A Bayesian Approach to Sparse plus Low rank Network Identification

Mattia Zorzi and Alessandro Chiuso

Abstract—We consider the problem of modeling multivariate time series with parsimonious dynamical models which can be represented as sparse dynamic Bayesian networks with few latent nodes. This structure translates into a sparse plus low rank model. In this paper, we propose a Gaussian regression approach to identify such a model.

I. INTRODUCTION

Prediction error method (PEM) is a consolidate paradigm in system identification [14], [21]. In the traditional setting, candidate models are described in fixed parametric model structures, e.g. ARMAX, whose complexity is determined using cross validation or information based criteria. Regularization has been recently introduced in the PEM framework, see [15], [16], [6], [17], [19], as an alternative approach to control complexity of the estimated models. This latter class of methods start with a large enough (in principle infinite dimensional) model class; the inverse (ill-posed) problem of determining a specific model from a finite set of measured data can be made into a well posed problem using a penalty term, whose duty is to select models with specific features. In the Bayesian view, this is equivalent to the introduction of a prior probability on the model to estimate. For instance, the prior should account the fact that the model is more likely a prior probability on the model to estimate. For instance, the prior should account the fact that the model is more likely to predict 

In Section IV, we introduce the S+L network identification paradigm which takes into account the presence of few latent variables. In this way, the manifest variables Granger causes each other mostly through latent variables. The corresponding network has a two layer structure: one layer denotes the manifest variables and the other one the latent variables. The presence of few latent variables should drastically reduce the edges in the manifest layer, therefore increasing the degree of conciseness and robustness of the model. The identification paradigm we propose relies on the Bayesian estimation of predictor impulse responses and exploits the sparse plus low rank structure which characterizes the above network.

The outline of the paper follows. In Section II we review the Gaussian regression approach to estimate the predictor impulse responses. In Section III we introduce the sparse plus low rank (S+L) models for multivariate time series. In Section IV we introduce the S+L network identification problem. In Section V and Section VI we present our S+L network identification procedure. In Section VII we provide some numerical examples to show the effectiveness of our method. Finally, the conclusions are in Section VIII.

Notation

Throughout the paper, we will use the following notation. \( S_m \) denotes the vector space of symmetric matrices of dimension \( m \times m \). \( S_m^+ \) denotes the cone of the positive definite matrices, and \( \overline{S}_m \) its closure. Given \( v \in \mathbb{R}^m \) and \( G \in \mathbb{R}^{n \times m} \), \( v_i \) denotes the \( i \)-th entry of \( v \) and \( [G]_{ij} \) denotes the entry of \( G \) in position \((i, j)\), \( \| G \|_Q \) denotes the weighted Frobenius norm of \( G \) with weight matrix \( Q \in S_m^+ \). Given a stochastic process \( y = \{ y(t) \}_{t \in \mathbb{Z}} \) with some abuse of notation, \( y(t) \) will both denote a random vector and its sample value. Finally,

\[
y^-(t) := \left[ y(t - 1)^T \ y(t - 2)^T \ \ldots \right]^T
\]

(1)

denotes the past data vectors of \( y \) at time \( t \). In similar way, \( y_i^-(t) \) denotes the past data of \( y_i \).

II. GAUSSIAN REGRESSION APPROACH TO SYSTEM IDENTIFICATION

Consider a zero mean stationary Gaussian stochastic process \( y \) of dimension \( m \). It is well known that \( y \) can be represented as the output of a \( m \times m \) stable and minimum phase filter \( W(z) = I_m + \sum_{k=1}^{\infty} W_k z^{-k} \)

\[
y(t) = W(z) e(t)
\]

(2)

where \( e \) is the innovation process with covariance matrix \( \Sigma \). Moreover, it is not difficult to see that

\[
y(t) = G(z)y(t) + e(t)
\]

(3)
where $G(z) = I_m - W(z)^{-1}$ is the transfer function of the one step ahead predictor $\hat{y}(t)$ of $y(t)$ based on the observations $y^-(t)$, i.e.

$$
\hat{y}(t) = G(z)y(t).
$$

In the PEM framework, system identification (i.e. estimation of the transfer matrix $W(z)$ from finite data $y(1) \ldots y(N)$) can be framed as estimation of $G(z)$. Note that, $G(z)$ admits the decomposition $G(z) = \sum_{k=1}^{\infty} G_k z^{-k}$. Since $G(z)$ is BIBO stable, we use the approximation

$$
G(z) \approx \sum_{k=1}^{T} G_k z^{-k}
$$

where $T$ is sufficiently large. The parameters of the truncated transfer matrix are stacked in the vector $\theta \in \mathbb{R}^{m^2 T}$ which is defined as follows

$$
\theta = \begin{bmatrix}
  (g^{[1]}_1)^\top & \ldots & (g^{[m]}_1)^\top \\
  \vdots & \ddots & \vdots \\
  (g^{[1]}_m)^\top & \ldots & (g^{[m]}_m)^\top
\end{bmatrix}^\top
$$

where

$$
g^{[ij]} = \begin{bmatrix}
  G_1_{ij} & G_2_{ij} & \ldots & G_T_{ij}
\end{bmatrix}^\top
$$

denotes the impulse response coefficients of the transfer function in position $(i,j)$ of the transfer matrix $G(z)$. Then, we stack the data in the vector $y$ as follows

$$
y = \begin{bmatrix}
  y_1(T+1)^\top & \ldots & y_1(N)^\top \\
  \vdots & \ddots & \vdots \\
  y_m(T+1)^\top & \ldots & y_m(N)^\top
\end{bmatrix}^\top
$$

In similar way, we define

$$
e = \begin{bmatrix}
  e_1(T+1)^\top & \ldots & e_1(N)^\top \\
  \vdots & \ddots & \vdots \\
  e_m(T+1)^\top & \ldots & e_m(N)^\top
\end{bmatrix}^\top
$$

Let $\Phi \in \mathbb{R}^{mN \times m^2 T}$ be the regression matrix defined as

$$
\Phi = I_m \otimes \begin{bmatrix}
  \phi_1 & \ldots & \phi_m
\end{bmatrix}
$$

with $\phi_j \in \mathbb{R}^{N \times T}$ such that

$$
\phi_j = \begin{bmatrix}
  y_{j1}(T) & y_{j1}(T-1) & \ldots & y_{j1}(1) \\
  y_{j1}(T+1) & y_{j1}(T) & \ldots & y_{j1}(2) \\
  \vdots & \vdots & \ddots & \vdots \\
  y_{j1}(N-1) & y_{j1}(N-2) & \ldots & y_{j1}(N-T)
\end{bmatrix}
$$

From (13) the vector of the observations can be expressed as follows

$$
y = \Phi \theta + e
$$

where $\Phi \theta$ is the one step ahead predictor of $y$. Accordingly, the usual PEM estimator is obtained solving

$$
\hat{\theta}_{PEM} = \arg \min_{\theta} \|y - \Phi \theta\|^2_{\Sigma^{-1} \otimes I_N}.
$$

In this paper, we consider the Gaussian regression framework in [20]. Therefore, we model $\theta$ as a zero mean Gaussian random vector with covariance matrix, or kernel, denoted by $K$. The matrix $K \in S_m^+$ is parametrized through some hyperparameters $\xi$, see below. A reasonable assumption is that the past data $y^-(T+1)$ neither affect the priori probability on $\theta$ nor carries information on $\xi$, [18], that is

$$
p(y, \theta, y^-(T+1)|\xi) = p(y|\theta, y^-(T+1), \xi)p(\theta|y^-(T+1)|\xi) \approx p(y|\theta, y^-(T+1), \xi)p(\theta|\xi)p(y^-(T+1)).
$$

Let $\hat{\theta}$ be the posterior mean of $\theta$ according to model (14). In [15] it has been proved that, under assumption (14), $\theta$ can also be written as solution to the Tikhonov regularization problem

$$
\hat{\theta} = \arg \min_{\theta} \|y - \Phi \theta\|^2_{\Sigma^{-1} \otimes I_N} + \|\theta\|^2_{K^{-1}}.
$$

Moreover, it is not difficult to see that

$$
\hat{\theta} = K\Phi^\top (\Phi K\Phi^\top + \Sigma \otimes I_N)^{-1} y.
$$

Remark 2.1: Using the theory of reproducing kernel Hilbert spaces, [3], it is possible to show that the above results still hold for $T \to \infty$, [15].

Remark 2.2: Although we assumed $K \in S_m^+$, the Bayes estimator also holds for $K$ singular. In that case, problem (15) is well defined provided that $\theta$ belongs to the range of $K$.

The optimal solution $\hat{\theta}$ highly depends on the choice of $K$. A typical assumption is that the transfer functions in $G(z)$ are independent, that is $g^{[ij]}$ are independent vectors. We shall also assume that $g^{[ij]}$, with $i,j = 1 \ldots m$, are identically distributed, so that

$$
K = I_m^2 \otimes \tilde{K}
$$

where $\tilde{K} \in \mathbb{R}^{T \times T}$ is the covariance matrix of $g^{[ij]}$, with $i,j = 1 \ldots m$. In [15], the authors proposed the following structure for $K$

$$
\tilde{K} = \lambda^{-1} F_0 K_\beta F_0^\top.
$$

The nonnegative parameter $\lambda$ is usually called "scaling factor". The kernel matrix $K_\beta$ enforces stability of the predictor impulse responses as $T \to \infty$. Many choices can be considered, see [16] and [6]. In this paper we consider the tuned/correlated (TC) kernel,

$$
[K_\beta]_{t,s} = \min\{e^{-\beta t}, e^{-\beta s}\}, \ t, s = 1 \ldots T
$$

where $\beta < 0$ represents the decay rate of the impulse responses. $F_0$ is the parametric component which models high frequency oscillations. Indeed, predictor impulse responses may have some fast dynamics which cannot be captured by $K_\beta$. The kernel matrix $K$ is therefore characterized by the hyperparameters $\xi = \{\lambda, \beta, \theta\}$. $\xi$ is computed minimizing the negative log-marginal likelihood of $y$, [20], denoted by $\ell(y)$. Under the assumption (14), we have

$$
\ell(y) = \frac{1}{2} \log \det(V) + \frac{1}{2} y^\top V^{-1} y
$$

where

$$
V = \Phi K\Phi^\top + \Sigma \otimes I_N.
$$
III. SPARSE PLUS LOW RANK MODELS

Consider two zero mean stationary Gaussian stochastic processes \( y \) and \( x \) of dimension \( m \) and \( n \), respectively. Let \( y \) be manifest, i.e. it can be measured, and \( x \) latent, i.e. it is not accessible for observation. We assume that \( y \in \mathbb{R}^m \) and \( x \in \mathbb{R}^n \) are described by the model

\[
\begin{align*}
y(t) &= Fx(t) + S(z)y(t) + v(t) \\
x(t) &= Hz(y(t)) + w(t)
\end{align*}
\]

where \( H(z) = \sum_{k=1}^{\infty} H_k z^{-k} \) is a BIBO stable transfer matrix of dimension \( n \times m \), \( F \in \mathbb{R}^{m \times n} \), \( S(z) = \sum_{k=1}^{\infty} S_k z^{-k} \) is a BIBO stable transfer matrix of dimension \( m \times m \), \( v \) and \( w \) are, respectively, \( m \) and \( n \) dimensional white Gaussian noises (WGN) with zero mean and covariance matrix \( \Sigma_v \) and \( \Sigma_w \).

It is possible to describe the structure of model (22) using a (Bayesian) network [13] as in Figure 1. Each node of that network corresponds to a component of \( y \) or \( x \) and there is a direct link from the node of \( y_j \) (\( x_j \)) to the node of \( y_i \) (\( x_i \)) if and only if \( y_j^* \) (\( x_j^* \)) is needed to predict \( y_i(t) \) (\( x_i(t) \)), that is \( y_j \) (\( x_j \)) Granger causes \( y_i \) (\( x_i \)) [11]. In Figure 1 we provide an example with \( m = 6 \) and \( n = 1 \).

![Fig. 1. Example of a sparse plus low rank model with \( m = 6 \) and \( n = 1 \).](image)

In particular \( \hat{y}_5(t) \) is a function of \( \hat{x}_1(t) \) and \( \hat{y}_j(t) \), the remaining predictors \( \hat{y}_i(t) \) with \( i \neq 5 \) are only functions of \( \hat{x}_1(t) \), and \( \hat{x}_1(t) \) is a function of \( y^-(t) \). Our main modeling assumptions in (22) is that manifest variables Granger causes each other mostly through few latent variables. Therefore, we have \( n \ll m \), i.e. the number of latent factors in \( x \) is small as compared to the cross-sectional dimension of \( y \), and \( S(z) \) is sparse, i.e. many of its entries are null transfer functions, so that few manifest variables directly Granger causes each other. From (22), we obtain the sparse plus low rank (S+L) model for \( y \):

\[
y(t) = S(z)y(t) + L(z)y(t) + e(t)
\]

where \( S(z) \) is a sparse transfer matrix by assumption, \( L(z) := FH(z) \) is a low rank transfer matrix because \( F \) and \( H^T(z) \) are tall matrices, and \( e := v + w \) is WGN with covariance matrix \( \Sigma = \Sigma_v + F\Sigma_w F^\top \).

Let \( \hat{y}(t) \) and \( \hat{x}(t) \) be the minimum variance one step ahead predictor of \( y(t) \) and \( x(t) \), respectively, based on the observations \( y^-(t) \). From (22), we have

\[
\begin{align*}
\hat{x}(t) &= H(z)y(t) \\
\hat{y}(t) &= F\hat{x}(t) + S(z)y(t)
\end{align*}
\]

Accordingly,

\[
\hat{y}(t) = S(z)y(t) + L(z)y(t)
\]

that is, the predictable part of \( y \) is a function of few predicted latent factors and of the pasts of few components of \( y \). It is worth noting that in the case that \( L(z) = 0 \), i.e. there is no need of latent factors to characterize the predictor of \( y \), we obtain the sparse model presented in [7]. In the case \( S(z) = 0 \), i.e. the predictor of \( y \) is completely characterized by the predictors of the latent factors, we obtain a quasi static factor model, see for instance [9].

IV. PROBLEM FORMULATION

Consider the model (22) and assume data \( y(1) \ldots y(N) \) are available. The latent variable \( x \) cannot be observed nor its dimension \( n \) is known. In this Section, we address the problem of estimating \( S(z) \) and \( L(z) \) from the given observations. We shall do so by adapting the Gaussian regression approach described in Section II.

Since \( S(z) \) and \( L(z) \) are BIBO stable, we can use the approximations

\[
S(z) = \sum_{k=1}^{T} S_k z^{-k}, \quad L(z) = \sum_{k=1}^{T} L_k z^{-k}
\]

where \( T \) is sufficiently large. We construct the vectors \( y \) and \( e \) as in (8) and (9). The parameters of the transfer matrix \( S(z) \) are stacked in the vector \( \theta_s \in \mathbb{R}^{m^2 T} \) which is defined as follows

\[
\theta_s = \begin{bmatrix}
(s^{[11]})^\top & \ldots & (s^{[1 m]})^\top & \vdots & \ldots & (s^{[m 1]})^\top & \ldots & (s^{[m m]})^\top
\end{bmatrix}^\top
\]

where

\[
s^{[ij]} = \begin{bmatrix} [S_1]_{ij} & [S_2]_{ij} & \ldots & [S_T]_{ij} \end{bmatrix}^\top.
\]

In similar way, the parameters of the transfer matrix \( L(z) \) are stacked in

\[
\theta_l = \begin{bmatrix}
(l^{[11]})^\top & \ldots & (l^{[1 m]})^\top & \vdots & \ldots & (l^{[m 1]})^\top & \ldots & (l^{[m m]})^\top
\end{bmatrix}^\top
\]

where

\[
l^{[ij]} = \begin{bmatrix} [L_1]_{ij} & [L_2]_{ij} & \ldots & [L_T]_{ij} \end{bmatrix}^\top.
\]

Therefore, from (23) the vector of the observations can be expressed as

\[
y = \Phi(\theta_l + \theta_s) + e.
\]

We model \( \theta_s \) and \( \theta_l \) as two zero mean Gaussian independent random vectors with covariance matrix \( K_S \) and \( K_L \), respectively. Let \( \xi = \{\Gamma, \Lambda\} \) where \( \Gamma \) and \( \Lambda \) denotes the hyperparameters of \( K_S \) and \( K_L \), respectively, see Section
Let $\hat{\theta}_s$ and $\hat{\theta}_l$ denote the Bayes estimator of $\theta_s$ and $\theta_l$, respectively. Under the assumption
\[
p(y, \theta_1, \theta_s, y^- | \xi) \approx p(y | \theta_1, \theta_s, y^-(T + 1), \xi)p(\theta_1, \theta_s | \xi)p(y^- (T + 1))\]
we have
\[
(\hat{\theta}_s, \hat{\theta}_l) = \arg \min_{\theta_s, \theta_l} \|y - \Phi(\theta_s + \theta_l)\|^2_{\Sigma^{-1}}} + \|\theta_s\|^2_{K^S} + \|\theta_l\|^2_{K^L}.
\]
(33)

In the next Section, we show how to choose $K_S$ and $K_L$ in such a way to enforce $S(z)$ to be sparse and $L(z)$ low rank.

V. CHOICE OF THE KERNELS

In this Section we characterize the prior probability density for $\theta_s$ and $\theta_l$ by using the maximum entropy principle, see [5]. Such principle states that among all the prior probability densities satisfying the desired constraints, the optimal one should maximize the differential entropy.

Let $p(\theta_s, \theta_l)$ denote the joint probability density of $\theta_s$ and $\theta_l$. Their marginal probability densities are denoted by $p_s(\theta_s)$ and $p_l(\theta_l)$, respectively. Let $E_p$ denote the integration over $\mathbb{R}^{2mT}$ with respect to the probability measure $p$. Moreover, $P$ denotes the space of probability measures which are Lebesgue integrable. The differential entropy of $p$ is, [8],
\[
H(p) = -E_p[\log(p(\theta_s, \theta_l))].
\]
(34)

Next, we characterize the constraints on $p_s$ and $p_l$ inducing sparsity on $S(z)$ and low rank on $L(z)$, respectively. We start with the sparse part. The transfer function in position $(i,j)$ of $S(z)$ is the null transfer function if and only if $s^{(ij)}$ is the null vector. We consider the constraint on $p_s$
\[
E_p[\|s^{(ij)}\|^2_{K^S}] \leq p_{ij},
\]
(35)

where $p_{ij} \geq 0$. If $p_{ij} = 0$, then the $s^{(ij)}$ is zero in mean square and so is its posterior mean. Moreover, simple algebraic manipulations show that the weighted second moment bound in (35) implies a bound on the variance of $k$-th element of $s^{(ij)}$ which decays as the $k$-th element in the main diagonal of $\hat{K}$. Therefore, condition (35) tends to favor $s^{(ij)}$s with the features described at the end of Section II.

Regarding the low rank part, let $A_l \in \mathbb{R}^{m \times mT}$ be the random matrix such that
\[
A_l = \begin{bmatrix} L_1 & L_2 & \cdots & L_T \end{bmatrix}.
\]
(36)

Consider the constraint on $p_l$
\[
E_p[A_l(\hat{K}^{-1} \otimes I_m)A_l^\top] \leq Q.
\]
(37)

If $Q \in \mathbb{S}^+_m$ has $m - n$ singular values equal to zero, then the posterior mean of $A_lA_l^\top$ has rank equal to $n$. Therefore, $A_l$ admits the decomposition
\[
A_l = \begin{bmatrix} FH_1 & FH_2 & \cdots & FH_T \end{bmatrix},
\]
(38)

where $F \in \mathbb{R}^{mn \times n}$ and the $H_k \in \mathbb{R}^{n \times m}$ as in Section III. Equivalently, $L(z)$ admits the decomposition $L(z) = FH(z)$. Similarly to the sparse part, the weight matrix $\hat{K}^{-1} \otimes I_m$ tends to favor random vectors $l^{(ij)}$ having the features described at the end of Section II.

Therefore, we obtain the following maximum entropy problem
\[
\max_{p \in P} H(p)
\]
\[
s.t. \quad E_p[\|s^{(ij)}\|^2_{\hat{K}^{-1}}] \leq p_{ij}, \quad i, j = 1 \ldots m
\]
\[
E_p[A_l(\hat{K}^{-1} \otimes I_m)A_l^\top] \leq Q
\]
(39)

where $p_{ij} \geq 0$, $i, j = 1 \ldots m$, and $Q \in \mathbb{S}^+_m$.

**Theorem 5.1:** Given
\[
P = \text{diag}(p_{11} \ldots p_{1m} \ldots p_{m1} \ldots p_{mm}) \in \mathbb{S}^+_m
\]
(40)

and $Q \in \mathbb{S}^+_m$, the optimal solution to the maximum entropy problem is $p = p_s, p_l$, i.e. $\theta_s$ and $\theta_l$ are independent, where $p_s$ and $p_l$ are Gaussian probability densities with zero mean and covariance matrix
\[
K_S = \Gamma^{-1} \otimes \hat{K}
\]
(41)
\[
K_L = \Lambda^{-1} \otimes I_m \otimes \hat{K}
\]
(42)

where $\Gamma = TP^{-1}$ and $\Lambda = mTQ^{-1}$.

**Proof:** We characterize the optimal solution by exploiting the duality theory. We consider the Lagrange function
\[
L(p) = H(p)
\]
\[
+ \frac{1}{2} \sum_{i,j=1}^m \gamma(i,j)m + \frac{1}{2} \text{tr}(\Lambda(Q - E_p[A_l(\hat{K}^{-1} \otimes I_m)A_l^\top]])
\]
(43)

where $\gamma_i \geq 0$, $i = 1 \ldots m$, and $\Lambda \in \mathbb{S}_m^+$ are the Lagrange multipliers. It is not difficult to see that $L$ is strictly concave over $P$. Moreover, its maximum point is given by annihilating its first derivative. Therefore, we obtain
\[
p = \frac{1}{c} \exp \left( -\frac{1}{2} \sum_{i,j=1}^m \gamma(i,j)m + \frac{1}{2} \text{tr}(\Lambda(A_l(\hat{K}^{-1} \otimes I_m)A_l^\top]) \right)
\]
(44)

where $c$ is the normalization constant. Let $e_i$, $i = 1 \ldots m$, denote the $j$-th vector of the canonical basis of $\mathbb{R}^m$. Then, we have
\[
\sum_{i,j=1}^m \|s^{(ij)}\|^2_{\hat{K}^{-1}}
\]
\[
= \sum_{i,j=1}^m \theta_s^\top e_{(i-1)m+j}e_{(i-1)m+j}^\top \theta_s
\]
\[
= \theta_s^\top (\Gamma \otimes \hat{K}^{-1})\theta_s
\]
(45)

where
\[
\Gamma = \sum_{i,j=1}^m \gamma(i-1)m + j e_{(i-1)m+j} e_{(i-1)m+j}^\top = \text{diag}(\gamma_1 \ldots \gamma_{m^2}).
\]
(46)
Moreover, it is not difficult to see that
\[
\text{tr}(\Lambda A_l(\tilde{K}^{-1} \otimes I_m)A_l^\top) = \theta_l^\top (\Lambda \otimes I_m \otimes \tilde{K}^{-1})\theta_l.
\] (47)

Therefore, the optimal solution is such that \( p = p_s p_l \), where
\[
p_s = \frac{1}{c_s} \exp \left( -\frac{1}{2} \theta_s^\top (\Gamma \otimes \tilde{K}^{-1})\theta_s \right)
\]
\[
p_l = \frac{1}{c_l} \exp \left( -\frac{1}{2} \theta_l^\top (\Lambda \otimes I_m \otimes \tilde{K}^{-1})\theta_l \right)
\] (48)
with \( c_s \) and \( c_l \) normalization constants. Therefore, \( p_s \) and \( p_l \) are Gaussian probability densities with zero mean and covariance matrix
\[
K_S = \Gamma^{-1} \otimes \tilde{K}
\]
\[
K_L = \Lambda^{-1} \otimes I_m \otimes \tilde{K}.
\] (49)

Moreover,
\[
H(p) = \frac{1}{2} \log \det(K_S) + \frac{1}{2} \log \det(K_L) + m^2 T (1 + \log(2\pi)).
\] (50)
Therefore, the dual problem is equivalent to minimize the function
\[
J(\Gamma, \Lambda) = -T \log \det(\Gamma) + \text{tr}(\Gamma P) - m T \log \det(\Lambda) + \text{tr}(\Lambda Q)
\]
over the set \( S_{m_2}^+ \times S_{m_1}^+ \). Note that, \( J \) is strictly convex and its minimum point is given by setting to zero its first variations
\[
\delta J(\Gamma, \Lambda; U) = \text{tr}((-TT^{-1} + P)U)
\]
\[
\delta J(\Gamma, \Lambda; V) = \text{tr}((-mT\Lambda^{-1} + Q)V)
\] (51)
for any direction \( U \in S_{m_2} \) and \( V \in S_m \). Therefore, we obtain \( \Gamma = TP^{-1} \) and \( \Lambda = mTQ^{-1} \).

We conclude that the prior probability density inducing sparsity on \( \theta_s \) and low rank on \( \theta_l \) is zero mean Gaussian, with \( \theta_s \) and \( \theta_l \) independent; the marginal probability densities of \( \theta_s \) and \( \theta_l \) have covariance matrices (41) and (42), respectively.

Proposition 5.1: Under assumption (32), the solution to (33) is
\[
\hat{\theta}_s = (\Gamma^{-1} \otimes \tilde{K}) \Phi^\top c
\]
\[
\hat{\theta}_l = (\Lambda^{-1} \otimes I_m \otimes \tilde{K}) \Phi^\top c
\] (52)
(53)
where
\[
c = (\Phi(\Lambda^{-1} \otimes I_m \otimes \tilde{K} \otimes I_N) \otimes \Sigma \otimes I_N)^{-1} y.
\] (54)

Proof: It is sufficient to observe that Problem (33) can be rewritten as (15), where the parameter is \( \tilde{\theta} = \begin{bmatrix} \theta_s^\top & \theta_l^\top \end{bmatrix}^\top \), the regression matrix is \( \Phi = \begin{bmatrix} \Phi_S & \Phi_L \end{bmatrix} \), and the kernel matrix is
\[
\tilde{K} = \begin{bmatrix} K_S & 0 \\ 0 & K_L \end{bmatrix}.
\] (55)
Hence, we obtain
\[
\hat{\theta}_s = K_S \Phi^\top c, \: \hat{\theta}_l = K_L \Phi^\top c
\] (56)
with
\[
c = (\Phi(K_L + K_S) \Phi^\top + \Sigma \otimes I_N)^{-1} y.
\] (57)
Finally, from (41) and (42) we obtain (52) and (53).

Clearly, we are interested in the limit cases wherein \( \Gamma \) and \( \Lambda \) have some eigenvalues approaching infinity, because in that cases \( S(z) \) and \( L(z) \) almost surely are sparse and low rank, respectively. Note that, if all the eigenvalues of \( \Gamma \) tends to infinity then we only have \( L(z) \), that is we obtain a quasi-static factor model. In the case that all the eigenvalues of \( \Lambda \) tends to infinity we only have \( S(z) \), that is we obtain the sparse model studied in [7].

Remark 5.1: The fact that the matrix kernel \( K_S \) induces sparsity have been also proved in [2] using a non-Bayesian point of view.

Remark 5.2: It is possible to derive the same structure for \( K_L \) by adopting the regularization point of view. In [23], it has been shown that the penalty term \( \log \det(P) \), with \( P \in S_{m_2} \), induces low rank on \( P \). Moreover, it admits the upper bound
\[
\log \det(P) \leq \text{tr}(\Lambda P) - \log \det(\Lambda) - m
\] (58)
where \( \Lambda \in S_{m_2}^+ \) and equality holds if and only if \( \Lambda = P^{-1} \). Thus, to induce low rank on \( P = A_l(\tilde{K}^{-1} \otimes I_m)A_l^\top \) it is sufficient to consider the penalty
\[
\text{tr}(\Lambda A_l(\tilde{K}^{-1} \otimes I_m)A_l^\top) - \log \det(\Lambda) - m
\] (59)
where \( \Lambda \) represents a rough estimate of \( (A_l(\tilde{K}^{-1} \otimes I_m)A_l^\top)^{-1} \). The unique term depending on \( \theta_l \) in (59) is \( \text{tr}(\Lambda A_l(\tilde{K}^{-1} \otimes I_m)A_l^\top) = ||\theta_l||_{K^{-1}_L}^2 \). Thus, the penalty term \( ||\theta_l||_{K^{-1}_L}^2 \) induces low rank on \( A_l \).

VI. ESTIMATION OF THE HYPERPARAMETERS

In order to compute \( \hat{\theta}_s \) and \( \hat{\theta}_l \), we need to estimate \( \hat{K} \) and \( \xi = \{ \Gamma, \Lambda \} \). The kernel \( \hat{K} \) is estimated in a preliminary step using the model (3). \( \Gamma \) and \( \Lambda \) are obtained minimizing the negative log-marginal likelihood of \( y \)
\[
\arg\min_{\Gamma, \Lambda} \ell(y, \Gamma, \Lambda)
\] (60)
where, under (32), \( \ell(y, \Gamma, \Lambda) \) is given by (20) with
\[
V = \Phi(\Gamma^{-1} \otimes \tilde{K} + \Lambda^{-1} \otimes I_m \otimes \tilde{K}) \Phi^\top + \Sigma \otimes I_N.
\] (61)

Notice that, (60) is a nonconvex constrained optimization problem. Accordingly, only local minima can be computed. Because of the large number of variables to optimize, the correspondence among the local minima and the optimal solution could be compromised. For this reason, we constrain the structure of \( \Lambda \) (which corresponds to introducing an hyper-regularized on \( \Lambda \)) as follows:
\[
\Lambda = \alpha(I - UU^\top) + U \text{diag}(\beta_1 \ldots \beta_r) U^\top
\] (62)
where \( U \in \mathbb{R}^{m \times r} \) and its columns are the first \( r \) singular vectors of an estimate of \( (A_l A_l^\top)^{-1} \). The rank \( r \) of \( U \) is gradually increased until the negative log likelihood \( \ell \) stops decreasing.

In this way, the constraints in \( \Lambda \) are decoupled along the “most reliable” singular vectors \( U \) and their orthogonal complement.
Finally, to efficiently solve the optimization problem of step 7 in Algorithm 1, we used the scaled gradient projection algorithm developed in [4].

VII. SIMULATION RESULTS

We consider three Monte Carlo studies of 50 runs where at any run a model with \( m = 6 \) manifest variables is randomly generated. For each run in the Monte Carlo experiments an identification data set and a test set, both of size 500, are generated. The noise covariance matrix \( \Sigma \) is always estimated via a preliminary step using a low-bias ARX-model, see [10].

In the first experiment, the models have McMillan degree equal to 20, and are perturbed versions of (23) with \( l = 1 \) latent variable and four non null transfer functions in \( S(z) \).

The second experiment is identical to the first one, with the exception that the latent variables are \( l = 2 \).

In the third experiment, the models have McMillan degree equal to 30, but without a special structure.

The following algorithms were used:

- **Algorithm 1**: Computation of \( \Gamma \) and \( \Lambda \)

```plaintext
input : y(1) \ldots y(N) and \( \hat{K} \)
output: \( \Gamma, \Lambda \)
1 Let \( \hat{G}_1, \ldots, \hat{G}_T \) be the coefficients estimated from (15) with \( \hat{K} = I_{m^2} \otimes \hat{K} \)
2 \( \Lambda(0) \leftarrow [\hat{G}_1, \ldots, \hat{G}_T] \)
3 \( r^{(-1)} \leftarrow 1, r^{(0)} \leftarrow 1, r^{(1)} \leftarrow 1 \)
4 \( k \leftarrow 1 \)
5 while \( r^{(k)} - r^{(k-2)} < 2 \) do
6 \( U^{(k)} \leftarrow \text{first } r \text{ singular vectors of } (A^{(k-1)}(A^{(k-1)})\Gamma)^{-1} \)
7 \( (\Gamma^{(k)}, \Lambda^{(k)}) \leftarrow \arg \min_{\Gamma,\Lambda} \ell(y, \Gamma, \Lambda) \) with \( \Gamma = \alpha(I - U^{(k)}U^{(k)T}) + (U^{(k)}\text{diag}(\beta_1 \ldots \beta_r)U^{(k)T}) \)
8 Let \( \hat{L}_1^{(k)}, \ldots, \hat{L}_T^{(k)} \) be the coefficients of \( L(z) \) estimated from (33) with \( K_S = (\Gamma^{(k)})^{-1} \otimes \hat{K} \) and \( K_L = (\Lambda^{(k)})^{-1} \otimes I_{m^2} \otimes \hat{K} \)
9 \( \Lambda^{(k)} \leftarrow [\hat{L}_1^{(k)}, \ldots, \hat{L}_T^{(k)}] \)
10 if \( \ell(y, \Lambda^{(k)}) \geq \ell(y, \Lambda^{(k-1)}) \) then
11 \( r^{(k+1)} \leftarrow r^{(k)} + 1 \)
12 else
13 \( r^{(k+1)} \leftarrow r^{(k)} \)
14 \( k \leftarrow k + 1 \)
```

Table I shows the percentage of the average relative complexity of the S+L Bayesian network. In particular, in the first two experiments our method is able to detect that the underlying model is close to have a simple S+L network. Figure 2 shows the COD in the first two experiments. One can see that SL provides a slightly better performance than TC.

![Table I: Average relative complexity of the S+L Bayesian network](image)

| Expt. # | #1  | #2  | #3  |
|---------|-----|-----|-----|
| AC      | 55.89 | 63.72 | 81.56 |

- average relative complexity of the S+L Bayesian network of the SL model (in percentage)

\[ AC = \frac{100}{500} \sum_{t=1}^{500} \frac{\#SL_k}{m^2T} \] (63)

where \( \#SL_k \) is the number of parameters of the estimated S+L model at the \( k \)-th run, whereas \( m^2T \) is the number of parameters of a nonstructured model

- one step ahead coefficient of determination (in percentage)

\[ \text{COD} = 100 \left( 1 - \frac{1}{500} \sum_{t=1}^{500} \frac{\|y^{\text{test}}(t) - \hat{y}^{\text{test}}(t)\|^2}{\|y^{\text{test}}(t) - \bar{y}^{\text{test}}\|^2} \right) \] (64)

where \( \bar{y}^{\text{test}} \) denotes the sample mean of the test set data \( y^{\text{test}}(1), \ldots, y^{\text{test}}(500) \) and \( \hat{y}^{\text{test}}(t) \) is the one step ahead prediction computed using the estimated model

- average impulse response fit (in percentage)

\[ \text{AIRF} = 100 \left( 1 - \frac{\sum_{k=1}^{50} \|G_k - \hat{G}_k\|^2}{\sum_{k=1}^{50} \|G_k - \hat{G}_k\|^2} \right) \] (65)

with \( \hat{G} = \frac{1}{50} \sum_{k=1}^{50} G_k \).

Fig. 2. One step ahead coefficient of determination in the first experiment (left panel) and in the second experiment (right panel).
On the other hand, SL provides better estimators for the predictor coefficients than the TC, Figure 3. Finally, Figure 4 shows the COD in the third experiment. The median of SL is slightly worse than the one of TC. On the other hand, the bottom whisker of SL is better than the one of TC. Indeed, SL simplified the S+L network, see Table I, increasing the robustness of the estimated predictor impulse response coefficients.

VIII. CONCLUSIONS

In this paper, we proposed a Gaussian regression approach to identify multivariate time series having sparse Bayesian network with few latent nodes. Simulations show that our approach is able to identify such S+L network without compromising the prediction performance.

REFERENCES

[1] A. Abdelwahab, O. Amor, and T. Abdelwahed. The analysis of the interdependence structure in international financial markets by graphical models. Int. Res. J. Finance Econ., 15:291–306, 2008.
[2] A. Aravkin, J. Burke, A. Chiuso, and G. Pillonetto. Convex vs non-convex estimators for regression and sparse estimation: the mean squared error properties of ard and glasso. Journal of Machine Learning Research, 15:217–252, 2014.
[3] N. Aronszajn. Theory of reproducing kernels. Trans. Amer. Math. Soc., 68:337–404, 1950.
[4] S. Bonettini, A. Chiuso, and M. Prato. A scaled gradient projection method for Bayesian learning in dynamical systems. SIAM Journal on Scientific Computing (accepted), 2015.
[5] F. Carli, T. Chen, and L. Ljung. Maximum entropy kernels for system identification. arXiv preprint arXiv:1411.5620, 2014.
[6] T. Chen, H. Ohlsson, and L. Ljung. On the estimation of transfer functions, regularizations and gaussian processes-revisited. Automatica, 48(8):1525–1535, 2012.
[7] A. Chiuso and G. Pillonetto. A Bayesian approach to sparse dynamic network identification. Automatica, 48(8):1553–1565, 2012.
[8] T. Cover and J. Thomas. Information Theory. Wiley, New York, 1991.
[9] M. Deistler and C. Zinner. Modelling high-dimensional time series by generalized linear dynamic factor models: An introductory survey. Communications in Information & Systems, 7(2):153–166, 2007.
[10] G. Goodwin, M. Gevers, and B. Ninness. Quantifying the error in estimated transfer functions with application to model order selection. IEEE Transactions on Automatic Control, 37(7):913–928, Jul 1992.
[11] C. Granger. Investigating causal relations by econometric models and cross-spectral methods. Econometrica, 37(3):424–438, Aug. 1969.
[12] E Kolaczyk. Statistical analysis of network data: methods and models. Springer, 2009.
[13] S.L. Lauritzen. Graphical Models. Oxford University Press, Oxford, 1996.
[14] L. Ljung, editor. System Identification (2ND Ed.): Theory for the User. Prentice Hall, New Jersey, 1999.
[15] G. Pillonetto, A. Chiuso, and G. De Nicolao. Prediction error identification of linear systems: A nonparametric gaussian regression approach. Automatica, 47(2):291–305, 2011.
[16] G. Pillonetto and G. De Nicolao. A new kernel-based approach for linear system identification. Automatica, 46:81–93, 2010.
[17] G. Pillonetto, F. Dinuzzo, T. Chen, G. De Nicolao, and L. Ljung. Kernel methods in system identification, machine learning and function estimation: A survey. Automatica, 50(3):657–682, 2014.
[18] T. Poggio and F. Girosi. Networks for approximation and learning. Proceedings of the IEEE, 78(9):1481–1497, 1990.
[19] G. Prando, A. Chiuso, and G. Pillonetto. Bayesian and regularization approaches to multivariable linear system identification: the role of rank penalties. In Proceedings of the IEEE Conference on Decision and Control, 2014.
[20] C. Rasmussen and C. Williams. Gaussian Processes for Machine Learning. The MIT Press, 2006.
[21] T. Söderström and P. Stoica, editors. System Identification. Prentice Hall, New Jersey, 1988.
[22] A. Werhli, M. Grzegorczyk, and D. Husmeier. Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical gaussian models and Bayesian networks. Bioinformatics, 22(20):2523–2531, 2006.
[23] D. Wipf. Non-convex rank minimization via an empirical Bayesian approach. In Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence, pages 914–923, 2012.