Sparse Learning in Gaussian Chain Graphs for State Space Models

Lasse Petersen LP@MATH.KU.DK
Department of Mathematical Sciences
University of Copenhagen, Denmark

Abstract

The graphical lasso is a popular method for estimating the structure of undirected Gaussian graphical models from data by penalized maximum likelihood. This paper extends the idea of structure estimation of graphical models by penalized maximum likelihood to Gaussian chain graph models for state space models. First we show how the class of linear Gaussian state space models can be interpreted in the chain graph set-up under both the LWF and AMP Markov properties, and we demonstrate how sparsity of the chain graph structure relates to sparsity of the model parameters. Exploiting this relation we propose two different penalized maximum likelihood estimators for recovering the chain graph structure from data depending on the Markov interpretation at hand. We frame the penalized maximum likelihood problem in a missing data set-up and carry out estimation in each of the two cases using the EM algorithm. The common E-step is solved by smoothing, and we solve the two different M-steps by utilizing existing methods from high dimensional statistics and convex optimization.

Keywords: state space models; chain graph models; high dimensional statistics; sparse learning; EM algorithm; convex optimization.

1. Introduction

The *graphical lasso* (Banerjee et al., 2008; Friedman et al., 2008) produces a sparse estimate of the concentration matrix $\Theta = \Sigma^{-1}$ of a regular multivariate Gaussian distribution by penalized maximum likelihood from independent samples $x_1, \dots, x_N \sim \mathcal{N}(0, \Sigma)$. The estimator is given by

$$\hat{\Theta} = \underset{\Theta \in \mathcal{S}_p^{++}}{\arg\min} \{\text{tr}(\Theta S) - \log\det\Theta + ||W \circ \Theta||_1\} \tag{1}$$

where $\mathcal{S}_p^{++}$ are the real, symmetric, positive definite $p \times p$ matrices, $S = \frac{1}{N}\sum_{i=1}^N x_i x_i^T$ is the empirical covariance matrix, $||A||_1 = \sum_{ij}|A_{ij}|$ for a matrix A, $\circ$ denotes elementwise multiplication and W is a matrix of non-negative tuning parameters, e.g. $W = \lambda 1_{p\times p}$ for $\lambda \geq 0$. This is related to undirected Gaussian graphical models through the fact that if $X = (X_v)_{v\in V} \sim \mathcal{N}(0, \Sigma)$, then X is Markov w.r.t. its *concentration graph* $\mathcal{G} = (V, E)$ with edges $E = \{(u, v) \mid u \neq v, \Theta_{uv} \neq 0\}$. See Lauritzen (1996). Hence a sparse estimated concentation matrix $\hat{\Theta}$ gives rise to a sparse associated concentration graph $\hat{\mathcal{G}}$, which gives simple model interpretations.

This paper is concerned with exploiting the principle of penalized maximum likelihood for structure estimation in Gaussian chain graph models. This has previously been studied in a multivariate regression framework, $Y = BX + \varepsilon$ with $Y \in \mathbb{R}^d$, $X \in \mathbb{R}^p$ and $\varepsilon \sim \mathcal{N}(0, \Theta^{-1})$, which corresponds to a chain graph with two chain components — one for covariates and one for responses. Rothman et al. (2010) and Lin et al. (2016) consider sparse estimation of B and Θ in this set-up, which results in an estimated chain graph in the Andersson-Madigan-Perlman (AMP) Markov interpretation (Andersson et al., 2001). However, this estimator gives rise to a non-convex optimization problem for which there are no guarenties of convergence to a global optimum.

Lee and Liu (2012) and McCarter and Kim (2014) also consider multivariate regression, but in an exponential family parametrization with sparsity inducing penalties on the canonical parameters $\Theta = \Sigma^{-1}$ and $\Lambda = \Theta B$, which gives an estimated chain graph in the Lauritzen-Wermuth-Frydenberg (LWF) Markov interpretation (Frydenberg, 1990; Lauritzen and Wermuth, 1989). Here the estimator gives rise to a convex optimization problem as a result of the exponential family parametrization.

The purpose of this paper is to extend the existing methodology to Gaussian chain graphs for state space models. This extends the usage from multivariate regressions to time series data and allows for the case where the observations are corrupted by additive noise. State space models with sparsity inducing penalties has previously been considered in Noor et al. (2012) and Hasegawa et al. (2014), and our inference approach is similar to what is employed in their work. However, we further give the problem a graphical modeling framework, and we relate the penalization strategy to the chain graph Markov interpretation at hand. The main contributions of this paper are two EM algorithms for performing simultaneous parameter estimation and structure learning of a state space model and its associated chain graph under both the LWF and AMP Markov interpretation.

The paper is organized as follows. First we introduce linear Gaussian state space models and motivate the necessity of chain graphs for giving a detailed description of conditional independence for this model class. Next we give a brief introduction to chain graphs and their Markov properties, and we demonstrate how state space models can be viewed in a chain graph framework. We then develop an E-step and two different M-steps according to the Markov interpretation at hand.

2. Model Formulation

Let us begin by introducing our model class of interest.

Definition 1 *We define a **linear Gaussian state space model** (LGSSM) to be a pair of discrete time stochastic processes (X_t, Y_t) with X_t and Y_t both taking values in $\mathbb{R}^p$ such that*

$$X_t \mid X_{t-1} = x_{t-1} \sim \mathcal{N}(Bx_{t-1}, \Sigma) \quad \textit{and} \quad Y_t \mid X_t = x_t \sim \mathcal{N}(x_t, \rho^2 I_p) \tag{2}$$

for $t = 1, \dots, N$ where X_0 is degenerate at $x_0 \in \mathbb{R}^p$. Here $\Sigma \in \mathcal{S}_p^{++}$ is a covariance matrix, $B \in \mathbb{R}^{p \times p}$ is a matrix of regression coefficients and $\rho^2 \geq 0$. The process (X_t) is assumed to be latent, while the process (Y_t) is observable.

From the distributional specification (2) we have the following factorization of the density of $(X_1, Y_1, \dots, X_N, Y_N)$ conditional on the initial value X_0 of the latent process:

$$f(x_1, y_1, \dots, x_N, y_N \mid x_0) = \prod_{t=1}^{N} f(x_t \mid x_{t-1}) \prod_{t=1}^{N} f(y_t \mid x_t). \tag{3}$$

Therefore the conditional independence structure of the process can be described by a directed acyclic graphical model as in Figure 1. From this DAG we can read of conditional independencies between the variables $X_1, Y_1, \dots, X_N, Y_N$ by using, e.g., d-separation. However, the DAG does not give information about conditional independencies between single coordinates of the processes, e.g., whether there are conditional independencies among $X_{t,1}, \dots, X_{t,p}$ when conditioning on $X_{t-1,1}, \dots, X_{t-1,p}$. In order to provide such a detailed description of the conditional independence structure of the model, we will describe the model in a chain graph setting.

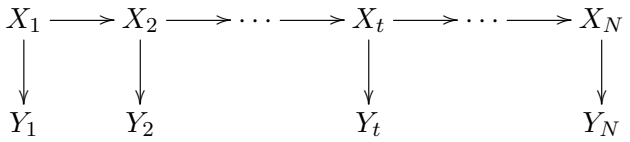

Figure 1: Directed acyclic graphical model for $(X_1, Y_1, \ldots, X_N, Y_N) \mid X_0 = x_0$.

3. Chain Graph Models

We now introduce the basic definition of a chain graph, the two different Markov properties that are usually associated with such graphs and their parametric restrictions in the Gaussian case.

Definition 2 *Let $\mathcal{G} = (V, E)$ be a graph where E is allowed to contain both undirected and directed edges. If $\mathcal{G}$ has no semi-directed cycles, i.e., cycles where all directed edges point in the same direction, then we call $\mathcal{G}$ a **chain graph**. Associated with a chain graph $\mathcal{G}$ we form the directed graph $\mathcal{D} = (\mathcal{T}, \mathcal{E})$, where $\mathcal{T}$ are the connected components of $\mathcal{G}$ after deleting all directed edges, and $\tau \to \tau' \in \mathcal{E}$ for $\tau, \tau' \in \mathcal{T}$ if there exists $u \in \tau$ and $u' \in \tau'$ such that $u \to u' \in E$. We call $\mathcal{D}$ the associated **graph of chain components** of $\mathcal{G}$ and note that the absence of semi-directed cycles in $\mathcal{G}$ ensures that $\mathcal{D}$ is a DAG.*

Chain graphs can be endowed with (at least) two different Markov interpretations, namely the AMP and LWF interpretation, which we will now describe. Let $Z = (Z_v)_{v \in V}$ be a collection of random variables indexed by the vertices of a chain graph $\mathcal{G} = (V, E)$. For a subset of vertices $A \subset V$, we denote by $\text{pa}_{\mathcal{G}}(A)$ and $\text{nb}_{\mathcal{G}}(A)$ the parents and neighbors of A relative to the graph $\mathcal{G}$. Consider the following four properties that Z can potentially fulfill w.r.t. $\mathcal{G}$:

C1) The distribution of Z satisfies the directed local Markov property w.r.t. $\mathcal{D}$.

C2) For each $\tau \in \mathcal{T}$, the distribution of $Z_\tau \mid Z_{\text{pa}_{\mathcal{D}}(\tau)} = z_{\text{pa}_{\mathcal{D}}(\tau)}$ is globally Markov w.r.t. $\mathcal{G}_\tau$.

C3) For each $\tau \in \mathcal{T}$ and $\sigma \subset \tau$ we have $\sigma \perp\!\!\!\perp (\text{pa}_{\mathcal{D}}(\tau) \setminus \text{pa}_{\mathcal{G}}(\sigma)) \mid \text{pa}_{\mathcal{G}}(\sigma) \cup \text{nb}_{\mathcal{G}}(\sigma)$.

C4) For each $\tau \in \mathcal{T}$ and $\sigma \subset \tau$ we have $\sigma \perp\!\!\!\perp (\text{pa}_{\mathcal{D}}(\tau) \setminus \text{pa}_{\mathcal{G}}(\sigma)) \mid \text{pa}_{\mathcal{G}}(\sigma)$.

Here $A \perp\!\!\!\perp B \mid C$ is shorthand for $Z_A \perp\!\!\!\perp Z_B \mid Z_C$ for disjoint $A, B, C \subset V$. From these conditions, we can formulate the two Markov properties that we will associate with chain graphs.

Definition 3 *Let $Z = (Z_v)_{v \in V}$ and $\mathcal{G} = (V, E)$ be as above. If Z satisfies C1, C2 and C3, then we say it has the **LWF Markov property** w.r.t. $\mathcal{G}$. If Z satisfies C1, C2 and C4, we say it has the **AMP Markov property** w.r.t. $\mathcal{G}$.*

As with undirected Gaussian graphical models, the LWF and AMP Markov properties impose certain parametric restrictions for the Gaussian distribution. To describe these, assume further that Z follows a regular multivariate Gaussian distribution $\mathcal{N}(0, \Sigma)$ on $\mathbb{R}^V$. If Z satisfies C1, which is common to the LWF and AMP Markov property, then the distribution of Z is determined by the

conditional distributions of $Z_\tau \mid Z_{\mathrm{pa}_\mathcal{D}(\tau)} = z_{\mathrm{pa}_\mathcal{D}(\tau)}$ for $\tau \in \mathcal{T}$, because the density of Z factorizes according to $\mathcal{D}$. We can write each of these conditional distributions as a multivariate regression

$$Z_\tau \mid Z_{\mathrm{pa}_\mathcal{D}(\tau)} = z_{\mathrm{pa}_\mathcal{D}(\tau)} \sim \mathcal{N}\left(B_\tau z_{\mathrm{pa}_\mathcal{D}(\tau)}, \Sigma_\tau\right) \tag{4}$$

where B_τ is a matrix of regression coefficients. Alternatively we can introduce the parameters $K_\tau = \Sigma_\tau^{-1}$ and $\Lambda_\tau = K_\tau B_\tau$, which are the canonical parameters in an exponential family representation of the distribution, such that

$$Z_\tau \mid Z_{\mathrm{pa}_\mathcal{D}(\tau)} = z_{\mathrm{pa}_\mathcal{D}(\tau)} \sim \mathcal{N}\left(K_\tau^{-1}\Lambda_\tau z_{\mathrm{pa}_\mathcal{D}(\tau)}, K_\tau^{-1}\right). \tag{5}$$

We then have the following description of the parametric restriction implied by the LWF and AMP Markov properties. See Andersson et al. (2001) for details.

Proposition 4 *Let $Z = (Z_v)_{v\in V}$ and $\mathcal{G} = (V, E)$ be as above with Z satisfying C1. Then it holds for any chain component $\tau \in \mathcal{T}$ that*

i) if Z satisfies C2, then $(K_\tau)_{uv} = 0$ for all $u, v \in \tau$ with $u - v \notin E$,

ii) if Z satisfies C3, then $(\Lambda_\tau)_{uv} = 0$ for $u \in \tau$ and $v \in pa_\mathcal{D}(\tau) \setminus pa_\mathcal{G}(u)$,

iii) if Z satisfies C4, then $(B_\tau)_{uv} = 0$ for $u \in \tau$ and $v \in pa_\mathcal{D}(\tau) \setminus pa_\mathcal{G}(u)$.

In conclusion, the AMP Markov property encodes zeros in B_τ and K_τ of the regressions (4), while the LWF Markov property implies zeros in the canonical parameters Λ_τ and K_τ corresponding to the parametrization (5).

4. Chain Graphs for State Space Models

Let us now describe how LGSSMs can be viewed in the chain graph model setting. Naturally, we associate the vertices V of our chain graph $\mathcal{G} = (V, E)$ with the random variables in our LGSSM, and the chain components of our chain graph are the variables $X_1, Y_1, \ldots, X_N, Y_N$. Due to the property C1 and the factorization (3), the graph in Figure 1 must necessarily be the DAG of chain components of the chain graph. Now introduce the canonical parameters $\Theta = \Sigma^{-1}$ and $\Lambda = \Theta B$ such that we can reparametrize our model as an exponential family:

$$X_t \mid X_{t-1} = x_{t-1} \sim \mathcal{N}(\Theta^{-1}\Lambda x_{t-1}, \Theta^{-1}) \quad \text{and} \quad Y_t \mid X_t = x_t \sim \mathcal{N}(\rho^{-2}x_t, (\rho^{-2}I_p)^{-1}). \tag{6}$$

With inspiration from concentration graphs for undirected Gaussian graphical model and the properties i), ii) and iii) we can define the following chain graphs to associate with a LGSSM.

Definition 5 *Let (X_t, Y_t) follow a LGSSM with parameters Λ (B resp.) $\in \mathbb{R}^{p\times p}$, $\Theta \in \mathcal{S}_p^{++}$ and $\rho^2 \geq 0$. Then we define the **LWF (AMP resp.) concentration graph** $\mathcal{G} = (V, E)$ associated with these parameters to have $\mathcal{D}$ in Figure 1 as its associated DAG of chain components and edges E as follows. If $\Theta_{uv} \neq 0$, then we include the undirected edge $X_{t,u} - X_{t,v} \in E$ for each $t = 1, \ldots, N$. If $\rho^2 > 0$, then we let $X_{t,u} \to Y_{t,u} \in E$ for each $t = 1, \ldots, N$ and $u = 1, \ldots, p$. Lastly, if Λ_{uv} (B_{uv} resp.) $\neq 0$, then we include the directed edge $X_{t-1,v} \to X_{t,u} \in E$ for each $t = 2, \ldots, N$.*

Example 1 *Consider a LGSSM in the LWF interpretation with parameters*

$$\Lambda = \begin{pmatrix} * & 0 & 0 \\ * & 0 & * \\ 0 & * & 0 \end{pmatrix}, \quad \Theta = \begin{pmatrix} * & * & * \\ * & * & 0 \\ * & 0 & * \end{pmatrix} \quad \textit{and} \quad \rho^2 > 0 \tag{7}$$

where $*$ refers to some non-zero value. The subgraph of the LWF concentration graph $\mathcal{G}$ containing the latent process (X_t) can be seen in Figure 2. The undirected graphical structure within each

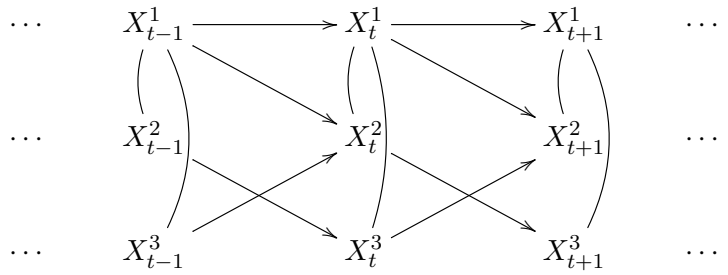

Figure 2: LWF concentration graph $\mathcal{G}$ associated with the parameters (7) restricted to (X_t).

chain component is constructed from Θ analogously with undirected Gaussian graphical models, while the directed edges between chain components are drawn using the zero-pattern of Λ. The full chain graph $\mathcal{G}$ simply has a directed edge from each coordinate of X_t to the corresponding coordinate of Y_t, since there is a non-trivial noise term in this particular example.

It is not hard to show that if (X_t, Y_t) follows a LGSSM with parameters Λ (B resp.) $\in \mathbb{R}^{p\times p}, \Theta \in \mathcal{S}_p^{++}$ and $\rho^2 \geq 0$, then $(X_1, Y_1, \ldots, X_N, Y_N) \mid X_0 = x_0$ will be LWF (AMP resp.) Markov with respect to its associated LWF (AMP resp.) concentration graph.

In conclusion, sparse parameters give rise to sparse chain graphs, which, in turn, give simple model interpretations through the properties C1-C4. In practise the parameters — and thus the graph structure — are unknown and must be estimated from data. However, we cannot expect to estimate entries of the parameters to be exactly zero, and so the need for sparse estimation procedures arise.

5. Sparse Learning via EM Algorithm

Given data $x_0, y_1, \ldots, y_N$ from a LGSSM we will carry out estimation by penalized maximum likelihood with sparsity inducing ℓ_1-penalties on Λ and Θ or B and Θ depending on the chain graph interpretation at hand.

We frame the estimation problem in a missing data set-up and perform inference using the EM algorithm. In this context, the complete data is $x_0, x_1, y_1, \ldots, x_N, y_N$, the missing data is $x_1, \ldots, x_N$ while the observed data is $x_0, y_1, \ldots, y_N$. We will here use a penalized version of the EM algorithm, where penalization is applied in the M-step (Green, 1990). First we derive the E-step, which involves computing the conditional expectation of the complete data log-likelihood given data and current EM estimate.

Proposition 6 *Let $x_0, y_1, \ldots, y_N$ be data from a LGSSM and $\theta^{(k)} = (\Lambda^{(k)}, \Theta^{(k)}, (\rho^2)^{(k)})$ the parameter estimate in the current EM iteration. Then the expected complete data log-likelihood given data and current parameter estimate is given by*

$$Q(\theta \mid \theta^{(k)}) = \log\det\Theta - \text{tr}(\Theta M_1) + 2\text{tr}(\Lambda M_2) - \text{tr}(\Lambda^T\Theta^{-1}\Lambda M_3) - p\log\rho^2 - \frac{1}{\rho^2}M_4$$

where $M_4 \in \mathbb{R}$ and $M_1, M_2, M_3 \in \mathbb{R}^{p\times p}$ depends on data and $\theta^{(k)}$ and are given by

$$\begin{aligned}
M_1 &= \frac{1}{N}\sum_{t=1}^{N} E(X_t X_t^T \mid Y_{1:N} = y_{1:N}, \theta^{(k)}),\\
M_2 &= \frac{1}{N}\sum_{t=1}^{N} E(X_{t-1} X_t^T \mid Y_{1:N} = y_{1:N}, \theta^{(k)}),\\
M_3 &= \frac{1}{N}\sum_{t=1}^{N} E(X_{t-1} X_{t-1}^T \mid Y_{1:N} = y_{1:N}, \theta^{(k)}),\\
M_4 &= \frac{1}{N}\sum_{t=1}^{N} y_t y_t^T - 2y_t^T E(X_t \mid Y_{1:N} = y_{1:N}, \theta^{(k)}) + E(X_t^T X_t \mid Y_{1:N} = y_{1:N}, \theta^{(k)}),
\end{aligned}$$

where we use the shorthand notation $Y_{1:N} = (Y_1, \ldots, Y_N)$.

Proof Due to the factorization (3) we can write the complete data log-likelihood as

$$\begin{aligned}
\ell(\Lambda, \Theta, \rho^2 \mid x, y) =& \frac{N}{2}\log\det\Theta - \frac{1}{2}\sum_{t=1}^{N}(x_t - \Theta^{-1}\Lambda x_{t-1})^T\Theta(x_t - \Theta^{-1}\Lambda x_{t-1})\\
&- \frac{Np}{2}\log\rho^2 - \frac{1}{2\rho^2}\sum_{t=1}^{N}(y_t - x_t)^T(y_t - x_t)
\end{aligned}$$

where we have ignored additive constants. By rescaling with $2/N$ and using the cyclic property of the matrix trace, i.e. $\text{tr}(AB) = \text{tr}(BA)$ for conformable matrices, we obtain

$$\begin{aligned}
\ell(\Lambda, \Theta, \rho^2 \mid x, y) \propto& \log\det\Theta - \frac{1}{N}\sum_{t=1}^{N}\text{tr}(\Theta x_t x_t^T) - 2\text{tr}(\Lambda x_{t-1} x_t^T) + \text{tr}(\Lambda^T\Theta^{-1}\Lambda x_{t-1} x_{t-1}^T)\\
&- p\log\rho^2 - \frac{1}{\rho^2}\frac{1}{N}\sum_{t=1}^{N} y_t^T y_t - 2y_t^T x_t + x_t^T x_t.
\end{aligned}$$

Taking conditional expectation with respect to data and current EM estimate and using linearity of the trace we obtain the wanted result. ■

Note that we have formulated the E-step in terms of the canonical parameters, but it is always possible to re-parametrize using $\Lambda = \Theta B$, and θ is simply a placeholder for the parameters under consideration in what follows. The conditional expectations that are needed when computing the

quantities $M_1, \ldots, M_4$ are the topic of smoothing in hidden Markov models, and one can use, e.g., the Rauch-Tung-Striebel smoother. See, e.g., Särkkä (2013) for details.

We now turn to the M-step. In the LWF interpretation we parametrize the expected complete data log-likelihood using canonical parameters and put sparsity inducing ℓ_1-penalties on Λ and Θ. The next EM iteration is produced by carrying out the optimization:

$$\hat{\theta}^{(k+1)} = \underset{(\Lambda,\Theta,\rho^2)\in\mathbb{R}^{p\times p}\times\mathcal{S}_p^{++}\times\mathbb{R}_+}{\arg\min} \left\{ f_1(\Lambda,\Theta) + f_3(\rho^2) + \lambda_1||\Lambda||_{1,\text{off}} + \lambda_2||\Theta||_{1,\text{off}} \right\}. \tag{8}$$

Here the function f_1 is the part of the negative expected complete data log-likelihood that depends on the parameters Λ and Θ,

$$f_1(\Lambda,\Theta) = \text{tr}(\Theta M_1) - 2\text{tr}(\Lambda M_2) + \text{tr}(\Lambda^T\Theta^{-1}\Lambda M_3) - \log\det\Theta,$$

and $f_3(\rho^2) = p\log\rho^2 + M_4/\rho^2$ is the part that depends on ρ^2. We let $||\Lambda||_{1,\text{off}} = \sum_{i\neq j}|\Lambda_{ij}|$, i.e. we choose not to penalize the diagonal. The numbers $\lambda_1, \lambda_2 \geq 0$ are tuning parameters determining the sparsity level of the estimates.

In the AMP interpretation we parametrize the expected complete data log-likelihood using the regression matrix and put ℓ_1-penalties on B and Θ. The M-step is then the optimization:

$$\hat{\theta}^{(k+1)} = \underset{(B,\Theta,\rho^2)\in\mathbb{R}^{p\times p}\times\mathcal{S}_p^{++}\times\mathbb{R}_+}{\arg\min} \left\{ f_2(B,\Theta) + f_3(\rho^2) + \lambda_1||B||_{1,\text{off}} + \lambda_2||\Theta||_{1,\text{off}} \right\}. \tag{9}$$

Here f_2 is the part that depends on the parameters B and Θ,

$$f_2(B,\Theta) = \text{tr}(\Theta M_1) - 2\text{tr}(\Theta B M_2) + \text{tr}(B^T\Theta B M_3) - \log\det\Theta,$$

and $f_3(\rho^2) = p\log\rho^2 + M_4/\rho^2$ as before.

Note that in both (8) and (9) there is variation independence between ρ^2 and the remaining parameters. Hence the optimization regarding ρ^2 can be performed separately, and is given by the conditional expectation of the empirical residual variance of the regression from X_t to Y_t:

$$(\hat{\rho}^2)^{(k+1)} = \frac{1}{Np}\sum_{t=1}^{N} E\left((Y_t - X_t)^T(Y_t - X_t) \mid Y_{1:N} = y_{1:N}, \theta^{(k)}\right) = \frac{1}{p}M_4. \tag{10}$$

Next we turn to the problem of performing the optimization in (8) regarding Λ and Θ. As we shall see, this task can be re-formulated into an equivalent optimization problem. Let the $2p \times 2p$ matrices $\mathbf{T}(\Lambda,\Theta)$ and $\mathbf{M}$ be given by

$$\mathbf{T}(\Lambda,\Theta) = \begin{pmatrix} \mathbf{T}_{11} & \mathbf{T}_{12} \\ \mathbf{T}_{12}^T & \mathbf{T}_{22} \end{pmatrix} = \begin{pmatrix} \Theta & -\Lambda \\ -\Lambda^T & I_p + \Lambda^T\Theta^{-1}\Lambda \end{pmatrix} \quad \text{and} \quad \mathbf{M} = \begin{pmatrix} M_1 & M_2^T \\ M_2 & M_3 \end{pmatrix}.$$

Consider the optimization problem

$$\min_{(\Lambda,\Theta)\in\mathbb{R}^{p\times p}\times\mathcal{S}_p^{++}} \{\text{tr}(\mathbf{T}(\Lambda,\Theta)\mathbf{M}) - \log\det\mathbf{T}(\Lambda,\Theta) + ||W\circ\mathbf{T}(\Lambda,\Theta)||_1\} \tag{11}$$

where W is the $2p \times 2p$ matrix

$$W = \begin{pmatrix} \lambda_2 E & \frac{1}{2}\lambda_1 E \\ \frac{1}{2}\lambda_1 E & 0_{p\times p} \end{pmatrix}$$

where E is the $p \times p$ matrix given by $E_{kk} = 0$ for $k = 1, \ldots, p$ and $E_{ij} = 1$ for $i \neq j$.

Proposition 7 *Performing the optimization in* (8) *regarding* Λ *and* Θ *is equivalent with solving the optimization problem* (11)*. Furthermore,* (11) *can be solved by applying graphical lasso* (1) *with optimization variable* $\boldsymbol{T}$*, empirical covariance matrix* $S = \boldsymbol{M}$ *and penalization matrix* W*, and afterwards extracting* $\hat{\Theta}^{(k+1)} = \hat{\boldsymbol{T}}_{11}$ *and* $\hat{\Lambda}^{(k+1)} = -\hat{\boldsymbol{T}}_{12}$*.*

Proof We show the equivalence by simply writing out the objective function of (11) and compare with (8). First we see that the trace term of (11) can be written

$$\begin{aligned}\operatorname{tr}(\mathbf{T}(\Lambda,\Theta)\mathbf{M}) &= \operatorname{tr}(\Theta M_1 - \Lambda M_2 - \Lambda^T M_2^T + M_3 + M_3\Lambda^T\Theta^{-1}\Lambda) \\ &= \operatorname{tr}(\Theta M_1) - 2\operatorname{tr}(\Lambda M_2) + \operatorname{tr}(\Lambda^T\Theta^{-1}\Lambda M_3) + \operatorname{tr}(M_3),\end{aligned}$$

and the determinant of $\mathbf{T}(\Lambda,\Theta)$ is equal to

$$\det \mathbf{T}(\Lambda,\Theta) = \det\Theta\det(I_p + \Lambda^T\Theta^{-1}\Lambda - (-\Lambda^T)\Theta^{-1}(-\Lambda)) = \det\Theta\det I_p = \det\Theta.$$

Lastly, we clearly have $||W \circ \mathbf{T}(\Lambda,\Theta)||_1 = \lambda_1||\Lambda||_{1,\text{off}} + \lambda_2||\Theta||_{1,\text{off}}$. Comparing to the part of the objective function of (8) concerning Λ and Θ, we see that the two objective function are equal up to the additive constant $\operatorname{tr}(M_3)$. This constant is computed using the current EM estimate $\theta^{(k)}$, but does not depend on the optimization variable θ, so it does not affect the optimization.

Let us argue that (11) can be solved by graphical lasso. First we note that the objective function has the correct functional form when comparing to (1). Secondly, we have $\mathbf{T}(\Lambda,\Theta) \in \mathcal{S}_p^{++}$ if and only if $\Theta \in \mathcal{S}_p^{++}$ so that the optimization domains are in fact equal. This is realized by using the Schur complement characterization of positive definiteness, i.e. that $\mathbf{T} \in \mathcal{S}_p^{++}$ if and only if $\mathbf{T}_{11} \in \mathcal{S}_p^{++}$ and $\mathbf{T}/\mathbf{T}_{11} \in \mathcal{S}_p^{++}$. Since $\mathbf{T}_{11} = \Theta$ and $\mathbf{T}/\mathbf{T}_{11} = I_p$ we conclude the wanted. ■

Input: Data $x_0, y_1, \ldots, y_N$, initial parameter values $\theta^{(0)}$ and $\lambda_1, \lambda_2 \geq 0$.
Output: Sparse parameter estimates $\hat{\Lambda}, \hat{\Theta}$ and $\hat{\rho}^2$.
begin
  $k \leftarrow 0$;
  repeat
    Compute $M_1, \ldots, M_4$ by smoothing using data and current estimate $\theta^{(k)}$;
    Update ρ^2 using (10);
    Solve the optimization (11) using graphical lasso to obtain $\hat{\mathbf{T}}$;
    Update Λ and Θ using estimate $\hat{\mathbf{T}}$ as described in Proposition 7;
    $k \leftarrow k + 1$;
  until *convergence criterion is met*;
  return $(\Lambda^{(k)}, \Theta^{(k)}, (\rho^2)^{(k)})$;
end

Algorithm 1: EM algorithm for sparse estimation of Λ, Θ and ρ^2 in a LGSSM.

We now turn to the optimization (9) regarding the parameters B and Θ. First note that the optimization problem (8) is convex, which is due to f_1 being the (expected) negative log-likelihood of an exponential family and that the ℓ_1-penalty is convex. However, the function f_2 is not jointly convex in B and Θ, but it is bi-convex, i.e. $B \mapsto f_2(B, \Theta_0)$ and $\Theta \mapsto f_2(B_0, \Theta)$ are convex for fixed $\Theta_0 \in \mathcal{S}_p^{++}$ and $B_0 \in \mathbb{R}^{p\times p}$ respectively. See Lee and Liu (2012) for a discussion. Therefore,

we will perform the optimization regarding B and Θ using an alternating convex search. More specifically, set $B_*^{(0)} := B^{(k)}$ and $\Theta_*^{(0)} := \Theta^{(k)}$, and then perform the optimizations

$$\Theta_*^{(i+1)} = \underset{\Theta \in \mathcal{S}_p^{++}}{\arg\min} \left\{ f_2(B_*^{(i)}, \Theta) + \lambda_2 ||\Theta||_{1,\text{off}} \right\}, \tag{12}$$

$$B_*^{(i+1)} = \underset{B \in \mathbb{R}^{p \times p}}{\arg\min} \left\{ f_2(B, \Theta_*^{(i+1)}) + \lambda_1 ||B||_{1,\text{off}} \right\} \tag{13}$$

for $i = 0, 1, \ldots$ until convergence. Then set $B^{(k+1)} := B_*^{(\infty)}$ and $\Theta^{(k+1)} := \Theta_*^{(\infty)}$ at convergence. The following proposition gives a way of solving (12) using existing methods.

Proposition 8 *The optimization* (12) *can be solved with graphical lasso with empirical covariance matrix* $S = M_1 - 2B_*^{(i)} M_2 + B_*^{(i)} M_3 (B_*^{(i)})^T$ *and penalty matrix* $W = \lambda_2 E$ *where* E *is as before.*

Proof We observe that

$$\begin{aligned} f_2(B_*^{(i)}, \Theta) &= \text{tr}(\Theta M_1) - 2\text{tr}(\Theta B_*^{(i)} M_2) + \text{tr}((B_*^{(i)})^T \Theta B_*^{(i)} M_3) - \log\det\Theta \\ &= \text{tr}(\Theta(M_1 - 2B_*^{(i)} M_2 + B_*^{(i)} M_3 (B_*^{(i)})^T)) - \log\det\Theta \end{aligned}$$

such that the objective function of (12) matches that of the graphical lasso problem (1) with $S = M_1 - 2B_*^{(i)} M_2 + B_*^{(i)} M_3 (B_*^{(i)})^T$. Note that this is a valid empirical covariance matrix since in fact

$$S = E\left(\frac{1}{N} \sum_{t=1}^{N} (X_t - B_*^{(i)} X_{t-1})(X_t - B_*^{(i)} X_{t-1})^T \mid Y_{1:N} = y_{1:N}, \theta^{(k)} \right),$$

i.e. S is the conditional expectation of the empirical covariance of the fitted residuals for the regression from X_{t-1} to X_t given data and current EM estimate. ■

Just as (12) turned out to be solvable by applying graphical lasso, also (13) can be solved by existing methods, namely the lasso estimator (Tibshirani, 1996). The lasso estimates $\beta \in \mathbb{R}^p$ in the general linear model $y = A\beta + \varepsilon$, where A is a $N \times p$ design matrix and $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_p)$, by

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \left\{ \frac{1}{2N} (y - A\beta)^T (y - A\beta) + ||\lambda \circ \beta||_1 \right\} \tag{14}$$

with $\lambda \in \mathbb{R}^p$ a vector of non-negative tuning parameters.

Proposition 9 *The optimization* (13) *can be solved by applying lasso regression in the following way. Let* $\hat{\beta}$ *be the result of a lasso regression with design matrix* A *and response vector* y *given by*

$$A = \sqrt{2N}(M_3 \otimes \Theta_*^{(i+1)})^{1/2} \quad \textit{and} \quad y = \sqrt{2N}(M_3 \otimes \Theta_*^{(i+1)})^{1/2} \textit{vec}(M_2^T M_3^{-1}),$$

and tuning parameter $\lambda = \lambda_1 R$ *where* $R \in \mathbb{R}^{p^2}$ *with* $R_1 = R_{p+2} = R_{2p+3} = \cdots = R_{p^2} = 0$ *and all other entries are* 1*. Here* $\otimes$ *denotes the Kronecker product, vec denotes vectorization of matrices and* $C^{1/2}$ *denotes the square root of a matrix* C*. Then* $B_*^{(i+1)}$ *is given by setting* $\textit{vec}(B_*^{(i+1)}) := \hat{\beta}$*.*

Proof Writing out the lasso objective function yields

$$\frac{1}{2N}(y - A\beta)^T(y - A\beta) + \lambda_1||\beta||_1 = \text{const.} + \frac{1}{2N}\beta^T A^T A\beta - \frac{1}{N}y^T A\beta + \lambda||\beta||_1, \qquad (15)$$

while writing out the objective function of (13) gives

$$f_2(B, \Theta_*^{(i+1)}) + \lambda_1||B||_{1,\text{off}} = \text{const.} + \text{tr}(B^T\Theta_*^{(i+1)}BM_3) - 2\text{tr}(\Theta_*^{(i+1)}BM_2) + \lambda_1||B||_{1,\text{off}}.$$

First note the useful relation between the matrix trace and the kronecker product and vectorization of matrices, $\text{tr}(ABCD) = \text{vec}(A^T)^T(D^T \otimes B)\text{vec}(C)$, where A, B, C and D are conformable matrices. Using this relation we can write

$$\text{tr}(B^T\Theta_*^{(i)}BM_3) = \text{vec}(B)^T(M_3 \otimes \Theta_*^{(i+1)})\text{vec}(B)$$

and also

$$2\text{tr}(\Theta_*^{(i)}BM_2) = 2\text{tr}(M_3^{-1}M_2\Theta_*^{(i+1)}BM_3) = 2\text{vec}(M_2^T M_3^{-1})^T(M_3 \otimes \Theta_*^{(i+1)})\text{vec}(B).$$

Letting A and y be as in the proposition and pluggin into the lasso objective function (15) we recover the objective function of (13) written in terms of vec and $\otimes$ as proposed. ■

Input: Data $x_0, y_1, \ldots, y_N$, initial parameter values $\theta^{(0)}$ and $\lambda_1, \lambda_2 \geq 0$.
Output: Sparse parameter estimates $\hat{B}, \hat{\Theta}$ and $\hat{\rho}^2$.
begin
  $k \leftarrow 0$;
  repeat
    Compute $M_1, \ldots, M_4$ by smoothing using data and current estimate $\theta^{(k)}$;
    Update ρ^2 using (10);
    Set $B_*^{(0)} \leftarrow B^{(k)}$, $\Theta_*^{(0)} \leftarrow B^{(k)}$ and $i \leftarrow 0$;
    repeat
      Update $\Theta_*^{(i+1)}$ using Proposition 8;
      Update $B_*^{(i+1)}$ using Proposition 9;
      $i \leftarrow i + 1$;
    until *convergence criterion is met*;
    Set $B^{(k+1)} \leftarrow B_*^{(\infty)}$, $\Theta^{(k+1)} \leftarrow \Theta_*^{(\infty)}$ and $k \leftarrow k + 1$;
  until *convergence criterion is met*;
  return $(B^{(k)}, \Theta^{(k)}, (\rho^2)^{(k)})$;
end

Algorithm 2: EM algorithm for sparse estimation of B, Θ and ρ^2 in a LGSSM.

6. Simulations

In this section we evaluate convergence of our proposed algorithms by means of simulation. For the case $p = 40$ and $N = 200$ we simulate valid true parameters for the LWF and AMP model such

that each of the matrices Θ, Λ and B has 75% zero entries, and ρ^2 is chosen to be 0.1 times the average of the diagonal of $\Sigma = \Theta^{-1}$. We then simulate 100 independent data set from each of the two LGSSMs and perform estimation using Algorithm 1 for LWF-data and Algorithm 2 for AMP-data. For tuning parameters λ_1, λ_2 we choose values ad hoc that do not produce neither completely sparse nor completely dense solutions. For each variable, say Θ, we track the relative difference from one EM iteration to the next by computing $\text{RD}_\Theta(k) := ||\Theta^{(k)} - \Theta^{(k-1)}||_F \cdot ||\Theta^{(k-1)}||_F^{-1}$, where $||\cdot||_F$ is the Frobenious norm. The results can be seen in Figure 3. We observe that the relative

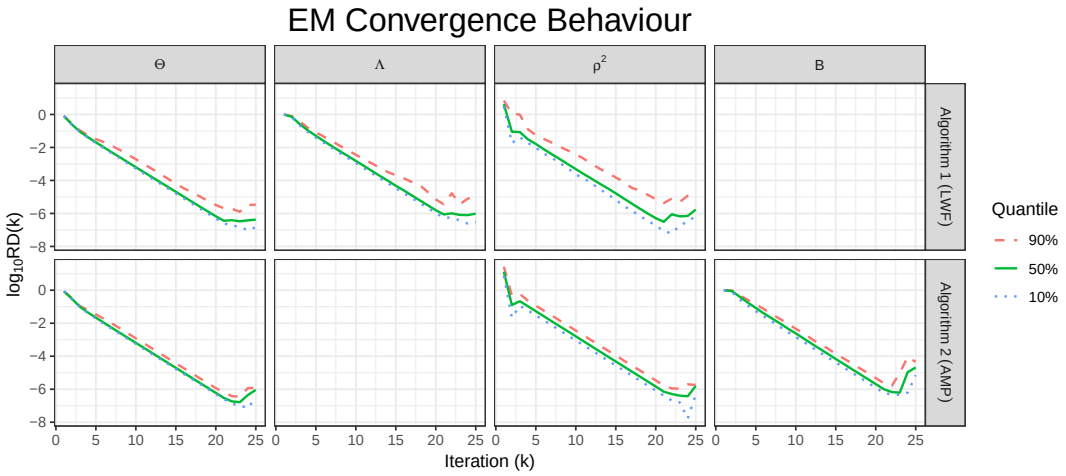

Figure 3: Selected quantiles of $\log_{10} \text{RD}(k)$ computed for each variable at each iteration k based on 100 runs of Algorithm 1 and Algorithm 2 on simulated data with fixed true parameters. The algorithms were terminated when each relative difference dropped below 10^{-6}.

difference decreases approximately linearly on $\log_{10}$-scale. Moreover, the convergence behaviour is stable over the 100 runs. On average Algorithm 1 took 22.23 iterations before convergence, while Algorithms 2 needed 22.61 iterations on average before convergence. On average Algorithm 2 took 11.23 times longer than Algorithm 1 before convergence.

7. Conclusion

The purpose of this paper was to give a chain graph model framework for linear Gaussian state space model and develop algorithms for performing parameter estimation and structure learning from empirical data. We have proposed two different EM algorithms for performing this task depending on the chain graph interpretation (LWF or AMP) at hand, and we have justified convergence of the algorithms empirically through simulation. Next steps include developing methods for choosing the tuning parameters of the algorithms. This will enable us to consider edge-recovery properties of the algorithms and, moreover, make the algorithms useful in real world applications of the models.

Acknowledgments

This work was supported by a research grant (13358) from VILLUM FONDEN.

References

S. A. Andersson, D. Madigan, and M. D. Perlman. Alternative Markov properties for chain graphs. *Scand. J. Statist.*, 28(1):33–85, 2001. ISSN 0303-6898.

O. Banerjee, L. E. Ghaoui, and A. d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *Journal of Machine learning research*, 9(Mar):485–516, 2008.

J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

M. Frydenberg. The chain graph Markov property. *Scand. J. Statist.*, 17(4):333–353, 1990. ISSN 0303-6898.

P. J. Green. On use of the EM algorithm for penalized likelihood estimation. *J. Roy. Statist. Soc. Ser. B*, 52(3):443–452, 1990. ISSN 0035-9246.

T. Hasegawa, R. Yamaguchi, M. Nagasaki, S. Miyano, and S. Imoto. Inference of gene regulatory networks incorporating multi-source biological knowledge via a state space model with l1 regularization. *PLoS One*, 9(8):e105942, 2014.

S. L. Lauritzen. *Graphical models*, volume 17 of *Oxford Statistical Science Series*. The Clarendon Press, Oxford University Press, New York, 1996. ISBN 0-19-852219-3. Oxford Science Publications.

S. L. Lauritzen and N. Wermuth. Graphical models for associations between variables, some of which are qualitative and some quantitative. *Ann. Statist.*, 17(1):31–57, 1989. ISSN 0090-5364.

W. Lee and Y. Liu. Simultaneous multiple response regression and inverse covariance matrix estimation via penalized gaussian maximum likelihood. *Journal of multivariate analysis*, 111:241–255, 2012.

J. Lin, S. Basu, M. Banerjee, and G. Michailidis. Penalized maximum likelihood estimation of multi-layered gaussian graphical models. *J. Mach. Learn. Res.*, 17(1):5097–5147, Jan. 2016. ISSN 1532-4435.

C. McCarter and S. Kim. On sparse gaussian chain graph models. In *Advances in Neural Information Processing Systems*, pages 3212–3220, 2014.

A. Noor, E. Serpedin, M. Nounou, and H. Nounou. Inferring gene regulatory networks via nonlinear state-space models and exploiting sparsity. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 9(4):1203–1211, 2012.

A. J. Rothman, E. Levina, and J. Zhu. Sparse multivariate regression with covariance estimation. *Journal of Computational and Graphical Statistics*, 19(4):947–962, 2010.

S. Särkkä. *Bayesian filtering and smoothing*, volume 3 of *Institute of Mathematical Statistics Textbooks*. Cambridge University Press, Cambridge, 2013. ISBN 978-1-107-61928-9.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.