Vector-Autoregressive models via tensor decomposition

Yvann Le Fay yvann.lefay@ensae.fr, Antoine Schoonaert antoine.schoonaert@etu.emse.fr

January 9, 2024

1 Introduction and contributions

A *Vector Autoregression* (VAR) model is the natural extension of the *Autoregressive* model (AR) to multivariate time series. This type of model is intensively used in the industry for describing and performing forecasts on time series with possible complex underlying dynamics. Some examples of time series drawn from the financial industry, which are modelled by VAR models, are GDP Growth Rate, Treasury Bond rates, CPI Index as well as Stock prices. Formally, a time series $\{y\} \in (\mathbb{R}^N)^{\mathbb{N}}$ is said to follow a VAR model of order $P \geq 0$ if the y_t's satisfy, for any $t \geq P$,

$$y_t = A_1 y_{t-1} + \dots A_P y_{t-P} + \varepsilon_t, \tag{1}$$

where the A_i's are $N \times N$ matrices, the ε's, called exogenous innovations, are independent and identically distributed random variables, typically Gaussian variables. VAR models provide a statistically tractable framework that allows path-dependent forecast, statistical inference and design of statistical tests. Moreover, VAR models have a high expressive power because of the freedom on the order P, the law of the innovations as well as the linear combinations given by the A_i's.

However, in a high-dimensional setting, whether it's in N or in P, the number of interest variables grows like N^2P, thus making the regression procedure both computationally and memory-expensive. Wang et al. [2020] introduce a rearrangement of (1) which leverages a tensor decomposition technique to cast the regression problem into a lower dimensional space, critically depending on the structure of the A_i's, i.e., their sparsity and their spanned vector spaces.

The rest of the report is organised as follows:

- In Section 2, we quickly review the tensor decomposition approach to make the VAR regression procedure more efficient. Furthermore, we give details on the implementation of the Alternating Least Square (ALS) and the Sparse Higher-Order Reduced Rank (SHORR) Algorithms.
- In Section 3, we describe the economic dataset FRED-QD [McCracken and Ng, 2020] on which we test the forecasting power of the VAR model.
- In section 4, we numerically validate the theoretical findings of Wang et al. [2020] on the asymptotic properties of the estimates under the stationarity assumption as well as the reasonable performance of the VAR regressions on the considered dataset.

Implementation The full JAX implementation is available at `https://github.com/ylefay/high_dimensional_vector_autoregression`.

Individual Contributions The ALS and SHORR regression routines were implemented from scratch by YLF. The implementation of the rank selection procedures and the experiments were conducted by AS. The introduction and model estimation sections inside the methodology were written by YLF. The rank selection section, the data and results sections were written by AS, after which the two authors reviewed the manuscript.

2 Method

2.1 Model estimation

Let us rewrite (1) using tensorial notations. Let $\mathbf{A} \in \mathbb{R}^{N\times N\times P}$ be defined by $\mathbf{A}_{\cdot,\cdot,j} = A_j$ for $1 \leq j \leq p$, thus the mode-(1) matricisation of $\mathbf{A}$, which we denote by $\mathbf{A}_{(1)}$, is defined by $\mathbf{A}_{(1)} = \begin{bmatrix} A_1 \mid \ldots \mid A_p \end{bmatrix}$. Let $x_t = (y_{t-1}^\top, \ldots, y_{t-p}^\top)^\top$, (1) becomes

$$y_t = \mathbf{A}_{(1)} x_t + \varepsilon_t. \tag{2}$$

It has been shown in Tucker [1966] that there exists a tensorial factorisation of $\mathbf{A}$

$$\mathbf{A} = G \times_1 U_1 \times_2 U_2 \times_3 U_3, \tag{3}$$

where $G \in \mathbb{R}^{r_1\times r_2\times r_3}$, $U_1 \in \mathbb{R}^{N\times r_1}$, $U_2 \in \mathbb{R}^{N\times r_2}$, $U_3 \in \mathbb{R}^{P\times r_3}$ and the r_i's are the rank of the matricization of A along its i-th mode. This decomposition is particularly well adapted to low-rank structures where the r_i's are such that $r_1, r_2 \ll N$ and $r_3 \ll P$ since the number of variables of interest is $r_1 r_2 r_3 + N(r_1 + r_2) + P r_3 \ll N^2 P$. Furthermore, minimising the least-square error $L(\mathbf{A}) = \sum_{i=1}^T \left\| y_t - \mathbf{A}_{(1)} x_t \right\|^2$ with respect to (w.r.t) A is equivalent to minimising it w.r.t (G, U_1, U_2, U_3). Using tensorial algebra, it can be shown that, for any of the factor involved in the factorisation, let it be B_i for some $i \in \{1, 4\}$, the loss can be rewritten as a simple least-square regression objective, i.e.,

$$L(B_i) = \|y - X_i \mathrm{vec}(B_i)\|^2, \tag{4}$$

where $y = \mathrm{vec}((y_t))$ and X_i depends on the B_j for $j \neq i$ as well as the y_t's. Thus, the optimal $\hat{B}_i$ is the OLS estimator given by $\mathrm{vec}(\hat{B}_i) = (X_i^\top X_i)^\dagger X^\top y$. Computing alternatingly the four estimated factors $\hat{G}, \hat{U}_1, \hat{U}_2, \hat{U}_3$ until the estimated tensor $\hat{\mathbf{A}} = \hat{G} \times_1 \hat{U}_1 \times_2 \hat{U}_2 \times_3 \hat{U}_3$ converges, gives rise to the ALS Algorithm for computing the multilinear low-rank (MLR) least squares estimate. It can be shown that under mild assumptions, the MLR estimate $\hat{\mathbf{A}}_{\mathrm{MLR}}$ is asymptotically normal and consistent [Wang et al., 2020, Th. 1.].

In the case the U_i's are sparse, a convenient regression tool is a Lasso-penalised regression. Adding to the loss a l_1-penalisation term, (4) becomes

$$L(B_i) = \|y - X_i \mathrm{vec}(B_i)\|^2 + \lambda \|U_1\|_1 \|U_2\|_1 \|U_3\|_1. \tag{5}$$

Under technical assumptions, in particular orthogonality and sparsity of the U_i's, as well as the orthogonality of the $G_{(i)}$, the SHORR estimate obtained by performing alternatingly the sparse orthogonal regressions and the regression on $G_{(1)}$ is shown to be asymptotically normal and consistent [Wang et al., 2020, Th. 2.].

To perform the successive sparse orthogonal regression,

$$\min_{B^\top B = \mathrm{Id}} \|y - X\mathrm{vec}(B)\|_2^2 + \lambda \|B\|_1, \tag{6}$$

we introduce a dummy W with the additional constraint that $W = B$ and solve two subproblems. This gives rise to a two-steps subroutine,

Algorithm 1: The ADMM subroutine Algorithm for sparse-orthogonal regression (6)

input : Initial guess $W = B$, $M = 0$, some path for the dual penalisation κ
output: Estimated solution B to (6)
while *stopping criterion* **do**

$B \leftarrow \text{argmin}_{B^\top B = \text{Id}} \|y - X\text{vec}(B)\|_2^2 + \kappa \|B - W + M\|_F^2$ `// using SOC algorithm`

$W \leftarrow \text{argmin}_W \|B - W + M\|_F^2 + \lambda \|W\|_1$ `// using explicit soft-thresholding method`

$M \leftarrow M + B - W$

end
return B

Since the two steps are not explicit in Wang et al. [2020], we further explicit them. The first problem can be solved by the splitting orthogonal constraint (SOC) Algorithm 2 due to Lai and Osher [2014] and given in Appendix 5.1. This method tackles the first step in Algorithm 1 in a similar fashion to the previous approach, by adding a dummy variable and performing successively the minimisation of the loss, then the minimisation of the distance between the dummy variable and B. The two first steps involved in Algorithm 2 are the solution to the first convex problem appearing in the SOC iteration, whose solution is given by a simple OLS estimator. The third step is an orthogonal quadratic program (QP) which has an explicit solution given by Lai and Osher [2014, Th. 1.]. Furthermore, the soft-thresholding step for the unconstrainted l_1-regression problem in Algorithm 1 is given in Chambolle et al. [1998]:

$$W \leftarrow (Q + M - 2\lambda/\kappa > 0)(Q + M - 2\lambda/\kappa) - (-Q - M - 2\lambda/\kappa > 0)(-Q - M - 2\lambda/\kappa). \quad (7)$$

2.2 Rank Selection

To perform the HOSVD factorisation (3), first, we must determine the rank of $A_{(1)}$, $A_{(2)}$ and $A_{(3)}$ since those determine the dimensions of the involved factors G, $U1$, U_2 and U_3. As clarified by Wang et al. [2020] and Xia et al. [2015], these ranks can be obtained by estimating $\hat{\mathbf{A}}$ from $\mathbf{A}$ using a non-decomposition method. Then, the estimated ranks are given by $\hat{r}_i = \underset{1 \leq j \leq p_i - 1}{\text{argmin}} \frac{\sigma_{j+1}(\hat{\mathbf{A}}_{(i)}) + c}{\sigma_j(\hat{\mathbf{A}}_{(i)}) + c}$, for $i \in \{1, 2, 3\}$, where $\sigma_j(\hat{\mathbf{A}}_{(i)})$ denotes the j-th singular value of $\hat{\mathbf{A}}_{(i)}$ rank in descending order, p_i corresponds to the maximum possible rank, i.e. N or P, so $p_1 = N, p_2 = N$ and $p_3 = P$ and c is a constant to avoid division by zero. We follow Wang et al. [2020] by assuming that $c > 0$, $\frac{\|\hat{\mathbf{A}} - \mathbf{A}\|_F}{c} = o(1)$ and $c \max_{1 \leq i \leq 3} (\frac{1}{\sigma_{r_i}(\mathbf{A}_i)} \max_{1 \leq j \leq r_i} (\frac{\sigma_j(\mathbf{A}_i)}{\sigma_{j+1}(\mathbf{A}_i)})) = o(1)$.

3 Data

For us to assess the asymptotic properties of the two estimates, we sample stationary VAR datasets. To do that, we could use a reparametrisation technique for the A_i's, given in Ansley and Kohn [1986]. However, for the sake of simplicity, we simply decided to sample $\mathbf{A}$ until stationarity is satisfied. The innovations are set to independent standard normal variables.

We assess the forecasting power of the VAR model on the FRED-QD (https://research.stlouisfed.org/econ/mccracken/fred-databases/ [McCracken and Ng, 2020]). This dataset contains quarterly statements from various fields. The records start in 1960 and the latest measurements are for 2010. The database contains 257 measurements for 247 dimensions, which can be divided into 14 major groups, ranging from "Money and Credit" to "Housing". We focus our analysis on "Industrial Production", which contains 14 dimensions out of 257 measurements. In this group, the various series are measures of industrial production in sectors such as Residential Utilities, Manufacturing, Fuels, Consumer Goods and Materials. We remove any dimension containing missing values, keeping only 6 dimensions out of 257 measurements, then normalise the data. To make it stationary, which we could test using an augmented Dickey-Fuller test, we differentiate it.

4 Results

In this section, we numerically validate the theoretical convergence properties derived in Wang et al. [2020] for the MLR and SHORR estimates. Furthermore, we showcase the reasonable forecasting power of the corresponding VAR models on the real data described in 3.

4.1 The asymptotic convergence of the MLR estimate

In this experiment, we assess the asymptotic consistency of the MLR estimate [Wang et al., 2020, Th. 1.]:

$$\sqrt{T}\{\mathrm{vec}((\hat{\mathbf{A}}_{\mathrm{MLR}})_{(1)}) - \mathrm{vec}(\mathbf{A}_{(1)})\} \to \mathcal{N}(0, \Sigma_{\mathrm{MLR}}), \tag{8}$$

where Σ_{MLR} is given in Wang et al. [2020, Th. 1]. We sample several $\mathbf{A}$'s with their associated series. We let $N = 10$ and $P = 5$ and let vary T from 2000 to 10000 with a step size of 2000. We analyse this convergence in two cases depending on the ranks of $\mathbf{A}$: the case (a) where $r_1 = r_2 = r_3 = 2$, the case (b) where $r_1 = r_2 = r_3 = 3$. Finally, we replicate this experiment 100 times. See Figure 1 for the results. We do not test for the normality of the errors. The error exhibits a $O(1/\sqrt{T})$ rate while the variance exhibits a $O(1/T)$ rate as predicted by (8).

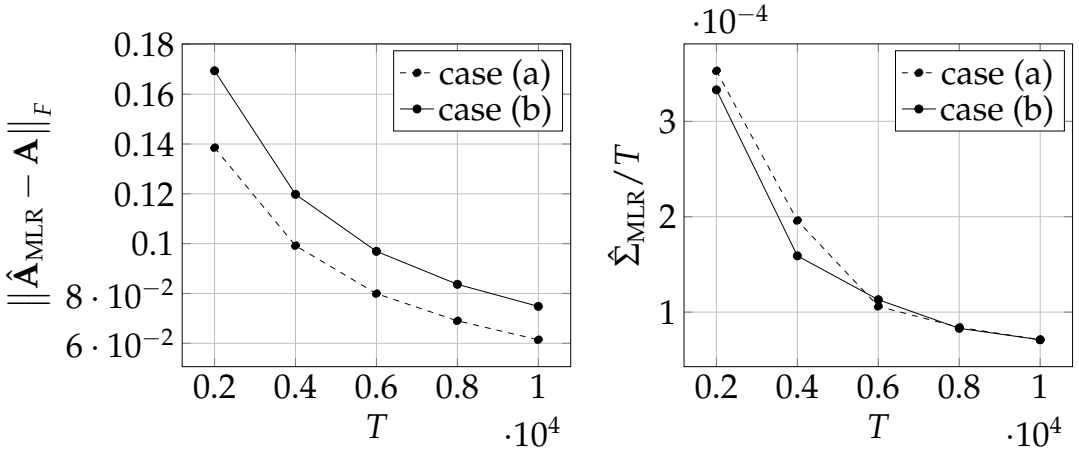

Figure 1: Left: average bias w.r.t T across the 100 experiments. Right: average estimated variance w.r.t to T across the 100 experiments.

4.2 The asymptotic convergence of the SHORR estimate

In this section, we validate the error bound for the SHORR estimate [Wang et al., 2020, Th. 2]:

$$\left\|\hat{\mathbf{A}}_{\text{SHORR}} - \mathbf{A}\right\|_F = O_P(\gamma), \tag{9}$$

where $\gamma = S\log(N^2P)/T$ and $S = s_1s_2s_3$ and each s_i is the maximum number of nonzero values in each column of U_i. For this experiment, we test this result in two different cases with 100 replications for each case. For case (a), we set $N = 3$, $P = 2$, $r_1 = r_2 = r_3 = 2$ and $s_1 = s_2 = s_3 = 2$, and for case (b) we keep the same r and s but we change N and P so that $N = 4$ and $P = 3$. To obtain sparse U matrices, we use a method detailed in Appendix 5.2. By varying γ in $[15, 30, 45, 60] \times 10^{-3}$ we obtain the results in Figure 4 given in Appendix 5.4. The error clearly exhibits a linear increase w.r.t γ as predicted by (9).

4.3 Rank Selection performance

We replicate the setup in Wang et al. [2020] by choosing the non-decomposition estimator $\hat{\mathbf{A}}$ to be given by the nuclear-penalised estimator, $\hat{\mathbf{A}}_{\text{NN}}$ given by $\hat{\mathbf{A}}_{\text{NN}} = \text{argmin}\ \frac{1}{T}\sum_{t=1}^{T}\left\|y_t - \mathbf{A}_{(1)}x_t\right\|_2^2 + \lambda\left\|\mathbf{A}_{(1)}\right\|_*$, where $\|\cdot\|_*$ is the nuclear norm, $\lambda = \alpha\log(N^2P)$ where α has been hand-chosen to be 10^{-2}. We make 50 rank predictions for series generated with $N = 4, P = 3$ and $r_1 = r_2 = r_3 = 2$. See Figure 4 given in Appendix 5.4 for the hit-ratio of the predicted ranks. The results coincide with Wang et al. [2020, Th. 3] which states the rank estimators converge in probability to the true ranks as T goes to ∞.

4.4 Forecasting Industrial Production

There are several methods to choose a consistent order P for the VAR model prediction, such as information criterion or choosing the best P to minimize the error prediction using a cross-validation procedure to avoid overfitting issues. The chosen P in our case is 4. We remove the last 8 points to compare with our forecasts. We then estimate the ranks on the differentiated series and obtain $r_1 = 1, r_2 = 2$ and $r_3 = 1$. The Lasso penalisation λ for the SHORR estimate is chosen to be the lower-bound on λ given by Wang et al. [2020, Th. 2]. See Figures 2 and 3 for the forecasted time-series.

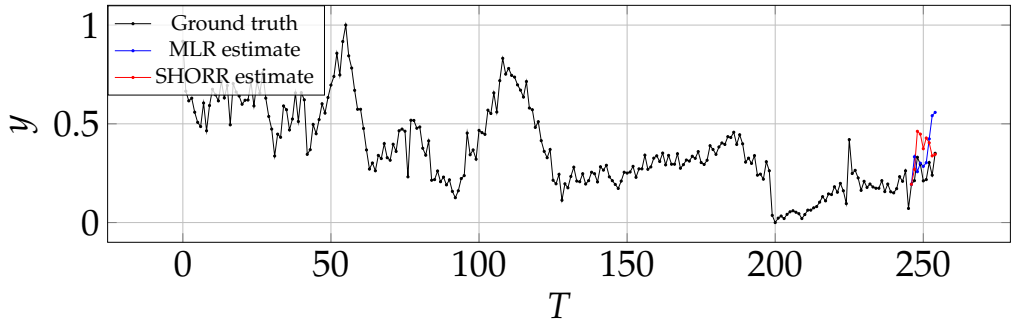

Figure 2: Forecast on one out of the 6-th dimension

5 Supplementary Material

5.1 The SOC Algorithm

To perform the first step regression appearing in 1, we use a specialisation of the splitting orthogonal constraint (SOC) Algorithm described in Lai and Osher [2014]:

Algorithm 2: The SOC Algorithm for the first subproblem in 1

output: Estimated B for the orthogonal minimisation problem in 1

Initialise $Z = 0$, B (from the previous iteration), some path for γ, $T = \kappa \text{vec}(W - M)$, $r = (X^\top X + (\kappa + \gamma)\text{Id})^\dagger$

while *B has not converged* **do**

reg $\leftarrow r(Xy + T + \gamma \text{vec}(B - Z))$ `// OLS`

Construct B' with shape B from reg

$B = \text{argmin}_{Q^\top Q = \text{Id}} \left\| B - (Z + B') \right\|_F^2$ `// orthogonal QP for which closed-formula are known`

$Z \leftarrow Z + B' - B$

end

return B

5.2 Sampling orthogonal matrices

We sample the orthogonal matrices of dimension (4, 2), (3, 2), and (2, 2) for experiments in Section 4.2 in the following way:

1. For (4, 2): We use the matrice $M = \begin{pmatrix} a_1/\sqrt{a_1^2 + a_2^2} & 0 \\ a_2/\sqrt{a_1^2 + a_2^2} & 0 \\ 0 & b_1/\sqrt{b_1^2 + a_b^2} \\ 0 & b_2/\sqrt{b_1^2 + b_2^2} \end{pmatrix}$, where a_1, a_2, b_1 and b_2 are independent standard normal variables.
2. For (3, 2): We do the same as before but we delete the last row and set the right coefficient to 1.
3. For (2, 2): We sample $\theta \sim \mathcal{U}(0, 2\pi)$ and we use the matrice $M = \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix}$.

5.3 The multivariate forecast of the Industrial Production

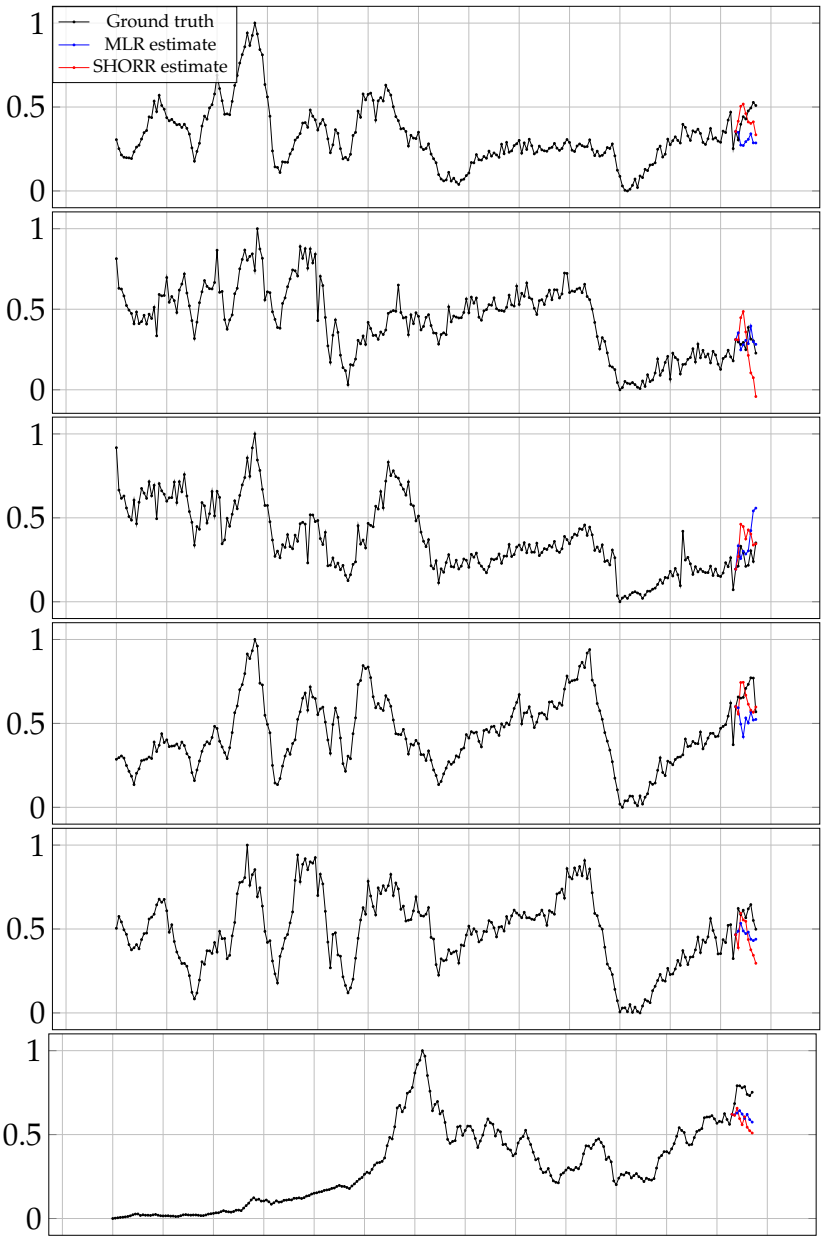

Figure 3: Sampled forecast for the VAR with order $P = 4$ using both the MLR and SHORR estimates.

5.4 Consistency of the SHORR and rank estimates

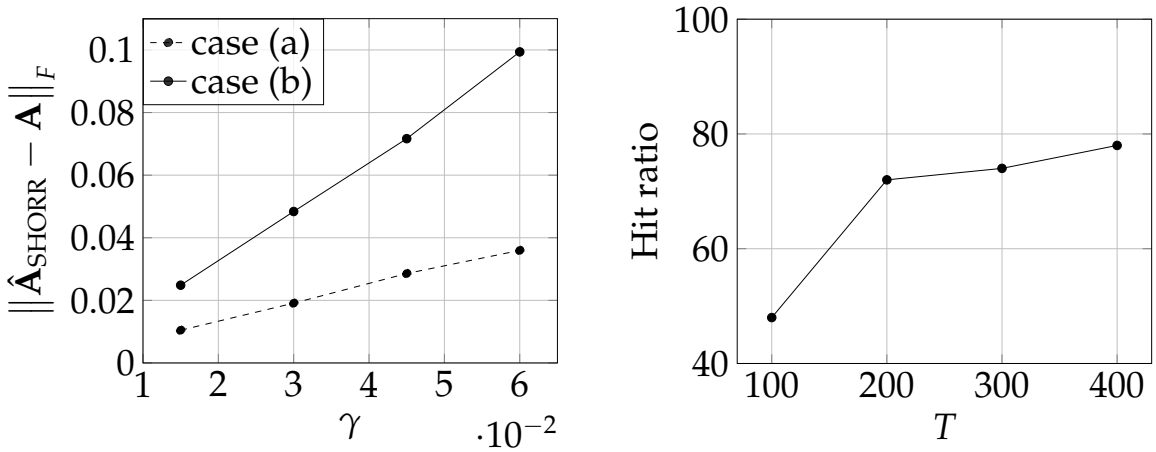

Figure 4: Left: Average bias w.r.t γ across 100 experiments. Right: Average number of times $(\hat{r}_1, \hat{r}_2, \hat{r}_3) = (r_1, r_2, r_3)$ across 50 replications.

References

C. F. Ansley and R. Kohn. A note on reparameterizing a vector autoregressive moving average model to enforce stationarity. *Journal of Statistical Computation and Simulation*, 24(2):99–106, 1986. doi: 10.1080/00949658608810893.

A. Chambolle, R. De Vore, N.-Y. Lee, and B. Lucier. Nonlinear wavelet image processing: variational problems, compression, and noise removal through wavelet shrinkage. *IEEE Transactions on Image Processing*, 7(3):319–335, 1998. doi: 10.1109/83.661182.

R. Lai and S. Osher. A splitting method for orthogonality constrained problems. *Journal of Scientific Computing*, 2014. doi: 10.1007/s10915-013-9740-x.

M. W. McCracken and S. Ng. Fred-qd: A quarterly database for macroeconomic research, 2020.

L. R. Tucker. Some mathematical notes on three-mode factor analysis. *Psychometrika*, 1966.

D. Wang, Y. Zheng, H. Lian, and G. Li. High-dimensional vector autoregressive time series modeling via tensor decomposition, 2020.

Q. Xia, W. Xu, and L. ZHU. Consistently determining the number of factors in multivariate volatility modelling. *Statistica Sinica*, 25(3):1025–1044, July 2015. ISSN 1017-0405. doi: 10.5705/ss.2013.252.