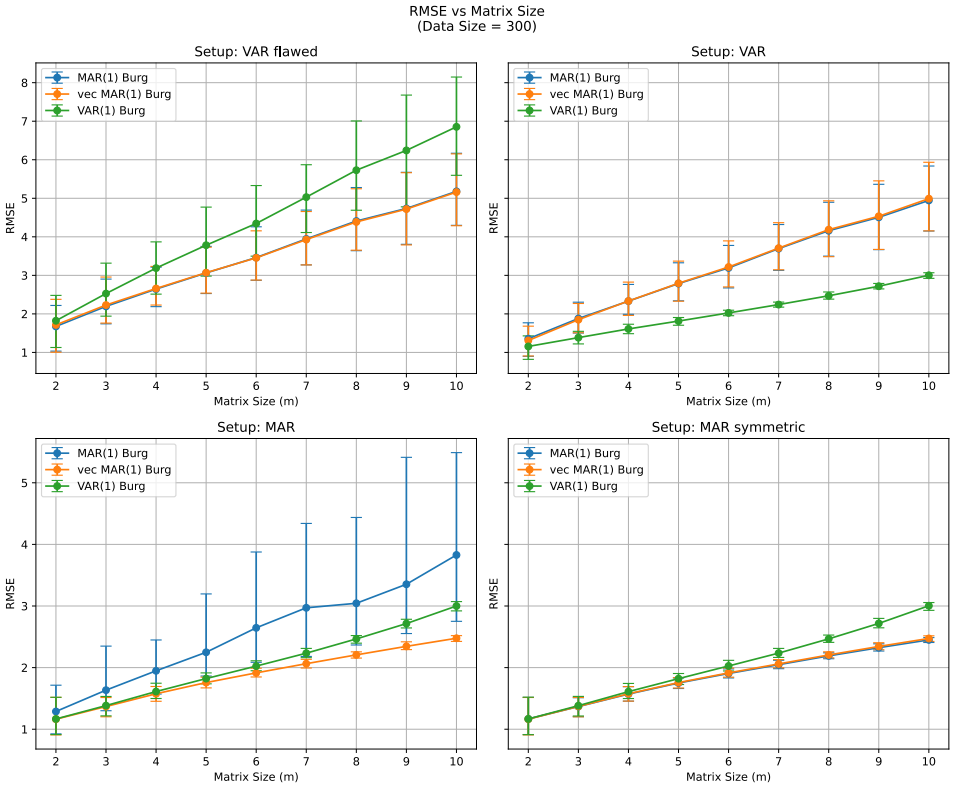

Figure 2. RMSE of the Burg approach for the MAR(1) and VAR(1) models for a time series of length 300 across the four setups. For details, see Section 3.1.

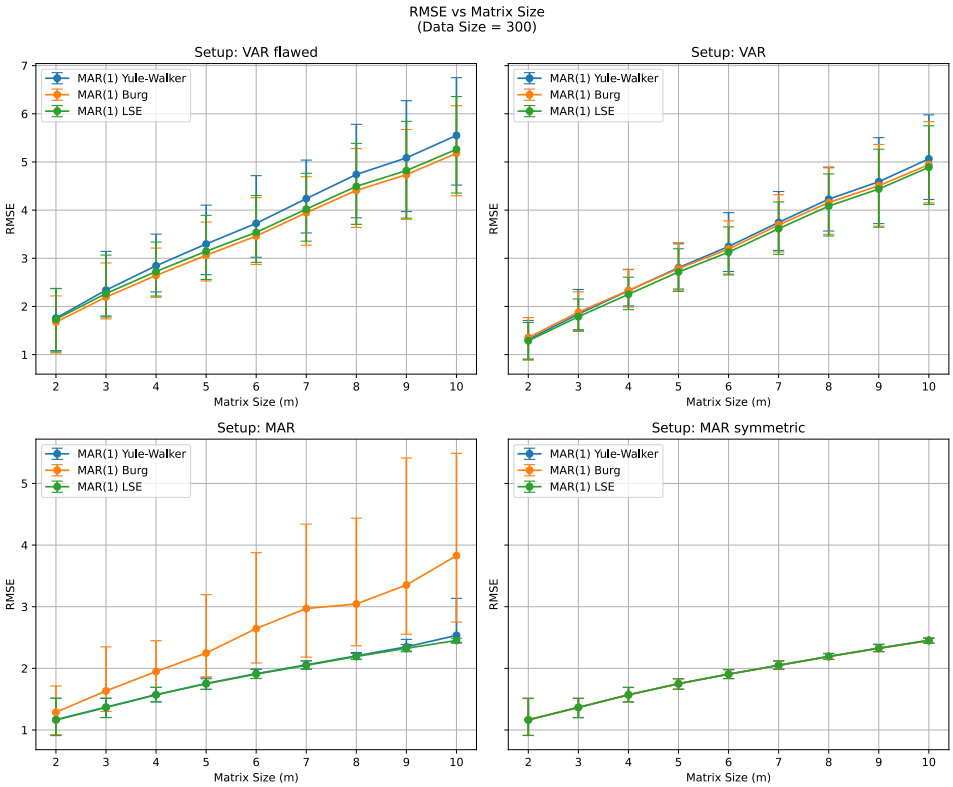

Figure 3. RMSE of the YW, Burg and LSE models for a time series of length 300 across the four setups. For details, see Section 3.1.

4.3 Comparison of YW, Burg, LSE

The comparative performance of the YW, Burg, and LSE estimators is shown in Figure (3). In all but the non-symmetric MAR setup, all three estimators perform similarly. However, in the non-symmetric MAR setup, Burg's method performs significantly worse than both the YW and LSE methods and has a higher variance in its RMSE. This further highlights the sensitivity of Burg's method to the symmetry assumption of the MAR(1) model. Together with the stability results discussed earlier, this suggests that LSE seems to be the most robust estimation method for MAR(1) processes. Even though LSE does not guarantee stable estimates, in our experiments, it produced stable models more consistently than both YW and Burg's method across all setups.

4.4 Comparing RMSE and Kronecker Energy

The relationship between structural misspecification and predictive accuracy is quantified in Figure (4). The results indicate a strong negative correlation between Kronecker Energy and the RMSE of the MAR(1) estimator. As the transition matrix moves closer to a perfect Kronecker structure (KE $\to$ 1.0), the MAR(1) RMSE decreases significantly but seems to only reach the baseline level of the VAR(1) model when KE $= 1$. In contrast, the VAR(1) estimator remains relatively indifferent to the Kronecker structure. This confirms that the MAR model's parsimony only provides a competitive advantage when the underlying process resides sufficiently close to the Kronecker subspace.

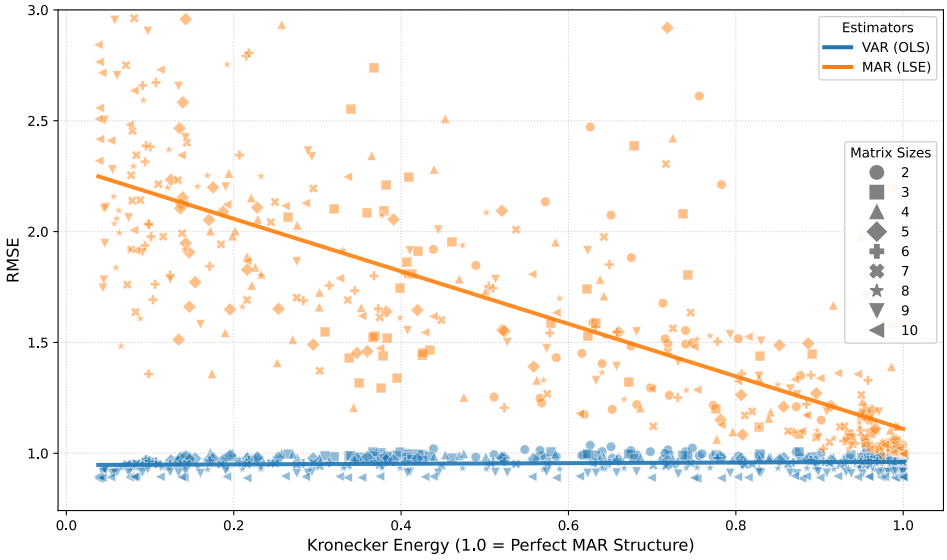

Figure 4. Comparison of Kronecker Energy vs. RMSE for the VAR(1) (OLS) and MAR(1) (LSE) models for matrix sizes $m = 2, 3, \ldots, 10$ and $T = 500$. The different symbols represent single experimental runs and the lines represent a regression fit through the points. For details, see Section 3.2.

5 Discussion

The results of this study highlight the critical trade-offs between model parsimony and structural accuracy in matrix-valued time series analysis. While MAR models significantly reduce parameter dimensionality from m^2n^2 in a full VAR to $m^2 + n^2$, their effectiveness is highly dependent on the underlying data-generating process. Our analysis of Kronecker Energy (KE) confirms that the MAR model's primary advantage only translates to superior predictive accuracy when the true transition matrix resides close to the Kronecker subspace. As shown in the sensitivity analysis, the RMSE of the MAR(1) estimator decreases significantly as the KE approaches 1.0, indicating a perfect MAR structure. Conversely, the VAR(1) estimator remains relatively indifferent to the Kronecker structure, maintaining stable performance across the KE spectrum. This suggests that in high-dimensional applications where row-column dependencies are weak, the bias introduced by the MAR specification may outweigh the benefits of reduced parameter variance.

A central finding of this reproduction study is the sensitivity of various estimation methods to their underlying assumptions, particularly regarding symmetry. Burg's method, while effective for short time series due to its minimization of both forward and backward prediction errors, proved highly sensitive to the symmetry of the coefficient matrices. In non-symmetric MAR setups, Burg's method performed significantly worse than both Yule-Walker and Least Squares estimators and exhibited higher variance in its RMSE. However, when the symmetry constraints $A = A^\top$ and $B = B^\top$ were met, the MAR(1) Burg estimator and its vectorized counterpart achieved similar performance.

Regarding numerical stability, the Yule-Walker estimator exhibited the highest frequency of unstable estimates, particularly within the general VAR setup, which we

argue can be attributed to the inherent limitations of the numerical optimization required for the MAR(1) case. The reliance on this iterative numerical optimization process introduces potential convergence issues, especially in high-dimensional settings where the objective function surface may be less well-behaved. Consequently, the lack of a direct analytical solution may lead the algorithm to converge on parameters that satisfy the normalization constraint $\|A\|_F = 1$ but fail to stay within the unit circle required for the causality condition $\rho(B \otimes A) < 1$. In contrast, the LSE method achieved the lowest overall instability. Notably, the imposition of symmetry in the MAR(1) setup eliminated unstable estimates across all estimators, suggesting that structural constraints can enhance numerical robustness.

Furthermore, our results provide a point of comparison for the findings reported in prior literature. In our attempts to align with the results of Kołodziejski (2025), we observed that the counterintuitive trend—where a MAR(1) model appears to outperform a correctly specified VAR(1) process—could have emerged primarily in a specific experimental configuration where unrelated samples were compared. When adhering to standard evaluation protocols, the VAR(1) estimator consistently outperformed the MAR(1) estimators on data generated from a VAR(1) process. This observation is consistent with the residual analysis; while our primary simulations yielded Mardia's test median p-values close to 1.0, indicating sound residuals, the lower median p-values (0.34) reported in (Kołodziejski, 2025), only appeared when using the flawed evaluation setup. These points of divergence suggest that the original findings may have been influenced by a methodological error rather than inherent model superiority.

6 Conclusion

This study evaluated Matrix Autoregressive (MAR) models against classical Vector Autoregressive (VAR) frameworks, identifying critical trade-offs between parameter parsimony and structural accuracy. Our findings demonstrate that while MAR estimators provide superior accuracy when data adheres to a Kronecker structure, this advantage is highly sensitive to model misspecification and diminishes as the data-generating process diverges from this assumption. Furthermore, our replication of prior results suggests that perceived MAR superiority in previous publications (Kołodziejski, 2025) may be influenced by specific evaluation configurations rather than inherent model advantages.

Regarding estimation techniques, Burg's method proved effective in symmetric environments but showed high sensitivity to non-symmetric processes. The Yule-Walker approach frequently exhibited numerical instability, likely due to the complexities of iterative L-BFGS-B optimization in the matrix-valued case. In contrast, the Least Squares Estimation method consistently achieved the lowest overall instability and demonstrated robust performance across both correctly and incorrectly specified environments. Consequently, due to its computational efficiency and reliable performance across all tested setups, we recommend the LSE method as the primary choice for matrix-valued autoregressive modeling even though it does not guarantee stable estimates.

References

Burg, J. P. (1975). *Maximum Entropy Spectral Analysis*. Stanford University.

Chen, R., Xiao, H., & Yang, D. (2021). Autoregressive models for matrix-valued time series. *Journal of Econometrics*, *222*(1, Part B), 539–560. doi: 10.1016/j.jeconom.2020.07.015

Kay, S. M. (1988). *Modern Spectral Estimation: Theory and Application*. Prentice Hall.

Kołodziejski, K. (2025). *Estimation methods of Matrix-valued AR model* (No. arXiv:2505.15220). arXiv. doi: 10.48550/arXiv.2505.15220

Li, Z., & Xiao, H. (2021). *Multi-linear Tensor Autoregressive Models* (No. arXiv:2110.00928). arXiv. doi: 10.48550/arXiv.2110.00928

Liu, D. C., & Nocedal, J. (1989). On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, *45*(1), 503–528. doi: 10.1007/BF01589116

Lütkepohl, H. (2005). *New Introduction to Multiple Time Series Analysis*. Berlin, Heidelberg: Springer. doi: 10.1007/978-3-540-27752-1

Mardia, K. V. (1970). Measures of Multivariate Skewness and Kurtosis with Applications. *Biometrika*, *57*(3), 519–530. doi: 10.2307/2334770

Samadi, S. Y., & Billard, L. (2025). On a matrix-valued autoregressive model. *Journal of Time Series Analysis*, *46*(1), 3–32. doi: 10.1111/jtsa.12748

Van Loan, C. F., & Pitsianis, N. (1993). Approximation with Kronecker Products. In M. S. Moonen, G. H. Golub, & B. L. R. De Moor (Eds.), *Linear Algebra for Large Scale and Real-Time Applications* (pp. 293–314). Dordrecht: Springer Netherlands. doi: 10.1007/978-94-015-8196-7_17

Appendix

A Numerical Verification of Autocovariance Mappings

This appendix provides a numerical verification of the permutation required to transform a vectorized autocovariance matrix Γ_1 into the Kronecker-structured matrix $\Gamma_1^{\otimes} = \mathbb{E}[X_t \otimes X_{t-1}^T]$ used in Matrix Yule-Walker equations.

Consider a $MAR(1)$ model with 2×2 matrices ($m = n = 2$):

$$X_t = \begin{bmatrix} 1 & 3 \\ 2 & 4 \end{bmatrix}, \quad X_{t-1} = \begin{bmatrix} 5 & 7 \\ 6 & 8 \end{bmatrix} \implies \Gamma_1^{\otimes} = X_t \otimes X_{t-1}^T = \begin{bmatrix} 5 & 6 & 15 & 18 \\ 7 & 8 & 21 & 24 \\ 10 & 12 & 20 & 24 \\ 14 & 16 & 28 & 32 \end{bmatrix}$$

Kołodziejski (2025) suggest the following indexing to map elements from Γ_1 to $\Gamma_1^{\otimes}$:

$$\Gamma_k^{\otimes} = [\gamma_{ij}^{\otimes}]_{i,j=0,\ldots,mn-1} = [\gamma_{(i\%m)\cdot n+(j\%n),(i//m)\cdot m+(j//m)}]_{i,j=0,\ldots,mn-1}$$

for a column-major vectorization. Here, % denotes the modulo operation and // denotes integer division. However, numerical tests indicate this mapping does not yield the correct $\Gamma_1^{\otimes}$. To illustrate, we compute Γ_1 using the column-major vectorizations, then apply the author's mapping and a corrected mapping to verify the results.

Flattening by columns, we have $\text{vec}_{\text{row}}(X_t) = [1, 3, 2, 4]^T$ and $\text{vec}_{\text{row}}(X_{t-1}) = [5, 7, 6, 8]^T$. The vectorized autocovariance $\Gamma_{1,\text{row}} = \text{vec}_{\text{row}}(X_t)\text{vec}_{\text{row}}(X_{t-1})^T$ and its corrected permutation are:

$$\Gamma_{1,\text{row}} = \begin{bmatrix} 5 & 7 & 6 & 8 \\ 15 & 21 & 18 & 24 \\ 10 & 14 & 12 & 16 \\ 20 & 28 & 24 & 32 \end{bmatrix} \xrightarrow{\text{Authors}} \begin{bmatrix} 5 & 10 & 6 & 12 \\ 15 & 20 & 18 & 24 \\ 7 & 14 & 8 & 16 \\ 21 & 28 & 24 & 32 \end{bmatrix} \xrightarrow{\text{Corrected}} \begin{bmatrix} 5 & 6 & 15 & 18 \\ 7 & 8 & 21 & 24 \\ 10 & 12 & 20 & 24 \\ 14 & 16 & 28 & 32 \end{bmatrix}$$

The corrected row-major mapping is:

$$\Gamma_k^{\otimes} = [\gamma_{ij}^{\otimes}]_{i,j=0,\ldots,mn-1} = [\gamma_{(i//m)\cdot m+(j//n),(i\%n)\cdot m+(j\%m)}]_{i,j=0,\ldots,mn-1}$$

B Symmetry in Burg Estimation

Lemma 1. *For the Matrix Autoregressive model of order 1 (MAR(1)), defined as* $\mathbf{X}_t = \mathbf{A}\mathbf{X}_{t-1}\mathbf{B}^\top + \mathbf{Z}_t$*, the Burg estimator is structurally consistent with the statistical time-reversibility of a stationary process if and only if the coefficient matrices are symmetric (*$\mathbf{A} = \mathbf{A}^\top$ *and* $\mathbf{B} = \mathbf{B}^\top$*).*

Proof. The Burg estimator is defined by minimizing the sum of forward and backward prediction errors. According to the MAR(1) definition in Eq. (1), the Burg method implicitly assumes a backward predictor of the form:

$$\hat{\mathbf{X}}_{t-1}^{\text{Burg}} = \mathbf{A}\mathbf{X}_t\mathbf{B}^\top. \tag{5}$$

To verify if this assumption holds for a general stationary process, we utilize the vectorization identity from Eq. (2). The process can be written as a VAR(1) model $\mathbf{y}_t = \mathbf{\Phi}\mathbf{y}_{t-1} + \mathbf{e}_t$, where $\mathbf{y}_t = \text{vec}(\mathbf{X}_t)$ and $\mathbf{\Phi} = \mathbf{B} \otimes \mathbf{A}$.

For a stationary process with normalized covariance, Lütkepohl (2005) establishes that the autocovariance at lag -1 is the transpose of the autocovariance at lag 1 ($\mathbf{\Gamma}_y(-1) = \mathbf{\Gamma}_y(1)^\top$). Consequently, the statistically optimal backward transition matrix is the transpose of the forward matrix ($\mathbf{\Phi}^\top$). Applying this to the Kronecker structure, the true statistical backward predictor is:

$$\hat{\mathbf{y}}_{t-1} = \mathbf{\Phi}^\top \mathbf{y}_t = (\mathbf{B} \otimes \mathbf{A})^\top \mathbf{y}_t = (\mathbf{B}^\top \otimes \mathbf{A}^\top)\mathbf{y}_t.$$

De-vectorizing this result using the inverse of the vectorization identity ($\text{vec}(\mathbf{LMR}) = (\mathbf{R}^\top \otimes \mathbf{L})\text{vec}(\mathbf{M})$) yields the true matrix-valued backward predictor:

$$\hat{\mathbf{X}}_{t-1}^{\text{Stat}} = \mathbf{A}^\top \mathbf{X}_t \mathbf{B}. \tag{6}$$

For the Burg estimator to be structurally consistent, its assumed backward dynamic (5) must match the statistical reality (6). Equating $\mathbf{A}\mathbf{X}_t\mathbf{B}^\top = \mathbf{A}^\top\mathbf{X}_t\mathbf{B}$ holds for general $\mathbf{X}_t$ if and only if $\mathbf{A} = \mathbf{A}^\top$ and $\mathbf{B} = \mathbf{B}^\top$. □

C Mardia's Test Results

The following tables present the p-values from Mardia's tests for multivariate normality of the residuals from the various estimated models across different setups.

Table 1. P-values of Mardia's tests for model residuals: **VAR FLAWED**

Model	Minimum	Quantile 1%	Quantile 5%	Median
MAR(1) Burg	0.0003	0.0115	0.0849	0.3459
MAR(1) LSE	0.0000	0.0141	0.0914	0.3475
MAR(1) Yule-Walker	0.0000	0.0152	0.0900	0.3471
VAR(1) Burg	0.0000	0.0146	0.0916	0.3454
VAR(1) Yule-Walker	0.0000	0.0138	0.0911	0.3453
vec MAR(1) Burg	0.0000	0.0195	0.0886	0.3464
vec MAR(1) Yule-Walker	0.0000	0.0188	0.0885	0.3462

Table 2. P-values of Mardia's tests for model residuals: **VAR**

Model	Minimum	Quantile 1%	Quantile 5%	Median
MAR(1) Burg	0.0000	0.0037	0.1072	0.9997
MAR(1) LSE	0.0000	0.0042	0.1256	0.9996
MAR(1) Yule-Walker	0.0000	0.0053	0.1314	0.9994
VAR(1) Burg	0.0000	0.0062	0.1206	0.9995
VAR(1) Yule-Walker	0.0000	0.0078	0.1203	0.9995
vec MAR(1) Burg	0.0000	0.0111	0.1183	0.9997
vec MAR(1) Yule-Walker	0.0000	0.0108	0.1199	0.9996

Table 3. P-values of Mardia's tests for model residuals: **MAR**

Model	Minimum	Quantile 1%	Quantile 5%	Median
MAR(1) Burg	0.0000	0.0094	0.1192	0.9996
MAR(1) LSE	0.0000	0.0038	0.1194	0.9996
MAR(1) Yule-Walker	0.0000	0.0022	0.1252	0.9993
VAR(1) Burg	0.0000	0.0058	0.1081	0.9995
VAR(1) Yule-Walker	0.0000	0.0040	0.1078	0.9994
vec MAR(1) Burg	0.0000	0.0031	0.1243	0.9996
vec MAR(1) Yule-Walker	0.0000	0.0027	0.1269	0.9996

Table 4. P-values of Mardia's tests for model residuals: **MAR SYMMETRIC**

Model	Minimum	Quantile 1%	Quantile 5%	Median
MAR(1) Burg	0.0000	0.0047	0.1112	0.9996
MAR(1) LSE	0.0000	0.0056	0.1122	0.9996
MAR(1) Yule-Walker	0.0000	0.0048	0.1096	0.9995
VAR(1) Burg	0.0000	0.0045	0.1080	0.9994
VAR(1) Yule-Walker	0.0000	0.0053	0.1083	0.9994
vec MAR(1) Burg	0.0000	0.0030	0.1144	0.9994
vec MAR(1) Yule-Walker	0.0000	0.0031	0.1139	0.9995

D Stability Analysis

We analyze the stability of the estimated models by computing the spectral radius of the estimated coefficient matrices. As shown in Figure 5, the Yule-Walker estimator exhibits a higher percentage of unstable estimates across all setups, while the LSE achieves the lowest instability. Interestingly, for the symmetric MAR(1) setup, no unstable estimates are observed for any estimator.

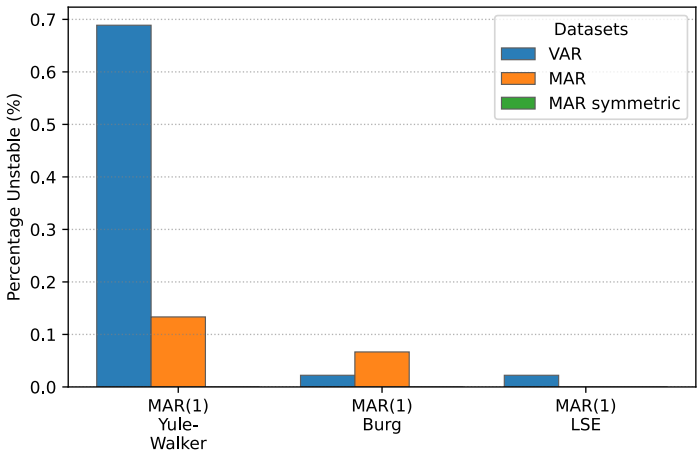

Figure 5. Percentage of unstable estimates (spectral radius ≥ 1) for different estimators and setups across all datasets. For details, see Section 3.