



Libraries and Learning Services

# University of Auckland Research Repository, ResearchSpace

## Copyright Statement

The digital copy of this thesis is protected by the Copyright Act 1994 (New Zealand).

This thesis may be consulted by you, provided you comply with the provisions of the Act and the following conditions of use:

- Any use you make of these documents or images must be for research or private study purposes only, and you may not make them available to any other person.
- Authors control the copyright of their thesis. You will recognize the author's right to be identified as the author of this thesis, and due acknowledgement will be made to the author where appropriate.
- You will obtain the author's permission before publishing any material from their thesis.

## General copyright and disclaimer

In addition to the above conditions, authors give their consent for the digital copy of their work to be used subject to the conditions specified on the [Library Thesis Consent Form](#) and [Deposit Licence](#).

# **Inferring the Conditional Independence Graph for a Multivariate Autoregressive Model via Convex Optimization**

Saïd Maanan

A thesis submitted in partial fulfillment for the  
degree of Doctor of Philosophy

in the

Department of Statistics  
University of Auckland

Supervisor: Dr. Ciprian Doru Giurcăneanu

April 2018



# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Fundamental notions . . . . .	1
1.2	Graphical models for time series . . . . .	2
1.3	Estimation methods . . . . .	5
1.4	Other families of graphical models . . . . .	8
1.5	Contributions . . . . .	9
<b>2</b>	<b>Renormalized Maximum Likelihood</b>	<b>11</b>
2.1	Preliminaries . . . . .	11
2.2	RNML for VAR models . . . . .	12
2.3	RNML formula for $p > 0$ . . . . .	14
2.4	RNML formula for $p = 0$ . . . . .	20
2.5	Asymptotic equivalence between RNML and SBC . . . . .	22
2.6	Experimental Results . . . . .	27
2.6.1	Example 1 . . . . .	27
2.6.2	Example 2 . . . . .	29
2.7	Summary . . . . .	32
<b>3</b>	<b>Efficient Algorithms for VAR Graph Estimation</b>	<b>39</b>
3.1	Preliminaries . . . . .	39
3.2	Main algorithmic steps and their computational complexity . . .	43
3.3	Faster algorithms . . . . .	44
3.3.1	Description . . . . .	44

3.3.2	Further refinements . . . . .	48
3.4	Comparison with other methods based on convex optimization . . . . .	51
3.5	Modified IT criteria . . . . .	52
3.6	Simulated data . . . . .	54
3.7	Air pollution data . . . . .	59
3.8	International stock markets . . . . .	63
3.9	Summary . . . . .	67
<b>4</b>	<b>Latent-Variable VAR Graphical Models</b>	<b>69</b>
4.1	Preliminaries . . . . .	69
4.2	Maximum Entropy Expectation-Maximization algorithm . . . . .	72
4.3	Extended IT criteria . . . . .	78
4.4	Experimental results . . . . .	81
4.4.1	Artificial data . . . . .	81
4.4.2	Settings for AlgoEM . . . . .	81
4.4.3	Comparison of AlgoEM and AlgoSL . . . . .	85
4.4.4	Real-world data . . . . .	91
4.5	Summary . . . . .	93
<b>5</b>	<b>Final remarks</b>	<b>95</b>
	<b>Bibliography</b>	<b>97</b>

# List of Figures

2.1	Example 1: Performance of various criteria in estimating the order of VAR-model. The value of "true" VAR-order is written on the vertical axes of the plots. . . . .	33
2.2	Example 2: Performance of various criteria in estimating the order of VAR-model. All graphical conventions are the same as Figure 2.1.	34
2.3	Example 2: Statistics for the maximum value of $I$ -divergences computed on the $\mathcal{G}$ -grid. For each IT criterion, we plot two error bars, each of which represents mean plus minus standard deviation. Note that $p^\circ = 10$ . . . . .	35
2.4	Example 2: Statistics of Itakura-Saito divergence. All graphical conventions are as in Figure 2.3. $J$ replaces $I_{\max}$ and $J^{\text{ME}}$ replaces $I_{\max}^{\text{ME}}$ .	36
2.5	Example 2: Statistics computed only for the cases when $\hat{p}_{\text{RNML}} = 10$ . All conventions are as in Figure 2.3: $I_{\max}$ is replaced by $I_c$ and $J_{\max}^{\text{ME}}$ is replaced by $J_c^{\text{ME}}$ . . . . .	37
3.1	Performance of estimation methods evaluated by the SIM index defined in (3.17): The closer is SIM to value one, the better is the estimation. In the legend, we indicate for each method the optimization used in Stage2 (NS or ME) as well as the IT criteria employed in Stage1 and Stage2. Note that $K = 5$ and $p^\circ = 10$ . . . . .	55
3.2	Results similar to those presented in Figure 3.1, except that this time $K = 10$ and $p^\circ = 5$ . . . . .	56

3.3	Experimental results for Stage2, obtained when $K = 10$ and $T = 1000$ . In estimation, the order of the model is assumed to be known ( $p^\circ = 5$ ). In Figure 3.3 (a), the “true” SP has ten zeros, whereas in Figure 3.3 (b) has forty zeros. Each vertical line corresponds to an estimation method, and the methods are ordered on the horizontal axis based on the average execution time (in sec. divided by $10^3$ ) computed from ten runs (2.6 GHz-processor). The significance of the acronyms is as follows: FL=Fast List, L = List, ME = Maximum Entropy, NS = Nearest Sparse, RME = Regularized Maximum Entropy and WS = Wermuth-Scheidt. . . . .	60
3.4	Number of times each pair of components of air pollution multivariate time series is found to be conditionally independent (divided by the number of analyzed years). The pairs which have never been selected are not shown in the figure. The results in panel (a) are obtained with the List and those in panel (b) are produced by applying Greedy. 62	
3.5	International stock markets: $\text{ITC}(\mathbf{Y}; \hat{p}, \text{SP}^i) - \min_{0 \leq q \leq \overline{K}} \text{ITC}(\mathbf{Y}; \hat{p}, \text{SP}^q)$ for $\text{ITC} \in \{\text{SBC}, \text{RNML}\}$ , when the index $i$ in Algorithm 2 takes values from zero (no zeros in the fitted SP) to $\overline{K}$ (maximum number of zeros in SP). Graphical conventions are the same for both plots. The vertical line marks the SP for which the ITC is minimized. Let $\{(a_i, b_i)\} = \text{SP}^i \setminus \text{SP}^{i-1}$ for $0 < i \leq \overline{K}$ . To the point whose abscissa is $i$ , we assign a marker according to the classification from Abdelwahab, Amor and Abdelwaheb (2008) of the PSC of marginal time series $\mathbf{y}_{a_i}$ and $\mathbf{y}_{b_i}$ (“strong PSC”, “weak PSC”, “null PSC”). . . . .	68
4.1	Sparsity patterns for the ISDM of the generated data: KS is the number of non-zero entries in the lower triangular part of SP. The black dots represent the locations of the non-zero entries, whereas the light grey dots are the zero entries. . . . .	82

4.2	Results for VAR-models with $KS = 15$ , for which the true sparsity pattern is shown in Figure 1c. With the convention that $\text{dist}$ denotes the distance between the estimated sparsity pattern and the true one, we plot mean $\pm 1$ standard deviation ( $\text{dist}$ ) versus the parameter $\lambda$ . The statistics are computed from $N_{tr} = 10$ trials, for both the <i>adaptive</i> and the <i>non-adaptive</i> case. . . . .	83
4.3	Impact of $N_{it}$ on the performance of AlgoEM: Evaluation is done by replacing in AlgoEM the IT criterion with an oracle having full knowledge about the true sparsity pattern. For each $KS$ and for each $N_{it}$ we run $N_{tr} = 10$ trials for calculating the average distance between the true sparsity pattern and the sparsity pattern selected by oracle. . . . .	84
4.4	Performance of IT criteria compared to that of an oracle: (a) Only the first major loop, MEEM(Pen), of AlgoEM is executed; (b) Both MEEM(Pen) and MEEM(Con) are executed. . . . .	86
4.5	Estimation results obtained when AlgoSL is applied to the same time series which have been used to evaluate the performance of AlgoEM in Figure 4.4. For selection of the sparsity pattern, we employ the score functions in (4.31)-(4.33) and the IT criteria from Section 4.3. Score function $SF_2$ is not shown in the graph because it leads to large values of the average distance: 102.4 for $KS = 2$ , 101.5 for $KS = 3$ , and 89.1 for $KS = 15$ . . . . .	88
4.6	Estimation results when sample size is small ( $T = 500$ ). . . . .	90
4.7	Distances computed for sparsity patterns which are estimated from a time series of length $T = 500$ ; the value of $KS$ for the true sparsity pattern is 15: AlgoEM ( $r$ is the number of latent variables used in estimation). . . . .	91
4.8	Distances computed for sparsity patterns which are estimated from a time series of length $T = 500$ ; the value of $KS$ for the true sparsity pattern is 15: AlgoSL. . . . .	91

4.9	International stock markets data: Comparison between the sparsity pattern for manifest variables from Zorzi and Sepulchre (2016) (red) and the pattern produced by AlgoEM (green) when either $SF_1$ or one of the following IT criteria is used: SBC, EBIC, $EBIC_{FD}$ , FPE, RNML, $RNML_{FD}$ . . . . .	93
-----	---	----

# List of Tables

2.1	Expressions of various IT criteria for a VAR of order $p$ fitted to a data set that contains $K$ time series of length $T$ . For all the criteria listed in the table, the goodness-of-term is $(T/2) \log \det \hat{\Sigma}_p$ . For simplicity, we remove from the penalty terms those quantities which do not depend on $p$ . We prefer to write the formula of log FPE instead of FPE for facilitating the comparison with other criteria. It is easy to see that the formula of SBC is the same with the one given in Section 2.5. . . . .	28
3.1	Computational complexity of various algorithms used to infer the conditional independence graph for VAR. . . . .	45
3.2	Comparison of parametric methods for inferring graphical models via convex optimization: For the optimization problems we use the acronyms NS (Nearest Sparse), ME (Maximum Entropy), RME (Regularized Maximum Entropy), and indicate the equations where the problems are explicitly written. . . . .	53
3.3	Year 2004 . . . . .	63
3.4	Year 2005 . . . . .	64
3.5	Year 2006 . . . . .	64
3.6	Year 2007 . . . . .	65
3.7	Year 2008 . . . . .	65
3.8	Year 2009 . . . . .	66
3.9	Year 2010 . . . . .	66



# Abbreviations

AIC Akaike Information Criterion

ADMM Alternating Direction Method of Multipliers

EIG Eigenvalue decomposition

EBIC Extended Bayesian Information Criterion

FPE Final Prediction Error

GOF Goodness of Fit

ITC Information Theoretic Criterion

ISDM Inverse of the spectral density matrix

KIC Kullback Information Criterion

KL Kullback-Leibler

ME Maximum Entropy

MDL Minimum Description Length

NS Near Sparse

NML Normalized Maximum Likelihood

PARCOR Partial autocorrelations

PSC Partial spectral coherence

## ABBREVIATIONS

---

PEN Penalty

RNML Renormalized Maximum Likelihood

SBC Schwarz Bayesian Criterion

SDP Semidefinite programming

SP Sparsity pattern

SDM Spectral density matrix

VARMA Vector autoregressive moving-average

VAR Vector autoregressive

# Abstract

This thesis is concerned with estimation problems in graphical models of time series. In the case of vector autoregressive (VAR) processes, the conditional independence relations between variables can be inferred by finding the zeros of the inverse spectral density matrix (ISDM). This generally involves two problems: (i) select the order of the VAR model and (ii) choose the sparsity pattern of its ISDM. We first introduce a novel information theoretic (IT) criterion for order selection, the Renormalized Maximum Likelihood (RNML). We prove that RNML criterion is strongly consistent. We also demonstrate empirically its good performance for examples of VAR which have been considered in recent literature because they possess a particular type of sparsity.

As a second contribution, we introduce novel algorithms for inferring the conditional independence graph of a VAR process. We propose a new family of convex optimization algorithms that can be used to solve this problem; in our approach the high-sparsity assumption is not needed. We conduct experiments with simulated data, air pollution data and stock market data for demonstrating that our algorithms are faster and more accurate than similar methods proposed in the previous literature.

Finally, we focus on latent-variable graphical models for multivariate time series. We generalize an algorithm introduced in the previous literature for finding zeros in the inverse of the covariance matrix such that to identify the sparsity pattern of the ISDM of a time series containing latent components.

When applied to a given time series, the algorithm produces a set of candidate models. Various IT criteria are employed for deciding the winner. We conduct an empirical study in which the method we propose is compared with the state-of-the-art.

# Acknowledgements

This thesis would not have been possible without the guidance from my advisor, Dr. Ciprian Giurcăneanu, who offered his continuous advice and encouragement throughout the course of this thesis.

The work for this thesis was ostensibly supported by the UoA Department of Statistics Doctoral Scholarship, it was entirely carried out from May 2015 to April 2018. The thesis is mainly based on the following publications:

1. **Maanan, S.**, Dumitrescu, B., and Giurcăneanu, C. D. “Renormalized maximum likelihood for multivariate autoregressive models.” *2016 24th European Signal Processing Conference (EUSIPCO)*, 150–154, 2016.
2. **Maanan, S.**, Dumitrescu, B., and Giurcăneanu, C. D. “Conditional independence graphs for multivariate autoregressive models by convex optimization: Efficient Algorithms.” *Signal Processing*, 133:122–134, 2017.
3. **Maanan, S.**, Dumitrescu, B., and Giurcăneanu, C. D. “Maximum Entropy Expectation-Maximization algorithm for fitting latent-variable graphical models to multivariate time series.” *Entropy*, 20(1), 2018(20 pages).



## Co-Authorship Form

This form is to accompany the submission of any PhD that contains published or unpublished co-authored work. **Please include one copy of this form for each co-authored work.** Completed forms should be included in all copies of your thesis submitted for examination and library deposit (including digital deposit), following your thesis Acknowledgements. Co-authored works may be included in a thesis if the candidate has written all or the majority of the text and had their contribution confirmed by all co-authors as not less than 65%.

Please indicate the chapter/section/pages of this thesis that are extracted from a co-authored work and give the title and publication details or details of submission of the co-authored work.		
Chapter 2 and some paragraphs in Chapter 1. Extracted from the following publication: "Renormalized maximum likelihood for multivariate autoregressive models." 2016 24 <sup>th</sup> European Signal Processing Conference (EUSIPCO), 150-154, 2016.		
Nature of contribution by PhD candidate	Wrote first draft of the paper, contributed to the derivation of the RNML criterion, wrote parts of the Matlab code, performed all the experiments.	
Extent of contribution by PhD candidate (%)	70%	



### CO-AUTHORS

Name	Nature of Contribution
Prof. Bogdan Dumitrescu	Contributed with part of the algorithm and the corresponding code, helped improving the quality of the text
Dr. Ciprian Doru Giurcaneanu	Coordinated the work for this publication, improved the quality of the text, wrote some code, did some of the mathematical derivations

### Certification by Co-Authors

The undersigned hereby certify that:

- ❖ the above statement correctly reflects the nature and extent of the PhD candidate's contribution to this work, and the nature of the contribution of each of the co-authors; and
- ❖ that the candidate wrote all or the majority of the text.

Name	Signature	Date
Prof. Bogdan Dumitrescu		April 28, 2018
Dr. Ciprian Doru Giurcaneanu		April 26, 2018

## Co-Authorship Form

This form is to accompany the submission of any PhD that contains published or unpublished co-authored work. **Please include one copy of this form for each co-authored work.** Completed forms should be included in all copies of your thesis submitted for examination and library deposit (including digital deposit), following your thesis Acknowledgements. Co-authored works may be included in a thesis if the candidate has written all or the majority of the text and had their contribution confirmed by all co-authors as not less than 65%.

Please indicate the chapter/section/pages of this thesis that are extracted from a co-authored work and give the title and publication details or details of submission of the co-authored work.		
Chapter 3 and some paragraphs from Chapter 1. Extracted from the following publication: "Conditional independence graphs for multivariate autoregressive models by convex optimization: Efficient algorithms." Signal Processing, 133: 122-134, 2017.		
Nature of contribution by PhD candidate	Wrote first draft of the paper, produced the figures, contributed to the development of the novel algorithms, wrote parts of the Matlab code, performed all the experiments.	
Extent of contribution by PhD candidate (%)	75%	



### CO-AUTHORS

Name	Nature of Contribution
Prof. Bogdan Dumitrescu	Contributed with part of the algorithm and the corresponding code, helped improving the quality of the text
Dr. Ciprian Doru Giurcaneanu	Coordinated the work for this publication, improved the quality of the text, made some suggestions for the novel algorithms

### Certification by Co-Authors

The undersigned hereby certify that:

- ❖ the above statement correctly reflects the nature and extent of the PhD candidate's contribution to this work, and the nature of the contribution of each of the co-authors; and
- ❖ that the candidate wrote all or the majority of the text.

Name	Signature	Date
Prof. Bogdan Dumitrescu		April 28, 2018
Dr. Ciprian Doru Giurcaneanu		April 26, 2018

## Co-Authorship Form

This form is to accompany the submission of any PhD that contains published or unpublished co-authored work. **Please include one copy of this form for each co-authored work.** Completed forms should be included in all copies of your thesis submitted for examination and library deposit (including digital deposit), following your thesis Acknowledgements. Co-authored works may be included in a thesis if the candidate has written all or the majority of the text and had their contribution confirmed by all co-authors as not less than 65%.

Please indicate the chapter/section/pages of this thesis that are extracted from a co-authored work and give the title and publication details or details of submission of the co-authored work.		
Chapter 4. Extracted from the following publication: "Maximum Entropy Expectation-Maximization Algorithm for Fitting Latent-Variable Graphical Models to Multivariate Time Series." Entropy 20(1), 2018 (20 pages).		
Nature of contribution by PhD candidate	Contributed to the development of the main algorithm and to its implementation in Matlab, performed the experiments, produced all the figures, wrote the first draft of the paper.	
Extent of contribution by PhD candidate (%)	80%	



### CO-AUTHORS

Name	Nature of Contribution
Prof. Bogdan Dumitrescu	Proposed the adaptation of the Expectation Maximization algorithm, wrote part of the code and contributed to polishing the paper
Dr. Ciprian Doru Giurcaneanu	Coordinated the work for this publication, improved the quality of the text, made some suggestions for the main algorithm

### Certification by Co-Authors

The undersigned hereby certify that:

- ❖ the above statement correctly reflects the nature and extent of the PhD candidate's contribution to this work, and the nature of the contribution of each of the co-authors; and
- ❖ that the candidate wrote all or the majority of the text.

Name	Signature	Date
Prof. Bogdan Dumitrescu		April 28, 2018
Dr. Ciprian Doru Giurcaneanu		April 26, 2018



# Chapter 1

## Introduction

### 1.1 Fundamental notions

Graphical models are instrumental in the analysis of multivariate data. Originally, these models have been employed for independently sampled data, but their use has been extended to multivariate, stationary time series [Brillinger \(1996\)](#), [Dahlhaus, Eichler and Sandkühler \(1997\)](#), which triggered their popularity in statistics, machine learning, signal processing and neuroinformatics.

For better understanding the significance of graphical models, let  $\mathbf{x}$  be a random vector having a Gaussian distribution with zero-mean and positive definite covariance matrix  $\mathbf{\Gamma}$ . A graph  $G = (V, E)$  can be assigned to  $\mathbf{x}$  in order to visualize the conditional independence between its components. The symbol  $V$  denotes the vertices of  $G$ , while  $E$  is the set of its edges. There are no loops from a vertex to itself, nor multiple edges between two vertices. Hence,  $E$  is a subset of  $\{(a, b) \in V \times V : a \neq b\}$ . Each vertex of the graph is assigned to an entry of  $\mathbf{x}$ . We conventionally draw an edge between two vertices  $a$  and  $b$  if the random variables  $\mathbf{x}_a$  and  $\mathbf{x}_b$  are *not* conditionally independent, given all other components of  $\mathbf{x}$ . The description above follows the main definitions from [Speed and Kiiveri \(1986\)](#) and assumes that the graph  $G$  is *undirected*. Proposition 1 from the same reference provides a set of equivalent conditions for conditional independence. The most interesting one claims that  $\mathbf{x}_a$  and  $\mathbf{x}_b$  are conditionally

independent if and only if the entry  $(a, b)$  of  $\mathbf{\Gamma}^{-1}$  is zero. This shows that the missing edges of  $G$  correspond to zero-entries in the inverse of the covariance matrix, which is called the concentration matrix.

There is an impressive amount of literature on graphical models. In this thesis, we focus on a generalization of this problem to time series. The main difference between the *static* case and the *dynamic* case is that the former relies on the sparsity pattern of the concentration matrix, whereas the latter is looking for zeros in the inverse of the spectral density matrix. One of the main difficulties stems from the fact that the methods developed in the static case cannot be applied straightforwardly to time series.

## 1.2 Graphical models for time series

We consider a zero-mean, Gaussian, stationary stochastic vector process  $\{\mathbf{y}_t\}$ . It is assumed that, for all  $t \in \mathbb{Z}$ , the vector  $\mathbf{y}_t$  has  $K$  real-valued entries ( $K > 2$ ). For any integer  $h$ , let  $\mathbf{R}(h) = \text{Cov}(\mathbf{y}_t, \mathbf{y}_{t-h})$  be the autocovariance matrix at lag  $h$ , for the time series  $\{\mathbf{y}_t\}$ . The spectral density matrix is defined as  $\mathbf{S}(\omega) = \sum_{h=-\infty}^{\infty} \mathbf{R}(h)e^{-jh\omega}$ , where  $j = \sqrt{-1}$ . For each  $\omega \in \mathbb{R}$ ,  $\mathbf{S}(\omega)$  is a Hermitian matrix of size  $K \times K$ . Additionally,  $\mathbf{S}(\omega)$  is a periodic function of period  $2\pi$ , and this explains why the values of  $\omega$  are restricted to  $(-\pi, \pi]$  in spectral analysis. Additionally, we assume that the eigenvalues of  $\mathbf{S}(\omega)$  are bounded and bounded away from zero, uniformly for all frequencies in  $(-\pi, \pi]$ .

All the definitions introduced so far can be found in most of the textbooks on time series. The interested reader can find them, for example, in [Dahlhaus \(2000\)](#). This is widely used in the analysis of multivariate time series. In [Dahlhaus \(2000\)](#), it is shown how PSC can be employed for constructing graphical models that, in the particular case of Gaussian time series, reduce to the conditional independence graph. Even if the extension to the non-Gaussian case is very interesting, we prefer to follow the traditional approach and use the Gaussian assumption. This is particularly convenient for the derivation of

a novel information theoretic criterion which is instrumental for the estimation methods that we propose. Before discussing the estimation problem, we provide some more details on the definition of PSC and its numerical evaluation.

Assume that the  $a$ -th and the  $b$ -th components of  $\{\mathbf{y}_t\}$  are denoted by  $\{\mathbf{y}_{t,a}\}$  and  $\{\mathbf{y}_{t,b}\}$ , respectively. The notation  $\{\mathbf{y}_{t,-ab}\}$  is employed for the rest of  $K - 2$  components. We are interested to compute the correlation between  $\{\mathbf{y}_{t,a}\}$  and  $\{\mathbf{y}_{t,b}\}$ , conditional on  $\{\mathbf{y}_{t,-ab}\}$ . To this end, the following steps are considered (Davis, Zang and Zheng, 2016): (i) Find the optimal linear filter which removes the linear effect of  $\{\mathbf{y}_{t,-ab}\}$  from  $\{\mathbf{y}_{t,a}\}$  and get the residual series which is the outcome of this filter; (ii) Do the same as in step (i) after replacing  $\{\mathbf{y}_{t,a}\}$  with  $\{\mathbf{y}_{t,b}\}$ ; (iii) Compute the correlations between the two residual series and then use them in order to compute the residual cross-spectral density; (iv) For  $\omega \in (-\pi, \pi]$ , define  $\text{PSC}_{ab}(\omega)$  to be the scaled cross-spectral density between the two residual series (see (Davis, Zang and Zheng, 2016, Eq. (2.4))).

From the algorithm outlined above, we have that  $\{\mathbf{y}_{t,a}\}$  and  $\{\mathbf{y}_{t,b}\}$  are uncorrelated (given  $\{\mathbf{y}_{t,-ab}\}$ ) if and only if  $\text{PSC}_{ab}(\omega) = 0$  for all  $\omega \in (-\pi, \pi]$ . This is equivalent with saying that the cross-spectral density for the two residual series is zero, for all values of  $\omega$ . Even if the definition introduced above is very intuitive, it is difficult to be used in practice because it involves the optimal filters. A more convenient expression of PSC can be obtained from the equation which relates the residual cross-spectral density and the entries of the spectral density matrix of the process  $\{\mathbf{y}(t)\}$ . More precisely, according to Brillinger (1996), we have that the cross-spectral density of the residual series corresponding to  $\{\mathbf{y}_{t,a}\}$  and  $\{\mathbf{y}_{t,b}\}$ , at frequency  $\omega$ , is given by

$$[\mathbf{S}(\omega)]_{ab} - [\mathbf{S}(\omega)]_{a,-ab} \{[\mathbf{S}(\omega)]_{-ab,-ab}\}^{-1} [\mathbf{S}(\omega)]_{-ab,b},$$

where the symbol  $[\mathbf{S}(\omega)]_{ab}$  is used for the entry located at the intersection of the  $a$ -th row and the  $b$ -th column of  $\mathbf{S}(\omega)$ . The matrix  $[\mathbf{S}(\omega)]_{-ab,-ab}$  is obtained from  $\mathbf{S}(\omega)$  after removing the rows and columns whose indices are  $a$  and  $b$ . The significance of the rest of symbols which appear in the equation above is

evident. The main drawback is that computing the PSC for one single pair of marginal series of  $\{\mathbf{y}_t\}$  requires to invert a matrix of size  $(K - 2) \times (K - 2)$ . The computational burden can be further lowered by using the following result from (Dahlhaus, 2000):

$$\text{PSC}_{ab}(\omega) = -\frac{[\mathbf{S}^{-1}(\omega)]_{ab}}{\sqrt{[\mathbf{S}^{-1}(\omega)]_{aa}[\mathbf{S}^{-1}(\omega)]_{bb}}}, \quad (1.1)$$

where the notation for the entries of  $\mathbf{S}^{-1}(\omega)$  is similar to the one used previously for the entries of  $\mathbf{S}(\omega)$ . Most interestingly, the formula involves the computation of the inverse for one single matrix of size  $K \times K$  instead of inverting many matrices of size  $(K - 2) \times (K - 2)$ . The most important consequence is that  $\{\mathbf{y}_{t,a}\}$  and  $\{\mathbf{y}_{t,b}\}$  are conditionally uncorrelated if and only if the  $(a, b)$ -entry of the inverse of the spectral density matrix (ISDM) is zero.

This result has a strong connection with the problem of inferring graphical models. For instance, the undirected graph which corresponds to  $\{\mathbf{y}_t\}$  has exactly  $K$  vertices, one for each component of the multivariate time series. The vertices of two distinct components, say  $a$  and  $b$ , are not connected with an edge if  $\{\mathbf{y}_{t,a}\}$  and  $\{\mathbf{y}_{t,b}\}$  are uncorrelated, conditional on all other components of the time series. Otherwise, they are connected. It follows that the edges which are missing from the undirected graph assigned to  $\{\mathbf{y}(t)\}$  correspond to the zero-entries of ISDM. This shows clearly that the problem of graphical models for times series is a generalization of the similar problem for the “static” case. As we already pointed out, the main difference in the case of time series is that we are looking for zeros in the ISDM and not in the inverse of the covariance matrix. When ISDM is estimated from a finite number of measurements, the  $(a, b)$ -entry will not be exactly zero even when  $\{\mathbf{y}_{t,a}\}$  and  $\{\mathbf{y}_{t,b}\}$  are conditionally uncorrelated. Hence, it is needed a test for deciding which entries are equal to zero. Next, we will briefly discuss various solutions proposed in the previous literature, for solving this problem.

## 1.3 Estimation methods

In [Dahlhaus \(2000\)](#), a non-parametric estimator is used to compute the spectral density matrix for all the points of a regular grid on  $(0, 2\pi]$ . For each  $\omega$  on the grid, the estimated  $\mathbf{S}(\omega)$  is inverted and rescaled in order to calculate the PSC. For any pair  $(a, b)$ , the test for deciding if the corresponding entry of ISDM is zero relies on the asymptotic distribution of  $\max_{\omega} |\text{PSC}_{ab}(\omega)|^2$  under the hypothesis that  $\{\mathbf{y}_{a,t}\}$  and  $\{\mathbf{y}_{b,t}\}$  are conditionally uncorrelated (see also [Dahlhaus, Eichler and Sandkühler \(1997\)](#)).

Another pioneering contribution to the problem of graphical models for time series is ([Bach and Jordan, 2004](#)). The focus of their article is on constructing *directed graphs* which are used in prediction. An important result is ([Bach and Jordan, 2004](#), Prop. 2), which gives the expression of Kullback-Leibler (KL) divergence between two zero-mean stationary vector-valued Gaussian processes whose spectral density matrices are invertible. The formula is a generalization of KL-divergence between two positive-definite matrices (see, for example, ([Speed and Kiiveri, 1986](#), Eq. (6))).

The expression of KL-divergence was also employed in [Matsuda \(2006\)](#), but in a slightly different context. More precisely, [Matsuda \(2006\)](#) uses an iterative algorithm initialized with the graph which has all possible edges. At each iteration, an edge is collapsed; the non-parametrically-estimated KL-divergence between the graphical model which contains the edge and the one which does not contain it is computed. This quantity is further employed for a test which decides if the model with the collapsed edge contains the “true” graph. A faster variant of the method in [Matsuda \(2006\)](#) was recently developed in [Wolstenholme and Walden \(2015\)](#).

From the recent literature, we mention [Jung \(2015\)](#), which addresses the problem of graphical models for high-dimensional time series. Their method assumes that the observed process is sufficiently smooth in the frequency domain and its ISDM is very sparse. These assumptions allow to find the pattern of zeros

of ISDM when the sample size is smaller than the number of the time series that are measured. The estimation is performed by combining the well-known Blackman-Tukey method (Stoica and Moses, 2005) with the multitask LASSO (Buhlmann and van der Geer, 2011).

The parametric methods rely on the fact that the observations are outcomes of a vector autoregressive (VAR) process. The well-known difference equation of the process is (Lutkepöhl, 2005):

$$\mathbf{y}_t = \mathbf{A}_1 \mathbf{y}_{t-1} + \cdots + \mathbf{A}_{p^\circ} \mathbf{y}_{t-p^\circ} + \mathbf{u}_t, \quad t = 1, 2, \dots, T, \quad (1.2)$$

where  $\mathbf{A}_1, \dots, \mathbf{A}_{p^\circ}$  are matrix coefficients of size  $K \times K$  and  $\{\mathbf{u}_t\}_{t=1}^T$  is a sequence of independently and identically distributed random  $K$ -vectors. The vectors  $\{\mathbf{u}_t\}_{t=1}^T$  are drawn from a  $K$ -variate Gaussian distribution with zero mean vector and covariance matrix  $\Sigma \succ 0$ . Additionally, the vectors  $\{\mathbf{y}_t\}_{t=1-p^\circ}^0$  are assumed to be constant. Hereafter, we use the notation  $\mathbf{Y} = [\mathbf{y}_1, \dots, \mathbf{y}_T]'$ , where  $(\cdot)'$  is the transposition operator.

It follows that ISDM has the expression (Songsiri, Dahl and Vandenberghe, 2010):

$$\mathbf{S}^{-1}(\omega) = \mathbf{A}^H(\omega) \Sigma^{-1} \mathbf{A}(\omega) = \sum_{m=-p^\circ}^{p^\circ} \mathbf{Q}_m e^{-j\omega m}, \quad (1.3)$$

where  $(\cdot)^H$  is the operator for conjugate transpose. We define  $\mathbf{A}_0 = -\mathbf{I}$  and  $\mathbf{A}(\omega) = -\sum_{m=0}^{p^\circ} \mathbf{A}_m e^{-j\omega m}$ . We make the convention that  $\mathbf{I}$  stands for the identity matrix of appropriate size. For  $m \geq 0$ , we have that  $\mathbf{Q}_m = \sum_{i=0}^{p^\circ-m} \mathbf{A}_i' \Sigma^{-1} \mathbf{A}_{i+m}$  and  $\mathbf{Q}_{-m} = \mathbf{Q}_m'$ .

The first work on the problem of graphical models for VAR is Eichler (2006), using Whittle's approximation for the log-likelihood function and introducing iterative algorithms which extend those previously applied to infer graphical models for variables that are sampled with independent replications (Lauritzen, 1996).

The parametric methods use various information theoretic (IT) criteria for finding the order of the VAR-model and for identifying the sparsity pattern (SP)

of ISDM. Interestingly enough, they do not assume that only very few of the entries of the ISDM are non-zero. As the high sparsity is not included in the set of assumptions, the authors of these works employ “classical” IT criteria: SBC - Schwarz’s Bayesian Criterion (Schwarz, 1978), AIC - Akaike Information Criterion (Akaike, 1974), AIC<sub>c</sub> - “corrected” AIC (Brockwell and Davis, 1991).

In Davis, Zang and Zheng (2016), a non-parametric estimator is used “to guess” the entries of  $\mathbf{S}^{-1}(\omega)$  which are likely to be non-zero. This leads to a list of competing models,  $\text{VAR}(p, \text{SP})$ , where  $p$  does not exceed a pre-defined  $p_{\max}$ -order and SP denotes the sparsity pattern of  $\mathbf{S}^{-1}(\omega)$ . SP is further converted into zeros of  $[\mathbf{A}_1 \cdots \mathbf{A}_p]$ , then each candidate model is fitted to the data and the winner is selected by using SBC. The results are refined in the second stage of the procedure, where SBC is applied again.

The approach from Songsiri, Dahl and Vandenberghe (2010) is more computationally intensive because a VAR-model is fitted to the data for each pair  $(p, \text{SP})$  when all possible SP’s are considered. The major contribution of Songsiri, Dahl and Vandenberghe (2010) consists in recasting the model fitting as a convex optimization problem. A faster method based on convex optimization is proposed in Songsiri, Dahl and Vandenberghe (2010) and is further developed in Avventi, Lindquist and Wahlberg (2013), where vector autoregressive moving-average (VARMA) models are considered instead of VAR. They show how the Maximum Entropy estimate of VARMA model can be produced as the solution of a convex optimization problem in which the ISDM is constrained to have a certain sparsity pattern. Avventi, Lindquist and Wahlberg (2013) provide an algorithm for generating a set of candidates for the ISDM sparsity pattern by exploiting the distributional properties of the estimated PSC. The best candidate is selected by evaluating a fitness function which takes into consideration the adherence of the model to the measurements as well as the sparsity of the model.

## 1.4 Other families of graphical models

So far the presentation was mainly focused on the idea that constructing an *undirected graph* for a multivariate time series  $\{\mathbf{y}_t\}$  reduces to find zeros in its ISDM. In the case of VAR-model, with the notation from (1.3), we have that  $\{\mathbf{y}_{a,t}\}$  and  $\{\mathbf{y}_{b,t}\}$  are conditionally independent if  $\sum_{i=0}^{p^\circ-m} \sum_{u=1}^K [\mathbf{A}_i]_{au} [\mathbf{A}_{i+m}]_{bu} = 0$  for  $m = 0, \dots, p^\circ$  (see also (Eichler, 2006, Sec. 3)). This clearly shows that the relationship between conditional independence and the matrix coefficients of VAR is complicated, even if we have assumed that  $\mathbf{\Sigma} = \mathbf{I}$ .

However, if the goal is to draw a *directed graph* for VAR, then one should be looking for zeros in  $\mathbf{A}_1, \dots, \mathbf{A}_p$ . The connection between this approach and Granger causality is well-known (see, for example, (Lutkepöhl, 2005, Sec. 2.3)). More importantly, VAR-models for which the coefficients  $\mathbf{A}_1, \dots, \mathbf{A}_p$  are sparse are known to have good predictive capabilities (Davis, Zang and Zheng, 2016).

The fundamental differences between the two family of graphical models for time series (directed graphs and undirected graphs) were already pointed out in the seminal work of Brillinger (Brillinger, 1996). In this thesis, we are only concerned with the sparsity in the frequency domain (zeros in the ISDM) and not with the sparsity in the time domain (zeros in the matrix coefficients of the model).

Chandrasekaran, Parrilo and Willsky (2012) extend the static case by allowing the presence of latent variables. The key point of their approach is to express the manifest concentration matrix as the sum of a sparse matrix and a low-rank matrix. Additionally, they provide conditions for the decomposition to be unique, in order to guarantee the identifiability. The two matrices are estimated by minimizing a penalized likelihood function, where the penalty involves both the  $\ell_1$ -norm and the nuclear norm. Interestingly enough, Lauritzen and Meinhausen (2012) pointed out that an alternative solution, which relies on the Expectation-Maximization algorithm, can be easily obtained.

A happy marriage between the approach from [Chandrasekaran, Parrilo and Willsky \(2012\)](#) and the use of Maximum Entropy led to the solution proposed in [Zorzi and Sepulchre \(2016\)](#) for the identification of graphical models of autoregressive processes with latent variables. Similar to [Chandrasekaran, Parrilo and Willsky \(2012\)](#), the estimation is done by minimizing a cost function whose penalty term is given by a linear combination of the  $\ell_1$ -norm and the nuclear norm. The two coefficients of this linear combination are chosen by the user and they have a strong influence on the estimated model. The method introduced in [Zorzi and Sepulchre \(2016\)](#) performs the estimation for various pairs of coefficients which yield a set of candidate models; the winner is decided by using a *score function*.

## 1.5 Contributions

In Chapter [2](#), we derive a new IT criterion, the Renormalized Maximum Likelihood (RNML) for VAR-models. After deriving the expression of RNML for the case when  $p > 0$  and the case when  $p = 0$ , we perform an asymptotic analysis of the RNML, where we prove that RNML and SBC reduce to the same formula when  $T \rightarrow \infty$ . An important consequence of this result is that RNML is a strongly consistent estimator for the order of the model. After that, we run a series of simulations in order to show the superiority of RNML with respect to other IT criteria.

In Chapter [3](#), we propose an alternative solution to the convex optimization-based estimation method for VAR graphical models from [Songsiri, Dahl and Vandenberghe \(2010\)](#). Our approach consists of two stages: (i) select the order of the model by using an IT criterion; (ii) for each sparsity pattern (SP) in a given list, solve a convex optimization problem which finds the VAR-model whose ISDM zeros are located according to SP and is closest from the VAR-model produced in the first stage. We introduce a family of efficient algorithms for generating the list of SP-candidates, and show how RNML and other IT

criteria can be altered for selecting the best SP from the list. A theoretical analysis demonstrates that the computational complexity of the newly proposed method is lower than the complexity of the existing algorithms. The estimation accuracy is tested in an experimental study that involves simulated and real-life data.

And lastly, in Chapter 4, we focus on latent-variable graphical models for multivariate time series. We show how the algorithm introduced in Lauritzen and Meinhausen (2012) can be generalized such that to identify the sparsity pattern of the ISDM of a time series containing latent components. When applied to a given time series, the algorithm produces a set of candidate models. As before, various IT criteria are employed for deciding the winner. We conduct an empirical study in which the algorithm is compared with the state-of-the-art.

Chapter 5 concludes the thesis.

## Chapter 2

# Renormalized Maximum Likelihood

### 2.1 Preliminaries

Model selection is concerned with choosing a statistical model from a set of candidate models. Naturally, simple models are preferred to complex ones. But formalizing what is meant by ‘simple’ and ‘complex’ is not a straightforward task, because it is not easy to decide how the number of parameters in the model should be balanced against its fit to data. To solve this problem, many information theoretic criteria have been developed. Those criteria belong either to the frequentist or the Bayesian frameworks, in both of which it is assumed that there exists a true probability distribution  $f(\cdot)$  from which the observed data were sampled, and the goal is to develop models that approximate this truth. It follows that the goal of model selection is to find the model that is closest to the truth. However, this approach has many shortcomings.

According to the Minimum Description Length (MDL) principle, the foundational assumption according to which there is a certain probability distribution  $f(\cdot)$  from which the observed data were sampled is incorrect. According to this view, the question of whether the true distribution  $f(\cdot)$  even exists is inherently unanswerable.

The MDL principle adopts a different approach to modelling. It proposes that the basic goal should be to find regularities in data, and use these regularities to compress the data set so as to unambiguously describe it using fewer symbols than the number of symbols needed to describe the data literally. The more a model permits data compression, the more it enables us to discover the regularities underlying the data.

One recent technique based on the MDL principle is the Normalized Maximum Likelihood (NML). The NML criterion was developed for both cases when the models are of the discrete type and when they are of the continuous type. The major problem in the continuous case stems from the fact that the denominator of the criterion is not finite because it involves an integral for which the integration space is infinite. To deal with this issue, [Rissanen \(2000\)](#) proposed to constrain the integration space. He also proposed a second normalization of the criterion so as to get rid of the hyper parameters. The resulting criterion was therefore named the Renormalized Maximum Likelihood (RNML).

Since the RNML has been introduced more recently compared to other traditional criteria, therefore it has greater potential than the traditional methods, and new applications of the idea are emerging continuously.

## 2.2 RNML for VAR models

Consider the scenario in which a VAR( $p$ )-model is fitted to the data  $\mathbf{Y} = [\mathbf{y}_1, \dots, \mathbf{y}_T]'$ . Similar to the notation used in (1.2),  $\Sigma$  is the covariance matrix of the driven noise. Assuming that  $\mathbf{A}_1, \dots, \mathbf{A}_p$  are the coefficients of the model, we define  $\mathbf{B} = [\mathbf{A}_1, \dots, \mathbf{A}_p]^\top$ . According to [Rissanen \(2000, 2007\)](#), the normalized maximum likelihood (NML) equals the negative logarithm of

$$\hat{f}(\mathbf{Y}; p) = \frac{f\left(\mathbf{Y}; \hat{\mathbf{B}}(\mathbf{Y}), \hat{\Sigma}(\mathbf{Y})\right)}{C_p}. \quad (2.1)$$

In our calculations, we use natural logarithms, denoted by  $\log(\cdot)$ . In (2.1), the numerator is the maximum value of the likelihood function, given the measure-

ments  $\mathbf{Y}$ .  $\hat{\mathbf{B}}(\mathbf{Y})$  and  $\hat{\mathbf{\Sigma}}(\mathbf{Y})$  are the maximum likelihood (ML) estimates of the parameters of the model. For the denominator, we have

$$C_p = \int f(\mathbf{Y}; \hat{\mathbf{B}}(\mathbf{Y}), \hat{\mathbf{\Sigma}}(\mathbf{Y})) d\mathbf{Y}, \quad (2.2)$$

where the domain of integration is the entire space of observations. Since the integral above diverges, we apply the same type of constraint as the one proposed in [Rissanen \(2000\)](#). This leads to a finite result which depends on some hyper-parameters. Because we do not want to choose subjectively the values of the hyper-parameters, we follow the recommendations from ([Rissanen, 2000](#)) and perform a second normalization step. The resulting formula is named  $\text{RNML}(\mathbf{Y}; p)$ .

According to the best of our knowledge, the expression of RNML for VAR-models was not obtained so far. The first attempt at estimating the order of *univariate* AR models by RNML is the one from [Giurcăneanu and Rissanen \(2006\)](#). The approach from [Giurcăneanu and Rissanen \(2006\)](#) was further extended in [Schmidt and Makalic \(2011\)](#), where the focus is still on the univariate case. We note in passing that, the method employed in [Schmidt and Makalic \(2011\)](#) for evaluating the criterion does not allow the use of the second normalization step. More interestingly, the work of [Schmidt and Makalic \(2011\)](#) relies on the re-parametrization of the AR model by partial autocorrelations (PARCOR).

The univariate PARCOR function was extended to vector time series by introducing (i) the partial autoregression matrix function, (ii) the partial lag autocorrelation matrix function and (iii) the partial autocorrelation matrix function [see ([Wei, 2006](#), Section 16.5) for a tutorial review]. The last one is best known for its use in the normalized Whittle-Wiggins-Robinson algorithm ([Morf, Vieira and Kailath, 1978](#)). However, none of these functions enables the calculation of the integral in (2.2).

In order to overcome the difficulties, we recast VAR in the form of a linear regression model, which means that the random vectors  $\{\mathbf{y}_t\}_{t=1}^{T-1}$  are treated as

fixed predictors. This technique is widely used in time series analysis (see, for example, Lutkepöhl (2005), Ting et al. (2015), Neumaier and Schneider (2001)). We are encouraged to apply it by the experimental results reported by Rissanen, T.Roos and Myllymäki (2010) which show, for the univariate case, that the RNML criterion devised for variable selection in linear regression works properly when is employed to estimate the order of autoregressions.

In our derivations, we will use some techniques from (Hirai and Yamanishi, 2013), which appears to be the only work that considers the problem of RNML-computation for the case when the measurements are vector-valued and not scalar-valued. However, their results for multidimensional data are confined to Gaussian mixture model.

In the next section, we show how can be derived the expression of  $\text{RNML}(\mathbf{Y}; p)$  for the case when  $p > 0$ . The formula for  $\text{RNML}(\mathbf{Y}; p = 0)$  is given in Section 2.4. In practice, these formulae are employed to evaluate  $\text{RNML}(\mathbf{Y}; p)$  for  $p \in \{p_{\min}, \dots, p_{\max}\}$ ; the estimate of the order ( $\hat{p}$ ) equals the value of  $p$  which minimizes the criterion.

## 2.3 RNML formula for $p > 0$

Assume that VAR( $p$ )-model with order  $p > 0$  (see eq. (1.2)) is fitted to the data  $\mathbf{Y}$ . We prove the following result:

**Proposition 2.3.1.** *Under the hypotheses that  $T \geq K(p + 1)$  and the vectors  $\{\mathbf{u}_t\}_{t=1}^T$  are Gaussian distributed, the expression of the RNML-criterion is*

$$\text{RNML}(\mathbf{Y}; p) = \text{GOF} + \sum_{i=1}^3 \text{PEN}_i,$$

where

$$\text{GOF} = [(T - Kp - K + 1)/2] \log \det \hat{\Sigma}_p \quad (2.3)$$

$$\text{PEN}_1 = -\log \Gamma_K[(T - Kp)/2] \quad (2.4)$$

$$\text{PEN}_2 = -\log \Gamma[(K^2 p)/2] \quad (2.5)$$

$$\text{PEN}_3 = [(K^2 p)/2] \log \text{tr} \left[ (\mathbf{Y}'\mathbf{Y})/T - \hat{\Sigma}_p \right]. \quad (2.6)$$

Here  $\Gamma[\cdot]$  is Gamma function and  $\Gamma_K[\cdot]$  is the multivariate Gamma function. The operators  $\det(\cdot)$  and  $\text{tr}(\cdot)$  stand for the determinant and the trace, respectively. By  $\hat{\Sigma}_p$  we denote the estimate of the error covariance matrix obtained when VAR( $p$ )-model is fitted to the measurements  $\mathbf{Y}$ .

*Proof.*

### Preliminary calculations

For ease of writing, we introduce the notation  $\ell = Kp$  and define:

$$\mathbf{Z}_t = [\mathbf{y}'_t, \dots, \mathbf{y}'_{t-p+1}]' \ (\ell \times 1),$$

$$\mathbf{Z} = [\mathbf{Z}_0, \dots, \mathbf{Z}_{T-1}]' \ (T \times \ell),$$

$$\mathbf{U} = [\mathbf{u}_1, \dots, \mathbf{u}_T]' \ (T \times K).$$

Remark that the size of each newly defined quantity is listed in the parentheses.

As  $\mathbf{Y} = [\mathbf{y}_1, \dots, \mathbf{y}_T]'$  and  $\mathbf{B} = [\mathbf{A}_1, \dots, \mathbf{A}_p]'$ , it follows that

$$\mathbf{Y} = \mathbf{ZB} + \mathbf{U}. \quad (2.7)$$

Under the hypotheses that  $T \geq K + \ell$  and the vectors  $\{\mathbf{u}_t\}_{t=1}^T$  are Gaussian distributed, the conditional ML estimators are given by (Mardia, Kent and Bibby, 1979):

$$\hat{\mathbf{B}} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y}, \quad (2.8)$$

$$\hat{\Sigma} = (\mathbf{Y}'\mathbf{P}_{\mathbf{Z}}^{\perp}\mathbf{Y})/T, \quad (2.9)$$

where  $\mathbf{P}_{\mathbf{Z}}^{\perp} = \mathbf{I} - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$  is the projection matrix onto the subspace orthogonal to the columns of  $\mathbf{Z}$ .

After applying the column stacking operator  $(\cdot)^V$  to both sides of the identity in (2.7), we obtain

$$\mathbf{Y}^V = (\mathbf{I} \otimes \mathbf{Z})\mathbf{B}^V + \mathbf{U}^V,$$

where  $\otimes$  denotes the Kronecker product. It is evident that  $\mathbf{U}^V \sim \mathcal{N}_{TK}(\mathbf{0}, \mathbf{\Sigma} \otimes \mathbf{I})$ . The generalized least-squares estimator for  $\mathbf{B}^V$  is (Mardia, Kent and Bibby, 1979)

$$\hat{\mathbf{B}}^V = [\mathbf{I} \otimes (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}']\mathbf{Y}^V$$

and coincides with the ML estimator for  $\mathbf{B}$  [see (2.8)].

From the standard properties of the ML estimators, we have (Mardia, Kent and Bibby, 1979): (P<sub>1</sub>)  $\hat{\mathbf{B}}^V \sim \mathcal{N}(\mathbf{B}^V, \mathbf{\Omega})$ , where  $\mathbf{\Omega} = \mathbf{\Sigma} \otimes (\mathbf{Z}'\mathbf{Z})^{-1}$ ; (P<sub>2</sub>)  $T\hat{\mathbf{\Sigma}} \sim \mathcal{W}_K(\mathbf{\Sigma}, T - \ell)$  [Wishart distribution with scale matrix  $\mathbf{\Sigma}$  and degrees of freedom parameter  $T - \ell$ ]; (P<sub>3</sub>)  $\hat{\mathbf{B}}^V$  is statistically independent of  $\hat{\mathbf{\Sigma}}$ .

### First normalization step

We introduce the supplementary notation  $\boldsymbol{\theta} = (\mathbf{B}^V, \mathbf{\Sigma})$ . For simplicity, we write  $\hat{\mathbf{B}}^V$  instead of  $\hat{\mathbf{B}}^V(\mathbf{Y}^V)$  and  $\hat{\mathbf{\Sigma}}^V$  instead of  $\hat{\mathbf{\Sigma}}^V(\mathbf{Y}^V)$ . Hence,  $\hat{\boldsymbol{\theta}} = (\hat{\mathbf{B}}, \hat{\mathbf{\Sigma}})$ .

Using the properties (P<sub>1</sub>)-(P<sub>3</sub>) along with the fact that the statistics  $\hat{\boldsymbol{\theta}}$  are sufficient for  $\boldsymbol{\theta}$  (Mardia, Kent and Bibby, 1979), we get the following chain of identities for the likelihood function:

$$\begin{aligned} f(\mathbf{Y}; \boldsymbol{\theta}) &= f(\mathbf{Y}|\boldsymbol{\theta})g(\boldsymbol{\theta}; \hat{\boldsymbol{\theta}}), \\ g(\boldsymbol{\theta}; \hat{\boldsymbol{\theta}}) &= g_1(\hat{\mathbf{B}}^V; \boldsymbol{\theta})g_2(\hat{\mathbf{\Sigma}}; \mathbf{\Sigma}), \\ g_1(\hat{\mathbf{B}}^V; \boldsymbol{\theta}) &= \frac{\exp[-(\boldsymbol{\delta}'_{\mathbf{B}}\mathbf{\Omega}^{-1}\boldsymbol{\delta}_{\mathbf{B}})/2]}{(2\pi)^{\ell K/2}(\det \mathbf{\Sigma})^{\ell/2}(\det \mathbf{Z}'\mathbf{Z})^{-K/2}}, \\ g_2(\hat{\mathbf{\Sigma}}; \mathbf{\Sigma}) &= \frac{T(\det T\hat{\mathbf{\Sigma}})^{(T-\ell-K-1)/2} \exp\left[-\frac{T}{2}\text{tr}(\mathbf{\Sigma}^{-1}\hat{\mathbf{\Sigma}})\right]}{2^{(T-\ell)K/2}(\det \mathbf{\Sigma})^{(T-\ell)/2}\Gamma_K[(T-\ell)/2]}, \end{aligned}$$

where  $\boldsymbol{\delta}_{\mathbf{B}} = \hat{\mathbf{B}}^V - \mathbf{B}^V$ . After little algebra, we get:

$$g(\hat{\boldsymbol{\theta}}; \hat{\boldsymbol{\theta}}) = G(\det \hat{\mathbf{\Sigma}})^{-(\ell+K+1)/2}, \quad (2.10)$$

where

$$G = \frac{T^{1+K(T-\ell-K-1)/2} \det(\mathbf{Z}'\mathbf{Z})^{K/2} \exp(-TK/2)}{(2^{TK/2})(\pi^{\ell K/2}) \Gamma_K[(T-l)/2]}, \quad (2.11)$$

$$\det \hat{\Sigma} = \prod_{j=1}^K \hat{\lambda}^{(j)}. \quad (2.12)$$

Additionally, we assume that the eigenvalues of  $\hat{\Sigma}$  are ordered as follows:  $\lambda^{(K)} \geq \dots \geq \lambda^{(1)} > 0$ .

As a preliminary step for constraining the integration domain in (2.2), we consider the hyper-parameters  $R > 0$  and  $\lambda_{\min}^{(K)} \geq \dots \geq \lambda_{\min}^{(1)} > 0$ . We take  $\Lambda_{\min} = \left\{ \lambda_{\min}^{(j)} \right\}_{j=1}^K$ . With the convention that  $\|\cdot\|$  denotes the Euclidean norm, we define:

$$\begin{aligned} \mathcal{B}(R) &= \left\{ \hat{\mathbf{B}}^V : \|(\mathbf{I} \otimes \mathbf{Z})\hat{\mathbf{B}}^V\|^2 / (TK) \leq R \right\}, \\ \mathcal{L}(\Lambda_{\min}) &= \left\{ \hat{\Sigma} : \hat{\lambda}^{(j)} \geq \lambda_{\min}^{(j)} \text{ for } j = \overline{1, K} \right\}, \\ \mathcal{T}(R, \Lambda_{\min}) &= \left\{ \hat{\theta} : \hat{\mathbf{B}}^V \in \mathcal{B}(R) \text{ and } \hat{\Sigma} \in \mathcal{L}(\Lambda_{\min}) \right\}, \\ \mathcal{Y}(R, \Lambda_{\min}) &= \left\{ \mathbf{Y}^V : \hat{\theta}(\mathbf{Y}^V) \in \mathcal{T}(R, \Lambda_{\min}) \right\}, \\ \mathcal{Y}(\hat{\theta}) &= \left\{ \mathbf{Y}^V : \hat{\theta}(\mathbf{Y}^V) = \hat{\theta} \right\}. \end{aligned}$$

The constrained normalization factor is:

$$\begin{aligned} C_p(R, \Lambda_{\min}) &= \int_{\mathcal{Y}(R, \Lambda_{\min})} f(\mathbf{Y}^V; \theta(\hat{\mathbf{Y}}^V)) d\mathbf{Y}^V \\ &= \int_{\mathcal{T}(R, \Lambda_{\min})} \left\{ \int_{\mathcal{Y}(\hat{\theta})} f(\mathbf{Y}^V | \hat{\theta}) d\mathbf{Y}^V \right\} g(\hat{\theta}; \hat{\theta}) d\hat{\theta} \end{aligned} \quad (2.13)$$

$$= G \left\{ \prod_{j=1}^K \int_{\lambda_{\min}^{(j)}}^{\infty} [\hat{\lambda}^{(j)}]^{-\frac{\ell+K+1}{2}} d\hat{\lambda}^{(j)} \right\} \int_{\mathcal{B}(R)} d\hat{\mathbf{B}}^V \quad (2.14)$$

$$= G \left\{ \prod_{j=1}^K \frac{[\lambda_{\min}^{(j)}]^{-(l+K-1)/2}}{(l+K-1)/2} \right\} \text{vol}[\mathcal{B}(R)]. \quad (2.15)$$

In (2.13), we used the fact that the inner integral equals one, while in (2.14) we applied the identities in (2.10)-(2.12). For the volume which appears in (2.15), it is easy to show [see, for example, (Giurcăneanu, Razavi and Liski, 2011)] that

$$\text{vol}[\mathcal{B}(R)] = \eta R^{\zeta K\ell},$$

where

$$\begin{aligned}\eta &= \frac{(TK\pi)^{lK/2}}{\Gamma[(lK)/2 + 1] \det((\mathbf{I} \otimes \mathbf{Z})'(\mathbf{I} \otimes \mathbf{Z}))^{1/2}}, \\ \zeta &= 1/2.\end{aligned}$$

One possible option for computing  $C_p(R, \Lambda_{\min})$  is to make subjective selections for the value of  $R$  and the entries of  $\Lambda_{\min}$ . However, we follow the recommendations from (Rissanen (2007)) and perform another normalization step. Before discussing this step, we observe that  $C_p(R, \Lambda_{\min})$  becomes smaller when  $R$  decreases. Keeping in mind that we want to minimize the “code length” given by

$$-\log \hat{f}(\mathbf{Y}^V; p) = -\log f\left(\mathbf{Y}^V; \hat{\mathbf{B}}(\mathbf{Y}^V), \hat{\mathbf{\Sigma}}(\mathbf{Y}^V)\right) + \log C_p(R, \Lambda_{\min}).$$

we take

$$R = \frac{\|(\mathbf{I} \otimes \mathbf{Z})\hat{\mathbf{B}}^V\|^2}{KT}, \tag{2.16}$$

$$= \frac{\text{tr}\left(\mathbf{Y}'\mathbf{Y} - T\hat{\mathbf{\Sigma}}\right)}{KT}. \tag{2.17}$$

The selection of the  $R$ -value in (2.16) is mainly determined by the definition of  $\mathcal{B}(R)$ . The identity in (2.17) is straightforwardly obtained from  $\hat{\mathbf{\Sigma}} = (\mathbf{Y}'\mathbf{P}_{\mathbf{Z}}^{\perp}\mathbf{Y})/T$ , where  $\mathbf{P}_{\mathbf{Z}}^{\perp} = \mathbf{I} - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$  is the projection matrix onto the subspace orthogonal to the columns of  $\mathbf{Z}$  [see (Mardia, Kent and Bibby, 1979)].

Similar considerations lead to

$$\lambda_{\min}^{(j)} = \hat{\lambda}_j, \quad j = 1, \dots, K.$$

Hence, we have:

$$\begin{aligned} \log C_p(R, \Lambda_{\min}) = & -\log \Gamma_K[(T - \ell)/2] + K \log \frac{2}{\ell + K - 1} \\ & - \frac{\ell + K - 1}{2} \log \det \hat{\Sigma} + \frac{\ell K}{2} \log \frac{\text{tr}(\mathbf{Y}'\mathbf{Y} - T\hat{\Sigma})}{T} + \log Ct, \end{aligned}$$

where

$$Ct = \frac{T^{1+K(T-K-1)/2} \exp(-TK/2)}{2^{TK/2}}.$$

Remark that  $\log Ct$  does not depend on the order  $p$  of the model, so it can be ignored in our future calculations.

### Second normalization step

We choose  $R_1, R_2, \lambda_1, \lambda_2$  such that  $R_2 > R_1 > 0$  and  $\lambda_2 > \lambda_1 > 0$ . As in the previous definitions, we have:

$$\begin{aligned} \mathcal{B}(R_1, R_2) &= \left\{ \hat{\mathbf{B}}^V : R_1 \leq \frac{\|(\mathbf{I} \otimes \mathbf{Z})\hat{\mathbf{B}}^V\|^2}{TK} \leq R_2 \right\}, \\ \mathcal{L}(\lambda_1, \lambda_2) &= \left\{ \hat{\Sigma} : \lambda_1 \leq \hat{\lambda}^{(j)} \leq \lambda_2 \text{ for } j = \overline{1, K} \right\}, \\ \mathcal{T}(R_1, R_2, \lambda_1, \lambda_2) &= \left\{ \hat{\boldsymbol{\theta}} : \hat{\mathbf{B}}^V \in \mathcal{B}(R_1, R_2) \text{ and } \hat{\Sigma} \in \mathcal{L}(\lambda_1, \lambda_2) \right\}, \\ \mathcal{Y}(R_1, R_2, \lambda_1, \lambda_2) &= \left\{ \mathbf{Y}^V : \hat{\boldsymbol{\theta}}(\mathbf{Y}^V) \in \mathcal{T}(R_1, R_2, \lambda_1, \lambda_2) \right\}. \end{aligned}$$

The normalization term  $\overline{C}_p(R_1, R_2, \lambda_1, \lambda_2)$  is computed as follows:

$$\begin{aligned} \overline{C}_p(R_1, R_2, \lambda_1, \lambda_2) &= \int_{\mathcal{Y}(R_1, R_2, \lambda_1, \lambda_2)} \frac{f(\mathbf{Y}^V; \boldsymbol{\theta}(\mathbf{Y}^V))}{C_p(R, \Lambda_{\min})} d\mathbf{Y}^V \\ &= \int_{\mathcal{T}(R_1, R_2, \lambda_1, \lambda_2)} \frac{g(\hat{\boldsymbol{\theta}}; \hat{\boldsymbol{\theta}})}{C_p(R, \Lambda_{\min})} d\hat{\boldsymbol{\theta}} \\ &= \left[ \frac{2}{\ell + K - 1} \right]^{-K} \left[ \prod_{j=1}^K \int_{\lambda_1}^{\lambda_2} \frac{1}{\hat{\lambda}^{(j)}} d\hat{\lambda}^{(j)} \right] \times \int_{R_1}^{R_2} \frac{K\ell}{2} \frac{1}{R} dR \end{aligned}$$

For proving the last identity, we have used the expression of  $g(\hat{\boldsymbol{\theta}}; \hat{\boldsymbol{\theta}})$  from (2.10)-(2.12) as well as the expression of  $C_p(R, \Lambda_{\min})$  from (2.15). For computing the integral of the inverse of the volume from  $C_p(R, \Lambda_{\min})$ -formula, we have observed

that the expression of the volume coincides with the one from (Giurcăneanu, Razavi and Liski, 2011, Eq. (10)) and we employed the identities from (Giurcăneanu, Razavi and Liski, 2011, Appendix A). Hence, we get:

$$\begin{aligned} \log \bar{C}_p(R_1, R_2, \lambda_1, \lambda_2) \\ = -K \log \frac{2}{\ell + K - 1} + \log \ell \\ + \log \frac{K}{2} \end{aligned} \quad (2.18)$$

$$+ K \log \log \frac{\lambda_2}{\lambda_1} \quad (2.19)$$

$$+ \log \log \frac{R_2}{R_1}. \quad (2.20)$$

We ignore the term in (2.18) because is constant. Due to the same reasons as those invoked in (Rissanen, 2007, Hirai and Yamanishi, 2013), we drop the terms in (2.19) and (2.20).

After collecting all the terms corresponding to  $\log C_p(R, \Lambda_{\min})$  and  $\log \bar{C}_p(R_1, R_2, \lambda_1, \lambda_2)$  along with negative log-likelihood, we obtain the RNML criterion which is presented in Proposition 2.3.1.  $\square$

## 2.4 RNML formula for $p = 0$

In this case, we only need to compute the estimate  $\hat{\Sigma}_0 = (\mathbf{Y}'\mathbf{Y})/T$ . All other calculations are similar to those from the previous section, but simpler. It is easy to show that the logarithm of the normalization factor which corresponds to  $\log C_p(R, \Lambda_{\min})$  is

$$\log C_0(\Lambda_{\min,0}) = -\frac{K-1}{2} \log \det \hat{\Sigma}_0 - \log \Gamma_K \left( \frac{T}{2} \right) + K \log \frac{2}{K-1},$$

where  $\Lambda_{\min,0}$  is the set of eigenvalues of  $\hat{\Sigma}_0$ . Furthermore, if we constrain these eigenvalues to the interval  $[\lambda_3, \lambda_4]$ , then we get the logarithm of the normalization factor which corresponds to  $\log \bar{C}_p(R_1, R_2, \lambda_1, \lambda_2)$ :

$$\log \bar{C}_0(\lambda_3, \lambda_4) = K \log \log \frac{\lambda_4}{\lambda_3} - K \log \frac{2}{K-1}.$$

It follows that

$$\begin{aligned} \text{RNML}(\mathbf{Y}; p = 0) &= \frac{T - K + 1}{2} \log \det \hat{\Sigma}_0 \\ &\quad - \log \Gamma\left(\frac{T}{2}\right) + K \log \log \frac{\lambda_4}{\lambda_3}. \end{aligned}$$

It is a simple exercise to verify that the expression above reduces to the one in (Rissanen, 2007, Eq. (9.40)) when  $K = 1$  (univariate case). The discussion on the role of the hyper-parameters,  $\lambda_3$  and  $\lambda_4$ , goes along the same lines as in (Rissanen, 2007). From identity in (2.17) and the definitions of  $\mathcal{B}(R_1, R_2)$  and  $\mathcal{L}(\lambda_1, \lambda_2)$  in Section 2.3, we have the double inequality:

$$R_1 + \lambda_1 \leq \text{tr}(\hat{\Sigma}_0)/K \leq R_2 + \lambda_2.$$

At the same time,  $\lambda_3 \leq \text{tr}(\hat{\Sigma}_0)/K \leq \lambda_4$ , which makes it natural to choose  $\lambda_3 = R_1 + \lambda_1$  and  $\lambda_4 = R_2 + \lambda_2$ . If we apply the same technique as in (Rissanen, 2007) by taking  $\lambda_1 = R_1 = a$  and  $\lambda_2 = R_2 = b$  ( $0 < a < b$ ), the contribution of the hyper-parameters to  $\text{RNML}(\mathbf{Y}; p)$  is  $(K + 1) \log \log(b/a)$  for  $p > 0$ . When comparing this result with  $K \log \log(b/a)$ , which is the contribution of the hyper-parameters to  $\text{RNML}(\mathbf{Y}; p = 0)$ , we can conclude that neglecting the hyper-parameters might have a negative impact on the selection of the model. The formula for  $\text{RNML}(\mathbf{Y}; p = 0)$  is not used in this thesis and its derivation was included only for the sake of completeness. The expression of RNML which we use is the one given in Proposition 2.3.1. For better understanding the significance of this formula, we demonstrate below that it is asymptotically equivalent to the well-known Bayesian Information Criterion (BIC). Because in the time series literature, the term Schwarz Bayesian Criterion (SBC) is employed instead of BIC, we prefer to use this acronym.

## 2.5 Asymptotic equivalence between RNML and SBC

This analysis aims to clarify the relationship between RNML and SBC, whose formula is (Schwarz, 1978)

$$\text{SBC}(\mathbf{Y}; p) = \frac{T}{2} \log \det \hat{\Sigma}_p + \frac{K^2 p}{2} \log T.$$

We are interested in their relative behavior when  $T \rightarrow \infty$ . Before presenting the main proposition of this section, we prove an auxiliary result.

**Lemma 2.5.1.** *Assume that  $T \rightarrow \infty$ ,  $K$  is fixed and  $p$  does not increase with  $T$ . With the notation from Proposition 2.3.1, we have:*

$$\text{GOF} = \left[ \frac{T}{2} \log \det \hat{\Sigma}_p \right] [1 - o(1)]$$

and

$$\text{PEN}_1 = \left[ \frac{K^2 p}{2} \log T \right] [1 - o(1)].$$

*Proof.*

The identity for GOF can be obtained straightforwardly. For  $\text{PEN}_1$ , we note that the multivariate Gamma function can be written as

$$\frac{\Gamma_K \left[ \frac{T-Kp}{2} \right]}{\pi^{\frac{K(K-1)}{4}}} = \prod_{i=1}^K \Gamma \left[ \frac{T - Kp + 1 - i}{2} \right]. \quad (2.21)$$

Furthermore, we use the Stirling approximation (Artin, 1964, p. 24):

$$\log \Gamma(z) = \frac{1}{2} \log(2\pi) + \left( z - \frac{1}{2} \right) \log z - z + \frac{\theta}{12z},$$

where  $z > 0$  and  $0 < \theta < 1$ .

We neglect the factor  $\pi^{K(K-1)/4}$  in (2.21) because it does not depend on the model order. With the convention that  $\ell = Kp$ , we have:

$$\begin{aligned}
 & \log \Gamma_K \left( \frac{T - \ell}{2} \right) \\
 &= \sum_{i=1}^K \log \Gamma \left( \frac{T - \ell + 1 - i}{2} \right) \\
 &= \sum_{i=1}^K \frac{T - \ell - i}{2} \log \frac{T - \ell + 1 - i}{2} \\
 &\quad - \sum_{i=1}^K \frac{T - \ell + 1 - i}{2} + \frac{K}{2} \log(2\pi) + \mathcal{O} \left( \frac{1}{T} \right).
 \end{aligned}$$

Because the terms which do not depend on  $p$  can be ignored, we write

$$\begin{aligned}
 & -\text{PEN}_1 \\
 &= \frac{1}{2} \sum_{i=1}^K (T - \ell - i) \log(T - \ell + 1 - i) + \frac{K\ell}{2} (1 + \log 2) \\
 &= \frac{1}{2} \sum_{i=1}^K (T - \ell - i) \log T \\
 &\quad + \frac{1}{2} \sum_{i=1}^K (T - \ell - i) \log \left( 1 - \frac{\ell - 1 + i}{T} \right) \\
 &\quad + \frac{K\ell}{2} (1 + \log 2).
 \end{aligned}$$

We drop the constant terms and use the Maclaurin series expansion for

$\log [1 - (\ell - 1 + i)/T]$  when  $1 \leq i \leq K$ :

$$\begin{aligned}
 & -\text{PEN}_1 \\
 &= -\frac{K\ell}{2} \log T + \frac{K\ell}{2} (1 + \log 2) \\
 &\quad + \frac{1}{2} \sum_{i=1}^K (T - \ell - i) \log \left( 1 - \frac{\ell - 1 + i}{T} \right) \\
 &= -\frac{K\ell}{2} \log T + \frac{K\ell}{2} (1 + \log 2) \\
 &\quad - \frac{1}{2} \sum_{i=1}^K (T - \ell - i) \frac{\ell - 1 + i}{T} - \mathcal{O} \left( \frac{1}{T} \right) \\
 &= -\frac{K\ell}{2} \log T + \frac{K\ell}{2} \log 2 \\
 &\quad + \frac{1}{2T} \sum_{i=1}^K (\ell + i)(\ell + i - 1) - \mathcal{O} \left( \frac{1}{T} \right) \\
 &= -\frac{K\ell}{2} \log T + \frac{K\ell}{2} \log 2 \\
 &\quad + \frac{1}{2T} \left[ \ell^2 K + \ell K^2 + \frac{K^3}{3} - \frac{K}{3} \right] - \mathcal{O} \left( \frac{1}{T} \right) \\
 &= -\left[ \frac{K\ell}{2} \log T \right] [1 - o(1)] \\
 &= -\left[ \frac{K^2 p}{2} \log T \right] [1 - o(1)].
 \end{aligned}$$

This concludes the proof.  $\square$

**Proposition 2.5.1.** *RNML and SBC reduce to the same formula when  $T \rightarrow \infty$ . Assuming that the measurements  $\{\mathbf{y}_t\}_{t=1}^T$  are outcomes from a stable and stationary VAR-process with zero-mean vector and order  $p^\circ > 0$ , RNML is a strongly consistent estimator for the order of the process.*

*Proof.*

From the definitions given in Proposition 2.3.1, it is obvious that  $\text{PEN}_2$  becomes negligible with respect to  $\text{PEN}_1$  as  $T \rightarrow \infty$ . We will show below (see Remark 1) that  $\text{PEN}_3$  is bounded when  $T \rightarrow \infty$ . Then, Lemma 2.5.1 gives the stated relation between RNML and SBC. The consistency property of RNML is hence the same as for SBC, proved in (Lutkepöhl, 2005, p. 150).  $\square$

We now analyze  $\text{PEN}_3$ , which is the most intriguing penalty term because it does not depend only on the number of variables ( $K$ ), sample size ( $T$ ), VAR-order ( $p$ ), but also on the measurements [see (2.6)]. We propose to find an upper bound for  $\text{PEN}_3$  when  $T \rightarrow \infty$ . For any integer  $h$ , let  $\mathbf{R}(h) = \text{Cov}(\mathbf{y}_t, \mathbf{y}_{t-h})$  be the autocovariance matrix at lag  $h$ , for the time series  $\mathbf{Y}$ . According to (Reinsel, 1993, Section 3.3), the model of the time series can be “approximated” by a  $\text{VAR}(p)$ . For convenience, we assume that both  $p$  and  $p^\circ$  are from the set  $\{1, \dots, p_{\max}\}$ .

*Remark 1.* We make the convention that  $\Sigma_0 = \mathbf{R}(0)$ . It is known from (Reinsel, 1993, p. 75) that  $(\mathbf{Y}'\mathbf{Y})/T$  converges to  $\Sigma_0$  almost surely as  $T \rightarrow \infty$ . Similarly, for  $p > 0$ ,  $\hat{\Sigma}_p \rightarrow \Sigma_p$  almost surely. Hence, asymptotically in  $T$  we have:  $\text{PEN}_3 \leq (K^2 p/2) \log \text{tr}(\Sigma_0)$ . Note that the upper bound for  $\text{PEN}_3$  does not depend on  $T$ .

We show in the next proposition how the upper bound for  $\text{PEN}_3$  can be further improved. More importantly, the proof of the proposition reveals the relationship between  $\text{PEN}_3$  and the partial autocorrelation matrix.

**Proposition 2.5.2.** *When  $T \rightarrow \infty$ , if  $\text{RNML}(\mathbf{Y}; p)$  is evaluated for a data matrix  $\mathbf{Y}$  produced by a  $\text{VAR}(p^\circ)$ -model, then the following inequality holds:*

$$\text{PEN}_3 \leq \frac{Kp}{2} \log \det \Sigma_p + \frac{K^2 p}{2} \log K + \frac{K^2 p}{2} \log \left[ \frac{\phi(\Sigma_0)}{\psi(p, p^\circ)} - 1 \right],$$

where  $\phi(\Sigma_0) = \frac{\text{tr}(\Sigma_0)}{K(\det \Sigma_0)^{1/K}}$  and  $\psi(p, p^\circ)$  has the properties: (i)  $\psi(p, p^\circ) \in (0, 1)$  for all  $p, p^\circ \in \{1, \dots, p_{\max}\}$ ; (ii)  $\psi(p+1, p^\circ) \leq \psi(p, p^\circ)$  for  $p = 1, \dots, p^\circ - 1$ ; (iii)  $\psi(p, p^\circ) = \psi(p^\circ, p^\circ)$  for  $p = p^\circ + 1, \dots, p_{\max}$ .

*Proof.*

Firstly we employ the result from (Reinsel, 1993, p. 75) which says that the sample covariance matrix  $\hat{\mathbf{R}}(h)$  converges to  $\mathbf{R}(h)$  almost surely as  $T \rightarrow \infty$ . This result along with Yule-Walker equations (Reinsel, 1993, Eq. (9.a)) allow us to replace, in our asymptotic analysis,  $\hat{\Sigma}_p$  with  $\Sigma_p$  for all  $p \geq 0$ .

As in (Reinsel, 1993, p. 69), we take  $\mathbf{u}_{p,t}$  and  $\overleftarrow{\mathbf{u}}_{p,t-p}$  to be the residual vectors from the multivariate regressions of  $\mathbf{y}_t$  and  $\mathbf{y}_{t-p}$  on the set of predictor variables  $\mathbf{y}_{t-1}, \dots, \mathbf{y}_{t-p+1}$ . Hence, we have  $\Sigma_{p-1} = \text{Cov}(\mathbf{u}_{p,t})$ . Additionally, we introduce the notation  $\overleftarrow{\Sigma}_{p-1} = \text{Cov}(\overleftarrow{\mathbf{u}}_{p,t-p})$ .

Furthermore, for  $p > 0$ , we define the *partial correlation matrix* (see also (Reinsel, 1993, p. 70)(Wei, 2006, Eq. (16.5.52))):

$$\check{\mathbf{Q}}(p) = \left[ \overleftarrow{\Sigma}_{p-1}^{1/2} \right]^{-1} \text{Cov}(\overleftarrow{\mathbf{u}}_{p,t-p} \mathbf{u}_{p,t}') \left[ \Sigma_{p-1}^{1/2} \right]^{-1}, \quad (2.22)$$

where  $\overleftarrow{\Sigma}_{p-1}^{1/2}$  and  $\Sigma_{p-1}^{1/2}$  are the symmetric square roots of  $\overleftarrow{\Sigma}_{p-1}$  and  $\Sigma_{p-1}$ , respectively. If in (2.22) we replace  $\overleftarrow{\Sigma}_{p-1}^{1/2}$  with the lower triangular root of  $\overleftarrow{\Sigma}_{p-1}$  and  $\Sigma_{p-1}^{1/2}$  with the upper triangular root of  $\Sigma_{p-1}$ , then we obtain the transpose of the matrix defined in (Morf, Vieira and Kailath, 1978, Eq. (14)).

Note that the entries of  $\check{\mathbf{Q}}(p)$  are correlation coefficients only in the particular case when  $K = 1$  [see (Wei, 2006, p. 413-414) for a more detailed discussion]. However, the eigenvalues of  $\check{\mathbf{Q}}'(p)\check{\mathbf{Q}}(p)$  belong to the interval  $[0, 1]$  and they are equal to “the (squared) partial canonical correlations between the vectors  $\mathbf{y}_t$  and  $\mathbf{y}_{t-p}$  after adjustment for the dependence of these variables on the intervening values  $\mathbf{y}_{t-1}, \dots, \mathbf{y}_{t-p+1}$ ” (Reinsel, 1993, p. 71).

It follows from (Morf, Vieira and Kailath, 1978, p. 646) that

$$\det \Sigma_p = \det \Sigma_0 \prod_{i=1}^p \det(\mathbf{I} - \check{\mathbf{Q}}'(i)\check{\mathbf{Q}}(i)). \quad (2.23)$$

Additionally, we have that  $\check{\mathbf{Q}}(i) = \mathbf{0}$  when  $i > p^\circ$ . Since we assume that  $\Sigma_p \succ 0$ , all squared partial canonical correlations are strictly smaller than one.

Using the inequality of arithmetic and geometric means, we readily obtain

$$\begin{aligned} \text{tr}(\Sigma_0 - \Sigma_p) &\leq \text{tr}(\Sigma_0) - K \det \Sigma_p^{1/K} \\ &= K \det \Sigma_p^{1/K} \left[ \frac{\phi(\Sigma_0)}{\psi(p, p^\circ)} - 1 \right], \end{aligned}$$

where  $\psi(p, p^\circ) = \prod_{i=1}^p \det(\mathbf{I} - \check{\mathbf{Q}}'(i)\check{\mathbf{Q}}(i))^{1/K}$ . All that remains is to employ the inequality above in conjunction with the definition of  $\text{PEN}_3$ .  $\square$

*Remark 2.* From (2.23), we have that  $\psi(p, p^\circ) = (\det \Sigma_p / \det \Sigma_0)^{1/K}$ , which leads to the identity  $\frac{K^2 p}{2} \log \text{tr}(\Sigma_0) = \frac{K p}{2} \log \det \Sigma_p + \frac{K^2 p}{2} \log K + \frac{K^2 p}{2} \log \frac{\phi(\Sigma_0)}{\psi(p, p^\circ)}$ . This demonstrates that the upper bound for  $\text{PEN}_3$  in Proposition 2.5.2 is sharper than the one in Remark 1.

*Remark 3.* Note that, in the proof of Proposition 2.5.2, we have shown the relationship between the upper bound for  $\text{PEN}_3$  and the eigenvalues of the partial correlation matrix. As we have already pointed out, the statistical significance of these eigenvalues is well-known in the literature focused on time series analysis.

## 2.6 Experimental Results

In this section, we conduct experiments with simulated data for comparing RNML with other six IT criteria. The formulae of the criteria as well as a short description of their derivations can be found in Table 2.1.

### 2.6.1 Example 1

We consider a VAR-model for which  $K$  is large ( $K = 20$ ) and the “true” order takes small values ( $p^\circ = 1$  or  $p^\circ = 2$ ). This model was originally proposed in (Ting et al., 2015) and assumes that the driven noise is Gaussian and  $\Sigma_{p^\circ} = \mathbf{I}$ . The matrix coefficients of the model  $(\mathbf{A}_{1p^\circ}, \dots, \mathbf{A}_{p^\circ p^\circ})$  are of the form  $[\mathbf{D}_1 \ \mathbf{0}; \mathbf{0} \ \mathbf{D}_2]$ , where both  $\mathbf{D}_1$  and  $\mathbf{D}_2$  are  $10 \times 10$ . Remark that we use notational conventions like those from Matlab. The entries of  $\mathbf{D}_1$  and  $\mathbf{D}_2$  are statistically independent and they are drawn from a uniform distribution on the interval  $(-1/2, 1/2)$ . Additionally, all entries of the matrix coefficients which do not belong to the main diagonal are divided by  $1.35^{p^\circ}$ . We emphasize that, for  $p^\circ = 2$ , the non-zero entries of  $\mathbf{A}_{12}$  and  $\mathbf{A}_{22}$  are statistically independent and they are statistically independent with respect to the driven noise. In our experiments, for each  $p^\circ$ , we consider  $10^4$  realizations of the matrix coefficients. For each realization, a set of 275 samples is produced by randomly generating the driven noise. The first 200 samples are employed to estimate  $\hat{\mathbf{A}}_{1p}, \dots, \hat{\mathbf{A}}_{pp}$  and  $\hat{\Sigma}_p$  for

## 2. RENORMALIZED MAXIMUM LIKELIHOOD

Acronym	Penalty Term	Reference
SBC	$\frac{K^2 p}{2} \log T$	<a href="#">Schwarz (1978)</a>
log FPE	$\frac{KT}{2} \log \frac{T + pK}{T - pK}$	<a href="#">Akaike (1971)</a>
Based on the approximate covariance matrix of one-step ahead forecast errors		
AIC	$pK^2$	<a href="#">Akaike (1974)</a>
Asymptotically unbiased estimator for Kullback directed divergence		
AIC <sub>c</sub>	$\frac{T}{T - pK - K - 1} \left[ pK^2 + \frac{K(K + 1)}{2} \right]$	<a href="#">Hurvich and Tsai (1993)</a>
AIC corrected for small sample sizes		
KIC	$\frac{3pK^2}{2}$	<a href="#">Cavanaugh (1999)</a>
Asymptotically unbiased estimator for Kullback symmetric divergence		
KIC <sub>c</sub>	$\frac{TK(2pK + K + 1)}{2(T - pK - K - 1)} + \frac{TK}{2(T - pK) - (K - 1)} + \frac{2pK^2 + K^2 - K}{4}$	<a href="#">Seghouane (2006)</a>
KIC corrected for small sample sizes		

Table 2.1: Expressions of various IT criteria for a VAR of order  $p$  fitted to a data set that contains  $K$  time series of length  $T$ . For all the criteria listed in the table, the goodness-of-term is  $(T/2) \log \det \hat{\Sigma}_p$ . For simplicity, we remove from the penalty terms those quantities which do not depend on  $p$ . We prefer to write the formula of log FPE instead of FPE for facilitating the comparison with other criteria. It is easy to see that the formula of SBC is the same with the one given in Section 2.5.

$p = \overline{1, 8}$  by using the ARFIT-algorithm ([Neumaier and Schneider, 2001](#)). Then the best order is selected with RNML and six other IT criteria which are presented in Table 2.1. The same is done for the first 225 samples, then for the first 250 samples and eventually the whole data is used in the estimation process.

According to the plots shown in Figure 2.1, when  $p^\circ = 1$ , we have: (i) SBC, RNML, AIC<sub>c</sub> and KIC<sub>c</sub> perfectly estimate the correct order in all trials; (ii) AIC overestimates the order in all runs for which the sample size is  $T = 200$ ; (iii) For

$T = 225$ , AIC estimates correctly the order only 40 times out of  $10^4$ ; (iv) The performance of KIC is only slightly worse than that of  $\text{KIC}_c$ ; (v) FPE and KIC have the same level of performance.

As we can see Figure 2.1, the ranking of criteria changes when  $p^\circ = 2$ : (i) SBC severely underestimates the order and achieves a moderate score of 60% correct estimations only when  $T = 275$ ; (ii) RNML is correct in at least 90% of the runs, disregarding the sample size; (iii)  $\text{AIC}_c$  is ranked the best and is much better than AIC; (iv)  $\text{KIC}_c$  is better than KIC; (v) FPE is very good, except for  $T = 200$ . We note that, in (Ting et al., 2015), FPE was not considered and the estimation results were reported only for  $T = 200$ . The results we report for this sample size are similar to those in (Ting et al., 2015) and we assume that the differences are due to the fact that we apply a different algorithm for estimating the matrix coefficients of the model.

### 2.6.2 Example 2

We simulate data according to a VAR-model for which  $K = 5$  and  $p^\circ \in \{1, 5, 10, 15\}$ . As we are interested in the sparsity of ISDM of the VAR-model, we define  $N_{\text{SP}} = 9$  sparsity patterns which are denoted  $\{\text{SP}_i\}_{i=0}^8$ . After setting  $\text{SP}_0 = \emptyset$  and  $(u, v) = (1, 2)$ , we apply the following recursions, for  $i = \overline{0, 7}$ : (i)  $\text{SP}_{i+1} \leftarrow \text{SP}_i \cup \{(u, v)\}$  and (ii) if  $v < K$ , then  $(u, v) \leftarrow (u, v + 1)$ , else  $(u, v) \leftarrow (u + 1, u + 2)$ . Remark that  $\text{SP}_0 \subset \text{SP}_1 \subset \dots \subset \text{SP}_8$ .

Inspired by (Songsiri and Vandenberghe, 2010, Example 2), we generate for each SP in  $\{\text{SP}_i\}_{i=0}^8$  an ISDM with the property that the entries of  $\{\mathbf{Q}_m\}_{m=1}^{p^\circ}$  [see (1.3)] are zero in the positions corresponding to SP, and all other entries are randomly drawn from the univariate Gaussian distribution with mean  $2 \times 10^{-1}$  and variance  $10^{-4}$ . The matrix  $\mathbf{Q}_0$  is similarly produced, except that integer multiples of the identity matrix are added to it until ISDM is positive definite. Furthermore, we use spectral factorization of ISDM [see (Dumitrescu, 2007, App.B.5)] for obtaining the matrix polynomial  $\mathbf{A}_{\text{SP}}$  of order  $p^\circ$ . The covariance matrix  $\mathbf{\Sigma}_{\text{SP}}$  is a byproduct of this procedure. We simulate  $N_r$  different

$K$ -variate time series of length  $T_{\max}$  by using  $\mathbf{A}_{\text{SP}}$  and  $\mathbf{\Sigma}_{\text{SP}}$  in (1.2).

In our settings,  $N_r = 100$  and  $T_{\max} = 27000$ . Hence, for each  $p^\circ$ -order, the number of simulated  $K$ -variate time series is  $N_{\text{SP}} \times N_r = 900$ . Each time series is used to estimate the matrix coefficients and  $\hat{\mathbf{\Sigma}}_p$  for  $p = \overline{1, 20}$  by employing again the implementation of ARFIT algorithm (Neumaier and Schneider, 2001), available at <http://climate-dynamics.org/software/#arfit>. As in Example 1, the order is selected by the RNML and the criteria listed in Table 2.1. Firstly a subset of measurements ( $\mathbf{Y} = [\mathbf{y}_1, \dots, \mathbf{y}_T]'$ ) with  $T = 600$  is employed for VAR-order estimation and then the value of  $T$  is increased as follows: (i)  $T \leftarrow T + 100$  when  $600 \leq T \leq 900$  and (ii)  $T \leftarrow 3T$  when  $1000 \leq T \leq 9000$ . We count how many times each criterion selects the correct order. The results are shown in Figure 2.2. For  $p^\circ = 1$ , all seven criteria correctly estimate the order of the model in all runs and for all sample sizes. However, the ability of the criteria to correctly estimate the order changes for higher orders. When  $p^\circ = 5$ , FPE, RNML, AIC and AIC<sub>c</sub> yield the best estimates when  $T \leq 900$  by correctly selecting the order in 70% to 100% of the cases, while KIC and KIC<sub>c</sub> are much weaker; SBC selects wrong orders in all runs for which  $T \leq 1000$ . When  $p^\circ = 10$  and  $T \leq 900$ , we can observe in Figure 2.2 that SBC, KIC, and KIC<sub>c</sub> fail to estimate correctly the order. For these experimental settings, RNML is ranked the best. It is remarkable that, for  $p^\circ = 10$  and  $T \in \{900, 1000\}$ , RNML is the only IT criterion which selects correctly the order in more than 50% of the cases. For  $p^\circ = 15$ , the performance of all IT criteria declines when  $T$  is small. RNML is the only criterion which, at least for some runs, selects the true order when  $T \leq 900$ . This is evident in Figure 2.2. We can conclude that RNML is superior to other criteria when  $p^\circ$  is large.

As this thesis is mainly focused on the estimation of the ISDM, we perform the following experiment. After the order  $\hat{p}$  of the model is selected with an IT criterion, the autocovariance matrices  $\hat{\mathbf{R}}(0), \dots, \hat{\mathbf{R}}(\hat{p})$  can be easily estimated from the data. Furthermore, an estimate of ISDM can be obtained by solving a convex optimization problem which maximizes the entropy rate subject to the

following constraints: (i) the spectral density matrix matches  $\hat{\mathbf{R}}(0), \dots, \hat{\mathbf{R}}(\hat{p})$  and (ii) the sparsity pattern of ISDM is SP. For details, we refer to (Songiri, Dahl and Vandenberghe, 2010). The optimization problem will be also presented in Section 3.1. Because we want to evaluate the impact of model order selection on the accuracy of this estimation, we assume that SP is known. More precisely, we generate data as described above, but only for  $p^\circ = 10$ . This time, we reduce the number of sample sizes by dropping  $T = 27000$  and the number of SP's is also diminished because we do not consider  $\text{SP}_0$ . The number of runs is  $N_r = 100$ , which means that the number of  $K$ -variate time series for each sample size is  $(N_{\text{SP}} - 1) \times N_r = 800$ .

In order to clarify the notation, let us assume that  $\mathbf{S}(\omega)$  and  $\hat{\mathbf{S}}(\omega)$  are the matrix spectral densities for the “true” model and the estimated model, respectively. Recall that the order of the “true” model is  $p^\circ$ , while the order of the estimated model is  $\hat{p}$ . The maximum entropy estimate,  $\hat{\mathbf{S}}^{\text{ME}}(\omega)$ , corresponds also to a model of order  $\hat{p}$  and has the property that the sparsity of its ISDM is the same as the “true” SP. We take  $N_{\text{grid}} = 1024$  and we evaluate  $\mathbf{S}(\omega), \hat{\mathbf{S}}(\omega), \hat{\mathbf{S}}^{\text{ME}}(\omega)$  for  $\omega \in \mathcal{G}$ , where  $\mathcal{G} = \left\{ \frac{0 \times \pi}{N_{\text{grid}}}, \frac{1 \times \pi}{N_{\text{grid}}}, \dots, \frac{N_{\text{grid}} \times \pi}{N_{\text{grid}}} \right\}$ . In order to investigate how far is  $\mathbf{S}(\omega)$  from  $\hat{\mathbf{S}}(\omega)$ , we calculate the  $I$ -divergence between them by applying the general formula for two positive-definite matrices  $\mathbf{F}$  and  $\mathbf{G}$  (Speed and Kiiveri, 1986):

$$D(\mathbf{F}||\mathbf{G}) = -\frac{1}{2} [\log \det(\mathbf{F}\mathbf{G}^{-1}) + \text{tr}(\mathbf{I} - \mathbf{F}\mathbf{G}^{-1})]$$

Given that  $I(\omega)$  is the  $I$ -divergence between  $\mathbf{S}(\omega)$  and  $\hat{\mathbf{S}}(\omega)$ , we compute  $I_{\max} = \max_{\omega \in \mathcal{G}} I(\omega)$ . Similarly,  $I_{\max}^{\text{ME}}$  is the maximum of the  $I$ -divergence between  $\mathbf{S}(\omega)$  and  $\hat{\mathbf{S}}^{\text{ME}}(\omega)$  when  $\omega \in \mathcal{G}$ . Statistics concerning  $I_{\max}$  and  $I_{\max}^{\text{ME}}$  are plotted in Figure 2.4. Observe that RNML is the best among all criteria because it minimizes the maximum for each of the two  $I$ -divergences. This is true for all sample sizes that we have considered in our experiment (see again Figure 2.4). We also give an interpretation of the estimation results by resorting to multivariate Itakura-Saito divergence.

The multivariate Itakura-Saito divergence between the “true”  $\text{VAR}(p^\circ)$  and

the estimated  $\text{VAR}(\hat{p})$  is given by  $J = \frac{1}{2\pi} \int_0^{2\pi} D(\mathbf{S}(\omega) || \hat{\mathbf{S}}(\omega)) d\omega$  (Bach and Jordan, 2004, Ferrante, Masiero and Pavon, 2012). An approximation of  $J$  can be easily obtained from the values of the  $I$ -divergence computed on the grid  $\mathcal{G}$ . The same method can be applied for the evaluation of  $J^{\text{ME}}$ , the multivariate Itakura-Saito divergence between the “true” model and its maximum-entropy estimate. The statistics for  $J$  and  $J^{\text{ME}}$  are shown in Figure 2.4. When  $T \leq 1000$ , RNML yields values of  $J$  and  $J^{\text{ME}}$  which are larger than the values of divergences produced by other IT criteria. Bearing in mind that, for these sample sizes, RNML is the best in selecting the model order, we calculate a new set of statistics only from those runs where RNML estimates correctly the order. For differentiating between these statistics and those computed previously, we use the notation  $J_c$  instead of  $J$  and  $J_c^{\text{ME}}$  instead of  $J^{\text{ME}}$ . Observe in Figure 2.5 that, when  $T$  is small and  $\hat{p}_{\text{RNML}} = p^\circ$ ,  $J_c$  computed for RNML exceeds the values of  $J_c$  corresponding to other IT criteria. At the same time, in all these cases, RNML yields the greatest improvement of  $J_c^{\text{ME}}$  in comparison with  $J_c$ .

As a final observation, we note for  $T = 1000$  that only SBC, KIC and  $\text{KIC}_c$  lead to large values of  $J$ ,  $J_c$ ,  $J^{\text{ME}}$  and  $J_c^{\text{ME}}$ . However, when  $T$  increases from 1000 to 3000, SBC is the only criterion which produces relatively large values of Itakura-Saito divergence.

## 2.7 Summary

In this chapter, we introduced the RNML criterion for VAR-order selection. In our theoretical analysis, we proved that the criterion is strongly consistent. The results reported for experiments with simulated data demonstrate its abilities in estimating properly the order when the sample size is small or moderate. It can be used as part of an algorithm which firstly estimates the order and then identifies the sparsity pattern of ISDM. This application will be further developed in the next chapter.

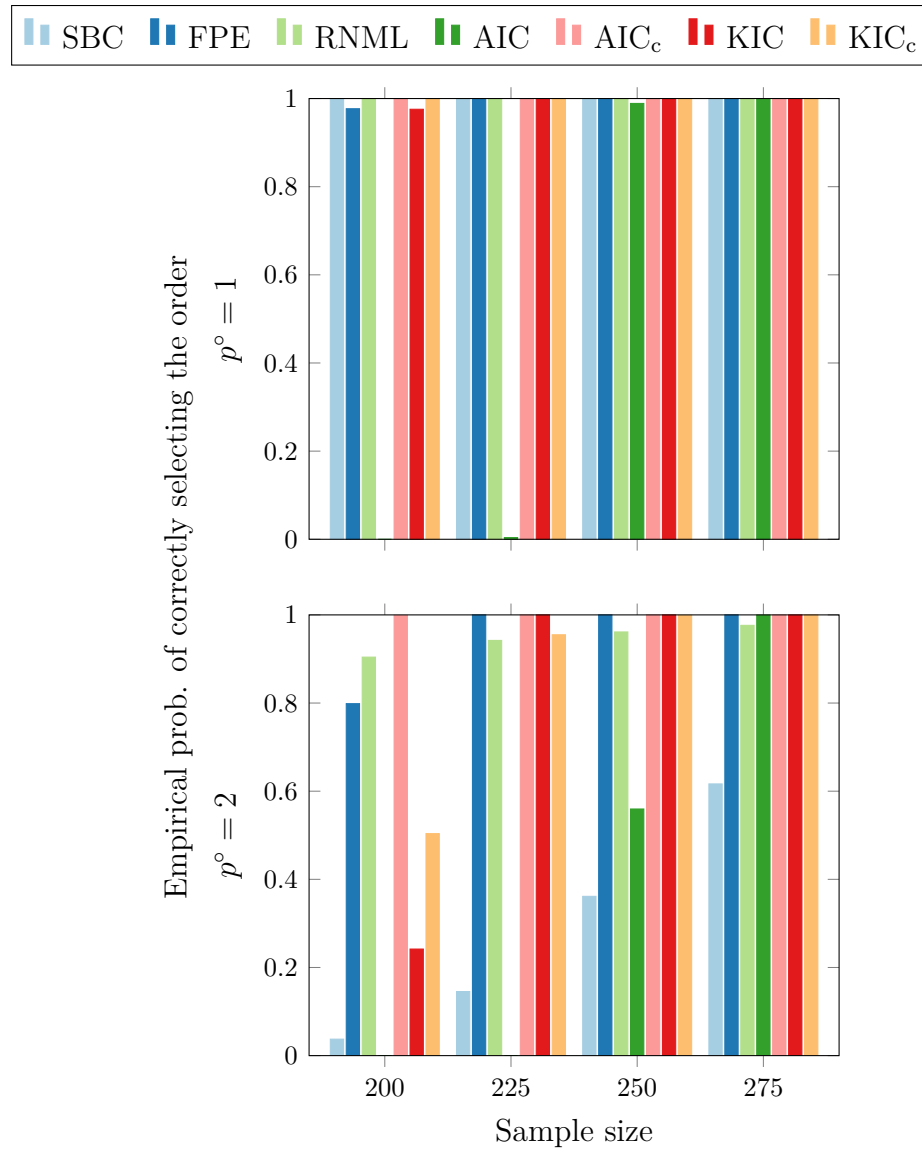


Figure 2.1: Example 1: Performance of various criteria in estimating the order of VAR-model. The value of "true" VAR-order is written on the vertical axes of the plots.

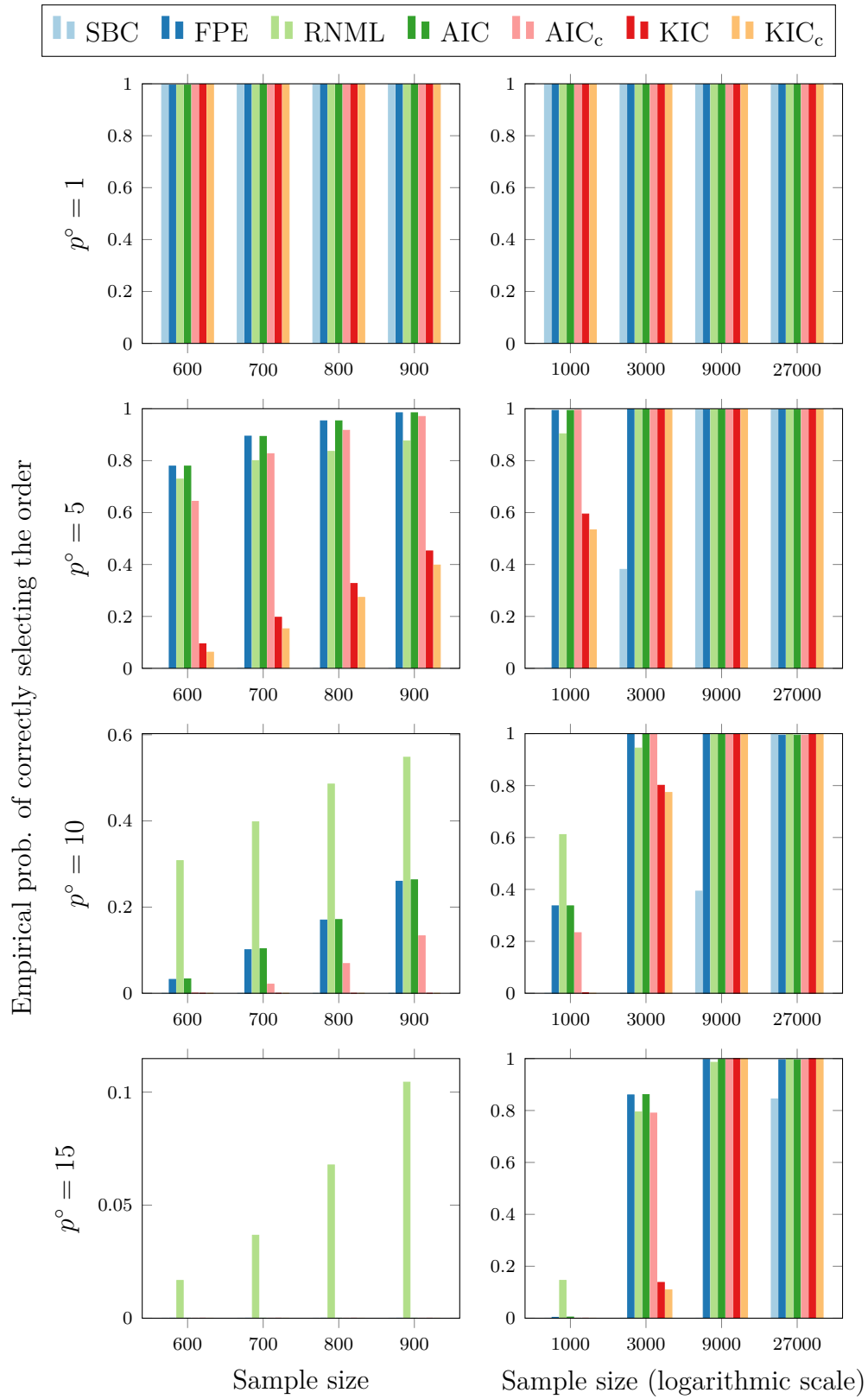


Figure 2.2: Example 2: Performance of various criteria in estimating the order of VAR-model. All graphical conventions are the same as Figure 2.1.

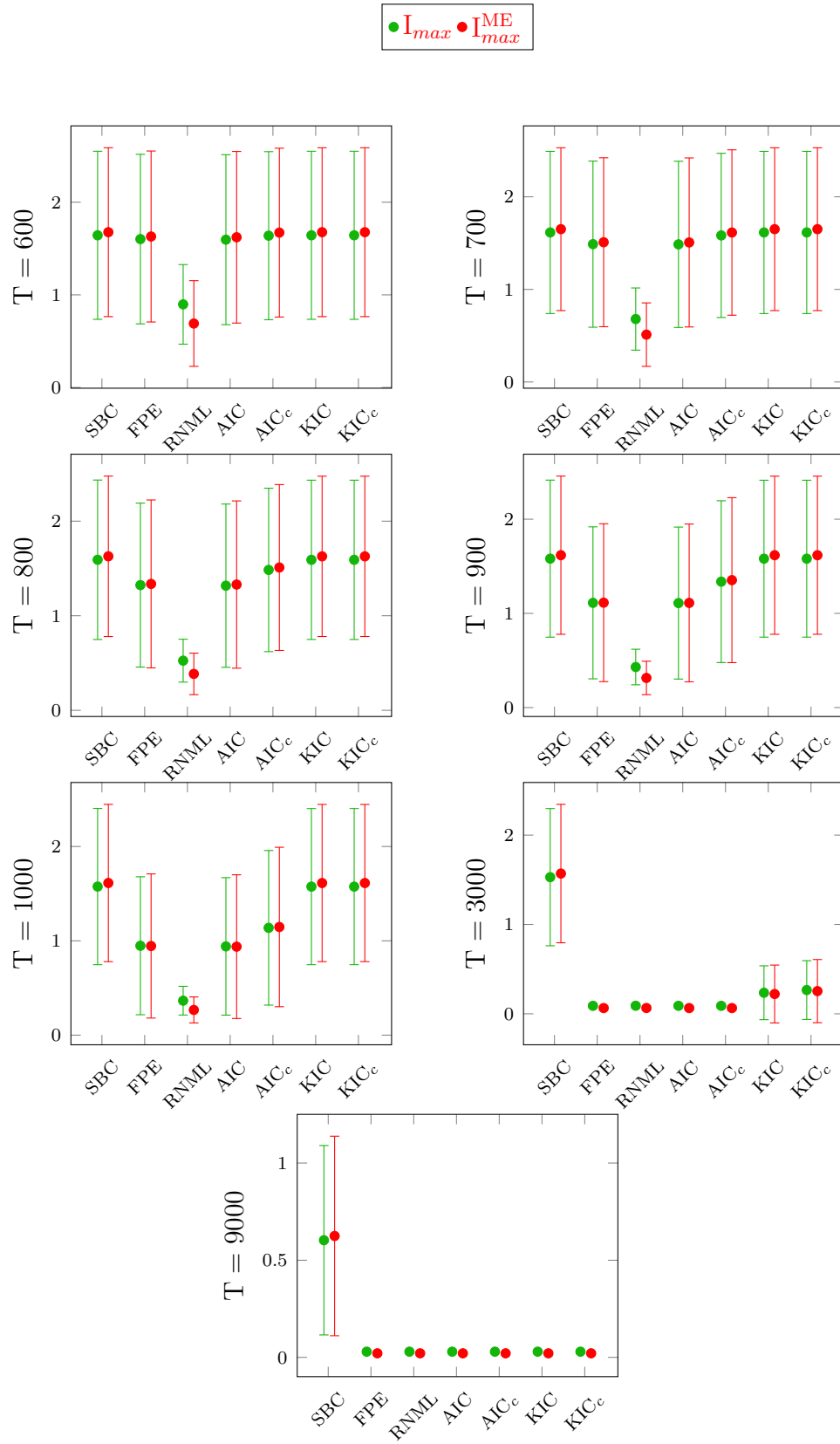


Figure 2.3: Example 2: Statistics for the maximum value of  $I$ -divergences computed on the  $\mathcal{G}$ -grid. For each IT criterion, we plot two error bars, each of which represents mean plus minus standard deviation. Note that  $p^\circ = 10$ . 35

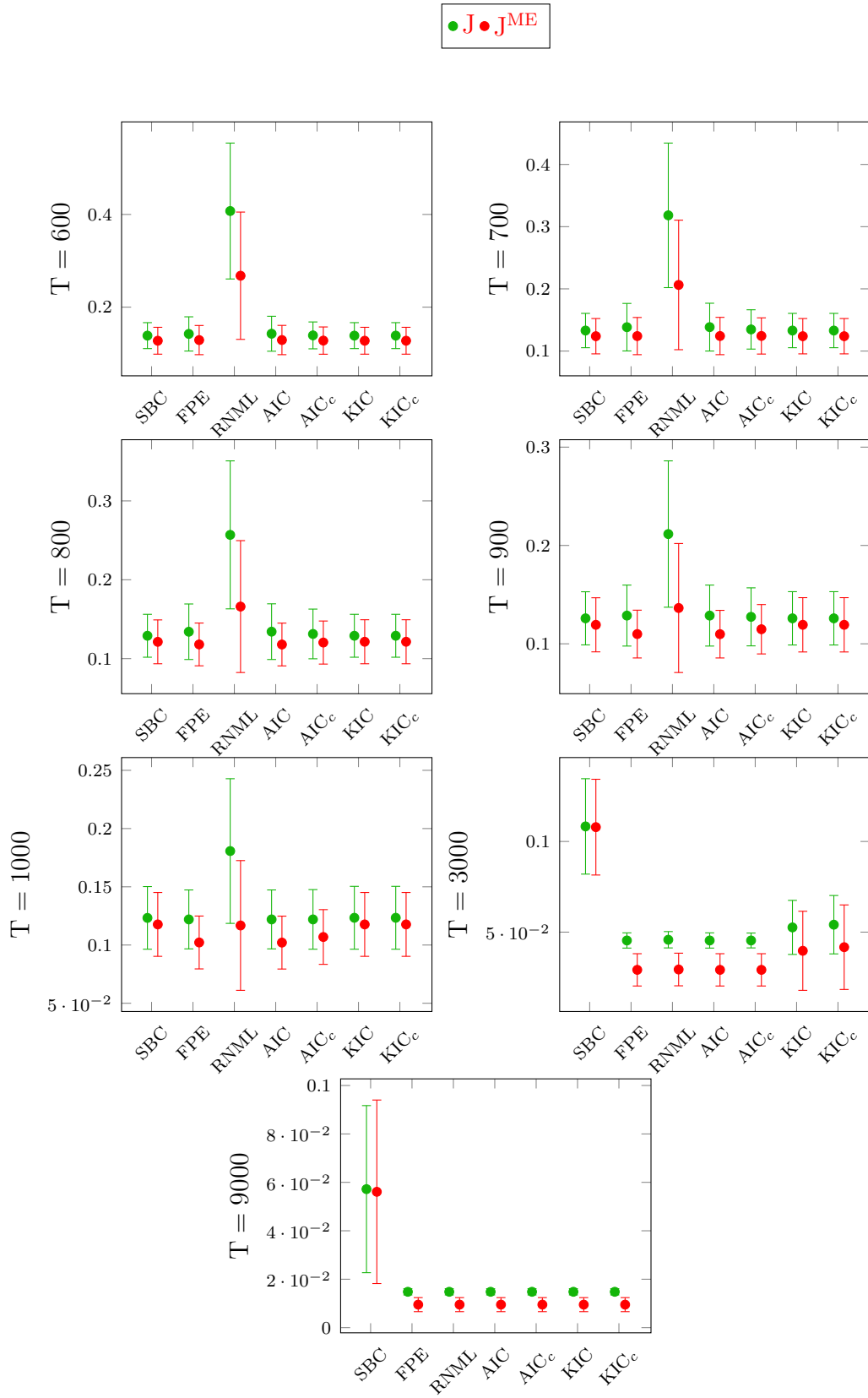


Figure 2.4: Example 2: Statistics of Itakura-Saito divergence. All graphical conventions are as in Figure 2.3.  $J$  replaces  $I_{\max}$  and  $J^{\text{ME}}$  replaces  $I_{\max}^{\text{ME}}$ .

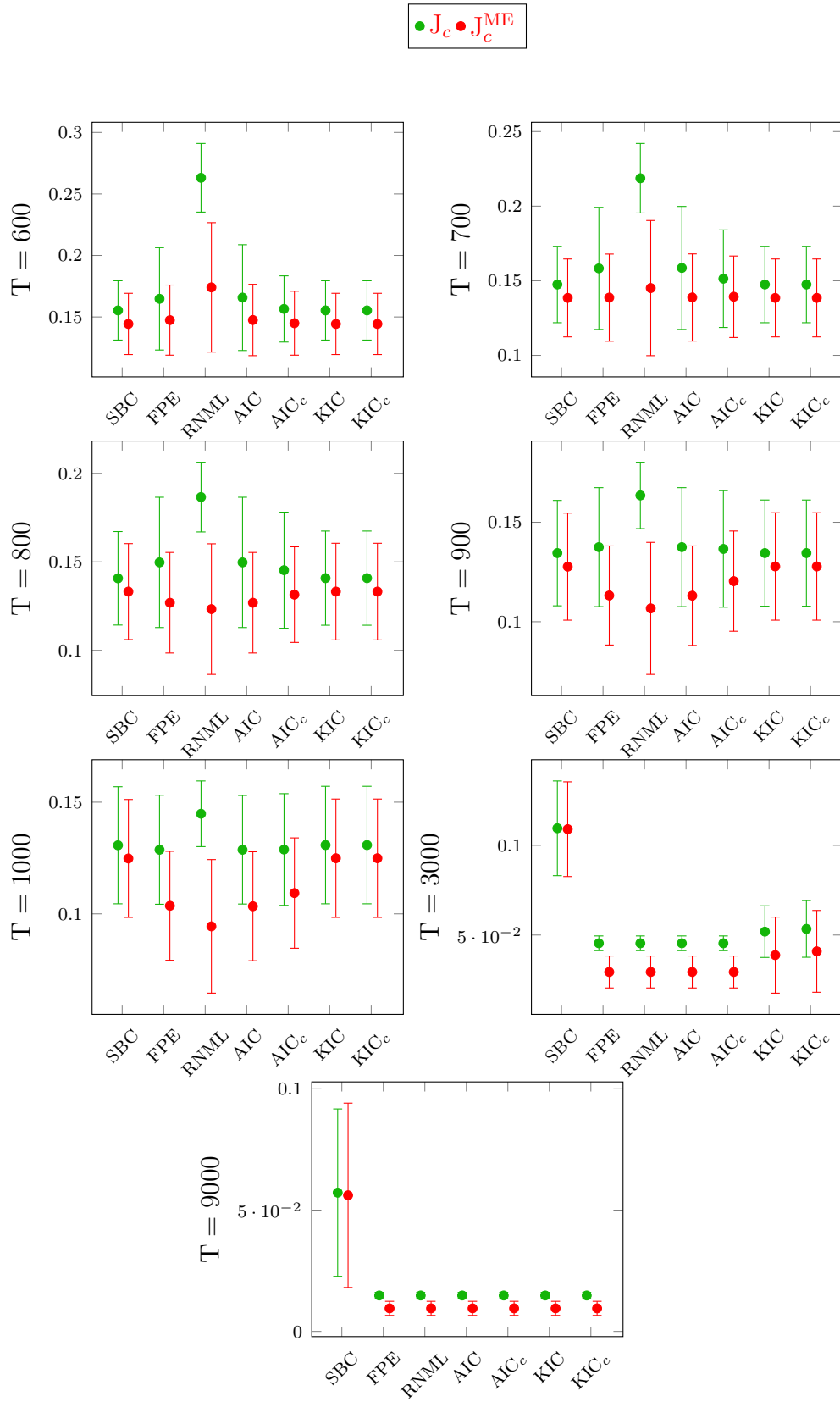


Figure 2.5: Example 2: Statistics computed only for the cases when  $\hat{p}_{\text{RNML}} = 10$ . All conventions are as in Figure 2.3:  $I_{\text{max}}$  is replaced by  $I_c$  and  $J_{\text{max}}^{\text{ME}}$  is replaced by  $J_c^{\text{ME}}$ .



## Chapter 3

# Efficient Algorithms for VAR Graph Estimation

### 3.1 Preliminaries

**Maximum entropy (ME) estimation** We consider the zero-mean, Gaussian, stationary stochastic vector process  $\{\mathbf{y}_t\}$  defined in (1.2). The order of the model ( $p^\circ$ ) as well as the sparsity pattern (SP) of the ISDM are assumed to be known. We have already mentioned in Section 2.6.2 that, for this  $K$ -variate model, an estimate of the ISDM can be obtained from the measurements  $\{\mathbf{y}_t\}_{t=1}^T$  by solving a convex optimization problem which maximizes the entropy rate subject to the following constraints: (i) the spectral density matrix matches the autocovariance matrices  $\hat{\mathbf{R}}(0), \dots, \hat{\mathbf{R}}(p^\circ)$  which are estimated from the data; (ii) the sparsity pattern of the ISDM is SP. The convex formulation is as follows (Songsiri, Dahl and Vandenberghe, 2010):

$$\begin{aligned} \min_{\mathbf{X}} \quad & \text{tr}(\mathcal{T}(\hat{\mathbf{R}})\mathbf{X}) - \log \det \mathbf{X}_{00} \\ \text{s.t.} \quad & \mathbf{X} \succeq 0 \\ & [\text{TR}_m(\mathbf{X})]_{ab} = 0, \quad \forall (a, b) \in \text{SP}, \quad m \in \{-p^\circ, \dots, p^\circ\} \end{aligned} \tag{3.1}$$

Note that

$$\hat{\mathbf{R}} = [\hat{\mathbf{R}}(0), \dots, \hat{\mathbf{R}}(p^\circ)] \tag{3.2}$$

and the operator  $\mathcal{T}(\cdot)$  constructs a symmetric block-Toeplitz matrix from its first block-row:

$$\mathcal{T}(\hat{\mathbf{R}}) = \begin{bmatrix} \hat{\mathbf{R}}(0) & \dots & \hat{\mathbf{R}}(p) \\ \vdots & \ddots & \vdots \\ \hat{\mathbf{R}}'(p) & \dots & \hat{\mathbf{R}}(0) \end{bmatrix}. \quad (3.3)$$

The variable  $\mathbf{X} \in \mathbb{R}^{K(p^\circ+1) \times K(p^\circ+1)}$  is symmetric and  $\mathbf{X}_{00} \succ 0$  is the  $K \times K$  subblock in the upper left corner of  $\mathbf{X}$ . The symbol  $\text{TR}_m(\cdot)$  stands for the block trace operator, with  $m$  the index of the block diagonal:

$$\text{TR}_m(\mathbf{X}) = \sum_{h=0}^{p^\circ-m} \mathbf{X}_{h,h+m}, \quad m \in \{0, \dots, p^\circ\}.$$

For negative indices, the relation  $\text{TR}_{-m}(\mathbf{X}) = [\text{TR}_m(\mathbf{X})]'$  holds. From (1.3), we get for all  $\omega \in (-\pi, \pi]$  that  $\hat{\mathbf{S}}_{\text{SP}}^{-1}(\omega) = \sum_{m=-\hat{p}}^{\hat{p}} \hat{\mathbf{Q}}_m e^{-j\omega m}$ , where  $\hat{\mathbf{Q}}_m = \text{TR}_m(\mathbf{X})$ . The coefficients of the model can be obtained from  $\mathbf{X}$  by spectral factorization (see (1.2) and (1.3)).

It is convenient to use CVX for finding the ME-estimate. According to Grant and Boyd (2010), CVX solves log det-problems with a successive approximation method which involves several semidefinite programming problems. Each such problem is initialized with the previous solution, hence needing less iterations than when solving from scratch. In our implementation, we prefer to employ CVX, even if the use of specialized algorithms could be better.

CVX is good especially because of its versatility. One can easily add or remove constraints to the optimization problem. With a special purpose algorithm like alternating direction method of multipliers (ADMM), the complexity would indeed scale better with size, but flexibility is difficult (Boyd et al. (2010)). Normally, we should test the solution with CVX and, when the optimization problem is well defined and becomes interesting for a large real application, then some more efficient algorithm can be implemented.

Because in practical situations, the order of the model and the sparsity pattern are *not* known, [Songsiri, Dahl and Vandenberghe \(2010\)](#) proposed to solve this problem for all pairs  $(p, \text{SP})$ , where  $p \leq p_{\max}$ . The value of  $p_{\max}$  is chosen by the user, and  $p_{\min}$  is set to zero. All possible SP's are considered. For each  $p \in \{p_{\min}, \dots, p_{\max}\}$ , and each SP in the list of sparsity patterns, a VAR-model is fitted to the data. The best pair  $(\hat{p}, \widehat{\text{SP}})$ , or equivalently the best model, is the one which minimizes an IT criterion. This method is called Full-Search.

As the criteria presented in Table 2.1 cannot be applied straightforwardly, [Songsiri, Dahl and Vandenberghe \(2010\)](#) modified three of them (AIC, AIC<sub>c</sub>, BIC-for which we use the name SBC) such that to be suitable for this problem. In Section 3.5, we will provide more information on how the criteria can be altered.

**Two-stage approach** Instead of the Full-Search from [Songsiri, Dahl and Vandenberghe \(2010\)](#), where each pair of  $p$  and SP are considered as the candidate models, we propose the two-stage approach, which is described in Algorithm 1. Remark that the stages for our estimation procedure are different from those in [Davis, Zang and Zheng \(2016\)](#). In our case, in Stage1, an IT criterion is employed to select the best order, say  $\hat{p}$  (see Chapter 2). The most important consequence is that, in Stage2, all the estimations are performed only for order  $\hat{p}$  and not for all the orders within the set  $\{1, \dots, p_{\max}\}$ .

As can be seen in Algorithm 1, at Step2.1, we find  $\hat{\mathbf{S}}_{\text{SP}}^{-1}(\omega)$  which has a given sparsity pattern (SP) and is nearest from  $\hat{\mathbf{S}}^{-1}(\omega)$ . We call this method Nearest-Sparse (NS). This is different from the approach used in [Songsiri, Dahl and Vandenberghe \(2010\)](#), where  $\hat{\mathbf{S}}_{\text{SP}}^{-1}(\omega)$  is obtained by using a convex optimization formulation of the ME-problem. In [\(Avventi, Lindquist and Wahlberg, 2013\)](#), NS was applied to get a good initialization for the ME-optimization problem. In Section 3.2, we show how the exact solution for NS can be computed efficiently. This allows us to use NS without applying the supplementary ME-optimization step. In contrast with [Avventi, Lindquist and Wahlberg \(2013\)](#), we do not

---

**Algorithm 1** Two-Stage Method

---

**Stage1** [Select  $\hat{p}$ ]:  
  **for all**  $p \in \{1, \dots, p_{\max}\}$  **do**  
    Fit  $\text{VAR}(p)$  to  $\mathbf{Y}$  and compute  $\text{ITC}(\mathbf{Y}; p)$ .  
  **end for**  
 $\hat{p} \leftarrow \arg \min_{1 \leq p \leq p_{\max}} \text{ITC}(\mathbf{Y}; p)$ . Save in  $\hat{\mathbf{S}}^{-1}(\omega)$  the ISDM of  $\text{VAR}(\hat{p})$ .  
**Stage2** [Select  $\widehat{\text{SP}}$ ]:  
  **for all** SP in the list of sparsity patterns for  $\text{VAR}(\hat{p})$ -models **do**  
    Step2.1: Find  $\hat{\mathbf{S}}_{\text{SP}}^{-1}(\omega)$  by minimizing  $\|\mathbf{E}(\omega)\|_{\infty} = \|\hat{\mathbf{S}}^{-1}(\omega) - \hat{\mathbf{S}}_{\text{SP}}^{-1}(\omega)\|_{\infty}$ ,  
    where  $\|\mathbf{E}(\omega)\|_{\infty}$  is the  $H_{\infty}$ -norm of the transfer function  $\mathbf{E}(\omega)$ , i.e. its  
    largest singular value for  $\omega \in (-\pi, \pi]$ .  
    Step2.2: Find the  $\text{VAR}(\hat{p}, \text{SP})$ -model by spectral factorization of  
     $\hat{\mathbf{S}}_{\text{SP}}^{-1}(\omega)$  and compute  $\text{ITC}(\mathbf{Y}; \hat{p}, \text{SP})$ .  
  **end for**  
 $\widehat{\text{SP}} \leftarrow \arg \min_{\text{SP}} \text{ITC}(\mathbf{Y}; \hat{p}, \text{SP})$ .

---

assume the order of the model to be given.

At Step2.2, we evaluate how suitable is  $\text{VAR}(\hat{p})$ , whose ISDM has the sparsity pattern SP, to model the data  $\mathbf{Y}$  by computing  $\text{ITC}(\mathbf{Y}; \hat{p}, \text{SP})$ . This implies the use of spectral factorization which is described in Section 3.2. Our approach has a solid justification and, more importantly, it does not involve a non-parametric estimator for the spectral density matrix.

In Stage2, the difficulty stems from the fact that all sparsity patterns for  $\text{VAR}(\hat{p})$ -models are evaluated. As the total number of SP's increases exponentially with  $K$ , it follows that the Exhaustive-Search can be applied only when  $K$  is small. In order to circumvent this difficulty, we propose two solutions: (i) Instead of enumerating from the very beginning all possible SP's, the list of candidates is generated dynamically; (ii) The number of SP's which compete is reduced by exploiting the peculiarities of ISDM. These solutions are presented in Section 3.3.1. In Section 3.3.2, we show how the combination of the two solutions leads to an entire family of algorithms.

## 3.2 Main algorithmic steps and their computational complexity

**Stage1** This is a classical problem, and for solving it we employ the ARFIT algorithm which guarantees that the complexity of computing the estimates for all considered orders is  $\mathcal{O}(TK^2p_{\max}^2)$  (see (Neumaier and Schneider, 2001, Section 3.3)).

**Stage2** At Step 2.1,  $\mathbf{E}(\omega)$  is a symmetric matrix polynomial, hence minimizing the  $H_\infty$ -norm amounts to minimizing the maximum over  $\omega \in (-\pi, \pi]$  of the absolute value of its eigenvalues. Denoting  $\mathbf{X}(\omega)$  the unknown sparse ISDM, the optimization problem becomes

$$\begin{aligned} \min_{\mathbf{X}} \quad & \lambda \\ \text{s.t.} \quad & -\lambda \mathbf{I} \preceq \hat{\mathbf{S}}^{-1}(\omega) - \mathbf{X}(\omega) \preceq \lambda \mathbf{I} \\ & \mathbf{X}(\omega) \succeq 0 \\ & [\mathbf{X}(\omega)]_{ab} = 0, \forall (a, b) \in \text{SP} \end{aligned} \tag{3.4}$$

The first constraint ensures the minimization of  $\|\mathbf{E}(\omega)\|_\infty$ . The second constraint imposes that the ISDM is positive semidefinite (symmetry is assumed by construction). Note that  $\hat{p}$  is fixed in (3.4) as it was already chosen in Stage 1. Since  $\mathbf{X}(\omega)$  is a trigonometric matrix polynomial, these constraints reduce to linear matrix inequalities as shown in (Dumitrescu, 2007, Section 3.10). The third constraint imposes the sparsity pattern SP and is a simple linear constraint. Overall, the optimization in (3.4) is a semidefinite programming (SDP) problem.

The problem (3.4) can be solved with CVX (Grant and Boyd, 2010) and the specialized library Pos3Poly (Şicleru and Dumitrescu, 2013), dedicated to optimization with positive polynomials. It is known that the complexity for such optimization problems is  $\text{DIM}^2 \times \text{CON}^2$ , where DIM is given by the size of the matrices involved and CON stands for the number of constraints (Dumitrescu, 2007). In our case,  $\text{DIM} = K(\hat{p} + 1)$  and  $\text{CON} = K^2(\hat{p} + 1)$ .

The Exhaustive-Search, which was mentioned in Section 3.1, assumes that all possible pairs  $(\hat{p}, \text{SP})$  are considered. With the convention that  $\overline{K} = K(K-1)/2$ , we have that the number of zeros which are located below the main diagonal in the sparsity pattern SP belongs to  $\{1, \dots, \overline{K}\}$ . Hence, in the case of Exhaustive-Search, the complexity of Step2.1 has the expression:

$$\sum_{q=1}^{\overline{K}} \binom{\overline{K}}{q} K^2(\hat{p}+1)^2 K^4(\hat{p}+1)^2 = \mathcal{O}(2^{K^2/2} K^6 \hat{p}^4). \quad (3.5)$$

For comparison, we also calculate the complexity for the situation when all models  $(p, \text{SP})$  with  $1 \leq p \leq p_{\max}$  are evaluated:

$$\sum_{p=1}^{p_{\max}} \sum_{q=1}^{\overline{K}} \binom{\overline{K}}{q} K^2(p+1)^2 K^4(p+1)^2 = \mathcal{O}(2^{K^2/2} K^6 p_{\max}^5). \quad (3.6)$$

At Step2.2, the spectral factorization of the positive matrix trigonometric polynomial  $\hat{\mathbf{S}}_{\text{SP}}^{-1}(\omega)$  can be computed by solving another SDP problem (see (Dumitrescu, 2007, App. B.5), McLean and Woerdeman (2002)). The estimated model is stable if  $\hat{\mathbf{S}}_{\text{SP}}^{-1}(\omega)$  is strictly positive definite for all  $\omega \in (-\pi, \pi]$ . A more cautious approach would be to replace the second constraint of (3.4) with  $\mathbf{X}(\omega) \succeq \epsilon \mathbf{I}$ , where  $\epsilon$  is a small positive constant.

Straightforward calculations lead to the conclusion that the complexity for Step2.2 is the same as for Step2.1. More importantly, according to (3.5) and (3.6), the computational burden becomes rapidly prohibitive when  $K$  raises. Below we propose some methods which can be applied when  $K$  is moderately large.

### 3.3 Faster algorithms

#### 3.3.1 Description

**Dynamic selection of  $\widehat{\text{SP}}$  (Greedy)** The complexity of Stage2 can be lowered as follows: We firstly run Step2.1 and Step2.2 for all SP's that contain exactly one zero below the main diagonal. Let  $\text{SP}^1$  be the sparsity pattern from

Full-Search Method : $\mathcal{O}(2^{K^2/2} K^6 p_{\max}^5)$		
Two-Stage Method		
Stage1 [Select $\hat{p}$ ]	Stage2 [Select $\widehat{\text{SP}}$ ]	
$\mathcal{O}(TK^2 p_{\max}^2)$	Exhaustive	$\mathcal{O}(2^{K^2/2} K^6 \hat{p}^4)$
	Greedy	$\mathcal{O}(K^{10} \hat{p}^4)$
	Static List	$\mathcal{O}(K^8 \hat{p}^4)$

Table 3.1: Computational complexity of various algorithms used to infer the conditional independence graph for VAR.

this family which minimizes a certain IT criterion we decide to apply. Then we perform Step2.1 and Step2.2 for all SP's of the form  $\text{SP}^1 \cup \{(u, v)\}$ , where  $1 \leq v < u \leq K$  and  $\{(u, v)\} \cap \text{SP}^1 = \emptyset$ . The one which minimizes the IT criterion is called  $\text{SP}^2$ . The procedure continues until we get  $\text{SP}^{\bar{K}}$ , which is a diagonal matrix. Remark that  $\text{SP}^1 \subset \dots \subset \text{SP}^{\bar{K}}$ . From these nested SP's, the winner is selected by comparing again the scores given by the IT criterion. When this algorithm is applied, the computational complexity for Stage2 is  $\sum_{q=1}^{\bar{K}} (\bar{K} - q + 1) K^6 (\hat{p} + 1)^4 = \mathcal{O}(K^{10} \hat{p}^4)$ .

**Selection of  $\widehat{\text{SP}}$  from a reduced list of candidates (Static List)** Greedy procedure outlined above is inspired from techniques applied in linear regression (see, for example, [Miller \(2002\)](#)). However, in the case of our problem, we can take advantage of the fact that an estimate  $\hat{\mathbf{S}}^{-1}(\omega)$  of ISDM can be easily computed after Stage1. Furthermore, this can be used to evaluate

$$M_{ab}[\hat{\mathbf{S}}^{-1}] = \max_{\omega \in (-\pi, \pi]} \hat{r}_{ab}[\hat{\mathbf{S}}^{-1}(\omega)], \text{ where} \quad (3.7)$$

$$\hat{r}_{ab}[\hat{\mathbf{S}}^{-1}(\omega)] = \frac{\left| \left[ \hat{\mathbf{S}}^{-1}(\omega) \right]_{ab} \right|}{\sqrt{\left[ \hat{\mathbf{S}}^{-1}(\omega) \right]_{aa} \left[ \hat{\mathbf{S}}^{-1}(\omega) \right]_{bb}}}, \quad (3.8)$$

for all pairs  $(a, b)$  with property  $1 \leq b < a \leq \bar{K}$  ([Dahlhaus, 2000](#)). The larger is  $M_{ab}[\hat{\mathbf{S}}^{-1}]$ , the higher is the chance that the marginal time series  $\mathbf{y}_a$  and  $\mathbf{y}_b$  are conditionally correlated ([Davis, Zang and Zheng, 2016](#)). This is why we

order the  $M$ -values increasingly and then define, based on this order, the list  $\text{SP}^1 \subset \dots \subset \text{SP}^{\bar{K}}$  of SP's which will compete in Stage2. If  $M_{a_1 b_1}[\hat{\mathbf{S}}^{-1}] \leq \dots \leq M_{a_{\bar{K}} b_{\bar{K}}}[\hat{\mathbf{S}}^{-1}]$ , then  $\text{SP}^i = \bigcup_{q=1}^i \{(a_q, b_q)\}$ . A possible modification might be to replace the expression of  $M_{ab}[\hat{\mathbf{S}}^{-1}]$  in (3.7) with  $\sum_{\omega \in \mathcal{G}} \left( \hat{r}_{ab}[\hat{\mathbf{S}}^{-1}(\omega)] \right)^2$ , where  $\mathcal{G}$  is a grid on  $[0, \pi]$  (see Section 2.6.2).

It is easy to observe that the cost of evaluating (3.8), for all pairs  $(a, b)$  which are needed, is  $\mathcal{O}(K^2 \hat{p}^4)$ . Considering the cost of all operations in Stage2, it follows that the computational complexity is  $\mathcal{O}(K^8 \hat{p}^4)$ . For ease of comparison, this result is shown in Table 3.1, along with the computational complexities of all estimation methods we discussed so far.

A scheme relying on a reduced list of candidates was already used in the literature (see [Avventi, Lindquist and Wahlberg \(2013\)](#) and the references therein). The main difference between our proposal and previous work stems from the estimation of ISDM. In our case,  $\hat{\mathbf{S}}^{-1}(\omega)$  is produced at the end of Stage1, whereas the other authors firstly calculate  $\hat{\mathbf{S}}_{\text{ker}}(\omega)$  by using a non-parametric method with some kernel function  $\text{ker}$  and then compute its inverse for all frequencies on a grid. The fact that we do not need to get the inverse for any matrix is a computational advantage.

Under the null hypothesis that  $r_{ab}[\mathbf{S}^{-1}(\omega)]$  is zero for all frequencies  $\omega \in (-\pi, \pi]$ , the real and imaginary part of  $\sqrt{T} \left( \hat{r}_{ab}[\hat{\mathbf{S}}_{\text{ker}}^{-1}(\omega)] - r_{ab}[\mathbf{S}^{-1}(\omega)] \right)$  are asymptotically independent and Gaussian distributed with zero-mean and variance which depends on  $\text{ker}$  ([Dahlhaus, Eichler and Sandkühler, 1997](#)). The proof of this result is based on  $\delta$ -method from [Lutkepöhl \(2005\)](#). The same method can be applied for finding the limiting distributions for real and imaginary part of  $[\hat{\mathbf{S}}^{-1}(\omega)]_{ab}$ , which we use in (3.8). Below we present the distributional properties for VAR(1), but they can be easily extended to VAR-models of any order. The key finding is that the real and imaginary part of  $\sqrt{T-1} \left( [\hat{\mathbf{S}}^{-1}(\omega)]_{ab} - [\mathbf{S}^{-1}(\omega)]_{ab} \right)$  are asymptotically Gaussian distributed with zero-mean and variance independent of  $T$ . The closed-form expression of the variance is generally given by a

complicated formula, but this is unimportant for us because we do not select the positions of the zero entries of ISDM by comparing  $M_{ab}[\hat{\mathbf{S}}^{-1}]$  with a threshold. Because we assume that the order of VAR equals one, the imaginary part of ISDM is given by

$$\Im \{ \mathbf{S}^{-1}(\omega) \} = \sin(\omega) [ -\mathbf{A}'_1 \boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}^{-1} \mathbf{A}_1 ]. \quad (3.9)$$

We define  $g \left( (\mathbf{A}'_1)^V \right) = [\Im \{ \mathbf{S}^{-1}(\omega) \}]^V$ , for a fixed value of  $\omega$ . The vec-operator  $(\cdot)^V$  denotes the vector obtained by stacking the columns of the matrix in the argument on top of one another. We want to get the asymptotic distribution of  $g \left( (\hat{\mathbf{A}}'_1)^V \right)$ , which is obtained from (3.9) by replacing  $\mathbf{A}_1$  with  $\hat{\mathbf{A}}_1$ . To this end, we employ a well-known result (Reinsel, 1993, p. 81):

$$N^{1/2} \left[ \left( (\hat{\mathbf{A}}'_1)^V - (\mathbf{A}'_1)^V \right) \right] \xrightarrow{D} \mathcal{N} \left( \mathbf{0}, \boldsymbol{\Sigma} \otimes \mathbf{R}(0)^{-1} \right),$$

where  $N = T - 1$ . The use of  $\delta$ -method (Lutkepöhl, 2005, p. 693) leads to

$$N^{1/2} \left[ g \left( (\hat{\mathbf{A}}'_1)^V \right) - g \left( (\mathbf{A}'_1)^V \right) \right] \xrightarrow{D} \mathcal{N} \left( \mathbf{0}, \frac{\partial g \left( (\mathbf{A}'_1)^V \right)}{\partial \left[ (\mathbf{A}'_1)^V \right]'} (\boldsymbol{\Sigma} \otimes \mathbf{R}(0)^{-1}) \frac{\partial g' \left( (\mathbf{A}'_1)^V \right)}{\partial (\mathbf{A}'_1)^V} \right).$$

Now we focus on the expression of the covariance matrix which appears in the equation above. After some algebra, we get

$$\frac{\partial \left( \mathbf{A}'_1 \boldsymbol{\Sigma}^{-1} \right)^V}{\partial \left[ (\mathbf{A}'_1)^V \right]'} = \boldsymbol{\Sigma}^{-1} \otimes \mathbf{I}, \quad (3.10)$$

$$\frac{\partial \left( \boldsymbol{\Sigma}^{-1} \mathbf{A}_1 \right)^V}{\partial \left[ (\mathbf{A}'_1)^V \right]'} = (\mathbf{I} \otimes \boldsymbol{\Sigma}^{-1}) \boldsymbol{\Xi}, \quad (3.11)$$

where  $\boldsymbol{\Xi}$  denotes the commutation matrix which has the property that  $(\mathbf{M}')^V = \boldsymbol{\Xi} \mathbf{M}^V$  for any  $K^2 \times K^2$  matrix  $\mathbf{M}$ .

In order to simplify the calculations, we further assume that  $\Sigma = \mathbf{I}$ . So,

$$\begin{aligned} & \frac{\partial g \left( (\mathbf{A}'_1)^V \right)}{\partial \left[ (\mathbf{A}'_1)^V \right]'} (\Sigma \otimes \mathbf{R}(0)^{-1}) \frac{\partial g' \left( (\mathbf{A}'_1)^V \right)}{\partial (\mathbf{A}'_1)^V} \\ &= \sin^2(\omega)(-\mathbf{I} + \Xi)(\mathbf{I} \otimes \mathbf{R}(0)^{-1})(-\mathbf{I} + \Xi) \\ &= \sin^2(\omega)(\mathbf{I} - \Xi) \left[ (\mathbf{I} \otimes \mathbf{R}(0)^{-1}) + (\mathbf{R}(0)^{-1} \otimes \mathbf{I}) \right]. \end{aligned}$$

Now we do the same type of analysis for the  $\Re\{\mathbf{S}^{-1}(\omega)\}$ . Firstly we note that

$$\Re\{\mathbf{S}^{-1}(\omega)\} = -\cos(\omega)[\mathbf{A}'_1 + \mathbf{A}_1] + \mathbf{A}'_1\mathbf{A}_1 + \mathbf{I}.$$

We take  $g \left( (\mathbf{A}'_1)^V \right) = [\Re\{\mathbf{S}^{-1}(\omega)\}]^V$  (for a fixed value of  $\omega$ ), and compute the asymptotic covariance matrix of  $N^{1/2} \left[ g \left( (\hat{\mathbf{A}}'_1)^V \right) - g \left( (\mathbf{A}'_1)^V \right) \right]$  by using the  $\delta$ -method. Note that

$$\frac{\partial (\mathbf{A}'_1\mathbf{A})^V}{\partial \left[ (\mathbf{A}'_1)^V \right]'} = (\mathbf{I} \otimes \mathbf{A}'_1)\Xi + (\mathbf{A}'_1 \otimes \mathbf{I}).$$

This result along with (3.10)-(3.11) lead to the following expression of the covariance matrix:

$$\begin{aligned} & (\mathbf{I} + \Xi) [-\cos(\omega)\mathbf{I} + (\mathbf{A}'_1 \otimes \mathbf{I})] (\mathbf{I} \otimes \mathbf{R}(0)^{-1}) [-\cos(\omega)\mathbf{I} + (\mathbf{A}_1 \otimes \mathbf{I})] (\mathbf{I} + \Xi) \\ &= (\mathbf{I} + \Xi) [-\cos(\omega)(\mathbf{I} \otimes \mathbf{R}(0)^{-1}) + (\mathbf{A}'_1 \otimes \mathbf{R}(0)^{-1})] [-\cos(\omega)\mathbf{I} + (\mathbf{A}_1 \otimes \mathbf{I})] (\mathbf{I} + \Xi) \\ &= (\mathbf{I} + \Xi) [\cos^2(\omega)(\mathbf{I} \otimes \mathbf{R}(0)^{-1}) - \cos(\omega)(\mathbf{A}_1 \otimes \mathbf{R}(0)^{-1}) \\ &\quad - \cos(\omega)(\mathbf{A}'_1 \otimes \mathbf{R}(0)^{-1}) + (\mathbf{A}'_1\mathbf{A}_1 \otimes \mathbf{R}(0)^{-1})]. \end{aligned}$$

### 3.3.2 Further refinements

**Refinement I** The strengths of Greedy and List can be combined by employing Algorithm 2 in Stage2. We make the convention that  $\text{NS}(\hat{\mathbf{S}}^{-1}, \text{SP})$  stands for the ISDM with sparsity pattern SP which is nearest from  $\hat{\mathbf{S}}^{-1}(\omega)$ . Even if not written explicitly in the algorithm, note that the evaluation of  $\text{ITC}[\mathbf{Y}; \hat{p}, \text{SP}]$  involves the spectral factorization of  $\text{NS}(\hat{\mathbf{S}}^{-1}, \text{SP})$ .

---

**Algorithm 2** Select  $\widehat{\text{SP}}$  Procedure
 

---

```

 $\text{SP}^0 \leftarrow \emptyset; \widehat{\text{SP}} \leftarrow \text{SP}^0; \hat{\mathbf{S}}_0^{-1} \leftarrow \hat{\mathbf{S}}^{-1}; \widehat{\text{ITC}} \leftarrow \text{ITC}[\mathbf{Y}; \hat{p}]$ 
for all  $i \in \{1, \dots, \overline{K}\}$  do
    Compute :  $M_{ab}[\hat{\mathbf{S}}_{i-1}^{-1}]$  for  $(a, b) \in (\text{SP}^{\overline{K}} \setminus \text{SP}^{i-1}); N_{\text{tot}}^i \leftarrow |\text{SP}^{\overline{K}} \setminus \text{SP}^{i-1}|$ 
    Sort:  $M_{a_1 b_1}[\hat{\mathbf{S}}_{i-1}^{-1}] \leq \dots \leq M_{a_{N_0^i} b_{N_0^i}}[\hat{\mathbf{S}}_{i-1}^{-1}] \leq \dots \leq M_{a_{N_{\text{tot}}^i} b_{N_{\text{tot}}^i}}[\hat{\mathbf{S}}_{i-1}^{-1}]$ 
    for all  $(a, b) \in \{(a_1, b_1), \dots, (a_{N_0^i}, b_{N_0^i})\}$  do
         $\text{SP}^{\text{temp}} \leftarrow \text{SP}^{i-1} \cup \{(a, b)\}; \hat{\mathbf{S}}_{\text{temp}}^{-1} \leftarrow \text{NS}(\hat{\mathbf{S}}^{-1}, \text{SP}^{\text{temp}})$ 
        if  $\text{ITC}[\mathbf{Y}; \hat{p}, \text{SP}^{\text{temp}}] < \widehat{\text{ITC}}$  then
             $\text{SP}^i \leftarrow \text{SP}^{\text{temp}}; \widehat{\text{SP}} \leftarrow \text{SP}^i; \hat{\mathbf{S}}_i^{-1} \leftarrow \hat{\mathbf{S}}_{\text{temp}}^{-1}; \widehat{\text{ITC}} \leftarrow \text{ITC}[\mathbf{Y}; \hat{p}, \text{SP}^{\text{temp}}]$ 
        end if
    end for
end for
return  $\widehat{\text{SP}}$ 
    
```

---

Algorithm 2 is initialized with the SP which does not contain any zero. At each step ( $i$ ), a zero is inserted below the main diagonal of SP. The location of the inserted zero is chosen from  $N_0^i$  possible positions by using an IT criterion. The parameter  $N_0^i$  is selected by the user. The most complex algorithm in this family corresponds to the case when  $N_0^i = N_{\text{tot}}^i$  for all  $i$ , and is the same as the Greedy procedure which was introduced above. When  $N_0^i = 1$  for all  $i$ , the algorithm has the lowest computational complexity and we call it List. It is worth pointing out that, in Algorithm 2, the list of candidates is not computed only in the beginning of Stage2, but it is renewed after each insertion of a zero below the main diagonal of ISDM. This makes List algorithm to be different from the Static List, which was discussed in the previous section. However, the computational complexity is about the same for both List and Static List.

**Refinement II** The results on complexity reported in Table 3.1 are based on the assumption that, in Stage2,  $\hat{\mathbf{S}}_{\text{SP}}^{-1}(\omega)$  is taken to be the nearest from  $\hat{\mathbf{S}}^{-1}(\omega)$ , in the class of ISDM that satisfy all the required constraints. A more advanced option is to find  $\hat{\mathbf{S}}_{\text{SP}}^{-1}(\omega)$  as the ME-solution corresponding to the sparsity pattern SP (see (3.1)). This assumes that the autocovariance matrices  $\hat{\mathbf{R}}(0), \dots, \hat{\mathbf{R}}(\hat{p})$  should be estimated from the data. Recall that  $\hat{p}$  was chosen in Stage1.

**Refinement III** Under the hypothesis that, at each step ( $i$ ), exactly one zero is inserted below the main diagonal of SP, the fastest algorithm is List (see again Refinement I). However, a Fast List algorithm can be obtained by allowing more zeros to be inserted at each step. For example, if a group of  $G_{N_0} > 1$  zeros is added to SP at each step  $i$ , then the total number of models which compete will be  $\lceil \bar{K}/G_0 \rceil + 1$ . As usual,  $\lceil \cdot \rceil$  is the smallest integer greater than or equal to the real number in the argument. The reduction in the number of calculations is evident, but the accuracy of estimation might deteriorate as well. We will investigate this aspect in Section 3.6.

### 3.4 Comparison with other methods based on convex optimization

The comparison is focused on three aspects: (i) if a non-parametric estimator is needed; (ii) which optimization problem should be solved; (iii) how many candidates are evaluated by using IT criteria. The outcome of this analysis is presented in Table 3.2. This can be used in conjunction with the results from previous sections for deriving the computational complexity of each method. For instance, it is obvious that the method we propose is much faster than the one in Songsiri, Dahl and Vandenberghe (2010).

However, the comparison with Songsiri and Vandenberghe (2010) is not straightforward because the method introduced in this reference involves the following optimization problem:

$$\begin{aligned} \min_{\mathbf{X}} \quad & \text{tr}(\mathcal{T}(\mathbf{R})\mathbf{X}) - \log \det \mathbf{X}_{00} + \lambda g(\mathbf{X}) \\ \text{s.t.} \quad & \mathbf{X} \succeq 0 \end{aligned} \tag{3.12}$$

Remark that the notation is the same as in (3.1). The value of  $\lambda$  as well as the sparsity promoter  $g(\cdot)$  are given by the user. An estimate of ISDM is obtained, and is further used to get an estimate of PSC. Even if some of the entries of the estimated PSC have small magnitudes, they are not exactly zero. In order to decide which of them are zeros, Songsiri and Vandenberghe (2010) compare the PSC entries with a threshold  $\text{Th}$ : All values smaller than  $\text{Th}$  lead to zeros in the sparsity pattern  $\text{SP}_\lambda$ . The final ISDM estimate corresponding to  $\lambda$  is the solution of the optimization problem in (3.1), for which  $\text{SP} = \text{SP}_\lambda$ . As it does not exist a particular value of  $\lambda$  which is best for all data sets, a method for generating a sequence of  $\lambda$ -values is proposed in Songsiri and Vandenberghe (2010). For each value in this sequence, the problems (3.1) and (3.12) should be solved. Hence, the complexity of the algorithm depends on the length of the sequence.

In [Avventi, Lindquist and Wahlberg \(2013\)](#), the model is assumed to be

$$\mathbf{y}_t = \sum_{m=1}^{p^\circ} \mathbf{A}_m \mathbf{y}_{t-m} + \sum_{m=0}^{p^\circ} b_m \mathbf{u}_{t-m}, \quad t = 1, 2, \dots, T, \quad (3.13)$$

where  $\{\mathbf{A}_m\}$  and  $\{\mathbf{u}_t\}$  have the same significance as in (1.2). Additionally,  $b_0, \dots, b_m$  are real numbers with property that  $b_0 = 1$ . All the poles and the roots of the model are assumed to be located inside the unit disc. As explained in [Avventi, Lindquist and Wahlberg \(2013\)](#), any VARMA model can be written like in (3.13). Remark that the value of  $p^\circ$  is assumed to be known. This makes difficult the comparison with our method in which the estimation of the order plays a central role. Moreover, the use of RNML-formula derived in Prop. 2.3.1 is restricted to VAR-models.

In the next section, we show how the IT criteria outlined in Table 2.1 can be altered such that to be used in Stage 2 of Algorithm 1.

### 3.5 Modified IT criteria

We have already pointed out in Chapter 2 that the IT criteria can be expressed as the sum of a goodness-of-fit (GOF) term and a penalty. They are derived on various grounds, but their expressions cannot be obtained easily for the problem we investigate. Due to this reason, we resort to the methodology applied previously to VAR topology selection problems, where the criteria originally proposed for model order selection have been modified such that to be employed for finding the best sparsity pattern. As the GOF term is obtained straightforwardly by fitting the model to the data, the difficult part is the alteration of the penalty term. Based on the observation that all the penalty terms of the criteria for model order selection involve the number of parameters of the model, [Songsiri, Dahl and Vandenberghe \(2010\)](#) proposed to replace it with the *effective number of parameters*:

$$N_{\text{ef}} = \frac{K(K+1)}{2} - N_0 + p(K^2 - 2N_0). \quad (3.14)$$

Methods from previous literature		
Ref.	Description	Computations
(Songsiri, Dahl and Vandenberghe, 2010)	Input: $p_{\max}$	NS: No
	ITC: SBC, AIC, AIC <sub>c</sub>	ME: Eq. (3.1) RME: No
	Enumeration: Full-Search	#Candidates: $2^{\bar{K}} \times p_{\max}$
(Songsiri and Vandenberghe, 2010)	Input: $p_{\max}$ , $\#\lambda$ , Th	NS: No
	ITC: SBC, AIC, AIC <sub>c</sub>	ME: Eq. (3.1) RME: Eq. (3.12)
	Enumeration: Each $\lambda$ leads to a candidate	#Candidates: $\#\lambda \times p_{\max}$
Newly proposed method		
	Description	Computations
	Input: $p_{\max}$ , $N_0^i$ or $G_0$ , option for NS/ME	NS: Eq. (3.4) [if chosen]
	ITC: RNML and others	ME: Eq. (3.1) [if chosen] RME: No
	Enumeration: Greedy	#Candidates: $\approx \bar{K}^2/2$
	Enumeration: Combination Greedy/List	#Candidates: $\approx \bar{K} N_0^i$
	Enumeration: List	#Candidates: $\approx \bar{K}$
	Enumeration: Fast List	#Candidates: $\approx \bar{K}/G_0$

Table 3.2: Comparison of parametric methods for inferring graphical models via convex optimization: For the optimization problems we use the acronyms NS (Nearest Sparse), ME (Maximum Entropy), RME (Regularized Maximum Entropy), and indicate the equations where the problems are explicitly written.

Note that  $N_0$  is the number of zeros in the lower triangular part of the ISDM. The expression above can be obtained straightforwardly by counting the number of non-zero entries in the  $\text{SP}^i$  matrices produced by Algorithm 2.

The formula in (3.14) was used in Songsiri, Dahl and Vandenberghe (2010) in order to modify three celebrated criteria mentioned in Section 3.1. Based on the empirical evidence from Songsiri, Dahl and Vandenberghe (2010), SBC is ranked best when the sample size is large, whereas AIC<sub>c</sub> works better for small sample sizes. This makes us to employ these two criteria in our experiments.

Their expressions are (Songsiri, Dahl and Vandenberghe, 2010):

$$\begin{aligned} \text{SBC} &= T \log \det \hat{\Sigma} + N_{\text{ef}} \log T, \\ \text{AIC}_c &= T \log \det \hat{\Sigma} + \frac{2N_{\text{ef}}T}{T - N_{\text{ef}} - 1}, \end{aligned} \quad (3.15)$$

where  $\hat{\Sigma}$  is the error covariance matrix.

Additionally, we show how the asymptotic analysis helps to alter log FPE in order to be used in Stage2 of the estimation procedure. According to Lutkepöhl (2005, p. 148), log FPE and AIC reduce to the same criterion when  $T \rightarrow \infty$ . Moreover, in AIC, the number of parameters per component is given by  $Kp$ . As we aim to maintain the relationship between FPE and AIC after alteration, we substitute the term  $Kp$  with  $\eta = N_{\text{ef}}/K$ , in the formula of log FPE (see Table 2.1). Hence, we get:

$$\log \text{FPE} = \log \det \hat{\Sigma} + K \log \frac{T + \eta}{T - \eta}$$

Similarly, the RNML-criterion is modified by putting  $\eta$  instead of  $Kp$  in (2.3)-(2.4) and  $N_{\text{ef}}$  instead of  $K^2p$  in (2.5)-(2.6):

$$\begin{aligned} \text{RNML} &= \frac{T - \eta - K + 1}{2} \log \det \hat{\Sigma} + \frac{N_{\text{ef}}}{2} \log \text{tr} \left( \hat{\mathbf{R}}(0) - \hat{\Sigma} \right) \\ &\quad - \log \Gamma_K \left( \frac{T - \eta}{2} \right) - \log \Gamma \left( \frac{N_{\text{ef}}}{2} \right), \end{aligned} \quad (3.16)$$

where  $\hat{\mathbf{R}}(0)$  has the same significance as in (3.1).

## 3.6 Simulated data

**Evaluation of the novel algorithms** The data are simulated according to the procedure described in Section 2.6.2. After the order  $\hat{p}$  of the model is selected with an IT criterion, an estimate of ISDM can be obtained by using either NS or ME. Because we want to evaluate the accuracy of this estimation, we assume that SP is known. More precisely, we generate data as in Section 2.6.2, but only for  $p^\circ = 10$ . As the number of runs is  $N_r = 10$ , it follows that the number of  $K$ -variate time series is  $N_{\text{SP}} \times N_r = 100$ . Based on the results shown in Figure

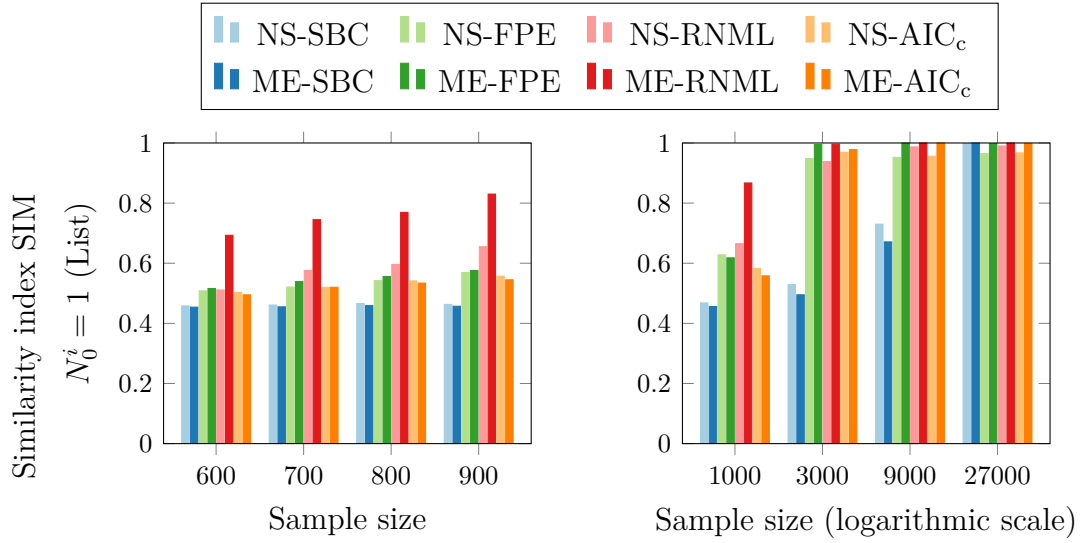


Figure 3.1: Performance of estimation methods evaluated by the SIM index defined in (3.17): The closer is SIM to value one, the better is the estimation. In the legend, we indicate for each method the optimization used in Stage2 (NS or ME) as well as the IT criteria employed in Stage1 and Stage2. Note that  $K = 5$  and  $p^\circ = 10$ .

2.2, we include in competition only four IT criteria: SBC, FPE, RNML and  $AIC_c$ . In experiments, we use Algorithm 2, with the following options for  $N_0^i$ ,  $i \in \{1, \dots, \bar{K}\}$ : (i)  $N_0^i = 1$  (List); (ii)  $N_0^i = \min(4, N_{\text{tot}}^i)$  and (iii)  $N_0^i = N_{\text{tot}}^i$  (Greedy). Interestingly enough, the performance is almost the same for all three options, which makes us to show in Figure 3.1 only the results for List, which has the lowest computational complexity.

The performance of an estimation method, met, in the  $r$ -th trial is evaluated by computing  $\mathcal{E}_{\text{SP},r}^{\text{met}}$ , which is the number of positions below the main diagonal where the estimated SP is different from the true SP. Remark that  $\mathcal{E}_{\text{SP},r}^{\text{met}} \in \{0, \dots, \bar{K}\}$ . This leads naturally to the definition of the following similarity index:

$$\text{SIM}_{\text{met}} = 1 - \frac{\sum_{i=1}^{N_{\text{SP}}} \sum_{r=1}^{N_r} \mathcal{E}_{\text{SP},r}^{\text{met}}}{N_{\text{SP}} \times N_r \times \bar{K}}. \quad (3.17)$$

Remark in Figure 3.1 that RNML is the best amongst all the tested criteria when the sample sizes are small or moderate. This is perfectly in line with what we already observed in Figure 2.2. Another interesting peculiarity of RNML

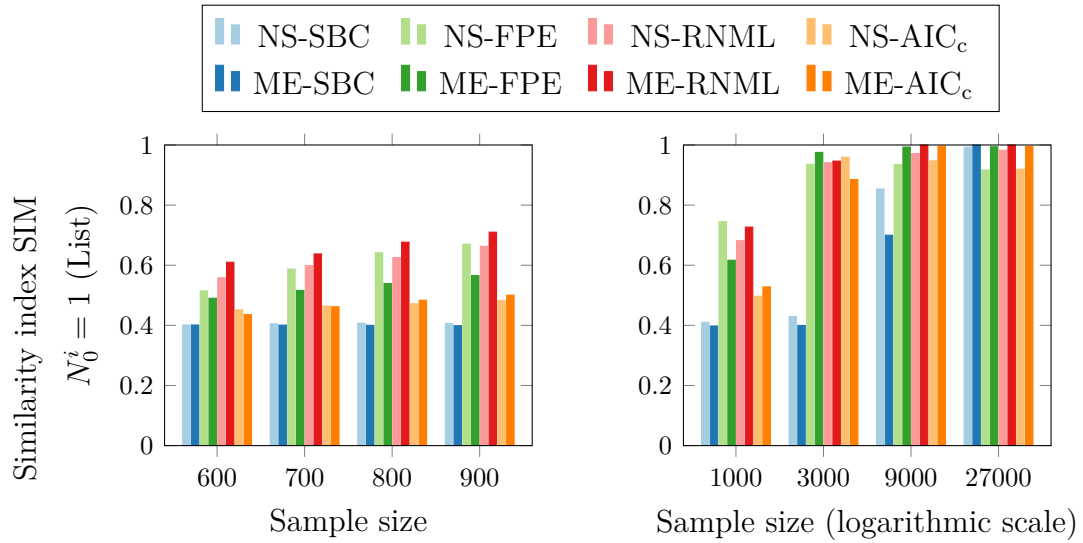


Figure 3.2: Results similar to those presented in Figure 3.1, except that this time  $K = 10$  and  $p^\circ = 5$ .

is that it is the only criterion for which the estimation performance shown in Figure 3.1 clearly improves when applying ME instead of NS.

The experiments are further extended by considering time series with  $K = 10$  components. We take the true order to be five and, when fitting the models, we select the order from the set  $\{1, \dots, 10\}$  by using SBC, FPE, RNML and AIC<sub>c</sub>. In what concerns the true SP's, we generate them randomly such that to have one SP for each  $\#0 \in \{0, 1, 5, 10, 15, 20, 25, 30, 35, 40\}$ .

In Stage2, we apply List for NS-ITC and ME-ITC, where ITC stands for the four criteria which have been also employed in Stage1. The estimation results are presented in Figure 3.2. Remark in Figure 3.2 that ME-RNML is the best, but the gap between the best and the second best method is smaller than in Figure 3.1. For  $T = 1000$ , the performance of NS-FPE in Figure 3.2 is surprisingly good.

**Comparisons with other methods** This makes us to investigate more carefully this case by considering various estimation methods in Stage2. Because we do not want our conclusions to be influenced by the outcome of Stage1, we assume that the order of the model is known ( $p^\circ = 5$ ). Furthermore, we select only

two SP's from the larger set used in the experiment for which the results are reported in Figure 3.2 because we want to study how the estimation accuracy depends on the sparsity of SP. Hence, we take SP's for which  $\#0$  is 10 and 40. Note that the maximum possible number of zeros is  $\overline{K} = 45$  because  $K = 10$ . For each SP, ten different data sets with  $T = 1000$  each are simulated.

Similar to Figure 3.2, the methods employed in estimation use List (L) algorithm for generating the set of candidates and, for optimization, either NS or ME are applied. Additionally, we consider the case when NS/ME are applied to a smaller set of competing models which are produced by Fast List (FL). More precisely, for the method described in Refinement III, we choose  $G_0 = 3$  and this makes all the competing SP's to have a number of zeros which is multiple integer of three. Therefore, the two "true" SP's are not amongst the candidates.

For the sake of comparison, the algorithms L and FL are also used in conjunction with the method from Eichler (2006), where an ISDM with a given sparsity pattern SP is fitted to the data by a cyclic algorithm. Each cycle begins with fitting a  $\text{VAR}(p^\circ)$ -model to the data by Yule-Walker equations (Reinsel, 1993). The number of the remaining steps for each cycle equals the number of zeros in SP because, at each such step, a zero is forced below the main diagonal of ISDM by Wermuth-Scheidt (WS) algorithm (Wermuth and Scheidt, 1977). We note that, for the very first cycle, the estimates in (3.2) with  $\hat{p} = p^\circ$  are used for solving the Yule-Walker equations. For all other cycles, the autocovariances used in Yule-Walker equations are calculated from the expression of SDM obtained at the end of the previous cycle. It was proven in Eichler (2006) that the algorithm converges to the ISDM with the desired SP. Based on this theoretical result, in our implementation we stop the iterations when the following condition is satisfied:

$$\frac{\sum_{(a,b) \in \text{SP}} \sum_{m=0}^{p^\circ} |[\hat{\mathbf{Q}}_m]_{ab}|}{(p^\circ + 1) \times (\#0)} < 10^{-4},$$

where  $\{\hat{\mathbf{Q}}_m\}_{m=0}^{p^\circ}$  are the estimates for the matrices in (1.3) and  $\#0$  is the number of zeros in SP. Additionally, the number of iterated cycles cannot be smaller

than three nor larger than twenty.

In this experiment, we also evaluate the method from [Songsiri and Vandenberghe \(2010\)](#), called here Regularized ME (RME), for which the sparsity promoter  $g(\cdot)$  in (3.12) is given by the formula in equation (22) from the aforementioned study. For generating the set of  $\lambda$ -values employed in the algorithm, we rely on theoretical results from [Basu and Michailidis \(2015\)](#) and choose  $\lambda(T) = \sqrt{\log(K)/T}$ . Furthermore, we calculate  $c = \lceil \log_{10}[1/\lambda(T)] \rceil$ , where  $\log_{10}(\cdot)$  denotes the logarithm with base 10 and  $\lceil \cdot \rceil$  is the smallest integer greater than or equal to the real number in the argument. Then we take  $\lambda_i = i \times 10^{-c}$ , for  $i = 1, \dots, 20$ . In this way, we obtain solutions that are denser than the true one as well as solutions that are sparser than the true one.

The results of the experiment are shown in Figure 3.3. The acronyms of the methods are recapitulated in the caption. The method called Oracle knows the “true” SP and selects from the list of SP’s created by each method the one which is the most similar with the truth (has the highest SIM index). Observe that the methods FL-NS and FL-ME which we propose are the fastest. The execution time for RME-ME, which is the combination of (3.12) and (3.1), is about the same as the time for FL-WS, which is the combination of FL with the method from [Eichler \(2006\)](#). If List (L) is employed together with NS or ME, the algorithm is still reasonably fast. However, L plus the method from [Eichler \(2006\)](#) leads to an average time of  $2.1 \times 10^3$  sec., without improving significantly the accuracy in selecting the structure. For the clarity of the picture, we do not include this method in Figure 3.3. In general, ME is slower than NS. Another important observation is that, disregarding how the family of candidate models is produced (FL, L or RME), the SIM-index computed for the Oracle is almost one and this demonstrates that all the approaches are able to include in this family models which are very close to the “true” SP. Therefore, in this case, the use of Full-Search, Exhaustive-Search or Greedy-Algorithm can only improve marginally the quality of the estimation. The capability of various IT criteria in selecting, from the family of candidates, the one which is closest to the “true” SP

depends very much on the sparsity of the model for which data were simulated. For example, in Figure 3.3 (b) where the number of zeros below the main diagonal in the ISDM of the simulated model is forty, all the criteria have good and very good performance (SIM-index greater than 0.84). When the number of zeros is only ten (see Figure 3.3 (a)), SIM can be as low as 0.2. In fact, the worst performance in Figure 3.3 (a) is that of SBC. In the same figure, we can see that the performance of RNML is reasonably good for all six methods which are compared.

### 3.7 Air pollution data

We investigate the conditional independence of the concentration levels for air pollutants. The data analyzed in the previous works are multivariate time series which consist of measurements of carbon monoxide (CO), nitrate monoxide (NO), nitrate dioxide (NO<sub>2</sub>), ozone (O<sub>3</sub>) and the solar radiation intensity (R) (Dahlhaus, 2000, Eichler, 2006, Songsiri, Dahl and Vandenberghe, 2010, Davis, Zang and Zheng, 2016, Avventi, Lindquist and Wahlberg, 2013). We use the same type of data, which are publicly available (California Environmental Protection Agency, 2016). They have been hourly recorded at Azusa, California, during 2004-2010. The subset corresponding to the year 2006 was already used in Songsiri, Dahl and Vandenberghe (2010), Davis, Zang and Zheng (2016).

In the pre-processing phase, we split the data into subsets, based on the year when they have been recorded. For the imputation of the missing data, we use the function `minimput` from the R-package `mtsdi` (Junger and de Leon, 2016). After this step, all negative values are turned to zero and the resulting time series are altered to have zero mean vector. The percentage of imputed data varies from 3.8% (in 2006) to 12.9% (in 2004). We get seven time series, with  $K = 5$ . For five of them,  $T = 8760$  and for the other two,  $T = 8783$  (leap years).

We run Stage1 of Algorithm 1 on these data sets, by allowing the pre-specified model orders to range from  $p_{\min} = 1$  to  $p_{\max} = 8$ . RNML selects the best order

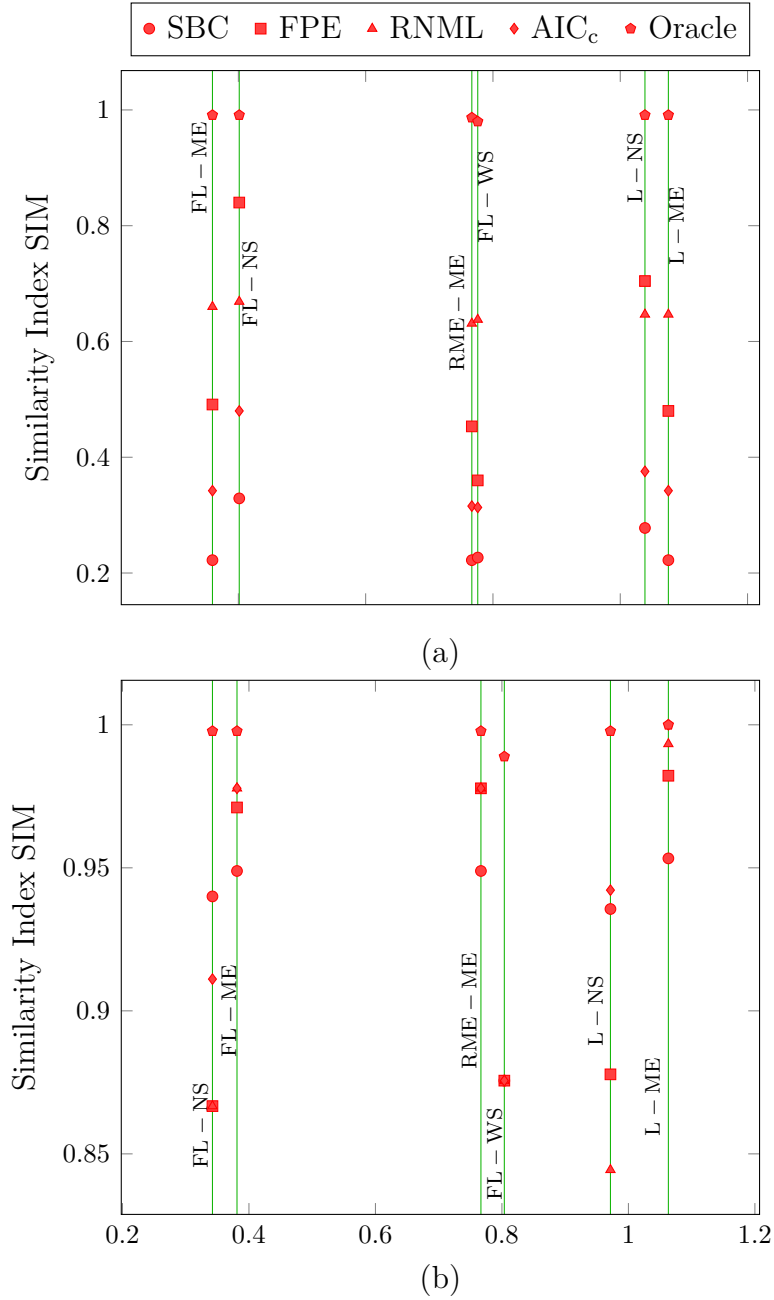


Figure 3.3: Experimental results for Stage2, obtained when  $K = 10$  and  $T = 1000$ . In estimation, the order of the model is assumed to be known ( $p^\circ = 5$ ). In Figure 3.3 (a), the “true” SP has ten zeros, whereas in Figure 3.3 (b) has forty zeros. Each vertical line corresponds to an estimation method, and the methods are ordered on the horizontal axis based on the average execution time (in sec. divided by  $10^3$ ) computed from ten runs (2.6 GHz-processor). The significance of the acronyms is as follows: FL=Fast List, L = List, ME = Maximum Entropy, NS = Nearest Sparse, RME = Regularized Maximum Entropy and WS = Wermuth-Scheidt.

to be 4 (for five years) and 5 (for two years). More compactly, we write [RNML : order4(5); order5(2)]. Similarly, the results obtained by the other criteria can be written [SBC : order4(4); order5(3)], [FPE : order8(7)] and [AIC<sub>c</sub> : order8(7)]. As it is almost generally accepted that the order should be four (Eichler, 2006, Songsiri, Dahl and Vandenberghe, 2010, Davis, Zang and Zheng, 2016), it means that the most accurate results are produced by RNML and SBC, whereas FPE and AIC<sub>c</sub> overestimate the order of the model.

We show in Figure 3.4 the pairs of components of the multivariate time series, which are found to be conditionally independent when running Stage2 of the algorithm. For each such pair, we plot the number of times it was deemed to be conditionally independent (divided by the number of years). It is important to emphasize that all pairs of pollutants presented in the figure are conditionally independent according to the ground knowledge from environmental chemistry (see Dahlhaus (2000) and the references therein). Additionally, “NIL” is used for the situation when the best ranked SP does not contain any zero. When NS is applied, all SP’s for which the smallest eigenvalue of ISDM over  $(-\pi, \pi]$  is not greater than  $10^{-6}$  are automatically disqualified from competition.

In Figure 3.4(a), which is produced by employing List, the large counts of NIL for ME-FPE and ME-AIC<sub>c</sub> demonstrate that these two methods fail to find conditional independence in the multivariate time series. ME-RNML is the most successful in identifying the conditional independence and its performance is followed closely by ME-SBC. In Figure 3.4(b), the results yield by ME-FPE and ME-AIC<sub>c</sub> are disappointing. It is interesting that the pair (Rad,CO) is always properly identified in Figure 3.4(b) if NS is used for optimization. In spite of the fact that the plot in this figure is different from the one in Figure 3.4(a), ME-RNML and ME-SBC remain superior to other methods.

For the sake of clarity, we present in Tables 3.3-3.9 all the results we obtained for air pollution data. Each table shows the pairs of components which are found to be statistically independent by various methods, for a specific year; the year when the measurements were collected is mentioned in the caption of the table.

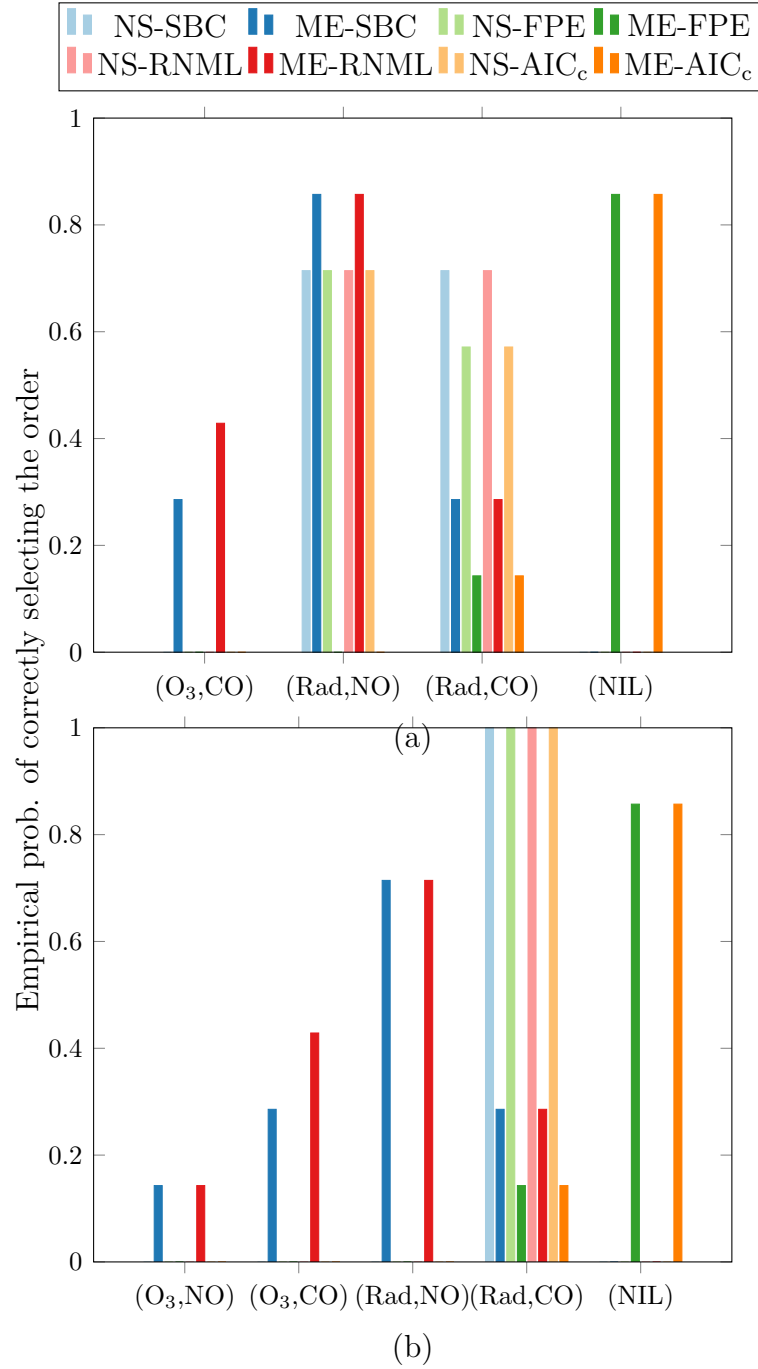


Figure 3.4: Number of times each pair of components of air pollution multivariate time series is found to be conditionally independent (divided by the number of analyzed years). The pairs which have never been selected are not shown in the figure. The results in panel (a) are obtained with the List and those in panel (b) are produced by applying Greedy.

The acronym L-G is used for denoting Algorithm 2 when  $N_0^i = \min(4, N_{\text{tot}}^i)$ . A very rapid inspection of the content of the tables leads immediately to the conclusion that, in all the cases, the selected ISDM has very few zeros. In fact, there are situations when the selected ISDM does not contain zeros at all, and these are counted as “NIL” in Figure 3.4. Moreover,  $(O_3, CO)$ ,  $(Rad, NO)$ ,  $(Rad, CO)$  and  $(O_3, NO)$  are the only pairs which, at least in one of the tables, are reported to be statistically independent. In light of this discussion, we conclude that none of the entries of ISDM are misclassified as zeros.

	Estimated orders for the VAR model			
	SBC	FPE	RNML	AICC
	4	8	4	8
Near Sparse				
	SBC	FPE	RNML	AICC
List	(Rad,CO)	(Rad,CO)	(Rad,CO)	(Rad,CO)
L-G	(Rad,CO)	(Rad,CO)	(Rad,CO)	(Rad,CO)
Greedy	(Rad,CO)	(Rad,CO)	(Rad,CO)	(Rad,CO)
Maximum Entropy				
	SBC	FPE	RNML	AICC
List	(Rad,CO)	(Rad,CO)	(Rad,CO)	(Rad,CO)
L-G	(Rad,CO)	(Rad,CO)	(Rad,CO)	(Rad,CO)
Greedy	(Rad,CO)	(Rad,CO)	(Rad,CO)	(Rad,CO)

Table 3.3: Year 2004

### 3.8 International stock markets

Another typical application of graphical models is in modeling the interdependence of financial markets. The undirected graph is obtained from the stock markets returns (at closing time) of  $K$  financial markets, for  $T$  consecutive days. Two such data sets, for which  $K = 5$ , were analyzed in [Songsiri, Dahl and Vandenberghe \(2010\)](#). We consider a data set with  $K = 10$  which was previously used to produce the graphical model in [\(Abdelwahab, Amor and Abdelwaheb, 2008, Figure 4\)](#) by applying the method from [Fried and Didelez \(2003\)](#). After downloading the data for the period Jan.2000 - Dec.2005 from

### 3. EFFICIENT ALGORITHMS FOR VAR GRAPH ESTIMATION

	Estimated orders for the VAR model			
	SBC	FPE	RNML	AICC
	4	8	4	8
Near Sparse				
	SBC	FPE	RNML	AICC
List	(Rad,NO)	(Rad,NO)	(Rad,NO)	(Rad,NO)
L-G	(Rad,CO)	(Rad,CO)	(Rad,CO)	(Rad,CO)
Greedy	(Rad,CO)	(Rad,CO)	(Rad,CO)	(Rad,CO)
Maximum Entropy				
	SBC	FPE	RNML	AICC
List	(Rad,NO)	None	(Rad,NO)	None
	(O <sub>3</sub> ,CO)		(O <sub>3</sub> ,CO)	
L-G	(Rad,NO)	None	(Rad,NO)	None
	(O <sub>3</sub> ,CO)		(O <sub>3</sub> ,CO)	
Greedy	(Rad,NO)	None	(Rad,NO)	None
	(O <sub>3</sub> ,CO)		(O <sub>3</sub> ,CO)	

Table 3.4: Year 2005

	Estimated orders for the VAR model			
	SBC	FPE	RNML	AICC
	5	8	4	8
Near Sparse				
	SBC	FPE	RNML	AICC
List	(Rad,NO)	(Rad,NO)	(Rad,NO)	(Rad,NO)
L-G	(Rad,CO)	(Rad,CO)	(Rad,CO)	(Rad,CO)
Greedy	(Rad,CO)	(Rad,CO)	(Rad,CO)	(Rad,CO)
Maximum Entropy				
	SBC	FPE	RNML	AICC
List	(Rad,NO)	None	(Rad,NO)	None
L-G	(Rad,NO)	None	(Rad,NO)	None
Greedy	(Rad,NO)	None	(Rad,NO)	None

Table 3.5: Year 2006

<http://finance.yahoo.com/> and applying the same pre-processing as in Abdelwahab, Amor and Abdelwaheb (2008), we get a 10-variate time series for which  $T = 1546$ .

In Stage1 of our algorithm, we take  $p_{\min} = 1$  and  $p_{\max} = 9$ ; SBC and RNML select  $\hat{p} = 1$ , whereas FPE and AIC<sub>c</sub> choose  $\hat{p} = 2$ . Selection of small orders is

	Estimated orders for the VAR model			
	SBC	FPE	RNML	AICC
	4	8	4	8
Near Sparse				
	SBC	FPE	RNML	AICC
List	(Rad,NO)	(Rad,NO)	(Rad,NO)	(Rad,NO)
	(Rad,CO)	(Rad,CO)	(Rad,CO)	(Rad,CO)
L-G	(Rad,CO)	(Rad,CO)	(Rad,CO)	(Rad,CO)
Greedy	(Rad,CO)	(Rad,CO)	(Rad,CO)	(Rad,CO)
Maximum Entropy				
	SBC	FPE	RNML	AICC
List	(Rad,NO)	None	(Rad,NO)	None
L-G	(Rad,NO)	None	(Rad,NO)	None
Greedy	(Rad,NO)	None	(Rad,NO)	None

Table 3.6: Year 2007

	Estimated orders for the VAR model			
	SBC	FPE	RNML	AICC
	5	8	5	8
Near Sparse				
	SBC	FPE	RNML	AICC
List	(Rad,NO)	(Rad,NO)	(Rad,NO)	(Rad,NO)
	(Rad,CO)		(Rad,CO)	
L-G	(Rad,CO)	(Rad,CO)	(Rad,CO)	(Rad,CO)
Greedy	(Rad,CO)	(Rad,CO)	(Rad,CO)	(Rad,CO)
Maximum Entropy				
	SBC	FPE	RNML	AICC
List	(Rad,NO)	None	(Rad,NO)	None
L-G	(O <sub>3</sub> ,NO)	None	(O <sub>3</sub> ,NO)	None
Greedy	(O <sub>3</sub> ,NO)	None	(O <sub>3</sub> ,NO)	None

Table 3.7: Year 2008

perfectly in line with the results reported previously (see, for example, [Songsiri, Dahl and Vandenberghe \(2010\)](#)). In Stage2, we apply List and Greedy for NS-ITC and ME-ITC, where  $ITC \in \{SBC, FPE, RNML, AIC_c\}$ . Each resulting graph is compared with the one in ([Abdelwahab, Amor and Abdelwaheb, 2008](#), Figure 4) by calculating the similarity index in (3.17) with  $N_{SP} = N_r = 1$ . The largest value of the index (0.84) is obtained when List is used in combination with

### 3. EFFICIENT ALGORITHMS FOR VAR GRAPH ESTIMATION

	Estimated orders for the VAR model			
	SBC	FPE	RNML	AICC
	5	8	5	8
Near Sparse				
	SBC	FPE	RNML	AICC
List	(Rad,NO)	(Rad,NO)	(Rad,NO)	(Rad,NO)
	(Rad,CO)	(Rad,CO)	(Rad,CO)	(Rad,CO)
L-G	(Rad,CO)	(Rad,CO)	(Rad,CO)	(Rad,CO)
Greedy	(Rad,CO)	(Rad,CO)	(Rad,CO)	(Rad,CO)
Maximum Entropy				
	SBC	FPE	RNML	AICC
List	(Rad,NO)	None	(Rad,NO)	None
			(O <sub>3</sub> ,CO)	
L-G	(Rad,NO)	None	(Rad,NO)	None
			(O <sub>3</sub> ,CO)	
Greedy	(Rad,NO)	None	(Rad,NO)	None
			(O <sub>3</sub> ,CO)	

Table 3.8: Year 2009

	Estimated orders for the VAR model			
	SBC	FPE	RNML	AICC
	4	8	4	8
Near Sparse				
	SBC	FPE	RNML	AICC
List	(Rad,CO)	(Rad,CO)	(Rad,CO)	(Rad,CO)
L-G	(Rad,CO)	(Rad,CO)	(Rad,CO)	(Rad,CO)
Greedy	(Rad,CO)	(Rad,CO)	(Rad,CO)	(Rad,CO)
Maximum Entropy				
	SBC	FPE	RNML	AICC
List	(Rad,NO)	None	(Rad,NO)	None
	(O <sub>3</sub> ,CO)		(O <sub>3</sub> ,CO)	
	(Rad,CO)		(Rad,CO)	
L-G	(Rad,NO)	None	(Rad,NO)	None
	(O <sub>3</sub> ,CO)		(O <sub>3</sub> ,CO)	
	(Rad,CO)		(Rad,CO)	
Greedy	(Rad,NO)	None	(Rad,NO)	None
	(O <sub>3</sub> ,CO)		(O <sub>3</sub> ,CO)	
	(Rad,CO)		(Rad,CO)	

Table 3.9: Year 2010

ME-SBC. The second largest is 0.78 and is produced by List plus ME-RNML. As these two results are the best, we show in Figure 3.5 the values of the two IT criteria for all SP's considered in List. Remark in Figure 3.5 (a) that the vertical line corresponding to the minimum value of SBC divides the panel into two parts. The red markers in the left part represent edges which, by mistake, have not been included in the ME-SBC graph. However, all of them correspond to pairs of components with “weak PSC” and, according to Abdelwahab, Amor and Abdelwaheb (2008), they represent “doubtful links”. Hence, not including them in the graph is not that severe. At the same time, the green markers in the right half of the panel represent four edges which are erroneously included in the ME-SBC graph. A similar interpretation can be done for the plot in Figure 3.5 (b).

### 3.9 Summary

In this chapter, we have proposed a family of algorithms for inferring the conditional independence graph of a VAR( $p$ )-model for  $K$ -variate time series. Our theoretical and empirical results demonstrate that the algorithms from this family can be used when  $p \leq 20$ ,  $K \leq 10$  and  $K \times p \leq 100$ . Thus far, the methods which rely on convex optimization and do not ask the user to make subjective choice of parameters have been suitable only for much smaller values of  $p$  and  $K$ . Another important feature of our method is the guaranteed stability of the fitted model. However, the algorithm does not include the case of latent variables, i.e. when one or more variables is not measurable. This will be the subject of the next chapter.

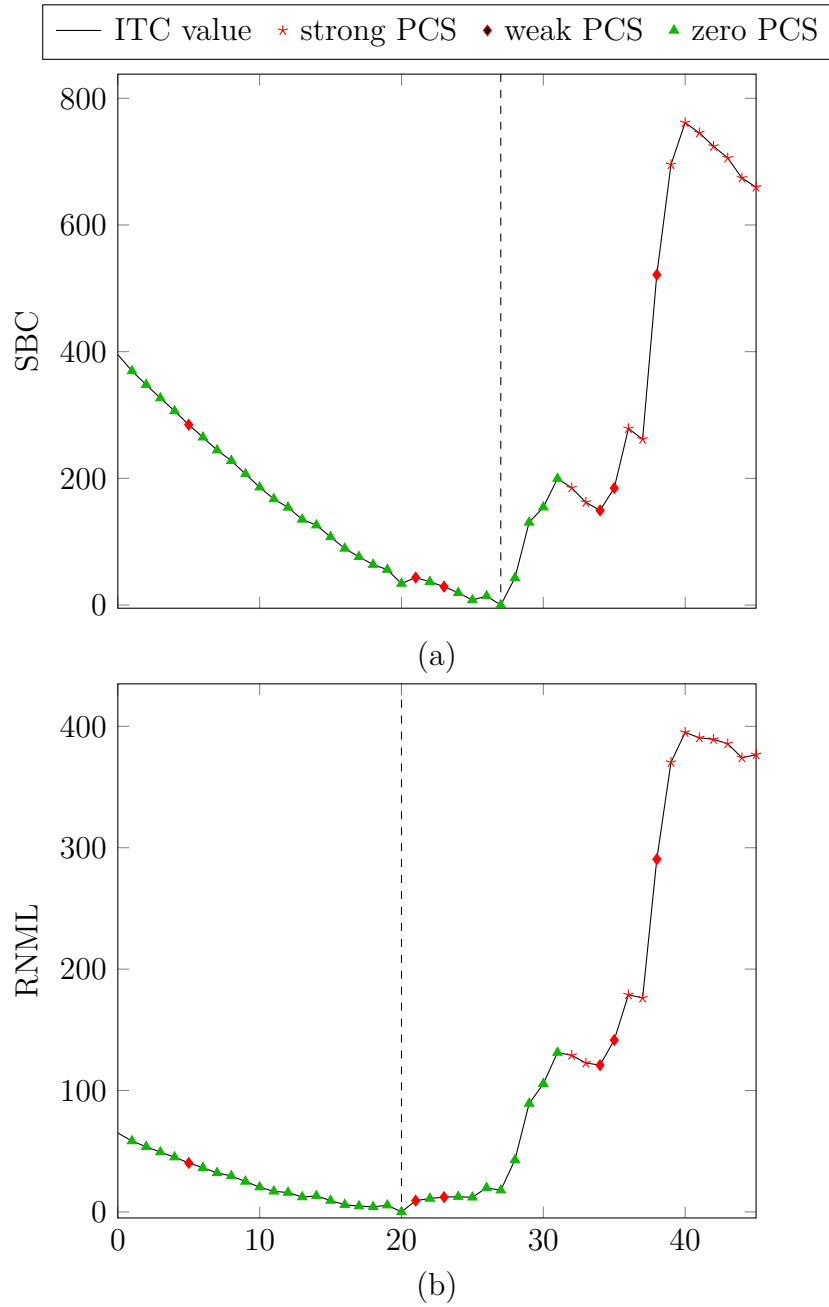


Figure 3.5: International stock markets:  $\text{ITC}(\mathbf{Y}; \hat{p}, \text{SP}^i) - \min_{0 \leq q \leq \bar{K}} \text{ITC}(\mathbf{Y}; \hat{p}, \text{SP}^q)$  for  $\text{ITC} \in \{\text{SBC}, \text{RNML}\}$ , when the index  $i$  in Algorithm 2 takes values from zero (no zeros in the fitted SP) to  $\bar{K}$  (maximum number of zeros in SP). Graphical conventions are the same for both plots. The vertical line marks the SP for which the ITC is minimized. Let  $\{(a_i, b_i)\} = \text{SP}^i \setminus \text{SP}^{i-1}$  for  $0 < i \leq \bar{K}$ . To the point whose abscissa is  $i$ , we assign a marker according to the classification from Abdelwahab, Amor and Abdelwaheb (2008) of the PSC of marginal time series  $\mathbf{y}_{a_i}$  and  $\mathbf{y}_{b_i}$  (“strong PSC”, “weak PSC”, “null PSC”).

# Chapter 4

## Latent-Variable VAR Graphical Models

### 4.1 Preliminaries

In the previous chapters we have addressed the problem of inferring the conditional independence graph for the VAR model given in (1.2). Here we focus on the identification of graphical models for VAR processes with latent variables. The definition for such processes is presented below.

Let  $\mathbf{x}_1, \dots, \mathbf{x}_T$  be a  $\kappa$ -dimensional ( $\kappa > 1$ ) time series generated by a stationary and stable VAR process of order  $p$ . We assume that the spacing of observation times is constant and  $\mathbf{x}_t = [\mathbf{x}_{1t}, \dots, \mathbf{x}_{\kappa t}]'$ . The difference equation of the process is

$$\mathbf{x}_t = \mathbf{A}_1 \mathbf{x}_{t-1} + \dots + \mathbf{A}_p \mathbf{x}_{t-p} + \boldsymbol{\epsilon}_t, \quad t = \overline{1, T}, \quad (4.1)$$

where  $\mathbf{A}_1, \dots, \mathbf{A}_p$  are matrix coefficients of size  $\kappa \times \kappa$  and  $\boldsymbol{\epsilon}_t$  is a sequence of independently and identically distributed random  $\kappa$ -vectors. We assume that the vectors  $\{\boldsymbol{\epsilon}_t\}_{t=1}^T$  are drawn from a  $\kappa$ -variate Gaussian distribution with zero mean vector and covariance matrix  $\boldsymbol{\Sigma} \succ 0$ . Additionally, the vectors  $\{\mathbf{x}_t\}_{t=1-p}^0$  are assumed to be constant.

As we already know from Section 1.2, the conditional independence relations between the variables in  $\mathbf{x}_t$  are provided by ISDM, which has the expression

$$\Phi^{-1}(\omega) = \mathbf{A}^H(\omega) \Sigma^{-1} \mathbf{A}(\omega) = \sum_{i=-p}^p \mathbf{Q}_i e^{-j\omega i}. \quad (4.2)$$

Similar to the notation used in (1.3), we define  $\mathbf{A}_0 = -\mathbf{I}$  and  $\mathbf{A}(\omega) = -\sum_{i=0}^p \mathbf{A}_i e^{-j\omega i}$ . For  $i \geq 0$ , we have that  $\mathbf{Q}_i = \sum_{k=0}^{p-i} \mathbf{A}'_k \Sigma^{-1} \mathbf{A}_{k+i}$  and  $\mathbf{Q}_{-i} = \mathbf{Q}'_i$ . The sparse structure of the ISDM contains conditional dependence relations between the variables of  $\mathbf{x}_t$ , i.e. two variables  $\mathbf{x}_a$  and  $\mathbf{x}_b$  are independent, conditional on the other variables, if and only if (Brillinger, 1996, Dahlhaus, 2000)

$$[\Phi^{-1}(\omega)]_{ab} = 0, \quad \forall \omega \in (-\pi, \pi]. \quad (4.3)$$

In a latent-variable graphical model it is assumed that  $\kappa = K + r$ , where  $K$  variables are accessible to observation (they are called manifest variables) and  $r$  variables are latent, i.e., not accessible to observation, but playing a significant role in the conditional independence pattern of the overall model. The existence of latent variables in a model can be described in terms of the ISDM by the block decomposition

$$\Phi(\omega) = \begin{bmatrix} \Phi_m(\omega) & \Phi'_{\ell m}(-\omega) \\ \Phi_{\ell m}(\omega) & \Phi_{\ell}(\omega) \end{bmatrix}, \quad \Phi^{-1}(\omega) = \begin{bmatrix} \Upsilon_m(\omega) & \Upsilon'_{\ell m}(-\omega) \\ \Upsilon_{\ell m}(\omega) & \Upsilon_{\ell}(\omega) \end{bmatrix}, \quad (4.4)$$

where  $\Phi_m(\omega)$  and  $\Phi_{\ell}(\omega)$  are the manifest and latent components of the spectral density matrix, respectively.

Using the Schur complement, the ISDM of the manifest component has the form (Zorzi and Sepulchre, 2016, Eq. (21)):

$$\Phi_m^{-1}(\omega) = \Upsilon_m(\omega) - \Upsilon'_{\ell m}(-\omega) \Upsilon_{\ell}^{-1}(\omega) \Upsilon_{\ell m}(\omega). \quad (4.5)$$

When building latent variable graphical models, we assume that  $r \ll K$ , i.e., few latent variables are sufficient to characterize the conditional dependence structure of the model. The second term in (4.5) has low rank because the middle matrix has small size. The previous formula can therefore be written

$$\Phi_m^{-1}(\omega) = \mathbf{S}(\omega) - \mathbf{\Lambda}(\omega), \quad (4.6)$$

where  $\mathbf{S}(\omega)$  is sparse, and  $\mathbf{\Lambda}(\omega)$  has (constant) low-rank almost everywhere in  $(-\pi, \pi]$ . Furthermore, we can write (Liegeois et al., 2015, Eq. (4)):

$$\mathbf{S}(\omega) - \mathbf{\Lambda}(\omega) = \mathbf{\Delta}(\omega)\mathbf{X}\mathbf{\Delta}(\omega)^H, \quad (4.7)$$

$$\mathbf{\Lambda}(\omega) = \mathbf{\Delta}(\omega)\mathbf{L}\mathbf{\Delta}(\omega)^H, \quad (4.8)$$

where  $\mathbf{\Delta}(\omega) = [\mathbf{I}, e^{j\omega}\mathbf{I}, \dots, e^{j\omega p}\mathbf{I}]$  is a shift matrix, and  $\mathbf{X}$  and  $\mathbf{L}$  are  $K(p+1) \times K(p+1)$  positive semidefinite matrices. We split all such matrices in  $K \times K$  blocks, e.g.,

$$\mathbf{L} = \begin{bmatrix} \mathbf{L}_{00} & \dots & \mathbf{L}_{0p} \\ \vdots & \ddots & \vdots \\ \mathbf{L}'_{0p} & \dots & \mathbf{L}_{pp} \end{bmatrix}. \quad (4.9)$$

According to the definition from Section 3.1, the block trace operator for such a matrix is

$$\text{TR}_i(\mathbf{L}) = \sum_{h=0}^{p-i} \mathbf{L}_{h, h+i}, \quad i = \overline{0, p}. \quad (4.10)$$

For negative indices, the relation  $\text{TR}_{-i}(\mathbf{L}) = \text{TR}_i(\mathbf{L})^\top$  holds. Note that (4.8) can be rewritten as

$$\mathbf{\Lambda}(\omega) = \sum_{i=-p}^p \text{TR}_i(\mathbf{L}) e^{j\omega i}.$$

The first  $p+1$  sample covariances of the VAR process are (Brockwell and Davis, 1991):

$$\hat{\mathbf{C}}_i = \frac{1}{T} \sum_{t=1}^{T-i} \mathbf{x}_{t+i} \mathbf{x}_t', \quad i = \overline{0, p}. \quad (4.11)$$

However, only the upper left  $K \times K$  blocks corresponding to the manifest variables can be computed from data; they are denoted  $\hat{\mathbf{R}}_i$ . With  $\hat{\mathbf{R}} = [\hat{\mathbf{R}}_0 \dots \hat{\mathbf{R}}_p]$ , we build the block Toeplitz matrix [see also (3.3)]

$$\mathcal{T}(\hat{\mathbf{R}}) = \begin{bmatrix} \hat{\mathbf{R}}_0 & \dots & \hat{\mathbf{R}}_p \\ \vdots & \ddots & \vdots \\ \hat{\mathbf{R}}'_p & \dots & \hat{\mathbf{R}}_0 \end{bmatrix}. \quad (4.12)$$

Zorzi and Sepulchre (2016) proposed to estimate the matrices  $\mathbf{X}$  and  $\mathbf{L}$  by solving the optimization problem

$$\begin{aligned} \min_{\mathbf{X}, \mathbf{L}} \quad & \text{tr}(\mathcal{T}(\hat{\mathbf{R}}) \mathbf{X}) - \log \det \mathbf{X}_{00} + \lambda \gamma f(\mathbf{X} + \mathbf{L}) + \lambda \text{tr}(\mathbf{L}) \\ \text{s.t.} \quad & \mathbf{X} \succeq 0 \quad \mathbf{L} \succeq 0 \end{aligned} \quad (4.13)$$

Minimizing  $\text{tr}(\mathbf{L})$  induces low rank in  $\mathbf{L}$  and  $\lambda, \gamma > 0$  are trade-off constants. The function  $f(\cdot)$  is a group sparsity promoter whose expression is given by

$$f(\mathbf{Z}) = \sum_{a=1}^K \sum_{b=1}^{a-1} \max_{i=0,p} |\text{TR}_i(\mathbf{Z})(a, b)|. \quad (4.14)$$

Note that  $\text{TR}_i(\mathbf{X} + \mathbf{L})(a, b)$  is the  $i$ -th degree coefficient of the polynomial that occupies the  $(a, b)$  position in the matrix polynomial  $\Phi_m^{-1}(\omega)$ . Sparsity is encouraged by minimizing the  $\ell_1$ -norm of the vector formed by the coefficients that are maximum for each position  $(a, b)$ . Once  $\mathbf{X}$  and  $\mathbf{L}$  are solved, the coefficients of the VAR model are obtained by spectral factorization.

## 4.2 Maximum Entropy

### Expectation-Maximization algorithm

The obvious advantage of the optimization problem in (4.13) is its convexity, which allows the safe computation of the solution. However, a possible drawback is the presence of two parameters,  $\lambda$  and  $\gamma$ , whose values should be chosen. A way to eliminate one of the parameters is to assume that the number  $r$  of latent variables is known. At least for parsimony reasons, it is natural to suppose that  $r$  is very small. Since a latent variable influences all manifest variables in the ISDM (4.5), there cannot be too many independent latent variables. So, giving  $r$  a fixed small value is likely to be not restrictive.

In this section, we describe an estimation method which is clearly different from the one in Zorzi and Sepulchre (2016). More precisely, we generalize the Expectation-Maximization algorithm from Lauritzen and Meinhausen (2012), developed there for independent and identically distributed random variables,

to a VAR process. To this purpose, we work with the full model (4.4) that includes the ISDM part pertaining to the  $r$  latent variables. Without loss of generality, we assume that  $\Upsilon_\ell(\omega)$  equals the identity matrix  $\mathbf{I}$ ; the effect of the latent variables on the manifest ones in (4.5) can be modelled by  $\Upsilon_{\ell m}$  alone. Combining with (4.2), the model is

$$\Phi^{-1}(\omega) = \begin{bmatrix} \Upsilon_m(\omega) & \Upsilon'_{\ell m}(-\omega) \\ \Upsilon_{\ell m}(\omega) & \mathbf{I} \end{bmatrix} = \sum_{i=-p}^p \mathbf{Q}_i e^{-j\omega i}, \quad (4.15)$$

where the matrices  $\mathbf{Q}_i$  have to be found.

The main difficulty of this approach is the unavailability of the latent part of the matrices (4.11). Were such matrices available, we could work with SDM  $\Phi(\omega)$  estimators (confined to order  $p$ ) of the form

$$\tilde{\Phi}(\omega) = \sum_{i=-p}^p \mathbf{C}_i e^{-j\omega i}, \quad (4.16)$$

where  $\mathbf{C}_i$  denotes the  $i$ -th covariance lag for the VAR process  $\{\mathbf{x}_t\}$  [see also (4.1) and (4.11)]. We split the matrix coefficients from (4.15) and (4.16) according to the size of manifest and latent variables, e.g.,

$$\mathbf{C}_i = \begin{bmatrix} \mathbf{C}_{m,i} & \mathbf{C}'_{\ell m,-i} \\ \mathbf{C}_{\ell m,i} & \mathbf{C}_{\ell,i} \end{bmatrix}. \quad (4.17)$$

To overcome the difficulty, the Expectation-Maximization algorithm alternatively keeps fixed either the model parameters  $\mathbf{Q}_i$  or the matrices  $\mathbf{C}_i$ , estimating or optimizing the remaining unknowns. The expectation step of Expectation-Maximization assumes that the ISDM  $\Phi^{-1}(\omega)$  from (4.15) is completely known. Standard matrix identities (Lauritzen and Meinhausen, 2012) can be easily extended to matrix trigonometric polynomials for writing down the formula

$$\Phi(\omega) = \begin{bmatrix} \Phi_m(\omega) & -\Phi_m(\omega)\Upsilon'_{\ell m}(-\omega) \\ -\Upsilon_{\ell m}(\omega)\Phi_m(\omega) & \mathbf{I} + \Upsilon_{\ell m}(\omega)\Phi_m(\omega)\Upsilon'_{\ell m}(-\omega) \end{bmatrix}. \quad (4.18)$$

Identifying (4.16) with (4.18) gives expressions for estimating the matrices  $\mathbf{C}_i$ , depending on the matrices  $\mathbf{Q}_i$  from (4.15). The upper left corner of (4.18) needs

no special computation, since the natural estimator is

$$\Phi_m(\omega) = \sum_{i=-p}^p \hat{\mathbf{R}}_i e^{-j\omega i},$$

where the sample covariances  $\hat{\mathbf{R}}_i$  are directly computable from the time series.

It results that

$$\mathbf{C}_{m,i} = \hat{\mathbf{R}}_i, \quad i = \overline{-p, p}. \quad (4.19)$$

The other blocks from (4.17) result from convolution expressions associated with the polynomial multiplications from (4.18). The lower left block of the coefficients is

$$\mathbf{C}_{\ell m,i} = - \sum_{k+s=i} \mathbf{Q}_{\ell m,k} \hat{\mathbf{R}}_s = - \sum_{k=\max(-p,i-p)}^{\min(p,i+p)} \mathbf{Q}_{\ell m,k} \hat{\mathbf{R}}_{i-k}, \quad i = \overline{-2p, 2p}. \quad (4.20)$$

Note that the trigonometric polynomial  $\Upsilon_{\ell m}(\omega)\Phi_m(\omega)$  has degree  $2p$ , since its factors have degree  $p$ . With (4.20) available, we can compute

$$\mathbf{C}_{\ell,i} = \delta_i \mathbf{I} + \sum_{k-s=i} \mathbf{C}_{\ell m,k} \mathbf{Q}'_{\ell m,-s} = \delta_i \mathbf{I} + \sum_{k=\max(-2p,i-p)}^{\min(2p,i+p)} \mathbf{C}_{\ell m,k} \mathbf{Q}'_{\ell m,i-k}, \quad i = \overline{-p, p}, \quad (4.21)$$

where  $\delta_i = 1$  if  $i = 0$  and  $\delta_i = 0$  otherwise. Although the degree of the polynomial from the lower right block of (4.18) is  $3p$ , we need to truncate it to degree  $p$ , since this is the degree of the ISDM  $\Phi^{-1}(\omega)$  from (4.15). This is the reason for computing only the coefficients  $i = \overline{-p, p}$  in (4.21). The same truncation is applied on (4.20); note that there we cannot compute only the coefficients that are finally needed, since all of them are required in (4.21).

In the maximization step of Expectation-Maximization, the covariance matrices  $\mathbf{C}_i$  are assumed to be known and are fixed; the ISDM can be estimated by solving an optimization problem that will be detailed below. The overall solution we propose is outlined in Algorithm 3, explained in what follows.

The initialization stage provides a first estimate for the ISDM, from which the Expectation-Maximization alternations can begin. An estimate for the left upper corner of  $\Phi^{-1}(\omega)$  is obtained by solving the *classical* Maximum Entropy problem

---

**Algorithm 3** Algorithm for Identifying SP of ISDM (AlgoEM)
 

---

**Input:** Data  $(\mathbf{x}_{1:K,1}, \dots, \mathbf{x}_{1:K,T})$ , VAR-order  $p$ , number of latent variables  $r$ , an information theoretic criterion (ITC).

**Initialization:**

Evaluate  $\hat{\mathbf{R}}_i$  for  $i = \overline{0, p}$ ; [see (4.11) and the discussion below it]

$\hat{\mathbf{R}} \leftarrow [\hat{\mathbf{R}}_0 \dots \hat{\mathbf{R}}_p]$ ;

$\hat{\Phi}_m(\omega) \leftarrow \sum_{i=-p}^p \hat{\mathbf{R}}_i e^{-j\omega i}$ ;

$\left\{ \check{\mathbf{Q}}_i^{(0)}(1 : K, 1 : K) \right\}_{i=0}^p \leftarrow \text{ME}_I(\hat{\mathbf{R}})$  [see (4.22)];

Compute  $\check{\mathbf{Y}}_{\ell m}^{(0)}(\omega)$  from EIG of  $\check{\mathbf{Q}}_0^{(0)}$ ;

**for all**  $\lambda \in \{\lambda_1, \dots, \lambda_L\}$  **do**

**Maximum Entropy Expectation-Maximization (penalized setting):**

**for**  $\text{it} = 1, \dots, N_{\text{it}}$  **do**

Use  $\hat{\Phi}_m(\omega)$  and  $\check{\mathbf{Y}}_{\ell m}^{(\text{it}-1)}(\omega)$  to compute  $\check{\mathbf{C}}^{(\text{it})}$  [see (4.16)-(4.18)];

$\left\{ \check{\mathbf{Q}}_i^{(\text{it})} \right\}_{i=0}^p \leftarrow \text{ME}_{\text{II}}(\check{\mathbf{C}}^{(\text{it})}, \lambda)$  [see (4.23)];

Get  $\check{\mathbf{Y}}_{\ell m}^{(\text{it})}(\omega)$  from  $\left\{ \check{\mathbf{Q}}_i^{(\text{it})} \right\}_{i=0}^p$  [see (4.15)];

**end for**

Use  $\left\{ \check{\mathbf{Q}}_i^{(N_{\text{it}})} \right\}_{i=0}^p$  to compute  $\check{\Phi}_\lambda^{-1}(\omega)$ ;

Determine  $\text{SP}_\lambda$  [see (4.24)];

**if** ADAPTIVE **then**

$\check{\mathbf{Y}}_{\ell m}^{(0)}(\omega) \leftarrow \check{\mathbf{Y}}_{\ell m}^{(N_{\text{it}})}(\omega)$

**end if**

$\check{\mathbf{Y}}_{\ell m}^{(0)}(\omega) \leftarrow \check{\mathbf{Y}}_{\ell m}^{(N_{\text{it}})}(\omega)$ ;

**Maximum Entropy Expectation-Maximization (constrained setting):**

**for**  $\text{it} = 1, \dots, N_{\text{it}}$  **do**

Use  $\hat{\Phi}_m(\omega)$  and  $\check{\mathbf{Y}}_{\ell m}^{(\text{it}-1)}(\omega)$  to compute  $\hat{\mathbf{C}}^{(\text{it})}$  [see (4.16)-(4.18)];

$\left\{ \hat{\mathbf{Q}}_i^{(\text{it})} \right\}_{i=0}^p \leftarrow \text{ME}_{\text{III}}(\hat{\mathbf{C}}^{(\text{it})}, \text{SP}_\lambda)$  [see (4.25)];

Get  $\hat{\mathbf{Y}}_{\ell m}^{(\text{it})}(\omega)$  from  $\left\{ \hat{\mathbf{Q}}_i^{(\text{it})} \right\}_{i=0}^p$  [see (4.15)];

**end for**

Use  $\left\{ \hat{\mathbf{Q}}_i^{(N_{\text{it}})} \right\}_{i=0}^p$  to compute  $\hat{\Phi}_\lambda^{-1}(\omega)$ ;

Find the matrix coefficients of the VAR-model by spectral factorization of

$\hat{\Phi}_\lambda^{-1}(\omega)$  and compute  $\text{ITC}(\text{Data}; \text{SP}_\lambda)$ .

**end for**

$\widehat{\text{SP}} \leftarrow \arg \min_{\lambda} \text{ITC}(\text{Data}; \text{SP}_\lambda)$ ;

---

for a VAR( $p$ )-model, using the sample covariances of the manifest variables [see (3.1)]. We present below the matrix formulation of this problem which allows an easy implementation in CVX (Grant and Boyd, 2010). The mathematical derivation of the matrix formulation from the information theoretic formulation can be found in Songsiri, Dahl and Vandenberghe (2010), Avventi, Lindquist and Wahlberg (2013).

**First Maximum Entropy Problem** [ME<sub>I</sub>( $\hat{\mathbf{R}}$ )]:

$$\begin{aligned} \min_{\mathbf{X}} \quad & \text{tr}(\mathcal{T}(\hat{\mathbf{R}})\mathbf{X}) - \log \det \mathbf{X}_{00} \\ \text{s.t.} \quad & \mathbf{X} \succeq 0 \end{aligned} \tag{4.22}$$

The block Toeplitz operator  $\mathcal{T}$  is defined in (4.12). The size of the positive semidefinite matrix variable  $\mathbf{X}$  is  $K(p+1) \times K(p+1)$ . For all  $i = \overline{0, p}$ , the estimate  $\check{\mathbf{Q}}_i^{(0)}(1 : K, 1 : K)$  of the ISDM (4.15) is given by  $\text{TR}_i(\mathbf{X})$ .

In order to compute an initial value for  $\Upsilon_{\ell m}(\omega)$ , we resort to the eigenvalue decomposition (EIG) of  $\check{\mathbf{Q}}_0^{(0)}(1 : K, 1 : K)$ . More precisely, after arranging the eigenvalues of  $\check{\mathbf{Q}}_0^{(0)}(1 : K, 1 : K)$  in the decreasing order of their magnitudes, we have  $\check{\mathbf{Q}}_0^{(0)}(1 : K, 1 : K) = \mathbf{U}\mathbf{D}\mathbf{U}'$ . Then, we set  $\check{\mathbf{Q}}_0^{(0)}(K+1 : K+r, 1 : K) = \mathbf{D}^{1/2}(1 : r, 1 : r)\mathbf{U}'(1 : K, 1 : r)$  and  $\check{\mathbf{Q}}_i^{(0)}(K+1 : K+r, 1 : K) = 0$  for  $i = \overline{1, p}$ .

When the covariances  $\mathbf{C}_i$  are fixed in the maximization step of the Expectation-Maximization algorithm, the coefficients of the matrix polynomial that is the ISDM (4.15) are estimated from the solution of the following optimization problem:

**Second Maximum Entropy Problem** [ME<sub>II</sub>( $\mathbf{C}, \lambda$ )]:

$$\begin{aligned} \min_{\mathbf{X}} \quad & \text{tr}(\mathcal{T}(\mathbf{C})\mathbf{X}) - \log \det \mathbf{X}_{00} + \lambda f(\mathbf{X}) \\ \text{s.t.} \quad & \mathbf{X} \succeq 0 \\ & \text{TR}_0(\mathbf{X})(K+1 : K+r, K+1 : K+r) = \mathbf{I} \\ & \text{TR}_i(\mathbf{X})(K+1 : K+r, K+1 : K+r) = \mathbf{0}, \quad i = \overline{1, p} \end{aligned} \tag{4.23}$$

Since now we work with the full model, the size of  $\mathbf{X}$  is  $(K+r)(p+1) \times (K+r)(p+1)$ . The function  $f(\cdot)$  is the sparsity promoter defined in (4.14) and depends only on the entries of the block corresponding to the manifest variables. The equality

constraints in (4.23) guarantee that the latent variables have variance one and they are independent, given the manifest variables, corresponding to the lower right block of (4.15).

The estimates  $\left\{\check{\mathbf{Q}}_i^{(N_{\text{it}})}\right\}_{i=0}^p$  obtained after these iterations are further employed to compute  $\check{\Phi}_\lambda^{-1}(\omega)$  by using (4.15). If  $\lambda$  is large enough, then  $\check{\Phi}_\lambda^{-1}(\omega)$  is expected to have a certain sparsity pattern,  $\text{SP}_\lambda$ . Since the objective of (4.23) does not ensure exact sparsification and also because of the numerical calculations, the entries of  $\check{\Phi}_\lambda^{-1}(\omega)$  that belong to  $\text{SP}_\lambda$  are small, but not exactly zero. In order to turn them to zero, we apply a method similar to the one from (Songsiri, Dahl and Vandenberghe, 2010, Sec. 4.1.3). We firstly compute the maximum of PSC [see also (3.7)-(3.8)]:

$$\max_{\omega \in (-\pi, \pi]} \frac{\left| \left[ \check{\Phi}_\lambda^{-1}(\omega) \right]_{ab} \right|}{\sqrt{\left[ \check{\Phi}_\lambda^{-1}(\omega) \right]_{aa} \left[ \check{\Phi}_\lambda^{-1}(\omega) \right]_{bb}}}, \quad (4.24)$$

for all  $a \neq b$  with  $1 \leq a, b \leq K$ . Then  $\text{SP}_\lambda$  comprises all the pairs  $(a, b)$  for which the maximum PSC is not larger than a threshold  $\text{Th}$ . The discussion on the selection of parameters  $N_{\text{it}}$  and  $\text{Th}$  is deferred to Section 4.4.

The regularized estimate of ISDM is further improved by solving a problem similar to (4.23), but with the additional constraint that the sparsity pattern of ISDM is  $\text{SP}_\lambda$ , more precisely:

**Third Maximum Entropy Problem**  $[\text{ME}_{\text{III}}(\mathbf{C}, \text{SP})]$ :

$$\begin{aligned} \min_{\mathbf{X}} \quad & \text{tr}(\mathcal{T}(\mathbf{C})\mathbf{X}) - \log \det \mathbf{X}_{00} \\ \text{s.t.} \quad & \mathbf{X} \succeq 0 \\ & \text{TR}_0(\mathbf{X})(K+1 : K+r, K+1 : K+r) = \mathbf{I} \\ & \text{TR}_i(\mathbf{X})(K+1 : K+r, K+1 : K+r) = \mathbf{0}, \quad i = \overline{1, p} \\ & \text{TR}_i(\mathbf{X})(a, b) = 0, \quad i = \overline{0, p}, \text{ if } \text{SP}(a, b) = 0 \end{aligned} \quad (4.25)$$

This step of the algorithm has a strong theoretical justification which stems from the fact that  $\hat{\Phi}^{-1}(\omega)$  is the Maximum Entropy solution for a covariance extension problem (see (Zorzi and Sepulchre, 2016, Remark 2.1)). The number of iterations,  $N_{\text{it}}$ , is the same as in the case of the first loop.

The spectral factorization of the positive matrix trigonometric polynomial  $\hat{\Phi}_\lambda^{-1}(\omega)$  is computed by solving a semidefinite programming problem. The implementation is the same as in Chapter 3, except that the model contains latent variables. Therefore, the matrix coefficients produced by spectral factorization are altered to keep only those entries that correspond to manifest variables. The resulting VAR model is fitted to the data and then various IT criteria are evaluated. The accuracy of the selected model depends on the criterion that is employed as well as on the strategy used for generating the  $\lambda$ -values that yield the competing models. In the next section, we list the model selection rules that we apply; the problem of generating the  $\lambda$ -values is treated in Section 4.4.

As already mentioned, the estimation problem is solved for several values of  $\lambda$ :  $\lambda_1 < \lambda_2 < \dots < \lambda_L$ . From the description above we know that, for each value of the parameter  $\lambda$ ,  $\Upsilon_{\ell m}(\omega)$  gets the same initialization, which is based on (4.22). It is likely that this initialization is poor. A better approach is an *ADAPTIVE* algorithm which takes into consideration the fact that the difference  $\lambda_i - \lambda_{i-1}$  is small for all  $i = \overline{2, L}$ . This algorithm initializes  $\Upsilon_{\ell m}(\omega)$  as explained above only when  $\lambda = \lambda_1$ . When  $\lambda = \lambda_i$  for  $i = \overline{2, L}$ , the initial value of  $\Upsilon_{\ell m}(\omega)$  is taken to be the estimate of this quantity that was previously obtained by solving the optimization problem in (4.23) for  $\lambda = \lambda_{i-1}$ . The effect of the *ADAPTIVE* procedure will be investigated empirically in Section 4.4.

The newly proposed estimation method outlined in Algorithm 3 is dubbed AlgoEM.

### 4.3 Extended IT criteria

For model selection purpose, we apply the same IT criteria as those we used in the previous chapter. The effective number of parameters is computed with formula:

$$N_{\text{ef}} = \frac{K(K+1)}{2} - N_0 + p(K^2 - 2N_0), \quad (4.26)$$

where  $N_0$  is the number of zeros in the lower triangular part of  $\text{SP}_\lambda$ . With the notational conventions from Algorithm 3, we write  $\text{ITC}(\text{Data}; \text{SP})$  instead of  $\text{ITC}$ .

A common feature of all the criteria presented in Section 3.5 is that they do not take into consideration how big is the family of the competing candidates. This might be problematic because the total number of possible sparsity patterns is as large as  $2^{\bar{K}}$ , where  $\bar{K} = K(K-1)/2$ . The solutions proposed in the previous literature for circumventing this difficulty are called *extended* IT criteria. We show below how these criteria can be applied for selecting the sparsity pattern. In this context, we also propose a novel variant of RNML.

First we write down the expression of the extended SBC proposed in Chen and Chen (2008). In the statistical literature, this criterion is named EBIC (Extended Bayesian Information Criterion):

$$\text{EBIC}(\text{Data}; \text{SP}) = \text{SBC}(\text{Data}; \text{SP}) + 2\gamma \log \binom{\bar{K}}{N_0}, \quad (4.27)$$

where  $\gamma \in [0, 1]$ . It is evident that EBIC is equivalent to SBC when  $\gamma = 0$ . More interestingly, Wallace (2005) and Roos, Myllymäki and Rissanen (2009) rely on arguments from information theory for justifying the use of a penalty term which counts the number of models that have the same number of parameters. For our problem, this is equivalent to choosing  $\gamma = 1$  in (4.27). Because this is also the value of  $\gamma$  that we use in this work for evaluating EBIC, we explain briefly the significance of the supplementary penalty term. The key point is to consider a scenario in which Data should be transmitted losslessly from an encoder to a decoder by employing the model given by SP. According to Roos, Myllymäki and Rissanen (2009), the first step is to transmit the value of  $N_0$ . The assumption that all possible values of  $N_0$  are equally probable leads to the conclusion that the code length for  $N_0$  is  $-\log(1/(\bar{K} + 1)) = \log(\bar{K} + 1)$ . As this quantity is the same for all models, it can be neglected. Then the decoder should be informed about the actual locations of the zeros in the sparsity pattern SP. Since the list of all sparsity patterns for a given  $N_0$  is known by both the

encoder and the decoder, all that remains is to send to the decoder the index of SP in this list. Under the hypothesis that all the sparsity patterns in the list are equally probable, the code length for the index is  $\log \binom{\bar{K}}{N_0}$ . In (4.27), this quantity is multiplied by two because of the scaling factor used in the definition of  $\text{SBC}(\text{Data}; \text{SP})$  [see (3.15)].

Another formulation of EBIC was introduced by Foygel and Drton (2010) for finding the graphical structure of a Gaussian model (static case), in the situation when the number of variables and the number of observations grow simultaneously. After modifying the criterion from Foygel and Drton (2010) by replacing the number of edges of the graph with  $N_{\text{ef}}$ , we get the following formula:

$$\text{EBIC}_{\text{FD}}(\text{Data}; \text{SP}) = \text{SBC}(\text{Data}; \text{SP}) + 4\gamma N_{\text{ef}} \log K, \quad (4.28)$$

where  $\gamma$  has the same significance as in (4.27), and again we take  $\gamma = 1$ . Remark that the term  $4\gamma N_{\text{ef}} \log K$  grows when  $N_0$  decreases; the term in (4.27),  $2\gamma \log \binom{\bar{K}}{N_0} = 2\gamma \log \left( \frac{\bar{K}}{\bar{K} - N_0} \right)$ , does not have the same property.

Relying on the asymptotic equivalence between RNML and SBC (see Section 2.5), we alter RNML by adding half of the extra penalty from (4.28); the scaling factor is needed because the ratio between the GOF term in (3.16) and the GOF term in (3.15) tends to 1/2 when  $T \rightarrow \infty$ . The new criterion, which is dubbed  $\text{RNML}_{\text{FD}}$ , has the following expression:

$$\text{RNML}_{\text{FD}}(\text{Data}; \text{SP}) = \text{RNML}(\text{Data}; \text{SP}) + 2N_{\text{ef}} \log K. \quad (4.29)$$

For all the selection rules listed above, the best model is the one which minimizes the value of the criterion. For evaluating the performance of IT criteria, we conduct an empirical study. The main results of this study are reported in the next section.

## 4.4 Experimental results

### 4.4.1 Artificial data

In our simulations, the order of the VAR model in (4.1) is taken to be one, the number of manifest variables is  $K = 15$ , and there is one single latent variable ( $r = 1$ ). Let KS denote the number of non-zero entries in the lower triangular part of the sparsity pattern of size  $K \times K$ . We consider three different values for KS: 2, 3 and  $K$ . Remark that the larger is KS, less sparse is the ISDM. The locations of the non-zero entries for each value of KS are graphically represented in Figure 4.1.

When generating the ISDM, all the matrices  $\{\mathbf{Q}_i\}_{i=0}^p$  in (4.2) have only ones on their main diagonals. The entries of the  $K \times K$  upper-left block of the matrix  $\mathbf{Q}_i$ , which should be non-zero according to Figure 4.1, are equal to  $0.5/(i + 1)$ . Additionally, the entries on the last row and on the last column of  $\mathbf{Q}_i$ , except the one on the main diagonal, are equal to  $0.3/(i + 1)$ . Integer multiples of the  $\kappa \times \kappa$  identity matrix are added to  $\mathbf{Q}_0$  until the resulting ISDM is positive definite. Furthermore, the spectral factorization is applied in order to obtain the matrix coefficients of the VAR-model from the ISDM (see Section 2.6.2 for more details). Hence, for each value of KS, one single VAR-model is produced and this is used for generating  $N_{\text{tr}} = 10$   $\kappa$ -variate time series of length  $T = 50000$ . To this end, we utilize Matlab functions from the package available at the address <http://climate-dynamics.org/software/#arfit>. After discarding from each time series the component corresponding to the latent variable, the simulated data are used for evaluating the performance of AlgoEM.

### 4.4.2 Settings for AlgoEM

The order of VAR, as well as the number of latent variables, is assumed to be known. The parameter  $\lambda$  takes values on a regular grid defined on the interval  $[10^{-3}, 10^{-1}]$ , for which the grid step is  $10^{-3}$ . It follows that the total number of

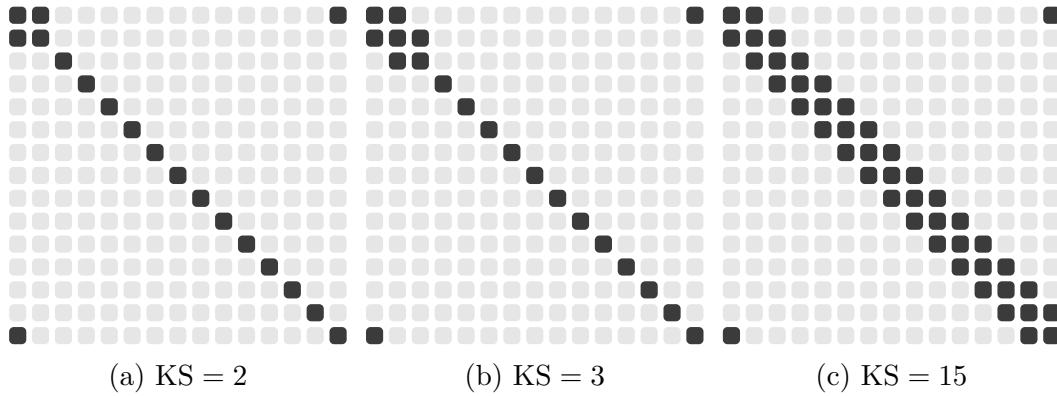


Figure 4.1: Sparsity patterns for the ISDM of the generated data: KS is the number of non-zero entries in the lower triangular part of SP. The black dots represent the locations of the non-zero entries, whereas the light grey dots are the zero entries.

values for  $\lambda$  is  $L = 100$ . The threshold  $\text{Th}$ , which is used in conjunction with (4.24) in order to get the estimated sparsity pattern, equals  $10^{-3}$ .

We are interested to evaluate the impact of the adaptive initialization procedure that was introduced in Section 4.2. This is why we run AlgoEM with and without this procedure, for all the time series we have generated. For each time series and for each value of  $\lambda$  on the grid, the estimated  $\text{SP}_\lambda$  is compared to the true sparsity pattern. The comparison reduces to computing the distance between the two sparsity patterns, which is given by the number of positions below the main diagonal where the patterns differ. For the case  $\text{KS} = 15$ , statistics related to this distance are presented in Figure 4.2. Remark that values of  $\lambda$  close to zero lead to estimated patterns which are not sparse. As expected, this happens disregarding if the adaptive procedure is applied or not. The use of the procedure has the positive effect that, for a large range of  $\lambda$ -values, the estimated patterns are close to the true one. The same is true for both  $\text{KS} = 2$  and  $\text{KS} = 3$ , which makes us apply the adaptive procedure in all the experiments outlined below.

The results reported in Figure 4.2 are obtained by taking the number of iterations to be  $N_{\text{it}} = 4$ . As the computational burden of the algorithm depends strongly on  $N_{\text{it}}$ , we investigate the effect of reducing the number of iterations

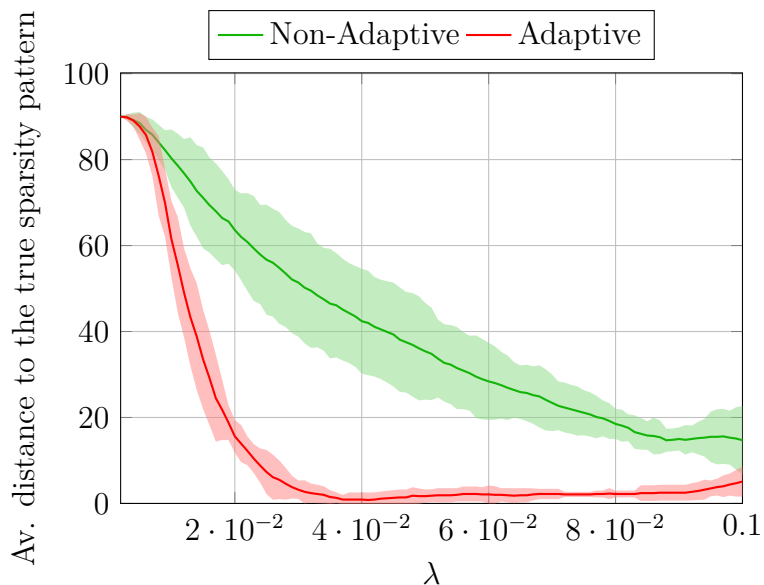


Figure 4.2: Results for VAR-models with  $KS = 15$ , for which the true sparsity pattern is shown in Figure 1c. With the convention that  $dist$  denotes the distance between the estimated sparsity pattern and the true one, we plot mean  $\pm 1$  standard deviation ( $dist$ ) versus the parameter  $\lambda$ . The statistics are computed from  $N_{tr} = 10$  trials, for both the *adaptive* and the *non-adaptive* case.

to  $N_{it} = 3$  and  $N_{it} = 2$ , respectively. In each case, the evaluation of performance is done by an oracle having complete knowledge about the true sparsity pattern. From the set of sparsity patterns produced when applying AlgoEM to a particular time series, the oracle selects the one which is closest to the true sparsity pattern. The closeness is measured by the distance defined above. The average distances computed from  $N_{tr} = 10$  trials are plotted in Figure 4.3. We can see in the figure that, in the case when  $KS = 2$  and AlgoEM performs only two iterations, the true sparsity pattern is always in the set of the candidates produced by the algorithm and this makes the average distance to be zero. In general, all the results shown in the figure are good as the average distance is smaller than one in all cases. Since the increase of  $N_{it}$  does not guarantee the improvement in performance, we take  $N_{it} = 2$  for reducing the complexity of the algorithm.

It is clear from the description of Algorithm 3 that  $N_{it}$  is the same for the two major loops of AlgoEM. We name the first loop MEEM(Pen) and the second

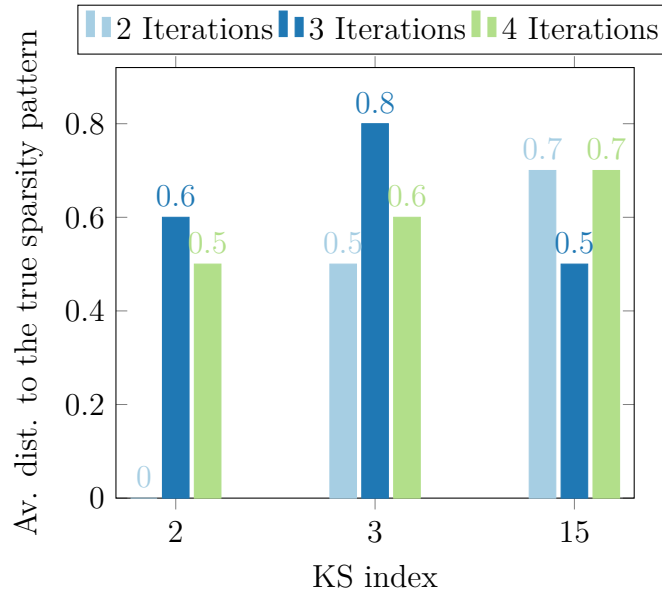


Figure 4.3: Impact of  $N_{it}$  on the performance of AlgoEM: Evaluation is done by replacing in AlgoEM the IT criterion with an oracle having full knowledge about the true sparsity pattern. For each KS and for each  $N_{it}$  we run  $N_{tr} = 10$  trials for calculating the average distance between the true sparsity pattern and the sparsity pattern selected by oracle.

one MEEM(Con). Obviously, MEEM is the acronym for Maximum Entropy Expectation-Maximization, Pen stands for *penalized* settings, and Con means *constrained* settings. Because we want to quantify the influence of MEEM(Con) on the accuracy of the estimation, we show in Figure 4.4 the average distances to the true pattern, when the estimates are produced by MEEM(Pen) and by MEEM(Con), respectively. This time we do not report results obtained only when an oracle is used for selecting the sparsity pattern, but also for the case when the selection is done with the IT criteria defined in Section 4.3. We can observe that  $AIC_c$  and FPE perform very poorly, whereas both  $EBIC_{FD}$  and  $RNML_{FD}$  are very good. The fact that  $RNML_{FD}$  is superior to  $RNML$  demonstrates the importance of the extra-term in (4.29). Remark also that  $EBIC_{FD}$  is more accurate than SBC, while EBIC has the same level of performance as SBC. The most important conclusion is that removing MEEM(Con) from AlgoEM does not deteriorate the final outcome.

The next step is to compare AlgoEM with the algorithm that solves the op-

timization problem in (4.13). We use the implementation from [Liegeois et al. \(2015\)](#), which is publicly available at the address <https://drive.google.com/file/d/0BykD206uX6KjSGlkQTRYWFBZGM/view>, and we call it AlgoSL. The name comes from the fact that the method in (4.13) is *sparse plus low-rank*.

### 4.4.3 Comparison of AlgoEM and AlgoSL

In [Liegeois et al. \(2015\)](#), the sparsity pattern for a given pair of parameters  $(\lambda, \gamma)$  is found by solving (4.13); [Zorzi and Sepulchre \(2016\)](#) uses further this sparsity pattern as an initialization for a constrained optimization problem [see also the discussion below equation (4.25)]. In both cases, a set of sparsity patterns is generated by choosing various values for the parameters  $\lambda$  and  $\gamma$ . For selecting the best one, they consider an alternative to IT criteria, which is dubbed *score function* (SF). The key point is that, when using SF, it is not needed to compute explicitly the matrix coefficients of the VAR-model. More details are provided below.

Let  $\hat{\Phi}_m$  be an estimate of  $\Phi_m$  in (4.4), which is constrained to have a certain sparsity pattern, SP. Using an idea from [Avventi, Lindquist and Wahlberg \(2013\)](#), [Zorzi and Sepulchre \(2016\)](#) suggests to employ Data for computing the correlogram  $\hat{\Phi}_m^c$  (with Bartlett window) ([Stoica and Moses, 2005](#)), and then to evaluate the relative entropy rate:

$$\mathbb{D}(\hat{\Phi}_m^c || \hat{\Phi}_m) = -\frac{1}{4\pi} \int_{-\pi}^{\pi} \log \det \Omega(\omega) + \text{tr}(\mathbf{I} - \Omega(\omega)) d\omega, \quad (4.30)$$

where  $\Omega(\omega) = \hat{\Phi}_m^c(\omega) \hat{\Phi}_m^{-1}(\omega)$  for all  $\omega \in (-\pi, \pi]$ . A discussion about this formula can be found in Section 2.6.2. It is worth noting that this index belongs to the Tau divergence family ([Zorzi, 2015b](#)) and to the Beta divergence family ([Zorzi, 2015a](#)). It has been proven in these references that the use of (4.30) in the formulation of the Maximum Entropy problem leads to the simplest solution, in the sense of the minimum McMillan degree.

The most important is that, in [Zorzi and Sepulchre \(2016\)](#),  $\mathbb{D}(\hat{\Phi}_m^c || \hat{\Phi}_m)$  is utilized for quantifying the adherence of the model to the data. The complexity

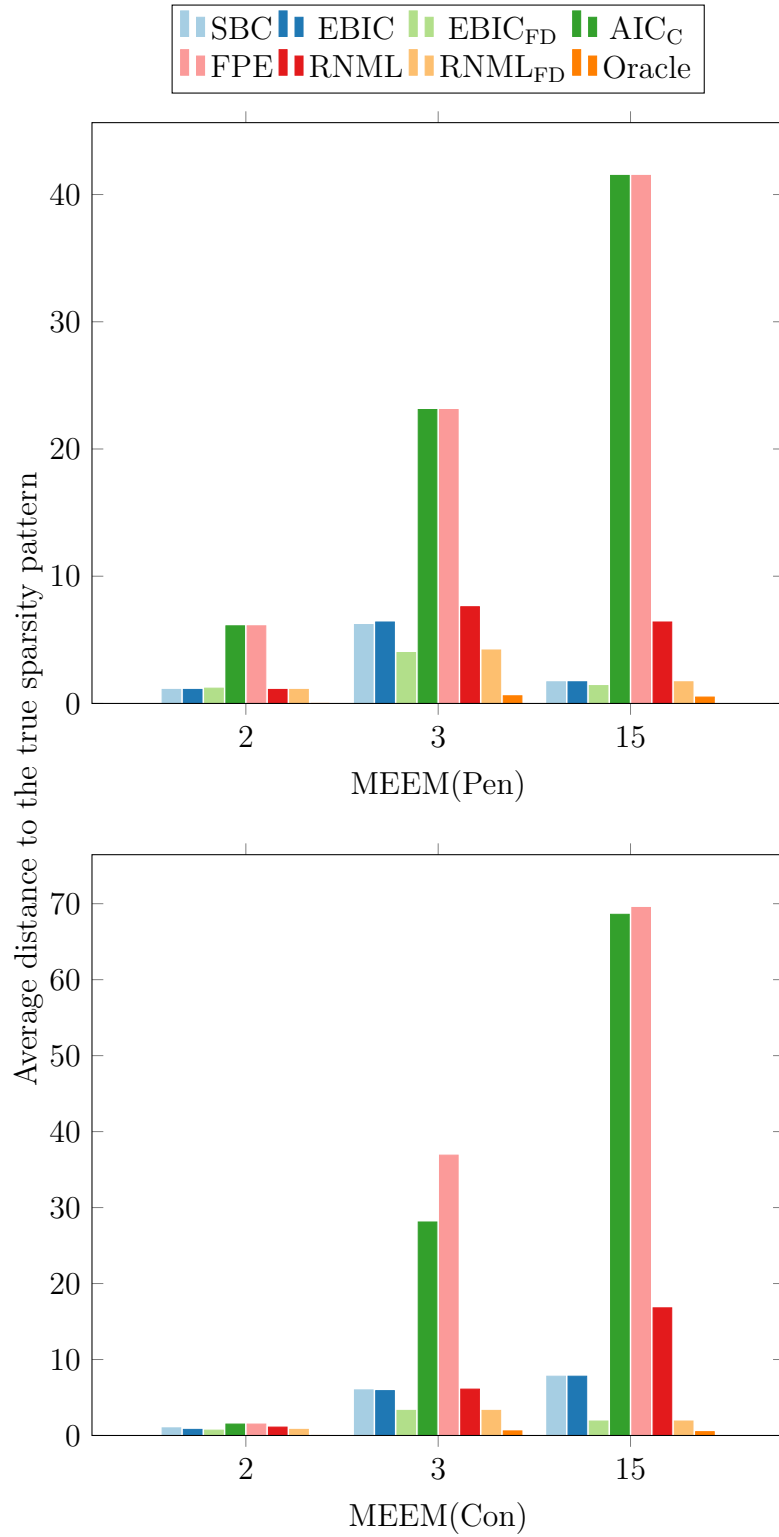


Figure 4.4: Performance of IT criteria compared to that of an oracle: (a) Only the first major loop, MEEM(Pen), of AlgoEM is executed; (b) Both MEEM(Pen) and MEEM(Con) are executed.

of the model is determined by  $N_e$ , the total number of edges in the graph (including the edges that connect the latent variables to the manifest variables). For instance, if the model has a single latent variable, then  $N_e = K(K+1)/2 - N_0$ . Note that  $N_0$  has the same significance as in (4.26). It follows that  $N_{ef} = N_e + p(K^2 - 2N_0)$ , which leads to the conclusion that  $N_{ef} > N_e$ , and the difference between the two quantities increases when the order of the model raises. If  $p = 1$  and the number of latent variables is at least three ( $r \geq 3$ ), then  $N_{ef} < N_e$  when  $N_0$  is large.

The score functions given in Zorzi and Sepulchre (2016) are:

$$\log \text{SF}_1(\text{Data}; \text{SP}) = \log \mathbb{D}(\hat{\Phi}_m^c || \hat{\Phi}_m) + \log N_e, \quad (4.31)$$

$$\text{SF}_2(\text{Data}; \text{SP}) = \mathbb{D}(\hat{\Phi}_m^c || \hat{\Phi}_m) + \frac{N_e}{T}, \quad (4.32)$$

$$\text{SF}_3(\text{Data}; \text{SP}) = \mathbb{D}(\hat{\Phi}_m^c || \hat{\Phi}_m) + \frac{N_e \log T}{T}, \quad (4.33)$$

where  $\mathbb{D}(\cdot || \cdot)$  is defined in (4.30). The formula of  $\text{SF}_1$  is log-transformed for ease of reading.

We apply AlgoSL to all the time series we have simulated (see again Section 4.4.1). In our experiments, we consider the pairs  $(\lambda, \gamma)$  for which  $\lambda \in \{0.1, 0.2, \dots, 0.6\}$  and  $\gamma \in \{0.01, 0.02, \dots, 0.5\}$ . For the selection of the sparsity pattern, we do not use only the score functions in (4.31)-(4.33), but we also employ the IT criteria from Section 4.3. The results are shown in Figure 4.5. The best criterion is  $\text{RNML}_{\text{FD}}$ , which is able to find the true sparsity pattern in all the experiments; the second best is  $\text{EBIC}_{\text{FD}}$ . The comparison of the plots in Figure 4.5 to those in Figure 4.4 leads to the conclusion that AlgoSL works better than AlgoEM when  $\text{RNML}_{\text{FD}}$  is employed for selecting the model. For understanding these results, we should take into consideration two important aspects: (i) Oracle gives perfect results for all KS in the case of AlgoSL, but not in the case of AlgoEM; to some extent this is due to the fact that 300  $(\lambda, \gamma)$ -pairs allow to produce a better set of candidates for AlgoSL than the one generated by 100  $\lambda$ -values for AlgoEM; (ii) The score functions have been used in Zorzi

and Sepulchre (2016) for time series of hundreds of samples, whereas the size of the time series we simulated is much larger ( $T = 50000$ ).

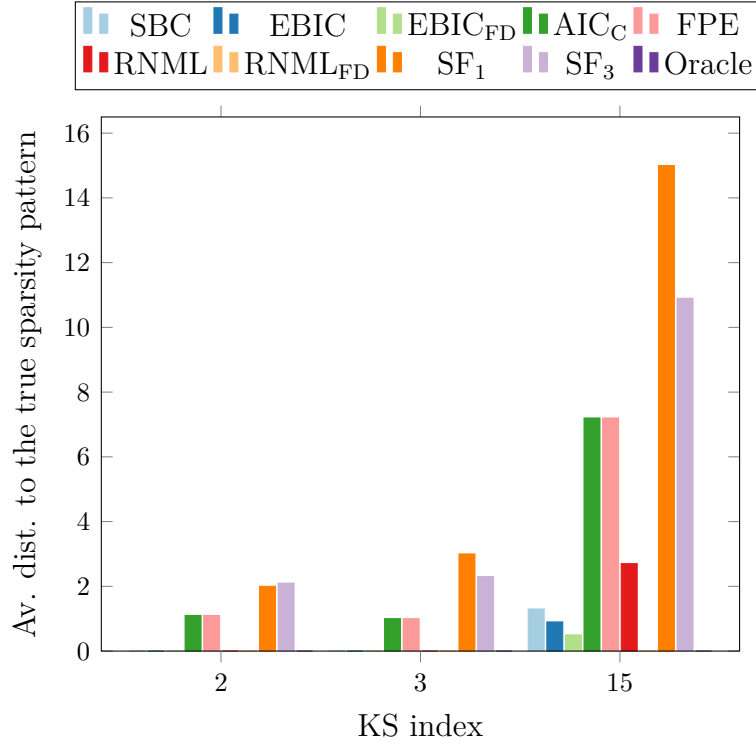


Figure 4.5: Estimation results obtained when AlgoSL is applied to the same time series which have been used to evaluate the performance of AlgoEM in Figure 4.4. For selection of the sparsity pattern, we employ the score functions in (4.31)-(4.33) and the IT criteria from Section 4.3. Score function  $SF_2$  is not shown in the graph because it leads to large values of the average distance: 102.4 for  $KS = 2$ , 101.5 for  $KS = 3$ , and 89.1 for  $KS = 15$ .

In order to clarify the second aspect, we conduct an experiment with simulated time series for which  $T = 500$ . The simulation procedure is the same as in Section 4.4.1 and all the settings are the same, except that the non-zero entries of the matrices  $\{\mathbf{Q}_i\}_{i=0}^p$  (which are not located on the main diagonal) have values that are fifty times larger. For AlgoEM, we only change the uniform grid for  $\lambda$ , which this time takes values on the interval  $[10^{-3}, 10^{-1}]$ ; the grid step is  $10^{-3}$ . Remark that the total number of values for  $\lambda$  is the same as before (100), and the new grid is obtained from the old one by applying a scaling factor suggested by Basu and Michailidis (2015). Based on some empirical evidence, we use the parameters  $\lambda \in \{0.02, 0.03, \dots, 0.3\}$  and  $\lambda\gamma \in \{0.01, 0.015, \dots, 0.2\}$

for AlgoSL. All IT criteria and all the score functions are used for both AlgoEM and AlgoSL, and the estimation results are reported in Figure 4.6. In general, they are worse than the results for large sample size ( $T = 50000$ ), and the difference is more evident when  $KS = 15$ . This shows that, for small  $T$ , it might be difficult to recover the true structure of ISDM when it is not sparse enough. It is encouraging that, even for  $KS = 15$ , oracle used in conjunction with AlgoSL finds the true pattern in all trials.

To gain more insight, we give in Figure 4.7 and Figure 4.8 a graphical representation of all the distances which are computed by oracle for a time series whose ISDM has  $KS = 15$ . It follows from the way in which we have chosen the experimental parameters that the total number of such distances calculated for AlgoSL is more than eleven times larger than the number of distances for AlgoEM. We should keep in mind that we need to compute the distance from the pattern estimated by each candidate in the list to the true sparsity pattern. After a laborious process of selecting the values of  $\lambda$  and  $\lambda\gamma$ , we ended-up with a two-dimensional grid which yields a good number of estimated patterns that are identical to the true pattern (see the black area inside the square shown in Figure 4.8). The behaviour of AlgoEM is different: For a large range of  $\lambda$ -values, the  $K \times K$  upper-left block of the estimated pattern is equal to the identity matrix, which makes the distance to the true pattern to be equal to  $KS$ . This happens when the number of latent variables used in AlgoEM equals the true number ( $r = 1$ ) as well as when  $r = 2$  is utilized in estimation. However, it is evident in Figure 4.7 that the estimation results are better when  $r = 1$ . If the number of latent variables is not known a priori, one possibility might be to use an ITC for selecting it. Note that the definition of  $N_{\text{ef}}$  in (4.26) should be modified. The score functions given in (4.31)-(4.33) do not need to be altered in order to be employed in selection of the number of latent variables.

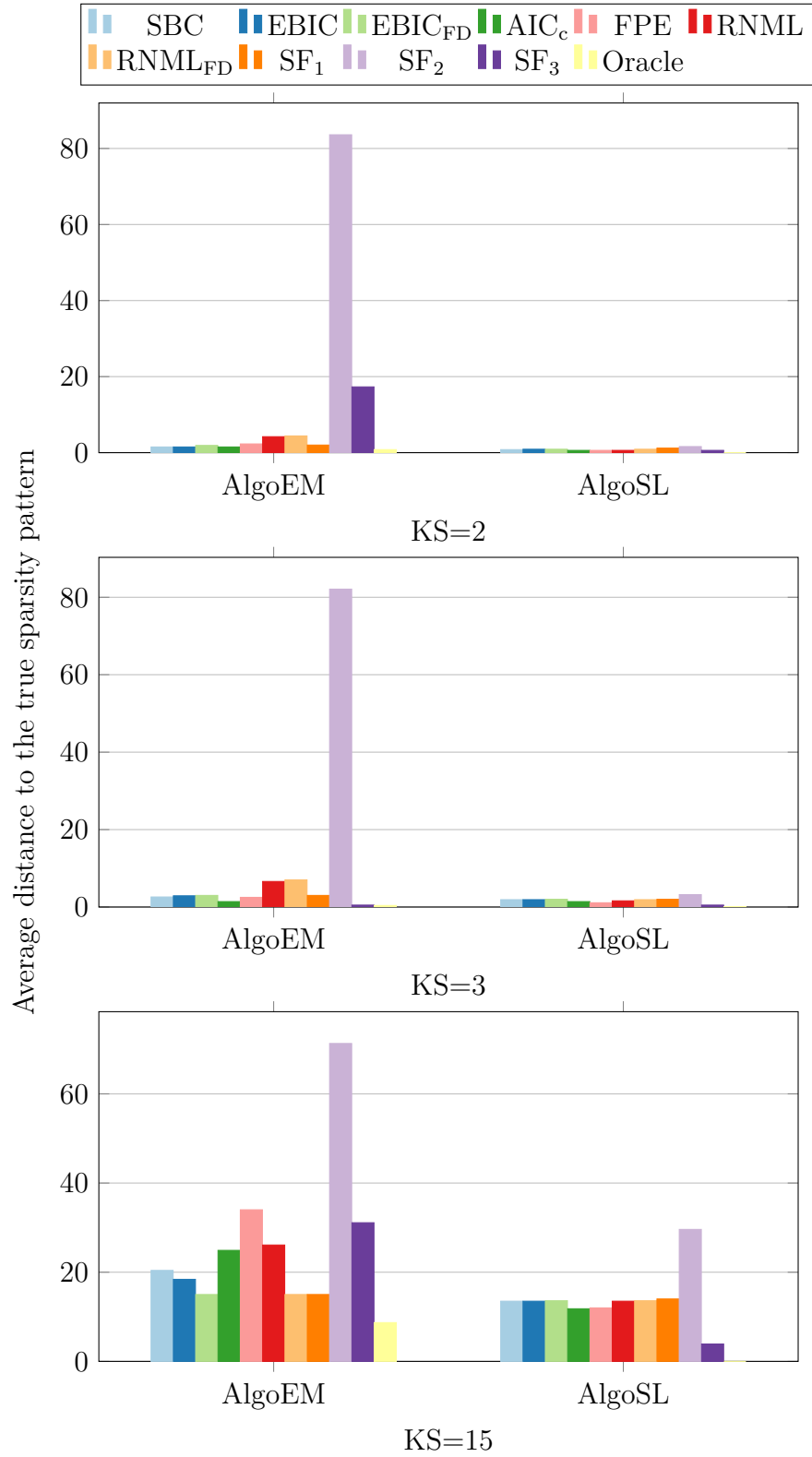


Figure 4.6: Estimation results when sample size is small ( $T = 500$ ).

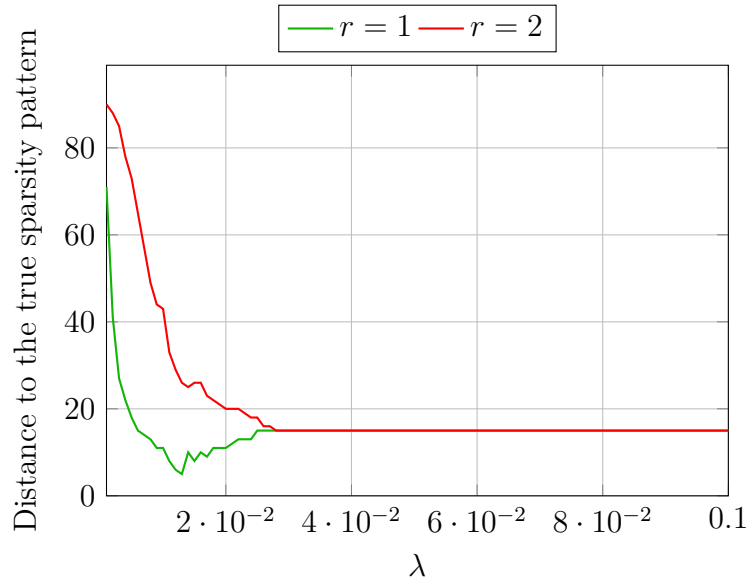


Figure 4.7: Distances computed for sparsity patterns which are estimated from a time series of length  $T = 500$ ; the value of KS for the true sparsity pattern is 15: AlgoEM ( $r$  is the number of latent variables used in estimation).

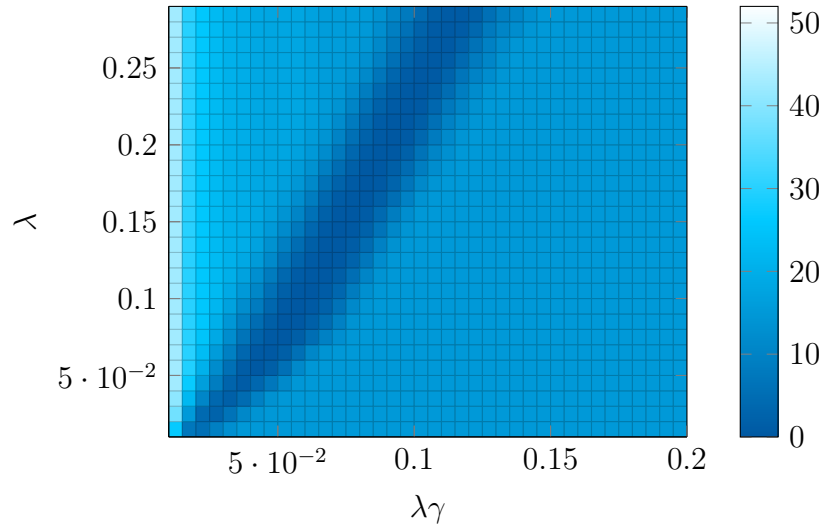


Figure 4.8: Distances computed for sparsity patterns which are estimated from a time series of length  $T = 500$ ; the value of KS for the true sparsity pattern is 15: AlgoSL.

#### 4.4.4 Real-world data

In addition to the experiments with simulated data, we test the capabilities of AlgoEM on multivariate time series of daily stock markets indices at closing time, which can be downloaded from the following address: <http://au.mathworks>.

[com/matlabcentral/fileexchange/48611-international-daily-stock-return-data-for-system-identification](http://matlabcentral/fileexchange/48611-international-daily-stock-return-data-for-system-identification). This dataset was produced by pre-processing the original data from <http://finance.yahoo.com>, which have been measured from 4th January 2012 to 31st December 2013. We refer to [Zorzi and Sepulchre \(2016\)](#) for details regarding the pre-processing. The time series is K-variate with  $K = 22$ , and the sample size is  $T = 518$ .

For each component of the time series, we give the name of the country where was measured and we write in parentheses the acronym used in this analysis: Australia (AU), New Zealand (NZ), Singapore (SG), Hong Kong (HK), China (CH), Japan (JA), Korea (KO), Taiwan (TA), Brazil (BR), Mexico (ME), Argentina (AR), Switzerland (SW), Greece (GR), Belgium (BE), Austria (AS), Germany (GE), France (FR), Netherlands (NL), United Kingdom (UK), United States (US), Canada (CA), Malaysia (MA). The official names of the price indices can be found in [Zorzi and Sepulchre \(2016\)](#).

Similar to [Zorzi and Sepulchre \(2016\)](#), we take the order of VAR-model to be  $p = 1$ , and we assume that there is a single latent variable ( $r = 1$ ). For AlgoEM, we have  $N_{it} = 4$  and  $\lambda$  takes values on a uniform grid on the interval  $[2 \times 10^{-3}, 2 \times 10^{-1}]$ , for which the grid step is  $2 \times 10^{-3}$ . When  $AIC_c$  is used for choosing the model, the results are disappointing because the selected sparsity pattern contains very few zeros. The same type of outcome is also obtained when applying  $SF_2$  or  $SF_3$ . It is interesting that  $SF_1$  as well as six different IT criteria ( $SBC$ ,  $EBIC$ ,  $EBIC_{FD}$ ,  $FPE$ ,  $RNML$ ,  $RNML_{FD}$ ) select exactly the same sparsity pattern. This one is compared in Figure 4.9 to the sparsity pattern given in [Zorzi and Sepulchre \(2016\)](#), for the same data set. Observe that the conditional independence graph yielded by AlgoEM has only four edges which connect manifest variables: Three of them connect vertices corresponding to Asian markets,  $(HK, CH)$ ,  $(HK, JA)$ ,  $(JA, KO)$ , and the fourth one connects two European markets,  $(GR, AS)$ . This is different from the conditional graph in [Zorzi and Sepulchre \(2016\)](#), where all the edges between manifest variables connect only the European markets (with the exception of Greece). An in-depth

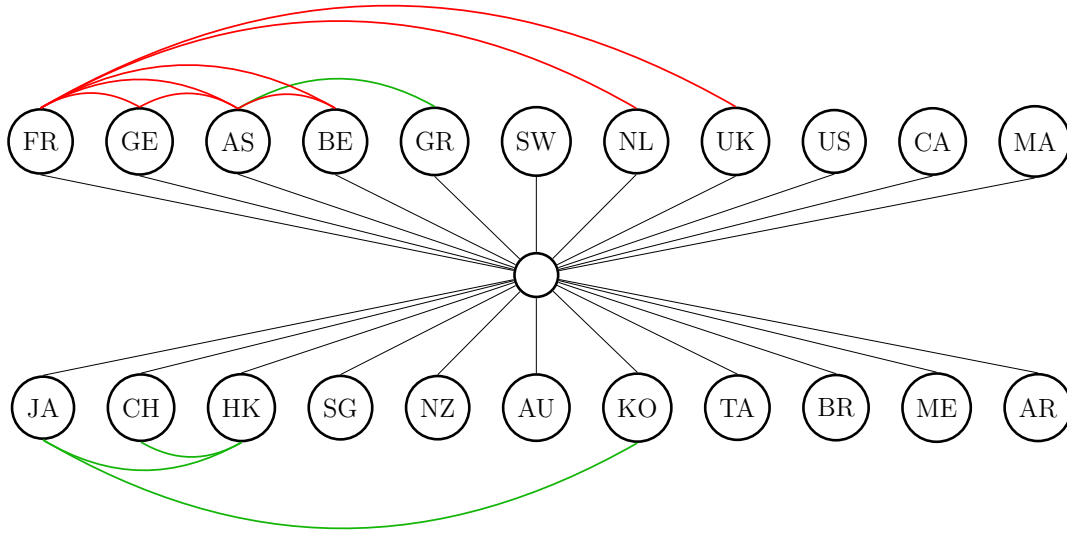


Figure 4.9: International stock markets data: Comparison between the sparsity pattern for manifest variables from [Zorzi and Sepulchre \(2016\)](#) (red) and the pattern produced by AlgoEM (green) when either  $SF_1$  or one of the following IT criteria is used: SBC, EBIC,  $EBIC_{FD}$ , FPE, RNML,  $RNML_{FD}$ .

analysis of the two graphs is beyond the interest of this thesis.

## 4.5 Summary

The main motivation for the work presented in this chapter is the solution proposed in [Lauritzen and Meinhausen \(2012\)](#) for the static case, as an alternative to the method from [Chandrasekaran, Parrilo and Willsky \(2012\)](#). We have shown how the estimation method from [Lauritzen and Meinhausen \(2012\)](#) can be generalized for the dynamic case. The resulting algorithm is dubbed AlgoEM. We have conducted an empirical study in which we have investigated the capabilities of AlgoEM and we have also compared it with the generalization of the method from [Chandrasekaran, Parrilo and Willsky \(2012\)](#), which was proposed in [Zorzi and Sepulchre \(2016\)](#). It is important to emphasize that the two methods for latent-variable autoregressive models, which we have compared, have some common features: apply the Maximum Entropy principle, use convex optimization, and generate a set of candidate models from which the best one is selected by a

certain rule. In the case of AlgoEM, the set of the candidates depends strongly on the parameter  $\lambda$ , which is chosen by the user. For the method in [Zorzi and Sepulchre \(2016\)](#), the user should choose two parameters,  $\lambda$  and  $\gamma$ . Based on our experience, the selection of the two parameters is much more difficult than the selection of the single parameter for AlgoEM. Another important aspect is how to pick-up the winner from the competing models. In [Zorzi and Sepulchre \(2016\)](#), this was restricted to the use of score functions. We have demonstrated empirically that the IT criteria might be an option to consider. Especially when the sample size is large, it is recommended to employ the criterion  $\text{RNML}_{\text{FD}}$  which we have introduced in this work.

# Chapter 5

## Final remarks

In the first chapter, we introduced the RNML criterion for VAR-order selection. In our theoretical analysis, we proved that the criterion is strongly consistent. The results reported for experiments with simulated data demonstrate its abilities in estimating properly the order when the sample size is small or moderate. It can be used as part of an algorithm which firstly estimates the order and then identifies the sparsity pattern of ISDM.

In the second chapter, we have proposed a family of algorithms for inferring the conditional independence graph of a  $\text{VAR}(p)$ -model for  $K$ -variate time series. Our theoretical and empirical results demonstrate that the algorithms from this family can be used when  $p \leq 20$ ,  $K \leq 10$  and  $K \times p \leq 100$ . Thus far, the methods which rely on convex optimization and do not ask the user to make subjective choice of parameters have been suitable only for much smaller values of  $p$  and  $K$ . Another important feature of our method is the guaranteed stability of the fitted model. As part of this work, the new RNML criterion for VAR- order selection was derived and its consistency property was proven. Given its good performance for small and moderate sample sizes, RNML can be also applied in other signal processing problems, not necessarily related to the inference of graphical models.

In the last chapter, we have shown how the estimation method from [Lauritzen and Meinhausen \(2012\)](#) can be generalized for the dynamic case. The resulting algorithm is dubbed AlgoEM. We have conducted an empirical study in which we have investigated the capabilities of AlgoEM and we have also compared it with the generalization of the method from [Chandrasekaran, Parrilo and Willsky \(2012\)](#), which was proposed in [Zorzi and Sepulchre \(2016\)](#). It is important to emphasize that the two methods for latent-variable autoregressive models, which we have compared, have some common features: apply the Maximum Entropy principle, use convex optimization, and generate a set of candidate models from which the best one is selected by a certain rule. In the case of AlgoEM, the set of the candidates depends strongly on the parameter  $\lambda$ , which is chosen by the user. For the method in [Zorzi and Sepulchre \(2016\)](#), the user should choose two parameters,  $\lambda$  and  $\gamma$ . Based on our experience, the selection of the two parameters is much more difficult than the selection of the single parameter for AlgoEM. Another important aspect is how to pick-up the winner from the competing models. In [Zorzi and Sepulchre \(2016\)](#), this was restricted to the use of score functions. We have demonstrated empirically that the IT criteria might be an option to consider. Especially when the sample size is large, it is recommended to employ the criterion  $\text{RNML}_{FD}$  which we have introduced in this work.

All experiments can be reproduced by using the Matlab code which can be downloaded from <https://www.stat.auckland.ac.nz/~cgiu216/PUBLICATIONS.htm>.

# Bibliography

- Abdelwahab, A., O. Amor and T. Abdelwaheb. 2008. “The analysis of the interdependence structure in international financial markets by graphical models.” *International Research Journal of Finance and Economics* 15:291–306.
- Akaike, H. 1971. “Autoregressive model fitting for control.” *Annals of the Institute of Statistical Mathematics* 23:163–180.
- Akaike, H. 1974. “A new look at the statistical model identification.” *IEEE Transactions on Automatic Control* AC-19:716–723.
- Artin, E. 1964. *The Gamma function*. Holt, Rinehart and Winston, Inc.
- Avventi, E., A.G. Lindquist and B. Wahlberg. 2013. “ARMA Identification of Graphical Models.” *IEEE Transactions on Automatic Control* 58:1167–1178.
- Bach, F.R. and M.I. Jordan. 2004. “Learning Graphical Models for Stationary Time Series.” *IEEE Transactions on Signal Processing* 52(8):2189–2199.
- Basu, S. and G. Michailidis. 2015. “Regularized estimation in sparse high-dimensional time series models.” *The Annals of Statistics* 43(4):1535 – 1567.
- Boyd, S., N. Parikh, E. Chu, B. Peleato and J. Eckstein. 2010. “Distributed optimization and statistical learning via the alternating direction method of multipliers.” *Foundations and Trends in Machine Learning* 3(1):1–122.
- Brillinger, D.R. 1996. “Remarks concerning graphical models for time series and point processes.” *Revista de Econometria* 16:1–23.

- Brockwell, P. J. and R. A. Davis. 1991. *Time Series: Theory and Method*. Springer Verlag.
- Buhlmann, Peter and Sara van der Geer. 2011. *Statistics for High-dimensional Data*. Springer Series in Statistics Springer Berlin Heidelberg.
- California Environmental Protection Agency. 2016. “Air Quality and Meteorological Information System (AQMIS).”  
**URL:** <http://www.arb.ca.gov/aqd/aqcd/aqcd.htm>
- Cavanaugh, J.E. 1999. “A large-sample model selection criterion based on Kullback’s symmetric divergence.” *Statistics and Probability Letters* 42:333—343.
- Chandrasekaran, V., P.A. Parrilo and A.S. Willsky. 2012. “Latent variable graphical model selection via convex optimization.” *The Annals of Statistics* 40(4):1935 – 1967.
- Chen, J. and Z. Chen. 2008. “Extended Bayesian information criteria for model selection with large model spaces.” *Biometrika* 95(3):759 – 771.
- Dahlhaus, R. 2000. “Graphical interaction models for multivariate time series.” *Metrika* 51:157–172.
- Dahlhaus, R., M. Eichler and J. Sandkühler. 1997. “Identification of synaptic connections in neural ensembles by graphical models.” *Journal of Neuroscience Methods* 77:93–107.
- Davis, R. A., P. Zang and T. Zheng. 2016. “Sparse Vector Autoregressive Modeling.” *Journal of Computational and Graphical Statistics* 25(4):1077–1096.  
**URL:** <https://doi.org/10.1080/10618600.2015.1092978>
- Dumitrescu, B. 2007. *Positive trigonometric polynomials and signal processing applications*. Springer.

- Eichler, M. 2006. Fitting graphical interaction models to multivariate time series. In *Proc. 22nd Conf. Uncertainty in Artif. Intell.* Arlington, VA, USA: AUAI Press pp. 147–154.
- Ferrante, A., C. Masiero and M. Pavon. 2012. “Time and Spectral Domain Relative Entropy: A New Approach to Multivariate Spectral Estimation.” *IEEE Transactions on Automatic Control* 57(10):2561–2575.
- Foygel, R. and M. Drton. 2010. Extended Bayesian information criteria for Gaussian graphical models. In *Advances in Neural Information Processing Systems*, ed. J.D. Lafferty, C.K.I. Williams, J. Shawe-Taylor, R.S. Zemel and A. Culotta. Vol. 23 pp. 604–612.
- Fried, R. and V. Didelez. 2003. “Decomposability and selection of graphical models for time series.” *Biometrika* 90:251–267.
- Giurcăneanu, C.D. and J. Rissanen. 2006. Estimation of AR and ARMA models by stochastic complexity. In *Time series and related topics*, ed. Hwai-Chung Ho, Ching-Kang Ing and Tze Leung Lai. Vol. 52 Institute of Mathematical Statistics Lecture Notes-Monograph Series pp. 48–59.
- Giurcăneanu, C.D., S.A. Razavi and A. Liski. 2011. “Variable selection in linear regression: Several approaches based on normalized maximum likelihood.” *Signal Processing* 91:1671–1692.
- Grant, M. and S. Boyd. 2010. “CVX: Matlab software for disciplined convex programming.” <http://cvxr.com/cvx>.
- Hirai, S. and K. Yamanishi. 2013. “Efficient Computation of Normalized Maximum Likelihood Codes for Gaussian Mixture Models With Its Applications to Clustering.” *IEEE Transactions on Information Theory* 59:7718–7727.
- Hurvich, C.M. and C.L. Tsai. 1993. “A corrected Akaike information criterion for vector autoregressive model selection.” *Journal of Time Series Analysis* 14:271–279.

- Jung, Alexander. 2015. “Learning the conditional independence structure of stationary time series: A multitask learning approach.” *IEEE Transactions on Signal Processing* 63(21):5677–5690.
- Junger, W.L. and A.P. de Leon. 2016. “Package: mtsdi,Version: 0.3.3.”  
**URL:** <http://www.inside-r.org/packages/cran/mtsdi/docs/mnimput>
- Lauritzen, S. and N. Meinhausen. 2012. “Discussion: Latent variable graphical model selection via convex optimization.” *The Annals of Statistics* 40(4):1973 – 1977.
- Lauritzen, S.L. 1996. *Graphical Models*. Oxford University Press.
- Liegeois, R., B. Mishra, M. Zorzi and R. Sepulchre. 2015. Sparse plus low-rank autoregressive identification in neuroimaging time series. In *54th IEEE Conference on Decision and Control (CDC)*. pp. 3965–3970.
- Lutkepöhl, H. 2005. *New Introduction to Multiple Time Series Analysis*. John Wiley & Sons.
- Mardia, K.V., J.T. Kent and J.M. Bibby. 1979. *Multivariate Analysis*. Academic Press.
- Matsuda, Y. 2006. “A test statistic for graphical model of multivariate time series.” *Biometrika* 93:399–409.
- McLean, J.W. and H.J. Woerdeman. 2002. “Spectral Factorizations and Sums of Squares Representations via Semidefinite Programming.” *SIAM J. Matrix Anal. Appl.* 23(3):646–655.
- Miller, A. 2002. *Subset selection in regression*. Chapman & Hall/CRC.
- Morf, M., A. Vieira and T. Kailath. 1978. “Covariance Characterization by Partial Autocorrelation Matrices.” *The Annals of Statistics* 6(3):643–648.
- Neumaier, A. and T. Schneider. 2001. “Estimation of parameters and eigenmodes of multivariate autoregressive models.” *ACM Trans. Math. Softw.* 27:27–57.

- Reinsel, G.C. 1993. *Elements of Multivariate Time Series Analysis*. Springer-Verlag.
- Rissanen, J. 2000. “MDL denoising.” *IEEE Transactions on Information Theory* 46(7):2537–2543.
- Rissanen, J. 2007. *Information and complexity in statistical modeling*. Springer Verlag.
- Rissanen, J., T. Roos and P. Myllymäki. 2010. “Model selection by sequentially normalized least squares.” *Journal of Multivariate Analysis* 101:839–849.
- Roos, T., P. Myllymäki and J. Rissanen. 2009. “MDL denoising revisited.” *IEEE Transactions on Signal Processing* 57(9):3347–3360.
- Schmidt, D.F. and E. Makalic. 2011. “Estimating the order of an autoregressive model using normalized maximum likelihood.” *IEEE Transactions on Signal Processing* 59(2):479–487.
- Schwarz, G. 1978. “Estimating the Dimension of a Model.” *Annals of Statistics* 6:461–464.
- Seghouane, A.K. 2006. “Vector autoregressive model-order selection from finite samples using Kullback’s symmetric divergence.” *IEEE Transactions on Circuits and Systems-I Regular Papers* 53:2327—2335.
- Şicleru, B.C. and B. Dumitrescu. 2013. “POS3POLY – A MATLAB Preprocessor for Optimization with Positive Polynomials.” *Optimization and Engineering* 14(2):251–273.
- Songsiri, J., J. Dahl and L. Vandenberghe. 2010. Graphical models of autoregressive processes. In *Convex Optimization in Signal Processing and Communications*, ed. D.P. Palomar and Y.C. Eldar. Cambridge Univ. Press pp. 89–116.

- Songsiri, J. and L. Vandenberghe. 2010. "Topology Selection in Graphical Models of Autoregressive Processes." *Journal of Machine Learning Research* 11:2671–2705.
- Speed, T.P. and H.T. Kiiveri. 1986. "Gaussian Markov distributions over finite graphs." *Annals of Statistics* 14(1):138–150.
- Stoica, P. and R. Moses. 2005. *Spectral analysis of signals*. Pearson Prentice Hall.
- Ting, C.-M., A.K. Seghouane, M.U. Khalid and S.-H. Salleh. 2015. "Is First-Order Vector Autoregressive Model Optimal for fMRI Data?" *Neural Computation* 27:1857–1871.
- Wallace, C.S. 2005. *Statistical and Inductive Inference by Minimum Message Length (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag.
- Wei, W.W.S. 2006. *Time Series Analysis. Univariate and Multivariate Methods*. Pearson Education, Inc.
- Wermuth, N. and E. Scheidt. 1977. "Algorithm AS 105: Fitting a covariance selection model to a matrix." *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 26(1):88 – 92.
- Wolstenholme, R.J. and A.T. Walden. 2015. "An Efficient Approach to Graphical Modeling of Time Series." *IEEE Transactions on Signal Processing* 63:3266–3276.
- Zorzi, M. 2015a. "An interpretation of the dual problem of the THREE-like approaches." *Automatica* 62:87 – 92.  
**URL:** <http://www.sciencedirect.com/science/article/pii/S0005109815003866>
- Zorzi, M. 2015b. "Multivariate Spectral Estimation Based on the Concept of Optimal Prediction." *IEEE Transactions on Automatic Control* 60(6):1647–1652.

Zorzi, M. and R. Sepulchre. 2016. “AR Identification of latent-variable graphical models.” *IEEE Transactions on Automatic Control* 61(9):2327–2340.