Graphical inference in non-Markovian linear-Gaussian state-space models

Emilie Chouzenoux, Víctor Elvira

▶ **To cite this version:**

Emilie Chouzenoux, Víctor Elvira. Graphical inference in non-Markovian linear-Gaussian state-space models. ICASSP 2024 - IEEE International Conference on Acoustics, Speech and Signal Processing, Apr 2024, Seoul, South Korea. hal-04417208

HAL Id: hal-04417208
https://inria.hal.science/hal-04417208

Submitted on 25 Jan 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

GRAPHICAL INFERENCE IN NON-MARKOVIAN LINEAR-GAUSSIAN STATE-SPACE MODELS

Émilie Chouzenoux[1] *and Víctor Elvira*[2]

[1] Université Paris-Saclay, CentraleSupélec, Inria, CVN, Gif-sur-Yvette, France
[2] School of Mathematics, University of Edinburgh, United Kingdom

ABSTRACT

State-space models (SSMs) are common tools in time-series analysis for inference and prediction. SSMs are versatile probabilistic models that allow for Bayesian inference by describing a (generally Markovian) latent process. However, the parameters of that latent process are often unknown and must be estimated. In this paper, we consider the parameter estimation in a SSM with a non-Markovian linear-Gaussian latent process. This process is described as a vector auto-regressive with p unknown matrices. Our algorithm LaGrangEM estimates these matrices through an expectation-maximization algorithm that exploits a graphical interpretation of the latent process in order to define prior knowledge about the unknown parameters. We connect the new algorithm with existing approaches such as Granger causality and graphical inference in SSMs. We discuss the strong potential of the algorithm to bring interpretability, e.g., in estimating causal relationships and their delays. The numerical experiments also show a superiority in performance.

***Index Terms*—** State-space modeling, Granger causality, graphical inference, expectation-maximization, sparsity.

1. INTRODUCTION

Dealing with multi-variate time series is central in countless discrete-time signal processing applications [1]. State-space models (SSMs) are probabilistic models that have shown a superior predictive performance while incorporating uncertainty quantification. Unlike other popular time-series approaches, such as vector auto-regressive models, SSMs consider a latent state that evolves over time and that explains the observed time series [2]. The linear-Gaussian SSM (LG-SSM) is arguably the most celebrated model, since it allows for exact inference when all model parameters are known [3, 4]. In practice, these parameters are rarely available and must be estimated [5]. Existing works based on graphical models time-series include [6–9], with applications in biology [10, 11], networks [12], and neuroscience [13]. Recent works have dealt with parameter estimation in Markovian SSMs through a graphical perspective, either point-wise [14–16] or fully probabilistic [17, 18].

In this paper, we consider a LG-SSM where the latent process follows an order-p vector auto-regressive process, VAR(p), generalizing standard LG-SSMs where $p = 1$. The p unknown state matrices, which govern the state dynamics, are estimated by our novel expectation-maximization algorithm, LaGrangEM. The E-step relies on a RTS algorithm in an extended space, while the M-step implements a modern non-convex optimization algorithm. This allows us to promote interesting properties in the sought matrices [19], e.g., sparsity. LaGrangEM can incorporate additional structure into the problem, e.g., that each dimension in the latent process is affected by at maximum another dimension, or that each dimension can affect to each other through at maximum one specific (unknown) lag. These two examples highlight that LaGrangEM provides an enhanced interpretability with close connections with graphical approaches and Granger causality methods [11, 20–22]. Moreover, the non-Markovianity in the latent process increases the dimension of the RTS algorithm by a factor of $p + 1$, which is a known limitation of these models. Thus, another advantage is that LaGrangEM can decrease in many cases the computational complexity in the RTS algorithm due to the structure and sparsity.

The rest of the paper is structured as follows. In Section 2, we provide background material. The novel LaGrangEM algorithm is introduced in Section 3. We provide numerical experiments in Section 4 and a conclusion in Section 5.

2. BACKGROUND

2.1. Linear non-Markovian state-space model

We consider an SSM with non-Markovian auto-regressive latent process of order $p \geq 1$, denoted as AR(p). For every time $k = 1, \ldots, K$, the evolving hidden state $\mathbf{x}_k \in \mathbb{R}^{N_x}$, $N_x \geq 1$, is linked to the observation $\mathbf{y}_k \in \mathbb{R}^{N_y}$, $N_y \geq 1$:

$$\begin{cases} \mathbf{x}_k = \sum_{i=1}^{p} \mathbf{A}_i \mathbf{x}_{k-i} + \mathbf{q}_k, \\ \mathbf{y}_k = \mathbf{H}\mathbf{x}_k + \mathbf{r}_k. \end{cases} \tag{1}$$

V.E. acknowledges support from ARL/ARO under grant W911NF-22-1-0235. E.C. acknowledges support from the European Research Council under Starting Grant MAJORIS ERC-2019-STG-850925.

Matrices $\mathbf{A}_i \in \mathbb{R}^{N_x \times N_x}$, for $i \in \{1, \dots, p\}$ describe the state transition for each lag i, $\mathbf{H} \in \mathbb{R}^{N_y \times N_x}$ is the observation matrix, $\{\mathbf{q}_k\}_{k=1}^K \sim \mathcal{N}(0, \mathbf{Q})$ is the i.i.d. state noise process with $\mathbf{Q} \in \mathbb{R}^{N_x \times N_x}$ symmetric definite positive (SDP), and $\{\mathbf{r}_k\}_{k=1}^K \sim \mathcal{N}(0, \mathbf{R})$ is the i.i.d. observation noise process with SDP $\mathbf{R} \in \mathbb{R}^{N_y \times N_y}$. The state process is initialized as $\mathbf{x}_i \sim \mathcal{N}(\mathbf{x}_i; \boldsymbol{\mu}_0, \mathbf{P}_0)$ for $i \in \{1-p, \dots, 0\}$, with known $\boldsymbol{\mu}_0 \in \mathbb{R}^{N_x}$ and SDP $\mathbf{P}_0 \in \mathbb{R}^{N_x \times N_x}$. In the remaining of the paper, we use the short notation $\mathbf{A} = [\mathbf{A}_1, \dots, \mathbf{A}_p] \in \mathbb{R}^{N_x \times pN_x}$, i.e., the rowwise concatenation of the state transition matrices.

2.2. Inference in non-Markovian LG-SSMs with known parameters

A standard approach for filtering and smoothing in (1) consists on stacking (columnwise) the p consecutive states into $\mathbf{z}_k = [\mathbf{x}_k; \mathbf{x}_{k-1}; \dots; \mathbf{x}_{k-p+1}] \in \mathbb{R}^{pN_x}$, and performing Kalman filtering (KF) and Rauch-Tung-Striebel (RTS) smoother on the equivalent SSM

$$\begin{cases} \mathbf{z}_k = \check{\mathbf{A}}\mathbf{z}_{k-1} + \check{\mathbf{q}}_k, \\ \mathbf{y}_k = \check{\mathbf{H}}\mathbf{z}_k + \mathbf{r}_k, \end{cases} \tag{2}$$

where we define

$$\check{\mathbf{A}} = \begin{bmatrix} \mathbf{A}_1 & \cdots & \cdots & \mathbf{A}_p \\ \mathbf{I} & 0 & \cdots & 0 \\ & \ddots & \ddots & \vdots \\ (0) & & \mathbf{I} & 0 \end{bmatrix} \in \mathbb{R}^{pN_x \times pN_x},$$

$$\check{\mathbf{H}} = [\mathbf{H}\ (0)] \in \mathbb{R}^{N_y \times pN_x},\ \check{\mathbf{Q}} = \begin{bmatrix} \mathbf{Q} & (0) \\ (0) & (0) \end{bmatrix} \in \mathbb{R}^{pN_x \times pN_x},$$

$\check{\mathbf{q}}_k \sim \mathcal{N}(0, \check{\mathbf{Q}})$, and $\mathbf{r}_k \sim \mathcal{N}(0, \mathbf{R})$ [23].

2.3. EM approach for parameter estimation

Running the the KF and RTS on model (2) requires knowing all parameters. The parameter estimation in LG-SSMs is often done via (a) gradient-based algorithms, using sensitivity equations or Fisher's identity to efficiently compute derivatives of the likelihood function, allowing iterative optimizers like quasi-Newton or Newton-Raphson to obtain the ML estimate [24, 25], or (b) EM-based algorithms, which optimize an iteratively improved bound on the marginal likelihood. See [4, 26] for more details.

In [15], we introduced GraphEM, an EM method for estimating matrix $\mathbf{A}$ in the case of a Markovian LG-SSM (i.e., $p = 1$). GraphEM incorporated a suitable prior on $\mathbf{A}$ following an original perspective relating the transition matrix to the adjacency matrix of a directed graph. In this work, we share a similar goal as in GraphEM, which unfortunately cannot be used since matrix $\check{\mathbf{Q}}$ in the extended space is singular, and the derivations of the E-step do not hold (we re-derive it in the next section). Also, we take a different graphical interpretation on the non-Markovian model which requires different priors and post-processing.

3. PROPOSED APPROACH

3.1. Graphical non-Markovian modeling

In this paper, we take a graphical modeling approach where the matrices $\mathbf{A}_i$ encode i-lagged dependencies among state dimensions. In particular, for any lag $i \in \{1, \dots, p\}$, a non-zero component $A_i(\ell, n) \neq 0$, implies that $\mathbf{x}_{k-i}(\ell)$ is affected by $\mathbf{x}_k(n)$, with $(n, \ell) \in \{1, \dots, N_x\}^2$. This can be interpreted as the discovery of conditional Granger causality [21] in the state space up to a p auto-regressive order. Thus, the temporal relations among the dimensions of the multivariate unobserved state process can be represented as a graphical model of p (possibly reflexive) directed graphs. Each graph is composed of N_x vertices, and its weights are the values of matrix $\mathbf{A}_i^\top$, for each lag $i \in \{1, \dots, p\}$.

Toy example. Figure 1 displays the latent state relations in the first equation of model (1), with $N_x = 3$, $p = 2$, and

$$\mathbf{A}_1 = \begin{pmatrix} 0.9 & 0.7 & 0 \\ 0 & 0 & -0.3 \\ 0 & 0 & 0 \end{pmatrix},\ \mathbf{A}_2 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0.8 & 0 \end{pmatrix},$$

i.e., with sparse matrices, which is key for interpretability. The left-hand-side in Figure 1 represents the probabilistic model of the non-Markovian SSM, with the 1-lag-ahead and 2-lag-ahead interactions depicted in blue and magenta, respectively. The right-hand-side plot shows the superposition of the $p = 2$ directed graphs. This example satisfies having (at maximum) one lag interaction per pair of dimensions, thus the model can be represented by a (coloured) directed graph that captures the relations in all p lags among dimensions.

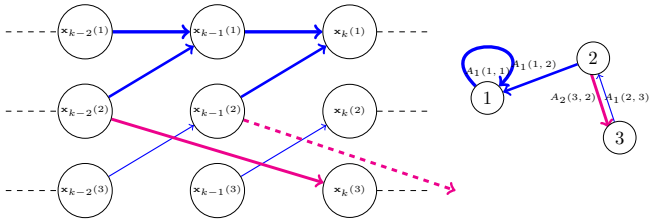

Fig. 1: Summary representation of the proposed graphical model, for the toy example. The edges in blue (resp. magenta) are defined as non-zero entries of $\mathbf{A}_1^\top$ (resp. $\mathbf{A}_2^\top$). Edge thickness is proportional to the absolute entries of these matrices. Left: Dynamic state evolution. Right: Graphs associated to matrices $\mathbf{A}_1$ and $\mathbf{A}_2$.

Connection with existing algorithms. LaGrangEM can be seen as an SSM with a Granger causality model in its latent process, which enhances interpretability due to the promoted structure and sparsity. Another advantage of LaGrangEM is that it keeps its probabilistic nature in both state and observation models, and could be easily extended to a fully Bayesian approach (e.g., as in SpaRJ [18]). Unlike in Granger-based fitting methods, LaGrangEM allows to promote structure in the latent process by incorporating prior knowledge. Moreover,

LaGrangEM can be seen as a generalization of GraphEM [14, 15], where the latent process follows a generic AR(p), instead of $p = 1$ in GraphEM. This requires re-deriving the E-step.

3.2. LaGrangEM algorithm

We now introduce LaGrangEM, an EM-based algorithm to estimate the maximum a posteriori (MAP) of $\mathbf{A}$ in model 1. This is equivalent to minimizing $\mathcal{L}(\mathbf{A}) \triangleq \mathcal{L}_0(\mathbf{A}) + \mathcal{L}_{1:K}(\mathbf{A})$ with $\mathcal{L}_0(\mathbf{A}) \triangleq -\log p(\mathbf{A})$ and $\mathcal{L}_{1:K}(\mathbf{A}) \triangleq -\log p(\mathbf{y}_{1:K}|\mathbf{A})$. Due to the intricate structure of the model in (1), the function $\mathcal{L}_{1:K}$ takes a recursive form that makes its minimization not straightforward (see discussion in [2, 15, 27] for the Markovian case).

The final algorithm is summarized in Alg. 1. Given an initial estimate $\mathbf{A}^{(0)} \in \mathbb{R}^{N_x \times pN_x}$, the EM algorithm alternates at each iteration $i \in \mathbb{N}$ between (a) the majorization step which builds $\mathcal{Q}(\mathbf{A}; \mathbf{A}^{(i-1)})$ as in Eq. (4), i.e., an upper bound on the neg-log-likelihood function (E-step), and (b) the minimization of this upper bound (M-step), which finds the new estimate as in (5). By construction, the sequence $(\mathbf{A}^{(i)})_{i\in\mathbb{N}}$ decreases monotonically the MAP loss $\mathcal{L}$, and convergence guarantees can be obtained under suitable assumptions [28].

Post-processing. Sophisticated constraints, such as the single lag assumption, described in Sec. 3.1, or the DAG constraint [29], can also be desirable. As these constraints are non-convex and non-continuous [30], they are not straightforward to be integrated inside the EM framework. Instead, we propose a post-processing step after the EM reached convergence, to project the estimated $\mathbf{A}$ on the desired constrained set. The projection $\mathbf{A}' = [\mathbf{A}'_1, \dots, \mathbf{A}'_p] \in \mathbb{R}^{N_x \times pN_x}$ over the single lag constraint of an $\mathbf{A} \in \mathbb{R}^{N_x \times pN_x}$ reads as

$$(\forall (n,\ell,i)) \quad A'_i(n,\ell) = \begin{cases} A_i(n,\ell) & \text{if } i = o(n,\ell), \\ 0 & \text{elsewhere,} \end{cases} \tag{3}$$

with $o(n,\ell) = \text{argmax}_{i\in 1,\dots,p}|A_i(n,\ell)|$ (with the convention $o(n,\ell) = 0$ if $A_i(n,\ell) = 0$ for all i). The matrix $\mathbf{O} = (o(n,\ell))_{1\le n,\ell\le N_x}$ defines the lag of the AR(p) model for the interaction (if any) of each pair of dimensions, improving the interpretability compared to Granger [20] or even conditional Granger models [21].

Prior on A. We are interested in priors on $\mathbf{A}$ that favor graphical interpretation (e.g.,, sparse $\mathbf{A}$). Moreover, our algorithm restricts to log-concave prior $p(\mathbf{A})$. As an example, we propose $\mathcal{L}_0(\mathbf{A}) = \gamma \sum_{i=1}^p \|\mathbf{A}_i\|_1$, where $\gamma \ge 0$ is the weight for the ℓ_1 term. Since $\mathcal{L}_0$ is a convex, lower-semicontinuous, and proper function ([31]), the existence of an explicit proximity operator ([32]) leads to a simple recursion in our M-step.

Algorithm 1 Proposed algorithm

Sequence $(\mathbf{y}_k)_{k=1}^K$, initial $\mathbf{A}^{(0)} \in \mathbb{R}^{N_x \times pN_x}$, $\gamma \ge 0$. Precision $\varepsilon > 0$.
For $i = 1, 2, \dots$

E-step:

Run KF and RTS smoother in extended space using $\check{\mathbf{A}}^{(i-1)}$.
Deduce $(\mathbf{m}^s_{k:k-p}, \mathbf{P}^s_{k:k-p})_{k=1}^K$ using [2, Chap.8], and define

$$(\forall \mathbf{A} \in \mathbb{R}^{N_x \times pN_x}) \quad \mathcal{Q}(\mathbf{A}; \mathbf{A}^{(i-1)}) = \frac{K}{2}\text{tr}\Big(\mathbf{Q}^{-1}(\mathbf{\Sigma} - \mathbf{\Delta}\mathbf{A}^\top - \mathbf{A}\mathbf{\Delta}^\top + \mathbf{A}\mathbf{\Phi}\mathbf{A}^\top)\Big), \tag{4}$$

with $\mathbf{\Sigma} = \sum_{k=1}^K \mathbf{\Pi}_k^{(0,0)}$, $\mathbf{\Delta} = \sum_{k=1}^K \mathbf{\Pi}_k^{(0,1)}$, $\mathbf{\Phi} = \sum_{k=1}^K \mathbf{\Pi}_k^{(1,1)}$.

M-step: Use DR solver to solve

$$\mathbf{A}^{(i)} = \text{argmin}_{\mathbf{A}} \Big(\mathcal{Q}(\mathbf{A}; \mathbf{A}^{(i-1)}) + \mathcal{L}_0(\mathbf{A})\Big). \tag{5}$$

Stopping condition:

If $\|\mathbf{A}^{(i)} - \mathbf{A}^{(i-1)}\|_F \le \varepsilon \|\mathbf{A}^{(i-1)}\|_F$, stop the recursion.

3.3. Derivations of LaGrangEM

E-step. For every $\mathbf{A} \in \mathbb{R}^{N_x \times pN_x}$ and $i \in \mathbb{N}$, the EM bound is defined as (see for instance [2, 28, 33])

$$\mathcal{Q}(\mathbf{A}; \mathbf{A}^{(i-1)}) = -\int p(\mathbf{x}_{1-p:K}|\mathbf{y}_{1:K}, \mathbf{A}^{(i-1)}) \times \log p(\mathbf{x}_{1-p:K}, \mathbf{y}_{1:K}|\mathbf{A}) \mathrm{d}\mathbf{x}_{1-\mathrm{p}:\mathrm{K}} + \mathrm{ct}_{/\mathbf{A}}, \tag{6}$$

where $\mathrm{ct}_{/\mathbf{A}}$ does not depend on $\mathbf{A}$, such that equality holds at $\mathbf{A} = \mathbf{A}^{(i-1)}$. Plugging Eq. (1), for every $\mathbf{A} \in \mathbb{R}^{N_x \times pN_x}$,

$$\log\, p(\mathbf{x}_{1-p:K}, \mathbf{y}_{1:K}|\mathbf{A}) = -\frac{K}{2}\sum_{k=1}^K (\mathbf{x}_k - \sum_{i=1}^p \mathbf{A}_i \mathbf{x}_{k-i})^\top \times \mathbf{Q}^{-1}(\mathbf{x}_k - \sum_{i=1}^p \mathbf{A}_i \mathbf{x}_{k-i}) + \mathrm{ct}_{/\mathbf{A}}. \tag{7}$$

We introduce the compact notations $\widetilde{\mathbf{A}} = [\mathbf{I}, -\mathbf{A}_1, \cdots, -\mathbf{A}_p] \in \mathbb{R}^{N_x \times (p+1)N_x}$ and $\mathbf{x}_{k:k-p} = [\mathbf{x}_k; \mathbf{x}_{k-1}; \cdots; \mathbf{x}_{k-p}] \in \mathbb{R}^{(p+1)N_x}$. Using (6)-(7), and marginalizing part of the variables,

$$\mathcal{Q}(\mathbf{A}; \mathbf{A}^{(i-1)}) = \frac{K}{2}\sum_{k=1}^K \int \mathbf{x}_{k:k-p}^\top (\widetilde{\mathbf{A}}^\top \mathbf{Q}^{-1} \widetilde{\mathbf{A}}) \mathbf{x}_{k:k-p} \times p(\mathbf{x}_{k:k-p}|\mathbf{y}_{1:K}, \mathbf{A}^{(i-1)}) d\mathbf{x}_{k:k-p} + \mathrm{ct}_{/\mathbf{A}}. \tag{8}$$

For every $k \in \{1, \dots, K\}$, $p(\mathbf{x}_{k:k-p}|\mathbf{y}_{1:K}, \mathbf{A}^{(i-1)})$ follows a multivariate Gaussian distribution with mean $\mathbf{m}^s_{k:k-p}$ and covariance $\mathbf{P}^s_{k:k-p}$, obtained by running KF/RTS algorithm with the extended transition matrix $\check{\mathbf{A}}^{(i-1)}$ [2, Chap.8]. Let

$$\mathbf{P}^s_{k:k-p} + \mathbf{m}^s_{k:k-p}(\mathbf{m}^s_{k:k-p})^\top := \begin{bmatrix} \mathbf{\Pi}_k^{(0,0)} & \mathbf{\Pi}_k^{(0,1)} \\ \mathbf{\Pi}_k^{(1,0)} & \mathbf{\Pi}_k^{(1,1)} \end{bmatrix} \tag{9}$$

with $\mathbf{\Pi}_k^{(0,0)} \in \mathbb{R}^{N_x \times N_x}$, $\mathbf{\Pi}_k^{(0,1)} = (\mathbf{\Pi}_k^{(1,0)})^\top \in \mathbb{R}^{N_x \times pN_x}$, $\mathbf{\Pi}_k^{(1,1)} \in \mathbb{R}^{pN_x \times pN_x}$. Hence, we can re-express (8) as

$$\mathcal{Q}(\mathbf{A}; \mathbf{A}^{(i-1)}) = \frac{K}{2} \sum_{k=1}^{K} \text{trace}(\mathbf{Q}^{-1}(\mathbf{\Pi}_k^{(0,0)} - \mathbf{A}(\mathbf{\Pi}_k^{(0,1)})^\top - \mathbf{\Pi}_k^{(0,1)}\mathbf{A}^\top + \mathbf{A}\mathbf{\Pi}_k^{(1,1)}\mathbf{A}^\top)) + \text{ct}_{/\mathbf{A}}. \quad (10)$$

Finally, using the additivity of the trace, and re-ordering terms, we derive the bound in Eq. (4) of LaGrangEM.

This bound shares a similar structure as in GraphEM [15], i.e., in the $p = 1$ case (see also [2, 34]). However, as mentioned above, it is not possible to directly apply GraphEM in model (2) since $\check{\mathbf{Q}}$ is not invertible and would lead to a singular bound. In contrast, here we depart from the E-step definition to obtain a valid bound.

M-step. We finally deduce an upper bound for the penalized loss, as $\mathcal{L}_0 + \mathcal{Q}(\cdot; \mathbf{A}^{(i-1)})$. We can solve (5) for a log-concave $p(\mathbf{A})$ by using the proximal splitting Douglas-Rachford (DR) algorithm [35]. The derivations can be found in [14] for a similar example, and skipped here due to lack of space.

4. NUMERICAL EVALUATION

We now evaluate our method on realistic graph datasets used in causal discovery in weather variability tracking. We consider five matrices $\mathbf{A}^* \in \mathbb{R}^{N_x \times pN_x}$, with $N_x = 5$, and $p = 6$, representing the ground truth state model used to produce `WEATH` datasets in the NeurIPS 2019 data challenge [36].[1] For each $\mathbf{A}^*$, we create times series $(\mathbf{x}_k, \mathbf{y}_k)_{k=1}^K$ using (1), with $K = 2 \times 10^3$, $\mathbf{H} = \mathbf{I}$ (i.e., $N_y = N_x$), $\mathbf{R} = \sigma_{\mathbf{R}}^2\mathbf{I}$, $\mathbf{Q} = \sigma_{\mathbf{Q}}^2\mathbf{I}$, and $\mathbf{P}_0 = \sigma_0^2\mathbf{I}$, with $(\sigma_{\mathbf{R}}, \sigma_{\mathbf{Q}}, \sigma_0) = (10^{-1}, 10^{-1}, 10^{-2})$. We evaluate the quality of our approximation, $\widehat{\mathbf{A}}$, in terms of RMSE$(\widehat{\mathbf{A}}, \mathbf{A}^*)$, as well as the accuracy in two detection tasks. Namely, we compute the macro-averaged accuracy between the (multi-class) ground truth lags $\mathbf{O}^*$ and their estimate $\widehat{\mathbf{O}}$ (see Sec. 3.1), and the accuracy on the (binary) detection of the transition matrix support (that is, the graph edges positions); a threshold value of 10^{-5} on the absolute entries is used.

We compare our approach LaGrangEM with the maximum likelihood (MLEM) estimation of $\mathbf{A}$ (i.e., running our Alg. 1 with $\mathcal{L}_0 \equiv 0$). In such case, an explicit expression is available for the M-step, not detailed here by lack of space. For both cases, we apply the projector (3), as a post-processing. We set precision $\varepsilon = 10^{-4}$ (with a maximum number of 50 EM iterations), and $\gamma = 200$ in all experiments. We also compare with conditional Granger causality (CGC) [21], a VAR(p) model extending generic Granger causality. Note that CGC estimates binary graphs, so that RMSE is not calculated in those cases.

[1] https://causeme.uv.es/static/datasets/TestWEATH/

Table 1: Results for proposed method and competitors. Mean (standard deviation) computed over 20 realizations.

Dataset	Method	RMSE	accur. (lag)	accur. (edge)
WeathN5_A	MLEM	0.210(0.034)	0.734(0.107)	$0.073(1.424 \times 10^{-17})$
	CGC	×	0.837(0.097)	**0.908**(**0.043**)
	LaGrangEM	**0.063**(**0.030**)	**0.868**(**0.097**)	0.903(0.039)
WeathN5_B	MLEM	0.177(0.076)	0.573(0.093)	$0.073(1.424 \times 10^{-17})$
	CGC	×	**0.764**(**0.060**)	0.858(0.030)
	LaGrangEM	**0.041**(**0.019**)	0.654(0.081)	**0.922**(**0.024**)
WeathN5_C	MLEM	0.135(0.037)	0.814(0.064)	0.067(0.001)
	CGC	×	0.806(0.103)	0.752(0.050)
	LaGrangEM	**0.107**(**0.030**)	**0.915**(**0.089**)	**0.919**(**0.023**)
WeathN5_D	MLEM	0.073(0.015)	0.869(0.018)	0.080(0.001)
	CGC	×	0.914(0.051)	0.716(0.052)
	LaGrangEM	**0.037**(**0.012**)	**0.942**(**0.030**)	**0.856**(**0.057**)
WeathN5_E	MLEM	0.172(0.027)	0.824(0.077)	0.067(0)
	CGC	×	0.897(0.065)	0.878(0.033)
	LaGrangEM	**0.059**(**0.015**)	**0.937**(**0.058**)	**0.937**(**0.027**)

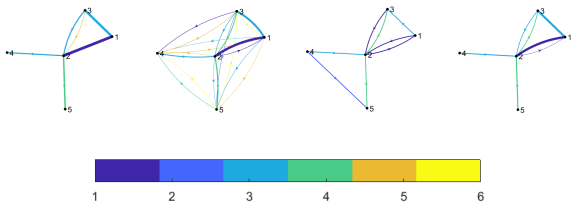

Fig. 2: From left to right: True graph, MLEM, CGC, and LaGrangEM estimates for WeathN5_A dataset. Edge color represents the lag for each edge. Self-loops are omitted for readibility.

The results are summarized in Tab. 1, averaged over 20 noise realizations in Eq. (1). LaGrangEM obtains a significant improvement in all cases, with CGC being a good competitor in two datasets. The sparsity promoting term in LaGrangEM especially shows its benefits for the edge detection task. Thus, LaGrangEM obtains excellent performance in all metrics. Fig. 2 illustrates this behavior, on the retrieval of WeathN5_A graph, showing the good performance of the proposed approach, in terms of edge retrieval (few spurious edges), edge weight estimation (accurate edge weights), and lag estimation (correct edge colors).

5. CONCLUSION

In this paper, we have addressed the problem of estimating transition matrices in LG-SSMs where the latent process follows a VAR(p) model. Moreover, we have adopted a graphical approach that interprets those parameters as p directed graphs, connecting to (conditional) Granger causality. Our derived EM-based algorithm allows us to introduce prior knowledge that promotes interesting properties such as sparsity, enhancing interpretability. Moreover, the sparsity can allow for computationally efficient RTS smoothing in the extended space that will be studied in the future. The numerical examples evidence an excellent performance on three different tasks, showing also a great potential of our approach.

References

[1] A. V. Oppenheim, *Discrete-Time Signal Processing*, Pearson Education India, 1999.

[2] S. Sarkka, *Bayesian Filtering and Smoothing*, 3 edition, 2013.

[3] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Journal of Basic Engineering*, vol. 82, pp. 35–45, 1960.

[4] S. Roweis and Z. Ghahramani, "A unifying review of linear gaussian models," *Neural computation*, vol. 11, no. 2, pp. 305–345, 1999.

[5] Z. Ghahramani and G. E. Hinton, "Parameter estimation for linear dynamical systems," 1996.

[6] M. Eichler, "Graphical modelling of multivariate time series," *Probability Theory and Related Fields*, vol. 153, no. 1, pp. 233–268, Jun. 2012.

[7] 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, Aug. 2004.

[8] D. Barber and A. T. Cemgil, "Graphical models for time-series," *IEEE Signal Processing Magazine*, vol. 27, no. 6, pp. 18–28, Nov 2010.

[9] V. N. Ioannidis, Y. Shen, and G. B. Giannakis, "Semi-blind inference of topologies and dynamical processes over dynamic graphs," *IEEE Transactions on Signal Processing*, vol. 67, no. 9, pp. 2263–2274, 2019.

[10] A. Pirayre, C. Couprie, L. Duval, and J.-C. Pesquet, "BRANE Clust: Cluster-assisted gene regulatory network inference refinement," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 15, no. 3, pp. 850–860, May 2018.

[11] D. Luengo, G. Rios-Munoz, V. Elvira, C. Sanchez, and A. Artes-Rodriguez, "Hierarchical algorithms for causality retrieval in atrial fibrillation intracavitary electrograms," *IEEE Journal of Biomedical and Health Informatics*, vol. 12, no. 1, pp. 143–155, Jan. 2019.

[12] C. Ravazzi, R. Tempo, and F. Dabbene, "Learning influence structure in sparse social networks," *IEEE Transactions on Control of Network Systems*, vol. PP, pp. 1–1, 12 2017.

[13] J. Richiardi, S. Achard, B. Horst, , and D. V. D. Ville, "Machine learning with brain graphs," *IEEE Signal Processing Magazine*, vol. 30, no. 3, pp. 58–70, 2013.

[14] E. Chouzenoux and V. Elvira, "GraphEM: EM algorithm for blind Kalman filtering under graphical sparsity constraints," in *Proceedings of the 45th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2020)*, 4–8 May 2020, pp. 5840–5844.

[15] V. Elvira and E. Chouzenoux, "Graphical inference in linear-Gaussian state-space models," *IEEE Transactions on Signal Processing*, vol. 70, pp. 4757–4771, 2022.

[16] E. Chouzenoux and V. Elvira, "Graphit: Iterative reweighted $\ell 1$ algorithm for sparse graph inference in state-space models," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP 2023)*, 2023.

[17] B. Cox and V. Elvira, "Parameter estimation in sparse linear-Gaussian state-space models via reversible jump Markov Chain Monte Carlo," in *Proceedings of the 30th European Signal Processing Conference (EUSIPCO 2022)*, 2022, pp. 797–801.

[18] B. Cox and V. Elvira, "Sparse Bayesian estimation of parameters in linear-Gaussian state-space models," *IEEE Transactions on Signal Processing (to appear in)*, 2023.

[19] J. Chiu, Y. Deng, and A. M. Rush, "Low-rank constraints for fast inference in structured models," in *Advances in Neural Information Processing Systems*, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., 2021.

[20] C. W. Granger, "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica: Journal of the Econometric Society*, pp. 424–438, 1969.

[21] J. F. Geweke, "Measures of conditional linear dependence and feedback between time series," *Journal of the American Statistical Association*, vol. 79, no. 388, pp. 907–915, 1984.

[22] D. Luengo, G. R. Muñoz, and V. Elvira, "Causality analysis of atrial fibrillation electrograms," in *2015 Computing in Cardiology Conference (CinC)*. IEEE, 2015, pp. 585–588.

[23] H. Lütkepohl, *New introduction to multiple time series analysis*, Springer Science & Business Media, 2005.

[24] M. Segal and E. Weinstein, "A new method for evaluating the log-likelihood gradient (score) of linear dynamic systems," *IEEE Transactions on Automatic Control*, vol. 33, no. 8, pp. 763–766, 1988.

[25] R. Olsson, K. Petersen, and T. Lehn-Schioler, "State-space models: from the EM algorithm to a gradient approach," *Neural Computation*, vol. 19, no. 4, pp. 1097–1111, 2007.

[26] Z. Ghahramani and G. Hinton, "Parameter estimation for linear dynamic systems," Tech. Rep., University of Toronto, 1996, http://mlg.eng.cam.ac.uk/zoubin/course04/tr-96-2.pdf.

[27] D. Nagakura, "Computing exact score vectors for linear Gaussian state space models," *Communications in Statistics - Simulation and Computation*, vol. 50, no. 8, pp. 2313–2326, 2021.

[28] C. F. J. Wu, "On the convergence properties of the EM algorithm," *The Annals of Statistics*, vol. 11, no. 1, pp. 95–103, 1983.

[29] S. Wei and Y. Xie, "Causal structural learning from time series: A convex optimization approach," Tech. Rep., 2023, https://arxiv.org/abs/2301.11336.

[30] D. Williams, "Beyond lasso: A survey of nonconvex regularization in gaussian graphical models," Tech. Rep., 2020, https://psyarxiv.com/ad57p/.

[31] H. H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, Springer, 2017.

[32] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM Journal on Imaging Sciences*, vol. 2, no. 1, pp. 183–202, 2009.

[33] O. Cappe, E. Moulines, and T. Ridden, *Inference in Hidden Markov Models*, Springer Series in Statistics. Springer New York, NY, 1st edition, 2005.

[34] R. H. Shumway and D. S. Stoffer, "An approach to time series smoothing and forecasting using the EM algorithm," *Journal of Time Series Analysis*, vol. 3, no. 4, pp. 253–264, 1982.

[35] P. L. Combettes and J.-C. Pesquet, "A Douglas-Rachford splitting approach to nonsmooth convex variational signal recovery," *IEEE Journal of Selected Topics in Signal Processing*, vol. 1, no. 4, pp. 564–574, Dec. 2007.

[36] J. Runge, X.-A. Tibau, M. Bruhns, J. Muñoz-Marí, and G. Camps-Valls, "The causality for climate competition," in *Proceedings of the NeurIPS 2019 Competition and Demonstration Track*, 2020, vol. 123, pp. 110–120.