RENORMALIZED MAXIMUM LIKELIHOOD FOR MULTIVARIATE AUTOREGRESSIVE MODELS

Saïd Maanan[1], *Bogdan Dumitrescu*[2], *Ciprian Doru Giurcăneanu*[1]

[1]Department of Statistics
University of Auckland
Private Bag 92019, Auckland 1142, New Zealand

[2]Department of Automatic Control and Computers
University Politehnica of Bucharest
313 Spl. Independenţei, 060042 Bucharest, Romania

ABSTRACT

Renormalized maximum likelihood (RNML) is a powerful concept from information theory. We show how it can be used to derive a criterion for selecting the order of vector autoregressive (VAR) processes. 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. In our experiments, we pay a special attention to models for which the inverse spectral density matrix (ISDM) has a specific sparsity pattern. The interest on these models is motivated by the relationship between sparse structure of ISDM and the problem of inferring the conditional independence graph for multivariate time series.

Index Terms— Renormalized maximum likelihood, vector autoregressive model, order selection, maximum entropy, convex optimization

1. INTRODUCTION AND PRELIMINARIES

Problem formulation: In this study, we address the fundamental problem of estimating the order of a vector autoregressive (VAR) process.

Let $\mathbf{y}_1, \ldots, \mathbf{y}_T$ be a K-dimensional ($K > 1$) time series generated by a stationary and stable VAR process of order p°. We assume that the spacing of observation times is constant and $\mathbf{y}_t = [y_{1\,t}, \ldots, y_{K\,t}]'$, for $t = \overline{1,T}$. The symbol $[\cdot]'$ denotes transposition. The well-known difference equation of the process is [1]:

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

where $\mathbf{A}_1, \ldots, \mathbf{A}_{p^\circ}$ are matrix coefficients of size $K \times K$ and $\mathbf{u}_t = [u_{1t}, \ldots, u_{Kt}]'$ is a sequence of independently and identically distributed random K-vectors. In our derivations, we need the supplementary hypothesis that the vectors $\{\mathbf{u}_t\}_{t=1}^{T}$ are drawn from a K-variate Gaussian distribution with zero mean vector and covariance matrix $\mathbf{\Sigma} \succ 0$. Additionally, the vectors $\{\mathbf{y}_t\}_{t=1-p^\circ}^{0}$ are assumed to be constant.

Motivation of the work: In a recent series of papers (see [2, 3, 4] and the references therein), various information theoretic criteria (ITC) have been used to select the order of VAR-models for which a sparsity pattern is assumed either for the matrix $\mathbf{B} = [\mathbf{A}_1, \ldots, \mathbf{A}_{p^\circ}]'$ or for the inverse spectral density matrix (ISDM) of the process. To make clear the last point, we denote the spectral density matrix of $\mathrm{VAR}(p^\circ)$-process in (1) by $\mathbf{S}(\omega)$, where $\omega \in (-\pi, \pi]$. Its eigenvalues are bounded and bounded away from zero, uniformly for all frequencies in $(-\pi, \pi]$. It follows that the ISDM $\mathbf{S}^{-1}(\omega)$ exists for $\omega \in (-\pi, \pi]$ and has the following expression [2]:

$$\mathbf{S}^{-1}(\omega) = \mathbf{A}^{\mathrm{H}}(\omega) \mathbf{\Sigma}^{-1} \mathbf{A}(\omega) = \sum_{m=-p^\circ}^{p^\circ} \mathbf{Q}_m e^{-j\omega m}, \quad (2)$$

where $j = \sqrt{-1}$ and $(\cdot)^{\mathrm{H}}$ 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 $\mathbf{Q}_m = \sum_{i=0}^{p^\circ - m} \mathbf{A}_i' \mathbf{\Sigma}^{-1} \mathbf{A}_{i+m}$ and $\mathbf{Q}_{-m} = \mathbf{Q}_m'$. The sparse structure of ISDM is especially important in connection with the problem of inferring the conditional independence graph for the observed time series [5].

Interestingly enough, the aforementioned studies do not assume that only very few of the entries of the considered matrices are non-zero. As the high sparsity is not included in the set of assumptions, the authors of these works employ "classical" ITC: SBC - Schwarz's Bayesian Criterion [6], AIC - Akaike Information Criterion [7], AICc - "corrected" AIC [8], KIC - Kullback Information Criterion [9], KICc - "corrected" KIC [10].

In [4], a nonparametric 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, $\mathrm{VAR}(p, \mathrm{SP})$, 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 $\mathbf{B}$, then each candidate model is fitted to the data and the win-

E-mails: s.maanan@auckland.ac.nz, bogdan.dumitrescu@acse.pub.ro, c.giurcaneanu@auckland.ac.nz. This work was supported by Dept. of Statistics (UOA) Doctoral Scholarship and the Romanian National Authority for Scientific Research, CNCS - UEFISCDI, project number PN-II-ID-PCE-2011-3-0400.

ner is selected by using SBC. The results are refined in the second stage of the procedure, where SBC is applied again. The approach from [2] is more computationally intensive because a VAR-model is fitted to the data for each pair (p, SP) when all possible SP's are considered and not only a short list of candidates like in [4]. The major contribution of [2] consists in recasting the model fitting as a convex optimization problem. In [3], selection of VAR-order is discussed in the context of analysis of fMRI data for which the ratio T/K is small, but not smaller than $p+1$.

Relation to prior work: When a $\mathrm{VAR}(p)$-model is fitted to the data $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_T]'$, the normalized maximum likelihood (NML) equals the negative logarithm of [11, 12]

$$\hat{f}(\mathbf{Y}; p) = \frac{f\left(\mathbf{Y}; \hat{\mathbf{B}}(\mathbf{Y}), \hat{\mathbf{\Sigma}}(\mathbf{Y})\right)}{C_p}. \tag{3}$$

In our calculations, we use only natural logarithms and denote them by $\log(\cdot)$. In (3), the numerator is the maximum value of the likelihood function, given the measurements $\mathbf{Y}$. $\hat{\mathbf{B}}(\mathbf{Y})$ and $\hat{\mathbf{\Sigma}}(\mathbf{Y})$ are the maximum likelihood (ML) estimates of the parameters of the model. For the denominator, we have:

$$C_p = \int f\left(\mathbf{Y}; \hat{\mathbf{B}}(\mathbf{Y}), \hat{\mathbf{\Sigma}}(\mathbf{Y})\right) \mathrm{d}\mathbf{Y}, \tag{4}$$

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 [11]. 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 [11] and perform a second normalization step. The resulting formula is named $\mathrm{RNML}(\mathbf{Y}; p)$.

According to the best of our knowledge, the expression of RNML for VAR-models was not obtained so far. The very first attempt at estimating the order of *univariate* AR models by RNML is the one from [13]. The approach from [13] was further extended in [14], where the focus is still on the univariate case. We note in passing that, the method employed in [14] for evaluating the criterion does not allow the use of the second normalization step. More interestingly, the work of Schmidt and Makalic 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 [15, Section 16.5] for a tutorial review). The last one is best known in the signal processing community for its use in the normalized Whittle-Wiggins-Robinson algorithm [16]. However, none of these functions enables the calculation of the integral in (4).

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, [1, 3, 17]). We are encouraged to apply it by the experimental results reported in [18] 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 [19], 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.

Significance of the paper: (i) Derivation of RNML-formula (see Section 2); (ii) Theoretical analysis of the new criterion which shows its asymptotic equivalence with SBC. As part of this analysis we find an upper bound for the penalty term of RNML which depends on the actual measurements and not only on the sample size and the number of parameters of the model. The partial autocorrelation matrix function is instrumental in proving this result (see Section 2); (iii) Numerical examples for demonstrating the performance of RNML (see Section 3).

Note: Because of the limited space, some of the results are presented in the supplemental material [20].

2. RNML CRITERION

Main formula: To fix the ideas, we assume that a $\mathrm{VAR}(p)$-model with order $p > 0$ is fitted to the data $\mathbf{Y}$. Our main result is the following.

Proposition 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*

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

where

$$\begin{aligned} \mathrm{GOF} &= [(T - Kp - K + 1)/2] \log |\hat{\mathbf{\Sigma}}_p| \\ \mathrm{PEN}_1 &= -\log \Gamma_K[(T - Kp)/2] \\ \mathrm{PEN}_2 &= -\log \Gamma[(K^2 p)/2] \\ \mathrm{PEN}_3 &= [(K^2 p)/2] \log \mathrm{tr}\left[(\mathbf{Y}'\mathbf{Y})/T - \hat{\mathbf{\Sigma}}_p\right]. \end{aligned} \tag{5}$$

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

Proof is outlined in [20, Section 1]. The acronym GOF is employed for the goodness-of-fit term, whereas PEN_1, PEN_2 and PEN_3 are penalty terms.

As in the case of other ITC, $\mathrm{RNML}(\mathbf{Y};p)$ is evaluated for $p=\overline{p_{\min},p_{\max}}$ and $\hat{p}$ is chosen to be that particular order which minimizes the criterion. We have assumed that $p_{\min}>0$, because this is typically the case for the problem addressed in this work. For completeness, we investigate in [20, Section 2] how RNML-formula can be derived for $p=0$ (see also [12]).

Asymptotic behavior: This analysis aims to clarify the relationship between RNML and SBC, whose formula is [6]

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

We are interested in their relative behavior when $T\to\infty$.

Lemma 1. *Assuming that $T\to\infty$, K is fixed and p does not increase with T, we have:* $\mathrm{GOF}=\left[\frac{T}{2}\log|\hat{\mathbf{\Sigma}}_p|\right][1-o(1)]$ *and* $\mathrm{PEN}_1=\left[\frac{K^2p}{2}\log T\right][1-o(1)]$.

Proof is outlined in [20, Section 3].

Proposition 2. RNML *and* SBC *reduce to the same formula when $T\to\infty$. 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. It is obvious that PEN_2 becomes negligible with respect to PEN_1 as $T\to\infty$. We will show below (see Remark 1) that PEN_3 is bounded when $T\to\infty$. Then, Lemma 1 gives the stated relation between RNML and SBC. The consistency property of RNML is hence the same as for SBC, proved in [1, p. 150]. □

We now analyze 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 (5)]. We propose to find an upper bound for PEN_3 when $T\to\infty$. For any integer h, let $\mathbf{R}(h)=\mathrm{Cov}(\mathbf{y}_t,\mathbf{y}_{t-h})$ be the autocovariance matrix at lag h, for the time series $\mathbf{Y}$. According to [21, Section 3.3], the model of the time series can be "approximated" by a $\mathrm{VAR}(p)$. For convenience, we assume that both p and p° are from the set $\{1,\dots,p_{\max}\}$.

Remark 1. *We make the convention that $\mathbf{\Sigma}_0=\mathbf{R}(0)$. It is known from [21, p. 75] that $(\mathbf{Y}'\mathbf{Y})/T$ converges to $\mathbf{\Sigma}_0$ almost surely as $T\to\infty$. Similarly, for $p>0$, $\hat{\Sigma}_p\to\Sigma_p$ almost surely. Hence, asymptotically in T we have:* $\mathrm{PEN}_3\le(K^2p/2)\log\mathrm{tr}(\mathbf{\Sigma}_0)$. *Note that the upper bound for* PEN_3 *does not depend on T.*

We show in the next proposition how the upper bound for PEN_3 can be further improved. More importantly, the proof of the proposition reveals the relationship between PEN_3 and the partial autocorrelation matrix.

Proposition 3. *When $T\to\infty$, if* $\mathrm{RNML}(\mathbf{Y};p)$ *is evaluated for a data matrix $\mathbf{Y}$ produced by a VAR(p°)-model, then the following inequality holds true:* $\mathrm{PEN}_3\le\frac{Kp}{2}\log|\mathbf{\Sigma}_p|+\frac{K^2p}{2}\log K+\frac{K^2p}{2}\log\left[\frac{\phi(\mathbf{\Sigma}_0)}{\psi(p,p^\circ)}-1\right]$, *where* $\phi(\mathbf{\Sigma}_0)=\dfrac{\mathrm{tr}(\mathbf{\Sigma}_0)}{K|\mathbf{\Sigma}_0|^{1/K}}$ *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) $\psi(p+1,p^\circ)\le\psi(p,p^\circ)$ for $p=\overline{1,p^\circ-1}$; (iii) $\psi(p,p^\circ)=\psi(p^\circ,p^\circ)$ for $p=\overline{p^\circ+1,p_{\max}}$.*

Proof is presented in [20, Section 4].

Remark 2. *From [20, Section 4], we have that $\psi(p,p^\circ)=(|\mathbf{\Sigma}_p|/|\mathbf{\Sigma}_0|)^{1/K}$, which leads to the identity $\frac{K^2p}{2}\log\mathrm{tr}(\mathbf{\Sigma}_0)=\frac{Kp}{2}\log|\mathbf{\Sigma}_p|+\frac{K^2p}{2}\log K+\frac{K^2p}{2}\log\frac{\phi(\mathbf{\Sigma}_0)}{\psi(p,p^\circ)}$. This demonstrates that the upper bound for* PEN_3 *in Proposition 3 is sharper than the one in Remark 1.*

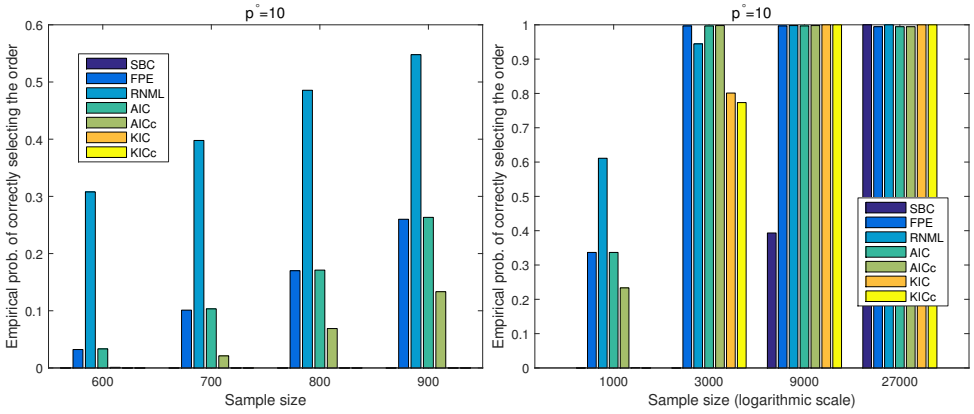
3. EXPERIMENTAL RESULTS

We compare RNML with SBC, AIC, AICc, KIC, KICc and FPE - Final Prediction Error criterion [22].

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 $N_{\mathrm{SP}}=9$ sparsity patterns which are denoted $\{\mathrm{SP}_i\}_{i=0}^8$. After setting $\mathrm{SP}_0=\emptyset$ and $(u,v)=(1,2)$, we apply the following recursions, for $i=\overline{0,7}$: (i) $\mathrm{SP}_{i+1}\leftarrow\mathrm{SP}_i\cup\{(u,v)\}$ and (ii) if $v<K$, then $(u,v)\leftarrow(u,v+1)$, else $(u,v)\leftarrow(u+1,u+2)$. Remark that $\mathrm{SP}_0\subset\mathrm{SP}_1\subset\cdots\subset\mathrm{SP}_8$.

Inspired by [23, Example 2], we generate for each SP in $\{\mathrm{SP}_i\}_{i=0}^8$ an ISDM with the property that the entries of $\{\mathbf{Q}_m\}_{m=1}^{p^\circ}$ [see (2)] are zero in the positions corresponding to SP, and all other entries are randomly drawn from the univariate Gaussian distribution with mean 2×10^{-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 [24, App.B.5]) for obtaining the matrix polynomial $\mathbf{A}_{\mathrm{SP}}$ of order p°. The covariance matrix $\mathbf{\Sigma}_{\mathrm{SP}}$ is a byproduct of this procedure. We simulate N_r different K-variate time series of length $T_{\max}$ by using $\mathbf{A}_{\mathrm{SP}}$ and $\mathbf{\Sigma}_{\mathrm{SP}}$ in (1).

In our settings, $N_r=100$ and $T_{\max}=27000$. Hence, for each p°-order, the number of simulated K-variate time series is $N_{\mathrm{SP}}\times N_r=900$. Each time series is used to estimate the matrix coefficients and $\hat{\mathbf{\Sigma}}_p$ for $p=\overline{1,20}$ by employing the implementation of ARFIT algorithm [17], available at `http://climate-dynamics.org/software/#arfit`. The order is selected by the seven ITC which were listed above. 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 $600\le T\le900$ and (ii) $T\leftarrow3T$ when $1000\le T\le9000$. We count how many times each criterion selects the correct



(a) (b)

Fig. 1: Performance of various criteria in estimating the order of VAR-model.

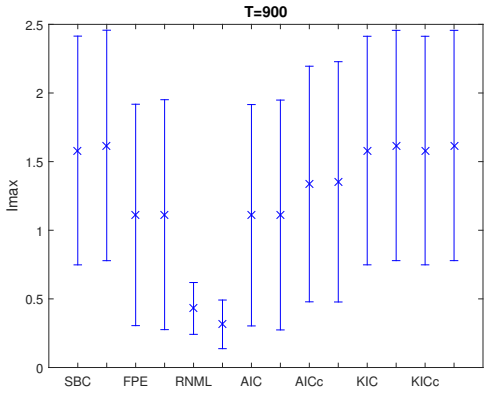


Fig. 2: Statistics for the maximum value of I-divergences computed on the $\mathcal{G}$-grid. For each ITC, we plot two error bars, each of which represents mean plus minus standard deviation: The first error bar is for $I_{\max}$, while the second one is for $I^{\mathrm{ME}}_{\max}$. The sample size is $T = 900$ and the "true" order is $p^\circ = 10$.

order. The results are shown in [20, Fig. 1], from which we excerpt here the case $p^\circ = 10$ in Fig. 1.

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 AICc yield the best estimates when $T \leq 900$ by correctly selecting the order in 70% to 100% of the cases, while KIC and KICc 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 Fig. 1 that SBC, KIC, and KICc 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 ITC which selects correctly the order in more than 50% of the cases. For $p^\circ = 15$, the performance of all ITC 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 property is well illustrated in [20, Fig. 1]. We can conclude that RNML is superior to other criteria when p° is large.

After the order $\hat{p}$ of the model is selected with an ITC, the autocovariance matrices $\hat{\mathbf{R}}(0), \ldots, \hat{\mathbf{R}}(\hat{p})$ 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 $\hat{\mathbf{R}}(0), \ldots, \hat{\mathbf{R}}(\hat{p})$ and (ii) the sparsity pattern of ISDM is SP. For details, we refer to [2]. 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 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 $(N_{\mathrm{SP}} - 1) \times N_r = 800$.

In order to clarify the notation, let us assume that $S(\omega)$ and $\hat{S}(\omega)$ are the matrix spectral densities for the "true" model and the estimated model, respectively. Recall that the order of the "true" model is p°, while the order of the estimated model is $\hat{p}$. The maximum entropy estimate, $\hat{S}^{\mathrm{ME}}(\omega)$, 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 $N_{\mathrm{grid}} = 1024$ and we evaluate $S(\omega), \hat{S}(\omega), \hat{S}^{\mathrm{ME}}(\omega)$ for $\omega \in \mathcal{G}$, where $\mathcal{G} = \left\{ \frac{0\times\pi}{N_{\mathrm{grid}}}, \frac{1\times\pi}{N_{\mathrm{grid}}}, \cdots, \frac{N_{\mathrm{grid}}\times\pi}{N_{\mathrm{grid}}} \right\}$. In order to investigate how far is $S(\omega)$ from $\hat{S}(\omega)$, we calculate the I-divergence between them by applying the general formula for two positive-definite matrices $\mathbf{F}$ and $\mathbf{G}$ [25]: $D(\mathbf{F}||\mathbf{G}) = -(1/2)\left[\log|\mathbf{F}\mathbf{G}^{-1}| + \mathrm{tr}(\mathbf{I} - \mathbf{F}\mathbf{G}^{-1})\right]$.

Given that $I(\omega)$ is the I-divergence between $S(\omega)$ and $\hat{S}(\omega)$, we compute $I_{\max} = \max_{\omega\in\mathcal{G}} I(\omega)$. Similarly, $I^{\mathrm{ME}}_{\max}$ is the maximum of the I-divergence between $S(\omega)$ and $\hat{S}^{\mathrm{ME}}(\omega)$

when $\omega \in \mathcal{G}$. Statistics concerning $I_{\max}$ and $I_{\max}^{\mathrm{ME}}$ are plotted in Fig. 2. Observe that RNML is the best among all criteria because it minimizes the maximum for each of the two I-divergences. This is true not only for $T = 900$, but for all sample sizes we have considered in our experiment (see [20, Fig. 2]). The outcome of this experiment is further analyzed by computing the multivariate Itakura-Saito divergence [26, 27] between the "true" model and the estimated model. The interested reader can find this analysis in [20, Section 5], where other numerical examples are presented as well.

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

4. FINAL REMARKS

In this paper, 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 a separate work.

5. REFERENCES

[1] H. Lutkepöhl, *New Introduction to Multiple Time Series Analysis*, John Wiley & Sons, 2005.

[2] J. Songsiri, J. Dahl, and L. Vandenberghe, "Graphical models of autoregressive processes," in *Convex Optimization in Signal Processing and Communications*, D.P. Palomar and Y.C. Eldar, Eds., pp. 89–116. Cambridge Univ. Press, 2010.

[3] C.-M. Ting, A.K. Seghouane, M.U. Khalid, and S.-H. Salleh, "Is first-order vector autoregressive model optimal for fMRI data?," *Neural Computation*, vol. 27, pp. 1857–1871, 2015.

[4] R.A. Davis, P. Zang, and T. Zheng, "Sparse vector autoregressive modeling," ArXiv preprint arXiv:1207.0520 [Online], 2012.

[5] D.R. Brillinger, "Remarks concerning graphical models for time series and point processes," *Revista de Econometria*, vol. 16, pp. 1–23, 1996.

[6] G. Schwarz, "Estimating the dimension of a model," *Annals of Statistics*, vol. 6, pp. 461–464, 1978.

[7] H. Akaike, "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, vol. AC-19, pp. 716–723, Dec. 1974.

[8] C.M. Hurvich and C.L. Tsai, "A corrected Akaike information criterion for vector autoregressive model selection," *Journal of Time Series Analysis*, vol. 14, pp. 271–279, 1993.

[9] J.E. Cavanaugh, "A large-sample model selection criterion based on Kullback's symmetric divergence," *Statistics and Probability Letters*, vol. 42, pp. 333–343, 1999.

[10] A.K. Seghouane, "Vector autoregressive model-order selection from finite samples using Kullback's symmetric divergence," *IEEE Transactions on Circuits and Systems-I Regular Papers*, vol. 53, pp. 2327–2335, 2006.

[11] J. Rissanen, "MDL denoising," *IEEE Transactions on Information Theory*, vol. 46, no. 7, pp. 2537–2543, Nov. 2000.

[12] J. Rissanen, *Information and complexity in statistical modeling*, Springer Verlag, 2007.

[13] C.D. Giurcăneanu and J. Rissanen, "Estimation of AR and ARMA models by stochastic complexity," in *Time series and related topics*, Hwai-Chung Ho, Ching-Kang Ing, and Tze Leung Lai, Eds., vol. 52, pp. 48–59. Institute of Mathematical Statistics Lecture Notes-Monograph Series, 2006.

[14] D.F. Schmidt and E. Makalic, "Estimating the order of an autoregressive model using normalized maximum likelihood," *IEEE Transactions on Signal Processing*, vol. 59, no. 2, pp. 479–487, 2011.

[15] W.W.S. Wei, *Time Series Analysis. Univariate and Multivariate Methods*, Pearson Education, Inc., 2006.

[16] M. Morf, A. Vieira, and T. Kailath, "Covariance characterization by partial autocorrelation matrices," *The Annals of Statistics*, vol. 6, no. 3, pp. 643–648, 1978.

[17] A. Neumaier and T. Schneider, "Estimation of parameters and eigenmodes of multivariate autoregressive models," *ACM Trans. Math. Softw.*, vol. 27, pp. 27–57, 2001.

[18] J. Rissanen, T. Roos, and P. Myllymäki, "Model selection by sequentially normalized least squares," *Journal of Multivariate Analysis*, vol. 101, pp. 839–849, 2010.

[19] S. Hirai and K. Yamanishi, "Efficient computation of normalized maximum likelihood codes for Gaussian mixture models with its applications to clustering," *IEEE Transactions on Information Theory*, vol. 59, pp. 7718–7727, 2013.

[20] S. Maanan, B. Dumitrescu, and C.D. Giurcăneanu, "Supplemental material to: Renormalized maximum likelihood for multivariate autoregressive models," `https://www.stat.auckland.ac.nz/~cgiu216/PUBLICATIONS.htm`, 2016.

[21] G.C. Reinsel, *Elements of Multivariate Time Series Analysis*, Springer-Verlag, 1993.

[22] H. Akaike, "Autoregressive model fitting for control," *Annals of the Institute of Statistical Mathematics*, vol. 23, pp. 163–180, 1971.

[23] J. Songsiri and L. Vandenberghe, "Topology selection in graphical models of autoregressive processes," *Journal of Machine Learning Research*, vol. 11, pp. 2671–2705, 2010.

[24] B. Dumitrescu, *Positive trigonometric polynomials and signal processing applications*, Springer, 2007.

[25] T.P. Speed and H.T. Kiiveri, "Gaussian Markov distributions over finite graphs," *Annals of Statistics*, vol. 14, no. 1, pp. 138–150, 1986.

[26] F.R. Bach and M.I. Jordan, "Learning graphical models for stationary time series," *IEEE Transactions on Signal Processing*, vol. 52, no. 8, pp. 2189–2199, 2004.

[27] A. Ferrante, C. Masiero, and M. Pavon, "Time and spectral domain relative entropy: A new approach to multivariate spectral estimation," *IEEE Transactions on Automatic Control*, vol. 57, no. 10, pp. 2561–2575, 2012.