SUPPLEMENTAL MATERIAL TO: RENORMALIZED MAXIMUM LIKELIHOOD FOR MULTIVARIATE AUTOREGRESSIVE MODELS

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

[1]Department of Statistics
University of Auckland

[2]Department of Automatic Control and Computers
Politehnica University of Bucharest

Important note: Hereafter, the main document will be referred to as [1].

1. PROOF OF PROPOSITION 1 FROM THE MAIN DOCUMENT

In this section, we apply techniques which are similar to those from [2, 3, 4, 5].

Preliminary calculations

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

$$\begin{aligned}\mathbf{Z}_t &= [\mathbf{y}_t', \ldots, \mathbf{y}_{t-p+1}']' \ (\ell \times 1),\\ \mathbf{Z} &= [\mathbf{Z}_0, \ldots, \mathbf{Z}_{T-1}]' \ (T \times \ell),\\ \mathbf{U} &= [\mathbf{u}_1, \ldots, \mathbf{u}_T]' \ (T \times K).\end{aligned}$$

Remark that the size of each newly defined quantity is listed in the parentheses. As $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_T]'$ and $\mathbf{B} = [\mathbf{A}_1, \ldots, \mathbf{A}_p]'$, it follows that

$$\mathbf{Y} = \mathbf{Z}\mathbf{B} + \mathbf{U}. \qquad (1)$$

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 [6]:

$$\begin{aligned}\hat{\mathbf{B}} &= (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y}, &(2)\\ \hat{\boldsymbol{\Sigma}} &= (\mathbf{Y}'\mathbf{P}_{\mathbf{Z}}^{\perp}\mathbf{Y})/T, &(3)\end{aligned}$$

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 $(\cdot)^{\mathrm{V}}$ to both sides of the identity in (1), we obtain

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

where $\otimes$ denotes the Kronecker product. It is evident that $\mathbf{U}^{\mathrm{V}} \sim \mathcal{N}_{TK}(\mathbf{0}, \boldsymbol{\Sigma} \otimes \mathbf{I})$. The generalized least-squares estimator for $\mathbf{B}^{\mathrm{V}}$ is [6]

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

and coincides with the ML estimator for $\mathbf{B}$ [see (2)].

From the standard properties of the ML estimators, we have:

(P$_1$) $\hat{\mathbf{B}}^{\mathrm{V}} \sim \mathcal{N}(\mathbf{B}^{\mathrm{V}}, \boldsymbol{\Omega})$, where $\boldsymbol{\Omega} = \boldsymbol{\Sigma} \otimes (\mathbf{Z}'\mathbf{Z})^{-1}$;

(P$_2$) $T\hat{\boldsymbol{\Sigma}} \sim \mathcal{W}_K(\boldsymbol{\Sigma}, T - \ell)$ [Wishart distribution with scale matrix $\boldsymbol{\Sigma}$ and degrees of freedom parameter $T - \ell$];

(P$_3$) $\hat{\mathbf{B}}^{\mathrm{V}}$ is statistically independent of $\hat{\boldsymbol{\Sigma}}$.

First normalization step

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

Using the properties (P$_1$)-(P$_3$) along with the fact that the statistics $\hat{\boldsymbol{\theta}}$ are sufficient for $\boldsymbol{\theta}$ [6], we get the following chain of identities for the likelihood function:

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

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

$$g(\hat{\boldsymbol{\theta}};\hat{\boldsymbol{\theta}}) = G|\hat{\boldsymbol{\Sigma}}|^{-(\ell+K+1)/2}, \qquad (4)$$

where

$$G = \frac{T^{1+K(T-\ell-K-1)/2}}{(2^{TK/2})(\pi^{\ell K/2})}\frac{|\mathbf{Z}'\mathbf{Z}|^{K/2}\exp(-TK/2)}{\Gamma_K\left[(T-l)/2\right]}, \qquad (5)$$

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.

$$|\hat{\mathbf{\Sigma}}| = \prod_{j=1}^{K} \hat{\lambda}^{(j)}. \tag{6}$$

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

As a preliminary step for constraining the integration domain in [1, Eq. (4)], we consider the hyper-parameters $R > 0$ and $\lambda_{\min}^{(K)} \geq \cdots \geq \lambda_{\min}^{(1)} > 0$. We take $\Lambda_{\min} = \left\{\lambda_{\min}^{(j)}\right\}_{j=1}^{K}$. With the convention that $\|\cdot\|$ denotes the Euclidean norm, we define:

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

The constrained normalization factor is:

$$\begin{aligned}
&C_p(R,\Lambda_{\min}) \\
&= \int_{\mathcal{Y}(R,\Lambda_{\min})} f\left(\mathbf{Y}^{\mathrm{V}}; \boldsymbol{\theta}(\hat{\mathbf{Y}}^{\mathrm{V}})\right) \mathrm{d}\mathbf{Y}^{\mathrm{V}} \\
&= \int_{\mathcal{T}(R,\Lambda_{\min})} \left\{ \int_{\mathcal{Y}(\hat{\boldsymbol{\theta}})} f(\mathbf{Y}^{\mathrm{V}}|\hat{\boldsymbol{\theta}})\mathrm{d}\mathbf{Y}^{\mathrm{V}} \right\} g(\hat{\boldsymbol{\theta}};\hat{\boldsymbol{\theta}})\mathrm{d}\hat{\boldsymbol{\theta}} \qquad (7) \\
&= G\left\{ \prod_{j=1}^{K} \int_{\lambda_{\min}^{(j)}}^{\infty} \left[\hat{\lambda}^{(j)}\right]^{-\frac{\ell+K+1}{2}} \mathrm{d}\hat{\lambda}^{(j)} \right\} \int_{\mathcal{B}(R)} \mathrm{d}\hat{\mathbf{B}}^{\mathrm{V}} \qquad (8) \\
&= G\left\{ \prod_{j=1}^{K} \frac{\left[\lambda_{\min}^{(j)}\right]^{-(l+K-1)/2}}{(l+K-1)/2} \right\} \mathrm{vol}\left[\mathcal{B}(R)\right]. \qquad (9)
\end{aligned}$$

In (7), we used the fact that the inner integral equals one, while in (8) we applied the identities in (4)-(6). For the volume which appears in (9), it is easy to show (see, for example, [5]) that

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

where

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

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 [2] 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

$$\begin{aligned}
-\log \hat{f}(\mathbf{Y}^{\mathrm{V}}; p) &= -\log f\left(\mathbf{Y}^{\mathrm{V}}; \hat{\mathbf{B}}(\mathbf{Y}^{\mathrm{V}}), \hat{\mathbf{\Sigma}}(\mathbf{Y}^{\mathrm{V}})\right) \\
&\quad + \log C_p(R,\Lambda_{\min}),
\end{aligned}$$

we take

$$\begin{aligned}
R &= \frac{\left\|(\mathbf{I}\otimes\mathbf{Z})\hat{\mathbf{B}}^{\mathrm{V}}\right\|^2}{KT}, \qquad (10) \\
&= \frac{\mathrm{tr}\left(\mathbf{Y}'\mathbf{Y} - T\hat{\mathbf{\Sigma}}\right)}{KT}. \qquad (11)
\end{aligned}$$

The selection of R-value in (10) is mainly determined by the definition of $\mathcal{B}(R)$. The identity in (11) is straightforwardly obtained from (3).

Similar considerations lead to

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

Hence, we have:

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

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:

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

$$\mathcal{T}(R_1,R_2,\lambda_1,\lambda_2)$$

$$
\begin{aligned}
&= \left\{\hat{\boldsymbol{\theta}} : \hat{\mathbf{B}}^{\mathrm{V}} \in \mathcal{B}(R_1, R_2) \text{ and } \hat{\boldsymbol{\Sigma}} \in \mathcal{L}(\lambda_1, \lambda_2)\right\}, \\
&\mathcal{Y}(R_1, R_2, \lambda_1, \lambda_2) \\
&= \left\{\mathbf{Y}^{\mathrm{V}} : \hat{\boldsymbol{\theta}}(\mathbf{Y}^{\mathrm{V}}) \in \mathcal{T}(R_1, R_2, \lambda_1, \lambda_2)\right\}.
\end{aligned}
$$

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

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

For proving the last identity, we have used the expression of $g(\hat{\boldsymbol{\theta}}; \hat{\boldsymbol{\theta}})$ from (4)-(6) as well as the expression of $C_p(R, \Lambda_{\min})$ from (9). 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 [5, Eq. (10)] and we employed the identities from [5, Appendix A]. Hence, we get:

$$
\begin{aligned}
&\log \overline{C}_p(R_1, R_2, \lambda_1, \lambda_2) \\
&= -K \log \frac{2}{\ell + K - 1} + \log \ell \\
&\quad + \log \frac{K}{2} \qquad (12) \\
&\quad + K \log \log \frac{\lambda_2}{\lambda_1} \qquad (13) \\
&\quad + \log \log \frac{R_2}{R_1}. \qquad (14)
\end{aligned}
$$

We ignore the term in (12) because is constant. Due to the same reasons as those invoked in [2, 4], we drop the terms in (13) and (14).

After collecting all the terms corresponding to $\log C_p(R, \Lambda_{\min})$ and $\log \overline{C}_p(R_1, R_2, \lambda_1, \lambda_2)$ along with negative log-likelihood, we obtain the RNML criterion which is presented in [1, Proposition 1].

2. RNML-CRITERION FOR ZERO-ORDER MODEL

In this case, we only need to compute the estimate $\hat{\boldsymbol{\Sigma}}_0 = (\mathbf{Y}'\mathbf{Y})/T$. All other calculations are similar to those from Section 1, but simpler. It is easy to show that the logarithm of the normalization factor which corresponds to $\log C_p(R, \Lambda_{\min})$ is

$$
\begin{aligned}
\log C_0(\Lambda_{\min,0}) &= -\frac{K-1}{2} \log |\hat{\boldsymbol{\Sigma}}_0| \\
&\quad - \log \Gamma_K\left(\frac{T}{2}\right) + K \log \frac{2}{K-1},
\end{aligned}
$$

where $\Lambda_{\min,0}$ is the set of eigenvalues of $\hat{\boldsymbol{\Sigma}}_0$. 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 $\log \overline{C}_p(R_1, R_2, \lambda_1, \lambda_2)$:

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

It follows that

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

It is a simple exercise to verify that the expression above reduces to the one in [2, Eq. (9.40)] when $K = 1$. The discussion on the role of the hyper-parameters, λ_3 and λ_4, goes along the same lines as in [2]. From the identity in (11) and the definitions of $\mathcal{B}(R_1, R_2)$ and $\mathcal{L}(\lambda_1, \lambda_2)$ in Section 1, we have the double inequality:

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

At the same time, $\lambda_3 \leq \operatorname{tr}(\hat{\Sigma}_0)/K \leq \lambda_4$, which leads to choosing $\lambda_3 = R_1 + \lambda_1$ and $\lambda_4 = R_2 + \lambda_2$. If we apply the same technique as in [2] by taking $\lambda_1 = R_1 = a$ and $\lambda_2 = R_2 = b$ $(0 < a < b)$, the contribution of the hyper-parameters to $\mathrm{RNML}(\mathbf{Y}; p)$ 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 $\mathrm{RNML}(\mathbf{Y}; p = 0)$, we can conclude that neglecting the hyper-parameters might have a negative impact on the selection of the model.

3. PROOF OF LEMMA 1 FROM THE MAIN DOCUMENT

The identity for GOF can be obtained straightforwardly. For PEN_1, we note that the the multivariate Gamma function can be written as

$$
\frac{\Gamma_K\left[\frac{T-Kp}{2}\right]}{\pi^{\frac{K(K-1)}{4}}} = \prod_{i=1}^{K} \Gamma\left[\frac{T - Kp + 1 - i}{2}\right]. \tag{15}
$$

Furthermore, we use the Stirling approximation [7, 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 (15) because it does not depend on the model order. With the convention that $\ell = Kp$, we have:

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

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

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

We drop the constant terms and use the Maclaurin series expansion for $\log\left[1-(\ell-1+i)/T\right]$ when $1 \leq i \leq K$:

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

This concludes the proof.

4. PROOF OF PROPOSITION 3 FROM THE MAIN DOCUMENT

The matrix coefficients of the VAR(p)-model are denoted $\mathbf{A}_{1\,p}, \ldots, \mathbf{A}_{p\,p}$. From Yule-Walker equations, we have [8]:

$$\begin{aligned}
\mathbf{A}_{(p)} &= \mathbf{R}_p^{-1}\mathbf{R}_{(p)},\\
\boldsymbol{\Sigma}_p &= \mathbf{R}(0) - \mathbf{R}'_{(p)}\mathbf{R}_p^{-1}\mathbf{R}_{(p)},
\end{aligned}$$

where

$$\begin{aligned}
\mathbf{A}_{(p)} &= [\mathbf{A}_{1\,p}, \ldots, \mathbf{A}_{p\,p}]',\\
\mathbf{R}_{(p)} &= [\mathbf{R}(1), \ldots, \mathbf{R}(p)]',\\
\mathbf{R}_p &= \begin{bmatrix} \mathbf{R}(0) & \mathbf{R}(1) & \cdots & \mathbf{R}(p-1)\\ \mathbf{R}'(1) & \mathbf{R}(0) & \cdots & \mathbf{R}(p-2)\\ \vdots & \vdots & \ddots & \vdots\\ \mathbf{R}(p-1)' & \mathbf{R}(p-2)' & \cdots & \mathbf{R}(0) \end{bmatrix}.
\end{aligned}$$

We employ the result from [8, p. 75] which says that the sample covariance matrix $\hat{\mathbf{R}}(h)$ converges to $\mathbf{R}(h)$ almost surely as $T \to \infty$. This result along with Yule-Walker equations allow us to replace, in our asymptotic analysis, $\hat{\boldsymbol{\Sigma}}_p$ with $\boldsymbol{\Sigma}_p$ for all $p \geq 0$.

As in [8, 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 $\mathbf{y}_{t-1}, \ldots, \mathbf{y}_{t-p+1}$. Hence, we have $\boldsymbol{\Sigma}_{p-1} = \mathrm{Cov}(\mathbf{u}_{p,t})$. Additionally, we introduce the notation $\overleftarrow{\boldsymbol{\Sigma}}_{p-1} = \mathrm{Cov}(\overleftarrow{\mathbf{u}}_{p,t-p})$.

Furthermore, for $p > 0$, we define the *partial correlation matrix* (see also [8, p. 70][9, Eq. (16.5.52)]):

$$\mathbf{Q}(p) = \left[\overleftarrow{\boldsymbol{\Sigma}}_{p-1}^{1/2}\right]^{-1} \mathrm{Cov}\left(\overleftarrow{\mathbf{u}}_{p,t-p}, \mathbf{u}_{p,t}\right)\left[\boldsymbol{\Sigma}_{p-1}^{1/2}\right]^{-1}, \quad (16)$$

where $\overleftarrow{\boldsymbol{\Sigma}}_{p-1}^{1/2}$ and $\boldsymbol{\Sigma}_{p-1}^{1/2}$ are the symmetric square roots of $\overleftarrow{\boldsymbol{\Sigma}}_{p-1}$ and $\boldsymbol{\Sigma}_{p-1}$, respectively. If in (16) we replace $\overleftarrow{\boldsymbol{\Sigma}}_{p-1}^{1/2}$ with the lower triangular root of $\overleftarrow{\boldsymbol{\Sigma}}_{p-1}$ and $\boldsymbol{\Sigma}_{p-1}^{1/2}$ with the upper triangular root of $\boldsymbol{\Sigma}_{p-1}$, then we obtain the transpose of the matrix defined in [10, Eq. (14)].

Note that the entries of $\mathbf{Q}(p)$ are correlation coefficients only in the particular case when $K = 1$ (see [9, p. 413-414] for a more detailed discussion). However, the eigenvalues of $\mathbf{Q}'(p)\mathbf{Q}(p)$ 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 $\mathbf{y}_{t-1}, \ldots, \mathbf{y}_{t-p+1}$" [8, p. 71].

It follows from [10, p. 646] that

$$|\mathbf{\Sigma}_p| = |\mathbf{\Sigma}_0| \prod_{i=1}^{p} |\mathbf{I} - \mathbf{Q}'(i)\mathbf{Q}(i)|\,.$$

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

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

$$\begin{aligned} &\mathrm{tr}(\mathbf{\Sigma}_0 - \mathbf{\Sigma}_p) \\ &\quad \leq \quad \mathrm{tr}(\mathbf{\Sigma}_0) - K|\mathbf{\Sigma}_p|^{1/K} \\ &\quad = \quad K|\mathbf{\Sigma}_p|^{1/K} \left[\frac{\phi(\mathbf{\Sigma}_0)}{\psi(p,p^\circ)} - 1 \right], \end{aligned}$$

where $\psi(p,p^\circ) = \prod_{i=1}^{p} |\mathbf{I} - \mathbf{Q}'(i)\mathbf{Q}(i)|^{1/K}$. All that remains is to employ the inequality above in conjunction with the definition of PEN_3.

5. ADDITIONAL INFORMATION ON EXPERIMENTS

5.1. Example 1

The complete description of this example as well as the interpretation of the outcome can be found in [1]. Here we display Figs. 1-4, in which the experimental results are shown. We also give an interpretation of the estimation results by resorting to multivariate Itakura-Saito divergence.

The multivariate Itakura-Saito divergence between the "true" $\mathrm{VAR}(p^\circ)$ and the estimated $\mathrm{VAR}(\hat{p})$ is given by $J = \frac{1}{2\pi}\int_0^{2\pi} D(S(\omega)||\hat{S}(\omega))\mathrm{d}\omega$ [11, 12]. An approximation of J can be easily obtained from the values of the I-divergence computed on the grid $\mathcal{G}$ (see [1]). The same method can be applied for the evaluation of J^{ME}, the multivariate Itakura-Saito divergence between the "true" model and its maximum-entropy estimate. The statistics for J and J^{ME} are shown in Fig. 3. When $T \leq 1000$, RNML yields values of J and J^{ME} which are larger than the values of divergences produced by other ITC. 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^{ME} instead of J^{ME}. Observe in Fig. 4 that, when T is small and $\hat{p}_{\mathrm{RNML}} = p^\circ$, J_c computed for RNML exceeds the values of J_c corresponding to other ITC. At the same time, in all these cases, RNML yields the greatest improvement of J_c^{ME} in comparison with J_c.

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

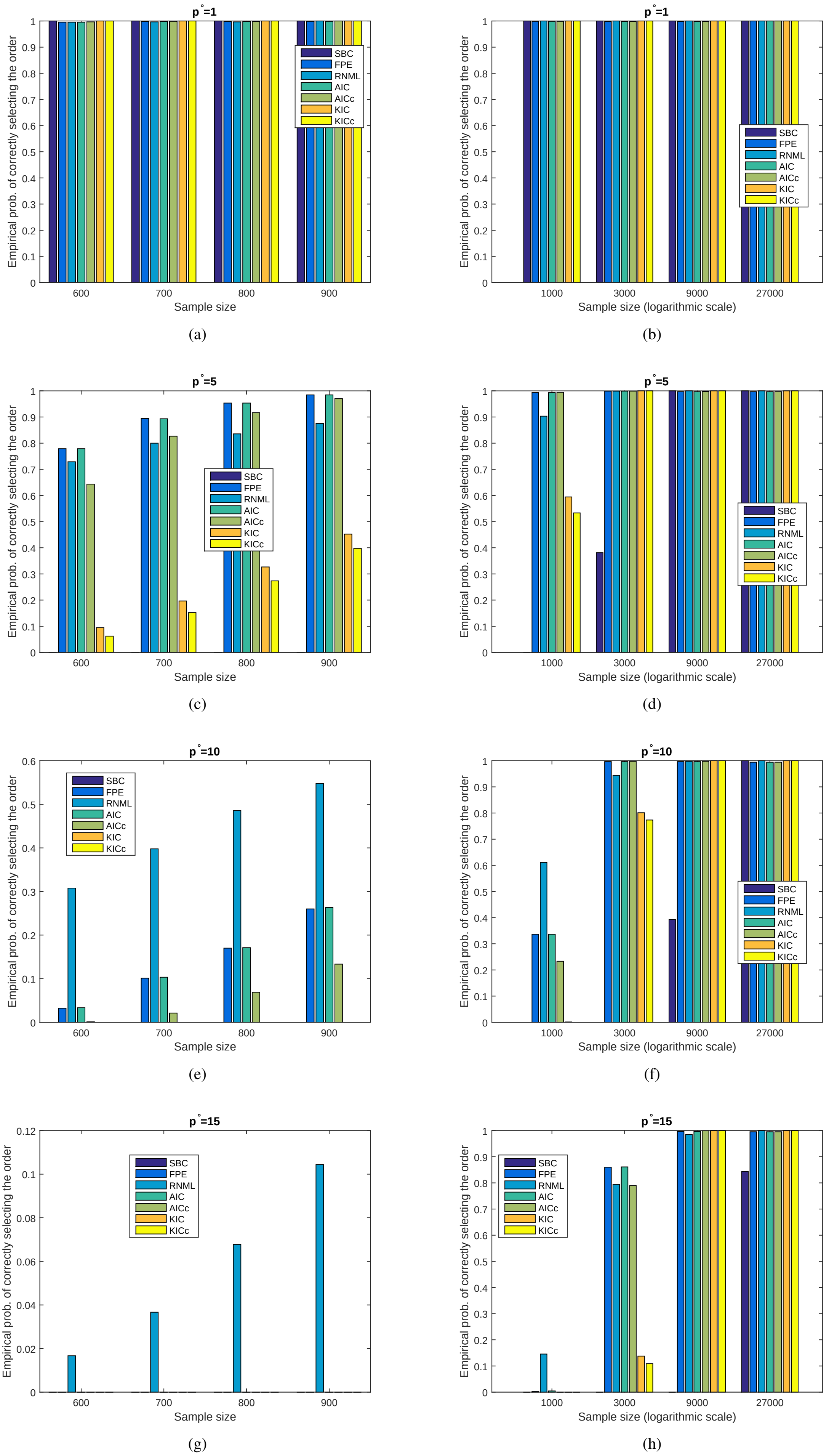
5.2. Example 2

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 [13] and assumes that the driven noise is Gaussian and $\mathbf{\Sigma}_{p^\circ} = \mathbf{I}$. The matrix coefficients of the model $(\mathbf{A}_{1\,p^\circ}, \ldots, \mathbf{A}_{p^\circ\,p^\circ})$ are of the form $[\mathbf{D}_1\ \mathbf{0}; \mathbf{0}\ \mathbf{D}_2]$, where both $\mathbf{D}_1$ and $\mathbf{D}_2$ are 10×10. 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°. We emphasize that, for $p^\circ = 2$, the non-zero entries of $\mathbf{A}_{1\,2}$ and $\mathbf{A}_{2\,2}$ are statistically independent and they are statistically independent with respect to the driven noise. In our experiments, for each p°, 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 $\hat{\mathbf{A}}_{1\,p}, \ldots, \hat{\mathbf{A}}_{p\,p}$ and $\hat{\mathbf{\Sigma}}_p$ for $p = \overline{1,8}$ by using the ARFIT-algorithm [14]. Then the best order is selected with seven ITC: SBC [15], AIC [16], AICc [17], KIC [18], KICc [19] and FPE [20]. 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 Fig. 5a, when $p^\circ = 1$, we have: (i) SBC, RNML, AICc and KICc 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 KICc; (v) FPE and KIC have the same level of performance.

As we can see in Fig. 5b, 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) AICc is ranked the best and is much better than AIC; (iv) KICc is better than KIC; (v) FPE is very good, except for $T = 200$.

We note that, in [13], 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 [13] 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.



(a)

(b)

(c)

(d)

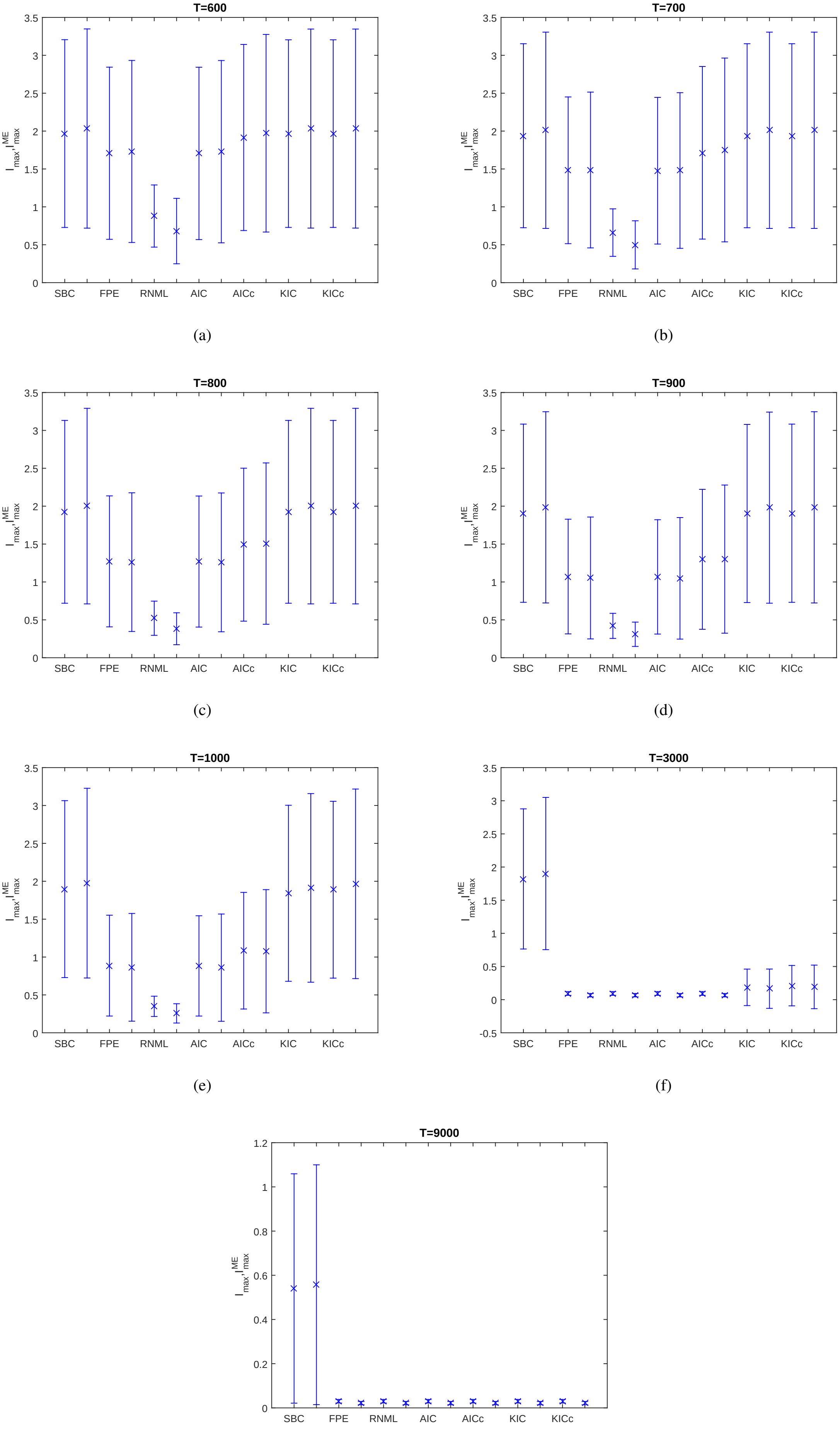
(e)

(f)

(g)

(h)

Fig. 1: Example 1: Performance of various criteria in estimating the order of VAR-model. The value of "true" VAR-order, p°, is written on the top of each plot.



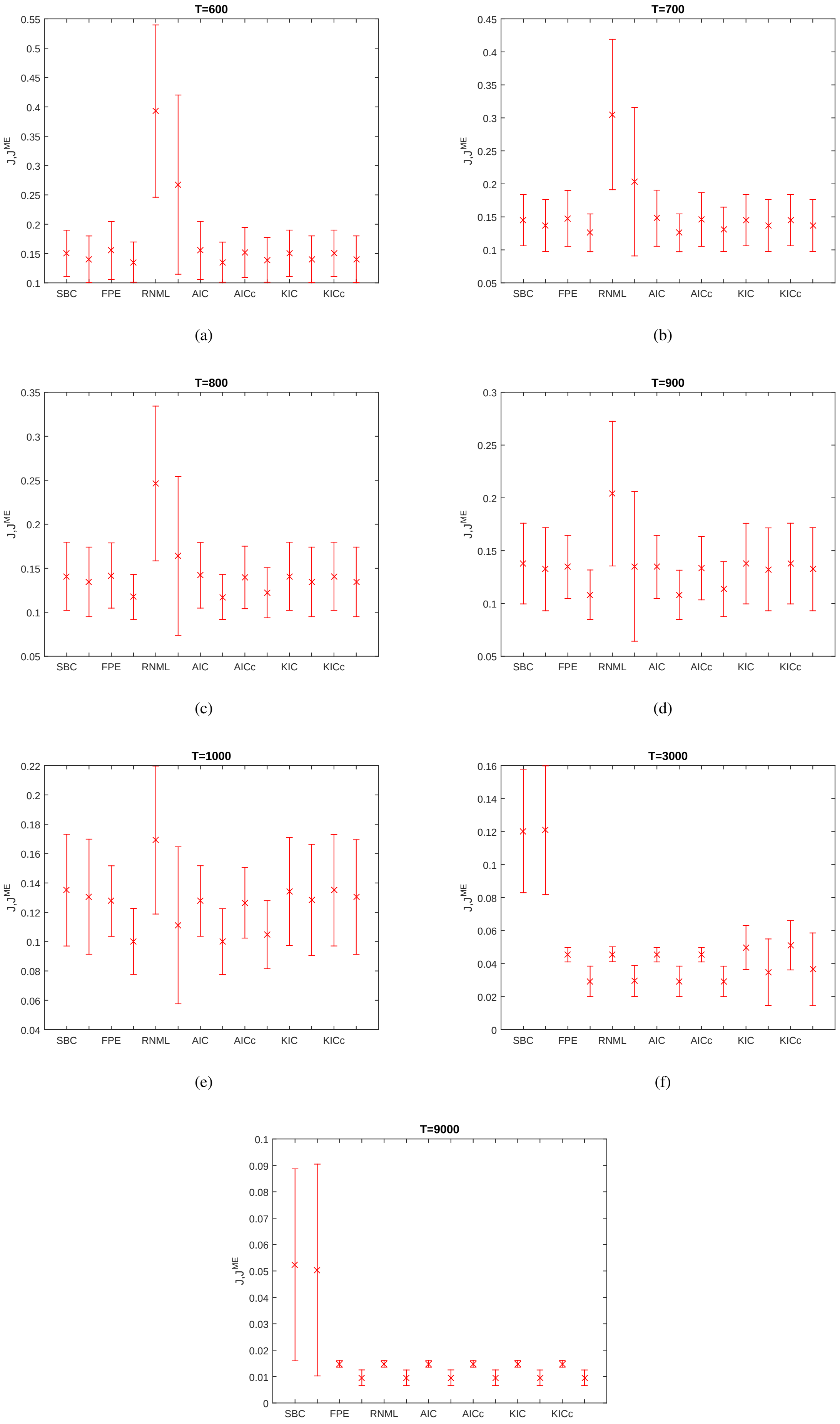
(a) (b)

(c) (d)

(e) (f)

(g)

Fig. 2: Example 1: 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 $I_{\max}^{\text{ME}}$. The sample size, T, is written on the top of each plot. Note that $p^{\circ} = 10$.



(a)

(b)

(c)

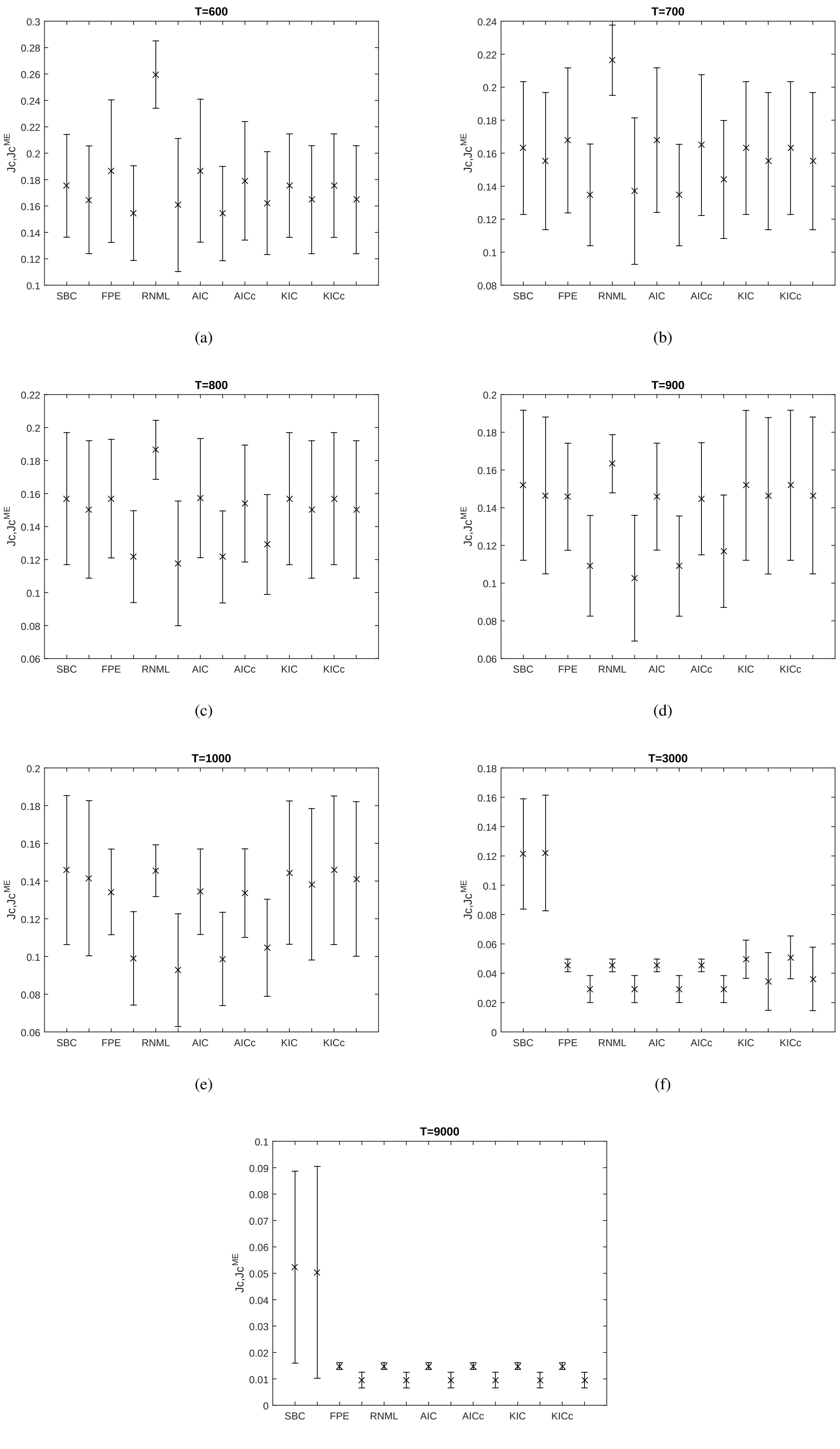
(d)

(e)

(f)

(g)

Fig. 3: Example 1: Statistics for Itakura-Saito divergence. All graphical conventions are the same like in Fig. 2, except that $I_{\max}$ is replaced by J and $I_{\max}^{\mathrm{ME}}$ is replaced by J^{ME}. Note that $p^{\circ} = 10$.



(a)

(b)

(c)

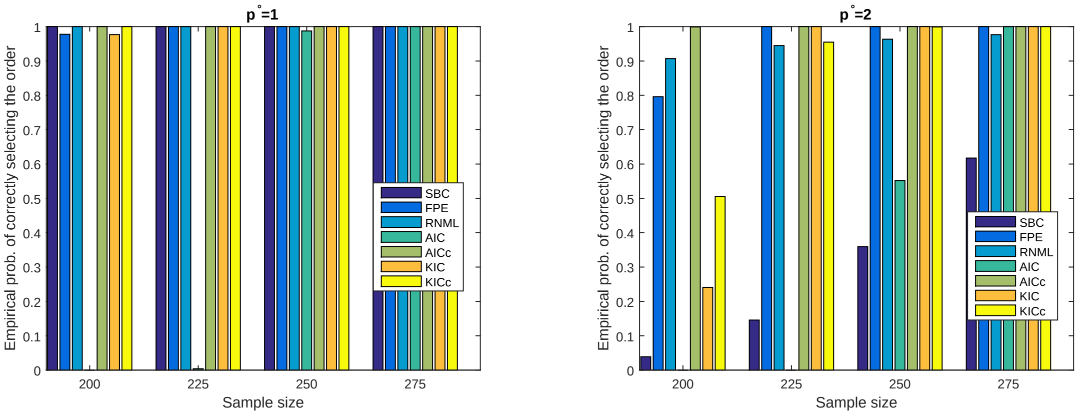
(d)

(e)

(f)

(g)

Fig. 4: Example 1: Statistics for Itakura-Saito divergence computed only for the cases when RNML estimates correctly the order of the model ($\hat{p}_{\mathrm{RNML}} = 10$). All graphical conventions are the same like in Fig. 2, except that $I_{\max}$ is replaced by J_c and $I_{\max}^{\mathrm{ME}}$ is replaced by J_c^{ME}. Note that $p^{\circ} = 10$.



(a)

(b)

Fig. 5: Example 2: Performance of various criteria in estimating the order of VAR-model. All graphical conventions are the same like in Fig. 1.

6. REFERENCES

[1] S. Maanan, B. Dumitrescu, and C.D. Giurcăneanu, “Renormalized maximum likelihood for multivariate autoregressive models,” EUSIPCO 2016.

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

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

[4] 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.

[5] C.D. Giurcăneanu, S.A. Razavi, and A. Liski, “Variable selection in linear regression: Several approaches based on normalized maximum likelihood,” *Signal Processing*, vol. 91, pp. 1671–1692, 2011.

[6] K.V. Mardia, J.T. Kent, and J.M. Bibby, *Multivariate Analysis*, Academic Press, 1979.

[7] E. Artin, *The Gamma function*, Holt, Rinehart and Winston, Inc., 1964.

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

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

[10] 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.

[11] 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.

[12] 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.

[13] 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.

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

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

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

[17] 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.

[18] 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.

[19] 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.

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