Paper submitted to 59th IEEE Conf. Decision and Control, Jeju Island, Republic of Korea, 8-11 December 2020.

A regularized kernel-based method for learning a module in a dynamic network with correlated noise

Venkatakrishnan C. Rajagopal, Karthik R. Ramaswamy and Paul M.J. Van den Hof

***Abstract*— In this paper, we consider the problem of identifying one system (module) embedded in a dynamic network that is disturbed by colored process noise sources, which can possibly be correlated. To achieve this using the direct method for single module identification, we need to formulate a Multi-Input-Multi-Output (MIMO) estimation problem which requires model order selection step for each module in the setup and estimation of large number of parameters. This results in larger variance in the estimates and increase in computation complexity. In this paper, we extend the Empirical Bayes Direct Method [1] that handles the above mentioned problems for a Multi-Input-Single-Output (MISO) setup to a MIMO setting. We keep a parametric model for the desired target module and model the impulse response of all the other modules as independent zero mean Gaussian process with the covariance matrix represented by the first-order stable spline kernel, accounting also for the noise model affecting the outputs of the MIMO setup. The parameters of the target module are obtained by maximizing the marginal likelihood of the output using the Empirical Bayes (EB) approach. To solve this, we use the Expectation Maximization (EM) algorithm which offers computational advantages. Numerical simulation illustrate the advantages of the developed method over existing classical methods.**

I. Introduction

Dynamic networks are multiple systems interconnected with each other and can be defined as a set of measurable signals (also called as node signals) interconnected through linear time-invariant (LTI) dynamic systems (modules), possibly driven by external excitation signals. Over the past decade, data-driven modeling in dynamic networks has garnered increased attention by many researchers in the field of system identification. Two major research problems in data-driven modeling of dynamic networks are the full network identification and single module identification. The former problem focuses on identifying the whole network dynamics [2]–[5], including aspects of identifiability [6]–[10], while the latter problem focuses on identifying a single module embedded in a dynamic network considering that the topology of the network is known [1], [11]–[19].

In this paper we focus on the problem of local module identification. In [11], the *direct method* for single module identification in dynamic networks has been introduced by extending the direct method for closed loop identification [20]. In this method, a Multi-Input-Single-Output (MISO) identification problem is formulated with inputs being all node signals directly connected to the output of the target module. However, the target module can be consistently estimated with limited number of inputs in the MISO problem and an algorithm for the limited predictor input selection has been presented in [14]. Limiting the inputs can lead to the situation of confounding variables[1] (see [18], [21] for details), which lead to biased estimates if not dealt with. This has been addressed in [21] by adding additional predictor inputs in the MISO setup. However, the above direct method approaches provide consistent (and Maximum Likelihood (ML)) estimates only under the situation of process noise acting on the nodes being uncorrelated.

The situation of correlation in process noise can be handled using the *indirect method* [17] and its variants like the *two stage method* [11], [14] and *instrumental variable* methods [12], [22]. However, the indirect method and its variants require a strong presence of measured external excitation signals to serve as predictor inputs, and might increase the cost of experiments.

On the contrary, the direct approaches use the entire information of the node signal (both excitation and noise signal), which makes it advantageous to use, but suffers from handling correlated noise. A solution to this problem has been provided in [18] where the *local direct method* has been introduced. In this method, we handle the effect of noise correlation in dynamic networks by moving from a MISO to Multi-Input-Multi-Output (MIMO) identification setup and obtain estimates with ML properties using the prediction error method (PEM). Therefore, the single module identification problem becomes embedded in a network MIMO identification problem, resulting in the problem of estimating high number of parameters that are of no prime interest to the experimenter. In addition, all these additional modules need to be suitably parameterized based on complexity criteria like AIC, BIC, or Cross Validation (CV) [20]. This step involves permuting candidate model orders for the modules which increases exponentially as the number of modules or orders increase. Also, algorithms to solve the network MIMO estimation problem for arbitrary model structures (except ARX, ARMAX [23]) are not available.

To eliminate the model order selection step and reduce the number of estimated parameters, we build on [1] and develop a regularized kernel based method (see [24] for a

The authors are with the Dept. of Electrical Engineering, Eindhoven University of Technology, The Netherlands v.comandoor.rajagopal@student.tue.nl, {k.r.ramaswamy, p.m.j.vandenhof}@tue.nl.

This work has received funding from the European Research Council (ERC), Advanced Research Grant SYSDYNET, under the European Union's Horizon 2020 research and innovation programme (grant agreement No 694504).

[1] unmeasured variables that directly or indirectly influence the input and output of an estimation problem.

survey) that extends the semi-parametric approach of [1] in a MISO setting to a MIMO setting.

We keep a parametric model for the target module in order to get an accurate description of the desired dynamics. To avoid model order selection and to reduce the number of estimated parameters, the additional modules are represented as independent zero mean Gaussian Process (GP) [25] whose covariance matrix is given by the *first-order stable spline kernel* [26] which enforces stability and smoothness of the impulse response coefficients. The parameter vector η containing the parameters of the target module, hyperparameters of the kernel and the covariance of the process noise are estimated by maximizing the marginal likelihood of the data, achieved by an Expectation-Maximization (EM) method having attractive computational properties.

This paper is organized as follows. In section II, the network setup and the problem is defined. In sections III and IV the method is explained, followed by numerical simulation and results in section V. Finally, the conclusions are presented

II. Problem statement

Following the setting in [11], we consider a dynamic network built up of L measurable internal variables or nodes $w_j(t), j = 1, \dots, L$. This network is defined as[2]

$$\underbrace{\begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_L \end{bmatrix}}_{w} = \underbrace{\begin{bmatrix} 0 & G_{12} & \cdots & G_{1L} \\ G_{21} & 0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & G_{L-1L} \\ G_{L1} & \cdots & G_{LL-1} & 0 \end{bmatrix}}_{G} \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_L \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_L \end{bmatrix} + \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_L \end{bmatrix}, \tag{1}$$

where q^{-1} is the delay operator i.e. $q^{-1}u(t) = u(t-1)$.

- G_{jl} is a strictly proper rational transfer function for $j = 1, \dots, L$ and $l = 1, \dots, L$, referred to as a *module*,
- There are no self loops in the network i.e. nodes are not directly related to itself, i.e. $G_{jj} = 0$,
- The topology of the network is known i.e. which entries of G are non-zero are known a priori.
- v_j is the process noise entering the node w_j. The vector process $v = \begin{bmatrix} v_1 & \dots & v_L \end{bmatrix}^\top$ is modelled as a stationary stochastic process with rational spectral density, $\Phi_v(\omega)$, such that $v = H(q)e$, where $e = \begin{bmatrix} e_1 & \dots & e_L \end{bmatrix}^\top$ is a Gaussian white noise process with covariance, $\Lambda > 0$, and $H(q)$ is square, stable, monic, minimum phase transfer matrix. The correlated noise situation, considered in this paper, refers to the situation of non-diagonal $\Phi_v(\omega)$ and $H(q)$.
- u_j is generated by the *external variables* r_k, that can be directly manipulated by the user and is given by $u_j = \sum_{k=1}^{K} R_{jk} r_k$, where R_{jk} are stable and proper rational transfer functions. Therefore, $u = \begin{bmatrix} u_1 & \dots & u_L \end{bmatrix}^\top$ can be represented as $u = Rr$ where, $r = \begin{bmatrix} r_1 & \dots & r_K \end{bmatrix}^\top$ and R is the matrix of rational transfer functions R_{jk}.

[2]time and frequency dependency is dropped for convenience.

Assumption 1: In a dynamic network represented by (1), we consider the following assumptions:

- The dynamic network is stable i.e. $(I - G)^{-1}$ is stable, and well posed (see [11] for details).
- The structure of process noise correlation is known i.e. we know a priori which entries of $\Phi_v(\omega)$ are nonzero.

According to the local direct method [18], a module G_{ji} embedded in a dynamic network with correlated noise can be consistently identified with a MIMO estimation setup $w_{\mathcal{D}} \rightarrow w_{\mathcal{Y}}$, where predictor inputs $w_{\mathcal{D}}$ and predicted outputs $w_{\mathcal{Y}}$ may have common signals, due to the handling of correlated disturbances (see [18] for more details on signal selection). The correlated disturbances that appear as confounding variables which affect both the input and output of an estimation problem, can be properly handled by including the related input signals as output too, and by exploiting a multivariate noise model to cover the correlated disturbances. The estimation setup results from the network equation

$$\underbrace{\begin{bmatrix} w_{\mathcal{Q}} \\ w_o \end{bmatrix}}_{w_{\mathcal{Y}}} = \underbrace{\begin{bmatrix} \bar{G}_{\mathcal{Q}\mathcal{Q}} & \bar{G}_{\mathcal{Q}\mathcal{U}} \\ \bar{G}_{o\mathcal{Q}} & \bar{G}_{a\mathcal{U}} \end{bmatrix}}_{\bar{G}} \underbrace{\begin{bmatrix} w_{\mathcal{Q}} \\ u_{\mathcal{U}} \end{bmatrix}}_{w_{\mathcal{D}}} + \underbrace{\begin{bmatrix} \bar{H}_{\mathcal{Q}\mathcal{Q}} & \bar{H}_{\mathcal{Q}\mathcal{U}} \\ \bar{H}_{o\mathcal{Q}} & \bar{H}_{a\mathcal{U}} \end{bmatrix}}_{\bar{H}} \underbrace{\begin{bmatrix} \xi_{\mathcal{Q}} \\ \xi_o \end{bmatrix}}_{\xi_{\mathcal{Y}}}, \tag{2}$$

where $w_{\mathcal{Q}}$ are the set of nodes that are common to both inputs and outputs that are needed to handle the noise correlations and confounding variable as discussed in [18], $u_{\mathcal{U}}$ and w_o are the sets of nodes that are exclusively inputs and outputs respectively. The vector $\xi_{\mathcal{Y}}$ is a Gaussian white noise process constructed by spectral decomposition and $\bar{H}$ is square, stable, monic and minimum phase. The desired target module is represented in $\bar{G}_{ji}$ i.e. $\bar{G}_{ji} = G_{ji}$ and $\bar{G}_{\mathcal{Q}\mathcal{Q}}$ is a hollow matrix and thus does not lead to transfers between signals that are the same. Also, the non-zero entries in $\bar{G}$ can be computed (refer to [18]). Without loss of generality, $r = 0$ is considered for simplicity.

We want to identify a parametric model for the module directly linking node w_i and w_j, represented as $G_{ji}(q, \theta)$ that describes the dynamics of the module of interest for a certain parameter vector $\theta \in \mathbb{R}^{n_\theta}$, from N measurements of the node signals $w_{\mathcal{D}}$ and $w_{\mathcal{Y}}$. In the local direct method, not only the target module G_{ji} but all the modules in $\bar{G}$ are parameterized, resulting in high number of parameters to estimate which causes a detrimental effect on the variance of the parameter estimates when N is not very large. Therefore, we focus on estimating a parametric model for the target module while reducing the number of parameters for the remaining modules in the MIMO identification setup.

III. Developing the Bayesian model

In this section, we discuss how we avoid parameterizing all but the target module using regularized kernel-based methods. As the starting point of the methodology in this paper, we use the MIMO structure in (2), as opposed to a MISO structure in the *Empirical Bayes Direct Method* (EBDM) [1]. Following (2), while maintaining the monicity

of the noise model, the equation can be re-ordered as

$$\underbrace{\begin{bmatrix} w_j \\ w_{\tilde{\mathcal{Y}}} \end{bmatrix}}_{\tilde{w}_{\mathcal{Y}}(t)} = \underbrace{\begin{bmatrix} G_{ji} & \bar{G}_{j\tilde{\mathcal{D}}} \\ \bar{G}_{\tilde{\mathcal{Y}}i} & \bar{G}_{\tilde{\mathcal{Y}}\tilde{\mathcal{D}}} \end{bmatrix}}_{\check{G}} \underbrace{\begin{bmatrix} w_i \\ w_{\tilde{\mathcal{D}}} \end{bmatrix}}_{\tilde{w}_{\mathcal{D}}(t)} + \underbrace{\begin{bmatrix} \bar{H}_{jj} & \bar{H}_{j\tilde{\mathcal{Y}}} \\ \bar{H}_{\tilde{\mathcal{Y}}j} & \bar{H}_{\tilde{\mathcal{Y}}\tilde{\mathcal{Y}}} \end{bmatrix}}_{\check{H}} \underbrace{\begin{bmatrix} \xi_j \\ \xi_{\tilde{\mathcal{Y}}} \end{bmatrix}}_{\tilde{\xi}_{\mathcal{Y}}(t)}, \quad (3)$$

where $\tilde{\mathcal{Y}} = \mathcal{Y} \setminus \{j\}$ and $\tilde{\mathcal{D}} = \mathcal{D} \setminus \{i\}$. The signals $\tilde{w}_{\mathcal{Y}}$, $\tilde{w}_{\mathcal{D}}$, and $\tilde{\xi}_{\mathcal{Y}}$ are suitably rearranged. To parameterize only G_{ji} in $\check{G}$, we first define the following quantities:

$$S(q) = I_{|\mathcal{Y}|} - \check{H}(q)^{-1}, \qquad \tilde{G}(q) = \begin{bmatrix} 0 & \bar{G}_{j\tilde{\mathcal{D}}} \\ \bar{G}_{\tilde{\mathcal{Y}}i} & \bar{G}_{\tilde{\mathcal{Y}}\tilde{\mathcal{D}}} \end{bmatrix},$$

$S_{\mathcal{D}}(q) = (I - S(q))\tilde{G}(q)$, where $|\mathcal{X}|$ denotes the cardinality of set $\mathcal{X}$. With these definitions, we build a predictor from (3) with a parameterized G_{ji} as

$$\tilde{w}_{\mathcal{Y}}(t) = (I - S(q)) \begin{bmatrix} G_{ji}(q,\theta) \\ \mathbf{0}_{(|\mathcal{Y}|-1)\times 1} \end{bmatrix} w_i(t) + S_{\mathcal{D}}(q)\tilde{w}_{\mathcal{D}}(t) + S(q)\tilde{w}_{\mathcal{Y}}(t) + \tilde{\xi}_{\mathcal{Y}}(t). \quad (4)$$

It is to be noted that the first element of $S_{\mathcal{D}}(q)$ is zero if $\bar{G}_{\tilde{\mathcal{Y}}i} = \mathbf{0}$, else $S_{\mathcal{D}}(q)$ becomes a full matrix due to the multiplication of $(I - S(q))$ and $\tilde{G}(q)$.

A. Vector description of network dynamics

Keeping a parametric model for the target module, we now need to model the other modules. First, we obtain a vector description of the network dynamics for the available N measurements using impulse response of the modules. We stack the first ℓ coefficients of the impulse response of each module in $S_{\mathcal{D}}(q)$ and $S(q)$ as

$$s_{\mathcal{D}} = \begin{bmatrix} s_{Y_1D_1}^\top, \ldots, s_{Y_{|\mathcal{Y}|}D_{|\mathcal{D}|}}^\top \end{bmatrix}^\top, \; s_{\mathcal{Y}} = \begin{bmatrix} s_{Y_1Y_1}^\top, \ldots, s_{Y_{|\mathcal{Y}|}Y_{|\mathcal{Y}|}}^\top \end{bmatrix}^\top,$$

where $Y_1, \ldots, Y_{|\mathcal{Y}|}$ and $D_1, \ldots, D_{|\mathcal{D}|}$ are elements of set $\mathcal{Y}$ and $\mathcal{D}$ respectively. ℓ is chosen sufficiently large to capture the impulse response dynamics. We also represent the target module $G_{ji}(q,\theta)$ as an impulse response, where the first N coefficients are collected in g_{ji} (the dependence on θ is implicit and dropped).

Next we introduce a vector notation for the signal $\tilde{w}_{\mathcal{Y}}(t)$:

$$\tilde{w}_{\mathcal{Y}} := \begin{bmatrix} \tilde{w}_{Y_1}(1) & \ldots & \tilde{w}_{Y_1}(N) & \tilde{w}_{Y_2}(1) & \ldots & \tilde{w}_{Y_{|\mathcal{Y}|}}(N) \end{bmatrix}$$

and we denote $G_\theta \in \mathbb{R}^{N\times N}$ as the Toeplitz matrix of g_{ji}, $\bar{W}_i \in \mathbb{R}^{N\times \ell}$ as the Toeplitz matrix of $\begin{bmatrix} 0 & 0 & w_i(1) & \ldots & w_i(N-2) \end{bmatrix}^\top$, and $W_i \in \mathbb{R}^{N\times N}$ as the Toeplitz matrix of $\begin{bmatrix} 0 & w_i(1) & \ldots & w_i(N-1) \end{bmatrix}^\top$, and $\breve{W}_k \in \mathbb{R}^{N\times \ell}$ as the Toeplitz of $\begin{bmatrix} 0 & w_k(1) & \ldots & w_k(N-1) \end{bmatrix}^\top$ where k belongs to the elements in $\mathcal{Y}$ and $\mathcal{D}$. We also define the following:

$$\begin{aligned} W_{\mathcal{Y}} &= \begin{bmatrix} W_{Y_1} & \ldots & W_{Y_{|\mathcal{Y}|}} \end{bmatrix} \quad W_{\mathcal{D}} = \begin{bmatrix} W_{D_1} & \ldots & W_{D_{|\mathcal{D}|}} \end{bmatrix} \\ \tilde{W}_i &= [G_\theta \bar{W}_i \;\; \mathbf{0}] \in \mathbb{R}^{N\times \ell|\mathcal{Y}|}, \\ \tilde{\boldsymbol{W}}_i &= \mathrm{diag}(\tilde{W}_i, \ldots, \tilde{W}_i) \in \mathbb{R}^{N|\mathcal{Y}|\times \ell|\mathcal{Y}|^2}, \\ \boldsymbol{W}_{\mathcal{D}} &= \mathrm{diag}(W_{\mathcal{D}}, \ldots, W_{\mathcal{D}}) \in \mathbb{R}^{N|\mathcal{Y}|\times \ell|\mathcal{D}|^2}, \\ \boldsymbol{W}_{\mathcal{Y}} &= \mathrm{diag}(W_{\mathcal{Y}}, \ldots, W_{\mathcal{Y}}) \in \mathbb{R}^{N|\mathcal{Y}|\times \ell|\mathcal{Y}|^2}. \end{aligned} \quad (5)$$

Having defined the above terms, (4) can be rewritten in vector form as

$$\tilde{w}_{\mathcal{Y}} = \begin{bmatrix} W_i \\ \mathbf{0} \end{bmatrix} g_{ji} - \tilde{\boldsymbol{W}}_i s_{\mathcal{Y}} + \boldsymbol{W}_{\mathcal{D}} s_{\mathcal{D}} + \boldsymbol{W}_{\mathcal{Y}} s_{\mathcal{Y}} + \xi, \quad (6)$$

where $\xi \in \mathbb{R}^{N|\mathcal{Y}|\times 1}$ is the vectorized noise.

B. Modeling the additional modules as GP

We now discuss our modeling strategy for the additional modules. Our aim is increase the accuracy of the desired parameter θ by limiting the number of parameters to be estimated to describe $\tilde{w}_{\mathcal{Y}}$ in (6). Therefore, we keep a parametric model for g_{ji} and model the remaining impulse responses in (6) as independent zero mean Gaussian Processes [25]. Gaussian processes are effective in reducing the variance of the impulse response estimate with suitable choice of a prior covariance matrix (kernel) [24], which we chose to be the *First order Stable Spline kernel* [26]. The kernel structure is given by $K := \lambda K_\beta$ with

$$[K_\beta]_{x,y} = \beta^{\max(x,y)},$$

where $\beta \in [0, 1)$ and $\lambda \geq 0$. λ and β are hyperparameters that govern the amplitude and exponential decay of the realization of the Gaussian vector respectively. The chosen kernel enforces smoothness and stability of the estimate of the impulse responses. Therefore, we have:

$$\begin{aligned} s_{Y_pD_k} &\sim \mathcal{N}(\mathbf{0}, \lambda_{pk}^D K_{\beta_{pk}^D}), p = 1, \ldots, |\mathcal{Y}|, k = 1, \ldots, |\mathcal{D}| \\ s_{Y_pY_k} &\sim \mathcal{N}(\mathbf{0}, \lambda_{pk}^Y K_{\beta_{pk}^Y}), p = 1, \ldots, |\mathcal{Y}|, k = 1, \ldots, |\mathcal{Y}|. \end{aligned} \quad (7)$$

Each impulse response prior is assigned with independent hyperparameters λ and β for flexibility of modeling. Let us now define, $\boldsymbol{s} = \begin{bmatrix} s_{\mathcal{Y}}^\top & s_{\mathcal{D}}^\top \end{bmatrix}^\top$, $\mathbf{W} = \begin{bmatrix} \boldsymbol{W}_{\mathcal{Y}} - \tilde{\boldsymbol{W}}_i & \boldsymbol{W}_{\mathcal{D}} \end{bmatrix}$ and let $\mathbf{K}$ be the block diagonal matrix constructed with the covariance of the impulse response priors. Using the above definitions, (6) can be written as,

$$\tilde{w}_{\mathcal{Y}} = \begin{bmatrix} W_i \\ \mathbf{0} \end{bmatrix} g_{ji} + \mathbf{W}\boldsymbol{s} + \xi. \quad (8)$$

In (8), $\boldsymbol{s}$ is modeled as Gaussian process. Therefore by considering a Gaussian distribution for noise ξ and also taking into account the noise correlations

$$\xi \sim \mathcal{N}(\mathbf{0}, \bar{\Sigma} \otimes I_N), \; \bar{\Sigma} := \begin{bmatrix} \sigma_{11}^2 & \sigma_{12}^2 & \ldots & \sigma_{1|\mathcal{Y}|}^2 \\ * & \sigma_{22}^2 & \ldots & \sigma_{2|\mathcal{Y}|}^2 \\ \vdots & \vdots & \ddots & \vdots \\ * & * & \ldots & \sigma_{|\mathcal{Y}||\mathcal{Y}|}^2 \end{bmatrix}$$

we can write a joint probabilistic description of $\boldsymbol{s}$ and $\tilde{w}_{\mathcal{Y}}$, which is jointly Gaussian, as:

$$p\left(\begin{bmatrix} \mathbf{s} \\ \tilde{w}_{\mathcal{Y}} \end{bmatrix}; \eta \right) \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{0} \\ \begin{bmatrix} W_i \\ \mathbf{0} \end{bmatrix} g_{ji} \end{bmatrix}, \begin{bmatrix} \mathbf{K} & \mathbf{K}\mathbf{W}^\top \\ \mathbf{W}\mathbf{K} & \mathbf{P} \end{bmatrix} \right) \quad (9)$$

where,

$$\mathbf{P} := \Sigma + \mathbf{W}\mathbf{K}\mathbf{W}^\top, \; \Sigma := \bar{\Sigma} \otimes I_N,$$

and,

$$\eta = [\theta\ \lambda_{11}^{D}\ \dots\ \lambda_{|\mathcal{Y}||\mathcal{D}|}^{D}\ \lambda_{11}^{Y}\ \dots\ \lambda_{|\mathcal{Y}||\mathcal{Y}|}^{Y}\ \beta_{11}^{D}\ \dots\ \beta_{|\mathcal{Y}||\mathcal{D}|}^{D}\ \beta_{11}^{Y}\ \dots\ \beta_{|\mathcal{Y}||\mathcal{Y}|}^{Y}\ \sigma_{11}^{2}\ \dots\ \sigma_{1|\mathcal{Y}|}^{2}\ \sigma_{22}^{2}\ \dots\ \sigma_{2|\mathcal{Y}|}^{2}\ \dots\ \sigma_{|\mathcal{Y}||\mathcal{Y}|}^{2}]^{\top}. \tag{10}$$

The parameter vector η governs the probability distribution function in (9). It consists of the parameters of $G_{ji}(\theta)$, the hyperparameters of the kernels of the impulse response models and the elements of the covariance of the noise acting on $\tilde{w}_{\mathcal{Y}}$. It is important to note that in EBDM we estimate the variance of the noise corrupting only the output of the target module $w_j(t)$, whereas here we need to estimate the elements of the full covariance matrix of the noise corrupting the signal vector $w_{\mathcal{Y}}(t)$, to take into account the effects of noise correlations. Therefore, if we estimate η, we get θ. To estimate the θ contained in η, we adopt an Empirical Bayes (EB) framework [27]. To this end, we consider the marginal pdf of $\tilde{w}_{\mathcal{Y}}$ by integrating out the effect of $\boldsymbol{s}$ and maximizing the marginal likelihood of $w_{\mathcal{Y}}$. The corresponding objective function is

$$\begin{aligned}\hat{\eta} &= \underset{\eta}{\operatorname{argmax}}\, p(\tilde{w}_{\mathcal{Y}};\eta)\\ &= \underset{\eta}{\operatorname{argmin}} \log|\mathbf{P}| + \left(\tilde{w}_{\mathcal{Y}} - \begin{bmatrix} W_i \\ \mathbf{0} \end{bmatrix} g_{ji}\right)^{\top} \mathbf{P}^{-1} \left(\tilde{w}_{\mathcal{Y}} - \begin{bmatrix} W_i \\ \mathbf{0} \end{bmatrix} g_{ji}\right).\end{aligned} \tag{11}$$

This optimization problem is complex and non-convex, and solving such a problem is cumbersome. Therefore, in the next section, we introduce a method to solve the marginal likelihood maximization problem through an iterative scheme.

IV. Maximizing Marginal Likelihood

For maximizing the marginal likelihood, we consider the iterative method of *Expectation Maximization* (EM) method [28] for obtaining the estimate of η. For this, we need to first define the latent variable whose estimaion simplifies the calculation of the marginal likelihood. In this case, we choose $\boldsymbol{s}$. The EM method guarantees convergence to a local minima [29] and the optimization problem is simplified as seen in *Lemma 1* compared to solving the original problem in (11). The EM method has two steps,

- *E-step*: Given an estimate of $\hat{\eta}^{(n)}$ at the n^{th} iteration, compute
$$Q^{(n)}(\eta) = \mathbb{E}_{p(\boldsymbol{s}|\tilde{w}_{\mathcal{Y}};\hat{\eta}^{(n)})}[\log p(\tilde{w}_{\mathcal{Y}}, \boldsymbol{s};\eta)], \tag{12}$$
- *M-step*: Compute $\hat{\eta}^{(n+1)}$ from
$$\hat{\eta}^{(n+1)} = \underset{\eta}{\operatorname{argmax}}\, Q^{(n)}(\eta). \tag{13}$$

The estimate $\hat{\eta}$ is obtained by iterating between (12) and (13) until the parameters converge. Although the procedure is iterative, the EM algorithm significantly simplifies solving (11), which are shown in our next steps.

The posterior distribution of $\boldsymbol{s}$ given $\tilde{w}_{\mathcal{Y}}$ and an estimate of η is Gaussian, and is given by $p(\boldsymbol{s}|\tilde{w}_{\mathcal{Y}};\eta) \sim \mathcal{N}(\boldsymbol{s}_m, P_{\boldsymbol{s}})$ [30] where

$$\begin{aligned} P_{\boldsymbol{s}} &= \mathbf{K} - \mathbf{K}\mathbf{W}^{\top}(\mathbf{W}\mathbf{K}\mathbf{W}^{\top} + \Sigma)^{-1}\mathbf{W}\mathbf{K},\\ \boldsymbol{s}_m &= (\mathbf{K}\mathbf{W}^{\top}(\mathbf{W}\mathbf{K}\mathbf{W}^{\top} + \Sigma)^{-1})(\tilde{w}_{\mathcal{Y}} - \begin{bmatrix} W_i \\ \mathbf{0} \end{bmatrix} g_{ji}). \end{aligned} \tag{14}$$

Let $\hat{\boldsymbol{s}}^{(n)}$ and $\hat{P}_{\boldsymbol{s}}^{(n)}$ be the posterior mean and covariance of $\boldsymbol{s}$ obtained from (14) using $\hat{\eta}^{(n)}$, we define $\hat{\boldsymbol{S}}^{(n)} := \hat{P}_{\boldsymbol{s}}^{(n)} + \hat{\boldsymbol{s}}^{(n)}\hat{\boldsymbol{s}}^{(n)\top}$ and each of its $\ell \times \ell$ diagonal block as $\hat{\boldsymbol{S}}_m^{(n)}$ which are the posterior second moment of $\hat{\boldsymbol{s}}_m^{(n)}$. Here, m corresponds to each combination of the impulse response in (7) and its respective hyperparameters.

The structure of $Q^{(n)}(\eta)$ in (12) for the setup in (11) is provided in the following lemma.

Lemma 1: Let $\hat{\eta}^{(n)}$ be the estimate of η at n^{th} iteration of the EM algorithm according to (13), then

$$Q^{(n)}(\eta) = Q_0^{(n)}(\theta, \Sigma) + \sum_m Q_{s_m}^{(n)}(\lambda_m, \beta_m) \tag{15}$$

where,

$$\begin{aligned} Q_0^{(n)}(\theta,\Sigma) = &-\log\det\Sigma - \operatorname{tr}\Big(\Sigma^{-1}\Big(\tilde{w}_{\mathcal{Y}}\tilde{w}_{\mathcal{Y}}^{\top}\\ &+ \begin{bmatrix} W_i \\ \mathbf{0} \end{bmatrix} g_{ji}g_{ji}^{\top}\begin{bmatrix} W_i^{\top} & \mathbf{0}^{\top} \end{bmatrix} + \mathbf{W}\hat{\boldsymbol{S}}^{(n)}\mathbf{W}^{\top} - \begin{bmatrix} W_i \\ \mathbf{0} \end{bmatrix} g_{ji}\tilde{w}_{\mathcal{Y}}^{\top}\\ &- \mathbf{W}\hat{\boldsymbol{s}}^{(n)}\tilde{w}_{\mathcal{Y}}^{\top} - \tilde{w}_{\mathcal{Y}}g_{ji}^{\top}\begin{bmatrix} W_i^{\top} & \mathbf{0}^{\top} \end{bmatrix} + \mathbf{W}\hat{\boldsymbol{s}}^{(n)}g_{ji}^{\top}\begin{bmatrix} W_i^{\top} & \mathbf{0}^{\top} \end{bmatrix}\\ &- \tilde{w}_{\mathcal{Y}}\hat{\boldsymbol{s}}^{(n)\top}\mathbf{W}^{\top} + \begin{bmatrix} W_i \\ \mathbf{0} \end{bmatrix} g_{ji}\hat{\boldsymbol{s}}^{(n)\top}\mathbf{W}^{\top}\Big)\Big), \end{aligned}$$

$$Q_{s_m}^{(n)}(\lambda_m,\beta_m) = -\log\det\lambda_m K_{\beta_m} - \frac{1}{\lambda_m}\operatorname{tr}\Big(K_{\beta_m}^{-1}\hat{\boldsymbol{S}}_m^{(n)}\Big).$$

■

The proof is provided in the appendix. It is indeed seen that (12) splits into a summation of simpler terms that depend on different elements of parameter vector η. Therefore, the update of η splits into several independent and simpler optimization problems, such that many parameters in η can be computed in parallel.

1) Update of kernel hyperparameters: It can be seen that the kernel hyperparameters can be updated independently of the rest of the parameetrs. The kernel hyperparameters are updated as per the *Theorem 1* [16], [31].

Theorem 1: Define

$$Q_{\beta_m}^{(n)}(\beta_m) = \ell\log\operatorname{tr}(K_{\beta_m}^{-1}\hat{\boldsymbol{S}}_m^{(n)}) + \log\det K_{\beta_m}. \tag{16}$$

Then,

$$\begin{aligned} \hat{\beta}_m^{(n+1)} &= \underset{\beta_n\in[0,1)}{\operatorname{argmin}}\, Q_{\beta_m}^{(n)}(\beta_m)\\ \hat{\lambda}_m^{(n+1)} &= \frac{1}{\ell}\operatorname{tr}(K_{\hat{\beta}_m^{(n+1)}}^{-1}\hat{\boldsymbol{S}}_m^{(n)}). \end{aligned} \tag{17}$$

■

The optimization problem in (16) is a scalar optimization in the domain [0,1) and computationally fast. The update of $\hat{\lambda}_m^{(n+1)}$ has a closed form solution, requiring no optimization and minimal computational effort. Therefore, the update of the kernel hyperparameters turns out to be simple.

2) Update of θ and noise covariance: The updates of θ and the noise covariance parameters in η are independent of the kernel hyperparameters. Following the similar reasoning in [32], the parameters θ and Σ are updated as per the *Theorem 2*.

Algorithm 1: Algorithm for identifying a local module in a dynamic network with correlated noise

Input:$\{w_k\}_{t=1}^N, k \in \mathcal{Y} \cup \mathcal{D}$
Output: $\hat{\theta}$

1) Set $n = 0$, Initialize $\hat{\eta}^{(0)}$.
2) Compute $\hat{\boldsymbol{P}}_s^{(n)}$, $\hat{\boldsymbol{s}}$, and $\hat{\boldsymbol{S}}^{(n)}$.
3) Update the kernel hyperparameters of all the impulse responses in (7), $\hat{\beta}_m^{(n+1)}$ and $\hat{\lambda}_m^{(n+1)}$ using (17).
4) Update $\hat{\theta}^{(n+1)}$ and $\hat{\Sigma}^{(n+1)}$ using (18).
5) Set $\hat{\eta}^{(n+1)}$ based on (10).
6) Set $n = n + 1$.
7) Repeat steps (2) to (6) until convergence.

Theorem 2: Define

$$Q_\theta^{(n)}(\theta) = \det\left(\sum_{t=1}^{N} \hat{P}_\xi^{(n)}(t)\right) .$$

Then

$$\begin{aligned} \hat{\theta}^{(n+1)} &= \underset{\theta}{\operatorname{argmin}}\, Q_\theta^{(n)}(\theta) \ , \\ \hat{\Sigma}^{(n+1)} &= \frac{1}{N}\left(\sum_{t=1}^{N} \hat{P}_\xi^{(n+1)}(t)\right) \otimes I_N \ . \end{aligned} \tag{18}$$

Here, $\hat{P}_\xi^{(n)}$ is computed based on the estimates of $\hat{\eta}^{(n)}$ and $\hat{\boldsymbol{s}}^{(n)}$, whereas $\hat{P}_\xi^{(n+1)}$ is computed based on $(n+1)^{\text{th}}$ estimate of θ, $(\hat{\theta}^{(n+1)})$, and n^{th} estimate of the hyperparameters $(\hat{\lambda}_m^{(n)}$, $\hat{\beta}_m^{(n)})$ and posterior mean $\hat{\boldsymbol{s}}^{(n)}$.

■

The expression for computing $\hat{P}_\xi$ is provided in the appendix. From *Theorem* 2, Σ is updated using a closed form expression, requiring minimal computation. Non-linear optimization is performed only for updating θ which is significantly more efficient compared to solving the non-linear optimization problem in PEM with all modules parameterized in the MIMO setup. Except for θ that requires solving a non-linear optimization problem at each iteration, all other updates are simple and computationally efficient. The steps for estimating $\hat{\eta}$ is provided in Algorithm 1. Initialization can be done by randomly choosing η subject to the constraints of the hyperparameters. For terminating the algorithm, the convergence criteria is defined as $\frac{\|\hat{\eta}^{(n)} - \hat{\eta}^{(n-1)}\|}{\|\hat{\eta}^{(n-1)}\|} < 10^{-5}$.

V. Numerical Simulations

Numerical simulations are performed to validate and illustrate the developed method. To this end, we consider the dynamic network shown in Figure 1 with 3 nodes. The network is excited using known external excitation signals $r_1(t)$ and $r_3(t)$ that are realizations of white noise with unit variance. The process noises of node 2 and 3 are correlated. In this network, we intend to identify the dynamics of the module G_{21}^0 (green module). The dynamics of all the modules and the noise models are given below.

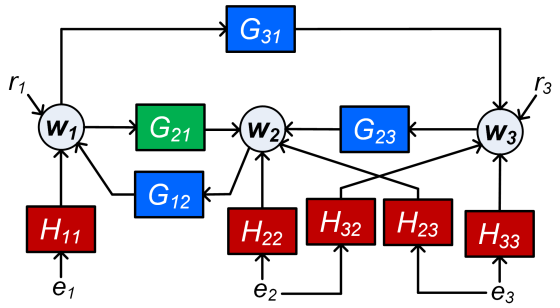

Fig. 1. A 3 node network with process noise correlated between the nodes 2 and 3: The target module is G_{21} (green box).

$$\begin{aligned} G_{21}^0 &= \tfrac{b_1 q^{-1} + b_2 q^{-2}}{1 + a_1 q^{-1} + a_2 q^{-2}} = \tfrac{1q^{-1} + 0.5q^{-2}}{1 + 0.8q^{-1} + 0.6q^{-2}} \\ G_{31}^0 &= \tfrac{-2.1q^{-1} + 2.4q^{-2}}{1 - 0.9q^{-1} - 0.1q^{-2}} \quad G_{12}^0 = \tfrac{0.03(q^{-1} + q^{-2})}{1 + 1.9q^{-1} + 0.9q^{-2}} \\ G_{23}^0 &= \tfrac{-0.2q^{-1} + 0.02q^{-2}}{1 - 0.2q^{-1} - 0.1q^{-2}} \quad H_{11}^0 = \tfrac{1 + 0.1q^{-1} - 0.03q^{-2}}{1 + 0.5q^{-1} + 0.1q^{-2}} \\ H_{22}^0 &= \tfrac{1 + 1.5q^{-1} - 0.2q^{-2}}{1 + 0.1q^{-1} - 0.01q^{-2}} \quad H_{33}^0 = \tfrac{1 - 0.4q^{-1} + 0.1q^{-2}}{1 - 0.4q^{-1} + 0.1q^{-2}} \\ H_{23}^0 &= \tfrac{0.3q^{-1} - 0.01q^{-2}}{1 - 0.4q^{-1} - 0.6q^{-2}} \quad H_{32}^0 = \tfrac{q^{-1} - q^{-2}}{1 - 1.9q^{-1} + 0.9q^{-2}} . \end{aligned}$$

We run 50 independent Monte Carlo simulations obtaining $N = 500$ data each time. The noise sources $e_1(t)$, $e_2(t)$ and $e_3(t)$ have variances of 0.1, 0.2 and 0.3 respectively. We assume that we know the model order of G_{21}^0. According to the local direct method [18], among the inputs $\{w_1, w_3\}$ that contributes to the output of the target module w_2, the noise correlation between the input w_3 and output w_2 can be handled by adding w_3 (common signal) to the output, and thereby covering the noise correlation by a $(2\times\ 2)$ noise modeling. Therefore, the input and output nodes of the MIMO estimation setup are given by $w_{\mathcal{D}} = \{w_1, w_3\}$ and $w_{\mathcal{Y}} = \{w_2, w_3\}$. We choose $\ell = 100$ for the length of impulse response vectors of the additional modules. To assess the performance of the developed method (named as *Empirical Bayes Local Direct Method* (EBLDM) for comparison), we compare it with the *Direct method* (DM) [11] and the *Two Stage Method* (TS) [11]. In the case of DM, we solve a 2-input/1- output MISO identification problem with $w_1(t)$ and $w_3(t)$ as inputs and $w_2(t)$ as output. In the two-stage method, the projection of the two inputs on the external signals $r_1(t)$ and $r_3(t)$ are used as inputs to the MISO identification problem. For both these methods, we use the *Akaike Information Criteria* (AIC) for selecting a suitable model order. Furthermore, to improve the accuracy of the estimate obtained by the Two Stage method, we also identify the noise model.

The box plot showing the fit of impulse response of G_{21}^0 is shown in Figure 2, where we have compared the performances of the direct method with true model order and the same method with model order selection step ('DM+TO' & 'DM+MOS'), the two stage method with model order selection step ('TS+MOS') and the developed EBLDM. The

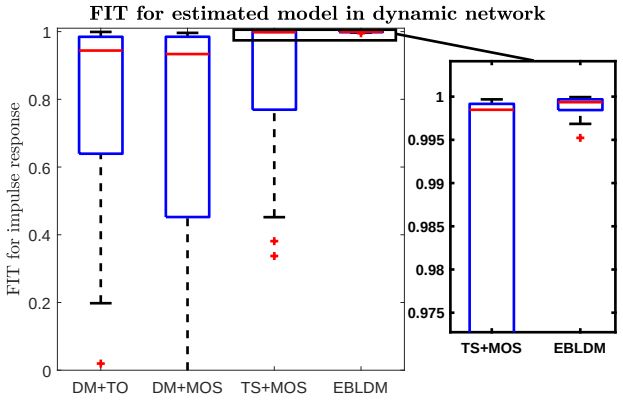

Fig. 2. Box plot of fit of the impulse response of $\hat{G}_{21}$ obtained by the two stage method, direct method and the developed method.

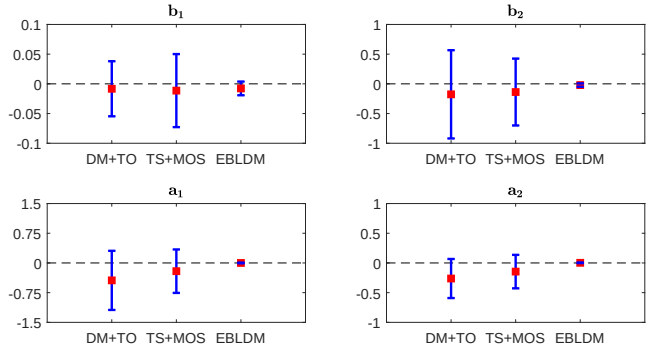

Fig. 3. Bias and standard deviation of the estimate of target module parameters

EBLDM has better overall fit of the impulse response than the classical methods. On comparing the bias and standard deviation plot of the parameters of $\hat{G}_{21}$, given in Figure 3, it is evident that the EBLDM provides smaller bias and substantially reduced variance of the estimated parameters. The variance reduction is attributed to the regularization approach used in the developed method. Among the other methods, the two stage method achieves smaller bias and variance than the direct method. A significant bias in the estimated parameters can be witnessed in the case of 'DM+TO' from Figure 3. This is in accordance with the theory that the direct method with the chosen MISO identification setup provides biased estimates under the situation of correlated noise, however, a MIMO identification setup (as in EBLDM) does not (see [18]). Overall, the developed EBLDM method proves effective for the considered relatively small network. As the size of the network grows, the results of the classical methods may further deteriorate due to the increase in number of parameterized modules and model order selection step that needs to be performed for it. Concerning this situation, EBLDM can stand out as an effective method by circumventing the model order selection step and providing reduced variance for large sized networks.

VI. Conclusion

Building on the EBDM, an effective approach for the network MIMO estimation problem that is required to identify a module in a dynamic network with correlated noise has been developed. The developed method circumvents the model order selection step for all the modules that are not of interest to the experimenter but needs to be identified for unbiased estimate of the target module. Furthermore, it uses the regularized non-parametric methods to reduce the number of estimated parameters, which reduces mean squared error of the estimated target module. Numerical simulation with an example network emphasize the potential of the introduced method in comparison with the available classical methods.

References

[1] K. R. Ramaswamy, G. Bottegal, and P. M. J. Van den Hof, "Local module identification in dynamic networks using regularized kernel-based methods," in *Proc. 57th IEEE Conf. on Decision and Control (CDC)*. IEEE, 2018, pp. 4713–4718.

[2] A. Haber and M. Verhaegen, "Subspace identication of large-scale interconnected systems," *IEEE Transactions on Automatic Control*, vol. 59, no. 10, pp. 2754–2759, 2014.

[3] P. Torres, J. W. van Wingerden, and M. Verhaegen, "Hierarchical PO-MOESP subspace identification for directed acyclic graphs," *Intern. J. Control*, vol. 88, no. 1, pp. 123–137, 2015.

[4] H. H. M. Weerts, P. M. J. Van den Hof, and A. G. Dankers, "Prediction error identification of linear dynamic networks with rank-reduced noise," *Automatica*, vol. 98, pp. 256–268, December 2018.

[5] M. Zorzi and A. Chiuso, "Sparse plus low rank network identification: a nonparametric approach," *Automatica*, vol. 76, pp. 355–366, 2017.

[6] J. Gonçalves and S. Warnick, "Necessary and sufficient conditions for dynamical structure reconstruction of LTI networks," *IEEE Trans. Automatic Control*, vol. 53, no. 7, pp. 1670–1674, Aug. 2008.

[7] H. H. M. Weerts, P. M. J. Van den Hof, and A. G. Dankers, "Identifiability of linear dynamic networks," *Automatica*, vol. 89, pp. 247–258, March 2018.

[8] J. Hendrickx, M. Gevers, and A. Bazanella, "Identifiability of dynamical networks with partial node measurements," *IEEE Trans. Autom. Control*, vol. 64, no. 6, pp. 2240–2253, 2019.

[9] H. J. van Waarde, P. Tesi, and M. K. Camlibel, "Topological conditions for identifiabaility of dynamical networks with partial node measurements," *IFAC-PapersOnLine*, vol. 51-23, pp. 319–324, 2018, proc. 7th IFAC Workshop on Distrib. Estim. and Control in Networked Systems.

[10] X. Cheng, S. Shi, and P. M. J. Van den Hof, "Allocation of excitation signals for generic identifiability of dynamic networks," in *Proc. 58th IEEE Conf. on Decision and Control (CDC)*. IEEE, 2019, pp. 5507–5512.

[11] P. M. J. Van den Hof, A. G. Dankers, P. S. C. Heuberger, and X. Bombois, "Identification of dynamic models in complex networks with prediction error methods - basic methods for consistent module estimates," *Automatica*, vol. 49, no. 10, pp. 2994–3006, 2013.

[12] A. G. Dankers, P. M. J. Van den Hof, X. Bombois, and P. S. C. Heuberger, "Errors-in-variables identification in dynamic networks – consistency results for an instrumental variable approach," *Automatica*, vol. 62, pp. 39–50, 2015.

[13] D. Materassi and M. Salapaka, "Identification of network components in presence of unobserved nodes," in *Proc. 2015 IEEE 54th Conf. Decision and Control, Osaka, Japan*, 2015, pp. 1563–1568.

[14] A. G. Dankers, P. M. J. Van den Hof, P. S. C. Heuberger, and X. Bombois, "Identification of dynamic models in complex networks with prediction error methods: Predictor input selection," *IEEE Trans. on Automatic Control*, vol. 61, no. 4, pp. 937–952, 2016.

[15] J. Linder and M. Enqvist, "Identification of systems with unknown inputs using indirect input measurements," *International Journal of Control*, vol. 90, no. 4, pp. 729–745, 2017.

[16] N. Everitt, G. Bottegal, and H. Hjalmarsson, "An empirical bayes approach to identification of modules in dynamic networks," *Automatica*, vol. 91, pp. 144–151, 5 2018.

[17] M. Gevers, A. Bazanella, and G. Vian da Silva, "A practical method for the consistent identification of a module in a dynamical network," *IFAC-PapersOnLine*, vol. 51-15, pp. 862–867, 2018, proc. 18th IFAC Symp. System Identif. (SYSID2018).

[18] K. R. Ramaswamy and P. M. J. Van den Hof, "A local direct method for module identification in dynamic networks with correlated noise," Tech. Rep., 2019, arXiv:1908.00976. Provisionally accepted by IEEE Trans. Automatic Control.
[19] K. R. Ramaswamy, P. M. J. Van den Hof, and A. G. Dankers, "Generalized sensing and actuation schemes for local module identification in dynamic networks," in *Proc. 58th IEEE Conf. on Decision and Control (CDC)*. IEEE, 2019, pp. 5519–5524.
[20] L. Ljung, *System Identification: Theory for the User*. Englewood Cliffs, NJ: Prentice-Hall, 1999.
[21] A. G. Dankers, P. M. J. Van den Hof, D. Materassi, and H. H. M. Weerts, "Conditions for handling confounding variables in dynamic networks," *IFAC-PapersOnLine*, vol. 50, no. 1, pp. 3983–3988, 2017, proc. 20th IFAC World Congress.
[22] T. Söderström and P. Stoica, *System Identification*. Hemel Hempstead, UK: Prentice-Hall International, 1989.
[23] H. H. M. Weerts, M. Galrinho, G. Bottegal, H. Hjalmarsson, and P. M. J. Van den Hof, "A sequential least squares algorithm for ARMAX dynamic network identification," *IFAC-PapersOnLine*, vol. 51-15, pp. 844 – 849, 2018, proc. 18th IFAC Symp. System Identif. (SYSID2018).
[24] 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*, vol. 50, no. 3, pp. 657–682, 2014.
[25] C. E. Rasmussen and C. K. I. Williams, *Gaussian processes for machine learning*. The MIT Press, 2006.
[26] T. Chen, H. Ohlsson, and L. Ljung, "On the estimation of transfer functions, regularizations and gaussian processes - revisited," *Automatica*, vol. 48, no. 8, pp. 1525–1535, 2012.
[27] J. S. Maritz and T. Lwin, *Empirical Bayes Methods*. Chapman and Hall, 1989.
[28] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 1, pp. 1–38, 1977.
[29] R. A. Boyles, "On the convergence of the EM algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 45, no. 1, pp. 47–50, 1983.
[30] B. D. O. Anderson and J. B. Moore, *Optimal filtering*. Englewood Cliffs, N.J., USA: Prentice-Hall, 1979.
[31] G. Bottegal, A. Y. Aravkin, H. Hjalmarsson, and G. Pillonetto, "Robust EM kernel-based methods for linear system identification," *Automatica*, vol. 67, pp. 114–126, 2016.
[32] K. J. Åström, "Maximum likelihood and prediction error methods," *Automatica*, vol. 16, no. 5, pp. 551 – 574, 1980.

APPENDIX

A. Proof of Lemma 1

Following Bayes' theorem, (12) can be written as follows,

$$Q^{(n)}(\eta) = \mathbb{E}[\log p(\tilde{w}_{\mathcal{Y}}|\boldsymbol{s};\eta)] + \mathbb{E}[\log p(\boldsymbol{s};\eta)] \tag{19}$$

Define

$$\mathcal{A} = -\frac{N|\mathcal{Y}|}{2}\log(2\pi) - \frac{1}{2}\log\det\Sigma - \frac{1}{2}\left(\tilde{w}_{\mathcal{Y}} - \begin{bmatrix} W_i \\ \mathbf{0} \end{bmatrix} g_{ji} - \mathbf{W}\boldsymbol{s}\right)^{\top} \Sigma^{-1} \left(\tilde{w}_{\mathcal{Y}} - \begin{bmatrix} W_i \\ \mathbf{0} \end{bmatrix} g_{ji} - \mathbf{W}\boldsymbol{s}\right), \tag{20}$$

$$\mathcal{B} = \sum_m \left(-\frac{\ell}{2}\log(2\pi) - \frac{1}{2}\log\det\lambda_m K_{\beta_m} - \frac{1}{2} s_m^{\top} (\lambda_m K_{\beta_m})^{-1} s_m \right) \tag{21}$$

Using properties of trace, removing constant terms, multiplying by 2 and taking the expectation with respect to posterior, (20) and (21) are written as,

$$\mathbb{E}[\mathcal{A}] = -\log\det\Sigma - \mathrm{tr}\Bigg(\Sigma^{-1} \Bigg(\tilde{w}_{\mathcal{Y}}\tilde{w}_{\mathcal{Y}}^{\top} + \begin{bmatrix} W_i \\ \mathbf{0} \end{bmatrix} g_{ji} g_{ji}^{\top} \begin{bmatrix} W_i^{\top} & \mathbf{0}^{\top} \end{bmatrix} + \mathbf{W}\boldsymbol{s}\boldsymbol{s}^{\top}\mathbf{W}^{\top} - \begin{bmatrix} W_i \\ \mathbf{0} \end{bmatrix} g_{ji}\tilde{w}_{\mathcal{Y}}^{\top} - \mathbf{W}\boldsymbol{s}\tilde{w}_{\mathcal{Y}}^{\top} - \tilde{w}_{\mathcal{Y}} g_{ji}^{\top} \begin{bmatrix} W_i^{\top} & \mathbf{0}^{\top} \end{bmatrix} + \mathbf{W}\boldsymbol{s} g_{ji}^{\top} \begin{bmatrix} W_i^{\top} & \mathbf{0}^{\top} \end{bmatrix} - \tilde{w}_{\mathcal{Y}}\boldsymbol{s}^{\top}\mathbf{W}^{\top} + \begin{bmatrix} W_i \\ \mathbf{0} \end{bmatrix} g_{ji}\boldsymbol{s}^{\top}\mathbf{W}^{\top} \Bigg) \Bigg), \tag{22}$$

$$\mathbb{E}[\mathcal{B}] = \sum_m -\log\det\lambda_m K_{\beta_m} - \mathrm{tr}\left((\lambda_m K_{\beta_m})^{-1} s_m s_m^{\top} \right) \tag{23}$$

By substituting, $\boldsymbol{s}$ as $\hat{\boldsymbol{s}}^{(n)}$, $\boldsymbol{s}\boldsymbol{s}^{\top}$ as $\hat{\boldsymbol{S}}^{(n)}$, s_m as $\hat{s}_m^{(n)}$ and $s_m s_m^{\top}$ as $\hat{S}_m^{(n)}$ in (22) and (23), (15) is obtained.

B. Proof of Theorem 1

We consider $Q_{s_m}(\lambda_m, \beta_m)$ in (15) and differentiate it with respect to λ_m. The derivative is then equated to 0 to obtain the expression for λ_m

$$\lambda_m = \frac{1}{\ell}\mathrm{tr}(K_{\beta_m})^{-1}\hat{S}_m. \tag{24}$$

(24) is then substituted in the $Q_{s_m}(\lambda_m, \beta_m)$ to eliminate λ_m, and with the change of sign, resulting in the following equation.

$$Q_B(\beta_m) = \ell\log\mathrm{tr}(K_{\beta_m}^{-1}\hat{S}_m) + \log\det K_{\beta_m} \tag{25}$$

Once β_m is obtained, then (24) is used to obtain λ_m.

C. Computation of $\hat{P}_{\xi}$

Let us define $W_1(t)$ and $W_2(t)$ as,

$$W_1(t) = \begin{bmatrix} \mathbf{W}(t,*) \\ \mathbf{W}(t+N,*) \\ \vdots \\ \mathbf{W}(t+(N_{\mathcal{Y}}-1)N,*) \end{bmatrix}, \tag{26}$$

$$W_2(t) = \begin{bmatrix} W_{ji}(t,*) \\ W_{ji}(t+N,*) \\ \vdots \\ W_{ji}(t+(N_{\mathcal{Y}}-1)N,*) \end{bmatrix} \tag{27}$$

where, $\mathbf{W}(t,*)$ corresponds to the t^{th} row of the matrix $\mathbf{W}$, and $W_{ji} = \begin{bmatrix} W_i^{\top} & \mathbf{0}^{\top} \end{bmatrix}^{\top}$. With the above definitions, we rewrite the cost function for updating θ and Σ as follows,

$$Q_0(\theta,\Sigma) = -N\log\det\bar{\Sigma} - \sum_{t=1}^{N} \mathrm{tr}\Big(\bar{\Sigma}^{-1} \big(\tilde{w}_{\mathcal{Y}} - W_1(t) g_{ji} - W_2(t)\boldsymbol{s}\big) \big(\tilde{w}_{\mathcal{Y}} - W_1(t) g_{ji} - W_2(t)\boldsymbol{s}\big)^{\top} \Big) \tag{28}$$

Since $\bar{\Sigma}$ is parameterized independently, (28) is differentiated with $\bar{\Sigma}^{-1}$ and equated to 0 (refer to [32] for details). The

obtained expression is substituted in (28) to get the following cost function for θ.

$$
\begin{aligned}
Q_{\theta}^{(n)}(\theta) =& \det\left(\sum_{t=1}^{N}\hat{P}_{\xi}^{(n)}(t)\right), \text{where,} \\
\hat{P}_{\xi}^{(n)}(t) =& \tilde{w}_{\mathcal{Y}}(t)\tilde{w}_{\mathcal{Y}}^{\top}(t) + W_2(t)\hat{g}_{ji}^{(n)}\hat{g}_{ji}^{(n)\top}W_2^{\top}(t) \\
&+ W_1(t)\hat{\boldsymbol{S}}^{(n)}W_1^{\top}(t) - W_2(t)\hat{g}_{ji}^{(n)}\tilde{w}_{\mathcal{Y}}^{\top}(t) \\
&- W_1(t)\hat{\boldsymbol{s}}^{(n)}\tilde{w}_{\mathcal{Y}}^{\top}(t) - \tilde{w}_{\mathcal{Y}}(t)\hat{g}_{ji}^{(n)\top}W_2^{\top}(t) \\
&+ W_1(t)\hat{\boldsymbol{s}}^{(n)}\hat{g}_{ji}^{(n)\top}W_2^{\top}(t) - \tilde{w}_{\mathcal{Y}}\hat{\boldsymbol{s}}^{(n)\top}W_1^{\top}(t) \\
&+ W_2(t)\hat{g}_{ji}^{(n)}\hat{\boldsymbol{s}}^{(n)\top}W_1^{\top}(t)
\end{aligned} \tag{29}
$$

$\hat{P}_{\xi}^{(n+1)}(t)$ is obtained by updating $\hat{g}_{ji}^{(n+1)}$ and recomputing (29).