Automatica 98 (2018) 256–268

Contents lists available at ScienceDirect

Automatica

journal homepage: www.elsevier.com/locate/automatica

Prediction error identification of linear dynamic networks with rank-reduced noise☆

Harm H.M. Weerts [a], Paul M.J. Van den Hof [a,*], Arne G. Dankers [b]

[a] Control Systems Group, Department of Electrical Engineering, Eindhoven University of Technology, The Netherlands
[b] Department of Electrical Engineering, University of Calgary, Canada

ARTICLE INFO

Article history:
Received 16 November 2017
Received in revised form 4 April 2018
Accepted 24 July 2018

Keywords:
System identification
Dynamic networks
Maximum likelihood
Rank-reduced noise
Consistency
Variance
Cramér–Rao lower bound

ABSTRACT

Dynamic networks are interconnected dynamic systems with measured node signals and dynamic modules reflecting the links between the nodes. We address the problem of identifying a dynamic network with known topology, on the basis of measured signals, for the situation of additive process noise on the node signals that is spatially correlated and that is allowed to have a spectral density that is singular. A prediction error approach is followed in which all node signals in the network are jointly predicted. The resulting joint-direct identification method, generalizes the classical direct method for closed-loop identification to handle situations of mutually correlated noise on inputs and outputs. When applied to general dynamic networks with rank-reduced noise, it appears that the natural identification criterion becomes a weighted LS criterion that is subject to a constraint. This constrained criterion is shown to lead to maximum likelihood estimates of the dynamic network and therefore to minimum variance properties, reaching the Cramér–Rao lower bound in the case of Gaussian noise. In order to reduce technical complexity, the analysis is restricted to dynamic networks with strictly proper modules.

1. Introduction

It is becoming more common to model complex dynamic systems as networks of interconnected dynamic modules, or *dynamic networks*. Data-driven modeling, or *identification*, of modules in these dynamic networks is then a natural problem to address. Applications range over many fields, for example identification of dynamics that connect different (MPC) control loops in industrial process control (Gudi & Rawlings, 2006; Van den Hof, Dankers, & Weerts, 2018), identification of biochemical networks (Yuan, Stan, Warnick, & Gonçalves, 2011), modeling of the dynamic behavior of a ship as a dynamic network (Linder, 2017), and modeling of stock prices in financial markets as a dynamic network (Materassi & Innocenti, 2010).

Various approaches have been developed for identification of dynamic networks, roughly divided into three categories. The first approach considers the identification of a single module in the dynamic network in the situation that the interconnection structure, or topology, of the network is known. The second approach focusses on identification of the full network dynamics for a given topology, and the last category deals with the identification of the topology (and dynamics) of the network. For identification of single modules, authors have used e.g. Wiener filters (Materassi & Salapaka, 2012), while the estimation of parametric transfer functions in a prediction error setting has been addressed in Dankers, Van den Hof, Bombois, and Heuberger (2015), Dankers, Van den Hof, Heuberger, and Bombois (2016), Gevers and Bazanella (2015), Linder and Enqvist (2017) and Van den Hof, Dankers, Heuberger, and Bombois (2013). Identification of the full network dynamics has been considered by modeling the network as a state–space system (Haber & Verhaegen, 2014), or as a network of transfer function modules (Weerts, Van den Hof, & Dankers, 2016b). Identifiability properties related to this problem have been addressed in Adebayo et al. (2012), Gevers, Bazanella, and Parraga (2017), Gonçalves and Warnick (2008), Weerts, Dankers, and Van den Hof (2015) and Weerts, Van den Hof, and Dankers (2018). Some different methods for topology detection can be found in literature, for example following a Bayesian approach (Chiuso & Pillonetto, 2012), a compressed sensing approach

☆ This project 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). The material in this paper was partially presented at the 20th World Congress of the International Federation of Automatic Control, July 9–14, 2017, Toulouse, France. This paper was recommended for publication in revised form by Associate Editor Martin Enqvist under the direction of Editor Torsten Söderström.

* Corresponding author.
E-mail addresses: h.h.m.weerts@tue.nl (H.H.M. Weerts), p.m.j.vandenhof@tue.nl (P.M.J. Van den Hof), adankers@hifieng.com (A.G. Dankers).

https://doi.org/10.1016/j.automatica.2018.09.033
0005-1098/

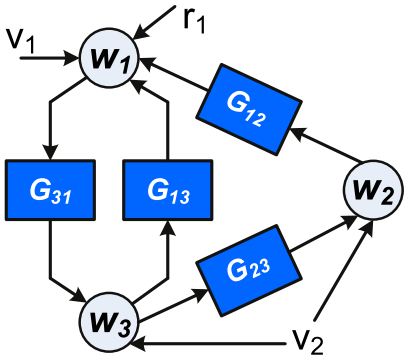

Fig. 1. Example of a network with rank-reduced noise. Node signals are w_i, being the outputs of the (circular) summation points, interconnected by modules G_{ij} and perturbed by non-measured disturbance signals v_i. Signals r_i are excitation signals available to the user.

(Hayden, Chang, Gonçalves, & Tomlin, 2016), or through one-step ahead prediction using Wiener filters (Materassi & Salapaka, 2012).

In this paper we consider networks that consist of measured node signals, which are interconnected by linear dynamic modules, as depicted in Fig. 1, and in line with the setup as defined in Van den Hof et al. (2013). We will address the problem of identifying, on the basis of measured node signals, the dynamics of all modules in a network, of which the topology is known, and where conditions on the disturbance signals v_i in the network are more general than typically considered. While in the current literature it is usually assumed that every node signal in the network has a non-zero process noise v_i that is uncorrelated to all other noises, i.e. for the vector noise process v it holds that $\Phi_v(\omega)$ is diagonal, we will address two steps of generalization:

- We will allow noise signals on the different node signals to be spatially correlated, i.e. $\Phi_v(\omega)$ is not necessarily diagonal, and
- We will allow $\Phi_v(\omega)$ to be singular, implying that node signals can be noise-free, or that disturbances are exactly related with each other through a linear filter.

Concerning the first step, this situation includes the handling of confounding variables, i.e. unmeasured variables that affect both inputs and outputs of an estimation problem. This notion is widely used in statistical estimation problems in networks and is also used in network identification problems, Dankers et al. (2016). The relation between confounding variables and correlated disturbances has been explained in Van den Hof, Dankers, and Weerts (2017).

Concerning the second step, note that modules in a network can also be implemented controllers, and controller outputs can be noise-free, as e.g. typically considered in a classical closed-loop identification problem (Ljung, 1999). In this case there is no process noise on a particular node signal. Alternatively, strong correlations between disturbance signals can occur e.g. if the network is a spatially distributed system affected by global disturbances, like a wind gust affecting wind turbines in a wind park. A deterministic relation between disturbance signals (like e.g. a delay) will cause the full disturbance spectrum to lose its full rank. A situation of loss of full rank is depicted in Fig. 1 where the process noises on nodes 2 and 3 are the same (perfect correlation). When identifying the full network dynamics, aiming not only at consistency of the module estimates, but also at minimum variance results, correlated disturbances will prevent the identification problem to be decomposable into separate multi-input single-output problems. The fact that the noise process is allowed to be rank-reduced causes some fundamental issues that need to be addressed in the prediction error identification setting.

Identification in the situation of rank-reduced noise is a topic that has not been widely addressed in the prediction error identification literature. Dynamic factor models have been developed in Deistler, Scherrer, and Anderson (2015) and Felsenstein (2014) to deal with rank-reduced noise. Maximum likelihood estimates with rank-reduced noise have been obtained for vector autoregressive systems (Kölbl, 2015) and linear regression (Srivastava & von Rosen, 2002). In a prediction error setting, the property of *network identifiability* has been defined in Weerts et al. (2015) and Weerts et al. (2018), covering also the situation of rank-reduced noise, while predictor models have been analyzed for the situation of noise-free nodes in Weerts, Van den Hof, and Dankers (2016a). In Weerts, Van den Hof, and Dankers (2017) a first analysis of consistent estimation of network models has been presented for the reduced-rank noise case, leading to the use of weighted and constrained least-squares identification criteria. This was a further extension of the preliminary work of Van den Hof, Weerts, and Dankers (2017b) where an open-loop one-input two-output situation with rank-reduced output noise was considered.

In this paper we are going beyond the consistency question, by including an analysis of the asymptotic variance of the prediction error method, and by developing the maximum likelihood estimator and the Cramér–Rao lower bound on the variance, for the situation of correlated and rank-reduced noise, while addressing networks with strictly proper modules. This paper builds on and further extends the preliminary results of Weerts et al. (2017).

First a definition of the dynamic network setup and the rank-reduced noise process is given in Section 2. Then, in Section 3, the prediction error identification setup is presented and a least squares identification criterion is shown to provide consistent estimates. In Section 4 the dependencies in the noise process are explicitly used to construct a constrained least squares identification criterion that is shown to lead to a maximum likelihood estimate under some conditions. An analysis of the asymptotic variance of the estimates is made in Section 5, where the variance expressions are related to the Cramér–Rao lower bound. Finally in Section 6 the theoretical results are illustrated in a numerical simulation example.

2. Dynamic network definition

Following the basic setup of Van den Hof et al. (2013), a dynamic network is built up out of L scalar *internal variables* or *nodes* w_j, $j = 1, \ldots, L$, and K *external variables* r_k, $k = 1, \ldots, K$. Each internal variable is described as:

$$w_j(t) = \sum_{\substack{l=1 \\ l \neq j}}^{L} G_{jl}^0(q) w_l(t) + \sum_{k=1}^{K} R_{jk}^0(q) r_k(t) + v_j(t) \tag{1}$$

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

- G_{jl}^0 are strictly proper rational transfer functions, and the single transfers G_{jl}^0 are referred to as *modules* in the network.
- r_k are *external variables* that can directly be manipulated by the user, and R_{jk}^0 are proper rational transfer functions;
- v_j is *process noise*, where the vector process $v = [v_1 \cdots v_L]^T$ is modeled as a stationary stochastic process with rational spectral density, such that there exists a p-dimensional white noise process $e := [e_1 \cdots e_p]^T$, $p \leq L$, with covariance matrix $\Lambda^0 > 0$ such that

  $$v(t) = H^0(q) e(t),$$

with $H^0(q)$ a proper rational transfer function.

When combining the L node signals we arrive at the full network expression

$$\begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_L \end{bmatrix} = \begin{bmatrix} 0 & G^0_{12} & \cdots & G^0_{1L} \\ G^0_{21} & 0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & G^0_{L-1\,L} \\ G^0_{L1} & \cdots & G^0_{L\,L-1} & 0 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_L \end{bmatrix} + R^0(q) \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_K \end{bmatrix} + H^0(q) \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_p \end{bmatrix}$$

Using obvious notation this results in the matrix equation:

$$w = G^0 w + R^0 r + H^0 e. \tag{2}$$

The network transfer function that maps the external signals r and e to the node signals w is denoted by

$$T^0(q) := \begin{bmatrix} T^0_{wr}(q) & T^0_{we}(q) \end{bmatrix}, \tag{3}$$

where

$$T^0_{wr}(q) := \left(I - G^0(q)\right)^{-1} R^0(q), \tag{4}$$

$$T^0_{we}(q) := \left(I - G^0(q)\right)^{-1} H^0(q). \tag{5}$$

Note that by choosing $G^0 = 0$ the network reduces to a multivariable open-loop system with inputs r and outputs w.

The noise component $\bar{v}(t)$ is defined according to $\bar{v}(t) := w(t) - T^0_{wr}(q) r(t)$ and satisfies

$$\bar{v}(t) = T^0_{we}(q) e(t), \tag{6}$$

with power spectral density

$$\Phi_{\bar{v}}(\omega) := T^0_{we}(e^{i\omega}) \Lambda^0 T^{0,T}_{we}(e^{-i\omega}). \tag{7}$$

This power spectral density can be determined using

$$\Phi_{\bar{v}}(\omega) = \Phi_w(\omega) - T^0_{wr}(e^{i\omega}) \Phi_r(\omega) T^{0,T}_{wr}(e^{-i\omega}), \tag{8}$$

where Φ_w and Φ_r are the power spectral densities of w and r respectively.

The noise model H^0 requires some further specification. For $p = L$, referred to as the *full-rank* noise case, H^0 is square, stable, monic and minimum-phase. The situation $p < L$ will be referred to as the *singular* or *rank-reduced* noise case.

For notational simplicity and without loss of generality the following assumption will be made.

Assumption 1. The L node signals w_j, $j = 1, \ldots, L$ are ordered in such a way that $\begin{bmatrix} v_1 \cdots v_p \end{bmatrix}^T$ is a full rank noise process. □

How this assumption can be dealt with in actual identification will be discussed later on in Remark 3.

The ordering of the noise signals gives rise to a representation for H^0 that satisfies

$$H^0(q) = \begin{bmatrix} H^0_a \\ H^0_b \end{bmatrix} \tag{9}$$

with H^0_a a proper rational transfer function which is square, monic, stable and stably invertible. For properties of H^0_b we need the following lemma, which is an adapted version of the spectral factorization theorem (Youla, 1961) that is also used in Weerts et al. (2018).

Lemma 1 (*Factorization of Reduced-rank Spectra*). *Consider an L-dimensional stationary stochastic process x with rational spectral density Φ_x and rank $p < L$, that satisfies the ordering property of Assumption 1. Then*

a. *Φ_x allows a unique spectral factorization*

$$\Phi_x = F \Delta F^*$$

with $F \in \mathbb{R}^{L\times p}(z)$, $F = \begin{bmatrix} F_a \\ F_b \end{bmatrix}$ with F_a square, monic, and F stable and having a stable left inverse $F^\dagger$ that satisfies $F^\dagger F = I_p$, and $\Delta \in \mathbb{R}^{p\times p}$, $\Delta > 0$;

b. *Based on the unique decomposition of Φ_x in (a.), there exists a unique factorization of Φ_x in the structure:*

$$\Phi_x = \breve{F} \breve{\Delta} \breve{F}^*$$

with $\breve{F} \in \mathbb{R}^{L\times L}(z)$ monic, stable with a stable inverse and $\breve{\Delta} \in \mathbb{R}^{L\times L}$, having the particular structure

$$\breve{F} = \begin{bmatrix} F_a & 0 \\ F_b - \Gamma & I \end{bmatrix}, \quad \breve{\Delta} = \begin{bmatrix} I \\ \Gamma \end{bmatrix} \Delta \begin{bmatrix} I \\ \Gamma \end{bmatrix}^T$$

and $\Gamma := \lim_{z\to\infty} F_b(z)$.

Proof. Part (a) is the standard spectral factorization theorem, see Youla (1961). The decomposition in part (b) can be verified by direct computation. Stability of $\breve{F}$ follows from stability of F. Stability of

$$\breve{F}^{-1} = \begin{bmatrix} F_a^{-1} & 0 \\ (F_b - \Gamma) F_a^{-1} & I \end{bmatrix} \tag{10}$$

follows since it contains only stable components. □

From Lemma 1 we know that H^0 is stable and has a stable left inverse $H^\dagger$, satisfying $H^\dagger H^0 = I_p$, the $p \times p$ identity matrix. The feedthrough term of H^0_b will throughout the paper be indicated with Γ^0, i.e. $\Gamma^0 := \lim_{z\to\infty} H^0_b(z)$.

When we apply Lemma 1b to $v(t)$ we can make a decomposition

$$v(t) = \breve{H}^0(q) \breve{e}(t) = \begin{bmatrix} H^0_a(q) & 0 \\ H^0_b(q) - \Gamma^0 & I \end{bmatrix} \begin{bmatrix} e \\ \Gamma^0 e \end{bmatrix} \tag{11}$$

where $\breve{H}^0$ satisfies the conditions in Lemma 1b, and with the L-dimensional white noise process $\breve{e}$ with covariance matrix $\breve{\Lambda}^0$ defined by:

$$\breve{\Lambda}^0 = \begin{bmatrix} I \\ \Gamma^0 \end{bmatrix} \Lambda^0 \begin{bmatrix} I \\ \Gamma^0 \end{bmatrix}^T. \tag{12}$$

From the definition of $\breve{e}$ we can see that there is a particular relation between the driving white noise process in the first p nodes and the last $L - p$ nodes. This particular relation is used throughout the paper.

Note that with (11) there are actually two different noise model representations:

$$v(t) = H_0(q) e(t) = \breve{H}_0(q) \breve{e}(t)$$

with $\breve{e}(t)$ and $v(t)$ being L-dimensional, and $e(t)$ being p-dimensional, with $p \leq L$. In the case of full-rank noise, $p = L$ and both representations are the same. Both expressions will be utilized.

The white noise process $e(t)$ is modeled as a stationary stochastic process. The probability density function (pdf) of the rank-reduced process $\breve{e}$ is defined by two equations (Rao, 1973), i.e. the pdf of e and the additional constraint

$$\begin{bmatrix} \Gamma^0 & -I \end{bmatrix} \breve{e} = 0. \tag{13}$$

An interpretation of this characterization of $\breve{e}$, is a p-dimensional pdf that lives on a plane described by (13). This interpretation is illustrated in Fig. 2 for an example of a 2-dimensional noise process $\breve{e}(t)$ having rank 1 with a Gaussian pdf.

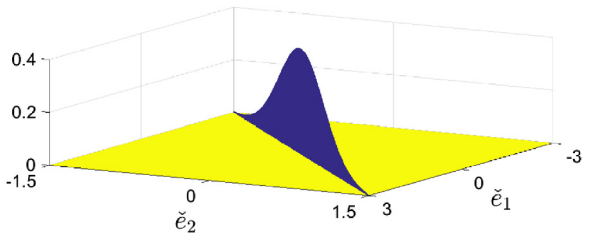

Fig. 2. pdf of rank-reduced noise $\check{e}(t) = [e(t)\ 0.5e(t)]^T$, with $e(t) \sim \mathcal{N}(0, \Lambda^0)$ a 1-dimensional random variable.

3. The joint-direct identification setup

After having defined the basic network properties and representations, the next step is to formulate the identification setting. Since our goal is identify the full network dynamics, i.e. all modules G_{ji}^0 that are actually present in the network, we are going to build a predictor model that predicts all measured node signals w in the network.

Definition 1. The one-step-ahead predictor for node signals $w(t)$ is defined as the conditional expectation

$$\hat{w}(t|t-1) := \mathbb{E}\left\{w(t) \mid w^{t-1},\ r^t\right\}, \tag{14}$$

conditioned on $w^{t-1} := \{w(0), w(1), \ldots, w(t-1)\}$ and $r^t := \{r(0), r(1), \ldots, r(t)\}$. □

We have shown that there are multiple ways to model the noise process v. In order to write a unique and explicit form for the predictor filters that generate the one-step-ahead prediction, we use the square version of the noise model (11), i.e. $v = \check{H}^0\check{e}$. This leads to the following result.

Proposition 1. *For a dynamic network considered in Section 2, the one-step-ahead predictor of the node signals $w(t)$ is given by*

$$\hat{w}(t|t-1) = W_w^0(q)w(t) + W_r^0(q)r(t), \tag{15}$$

with the predictor filters

$$W_w^0(q) = I - (\check{H}^0(q))^{-1}(I - G^0(q)), \tag{16}$$

$$W_r^0(q) = (\check{H}^0(q))^{-1}R^0(q). \tag{17}$$

Proof. Collected in the Appendix. □

Remark 1. In earlier work (Weerts et al., 2016a) the alternative noise model, determined by $H^0(q)$ was used as a basis for formulating the predictor filters. However due to intrinsic non-uniqueness of the corresponding filter expressions, the use of the square noise model $\check{H}^0$ is more attractive. Note that a subtle difference between the noise models $\check{H}^0$ and H^0 is that in $\check{H}^0$ (11) the feedthrough term $\Gamma^0 = \lim_{z\to\infty} H_b(z)$ is actually removed from $\check{H}^0$ and is represented now in $cov(\check{e})$ (12).

In order to arrive at a network identification setup we need to specify a network model and a network model set.

Definition 2 (*Network Model*)**.** A network model of a network with L nodes, and K external excitation signals, with a noise process of rank $p \leq L$ is defined by the quadruple:

$$M = (G, R, H, \Lambda)$$

with

- $G \in \mathbb{R}^{L\times L}(z)$, diagonal entries 0, all modules strictly proper and stable;
- $R \in \mathbb{R}^{L\times K}(z)$, proper and stable;
- $H \in \mathbb{R}^{L\times p}(z)$, satisfying the properties for $H^0(z)$.
- $\Lambda \in \mathbb{R}^{p\times p}$, $\Lambda > 0$;
- the network is well-posed[1] (Dankers, 2014), with $(I - G)^{-1}$ proper and stable. □

The data generating system is indicated by the model $\mathcal{S} = (G^0, R^0, H^0, \Lambda^0)$.

Definition 3 (*Network Model Set*)**.** A network model set for a network of L nodes, K external excitation signals, and a noise process of rank $p \leq L$, is defined as a set of parameterized matrix-valued functions:

$$\mathcal{M} := \left\{M(\theta) = \left(G(q,\theta), R(q,\theta), H(q,\theta), \Lambda(\theta)\right), \theta \in \Theta\right\},$$

with all models $M(\theta)$ satisfying the properties as listed in Definition 2. □

The data generating system $\mathcal{S}$ is represented by parameter θ_0, so $\mathcal{S} = M(\theta_0)$. In the parameterization the feedthrough of H_b is modeled by $\Gamma(\theta)$ defined as $\Gamma(\theta) := \lim_{z\to\infty} H_b(z,\theta)$.

Predictor (15) will be parameterized through $G(q,\theta)$, $R(q,\theta)$, $\check{H}(q,\theta)$, chosen according to a particular model set $\mathcal{M}$, to create the parameterized predictor

$$\begin{aligned}\hat{w}(t|t-1,\theta) = w(t) + \\ - \left(\check{H}(q,\theta)\right)^{-1}\{(I - G(q,\theta))w(t) - R(q,\theta)r(t)\},\end{aligned} \tag{18}$$

with

$$\check{H}(q,\theta) = \begin{bmatrix} H_a(q,\theta) & 0 \\ H_b(q,\theta) - \Gamma(\theta) & I \end{bmatrix}. \tag{19}$$

The L dimensional prediction error is then defined as

$$\varepsilon(t,\theta) := w(t) - \hat{w}(t|t-1,\theta). \tag{20}$$

For estimating the model parameters a Weighted Least Squares (WLS) criterion is considered:

$$\hat{\theta}_N^{WLS} = \arg\min_{\theta\in\Theta} \frac{1}{N}\sum_{t=1}^{N} \varepsilon^T(t,\theta)\, Q\, \varepsilon(t,\theta), \tag{21}$$

with $Q \geq 0$. The weight Q in the WLS criterion is necessary in order to show maximum likelihood properties, and thus asymptotic minimum variance properties of our estimated models, in Section 4. For analysis of consistency, first we need to introduce the notion of network identifiability. Network identifiability ensures that we are able to distinguish between different network topologies and dynamics.

Definition 4 (*Network Identifiability, Weerts et al., 2018*)**.** The network model set $\mathcal{M}$ is globally network identifiable at $M_0 := M(\theta_0)$ if for all models $M(\theta_1) \in \mathcal{M}$,

$$\left.\begin{aligned} T_{wr}(q,\theta_1) &= T_{wr}(q,\theta_0) \\ \Phi_{\bar{v}}(\omega,\theta_1) &= \Phi_{\bar{v}}(\omega,\theta_0) \end{aligned}\right\} \Rightarrow M(\theta_1) = M(\theta_0). \tag{22}$$

$\mathcal{M}$ is globally network identifiable if (22) holds for all $M_0 \in \mathcal{M}$. □

Since all modules in $G(q,\theta)$ are assumed to be strictly proper, a result of Weerts et al. (2018) can be employed to show that condition (22) in Definition 4 is equivalently formulated as:

$$\begin{aligned}&\{T(q,\theta_1) = T(q,\theta_0)\} \Rightarrow \\ &\{(G(\theta_1), R(\theta_1), H(\theta_1)) = (G(\theta_0), R(\theta_0), H(\theta_0))\},\end{aligned} \tag{23}$$

with $T(q,\theta)$ the parameterized version of $T^0(q)$ (3). A sufficient condition for a model structure to be network identifiable is that

[1] The implies that all principal minors of $\lim_{z\to\infty}(I - G(z))^{-1}$ are nonzero.

every node has an independent excitation source, which can be either process noise or an external excitation r. This is a result of the condition that the columns of $[H(q,\theta)\ R(q,\theta)]$ can be permuted to arrive at a matrix with a leading diagonal (Weerts et al., 2018). An alternative –and stronger – condition that takes into account the structure of $G(\theta)$, is presented in the following Proposition.

Proposition 2 (*Weerts et al., 2018*)**.** *Let $\mathcal{M}$ be a network model set that satisfies the following properties:*

a. *Every parameterized entry in the model $\{M(z,\theta), \theta \in \Theta\}$ covers the set of all proper rational transfer functions;*
b. *All parameterized transfer functions in the model $M(z,\theta)$ are parameterized independently (i.e. there are no common parameters).*

Then $\mathcal{M}$ is globally network identifiable at $M(\theta_0)$ if and only if

- *each row i of the transfer function matrix $\begin{bmatrix} G(\theta) & H(\theta) & R(\theta)\end{bmatrix}$ has at most $K+p$ parameterized entries, and*
- *for each i, the transfer function from all external signals (r_m, e_n) that are input to a non-parameterized module $R_{jm}(q)$ or $H_{in}(q)$, to node signals w_k that are input to parameterized modules $G_{ik}(q;\theta)$, is full row rank in $\theta=\theta_0$.* □

A detailed interpretation of the listed conditions is provided in Weerts (2018) and Weerts et al. (2018). For analysis of the asymptotic properties of the parameter estimate (21) it is attractive to consider the asymptotic criterion

$$\theta^\star = \arg\min_{\theta\in\Theta} \bar{V}(\theta), \tag{24}$$

with

$$\bar{V}(\theta) = \bar{\mathbb{E}}\ \varepsilon^T(t,\theta)\ Q\ \varepsilon(t,\theta), \tag{25}$$

and $\bar{\mathbb{E}}$ defined as $\lim_{N\to\infty}\sum_{t=1}^{N}\mathbb{E}$, according to Ljung (1999). In classical literature it has been shown that the solution of the weighted least squares criterion converges to the solution of the asymptotic criterion under some mild conditions (Ljung, 1999). Based on this result we can state that, under the condition that $w(t)$ and $r(t)$ are jointly quasi-stationary, $r(t)$ is bounded, and $e(t)$ has bounded moments of order ≥ 4, it holds that

$$\hat{\theta}_N^{WLS} \to \theta^\star \text{ w.p. 1 as } N\to\infty. \tag{26}$$

A consistent estimate is obtained if $\theta^\star$ is equal to θ_0. Conditions for this to hold are formulated in the next Proposition, which was introduced in Weerts et al. (2017).

Proposition 3. *Let $\mathcal{M}$ be a model set according to Definition 3, and let $\theta^\star$ be defined by (24). Then under the conditions*

(1) *The data generating system is in the model set, i.e. $\exists\theta_0\in\Theta$ such that $M(\theta_0)=\mathcal{S}$, and*
(2) *$\mathcal{M}$ is globally network identifiable at $\mathcal{S}$, and*
(3) *the external excitation signal r, if present, is persistently exciting of sufficiently high order and uncorrelated with e,*

it holds that[2]

$$\begin{aligned}&\{G(q,\theta^\star), H_a(q,\theta^\star), H_b(q,\theta^\star) - \Gamma(\theta^\star), R(q,\theta^\star)\}\\ &\qquad = \{G^0(q), H_a^0(q), H_b^0(q) - \Gamma^0, R^0(q).\}\end{aligned} \tag{27}$$

Proof. Collected in the Appendix. □

The matrix Γ^0 that appears in $cov(\check{e})$ (12) is not estimated through the criterion (21), but information on Γ^0 exists in the residual $\varepsilon(t,\theta^\star)$. Based on the dependencies in the innovation signal $\check{e}(t)$ (11) we split the prediction error into two parts:

$$\varepsilon(t,\theta) = \begin{bmatrix}\varepsilon_a(t,\theta)\\ \varepsilon_b(t,\theta)\end{bmatrix}, \tag{28}$$

where $\varepsilon_a \in \mathbb{R}^p$, and $\varepsilon_b \in \mathbb{R}^{L-p}$. Under zero initial conditions in the system and the predictor filters, the prediction error, when evaluated at $\theta=\theta_0$, has the same dependencies as the innovation, i.e.

$$\varepsilon_a(t,\theta_0) = e(t), \text{ and } \quad \varepsilon_b(t,\theta_0) = \Gamma^0 e(t),$$

such that $\Gamma^0\varepsilon_a(t,\theta^0) = \varepsilon_b(t,\theta^0)$. Based on this equation an estimate of Γ^0 can be constructed according to

$$\hat{\Gamma}_N = \left(\frac{1}{N}\sum_{t=1}^{N}\varepsilon_b(\hat{\theta}_N)\varepsilon_a^T(\hat{\theta}_N)\right)\left(\frac{1}{N}\sum_{t=1}^{N}\varepsilon_a(\hat{\theta}_N)\varepsilon_a^T(\hat{\theta}_N)\right)^{-1}. \tag{29}$$

If $\hat{\theta}_N$ is a consistent estimate, this estimate $\hat{\Gamma}_N$ will converge (w.p. 1) to

$$\Gamma^\star = \left(\mathbb{E}\ \varepsilon_b(\theta^\star)\varepsilon_a^T(\theta^\star)\right)\left(\mathbb{E}\ \varepsilon_a(\theta^\star)\varepsilon_a^T(\theta^\star)\right)^{-1} \tag{30}$$

which is

$$\Gamma^\star = \Gamma^0\Lambda^0(\Lambda^0)^{-1} = \Gamma^0, \tag{31}$$

showing that also $\hat{\Gamma}_N$ is consistent.

For full-rank noise the weight $Q = (\Lambda^0)^{-1}$ typically leads to minimum variance estimates, but for rank-reduced noise, $\check{\Lambda}^0$ is not invertible. In order to obtain minimum variance properties, further analysis is required to determine an optimal weighting matrix Q in the identification criterion (21).

The identification method that has been presented here is termed as "joint-direct method", as it combines elements from two classical methods for closed-loop identification (Ljung, 1999), i.e. the joint-io method that is based on treating all measured signals jointly and starts with estimating closed-loop transfer function objects, and the direct method in which plant and noise dynamics are parameterized directly.

4. Constrained least squares and maximum likelihood

The WLS criterion (21) does not take into account the fact that there are dependencies in the innovation process $\check{e}(t)$, as represented in (13). In this section we introduce an identification criterion which properly takes these dependencies into account. It is shown that this approach leads to Maximum Likelihood estimates and to an appropriate choice of weight for the WLS criterion. Based on the dependencies in the innovation we define

$$Z(t,\theta) := \Gamma(\theta)\varepsilon_a(t,\theta) - \varepsilon_b(t,\theta), \tag{32}$$

and introduce the Constrained Least Squares (CLS) estimator:

$$\begin{aligned}\hat{\theta}_N^{CLS} &= \arg\min_\theta \frac{1}{N}\sum_{t=1}^{N}\varepsilon_a^T(t,\theta)\ Q_a\ \varepsilon_a(t,\theta)\\ &\text{subject to } \frac{1}{N}\sum_{t=1}^{N}Z^T(t,\theta)Z(t,\theta) = 0,\end{aligned} \tag{33}$$

with $Q_a > 0$. For finite N, the quadratic constraint is equivalent to the constraint $Z(t,\theta)=0\ \forall t$, which was introduced in Weerts et al. (2017). We have chosen for a quadratic constraint as this facilitates the convergence and consistency result in the next proposition, and because it is less computationally demanding.

While the term $\Gamma(\theta)$ is not present in the quadratic cost function of (33), it enters the estimation procedure now through the

[2] Strictly speaking $\theta^\star$ can be a set and the equation holds for all $\theta\in\theta^\star$.

constraint. Consistency of the CLS estimate is formulated in the next proposition of which a preliminary version was presented in Weerts et al. (2017).

Proposition 4. *Let $\mathcal{M}$ be a model set according to Definition 3, let $\hat{\theta}_N^{CLS}$ be defined by (33) and let θ^* be defined by*

$$\theta^* = \arg\min_\theta \bar{\mathbb{E}}\ \varepsilon_a^T(t,\theta)\ Q_a\ \varepsilon_a(t,\theta) \quad \text{subject to } \bar{\mathbb{E}}\ Z^T(t,\theta)Z(t,\theta) = 0. \tag{34}$$

(1) *Under the conditions that $w(t)$ and $r(t)$ are jointly quasi-stationary, $r(t)$ is bounded, and $e(t)$ has bounded moments of order ≥ 4, it holds that*

$$\hat{\theta}_N^{CLS} \to \theta^* \text{ w.p. 1 as } N \to \infty. \tag{35}$$

(2) *Under the conditions that*

(a) *The data generating system is in the model set, i.e. $\exists \theta_0 \in \Theta$ such that $M(\theta_0) = \mathcal{S}$, and*
(b) *$\mathcal{M}$ is globally network identifiable at $\mathcal{S}$, and*
(c) *the external excitation signal r, if present, is persistently exciting of sufficiently high order and uncorrelated to e,*

it holds that[3]

$$\{G(q,\theta^*), H(q,\theta^*), R(q,\theta^*)\} = \{G^0(q), H^0(q), R^0(q)\}. \tag{36}$$

Proof. Collected in the Appendix. □

As opposed to the consistency result for the WLS estimate in Proposition 3, now the term $\Gamma(\theta)$, which is included in $H_b(q,\theta)$, is also estimated consistently, as it is part of the constraint equation in (34).

Since the dependencies in the noise terms are properly handled, it can be expected that the CLS estimate has a close resemblance with the maximum likelihood estimate, that asymptotically reaches the Cramér–Rao lower bound. This is analyzed next.

Theorem 1. *Let $e(t)$ be normally distributed and zero mean, i.e. $e(t) \sim \mathcal{N}(0, \Lambda^0)$, and consider a parameterized model set as in Definition 3. Then under zero initial conditions*[4] *:*

(1) *The Maximum Likelihood estimate of θ^0 is*

$$\hat{\theta}_N^{ML} = \arg\max_\theta \log L_a(\theta) \quad \text{subject to } \frac{1}{N}\sum_{t=1}^N Z^T(t,\theta)Z(t,\theta) = 0, \tag{37}$$

with

$$\log L_a(\theta) = c - \frac{N}{2}\log\det\Lambda(\theta) - \frac{1}{2}\sum_{t=1}^N \varepsilon_a^T(t,\theta)\Lambda^{-1}(\theta)\varepsilon_a(t,\theta). \tag{38}$$

(2) *Under the condition that $\Lambda(\theta)$ does not share parameters with $\varepsilon(t,\theta)$ the Maximum Likelihood estimate can alternatively be written as*

$$\hat{\theta}_N^{ML} = \arg\min_\theta \det\left(\frac{1}{N}\sum_{t=1}^N \varepsilon_a(t,\theta)\varepsilon_a^T(t,\theta)\right) \quad \text{subject to } 0 = \frac{1}{N}\sum_{t=1}^N Z^T(t,\theta)Z(t,\theta), \tag{39}$$

$$\Lambda(\theta) = \frac{1}{N}\sum_{t=1}^N \varepsilon_a(t,\theta)\varepsilon_a^T(t,\theta).$$

Proof. Collected in the Appendix. □

In (39) the last equation does not involve an actual constraint that limits the optimization problem, but it is merely there to calculate the estimated Λ.

Note that when a model set with fixed (non-parameterized) Λ is used, then the Maximum Likelihood estimate (37) reduces to the Constrained Least Squares estimate (33) with $Q_a = \Lambda^{-1}$. This implies that the CLS equipped with the appropriate weight $Q_a = (\Lambda^0)^{-1}$ is a maximum likelihood estimator in case of Gaussian disturbances.

Remark 2. If initial conditions are non-zero and not explicitly dealt with in the parameterized model, then part of the prediction error is caused by the initial conditions. Although this effect asymptotically goes to 0, the signal ε_b does not have to be linearly dependent on ε_a, and consequently there do not exist parameters for which $Z(t,\theta) = 0$ for all t. Similarly in the situation where $\mathcal{M}$ does not contain $\mathcal{S}$, it is possible that there do not exist parameters for which $Z(t,\theta) = 0$ for all t. When $Z(t,\theta)$ cannot be made 0 the constraint in (33) and (37) is not feasible and the solution set of the criterion is empty.

In order to deal with situations where there are non-zero initial conditions, or where the system is not in the model set, we introduce a relaxed criterion. This relaxed criterion has a relaxed constraint, which appears as an additional penalty term, weighted by the real-valued penalty weight $\lambda > 0$:

$$\hat{\theta}_N^{rel} = \arg\min_\theta \frac{1}{N}\sum_{t=1}^N \Big(\varepsilon_a^T(t,\theta)Q_a\varepsilon_a(t,\theta) + \lambda Z^T(t,\theta)Z(t,\theta)\Big). \tag{40}$$

The above criterion is equivalent to the CLS (33) for $\lambda \to \infty$. Another way to write the relaxed criterion is as the WLS (21) with parameterized weight

$$Q(\theta) = \begin{bmatrix} Q_a + \lambda\Gamma^T(\theta)\Gamma(\theta) & -\lambda\Gamma^T(\theta) \\ -\lambda\Gamma(\theta) & \lambda I \end{bmatrix}. \tag{41}$$

The optimal choice for λ will depend on the contribution of initial conditions, the contribution of unmodeled dynamics and the length of the data record. A further analysis of this optimal value is beyond the scope of this paper.

Remark 3. So far we have assumed that the node signals are ordered in such a way that the first p nodes are affected by a full-rank noise process, while the remaining $L - p$ nodes are affected by "dependent" noise. In Weerts et al. (2018) conditions have been derived for appropriately ordering the noise components. For the current situation where the modules $G_{jl}(q)$ are strictly proper, the rank p and the ordering of signals can be retrieved from $T_{wr}^\infty := \lim_{z\to\infty} T_{wr}(z)$ and $\Phi_{\tilde{v}}^\infty := \lim_{z\to\infty}\Phi_{\tilde{v}}(z)$. This information can be estimated from data. In this paper we will assume that the requested ordering has been performed prior to the identification of the dynamics of the network.

[3] Strictly speaking θ^* can be a set and the equation holds for all $\theta \in \theta^*$.

[4] The zero initial conditions reflect values of input and output values of the predictor filters, prior to the time interval $[1, N]$, that are required to calculate the predicted node signal within the time interval.

Remark 4. So far we have considered the situation that all modules in $G(q,\theta)$ are strictly proper. This situation can be extended to the situation of having proper modules in $G(q,\theta)$, thus allowing direct feedthrough terms, as long as there are no algebraic loops in the network. The network identifiability results of Weerts et al. (2018) include this case. The 2-node situation has been solved even for the situation of having algebraic loops (Weerts et al., 2016b). Since the formulation of the ML result will become technically more involved for non-strictly proper modules, we have preferred to restrict to the strictly proper module situation here.

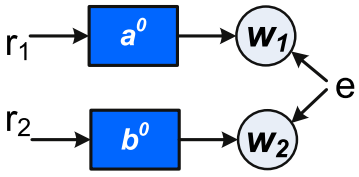

Fig. 3. System with 2 nodes, no dynamics and 1 noise disturbance. It is excited by the quasi-stationary excitation signals r_1, r_2 and the stochastic process e which are all mutually uncorrelated and have unit variance.

5. Minimum variance and the Cramér–Rao lower bound

5.1. Variance of weighted least squares estimates

In the situation that the noise is full rank the classical reasoning on parameter variance holds (Ljung, 1999). For $N \to \infty$ and $\mathcal{S} \in \mathcal{M}$ the estimate converges under weak conditions to a normal distribution given by

$$\underbrace{\sqrt{N}(\hat{\theta}_{CLS} - \theta^0)}_{:=\tilde{\theta}} \sim \mathcal{N}(0, P_\theta), \tag{42}$$

with P_θ positive definite. For full-rank noise processes, P_θ is defined by

$$P_\theta = \left[\bar{\mathbb{E}}\psi(t)Q\psi^T(t)\right]^{-1}\left[\bar{\mathbb{E}}\psi(t)Q\Lambda^0Q\psi^T(t)\right]\cdot \cdot\left[\bar{\mathbb{E}}\psi(t)Q\psi^T(t)\right]^{-1}, \tag{43}$$

with

$$\psi(t) := -\frac{d}{d\theta}\varepsilon^T(t,\theta)|_{\theta=\theta_0}. \tag{44}$$

For rank-reduced noise and the WLS criterion (21) the expression for P_θ is similar to the expression above, with Λ^0 replaced by $\check{\Lambda}^0$. This can be shown by following its derivation in Söderström and Stoica (1989) and using $\check{\Lambda}^0$ instead of Λ^0. The covariance matrix has a lower bound P_θ^0, leading to $P_\theta \geq P_\theta^0$, which for full-rank noise is given by

$$P_\theta^0 = \left[\bar{\mathbb{E}}\psi(t)(\Lambda^0)^{-1}\psi^T(t)\right]^{-1}. \tag{45}$$

However for the rank-reduced case, Λ^0 would have to be replaced by $\check{\Lambda}^0$, which, however, is singular, and so its inverse does not exist. Therefore a lower bound like (45) is not valid in this case.

The question is now what the minimum variance is when noise is rank-reduced. In the following example we will illustrate this problem.

Example 1. Consider the system in Fig. 3, where 2 parameters are to be estimated, θ_a and θ_b. The system is governed by

$$\begin{bmatrix} w_1(t) \\ w_2(t) \end{bmatrix} = \begin{bmatrix} a^0 & 0 \\ 0 & b^0 \end{bmatrix}\begin{bmatrix} r_1(t) \\ r_2(t) \end{bmatrix} + \underbrace{\begin{bmatrix} e(t) \\ e(t) \end{bmatrix}}_{\check{e}(t)}.$$

The disturbance process $\check{e}$ has covariance matrix $\check{\Lambda}^0 = \left[\begin{smallmatrix} 1 & 1 \\ 1 & 1 \end{smallmatrix}\right]$, which is singular. When the WLS (21) is used with a weight Q defined by (41) and $\Gamma(\theta)$ set to 1,

$$Q = \begin{bmatrix} 1+\lambda & -\lambda \\ -\lambda & \lambda \end{bmatrix} \tag{46}$$

with $\lambda > 0$, and prediction errors

$$\varepsilon_1 = w_1 - \theta_a r_1, \quad \varepsilon_2 = w_2 - \theta_b r_2, \tag{47}$$

we arrive at a consistent estimate. The identification criterion to be minimized becomes

$$\frac{1}{N}\sum_{t=1}^{N}\left\{\frac{1}{\lambda}\varepsilon_a^2(\theta_a) + \left(\begin{bmatrix} 1 & -1 \end{bmatrix}\begin{bmatrix} \varepsilon_a(\theta_a) \\ \varepsilon_b(\theta_b) \end{bmatrix}\right)^2\right\}. \tag{48}$$

In the limit as $\lambda \to \infty$, Q becomes singular, and the expression for the identification criterion reduces to

$$\frac{1}{N}\sum_{t=1}^{N}\left\{(a^0 - \theta_a)r_1 + e - (b^0 - \theta_b)r_2 - e\right\}^2. \tag{49}$$

In this expression the disturbance e drops out, and variance-free estimates of a^0 and b^0 are obtained.

This phenomenon also appears in the calculation of the covariance matrix (43). If we denote $\theta = [\theta_a\ \theta_b]^T$ then

$$\psi(t) = \begin{bmatrix} r_1(t) & 0 \\ 0 & r_2(t) \end{bmatrix}, \tag{50}$$

and we obtain

$$\bar{\mathbb{E}}\psi Q\psi^T = \bar{\mathbb{E}}\begin{bmatrix} r_1^2(1+\lambda) & -r_1r_2\lambda \\ -r_1r_2\lambda & r_2^2\lambda \end{bmatrix} = \begin{bmatrix} (1+\lambda) & 0 \\ 0 & \lambda \end{bmatrix} \tag{51}$$

and

$$\bar{\mathbb{E}}\psi Q\check{\Lambda}^0Q\psi^T = \bar{\mathbb{E}}\begin{bmatrix} r_1^2 & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}. \tag{52}$$

We can compute P_θ of (43) as

$$P_\theta = \begin{bmatrix} (1+\lambda) & 0 \\ 0 & \lambda \end{bmatrix}^{-1}\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} (1+\lambda) & 0 \\ 0 & \lambda \end{bmatrix}^{-1} = \begin{bmatrix} \frac{1}{(1+\lambda)^2} & 0 \\ 0 & 0 \end{bmatrix}. \tag{53}$$

Here we can see that as $\lambda \to \infty$ the covariance goes to 0. This phenomenon of variance-free estimation has also been observed in Everitt, Bottegal, Rojas, and Hjalmarsson (2015).

It could be tempting to use an expression like (45) for the lower bound on the covariance matrix, with the inverse covariance $(\Lambda^0)^{-1}$ replaced by a pseudo-inverse of Λ^0. In this example $(\check{\Lambda}^0)^\dagger = \frac{1}{4}\left[\begin{smallmatrix} 1 & 1 \\ 1 & 1 \end{smallmatrix}\right]$ and substituting this for $(\Lambda^0)^{-1}$ into (45) delivers

$$\left[\bar{\mathbb{E}}\psi(t)(\check{\Lambda}^0)^\dagger\psi^T(t)\right]^{-1} \neq 0, \tag{54}$$

which cannot be the expression for the minimum variance.

Note that this example is fully symmetric in nodes w_1 and w_2, or equivalently in systems a^0 and b^0. Nevertheless, for finite values of λ, one of the parameters θ_b is estimated variance-free, while θ_a is not. This is the result of the particular choice of weighting function, that according to (41) reflects the choice of w_1 as the full-rank noise node. Choosing the alternative weight $Q = \left[\begin{smallmatrix} \lambda & -\lambda \\ -\lambda & 1+\lambda \end{smallmatrix}\right]$ would resemble the situation of choosing w_2 as the full rank noise node, For both the weights, when we let $\lambda \to \infty$ the variance-free maximum likelihood estimate is obtained, which is again symmetric in θ_a and θ_b.

In this example it is possible to choose a weight beyond the structure of (41), e.g. $Q = \left[\begin{smallmatrix} 1 & -1 \\ -1 & 1 \end{smallmatrix}\right]$, in which case we arrive at a variance-free estimate, since $Q\Lambda^0 Q = 0$. For this choice of Q we are essentially only modeling the 'constraint', and we dropped the 'original cost function' ε_a^2. Such a weight Q is useful when all parameters in the model can be estimated using just the constraint.

We have shown that, also in the reduced-rank noise case, the choice of weighting matrix Q in the WLS criterion affects the covariance matrix of the estimate. A formal analysis of the asymptotic variance in the case of CLS follows next.

5.2. Variance of constrained least squares estimates

In order to find a closed-form expression for the minimum variance in the case of the CLS estimate, we have to address the full impact of the constraint, that typically reduces the effective parameter space in the criterion.

For the CLS situation we will write the asymptotic identification criterion as follows:

$$\begin{aligned} \theta^* &= \arg\min_{\theta} \bar{\mathbb{E}}\, \varepsilon_a^T(t,\theta)\, Q_a\, \varepsilon_a(t,\theta) \\ &\quad \text{subject to: } f(\theta) = 0, \end{aligned} \tag{55}$$

with $f(\theta)$ defined by $f(\theta) = \bar{\mathbb{E}}\, Z^T(t,\theta) Z(t,\theta)$. For variance analysis we will assume that $\theta^* = \theta^0$.

In a neighborhood around $\theta = \theta^0$ the constraint can be approximated using a first order Taylor series

$$Z(t,\theta) \approx Z(t,\theta^0) + \frac{\partial Z(t,\theta)}{\partial \theta}\Big|_{\theta=\theta^0} (\theta - \theta^0) = A(t)(\theta - \theta^0), \tag{56}$$

where

$$A(t) := \frac{\partial Z(t,\theta)}{\partial \theta}\Big|_{\theta=\theta^0} \tag{57}$$

with $A \in \mathbb{R}^{(L-p)\times n_\theta}$. The approximated constraint is then

$$\bar{\mathbb{E}}\, (\theta - \theta^0)^T A^T(t) A(t) (\theta - \theta^0) = 0, \tag{58}$$

where $\bar{\mathbb{E}}\, A^T(t)A(t)$ is of dimension $n_\theta \times n_\theta$. Note that $Z(t,\theta^0) = 0$, but that $Z(t,\theta)$ with θ in the neighborhood of θ^0 is non-zero. We can define a matrix Π of dimension $(n_\theta - n_\rho) \times n_\theta$ of full row rank such that

$$\bar{\mathbb{E}}\, A^T(t)A(t) = \Pi^T \Pi. \tag{59}$$

Then in the neighborhood of θ^0 the constraint is approximated by a quadratic constraint, leading to

$$\begin{aligned} \theta^* &= \arg\min_{\theta} \bar{\mathbb{E}}\, \varepsilon_a^T(t,\theta)\, Q_a\, \varepsilon_a(t,\theta) \\ &\quad \text{subject to: } (\theta - \theta^0)^T \Pi^T \Pi (\theta - \theta^0) = 0. \end{aligned} \tag{60}$$

In order to take the constraint into account in the variance analysis, a re-parameterization will be introduced, using a parameter ρ with $\dim(\rho) = n_\rho < \dim(\theta)$. The two parameters will be related through a mapping induced by the constraint, such that the new parameterization trivially satisfies the constraint.

Lemma 2. *The constrained parameter space determined by $(\theta - \theta^0)^T \Pi^T \Pi (\theta - \theta^0) = 0$, with Π defined in (59), is equivalently described by*

$$\theta = S\rho + C, \quad \text{with } \rho \in \mathbb{R}^{n_\rho}, \tag{61}$$

where $S \in \mathbb{R}^{n_\theta \times n_\rho}$ satisfies $\Pi S = 0$ and is full column rank, i.e. S characterizes the right nullspace of Π, and $C = \Pi^\dagger \Pi \theta^0$, where $\Pi^\dagger$ satisfies $\Pi \Pi^\dagger = I$.

Proof. Collected in the Appendix.

The unconstrained parameter ρ can now be used to rewrite the criterion (60) into a form that trivially satisfies the constraint. The resulting criterion is then essentially an unconstrained criterion operating on a lower dimensional parameter ρ.

Proposition 5. *The optimization problem (60) can equivalently be written as*

$$\theta^* = S\rho^* + C, \tag{62}$$

with

$$\rho^* = \arg\min_{\rho} \bar{\mathbb{E}}\, \varepsilon_a(t, S\rho + C)\, Q_a\, \varepsilon_a(t, S\rho + C). \tag{63}$$

Proof. Collected in the Appendix. □

Since (63) is an unconstrained identification criterion, we know that the asymptotic variance of the estimate $\hat{\rho}_N$ that corresponds to the asymptotic estimate ρ^* is given by

$$\begin{aligned} P_\rho = \left[\bar{\mathbb{E}}\psi_\rho(t) Q_a \psi_\rho^T(t)\right]^{-1} \left[\bar{\mathbb{E}}\psi_\rho(t) Q_a \Lambda^0 Q_a \psi_\rho^T(t)\right] \cdot \\ \cdot \left[\bar{\mathbb{E}}\psi_\rho(t) Q_a \psi_\rho^T(t)\right]^{-1}, \end{aligned} \tag{64}$$

with

$$\psi_\rho(t) = -\frac{d}{d\rho} \varepsilon_a^T(t, S\rho + C)|_{\rho=\rho^*}. \tag{65}$$

Combining this expression with (61) now provides an expression for P_θ, as formulated next.

Proposition 6. *The covariance matrices P_ρ and P_θ satisfy the following relation*

$$P_\theta = S P_\rho S^T. \tag{66}$$

Proof. Collected in the Appendix. □

It is well known that the lower bound of P_ρ is achieved when $Q_a = (\Lambda^0)^{-1}$, such that

$$P_\rho \geq P_\rho^0 = \left[\bar{\mathbb{E}}\psi_\rho(t)(\Lambda^0)^{-1}\psi_\rho^T(t)\right]^{-1}. \tag{67}$$

Then by Proposition 6 the lower bound of P_θ is achieved by

$$P_\theta \geq P_\theta^0 = S P_\rho^0 S^T. \tag{68}$$

So the lower bound of P_θ is achieved for $Q_a = (\Lambda^0)^{-1}$. Matrix S characterizes the right-nullspace of Π, so it is not a unique matrix, but all possible S matrices lead to the same P_θ and lower bound. As an illustration of the results, an example is shown for the CLS estimate.

Example 2. In this example, depicted in Fig. 4, the system is given by

$$w_1(t) = r_1(t) + 0.5 r_2(t) + e(t), \quad w_2(t) = r_2(t) + e(t). \tag{69}$$

The noise is rank-reduced, and has covariance matrix $\breve{\Lambda}^0 = \left[\begin{smallmatrix} 1 & 1 \\ 1 & 1 \end{smallmatrix}\right]$, which is singular. When the CLS (33) is used with knowledge of $\Gamma^0 = 1$ and prediction errors

$$\varepsilon_1 = w_1 - \theta_{a1} r_1 - \theta_{a2} r_2, \quad \varepsilon_2 = w_2 - \theta_b r_2, \tag{70}$$

where $\varepsilon_1 = \varepsilon_a$ and $\varepsilon_2 = \varepsilon_b$, we get a consistent estimate. The constraint consists of $\bar{\mathbb{E}}\, Z^2(t,\theta) = 0$ with

$$Z(t,\theta) = \varepsilon_1(t,\theta) - \varepsilon_2(t,\theta). \tag{71}$$

With $Z(t,\theta)$ being linear in θ it follows that

$$A(t) = \frac{\partial Z(t,\theta)}{\partial \theta}\Big|_{\theta=\theta^*} = \left[-r_1(t) \quad -r_2(t) \quad r_2(t)\right], \tag{72}$$

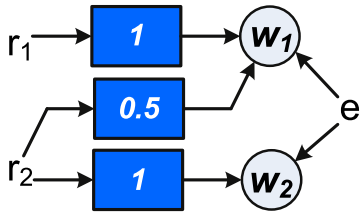

Fig. 4. System with 2 nodes, no dynamics and 1 noise disturbance. It is excited by the quasi-stationary excitation signals r_1, r_2 and the stochastic process e which are all mutually uncorrelated and have unit variance.

leading to

$$\bar{\mathbb{E}}A^T(t)A(t) = \bar{\mathbb{E}}\begin{bmatrix} r_1^2(t) & 0 & 0 \\ 0 & r_2^2(t) & -r_2^2(t) \\ 0 & -r_2^2(t) & r_2^2(t) \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & -1 \\ 0 & -1 & 1 \end{bmatrix}, \tag{73}$$

and with (59) that $\Pi = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & -1 \end{bmatrix}$.

Vectors S and C can now be determined based on $\Pi S = 0$ and $C = \Pi^\dagger \Pi \theta^*$, leading to:

$$S = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}, \quad C = \begin{bmatrix} 1 \\ -0.5 \\ 0 \end{bmatrix}.$$

With this choice of S, we can determine ψ_ρ using (65) as

$$\psi_\rho = -\frac{d}{d\rho}\left(w_1 - \begin{bmatrix} r_1 & r_2 & 0 \end{bmatrix}(S\rho + C)\right) = r_2. \tag{74}$$

Then P_ρ of (64) is given by:

$$P_\rho = (\bar{\mathbb{E}}\, r_2^2)^{-1}(\bar{\mathbb{E}}\, r_2^2)(\bar{\mathbb{E}}\, r_2^2)^{-1} = 1,$$

where $\Lambda^0 = Q_a = 1$. Applying Proposition 6, the covariance matrix of θ is then determined as

$$P_\theta = SP_\rho S^T = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix}.$$

Since we used the optimal weighting $Q_a = (\Lambda^0)^{-1}$ this is also the lower bound on the covariance matrix in the given situation. Note that in the considered situation the first parameter θ_{a_1} is estimated variance-free.

Remark 5. The lower bound on the covariance matrix can be 0, in particular situations. When the matrix $\bar{\mathbb{E}}A^T(t)A(t)$ in the constraint is square and full rank, then the constraint uniquely determines all parameters, and *all* parameters are determined variance-free.

Following a different line of reasoning, in Stoica and Ng (1998) the Cramér–Rao lower bound on the variance under parametric constraints has been derived for Gaussian distributed noise. That result can be linked to the lower bound that we just have obtained. In Stoica and Ng (1998) it is stated that first the Fisher information matrix J of the unconstrained part of the criterion (33) is obtained, which is

$$J = \bar{\mathbb{E}}\, \psi_a(t)\Lambda_0^{-1}\psi_a^T(t), \tag{75}$$

with $\psi_a(t) = \psi(t)\begin{bmatrix} I \\ 0 \end{bmatrix}$. This unconstrained part of the criterion does not contain all parameters, meaning that ψ_a contains rows that are 0, and J is singular. The lower bound on the covariance matrix cannot be given by J^{-1} since it does not exist. In Stoica and Ng (1998) it has been proven that the lower bound is given by

$$P_\theta^0 = S\left(S^T\bar{\mathbb{E}}\, \psi_a(t)\Lambda_0^{-1}\psi_a^T(t)S\right)^{-1}S^T, \tag{76}$$

with S as defined before. The above expression is equal to the lower bound in (68) that we obtained using a different reasoning, since by the chain rule for differentiation we have that

$$\psi_\rho(t) = S^T\psi_a(t) \tag{77}$$

which can be substituted in (76) to arrive at (68).

6. Simulation example

In this simulation example a 3 node network will be identified from data using the WLS and CLS criteria. We use the network in Fig. 1 with $r_1 = 0$ and v a 2-dimensional white noise process with $\Lambda^0 = I$, such that

$$G^0 = \begin{bmatrix} 0 & G_{12}^0 & G_{13}^0 \\ 0 & 0 & G_{23}^0 \\ G_{31}^0 & 0 & 0 \end{bmatrix}, \quad H^0 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}.$$

The dynamic modules are finite impulse responses with the following coefficients

$$\begin{bmatrix} G_{12}^0(q) \\ G_{13}^0(q) \\ G_{23}^0(q) \\ G_{31}^0(q) \end{bmatrix} = \begin{bmatrix} 0.33 & -0.2 & 0.13 & -0.08 & 0.05 \\ 0.2 & -0.45 & -0.73 & -0.54 & -0.25 \\ -0.15 & 0.12 & -0.9 & 0.6 & 0.3 \\ -0.5 & 0.06 & -0.1 & 0.03 & 0 \end{bmatrix} \times \begin{bmatrix} q^{-1} \\ q^{-2} \\ q^{-3} \\ q^{-4} \\ q^{-5} \end{bmatrix}.$$

In total 100 Monte-Carlo simulations are performed with the above network with $N = 1000$ samples taken for each data set.

A model structure is used with $G(q, \theta)$ having the same structure as G^0, $H(q, \theta) = \begin{bmatrix} I \\ \Gamma(\theta_\Gamma) \end{bmatrix}$, and with $\Lambda = I$. Parameters are collected in the vector

$$\theta^T = \begin{bmatrix} \theta_{12}^T & \theta_{13}^T & \theta_{23}^T & \theta_{31}^T & \theta_\Gamma^T \end{bmatrix} \in \mathbb{R}^{22}, \tag{78}$$

where θ_{ij} correspond to module $G_{ij}(\theta_{ij})$. The prediction error can be denoted by

$$\begin{bmatrix} \varepsilon_1(t,\theta) \\ \varepsilon_2(t,\theta) \\ \varepsilon_3(t,\theta) \end{bmatrix} = \begin{bmatrix} w_1(t) \\ w_2(t) \\ w_3(t) \end{bmatrix} - \begin{bmatrix} 0 & \phi_2(t) & \phi_3(t) & 0 \\ 0 & 0 & \phi_3(t) & 0 \\ \phi_1(t) & 0 & 0 & 0 \end{bmatrix}\theta, \tag{79}$$

with appropriately chosen regressors $\phi_i(t)$.

The WLS is applied as the relaxed CLS with weight (41) parameterized with $\Gamma(\theta)$. Two different choices for λ are used to illustrate the effect of increasing values of λ. Results of the WLS estimates, and of the CLS estimates, are plotted in Fig. 5.

It can be observed that the parameters of modules G_{12} and G_{13} do not change with different criteria, as the process noise on node 1 is independent of the process noises on nodes 2 and 3. The parameters of G_{23} and G_{31} are estimated with smaller variance when λ increases, since the estimate gets closer to the ML estimate. The parameters of Γ (indexed by numbers 21 and 22) are estimated with very small variance, even for small λ.

For this estimation the lower bound on the variance can be computed. The constraint is based on

$$Z(t,\theta) = \Gamma_1(\theta)\varepsilon_1(t,\theta) + \Gamma_2(\theta)\varepsilon_2(t,\theta) - \varepsilon_3(t,\theta), \tag{80}$$

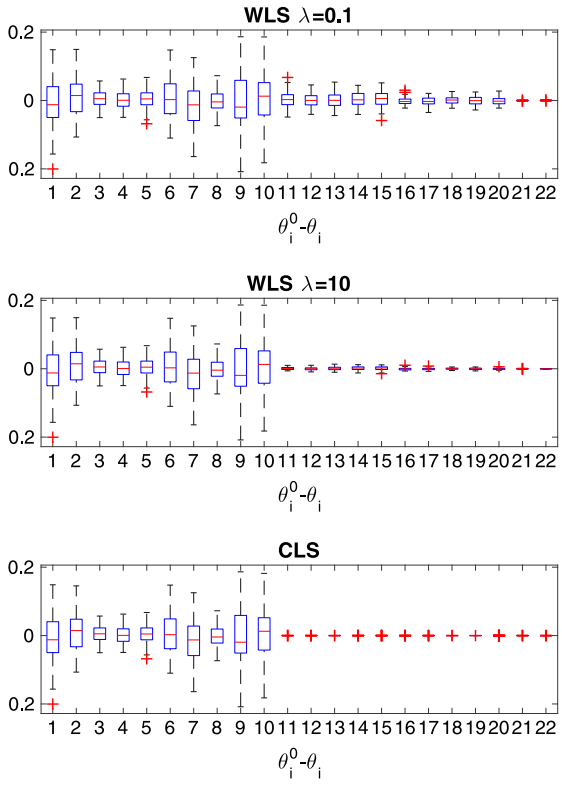

Fig. 5. Boxplot of parameter estimation errors for the 22 different parameters over 100 Monte-Carlo runs. The top and middle figures are the WLS estimates (21) with weight (41) and $\lambda = 0.1$ and $\lambda = 10$ respectively, the bottom figure is the CLS estimate (33).

and its derivative with respect to the parameters:

$$A(t) := \frac{d}{d\theta} Z(t,\theta)\Big|_{\theta=\theta^0} = \\ -\begin{bmatrix} \Gamma_1^0\phi_2 & \Gamma_1^0\phi_3 & \Gamma_2^0\phi_3 & -\phi_1 & \phi_2\theta_{12}^0 + \phi_3\theta_{13}^0 & \phi_3\theta_{23}^0 \end{bmatrix}, \tag{81}$$

where $\Gamma^0 = \begin{bmatrix} \Gamma_1^0 & \Gamma_2^0 \end{bmatrix}$.

For constructing the constraint based on finite time data, the expression $\bar{\mathbb{E}} A^T(t)A(t)$ is replaced by its sampled version, $\frac{1}{N}\bar{A}^T\bar{A}$ with $\bar{A}^T := \begin{bmatrix} A(1)^T & \cdots A(N)^T \end{bmatrix}$. The full rank matrix Π can then be constructed by taking an $(n_\theta - n_\rho) \times n_\theta$ full row rank sub-matrix of $\bar{A}$, according to

$$\Pi = \begin{bmatrix} A(1) \\ \vdots \\ A(12) \end{bmatrix}. \tag{82}$$

Because of the fact that $\Gamma_1^0 = 0$, the Π matrix is structured such that the left most 10 columns are 0. The other 12 columns constitute a 12×12 matrix of full rank. Matrix S is then defined by the right-nullspace of Π, and S has the particular structure that the first 10 rows are non-zero and form a 10×10 matrix of full rank, and the other 12 rows are 0 such that $\Pi S = 0$. When we consider $P_\theta^0 = SP_\rho^0 S^T$ and the structure of S, it is immediately observed that the lower bound on the variance of parameters 11 to 22 is 0.

The example shows a similar phenomenon as the static Example 1 in Fig. 3, i.e. two modules that map into node variables that are subject to the same disturbance, and as a result of this are estimated variance-free.

7. Conclusions

For dynamic networks with rank-reduced noise, an appropriately parameterized model combined with a weighted least squares criterion leads to consistent estimates under standard conditions. However for arriving at minimum variance and maximum likelihood results (under Gaussian disturbances), the required identification criterion becomes a weighted quadratic criterion subject to a constraint. A classical variance expression can be derived for the weighted least squares estimator, but for the constrained criterion the expressions need to be modified to appropriately deal with the constraint. For this latter situation explicit expressions for the variance have been derived, as well as expressions for the lower bound of this variance, reaching the Cramér–Rao lower bound for normally distributed noise. It has been observed and explained that parameters can be estimated variance-free. The analytical results have been illustrated with simulation examples. The results in this paper suggest that in situations where process noise is dominated by only a few white noise sources, it is beneficial to include a rank-reduced noise model, so as to reduce the variance of estimated models.

Appendix A. Proof of Proposition 1

First one predictor expression is derived using the square and monic noise model $\check{H}^0$, then it is shown that this is unique. We write the network equation (2) as

$$w = G^0 w + R^0 r + (\check{H}^0 - I)\check{e} + \check{e}.$$

Then we substitute using $He = \check{H}\check{e}$ and (2) the expression

$$\check{e} = (\check{H}^0)^{-1}[(I - G^0)w - R^0 r]$$

into the expression $(\check{H}^0 - I)\check{e}$, leading to

$$w = [I - (\check{H}^0)^{-1}(I - G^0)]w + (\check{H}^0)^{-1}R^0 r + \check{e}. \tag{A.1}$$

Since we assume that G^0 is strictly proper, $[I - (\check{H}^0)^{-1}(I - G^0)]$ is strictly proper, and evaluating the conditional expectation (14) leads to (15).

Appendix B. Proof of Proposition 3

First it will be shown that θ_0 is a minimum of the criterion, i.e. $\theta^0 \in \theta^\star$, after which it will be shown that $M(\theta_0)$ is the only minimum, i.e. $M(\theta_0) = M(\theta)\ \forall\ \theta \in \theta^\star$.

When combining (20), (18) and (2) it can be shown that the prediction error can be rewritten in terms of e and r

$$\begin{bmatrix} \varepsilon_a(\theta) \\ \varepsilon_b(\theta) \end{bmatrix} = F_e(q,\theta)e + \begin{bmatrix} I \\ \Gamma^0 \end{bmatrix} e + F_r(q,\theta)r, \tag{B.1}$$

with

$$F_e(\theta) := \check{H}^{-1}(\theta)(I - G(\theta))(I - G^0)^{-1}H^0 - \begin{bmatrix} I \\ \Gamma^0 \end{bmatrix},$$

$$F_r(\theta) := \check{H}^{-1}(\theta)\Big((I - G(\theta))(I - G^0)^{-1}R^0 - R(\theta)\Big),$$

where F_e is strictly proper since the innovation $\begin{bmatrix} I \\ \Gamma^0 \end{bmatrix} e$ has been written as a separate term.

The first term has a strictly proper filter, the innovation (second) term does not have delay, and since e is a white noise, the first 2 terms are uncorrelated with each other. By condition *2* the r term is uncorrelated with the e terms. In the quadratic function $\bar{V}(\theta)$ defined by (25) any cross-term between the 3 terms is 0 due to uncorrelatedness, therefore each of the terms can be minimized individually. Due to condition *1* the first and third terms are minimized by θ_0 and become 0. The second term does not contain

parameters, so it is trivially minimized. Then we can conclude that $\theta^0 \in \theta^\star$.

Now it will be shown that any parameter θ_1 which reaches the minimum of the cost function must result in $M(\theta_0) = M(\theta_1)$. It can be shown (Ljung, 1999 proof of Theorem 8.3) that

$$\begin{aligned} 0 &= \bar{V}(\theta_0) - \bar{V}(\theta_1) \\ &= \bar{\mathbb{E}}(\varepsilon(t,\theta_0) - \varepsilon(t,\theta_1))^T Q(\varepsilon(t,\theta_0) - \varepsilon(t,\theta_1)). \end{aligned} \tag{B.2}$$

Since $Q > 0$ we must have $\varepsilon(t,\theta_0) = \varepsilon(t,\theta_1)$, up to a possible transient term due to initial conditions, which decays to zero and therefore can be neglected in our asymptotic criterion. By condition *2*, $\begin{bmatrix} e(t) \\ r(t) \end{bmatrix}$ is a full rank process, such that

$$F_e(q,\theta_0) + \begin{bmatrix} I \\ \Gamma^0 \end{bmatrix} = F_e(q,\theta_1) + \begin{bmatrix} I \\ \Gamma^0 \end{bmatrix} \tag{B.3}$$

and

$$F_r(q,\theta_0) = F_r(q,\theta_1). \tag{B.4}$$

Since $F_e(q,\theta_0) = 0$ and $F_r(q,\theta_0) = 0$ we can write

$$\begin{bmatrix} I & 0 \\ \Gamma^0 & 0 \end{bmatrix} = \begin{bmatrix} I & 0 \\ \Gamma^0 & 0 \end{bmatrix} + \begin{bmatrix} F_e(q,\theta_1) & F_r(q,\theta_1) \end{bmatrix}. \tag{B.5}$$

When we use the expressions for F_e and F_r, then pre-multiply both sides of (B.5) with $(I - G(q,\theta_1))^{-1}\check{H}(q,\theta_1)$, and finally add $\begin{bmatrix} 0 & (I - G(q,\theta_1))^{-1}R(q,\theta_1) \end{bmatrix}$ to both sides, then

$$\begin{aligned} &(I - G(\theta_0))^{-1}\begin{bmatrix} H(\theta_0) & R(\theta_0) \end{bmatrix} = \cdot \\ &\underbrace{\cdot (I - G(\theta_1))^{-1}\begin{bmatrix} H_a(\theta_1) & R_a(\theta_1) \\ H_b(\theta_1) - \Gamma(\theta_1) + \Gamma^0 & R_b(\theta_1) \end{bmatrix}}_{:=T'(\theta_1)} \end{aligned} \tag{B.6}$$

is obtained, where R_a and R_b are defined by $R(q,\theta) = \begin{bmatrix} R_a(q,\theta) \\ R_b(q,\theta) \end{bmatrix}$. Note that $\Gamma(\theta_1)$ is the feedthrough of $H_b(\theta_1)$, such that the feedthrough of H_b is being 'replaced' with the true values Γ^0, and $\Gamma(\theta_1)$ does not appear in the equation.

In (27) we make no claims on the feedthrough of H_b, we have to show that

$$\begin{aligned} &T'(\theta_1) = T'(\theta_0) \Rightarrow \\ &\quad \{G(q,\theta^\star), H_a(q,\theta^\star), H_b(q,\theta^\star) - \Gamma(\theta^\star), R(q,\theta^\star)\} \\ &\qquad = \{G^0(q), H_a^0(q), H_b^0(q) - \Gamma^0, R^0(q).\} \end{aligned} \tag{B.7}$$

If we consider $\Theta' \in \Theta$ defined by all θ for which $\Gamma(\theta) = \Gamma^0$, then using the model set

$$\mathcal{M}' := \{M(\theta),\ \theta \in \Theta'\} \subseteq \mathcal{M},$$

we have that $T'(\theta) = T(\theta)$ for all $\theta \in \Theta'$. This means that we can apply the network identifiability reasoning to this situation. Since $\mathcal{M}'$ is a subset of $\mathcal{M}$, $\mathcal{M}'$ is globally network identifiable at $M(\theta_0)$ if $\mathcal{M}$ is globally network identifiable at $M(\theta_0)$. Using condition *3* we then have that

$$\begin{aligned} &T'(q,\theta_1) = T'(q,\theta_0) \\ &\qquad \Downarrow \\ &\begin{Bmatrix} G(q,\theta_1) & = G^0(q) \\ H_a(q,\theta_1) & = H_a^0(q) \\ H_b(q,\theta_1) - \Gamma(\theta_1) & = H_b^0(q) - \Gamma^0 \\ R(q,\theta_1) & = R^0(q). \end{Bmatrix}. \end{aligned} \tag{B.8}$$

□

Appendix C. Proof of Proposition 4

The convergence proof in Ljung (1999) needs to be adapted slightly in order to prove (35). Under the conditions in part *(1)* the cost function converges

$$\sup_{\theta\in\Theta}\left|\frac{1}{N}\sum_{t=1}^{N}\varepsilon_a^T(t,\theta)Q_a\varepsilon_a(t,\theta) - \bar{\mathbb{E}}\varepsilon_a^T(t,\theta)Q_a\varepsilon_a(t,\theta)\right| \to 0 \tag{C.1}$$

$w.p.1$ as $N \to \infty$. Similarly the constraint converges

$$\sup_{\theta\in\Theta}\left|\frac{1}{N}\sum_{t=1}^{N}Z^T(t,\theta)Z(t,\theta) - \bar{\mathbb{E}}Z^T(t,\theta)Z(t,\theta)\right| \to 0 \tag{C.2}$$

$w.p.\ 1$ as $N \to \infty$. Since the cost and constraint in (33) both converge (35) must hold.

Using the same reasoning as the proof of Proposition 3, θ_0 is a minimum of the cost function, and θ_0 satisfies the constraint. Now it is shown that $M(q,\theta_0)$ is the only model that is a minimum of the cost function that satisfies the constraint, i.e. $M(q,\theta_0) = M(\theta)\ \forall\ \theta \in \theta^*$.

It can be shown (Ljung, 1999 proof of Theorem 8.3) that

$$0 = \bar{\mathbb{E}}\varepsilon_a(t,\theta_0)^T Q_a\varepsilon_a(t,\theta_0) - \bar{\mathbb{E}}\varepsilon_a(t,\theta_1)^T Q_a\varepsilon_a(t,\theta_1) \tag{C.3}$$

if and only if

$$0 = \bar{\mathbb{E}}(\varepsilon_a(t,\theta_0) - \varepsilon_a(t,\theta_1))^T Q_a(\varepsilon_a(t,\theta_0) - \varepsilon_a(t,\theta_1)). \tag{C.4}$$

For the constraint we can use the fact that

$$Z(t,\theta_0) = \Gamma(\theta_0)\varepsilon_a(t,\theta_0) - \varepsilon_b(t,\theta_0) = 0, \quad \forall t \tag{C.5}$$

up to a possible transient term due to initial conditions that can be neglected in our asymptotic analysis. We can then rewrite the asymptotic constraint

$$0 = \bar{\mathbb{E}}Z^T(\theta_1)Z(\theta_1) \tag{C.6}$$

into the same form as (C.4)

$$0 = \bar{\mathbb{E}}(Z(\theta_0) - Z(\theta_1))^T(Z(\theta_0) - Z(\theta_1)). \tag{C.7}$$

Due to condition *(b)* and $Q_a > 0$ the predictor filters are identified from the above two equations, using the definitions of F_e and F_r from the proof of Proposition 3

$$\begin{aligned} &\begin{bmatrix} I & 0 \\ \Gamma(\theta_0) & -I \end{bmatrix}\left(F_e(\theta_0) + \begin{bmatrix} I \\ \Gamma^0 \end{bmatrix}\right) = \begin{bmatrix} I & 0 \\ \Gamma(\theta_1) & -I \end{bmatrix}\left(F_e(\theta_1) + \begin{bmatrix} I \\ \Gamma^0 \end{bmatrix}\right), \\ &\begin{bmatrix} I & 0 \\ \Gamma(\theta_0) & -I \end{bmatrix}F_r(\theta_0) = \begin{bmatrix} I & 0 \\ \Gamma(\theta_1) & -I \end{bmatrix}F_r(\theta_1). \end{aligned}$$

In these equations $F_e(\theta_0) = 0$ and $F_r(\theta_0) = 0$, such that the combination is

$$\begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} I & 0 \\ \Gamma(\theta_1) & -I \end{bmatrix}[F_e(\theta_1) + \begin{bmatrix} I \\ \Gamma^0 \end{bmatrix}\ F_r(\theta_1)]. \tag{C.8}$$

When this equation is pre-multiplied with $(I - G(\theta_1))^{-1}\check{H}(q,\theta_1)\begin{bmatrix} I & 0 \\ \Gamma(\theta_1) & -I \end{bmatrix}$ on both sides, and then $\begin{bmatrix} 0 & (I - G(\theta_1))^{-1}R(q,\theta_1) \end{bmatrix}$ is added on both sides, it is obtained that

$$T(q,\theta_0) = T(q,\theta_1), \tag{C.9}$$

By condition *(c)* the model set is globally network identifiable at θ_0 such that

$$T(\theta_0) = T(\theta_1) \Rightarrow M(\theta_0) = M(\theta_1). \quad \square \tag{C.10}$$

Appendix D. Proof of Theorem 1

First the proof of part *1* is given. The pdf of the innovation $\check{e}$ is given by 2 equations: there is the normal distribution of $e = [I \;\; 0]\check{e}$

$$f(e) = \frac{(2\pi)^{-\frac{p}{2}}}{|\Lambda|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} e^T \Lambda^{-1} e\right), \tag{D.1}$$

and

$$\begin{bmatrix}\Gamma^0 & -I\end{bmatrix}\check{e} = 0 \text{ w.p. } 1. \tag{D.2}$$

The likelihood for N datapoints is then also given by 2 equations (Khatri, 1968; Srivastava & von Rosen, 2002)

$$L_a(\theta) = \frac{(2\pi)^{-\frac{pN}{2}}}{|\Lambda(\theta)|^{\frac{N}{2}}} \exp\left(-\frac{1}{2}\varepsilon_a^T(t,\theta)\Lambda^{-1}(\theta)\varepsilon_a(t,\theta)\right), \tag{D.3}$$

and

$$\begin{bmatrix}\Gamma(\theta) & -I\end{bmatrix}\varepsilon(t,\theta) = 0 \text{ w.p. } 1 \quad \forall t. \tag{D.4}$$

Then taking the natural logarithm results in

$$\begin{aligned}\log L_a(\theta) =& c - \frac{N}{2}\log\det\Lambda(\theta) \\ &- \frac{1}{2}\sum_{t=1}^{N}\varepsilon_a^T(t,\theta)\Lambda^{-1}(\theta)\varepsilon_a(t,\theta).\end{aligned} \tag{D.5}$$

$\log L_a(\theta)$ is the criterion to be maximized combined with (D.4)

$$\begin{aligned}\theta_N^{ML} =& \arg\max_{\theta} \log L_a(\theta) \\ &\text{subject to } 0 = \varepsilon_b(t,\theta) - \Gamma(\theta)\varepsilon_a(t,\theta) \quad \forall t.\end{aligned} \tag{D.6}$$

Taking the sum of squares for each time t gives the equivalent constraint

$$\text{subject to } \frac{1}{N}\sum_{t=1}^{N} Z^T(t,\theta)Z(t,\theta) = 0, \tag{D.7}$$

with Z defined by (32).

Now part *2* is proven in a similar way as the maximum likelihood proof in Åström (1980) for full rank noise. Under the condition that $\Lambda(\theta)$ and $\varepsilon(\theta)$ do not share parameters, the cost function $\log L(\theta)$ is maximized at

$$\Lambda(\theta) = \frac{1}{N}\sum_{t=1}^{N}\varepsilon_a(t,\theta)\varepsilon_a^T(t,\theta) \tag{D.8}$$

In this maximum the constraint of (37) is satisfied. Then (D.8) is substituted into the objective of (37), and added as additional constraint, to obtain (39). $\square$

Appendix E. Proof of Lemma 2

The constraint is satisfied when $\Pi(\theta - \theta^0) = 0$ holds. When substituting (61) then we have

$$\Pi(S\rho + C - \theta^0) = 0, \tag{E.1}$$

where we have $\Pi S\rho = 0$, such that the constraint is independent of ρ. Substituting $C = \Pi^\dagger \Pi\theta^0$ then satisfies the equation. $\square$

Appendix F. Proof of Proposition 5

Proof is by substituting $\theta = S\rho - K^\dagger K^0$ into the CLS (33). Lemma 2 shows that this parameter mapping satisfies the constraint for all ρ, and thus can be removed. Equivalence of the cost function is trivial. $\square$

Appendix G. Proof of Proposition 6

With $P_\theta = \mathbb{E}(\theta^\star - \hat{\theta}_N)(\theta^\star - \hat{\theta}_N)^T$ and using the mapping (61) we get

$$P_\theta = \mathbb{E}S(\rho^\star - \hat{\rho}_N)(\rho^\star - \hat{\rho}_N)^T S^T, \tag{G.1}$$

such that $P_\theta = SP_\rho S^T$. $\square$

References

Adebayo, J., Southwick, T., Chetty, V., Yeung, E., Yuan, Y., Gonçalves, J., et al. (2012). Dynamical structure function identifiability conditions enabling signal structure reconstruction. In *51st IEEE Conf. Decision and Control (CDC)* (pp. 4635–4641). IEEE.

Åström, K. J. (1980). Maximum likelihood and prediction error methods. *Automatica*, *16*(5), 551–574.

Chiuso, A., & Pillonetto, G. (2012). A Bayesian approach to sparse dynamic network identification. *Automatica*, *48*(8), 1553–1565.

Dankers, A. G. (2014). *System identification in dynamic networks* (Ph.D dissertation), Delft University of Technology.

Dankers, A. G., Van den Hof, P. M. J., Bombois, X., & Heuberger, P. S. C. (2015). Errors-in-variables identification in dynamic networks –Consistency results for an instrumental variable approach. *Automatica*, *62*, 39–50.

Dankers, A. G., Van den Hof, P. M. J., Heuberger, P. S. C., & Bombois, X. (2016). Identification of dynamic models in complex networks with prediction error methods: predictor input selection. *IEEE Transactions on Automatic Control*, *61*(4), 937–952.

Deistler, M., Scherrer, W., & Anderson, B. D. O. (2015). The structure of generalized linear dynamic factor models. In J. Beran, Y. Feng, & H. Hebbel (Eds.), *Empirical economic and financial research* (pp. 379–400). Springer.

Everitt, N., Bottegal, G., Rojas, C. R., & Hjalmarsson, H. (2015). On the effect of noise correlation in parameter identification of simo systems. In *Proc. 17th IFAC Symposium on System Identification*. IFAC, (IFAC-PapersOnLine, 48(28): 326-331).

Felsenstein, E. (2014). *Regular and singular AR and ARMA models: the single and the mixed frequency case* (Ph.D. dissertation), Vienna University of Technology.

Gevers, M., & Bazanella, A. S. (2015). Identification in dynamic networks: identifiability and experiment design issues. In *Proc. 54th IEEE conference on decision and control (CDC)* (pp. 4005–4010). IEEE.

Gevers, M., Bazanella, A. S., & Parraga, A. (2017). On the identifiability of dynamical networks. *IFAC-PapersOnLine*, *50*(1), 10580–10585. Proc. 20th IFAC World Congress.

Gonçalves, J., & Warnick, S. (2008). Necessary and sufficient conditions for dynamical structure reconstruction of LTI Networks. *IEEE Transactions on Automatic Control*, *53*(7), 1670–1674.

Gudi, R. D., & Rawlings, J. B. (2006). Identification for decentralized model predictive control. *AIChE Journal*, *52*(6), 2198–2210.

Haber, A., & Verhaegen, M. (2014). Subspace identification of large-scale interconnected systems. *IEEE Transactions on Automatic Control*, *59*(10), 2754–2759.

Hayden, D., Chang, Y. H., Gonçalves, J., & Tomlin, C. J. (2016). Sparse network identifiability via compressed sensing. *Automatica*, *68*, 9–17.

Khatri, C. G. (1968). Some results for the singular normal multivariate regression models. *Sankhyā: The Indian Journal of Statistics, Series A*, 267–280.

Kölbl, L. (2015). *VAR systems: g-identifiability and asymptotic properties of parameter estimates for the mixed-frequency case* (Ph.D. dissertation), Vienna University of Technology.

Linder, J. (2017). *Indirect system identification for unknown input problems with applications to ships* (Ph.D. dissertation), Linköping University.

Linder, J., & Enqvist, M. (2017). Identification of systems with unknown inputs using indirect input measurements. *Internation Journal of Control*, *90*(4), 729–745.

Ljung, L. (1999). *System identification: theory for the user*. Englewood Cliffs, NJ: Prentice-Hall.

Materassi, D., & Innocenti, G. (2010). Topological identification in networks of dynamical systems. *IEEE Transactions on Automatic Control*, *55*(8), 1860–1871.
Materassi, D., & Salapaka, M. (2012). On the problem of reconstructing an unknown topology via locality properties of the wiener filter. *IEEE Transactions on Automatic Control*, *57*(7), 1765–1777.
Rao, C. R. (1973). *Linear statistical inference and its applications* (2nd ed.). London: Wiley.
Söderström, T., & Stoica, P. (1989). *System identification*. Hemel Hempstead, UK: Prentice Hall.
Srivastava, M. S., & von Rosen, D. (2002). Regression models with unknown singular covariance matrix. *Linear Algebra and its Applications*, *354*(1), 255–273.
Stoica, P., & Ng, B. C. (1998). On the Cramér-Rao bound under parametric constraints. *IEEE Signal Processing Letters*, *5*(7), 177–179.
Van den Hof, P. M. J., Dankers, A. G., Heuberger, P. S. C., & Bombois, X. (2013). Identification of dynamic models in complex networks with prediction error methods - basic methods for consistent module estimates. *Automatica*, *49*(10), 2994–3006.
Van den Hof, P. M. J., Dankers, A. G., & Weerts, H. H. M. (2017). From closed-loop identification to dynamic networks: generalization of the direct method. In *Proc. 56th IEEE Conf. Decision and Control (CDC)* (pp. 5845–5850). IEEE.
Van den Hof, P. M. J., Dankers, A. G., & Weerts, H. H. M. (2018). Identification in dynamic networks. *Computers & Chemical Engineering*, *109*, 23–29.
Van den Hof, P. M. J., Weerts, H. H. M., & Dankers, A. G. (2017). Prediction error identification with rank-reduced output noise. In *Proc. 2017 American control conference (ACC)* (pp. 382–387). IEEE.
Weerts, H. (2018). *Identifiability and identification methods for dynamic networks* (Ph.D. dissertation), Eindhoven University of Technology.
Weerts, H. H. M., Dankers, A. G., & Van den Hof, P. M. J. (2015). Identifiability in dynamic network identification. *IFAC-PapersOnLine*, *48–28*, 1409–1414. Proc. 17th IFAC Symposium on System Identification, Beijing, China.
Weerts, H. H. M., Van den Hof, P. M. J., & Dankers, A. G. (2016a). Identifiability of dynamic networks with part of the nodes noise-free. *IFAC-PapersOnLine*, *49*(13), 19–24. Proc. 12th IFAC Workshop ALCOSP, Eindhoven, the Netherlands.
Weerts, H. H. M., Van den Hof, P. M. J., & Dankers, A. G. (2016b). Identification of dynamic networks operating in the presence of algebraic loops. In *Proc. 55nd IEEE Conf. on Decision and Control (CDC)* (pp. 4606–4611). IEEE.
Weerts, H. H. M., Van den Hof, P. M. J., & Dankers, A. G. (2017). Identification of dynamic networks with rank-reduced process noise. *IFAC-PapersOnLine*, *50–1*, 10562–10567. Proc. 20th IFAC World Congress.
Weerts, H. H. M., Van den Hof, P. M. J., & Dankers, A. G. (2018). Identifiability of linear dynamic networks. *Automatica*, *89*, 247–258.
Youla, D. C. (1961). On the factorization of rational matrices. *IRE Transaction on Information Theory*, *7*, 172–189.
Yuan, Y., Stan, G. B., Warnick, S., & Gonçalves, J. (2011). Robust dynamical network structure reconstruction. *Automatica*, *47*(6), 1230–1235.

Harm H.M. Weerts was born in Bergen, The Netherlands, in 1989. He received his BSc degree from the Fontys University of Applied Sciences in Venlo, The Netherlands, in 2010, and his MSc degree in Systems & Control from Eindhoven University of Technology, The Netherlands, in 2014. Currently, he is working towards a Ph.D. degree at Eindhoven University of Technology. His Ph.D. research is on system identification applied to dynamic networks with a multivariable approach.

Paul M.J. Van den Hof received the M.Sc. and Ph.D. degrees in electrical engineering from Eindhoven University of Technology, Eindhoven, The Netherlands, in 1982 and 1989, respectively. In 1986 he moved to Delft University of Technology, where he was appointed as Full Professor in 1999. From 2003 to 2011, he was founding co-director of the Delft Center for Systems and Control (DCSC). As of 2011, he is a Full Professor in the Electrical Engineering Department, Eindhoven University of Technology. His research interests include system identification, identification for control, and model-based control and optimization, with applications in industrial process control systems, including petroleum reservoir engineering systems, and high-tech systems. He holds an ERC Advanced Research grant for a research project on identification in dynamic networks. Paul Van den Hof is an IFAC Fellow and IEEE Fellow, and Honorary Member of the Hungarian Academy of Sciences. He has been a member of the IFAC Council (1999–2005, 2017–2020), the Board of Governors of IEEE Control Systems Society (2003–2005), and an Associate Editor and Editor of Automatica (1992–2005). In the triennium 2017–2020 he is Vice-President of IFAC.

Arne G. Dankers received B.Sc. and M.Sc. degrees from the Department of Electrical and Computer Engineering at the University of Calgary in Calgary, Canada, and a Ph.D. degree from the Delft Center for Systems and Control at the Delft University of Technology in Delft, The Netherlands. He completed a Post-Doc position at the University of Calgary in partnership with Hifi Engineering Inc.. Currently he is employed full time at Hifi Engineering where he applies system identification for leak detection in pipelines. His current research interests include system identification, dynamic networks, acoustic modeling and leak detection in pipelines.