An empirical Bayes approach to identification of modules in dynamic networks

Niklas Everitt [a], Giulio Bottegal [b], Håkan Hjalmarsson [a]

[a] *ACCESS Linneaus Center, School of Electrical Engineering, KTH Royal Institute of Technology, Sweden*

[b] *Department of Electrical Engineering, TU Eindhoven, The Netherlands*

Abstract

We present a new method of identifying a specific module in a dynamic network, possibly with feedback loops. Assuming known topology, we express the dynamics by an acyclic network composed of two blocks where the first block accounts for the relation between the known reference signals and the input to the target module, while the second block contains the target module. Using an empirical Bayes approach, we model the first block as a Gaussian vector with covariance matrix (kernel) given by the recently introduced stable spline kernel. The parameters of the target module are estimated by solving a marginal likelihood problem with a novel iterative scheme based on the Expectation-Maximization algorithm. Additionally, we extend the method to include additional measurements downstream of the target module. Using Markov Chain Monte Carlo techniques, it is shown that the same iterative scheme can solve also this formulation. Numerical experiments illustrate the effectiveness of the proposed methods.

Key words: system identification, dynamic network, empirical Bayes, expectation-maximization.

1 Introduction

Networks of dynamical systems are everywhere, and applications are in different branches of science, e.g., econometrics, systems biology, social science, and power systems. Identification of these networks, usually referred to as *dynamic networks*, has been given increasing attention in the system identification community, see e.g., [1], [2], [3]. In this paper, we use the term "dynamic network" to mean the interconnection of *modules*, where each module is a linear time-invariant (LTI) system. The interconnecting signals are the outputs of these modules; we also assume that exogenous measurable signals may affect the dynamics of the network.

Two main problems arise in dynamic network identification. The first is topology detection, that is, understanding the interconnection between the modules. The second problem is the identification of one or more specific modules in the network. Some recent papers deal with both the aforementioned problems [4,5,1], whereas others are mainly focused on the identification of a single module in the network [6,7,8,9,10]. As observed in [2], dynamic networks with known topology can be seen as a generalization of simple compositions, such as systems in parallel, series or feedback connection. Therefore, identification techniques for dynamic networks may be derived by extending methods already developed for simple structures. This is the idea underlying the method proposed in [7], which generalized the results of [11] for the identification of cascaded systems to the context of dynamic networks. In that work, the underlying idea is that a dynamic network can be transformed into an acyclic structure, where any reference signal of the network is the input to a cascaded system consisting of two LTI blocks. In this alternative system description, the first block captures the relation between the reference and the noisy input of the target module, the second block contains the target module. The two LTI blocks are identified simultaneously using the prediction error method (PEM) [12]. In this setup, determining the model structure of the first block of the cascaded structure may be

⋆ This work was supported by the Swedish Research Council under contracts 2015-05285 and NewLEADS 2016-06079, by the European Research Council under the advanced grant LEARN, contract 267381, and by the European Research Council (ERC), Advanced Research Grant SYSDYNET, under the European Union's Horizon 2020 research and innovation programme (grant agreement No 694504).

Email addresses: `neveritt@kth.se` (Niklas Everitt), `g.bottegal@tue.nl` (Giulio Bottegal), `hjalmars@kth.se` (Håkan Hjalmarsson).

complicated, due to the possibly large number of interconnections in the dynamic network. Furthermore, it requires knowledge of the model structure of essentially all modules in the feedback loop. Therefore, in [7], the first block is modeled by an unstructured finite impulse response (FIR) model of high order. The major drawback of this approach is that, as is usually the case with estimated models of high order, the variance of the estimated FIR model is high. The uncertainty in the estimate of the FIR model of the first block will in turn decrease the accuracy of the estimated target module.

The objective of this paper is to propose a method for the identification of a module in dynamic networks that circumvents the high variance that is due to the high order model of the first block. Our main focus is on the identification of a specific module, which we assume is well described through a low-order parametric model. Following a recent trend in system identification [13], we use regularization to control the covariance of the identified sensitivity path by modeling its impulse response as a zero-mean stochastic process. The covariance matrix of this process is given by the recently introduced *stable spline kernel* [14], whose structure is parametrized by two *hyperparameters*.

We also consider the case where more sensors spread in the network are used in the identification of the target module, motivated by the fact that adding information through addition of measurements used in the identification process has the potential to further reduce the variance of the estimated module [15]. To avoid too many additional parameters to estimate, we model also the impulse response of the path linking the target module to any additional sensor as a Gaussian process.

An estimate of the target module is obtained by empirical Bayes (EB) arguments, that is, by maximization of the marginal likelihood of the available measurements [13]. This likelihood does not admit an analytical expression and depends not only on the parameter of the target module, but also on the kernel hyperparameters and the variance of the measurement noise. To estimate all these quantities, we design a novel iterative scheme based on an EM-type algorithm [16], known as the Expectation/Conditional-Maximization (ECM) algorithm [17]. This algorithm alternates the so called expectation step (E-step) with a series of conditional-maximization steps (CM-steps) that, in the problem under analysis, consist of relatively simple optimization problems, which either admit a closed form solution, or can be efficiently solved using gradient descent strategies. As for the E-step, we are required to compute an integral that, in the general case of multiple downstream sensor, does not admit an analytical expression. To overcome this issue, we use Markov Chain Monte Carlo (MCMC) techniques [18] to solve the integral associated with the E-step. In particular, we design an integration scheme based on the Gibbs sampler [19] that, in combination with the ECM method, builds up a novel identification method for the target module.

A part of this paper has previously been presented in [20]. More specifically, the case where only the sensors directly measuring the input and the output of the target module are used in the identification process were partly covered in [20], whereas, the general case where more sensors spread in the network are used in the identification of the target module is completely novel.

2 Problem Statement

We consider dynamic networks that consist of L scalar *internal variables* $w_j(t)$, $j = 1, \dots, L$ and L scalar external *reference signals* $r_l(t)$, $l = 1, \dots, L$. We do not state any specific requirement on the reference signals (i.e., we do not assume any condition on persistent excitation in the input [12]), requiring only $r_l(t) \neq 0$, $l = 1, \dots, L$, for some t. Notice, however, that even though the method presented in this paper does not require any specifics of the input, the resulting estimate is of course highly dependent on the properties of the input [12]. Some of the reference signal may not be present, i.e., they may be identically zero. Define $\mathcal{R}$ as the set of indices of reference signals that are present. In the dynamic network, the internal variables are considered nodes and transfer functions are the edges. Introducing the vector notation $w(t) := [w_1(t) \dots w_L(t)]^\top$, $r(t) := [r_1(t) \dots r_L(t)]^\top$, the dynamics of the network are defined by the equation

$$w(t) = \mathcal{G}(q)w(t) + r(t), \tag{1}$$

with

$$\mathcal{G}(q) = \begin{bmatrix} 0 & G_{12}(q) & \cdots & G_{1L}(q) \\ G_{21}(q) & 0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & G_{(L-1)L}(q) \\ G_{L1}(q) & \cdots & G_{L(L-1)}(q) & 0 \end{bmatrix},$$

where $G_{ji}(q)$ is a proper rational transfer function for $j = 1, \dots, L$, $i = 1, \dots, L$. The internal variables $w(t)$ are measured with additive white noise, that is

$$\tilde{w}(t) = w(t) + e(t),$$

where $e(t) \in \mathbb{R}^L$ is a stationary zero-mean Gaussian white-noise process with diagonal noise covariance matrix $\Sigma_e = \operatorname{diag}\{\sigma_1^2, \dots, \sigma_L^2\}$. We assume that the σ_i^2 are unknown. To ensure stability and causality of the network, the following assumptions hold for all networks considered in this paper.

Assumption 2.1 *The network is well posed in the sense that all principal minors of* $\lim_{q\to\infty}(I - \mathcal{G}(q))$ *are non-*

zero [2]. Furthermore, the sensitivity path $S(q) = (I - \mathcal{G}(q))^{-1}$ is stable

Assumption 2.2 *The reference variables $\{r_l(t)\}$ are mutually uncorrelated and uncorrelated with the measurement noise $e(t)$.*

Remark 2.1 *We note that, compared to e.g. [2], the dynamic network model treated in this paper does not include process noise (but in turn includes sensor noise). Process noise complicates the analysis and the derivation of the method, and will be treated in future publications.*

The network dynamics can then be rewritten as

$$\tilde{w}(t) = S(q)r(t) + e(t). \tag{2}$$

We define $\mathcal{N}_j$ as the set of indices of internal variables that have a direct causal connection to w_j, i.e., $i \in \mathcal{N}_j$ if and only if $G_{ji}(q) \neq 0$. Without loss of generality, we assume that $\mathcal{N}_j = \{1, 2, \ldots, p\}$, where p is the number of direct causal connections to w_j (we may always rename the nodes so that this holds). The goal is to identify module $G_{j1}(q)$ given N measurements of the reference $r(t)$, the "output" $\tilde{w}_j(t)$ and the set of p neighbor signals in $\mathcal{N}_j$. To this end, we express $\tilde{w}_j$, the measured output of module $G_{j1}(q)$ as

$$\tilde{w}_j(t) = \sum_{i\in\mathcal{N}_j} G_{ji}(q)w_i(t) + r_j(t) + e_j(t). \tag{3}$$

The above equation depends on the internal variables $w_i(t)$, $i \in \mathcal{N}_j$, which we we only have noisy measurement of; these can be expressed as

$$\tilde{w}_i(t) = w_i(t) + e_i(t) = \sum_{l\in\mathcal{R}} S_{il}(q)r_l(t) + e_i(t)\,. \tag{4}$$

where $S_{il}(q)$ is the transfer function path from reference $r_l(t)$ to output $\tilde{w}_i(t)$. Together, (3) and (4) allow us to express the relevant part of the network, possibly containing feedback loops, as a direct acyclic graph with two blocks connected in cascade. Note that, in general, the first block depends on all other blocks in the network. Therefore, accurate low order parameterization of this block depends on global knowledge of the network.

Example 2.1 *As an example, consider the network depicted in Figure 1, where, using (3) and (4), the acyclic graph of Figure 2 can describe the relevant dynamics, when $w_j = w_3$ is the output and we wish to identify $G_{31}(q)$.*

In the following, we briefly review two standard methods for closed-loop identification that we will use as starting point to derive the methodology described in the paper.

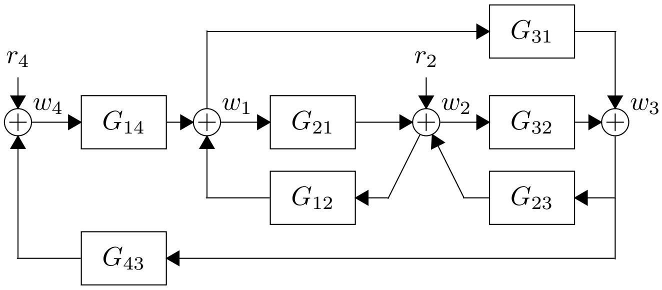

Fig. 1. Network example of 4 internal variables and 2 reference signals.

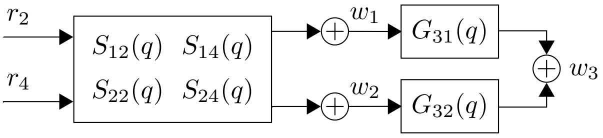

Fig. 2. Direct acyclic graph of part of the network in Figure 1.

Two-stage method: The first stage of the two-stage method [2], proceeds by finding a consistent estimate $\hat{w}_i(t)$ of all nodes $w_i(t)$ in $\mathcal{N}_j$. This is done by high-order modeling of $\{S_{il}\}$ and estimating it from (4) using the prediction error method. The prediction errors are constructed as

$$\varepsilon_i(t,\alpha) = \tilde{w}_i(t) - \sum_{l\in\mathcal{R}} S_{il}(q,\alpha)r_l(t), \tag{5}$$

where α is a parameter vector. The resulting estimate $S_{il}(q,\hat{\alpha})$ is then used to obtain the node estimate as

$$\hat{w}_i(t) = \sum_{l\in\mathcal{R}} S_{il}(q,\hat{\alpha})r_l(t). \tag{6}$$

In a second stage, the module of interest $G_{j1}(q)$ (and the other modules in $\mathcal{N}_j$) is parameterized by θ and estimated from (3), again using the prediction error method. The prediction errors are now constructed as

$$\varepsilon_j(t,\theta) = \tilde{w}_j(t) - r_j(t) - \sum_{i\in\mathcal{N}_j} G_{ji}(q,\theta)\hat{w}_i(t). \tag{7}$$

Simultaneous minimization of prediction errors: It is useful to briefly introduce the simultaneous minimization of prediction error method (SMPE) [7]. The main idea underlying SMPE is that if the two prediction errors (5) and (7) are simultaneously minimized, the variance will be decreased [11]. In the SMPE method, the prediction error of the measurement $\tilde{w}_j$ depends explicitly on α and is given by

$$\varepsilon_j(t,\theta,\alpha) = \tilde{w}_j(t) - \sum_{i\in\mathcal{N}_j} G_{ji}(q,\theta) \sum_{l\in\mathcal{R}} S_{il}(q,\alpha)r_l(t). \tag{8}$$

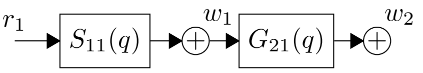

Fig. 3. Basic network of 1 reference signal and 2 internal variables.

The method proceeds to minimize

$$V_N(\theta, \alpha) = \frac{1}{N} \sum_{t=1}^{N} \left[\frac{\varepsilon_j^2(t, \theta, \alpha)}{\lambda_j} + \sum_{i \in \mathcal{N}_j} \frac{\varepsilon_i^2(t, \alpha)}{\lambda_i} \right]. \quad (9)$$

In [7], the noise variances are assumed known, and how to estimate the noise variances is not analyzed. As an initial estimate of the parameters θ and α, the minimizers of the two-stage method can be taken.

The main drawback is that the least-squares estimation of S may still induce high variance in the estimates. Additionally, if each of the n_s estimated transfer functions in S is estimated by the first n impulse response coefficients, the number of estimated parameters in S alone is $n_s \cdot n$. Already for relatively small dimensions of S the SMPE method is prohibitively expensive. To handle this, a frequency domain approach is taken in [21]. In this paper, we will instead use regularization to reduce the variance and the complexity.

3 Empirical Bayes estimation of the module

In this section we derive our approach to the identification of a specific module based on empirical Bayes (EB). For ease of exposition, we give a detailed derivation in the one-reference-one-module case. The extension to general dynamic networks follows along similar arguments and can be found in [22]. We first describe the proposed method in the setup where only one sensor downstream the target module is used. In the next section, we will focus on the general multi-sensor case.

We consider a dynamic network with one non-zero reference signal $r_1(t)$. Without loss of generality, we assume that the module of interest is $G_{21}(q)$, and hence $G_{22}(q), \ldots, G_{2L}(q)$ are assumed zero (We can always rename the signals such that this holds). The setting we consider has been illustrated in Figure 3. We parametrize the target module by means of a parameter vector $\theta \in \mathbb{R}^{n_\theta}$. Using the vector notation introduced in the previous section, we denote by $\tilde{w}_1$ the stacked measurements $\tilde{w}_1(t)$ before the module of interest $G_{21}(q, \theta)$, and by $\tilde{w}_2$ the stacked output of this module $\tilde{w}_2(t)$. We define the impulse response coefficients of $G_{21}(q, \theta)$ by the inverse discrete-time Fourier transform, and denote it by $g_\theta(t)$. Similarly we define s_{11} as the impulse response coefficients of $S_{11}(q)$, where $S_{11}(q)$ is, as before, the sensitivity path from $r_1(t)$ to $w_1(t)$, and $e_1(t)$ and $e_2(t)$ are the measurement noise sources (which we have assumed white and Gaussian). Their variance is denoted by σ_1^2 and σ_2^2, respectively. We rewrite the dynamics as

$$\begin{aligned} \tilde{w}_1 &= \quad R_1 s_{11} + e_1 \,, \\ \tilde{w}_2 &= G_\theta R_1 s_{11} + e_2 \,. \end{aligned} \quad (10)$$

where G_θ is the $N \times N$ lower triangular Toeplitz matrix of the N first impulse response samples g_θ, and $R_1$is the Toeplitz matrix of r_1. For computational purposes, we only consider the first n samples of s_{11}, where n is large enough such that the truncation captures the dynamics of the sensitivity $S_{11}(q)$ well enough. Let $z := [\tilde{w}_1^\top \; \tilde{w}_2^\top]^\top$ and let e be defined similarly; we rewrite (10) as

$$z = W_\theta s_{11} + e \,, \qquad W_\theta = \begin{bmatrix} R_1^\top & R_1^\top G_\theta^\top \end{bmatrix}^\top \quad (11)$$

Note that e is a random vector such that $\Sigma_e := \mathbb{E}[ee^\top] = \operatorname{diag}\{\sigma_1^2 I, \, \sigma_2^2 I\}$.

3.1 Bayesian model of the sensitivity path

To reduce the variance in the sensitivity estimate (and also reduce the number of estimated parameters), we cast our problem in a Bayesian framework and model the sensitivity function as a zero-mean Gaussian stochastic vector [23], i.e., $p(s_{11}; \lambda, K_\beta) \sim \mathcal{N}(0, \lambda K_\beta)$. The structure of the covariance matrix is given by the *first-order stable spline kernel* [14], whose structure obeys $\{K_\beta\}_{i,j} = \beta^{\max(i,j)}$. The parameter $\beta \in [0, 1)$ regulates the decay velocity of the realizations from the prior, whereas, λ tunes their amplitude. In this context, K_β is usually called a *kernel* (due to the connection between Gaussian process regression and the theory of reproducing kernel Hilbert space, see e.g. [23] for details) and determines the properties of the realizations of s. In particular, the stable spline kernel enforces smooth and BIBO stable realizations [14].

3.2 The marginal likelihood estimator

Since s_{11} is assumed stochastic, it admits a probabilistic description jointly with the vector of observations z, parametrized by the vector $\eta = [\sigma_1^2 \, \sigma_2^2 \, \lambda \, \beta \, \theta]$. The posterior distribution of s_{11} given the measurement vector z is also Gaussian, given by (see e.g. [24])

$$p(s_{11} | z; \eta) \sim \mathcal{N}(P W_\theta^\top \Sigma_e^{-1} z, P) \,, \quad (12)$$

$$P = (W_\theta^\top \Sigma_e^{-1} W_\theta + (\lambda K_\beta)^{-1})^{-1} \,, \quad (13)$$

and it is parametrized by the vector η. The module identification strategy we propose in this paper relies on an empirical Bayes approach. We introduce the marginal probability density function (pdf) of the measurements

$$p(z; \eta) = \int p(z, \, s_{11}) \, ds_{11} \sim \mathcal{N}(0, \, \Sigma_z) \,, \quad (14)$$

where $\Sigma_z = W_\theta \lambda K_\beta W_\theta^\top + \Sigma_e$. Then, we can define the maximum (log) marginal likelihood (ML) criterion as the maximum of the (log) marginal pdf $p(z;\, \eta)$ defined above, whose solution provides also an estimate of θ and thus of the module of interest.

4 Computation of the solution of the marginal likelihood criterion

Maximization of the marginal likelihood is a nonlinear problem that may involve a large number of decision variables, if n_θ is large. In this section, we derive an iterative solution scheme based on the Expectation/Conditional-Maximization (ECM) algorithm [17], which is a generalization of the standard Expectation-Maximization (EM) algorithm and will in our case converge to a stationary point just as EM does. In order to employ EM-type algorithms, one has to define a *latent variable*; in our problem, a natural choice is s_{11}. Then, a (local) solution to the log ML criterion is achieved by iterating over the following steps:

(E-step) Given an estimate $\hat{\eta}^{(k)}$ (computed at the k-th iteration of the algorithm), compute

$$Q^{(k)}(\eta) := \mathbb{E}\left[\log p(z,\, s_{11};\, \eta)\right] , \tag{15}$$

where the expectation is taken with respect to the posterior of s_{11} when the estimate $\eta^{(k)}$ is used, i.e., $p(s_{11}|z,\, \hat{\eta}^{(k)})$;

(M-step) Solve the problem $\hat{\eta}^{(k+1)} = \arg\max_\eta Q^{(k)}(\eta)$.

First, we turn our attention on the computation of the E-step, i.e., the derivation of (15). Let $\hat{s}_{11}^{(k)}$ and $\hat{P}^{(k)}$ be the posterior mean and covariance matrix of s_{11}, computed from (12) using $\hat{\eta}^{(k)}$. Define $\hat{S}_{11}^{(k)} := \hat{P}^{(k)} + \hat{s}_{11}^{(k)} \hat{s}_{11}^{(k)T}$. The following lemma provides an expression for the function $Q^{(k)}(\eta)$.

Lemma 4.1 *Let $\hat{\eta}^{(k)} = [\hat{\sigma}_1^{2(k)}\; \hat{\sigma}_2^{2(k)}\; \hat{\lambda}^{(k)}\; \hat{\beta}^{(k)}\; \hat{\theta}^{(k)}]$ be an estimate of η after the k-th iteration of the EM method. Then $Q^{(k)}(\eta) = -\frac{1}{2} Q_0^{(k)}(\sigma_1^2,\, \sigma_2^2,\, \theta) - \frac{1}{2} Q_s^{(k)}(\lambda,\, \beta)$, where*

$$\begin{aligned} Q_0^{(k)}(\sigma_1^2,\, \sigma_2^2,\, \theta) = \Big(&\log\det\{\Sigma_e\} + z^\top \Sigma_e^{-1} z - 2 z^\top W_\theta \hat{s}_{11}^{(k)} \\ &+ \mathrm{Tr}\Big\{ W_\theta^\top \Sigma_e^{-1} W_\theta \hat{S}_{11}^{(k)} \Big\}\Big) , \\ Q_s^{(k)}(\lambda,\, \beta) = \log&\det\{\lambda K_\beta\} + \mathrm{Tr}\Big\{ (\lambda K_\beta)^{-1}\, \hat{S}_{11}^{(k)} \Big\} . \end{aligned}$$

Having computed the function $Q^{(k)}(\eta)$, we now focus on its maximization. We first note that the decomposition of $Q^{(k)}(\eta)$ shows that the kernel hyperparameters can be updated independently of the rest of the parameters as shown in the following proposition (see [25] for a proof).

Proposition 4.1 *Define $Q_\beta(\beta) = \log\det\{K_\beta\} + n \log \mathrm{Tr}\Big\{ K_\beta^{-1} \hat{S}_{11}^{(k)} \Big\}$. Then*

$$\hat{\beta}^{(k+1)} = \arg\min_{\beta \in [0,1)} Q_\beta(\beta) \,, \; \hat{\lambda}^{(k+1)} = \frac{\mathrm{Tr}\Big\{ K_{\hat{\beta}^{(k+1)}}^{-1} \hat{S}_{11}^{(k)} \Big\}}{n} . \tag{16}$$

Therefore, the update of the scaling hyperparameter is available in closed-form, while the update of β requires the solution of a scalar optimization problem in the domain [0, 1), an operation that requires little computational effort, see [25] for details.

We are left with the maximization of the function $Q_0^{(k)}(\sigma_1^2,\, \sigma_2^2,\, \theta)$. In order to simplify this step, we split the optimization problem into constrained subproblems that involve fewer decision variables. This operation is justified by the ECM paradigm, which, under mild conditions [17], guarantees the same convergence properties of the EM algorithm even when the optimization of $Q^{(k)}(\eta)$ is split into a series of constrained subproblems. In our case, we decouple the update of the noise variances from the update of θ. By means of the ECM paradigm, we split the maximization of $Q_0^{(k)}(\sigma_1^2,\, \sigma_2^2,\, \theta)$ in a sequence of two constrained optimization subproblems:

$$\begin{aligned} \hat{\theta}^{(k+1)} &= \arg\max_\theta Q_0^{(k)}(\sigma_1^2,\, \sigma_2^2,\, \theta) && (17) \\ &\quad \text{s.t. } \sigma_1^2 = \hat{\sigma}_1^{2(k)},\; \sigma_2^2 = \hat{\sigma}_2^{2(k)} , \\ \hat{\sigma}_1^{2(k+1)},\, \hat{\sigma}_2^{2(k+1)} &= \arg\max_{\sigma_1^2,\, \sigma_2^2} Q_0^{(k)}(\sigma_1^2,\, \sigma_1^2,\, \theta) && (18) \\ &\quad \text{s.t. } \theta = \hat{\theta}^{(k+1)} . \end{aligned}$$

The following result provides the solution of the above problems.

Proposition 4.2 *Introduce the matrix $D \in \mathbb{R}^{N^2 \times N}$ such that $Dv = \mathrm{vec}\{\mathcal{T}_N\{v\}\}$, for any $v \in \mathbb{R}^N$. Define*

$$\hat{A}^{(k)} = D^\top (R_1 \hat{S}_{11}^{(k)} R_1^\top \otimes I_N) D \,,\; \hat{b}^{(k)} = \mathcal{T}_N\Big\{ R_1 \hat{s}_{11}^{(k)} \Big\}^\top \tilde{w}_2 .$$

Then

$$\hat{\theta}^{(k+1)} = \arg\min_\theta g_\theta^\top \hat{A}^{(k)} g_\theta - 2 \hat{b}^{(k)\top} g_\theta . \tag{19}$$

The closed form updates of the noise variances are as follows

$$\begin{aligned} \hat{\sigma}_1^{2(k+1)} &= \frac{1}{N}\Big(\|\tilde{w}_1 - R_1 \hat{s}_{11}^{(k)}\|_2^2 + \mathrm{Tr}\Big\{ R_1 \hat{P}^{(k)} R_1^\top \Big\} \Big) , \\ \hat{\sigma}_2^{2(k+1)} &= \frac{1}{N}\Big(\|\tilde{w}_2 - G_{\hat{\theta}^{(k+1)}} R_1 \hat{s}_{11}^{(k)}\|_2^2 \\ &\quad + \mathrm{Tr}\Big\{ G_{\hat{\theta}^{(k+1)}} R_1 \hat{P}^{(k)} R_1^\top G_{\hat{\theta}^{(k+1)}}^\top \Big\} \Big) . \end{aligned} \tag{20}$$

Each variance is the result of the sum of one term that measures the adherence of the identified systems to the data and one term that compensates for the bias in the estimates introduced by the Bayesian approach. The update of the parameter θ involves a (generally) nonlinear least-squares problem, which can be solved using gradient descent strategies. Note that, in case the impulse response g_θ is linearly parametrized (e.g., it is an FIR system or orthonormal basis functions are used [26]), then the update of θ is also available in closed-form.

Example 4.1 *Assume that the linear parametrization* $g_\theta = L\theta$, $L \in \mathbb{R}^{N\times n_\theta}$, *is used, then* $\hat{\theta}^{(k+1)} = \left(L^\top \hat{A}^{(k)} L\right)^{-1} L^\top \hat{b}^{(k)}$.

The proposed method for module identification is summarized in Algorithm 1.

Algorithm 1 Network empirical Bayes.
Initialization: Find an initial estimate of $\hat{\eta}^{(0)}$*, set* $k = 0$.

(1) Compute $\hat{s}_{11}^{(k)}$ *and* $\hat{P}^{(k)}$ *from* (12).
(2) Update the kernel hyperparameters using (16).
(3) Update the vector θ *by solving* (19).
(4) Update the noise variances using (20).
(5) Check if the algorithm has converged. If not, set $k = k+1$ *and go back to step 1.*

The method can be initialized in several ways. One option is to first estimate $\hat{S}_{11}(q)$ by an empirical Bayes method using only r_1 and $\tilde{w}_1$. Then, $\hat{w}_1$ is constructed from (6), using the obtained $\hat{S}_{11}(q)$. Finally, G is estimated using the prediction error method, using $\hat{w}_1$ as input and $\tilde{w}_2$ as output.

5 Including additional sensors

As reference signals can be added with little effort, a natural question is if also output measurements “downstream” of the module of interest can be added with little effort. In Example 2.1 the measurement w_4 is such a measurement that, with the same strategy as before, can be expressed as

$$w_4(t) = G_{43}(q)w_3(t) + r_4(t)\,. \tag{21}$$

Using this measurement for the purpose of identification would require the identification of $G_{43}(q)$ in addition to the previously considered modules. The signal $w_4(t)$ contains information about $w_3(t)$, and thus information about the module of interest. The price we have to pay for this information is the additional parameters to estimate and, as we will see, another layer of complexity. It is therefore advantageous to include the additional sensor when it is of high quality, i.e., subject to noise with low variance.

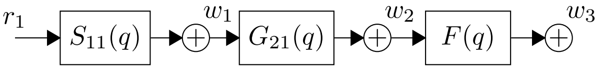

Fig. 4. Basic network of 1 reference signal and 3 internal variables.

To extend the previous framework to include additional measurements after the module of interest, let us consider the case where we would like to include only one additional measurement, in this context denoted by $\tilde{w}_3(t)$; the generalization to more sensors is straightforward but notationally heavy. Let the path linking the target module to the additional sensor be denoted by $F(q)$, with impulse response f. Furthermore, let us for simplicity consider the one-reference-signal-one-input case again, i.e., (10). The setting we consider has been illustrated in Figure 4. We model also $F(q)$ using a Bayesian framework by interpreting f as a zero-mean Gaussian stochastic vector, i.e., $p(f; \lambda_f, K_{\beta_f}) \sim \mathcal{N}(0, \lambda_f K_{\beta_f})$, where again K_{β_f} is the first-order stable spline kernel introduced in Section 3.1. We introduce the following variables

$$\sigma = \begin{bmatrix}\sigma_1^2 & \sigma_2^2 & \sigma_3^2\end{bmatrix}\,,\; z = \begin{bmatrix}\tilde{w}_1^\top & \tilde{w}_2^\top & \tilde{w}_3^\top\end{bmatrix}^\top\,,\; z_f = \tilde{w}_3\,. \tag{22}$$

For given vales of θ, s_{11} and f, we construct

$$W_s = \begin{bmatrix}R^\top & r^\top G_\theta^\top & R^\top G_\theta^\top F^\top\end{bmatrix}^\top\,, \tag{23}$$

$$W_f = \mathcal{T}_N\{G_\theta R s_{11}\}\,, \quad \Sigma = \operatorname{diag}\{\sigma\} \otimes I_N\,. \tag{24}$$

Notice that the last internal variable w_3 can be expressed as $w_3 = G_\theta R v$, where $v := F s_{11}$.

The key difficulty in this setup is that the description of the measurements and the system description with both s_{11} and f no longer admit a jointly Gaussian probabilistic model, because the above-defined v is the result of the convolution of two Gaussian vectors. In fact, a closed-form expression is not available. This fact has a detrimental effect in our empirical Bayes approach, because the marginal likelihood estimator of $\eta = [\sigma\ \lambda_s\ \beta_s\ \lambda_f\ \beta_f\ \theta]$, where λ_s, β_s are the hyperparameters of the prior of s_{11}, that is

$$\hat{\eta} = \arg\max_{\eta} \int p(z, s_{11}, f; \eta)\, \mathrm{d}s_{11}\, \mathrm{d}f, \tag{25}$$

does not admit an analytical expression, since the integral (25) is intractable. To treat this problem, again we resort to the ECM scheme introduced in Section 3. In this case, while the M-Step remains substantially unchanged, the E-step requires to compute

$$Q^{(k)}(\eta) := \mathbb{E}\left[\log p(z,\, s_{11},\, f;\, \eta)\right] = \int \log p(z,\, s_{11},\, f;\, \eta) p(s_{11}, f | z,\, \hat{\eta}^{(k)})\, \mathrm{d}s_{11}\, \mathrm{d}f. \tag{26}$$

As can be seen, this integral does not admit an analytical solution, because the posterior distribution $p(s_{11}, f|z, \hat{\eta}^{(k)})$ is non-Gaussian (it does not have an analytical form, in fact). However, using Monte Carlo techniques we can compute an approximation of the integral by sampling from the joint posterior density $p(s_{11}, f|z;\eta)$ (also called a target distribution). Direct sampling from the target distribution can be hard, because, as pointed out before, it does not admit a closed-form expression. If it is easy to draw samples from the conditional probability distributions, samples of the target distribution can be easily drawn using the Gibbs sampler. In Gibbs sampling, each conditional is considered the state of a Markov chain; by iteratively drawing samples from the conditionals, the Markov chain will converge to its stationary distribution, which corresponds to the target distribution. In our problem, the conditionals of the target distribution are as follows

- $p(s_{11}|f, z;\eta)$. Using W_s defined in (23), we write the linear model
$$z = W_s s_{11} + e, \tag{27}$$
where $e = [e_1^\top\, e_2^\top\, e_3^\top]^\top$. Then, given f, the vectors s_{11} and z are jointly Gaussian, so that $p(s_{11}|f, z;\eta) \sim \mathcal{N}(m_s, P_s)$, with
$$P_s = \left(W_s^\top \Sigma^{-1} W_s + (\lambda_s K_{\beta_s})^{-1}\right)^{-1}, \; m_s = P_s W_s^\top \Sigma^{-1} z\,.$$
- $p(f|s_{11}, z;\eta)$. Given s_{11} and r, all sensors but the last becomes redundant. Using (24) we write the linear model $z_f = W_f f + e_3$, which shows that $p(f|s_{11}, z;\eta) \sim \mathcal{N}(m_f, P_f)$, with
$$P_f = \left(\frac{W_f^\top W_f}{\sigma_3^2} + (\lambda_f K_{\beta_f})^{-1}\right)^{-1}, \; m_f = P_f \frac{W_f^\top}{\sigma_3^2} z_f\,.$$

The following algorithm summarizes the Gibbs sampler used for dynamic network identification.

Algorithm 2 Dynamic network Gibbs sampler. *Initialization: compute initial value of s_{11}^0 and f^0. For $k = 1$ to $M + M_0$:*

```
(1) Draw the sample $s_{11}^k$ from $p(s_{11}|f^{k-1}, z;\eta)$;
(2) Draw the sample $f^k$ from $p(f|s_{11}^k, z;\eta)$;
```

In this algorithm, M_0 is the number of initial samples that are discarded, which is also known as the *burn-in* phase [27]. These samples are discarded since the Markov chain needs a certain number of samples to converge to its stationary distribution.

We now discuss the computation of the E-step and the CM-steps using the Gibbs sampler scheme introduced above.

Proposition 5.1 *Introduce the mean and covariance quantities*
$$s_s^M = M^{-1}\sum_{k=M_0+1}^{M_0+M} s_{11}^k\,, P_s^M = M^{-1}\sum_{k=M_0+1}^{M_0+M}(s_{11}^k - s_s^M)(\cdot)^\top, \tag{28}$$
where $(\cdot)$ denotes the previous argument and f_s^M, P_f^M, v_s^M and P_v^M are defined similarly and where s_{11}^k, f^k and $v^k = s_{11}^k * f^k$ are samples drawn using Algorithm 2.

Define
$$\tilde{Q}_s(\lambda,\beta,x,X) := \log\det\{\lambda K_\beta\} + \mathrm{Tr}\{(\lambda K_\beta)^{-1}(xx^\top + X)\}$$
$$\tilde{Q}_z(\sigma^2,z,x,X) := N\log\sigma^2 + \frac{\|z - Rx\|_2^2 + \mathrm{Tr}\{RXR^\top\}}{\sigma^2}$$
$$\tilde{Q}_f(\sigma^2,z,\theta,x,X) := N\log\sigma^2 + \frac{1}{\sigma^2}\|z - G_\theta Rx\|_2^2 + \frac{1}{\sigma^2}\mathrm{Tr}\{G_\theta RXR^\top G_\theta^\top\}\,.$$

Then
$$\begin{aligned}-2Q^{(k)}(\eta) &= \lim_{M\to\infty} \tilde{Q}_s(\lambda_s,\beta_s,s_s^M,P_s^M) \\ &+\tilde{Q}_s(\lambda_f,\beta_f,f_s^M,P_f^M) + \tilde{Q}_z(\sigma_1^2,\tilde{w}_1,s_s^M,P_s^M) \\ &+\tilde{Q}_f(\sigma_2^2,\tilde{w}_2,\theta,s_s^M,P_s^M) + \tilde{Q}_f(\sigma_3^2,\tilde{w}_3,\theta,v_s^M,P_v^M)\,.\end{aligned} \tag{29}$$

The CM-steps are now very similar to the previous case and are reported in the following Proposition (the proof follows by similar reasoning as in the proof of Proposition 4.2).

Proposition 5.2 *Let $\hat{\eta}^{(k)}$ be the parameter estimate obtained at the k:th iteration. Define $S_s^M = s_s^M(s_s^M)^\top + P_s^M$, $S_v^M = v_s^M(v_s^M)^\top + P_v^M$,*
$$\hat{A}_s = D^\top(RS_s^M R^\top \otimes I_N)D\,, \quad \hat{b}_s = \mathcal{T}_N\{Rs_s^M\}^\top \tilde{w}_2\,,$$
$$\hat{A}_v = D^\top(RS_v^M R^\top \otimes I_N)D\,, \quad \hat{b}_v = \mathcal{T}_N\{Rv_s^M\}^\top \tilde{w}_3\,.$$

Then the updated parameter vector $\hat{\eta}^{(k+1)}$ is obtained as follows
$$\hat{\theta}^{(k+1)} = \arg\min_\theta \frac{g_\theta^\top \hat{A}_s g_\theta}{\sigma_2^2} + \frac{g_\theta^\top \hat{A}_v g_\theta}{\sigma_3^2} - 2\frac{\hat{b}_s^\top g_\theta}{\sigma_2^2} - 2\frac{\hat{b}_v^\top g_\theta}{\sigma_3^2}\,.$$

The closed form updates of the noise variances are
$$\hat{\sigma}_1^{2(k+1)} = \frac{1}{N}\left(\|\tilde{w}_1 - Rs_s^M\|_2^2 + \mathrm{Tr}\{RP_s^M R^\top\}\right),$$
$$\hat{\sigma}_2^{2(k+1)} = \frac{1}{N}\Big(\|\tilde{w}_2 - G_{\hat{\theta}^{(k+1)}}Rs_s^M\|_2^2 + \mathrm{Tr}\Big\{G_{\hat{\theta}^{(k+1)}}RP_s^M R^\top G_{\hat{\theta}^{(k+1)}}^\top\Big\}\Big),$$
$$\hat{\sigma}_3^{2(k+1)} = \frac{1}{N}\Big(\|\tilde{w}_3 - G_{\hat{\theta}^{(k+1)}}Rv_s^M\|_2^2$$

$$+ \operatorname{Tr}\Big\{G_{\hat{\theta}^{(k+1)}} R P_v^M R^\top G_{\hat{\theta}^{(k+1)}}^\top\Big\}\Big) \,.$$

The kernel hyperparameters are updated through (16) *for both* s_{11} *and* f*.*

The proposed method for module identification is summarized in Algorithm 3.

Algorithm 3 *Network empirical Bayes extension. Initialization: Find an initial estimate of* $\hat{\eta}^{(0)}$*, set* $k = 0$*.*

```
(1) Compute the quantities (28) using Algorithm 2.
(2) Update the kernel hyperparameters using (16).
(3) Update the vector $\theta$ and the noise variances according to Proposition 5.2.
(4) Check if the algorithm has converged. If not, set $k = k + 1$ and go back to step 1.
```

As can be seen, the main difference with Algorithm 3 compared to Algorithm 1 is that Step 2 of the algorithm requires a heavier computational burden because of the Gibbs sampling. Essentially, a posterior mean and covariance matrix is computed for each sample drawn in the Gibbs sampler, whereas they are computed only once in the previous algorithm. Nevertheless, as will be seen in the next section, this pays off in terms of performance in identifying the target module.

Remark 1 *The method presented in this section postulates Gaussian models for the sensitivity path s and the for the path to the additional sensor f, while the target module* G_θ *is modeled using a parametric approach. It may be tempting, especially in the multiple-sensor case presented in this section, to model also the target module using Gaussian processes. However, there are two main reasons for us not to doing so. First, our concept of dynamic network is that it is the result of the composition of a large number of* simple *modules, i.e., modules that can be modeled using few parameters (e.g., a DC motor having only one mode). Therefore, the use of parametric models seem more appropriate in this context. Second, in the case where only one sensor downstream is used for module identification (i.e., the case of Section 3), using Gaussian processes to model the target module would require to employ the Gibbs sampler also in that case.*

6 Numerical experiments

In this section, we present the result from a Monte Carlo simulation to illustrate the performance of the proposed method, which we abbreviate as *Network Empirical Bayes* (NEB) and its extension NEBX outlined in Section 5. We compare the proposed methods with SMPE (see Section 2) and the two-stage method on data simulated from the network described in Example 2.1. The reference signals used are zero-mean unit-variance Gaussian white noise. The noise signals e_k are zero-mean Gaussian white noise with variances such that noise to signal ratios $\mathbf{Var}\{w\}_k / \mathbf{Var}\{e\}_k$ are constant. The setting of the compared methods are provided in some more details below, where the model order of the plant $G(q)$ is known for both the SMPE method and the proposed NEB method.

2ST: The two-stage method as presented in Section 2.

SMPE: The method is initialized by the two-stage method. Then, the cost function (9), with a slight modification, is minimized. The modification of the cost function comes from that, as mentioned before, the SMPE method assumes that the noise variances are known. To make the comparison fair, also the noise variances need to be estimated. By maximum likelihood arguments, the logarithm of the determinant of the complete noise covariance matrix is added to the cost function (9) and the noise variances are included in θ, the vector of parameters to estimate. The tolerance is set to $\|\hat{\theta}^{(k+1)} - \hat{\theta}^{(k)}\| / \|\hat{\theta}^{(k)}\| < 10^{-6}$.

NEB: The method is initialized by the two-stage method. First, $\hat{S}(q)$ is estimated by least-squares. Second, G is estimated using MORSM [28] from the simulated signal $\hat{w}$ obtained from (6) and $\tilde{w}_j$. MORSM is an iterative method that is asymptotically efficient for open loop data. Then, Algorithm 1 is employed with the stopping criterion $\|\hat{\eta}^{(k+1)} - \hat{\eta}^{(k)}\| / \|\hat{\eta}^{(k)}\| < 10^{-6}$.

NEBX: The method is initialized by NEB. f^0 is obtained by an empirical Bayes method using simulated input and measured output of f. Then, Algorithm 3 is employed with the stopping criterion $\|\hat{\eta}^{(k+1)} - \hat{\eta}^{(k)}\| / \|\hat{\eta}^{(k)}\| < 10^{-6}$, or a maximum of 5 iterations.

The simulations were run in Julia, a high-level, high-performance dynamic programming language for technical computing [29] (the code is available [1]).

The Monte Carlo simulation compares the NEB method and NEBX with the 2ST and SMPE method on data from the network of Example 2.1, illustrated in Figure 1, where each of the modules are of second order, i.e.,

$$G_{ij}(q) = \frac{b_1 q^{-1} + b_2 q^{-2}}{1 + a_1 q^{-1} + a_2 q^{-2}},$$

for a set of parameters that were chosen such that all modules are stable and $\{S_{12}(q), S_{24}(q), S_{22}(q), S_{24}(q)\}$ are stable and can be well approximated with 60 impulse response coefficients. Two reference signals, $r_2(t)$ and $r_4(t)$ are available and $N = 150$ data samples are used with the goal to estimate $G_{31}(q)$ and G_{32}. In total 6 transfer functions are estimated, $\{S_{12}(q), S_{24}(q), S_{22}(q), S_{24}(q), G_{31}(q)$ and

[1] `https://github.com/neveritt/NEB-Example.jl`

$G_{32}(q)\}$, where $\{S_{12}(q), S_{24}(q), S_{22}(q), S_{24}(q)\}$ are each parametrized by $n = 60$ impulse response coefficients in all methods. For NEBX also $G_{43}(q)$ is estimated by $n = 60$ impulse response coefficients. The noise to signal ratio at each measurement is set to $\mathbf{Var}\{e\}_k / \mathbf{Var}\{w\}_k = 0.1$ and the additional measurement used in NEBX has a lower noise to signal ratio of $\mathbf{Var}\{e\}_4 / \mathbf{Var}\{w\}_4 = 0.01$.

The fits of the impulse responses of G_{31} for the experiment are shown as a boxplot in Figure 5. Comparing the fits obtained, the proposed NEB and NEBX methods are competitive with the SMPE method for this network. NEBX outperformed NEB in this simulation. However, NEBX is significantly more computationally expensive than NEB.

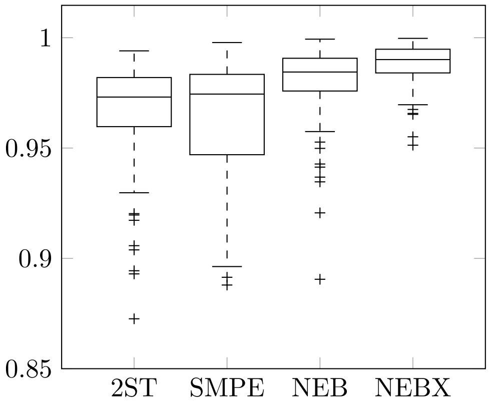

Fig. 5. Box plot of the fit of the impulse response of G_{31} obtained from 100 Monte Carlo runs by the methods 2ST, SMPE, NEB and NEBX respectively.

7 Conclusion

In this paper, we have addressed the identification of a module in dynamic networks with known topology. The problem is cast as the identification of a set of systems in series connection. The second system corresponds to the target module, while the first represents the dynamic relation between exogenous signals and the input and the target module. This system is modeled following a Bayesian kernel-based approach, which enables the identification of the target module using empirical Bayes arguments. In particular, the target module is estimated using a marginal likelihood criterion, whose solution is obtained by a novel iterative scheme designed through the ECM algorithm. The method is extended to incorporate measurements downstream of the target module, which numerical experiments suggest increases performance. The main limitation with the proposed algorithms is the restrictive assumptions on the noise. Generalizing the noise assumptions would improve the applicability of the method and is considered for future work.

A Appendix

Proof of Lemma 4.1: From Bayes' rule it follows that $\log p(z, s_{11};\, \hat{\eta}^{(k)}) = \log p(z|s_{11},\ \hat{\eta}^{(k)}) + \log p(s_{11};\, \hat{\eta}^{(k)})$, with (neglecting constant terms)

$$\log p(z|s_{11},\ \eta) \propto -\frac{1}{2}\log\det\{\Sigma_e\} - \frac{1}{2}\|z - W_\theta s_{11}\|^2_{\Sigma_e^{-1}}$$
$$\log p(s_{11};\, \eta) \propto -\frac{1}{2}\log\det\{\lambda K_\beta\} - \frac{1}{2}s_{11}^\top(\lambda K_\beta)^{-1}s_{11}\,.$$

Now we have to take the expectation w.r.t. the posterior $p(s_{11}|\tilde{w}_2;\, \hat{\eta}^{(k)})$. Developing the second term in the first equation above and recalling that $\mathbf{E}_{p(s_{11}|\tilde{w}_2;\, \hat{\eta}^{(k)})}[s_{11}^\top A s_{11}] = \mathrm{Tr}\Big\{A\hat{S}_{11}^{(k)}\Big\}$, the statement of the lemma readily follows.

Proof of Proposition 4.2: In $Q_0^{(k)}(\sigma_1^2,\, \sigma_2^2,\, \hat{\theta}^{(k+1)})$, fix Σ_e to the value $\hat{\Sigma}_e^{(k)}$ (computed inserting $\hat{\sigma}_1^{2(k)}$ and $\hat{\sigma}_2^{2(k)}$). We obtain the θ-dependent terms (after multiplying by a factor -2),

$$\begin{aligned}
-2z^\top\left(\hat{\Sigma}_e^{(k)}\right)^{-1}W_\theta\hat{s}_{11}^{(k)} &= -\frac{2}{\hat{\sigma}_2^{2(k)}}y^\top G_\theta R_1\hat{s}_{11}^{(k)} + k_1\\
&= -\frac{2}{\hat{\sigma}_2^{2(k)}}y^\top \mathcal{T}_N\Big\{R_1\hat{s}_{11}^{(k)}\Big\}g_\theta + k_1\\
\mathrm{Tr}\Big\{W_\theta^\top\Big(\Sigma_e^{(k)}\Big)^{-1}W_\theta\hat{S}_{11}^{(k)}\Big\} &= \frac{\mathrm{Tr}\Big\{G_\theta R_1\hat{S}_{11}^{(k)}R_1^\top G_\theta^\top\Big\}}{\hat{\sigma}_2^{2(k)}} + k_2\\
&= \frac{1}{\hat{\sigma}_2^{2(k)}}\,\mathrm{vec}\{G_\theta\}^\top(R_1\hat{S}_{11}^{(k)}R_1^\top \otimes I_N)\,\mathrm{vec}\{G_\theta\} + k_2\\
&= \frac{1}{\hat{\sigma}_2^{2(k)}}g_\theta^\top D^\top(R_1\hat{S}_{11}^{(k)}R_1^\top \otimes I_N)Dg_\theta + k_2,
\end{aligned}$$

where k_1 and k_2 contain terms independent of θ. Recalling the definitions of $\hat{A}^{(k)}$ and $\hat{b}^{(k)}$, (19) readily follows. Now, let θ be fixed at the value $\hat{\theta}^{(k+1)}$. The function (16) can be rewritten as (after multiplying by a factor -2).

$$\begin{aligned}
Q_0^{(k)}(\sigma_1^2,\, \sigma_2^2,\, \hat{\theta}^{(k+1)}) &= N(\log\sigma_1^2 + \log\sigma_2^2) + \frac{\|\tilde{w}_1\|_2^2}{\sigma_1^2}\\
&+\frac{1}{\sigma_1^2}\,\mathrm{Tr}\Big\{R_1^\top R_1\hat{S}_{11}^{(k)}\Big\} + \frac{\|\tilde{w}_2\|_2^2}{\sigma_2^2} - \frac{2\tilde{w}_2^\top}{\sigma_2^2}G_{\hat{\theta}^{(k+1)}}R_1\hat{s}_{11}^{(k)}\\
&-\frac{2\tilde{w}_1^\top}{\sigma_1^2}R_1\hat{s}_{11}^{(k)} + \frac{1}{\sigma_2^2}\,\mathrm{Tr}\Big\{R_1^\top G_{\hat{\theta}^{(k+1)}}^\top G_{\hat{\theta}^{(k+1)}}R_1\hat{S}_{11}^{(k)}\Big\}
\end{aligned}$$

The results (20) follow by minimizing this expression with respect to σ_1^2 and σ_2^2.

Proof of Proposition 5.1: Using Bayes' rule we can decompose the complete likelihood as $\log p(z, s_{11}, f;\eta) =$

$\log p(z|s_{11}, f;\eta) + \log p(s_{11};\eta) + \log p(f;\eta)$, and we will analyze each term in turn. First, note that

$$\begin{aligned}-2\log p(s_{11}|\eta) &= \log\det\{\lambda_s K_{\beta_s}\} + s_{11}^\top(\lambda_s K_{\beta_s})^{-1}s_{11}\\ &= \log\det\{\lambda_s K_{\beta_s}\} + \mathrm{Tr}\big\{(\lambda_s K_{\beta_s})^{-1}s_{11}s_{11}^\top\big\}\end{aligned}$$

Replacing $s_{11}s_{11}^\top$ with its sample estimate yields the first term in (29). Similarly, $-2\log p(f|\eta) = \log\det\big\{\lambda_f K_{\beta_f}\big\} + \mathrm{Tr}\big\{(\lambda_f K_{\beta_f})^{-1}ff^\top\big\}$. Replacing $ff^\top$ with its sample estimate yields the second term in (29). Finally, $-2\log p(z|t, s_{11};\eta) = \log\det\{\varSigma\} + (z-\hat{z})^\top\varSigma^{-1}(z-\hat{z})$, with $\hat{z} := [(Rs_{11})^\top\ (G_\theta Rs_{11})^\top\ (G_\theta Rv)^\top]^\top$. The first term is N times the sum of the logarithms of the noise variances squared. The second term decomposes into a sum of the (weighted) error of each signal. Then, the first weighted error is given by $\sigma_1^2\|\tilde{w}_1 - Rs_{11}\|_2^2 = \|\tilde{w}_1\|_2^2 - 2\tilde{w}_1^\top Rs_{11} + \mathrm{Tr}\big\{Rs_{11}s_{11}^\top R^\top\big\}$. Replacing s_{11} and $s_{11}s_{11}^\top$ with their respective estimates gives the third term in (29), with the corresponding noise variance term added. Similar calculations on the remaining two weighted errors gives the last two terms in (29). This concludes the proof.

References

[1] D. Materassi and G. Innocenti, "Topological identification in networks of dynamical systems," *IEEE Transactions on Automatic Control*, vol. 55, no. 8, pp. 1860–1871, 2010.

[2] P. M. J. Van den Hof, A. 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.

[3] H. Hjalmarsson, "System identification of complex and structured systems," *European J. of Control*, vol. 15, no. 3-4, pp. 275–310, 2009.

[4] D. Materassi and M. V. Salapaka, "On the problem of reconstructing an unknown topology via locality properties of the Wiener filter," *IEEE Transactions on Automatic Control*, vol. 57, no. 7, pp. 1765–1777, 2012.

[5] A. Chiuso and G. Pillonetto, "A Bayesian approach to sparse dynamic network identification," *Automatica*, vol. 48, no. 8, pp. 1553–1565, 2012.

[6] A. Dankers, P. M. J. Van den Hof, and P. S. C. Heuberger, "Predictor input selection for direct identification in dynamic networks," in *Proceedings of the 52nd IEEE Annual Conference on Decision and Control*. IEEE, 2013, pp. 4541–4546.

[7] B. Gunes, A. Dankers, and P. M. J. Van den Hof, "A variance reduction technique for identification in dynamic networks," in *Proceedings of the 19th IFAC World Congress*, 2014.

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

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

[10] P. Torres, J. W. van Wingerden, and M. Verhaegen, "Output-error identification of large scale 1D-spatially varying interconnected systems," *IEEE Transactions on Automatic Control*, vol. 60, no. 1, pp. 130–142, 2014.

[11] B. Wahlberg, H. Hjalmarsson, and J. Mårtensson, "Variance results for identification of cascade systems," *Automatica*, vol. 45, no. 6, pp. 1443–1448, 2009.

[12] L. Ljung, *System identification.* Springer, 1998.

[13] 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.

[14] G. Pillonetto and G. De Nicolao, "A new kernel-based approach for linear system identification," *Automatica*, vol. 46, no. 1, pp. 81–93, 2010.

[15] N. Everitt, G. Bottegal, C. R. Rojas, and H. Hjalmarsson, "Variance analysis of linear simo models with spatially correlated noise," *Automatica*, vol. 77, pp. 68–81, 2017.

[16] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. of the royal statistical society. Series B (methodological)*, pp. 1–38, 1977.

[17] X. L. Meng and D. B. Rubin, "Maximum likelihood estimation via the ECM algorithm: A general framework," *Biometrika*, vol. 80, no. 2, pp. 267–278, 1993.

[18] W. Gilks, S. Richardson, and D. Spiegelhalter, *Markov Chain Monte Carlo in Practice*, ser. Chapman & Hall/CRC Interdisciplinary Statistics. Taylor & Francis, 1995.

[19] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Transactions on pattern analysis and machine intelligence*, no. 6, pp. 721–741, 1984.

[20] N. Everitt, G. Bottegal, C. R. Rojas, and H. Hjalmarsson, "Identification of modules in dynamic networks: An empirical bayes approach," in *Proceedings of the 55th IEEE Annual Conference on Decision and Control*. IEEE, 2016, pp. 4612–4617.

[21] A. Dankers and P. M. J. Van den Hof, "Non-parametric identification in dynamic networks," in *Proceedings of the 54th IEEE Conference on Decision and Control*, 2015, pp. 3487–3492.

[22] N. Everitt, "Module identification in dynamic networks: parametric and empirical bayes methods," Ph.D. dissertation, KTH Royal Institute of Technology, 2017.

[23] C. Rasmussen and C. Williams, *Gaussian Processes for Machine Learning.* The MIT Press, 2006.

[24] B. Anderson and J. Moore, *Optimal Filtering.* Englewood Cliffs, N.J., USA: Prentice-Hall, 1979.

[25] 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.

[26] B. Wahlberg, "System identification using Laguerre models," *IEEE Transactions on Automatic Control*, vol. 36, pp. 551–562, 1991.

[27] S. Meyn and R. L. Tweedie, *Markov chains and stochastic stability; 2nd ed.*, ser. Cambridge Mathematical Library. Leiden: Cambridge Univ. Press, 2009.

[28] N. Everitt, M. Galrinho, and H. Hjalmarsson, "Optimal model order reduction with the steiglitz-mcbride method," *submitted to Automatica (arXiv:1610.08534)*, 2016.

[29] J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah, "Julia: A fresh approach to numerical computing," *SIAM Review*, vol. 59, no. 1, pp. 65–98, jan 2017.