Eindhoven University of Technology

Department of Electrical Engineering,
Control Systems Group

Multi-step scalable least squares method for network identification with unknown noise topology

Master Thesis
MSc Systems and Control

Stefanie Fonken 0833058

Supervisor: prof. dr. ir. Paul Van den Hof
Daily supervisor: ir. Karthik Ramaswamy

Eindhoven, November 2020

TU/e EINDHOVEN UNIVERSITY OF TECHNOLOGY

Declaration concerning the TU/e Code of Scientific Conduct for the Master's thesis

I have read the TU/e Code of Scientific Conduct[i].

I hereby declare that my Master's thesis has been carried out in accordance with the rules of the TU/e Code of Scientific Conduct

Date 11/5/20

Name Stefanie Fonken

ID-number 0833058

Signature

Insert this document in your Master Thesis report (2nd page) and submit it on Sharepoint

[i] See: http://www.tue.nl/en/university/about-the-university/integrity/scientific-integrity/
The Netherlands Code of Conduct for Academic Practice of the VSNU can be found here also.
More information about scientific integrity is published on the websites of TU/e and VSNU

Version 202007

Multi-step scalable least squares method for network identification with unknown noise topology ⋆

Stefanie J.M. Fonken

Control Systems Group, Department of Electrical Enginnering, Eindhoven University of Technology, The Netherlands (e-mail: s.j.m.fonken@ student.tue.nl).

Abstract:
Identification methods for dynamic networks require knowledge of the network and noise topology, and often rely on a non-convex optimization problem. However, detecting the noise topology that defines the noise correlation structure and the noise rank has not been addressed in literature. In this work we address the problem of detecting the noise topology and identifying a full dynamic network with known network topology, and where the noise can be correlated and of reduced rank. To this end we extend the convex Sequential Linear Regressions and Weighted Null Space Fitting methods to deal with reduced rank noise, and use these methods to estimate the noise topology and the network dynamics. Consequently we consistently estimate dynamic networks of Box Jenkins model structure, while keeping the computational burden low. We provide the consistency proof that includes the path based data informativity conditions which indicate where excitation signals must be present to guide the experimental design. We show that the presented method obtains a smaller variance compared to the Sequential Least Squares method for networks of Box Jenkins model structure.

Keywords: System identification, dynamic networks, least-squares, noise topology detection, reduced rank noise

1. INTRODUCTION

Data driven modeling of dynamic networks has received considerable attention in recent years. Dynamic networks represent large scale interconnected systems. Modeling of these networks plays an important role in biological systems, financial systems, electrical networks, and many other fields in science and engineering. In dynamic networks the nodes are the measurement points, and are interconnected via modules that contain the dynamics. The interconnections are also referred to as edges in the network. A simple representation of a network is shown in Figure (1). The nodes in a network are generally driven by external excitation signals such as process noise and measurable signals. The challenges addressed in identification of dynamic networks can roughly be divided into three categories. The first is identifying the interconnection structure of the nodes in a dynamic network referred to as network topology detection. The second is identification of the full network dynamics, and the third is identification of a specific module in a network, referred to as local module identification. While dynamic networks increase in complexity and size, measurement data is also increasingly accessible. This has given rise to accurate and scalable data driven identification methods.

⋆ This project has received funding from the European Research Council (ERC), Advanced Research Grant SYSDYNET, under the European Unions Horizon 2020 research and innovation programme (Grant Agreement No. 694504).

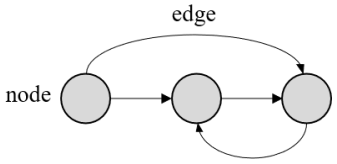


Fig. 1. Simple graph representation of a three node network, the interconnections between the nodes are the edges that contain dynamics

The classic closed loop identification methods such as the direct method (Ljung, 1999) and the indirect two-stage method (Van den Hof and Schrama, 1993) are not suitable for identification of large scale networks. These classic identification methods are generalized to a dynamic network framework in Van den Hof et al. (2013). This latter paper focuses on local module identification, where the optimization problem typically is formulated as a multi-input-single-output (MISO) problem. Local module identification is also studied in Dankers et al. (2015); Ramaswamy et al. (2018); Everitt et al. (2018) to mention a few.

In full network identification the network is formulated as a multi-input-multi-output (MIMO) system. Examples of full network identification can be found in Weerts et al. (2018b); Weerts et al. (2018); Dankers (2019); Fonken et al. (2020). The joint direct method (Weerts et al., 2018b) draws on the work of Weerts et al. (2017) and Van den Hof et al. (2017), and predicts all node signals

in the network jointly. This method generally has a non-convex optimization criterion. The Weighted Null Space fitting (WNSF) method (Galrinho et al., 2019) is a convex multi-step least squares method originally developed for single-input-single-output (SISO) systems, of which extensions to full network identification have been described in Weerts et al. (2018); Fonken et al. (2020). The Sequential Least Squares (SLS) method (Weerts et al., 2018) and the Sequential Linear Regressions (SLR) method (Dankers, 2019) are convex identification methods that split the optimization problem in smaller multi-input-single-output (MISO) problems. For large scale networks non-convex optimization problems become increasingly difficult to solve. Non-convex methods such as the joint direct method suffer from issues with local minima, where the number of local minima increases with the complexity and size of the network. The use of analytical solutions in convex methods and splitting the optimization problem contribute to a low computational burden. Convex methods such as SLS, SLR and WNSF are therefore more suitable to estimate large scale networks, especially when the optimization problem is divided over smaller more manageable optimization problems.

Available convex identification methods are generally limited to a certain model structure. For example the SLS is developed for networks of autoregressive moving average exogenous (ARMAX) model structure, and the SLR is suitable for networks where the dynamics can be represented by finite impulse response (FIR) functions. Moreover, extensions of WNSF to dynamic networks show that the method is suitable for networks of ARMAX and output error (OE) model structure. The non-convex joint direct method handles more general model structures, and is able to parametrize a Box Jenkins (BJ) network model. The BJ model structure is a more compact model structure, i.e. methods that parametrize a BJ model require less parameters to estimate and therefore obtain a reduced variance with respect to methods that parametrize a different model structure to estimate a BJ network. However, there is no convex full network identification method available that is able to parametrize a BJ model structure.

A common assumption in identification methods is that the network and noise topology are known. This assumption also holds for the local and full identification methods motioned before. Network topology detection literature shows a variety of available methods. Various network topology detection methods employ Wiener filters (Materassi and Innocenti, 2010; Materassi et al., 2011; Materassi and Salapaka, 2012), Bayesian model selection techniques (Wasserman, 2000; Chiuso and Pillonetto, 2012; Shi et al., 2019), or infer the topology from parametric estimates (Bolstad et al., 2011; Yuan et al., 2011; Dankers et al., 2012). However, noise topology detection has not been properly addressed in literature. Therefore the assumption of known noise topology in identification methods is not reasonable.

The noise topology defines, among others the interconnection or correlation structure of the process noise that disturbs the nodes. The noise spectrum is defined as $\Phi_v(\omega)$, with $v = He(t)$ where H is the noise model matrix, and vector $e(t)$ is a white noise process. The topology of H indicates where in the noise model matrix dynamics is present. If there are no noise correlations present the noise spectrum $\Phi_v(\omega)$ and noise model H are diagonal. When noise correlations are present both the noise spectrum $\Phi_v(\omega)$ and noise model H also contain off-diagonal elements.

Knowing the noise correlation structure that is derived from the noise topology, is of importance to obtain unbiased results for identification methods such as the local direct method (Van den Hof et al., 2013), that has restrictive assumptions on the noise topology. This local identification method loses consistency when noise correlations are present, which is remedied in a generalization of the direct method, the joint direct method (Weerts et al., 2018b; Van den Hof et al., 2017). Direct methods such as the joint direct method reduce variance and remove bias by modeling the noise correlations, for which the noise correlations need to be known. Furthermore, the two-stage (Van den Hof et al., 2013) and other indirect approaches that do not take the noise model into account can also provide unbiased estimates. However, using a parametric noise model can give maximum likelihood results, i.e. minimum variances can be achieved.

Another common assumption in identification methods is that the noise is full rank, where there are as many independent noise sources $e(t)$ as there are nodes. However, for large networks the assumption of full rank noise can become improbable. For example a global disturbance acting on (a part of) the network causes noise sources to be linearly dependent. If nodes are noise free or when noise sources in vector $e(t)$ are linearly dependent to each other, the noise is of reduced rank, i.e $\Phi_v(\omega)$ is singular. Moreover, the number of columns of the noise model H equals the noise rank, thus for reduced rank noise the noise model H and its noise topology are no longer a square matrix. The noise topology defines both the noise correlation structure and the noise rank. Consequently, for identification in large scale networks it is not reasonable to assume the noise is full rank.

For prediction error identification methods the reduced noise rank case has been addressed in Everitt et al. (2015); Van den Hof et al. (2017); Weerts et al. (2017, 2018b). The joint direct method can handle both correlated and reduced rank noise, whereas the available convex methods typically assume the noise is full rank. By utilizing the noise rank we can restrict the degree of freedom, that has the same effect as reducing the number of parameters to estimate, namely we obtain a reduced variance with respect to methods that do not appropriately treat reduced rank noise.

In this paper we assume we do not know the noise topology, but the network topology is known. We allow the process noise to be correlated, i.e. the noise spectrum $\Phi_v(\omega)$ is not necessarily diagonal. Additionally the noise is allowed to be of reduced rank, i.e. $\Phi_v(\omega)$ can be singular.

We aim to reduce the variance of scalable identification tools and consistently estimate large scale networks. As discussed, the available methods do not address noise topology detection and they commonly assume the noise is full rank. Knowing the noise correlations is of importance for consistency results, and appropriate modeling of reduced rank noise contributes to reducing the variance.

The joint direct method can handle both correlated and reduced rank noise, but is less suitable for large networks due to its non-convexity. The available convex identification methods are more appropriate for large scale networks, however they assume the noise is full rank and are limited to certain model structures.

To this end we consider convex algorithms that are scalable to large networks, and extend them to deal with reduced rank noise. We follow a step wise procedure where we first detect the noise topology and thereafter apply convex algorithms to identify a network of general model structure.

The paper proceeds with a definition of the dynamic network in Section 2. Section 3 provides background information on available convex methods. The developed noise topology detection method is presented in Section 4, followed by the developed identification method in Section 5. Section 6 presents the path based data informativity conditions. In Section 7 we give additional notes and compare the presented method to Sequential Least Squares, followed by the numerical results in Section 8. The conclusion and directions for future work are given in Sections 9 and 10. The consistency proofs are collected in the Appendix.

2. DYNAMIC NETWORKS

Following the setting of Van den Hof et al. (2013) a dynamic network is defined by L nodes or internal variables $w_j(t)$, $j = 1, \dots, L$, that are scalar measured signals. The underlying network is linear time invariant (LTI), and the nodes of the network can be expressed as

$$w_j(t) = \sum_{l \in \mathcal{N}_j} G^0_{jl}(q) w_l(t) + \sum_{k \in \mathcal{R}_j} R^0_{jk}(q) r_k(t) + v_j(t), \quad (1)$$

where

- q^{-1} the delay operator, i.e. $q^{-1}w_j(t) = w_j(t-1)$,
- $\mathcal{N}_j$ defines the set of indices of measured node signals w_l, $l \neq j$, for which $G^0_{jl}(q) \neq 0$, where $G^0_{jl}(q)$ is a strictly proper rational transfer function,
- $\mathcal{R}_j$ defines the set of indices of measured external excitation signals r_k, for which $R^0_{jk}(q) \neq 0$, where $R^0_{jk}(q)$ is a known proper rational transfer function,
- $v_j(t)$ is unmeasured process noise, where the noise vector $v = [v_1 \cdots v_L]$ is modeled as a stationary stochastic process represented by $v(t) = He(t)$. The $e = [e_1 \cdots e_p]$ is a white noise process of rank $p \leq L$ with covariance matrix Λ^0. $H^0(q)$ is a rational transfer function matrix that is monic and minimum-phase for $p = L$.

The full network expression, with omitted q and t, is

$$\begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_L \end{bmatrix} = \begin{bmatrix} 0 & G^0_{12} & \cdots & G^0_{1L} \\ G^0_{21} & 0 & \ddots & G^0_{2L} \\ \vdots & \ddots & \ddots & \vdots \\ G^0_{L1} & G^0_{L2} & \cdots & 0 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_L \end{bmatrix} + R^0 \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_K \end{bmatrix} + H^0 \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_p \end{bmatrix} \quad (2)$$

with the matrix notation given by

$$w = G^0 w + R^0 r + H^0 e, \quad (3a)$$

$$w = (I - G^0)^{-1}(R^0 r + H^0 e), \quad (3b)$$

where we assume the inverse $(I - G^0)^{-1}$ exists and the network is well-posed.

Since we allow the noise to be of reduced rank, the noise model is further specified according to (Weerts et al., 2018b). The noise is full rank if the noise rank $p = L$, with H^0 a square matrix. The noise is of reduced rank or singular if $p < L$, where $H^0 \in \mathbb{R}^{L \times p}$. Under the assumption that the nodes are ordered (Gevers et al., 2019; Weerts et al., 2018b) the noise disturbance $v(t)$ is represented by

$$\begin{bmatrix} v_a \\ v_b \end{bmatrix} = H^0 e = \begin{bmatrix} H^0_a \\ H^0_b \end{bmatrix} e, \quad (4)$$

where correct ordering of the nodes ensures that $v_a(t) = [v_1(t) \dots v_p(t)]^\top$ is a full rank noise process, i.e. the spectrum Φ_{v_a} is full rank. The properties of reduced rank H^0 are as follows

- $H^0 \in \mathbb{R}^{L \times p}$ is stable and has a stable left inverse $H^\dagger$, that satisfies $H^\dagger H = I \in \mathbb{R}^{p \times p}$,
- $H^0_a \in \mathbb{R}^{p \times p}$ is a monic rational transfer function matrix that is minimum-phase,
- $H^0_b \in \mathbb{R}^{(L-p) \times p}$ is a stable proper rational transfer function, where $(H^0_b - \Gamma^0)(H^0_a)^{-1}$ is stable.

The reduced rank noise can be modeled by different expressions (Weerts et al., 2016, 2017, 2018a,b). Following the factorization Lemma from Weerts et al. (2018a,b) the noise can also uniquely be modeled as

$$\begin{bmatrix} v_a \\ v_b \end{bmatrix} = \breve{H}^0 \breve{e} = \begin{bmatrix} H^0_a & 0 \\ H^0_b - \Gamma^0 & I \end{bmatrix} \begin{bmatrix} e \\ \Gamma^0 e \end{bmatrix}, \quad (5)$$

where $\breve{H}^0$ is square and monic, and where $v_a(t)$ is the full rank noise process and $v_b(t)$ contains the noise sources $\Gamma^0 e$ that are dependent to $e(t)$ through Γ^0. Moreover, Γ^0 is the direct feedthrough term of H_b, i.e. $\Gamma^0 = \lim_{z \to \infty} H^0_b(z)$. The direct feedthrough term Γ^0 is no longer present in the noise model, but instead in the covariance matrix

$$\breve{\Lambda}^0 = \begin{bmatrix} I \\ \Gamma^0 \end{bmatrix} \Lambda^0 \begin{bmatrix} I \\ \Gamma^0 \end{bmatrix}^\top = \begin{bmatrix} \Lambda^0 & \Lambda^0 \Gamma^{0\top} \\ \Gamma^0 \Lambda^0 & \Gamma^0 \Lambda^0 \Gamma^{0\top} \end{bmatrix}, \quad (6)$$

where

$$\Lambda^0 = \begin{bmatrix} \sigma^2_{e_1} & & 0 \\ & \ddots & \\ 0 & & \sigma^2_{e_p} \end{bmatrix}. \quad (7)$$

From here on we let subscript $\{a\}$ indicate the nodes subject to the full rank noise process $v_a(t)$, and subscript $\{b\}$ the nodes subject to disturbance noise $v_b(t)$ that contains the dependent noise sources $\Gamma^0 e(t)$.

In this paper we will consider networks with a BJ model structure. The rational transfer functions are defined as

$$\begin{aligned} G_{jl}(q,\theta) &= \frac{l^{jl}_1 q^{-1} + \cdots + l^{jl}_{m_l} q^{-m_l}}{1 + f^{jl}_1 q^{-1} + \cdots + f^{jl}_{m_f} q^{-m_f}}, \\ H_{jj}(q,\theta) &= \frac{1 + c^{jj}_1 q^{-1} + \cdots + c^{jj}_{m_c} q^{-m_c}}{1 + d^{jj}_1 q^{-1} + \cdots + d^{jj}_{m_d} q^{-m_d}}, \\ H_{js}(q,\theta) &= \frac{c^{js}_1 q^{-1} + \cdots + c^{js}_{m_c} q^{-m_c}}{1 + d^{js}_1 q^{-1} + \cdots + d^{js}_{m_d} q^{-m_d}}, \end{aligned} \quad (8)$$

that can be rewritten to

$$G_{jl}(q,\theta) = \frac{L_{jl}(q,\theta)}{F_{jl}(q,\theta)}, \quad H_{js}(q,\theta) = \frac{C_{js}(q,\theta)}{D_{js}(q,\theta)}. \quad (9)$$

If H^0 is reduced rank the elements of H_b^0 that contain a direct feedthrough term are also defined as $H_{jj}(q,\theta)$.

A number of assumptions is made which are clearly defined in the text. We start by assuming the data generating network satisfies the following additional properties

Assumption 1: We consider full dynamic networks that have properties

a. The network is well-posed, i.e. all principle minors of $\big(I - G^0(\infty)\big)$ are nonzero (Araki and Saeki, 1983),
b. $(I - G^0)^{-1}$ is stable and causal,
c. For $p = L$, H^0 is stable and has a stable inverse,
d. For $p < L$, $H^0 \in \mathbb{R}^{L\times p}$ satisfies the properties of the reduced rank noise model,
e. Known topology of G^0 and R^0, where R^0 is fixed and known,
f. Measurements of all node signals $w(t)$ are available,
g. External known excitation $r(t)$ is persistently exciting of a sufficiently high order,
h. Actual model orders m_i with $i = f, l, c, d$ are known.

The main objective of this paper is to present a consistent method that is suitable for parametric identification of dynamic networks of BJ structure with unknown noise topology, where process noise can be correlated and of reduced rank. Our emphasis is on keeping the computational burden low by utilizing available convex identification methods, while obtaining a reduced variance with respect to indirect methods by including noise topology detection and parametrization in the method. We use the ideas of the joint direct method to incorporate the reduced noise rank case in the convex methods. Additionally we provide path based data informativity conditions, that indicates where external excitation signals need to be allocated in the experimental design. In this paper we extend the SLR method to detect the noise topology. We combine SLR with WNSF to identify a BJ model structure consistently, where the developed method is scalable to larger networks.

The two main steps of the developed method are

- Detecting the noise topology,
- Parametric estimation of the BJ model structure.

In the next section we present background information on the SLR and WNSF methods. Our contribution will be given in the sections thereafter.

3. AVAILABLE IDENTIFICATION METHODS

The SLR and WNSF methods are convex methods that only employ analytical solutions. Both these methods assume network topology and noise topology are known. This section first provides background information on the SLR method and thereafter the WNSF method.

3.1 Sequential Linear Regressions

The SLR method (Dankers, 2019) is closely related to RLS ARMAX method (Ljung, 1999) and the SLS (Weerts et al., 2018), and is suitable to estimate full dynamic networks. The method parametrizes the G^0 and H^0 as finite impulse response (FIR) functions with a low computational burden, and is scalable to larger networks.

We consider dynamic network (1), discarding the external excitation $Rr(t)$. The SLR aims to parametrize the FIR functions of G_{jl}^0 and H_{js}^0 as

$$\begin{aligned} G_{jl}(\eta) &= \quad g_1^{jl}q^{-1} + \cdots + g_{n_g}^{jl}q^{-n_g}, \\ H_{jj}(\eta) &= 1 + h_1^{jj}q^{-1} + \cdots + h_{n_h}^{jj}q^{-n_h}, \\ H_{js}(\eta) &= \quad h_1^{js}q^{-1} + \cdots + h_{n_h}^{js}q^{-n_h}, \end{aligned} \tag{10}$$

assuming η_0 exists. The SLR method can be roughly divided in three main steps

- SLR Step 1: Initialization
- SLR Step 2: Reconstruct innovation and parametric FIR functions
- SLR Step 3: Re-estimate FIR functions

The SLR is an iterative procedure where in each iteration the estimates improve until a certain stopping criterion is reached.

We proceed to describe the SLR steps 2 and 3.

SLR Step 2: Reconstruct innovation and parametric FIR functions For this step we first define predictor

$$\hat{w}(t|t-1) := \bar{\mathbb{E}}\{w(t)|w^{t-1}\}. \tag{11}$$

and express it as

$$\hat{w}(t|t-1,\eta) = w - H(\eta)^{-1}\big(I - G(\eta)\big)w. \tag{12}$$

With predictor (12) we define the following prediction error

$$\begin{aligned} \varepsilon(t,\eta) &= w - \hat{w}(t|t-1,\eta) \\ &= H(\eta)^{-1}\big(I - G(\eta)\big)w, \end{aligned} \tag{13}$$

that is not linear in parameters η, but can be rewritten to

$$\varepsilon(t,\eta) = \big(I - G(\eta)\big)w + \big(I - H(\eta)\big)\varepsilon(t,\eta), \tag{14}$$

where $\varepsilon(t,\eta)$ now appears on both the left and right hand side of the equation. If an estimate of η, is available we obtain $\varepsilon(t,\hat{\eta}_N)$ that we can substitute in the right hand side of (14) such that

$$\varepsilon(t,\eta) = \big(I - G(\eta)\big)w + \big(I - H(\eta)\big)\varepsilon(t,\hat{\eta}_N) \tag{15}$$

is linear in the parameters η. In this setting $\varepsilon(t,\hat{\eta}_N)$ is the reconstructed innovation, and acts as an additional predictor input. The new predictor is therefore defined as

$$\hat{w}(t|t-1) := \bar{\mathbb{E}}\{w(t)|w^{t-1}, \varepsilon(\hat{\eta}_N)^{t-1}\}. \tag{16}$$

Due to the additional predictor input $\varepsilon(\hat{\eta}_N)$ the prediction error (15) is linear in parameters η. Therefore, we can split the MIMO predictor into L MISO predictors, where the parameter vector η_j for row j is obtained with least squares that minimizes the prediction error according to identification criterion

$$\hat{\eta}_{j_N} = \underset{\eta}{\operatorname{argmin}} \frac{1}{N}\sum_{t=1}^{N} \varepsilon_j(t,\eta)^2 + \eta_j^\top \Omega_j \eta_j, \tag{17}$$

where Ω is the regularization matrix that imposes stability on the estimates of η. Moreover, even if H^0 contains off-diagonal elements it is still possible to minimize the prediction error row by row, while employing analytical solutions.

Because finite data is used it is beneficial to iterate the estimation procedure. In the next step we reduce the variance by iteratively re-estimating the estimates of η.

SLR Step 3: Re-estimate FIR function The final estimates of η are obtained by iteratively re-estimating $\hat{\eta}_N$ (17), and by updating the predictor input $\varepsilon(t, \hat{\eta}_N)$ in (16) according to (15) until a certain stopping criterion is reached.

An advantage of the SLR method is that the MIMO optimization problem can be split up in L smaller more manageable MISO linear regressions, which is computationally attractive for large networks. The convexity of the method also contributes to a lower computational burden, since we do not have to worry about local minima. For the SLR method there is no theoretical proof of convergence. However, the SLR simulation results show that the estimates converge to (17).

3.2 Weighted Null Space Fitting

WNSF (Galrinho et al., 2019) is a multi-step least squares method originally developed for SISO systems. The semi-parametric method employs intermediate high order models, inspired by Durbin (1959, 1960). The asymptotic properties of the least squares method are derived using Ljung and Wahlberg (1992), where the model order $n \to \infty$ as the data length $N \to \infty$, denoted as $n(N)$. Extensions of the WNSF method to dynamic networks are available for ARMAX model structures (Weerts et al., 2018) and for OE model structures (Fonken et al., 2020). The three main steps of WNSF are

- WNSF Step 1: High order FIR or ARX model
- WNSF Step 2: Parametric model
- WNSF Step 3: Re-estimation of the parametric model

The first step of WNSF is the intermediate step that aims to obtain estimates with a negligible bias. Due to the large number of parameters that is estimated in the intermediate step, the variance will be high. WNSF Step 2 reduces the variance by reducing the number of parameters to estimate. WNSF Step 3 reduces the variance even further by re-estimating the parametric model. The method gives consistent and asymptotically efficient estimates under suitable assumptions.

For understanding the WNSF method we consider a SISO model with an OE model structure

$$y(t) = G^0(q)u(t) + e(t), \tag{18}$$

where $y(t)$ is the output, $u(t)$ the input, $e(t)$ is a white noise signal, and

$$G(q, \theta) = \frac{L(q, \theta)}{F(q, \theta)} \tag{19}$$

is a strictly proper transfer function as defined in (8), thus contains at least one delay. We continue to describe the three steps of WNSF in more detail, starting with how the intermediate model in WNSF is obtained.

WNSF Step 1: High order FIR model
We capture the impulse response of $G^0(q)$ with a FIR model $y(t) = G(\eta)u(t) + e(t)$, with $G(\eta) = \sum_{k=1}^{n} g_k q^{-k}$ of model order $n = n(N)$. We obtain the consistent least squares estimate according to

$$\hat{\eta}_N^n = \left[\frac{1}{N}\sum_{t=1}^{N}\varphi(t)\varphi^\top(t)\right]^{-1}\frac{1}{N}\sum_{t=1}^{N}\varphi(t)y(t), \tag{20}$$

where $\varphi(t) = [u(t-1) \cdots u(t-n)]^\top$. This high order estimate will have a low negligible bias when the model order n is chosen sufficiently large. However, the variance on the estimates is high.

In the next step we aim to reduce the high order model with parameter vector η to a parametric model with parameter vector θ, reducing the variance by reducing the number of parameters to estimate.

WNSF Step 2: Parametric model
To parametrize parameter vector θ we define the relation between the high order and parametric model

$$\begin{aligned} G(\eta) &= G(\theta), \\ \sum_{k=1}^{n} g_k q^{-k} &= \frac{L(\theta)}{F(\theta)}, \end{aligned} \tag{21}$$

that can be rewritten to

$$F(\theta)\sum_{k=1}^{n} g_k q^{-k} - L(\theta) = 0, \tag{22}$$

such that we can rewrite the relation linear in θ according to

$$\eta - Q(\eta)\theta = 0, \text{ with } Q(\eta) = \left[-\mathcal{T}_{n\times m}[G(\eta)]\ \bar{I}_{n\times m}\right], \tag{23}$$

where the top left corner of $\bar{I}_{n\times m}$ is $I_{m\times m}$ and has zeros otherwise, and with $\mathcal{T}_{n\times m}[G(\eta)]$ a lower triangular Toeplitz matrix, that has $[0\ g_1 \cdots g_{n-1}]^\top$ as first column. We use (23) to obtain an initial estimate of θ according to

$$\hat{\theta}_N^{[0]} = \left(Q^\top(\hat{\eta}_N^n)Q(\hat{\eta}_N^n)\right)^{-1}Q^\top(\hat{\eta}_N^n)\hat{\eta}_N^n \tag{24}$$

where estimates $\hat{\eta}_N^n$ are substituted, and $\hat{\theta}_N^{[0]}$ is consistent.

In the next step we reduce the variance further by re-estimating the parametric model.

WNSF Step 3: Re-estimation of parametric model
For the final WNSF step we revisit relation (23), that is used in the second WNSF step to initially obtain estimates of θ. However, (23) does not equal zero when the estimates $\hat{\eta}_N^n$ are substituted. Therefore the final step of WNSF will re-estimate the model, by correcting for

$$\hat{\eta}_N^n - Q(\hat{\eta}_N^n)\theta = T(\theta)(\hat{\eta}_N^n - \eta_0^n), \tag{25}$$

with $T(\theta)$ a matrix with the denominator polynomial as entry

$$T(\theta) = \mathcal{T}_{n\times n}[F(\eta)], \tag{26}$$

that is a lower triangular matrix where the first column is $[1\ f_1 \cdots f_m\ 0_{n-m-1}]^\top$ with $F(\theta) = 1 + \sum_{k=1}^{n} f_k q^{-k}$.

Although the error $(\hat{\eta}_N^n - \eta_0^n)$ is unknown, we know its distribution is approximately zero mean with estimated covariance

$$P = \hat{\sigma}_e^2\Big[\frac{1}{N}\sum_{t=1}^{N}\varphi(t)\varphi^\top(t)\Big]^{-1}, \tag{27}$$

where $\varphi(t)$ contains input signals $u(t)$. Then for a sufficiently large model order n the covariance of (25) is denoted as $T(\theta)PT^\top(\theta)$. This covariance is used as weighting for the re-estimation of the parameter vector θ, where $\hat{\sigma}_e^2$ is omitted because the scalar value does not contribute to the weighting. Moreover, because the data is finite it makes sense to iteratively re-estimate parameter vector θ

and improve the estimation results, according to weighted least squares

$$\hat{\theta}_N^{[k+1]} = \left(Q^\top(\hat{\eta}_N^n)W(\hat{\theta}_N^{[k]})Q(\hat{\eta}_N^n)\right)^{-1}Q^\top(\hat{\eta}_N^n)W(\hat{\theta}_N^{[k]})\hat{\eta}_N^n, \tag{28}$$

that is an analytical solution, and where

$$W(\theta^{[k]}) = T^{-\top}(\theta^{[k]})P^{-1}T^{-1}(\theta^{[k]}). \tag{29}$$

For iteration $k \geq 1$, the estimate $\hat{\theta}_N^{[k]}$ is asymptotically efficient, i.e. it reaches minimum variance for sufficiently large data lengths N. The analytical properties of the WNSF are proven in Galrinho et al. (2019).

In the next section we continue with the developed noise topology detection approach. Here we describe how the SLR approach is extended to detect the noise topology.

4. NOISE TOPOLOGY DETECTION

By including the noise model parametrization in the identification method, we reduce the variance compared to the two-stage and other indirect methods that do not include the noise model. If the noise correlation structure is known we can reduce the number of free parameters to estimate and obtain a reduced variance compared to unstructured identification methods. In addition, appropriate modeling of the reduced rank noise contributes to reducing the variance. Moreover, appropriate modeling of the noise correlations and noise rank removes bias on the estimates, for which the noise topology needs to be known.

The noise topology of noise model $H^0 \in \mathbb{R}^{L\times p}$ defines the noise correlations and the noise rank. We view the noise topology as a boolean matrix, where its matrix element ji is 1 if the corresponding element in the noise model H_{ji}^0 contains dynamics, and a 0 otherwise. Off-diagonal elements in the noise topology indicate the noise correlation structure, and the noise rank p defines the number of columns of the noise topology.

To estimate the noise topology we turn to methods available for network topology detection. Since we are interested in eventually obtaining the parametric estimates of θ it makes sense to apply network topology detection techniques on parametric estimates of the noise model. In network topology detection methods employ model selection techniques such as AIC and BIC (Yuan et al., 2011) on parametric estimates, or use sparsity inducing regularization techniques such as lasso and group lasso (Glasso) (Yuan and Lin, 2006; Friedman et al., 2010; Bolstad et al., 2011). We can obtain the parametric noise model using the SLR method that parametrizes the FIR functions of G^0 and H^0 separately. With a known network topology that defines the interconnection structure of G^0, we can infer the noise topology from the noise model estimates $H(\eta)$ obtained with analytical solutions. We will apply AIC, BIC, CV and Glasso on the parametric estimates of the noise model $H(\eta)$ obtained with the SLR approach.

To describe the developed method we focus on the reduced noise rank case $p < L$ with $H^0 \in \mathbb{R}^{L\times p}$ (4) or $\breve{H}^0 \in \mathbb{R}^{L\times L}$ (5). The full rank noise case $H^0 \in \mathbb{R}^{L\times L}$ will be considered as a special case. For the full noise rank case $p = L$ the noise model $H^0 \in \mathbb{R}^{L\times L} = \breve{H}^0 \in \mathbb{R}^{L\times L}$. To be more specific $H^0 \in \mathbb{R}^{L\times L} = H_a^0 \in \mathbb{R}^{p\times p}$. Thus for the full rank noise the same steps can be followed as for the reduced rank case, where the nodes w_b are discarded for $p = L$.

We start by assuming the nodes are un-ordered, where ordering refers to the order in which the nodes $w_j(t)$ are labeled with j. We define the un-ordered network with accent ($\tilde{}$). Incorrect ordering causes the noise process $\tilde{v}_a(t) = \tilde{H}_a^0\tilde{e}_a(t)$ to lose rank because $\tilde{e}_a(t)$ contains linearly dependent noise signals. If ordering is in place we ensure H_a^0 is a monic matrix that is driven by full rank noise $e(t) = [e_1(t)\ \cdots\ e_p(t)]^\top$. Note that ordering is not required for the full rank case $p = L$ because in this situation there are no linearly dependent noises present. The noise rank p gives us insight on matrix $\breve{H}$ (5), i.e. we know the last $L - p$ columns of $\breve{H}$ are $[0\ I]^\top$ and $\breve{H}$ is monic if the nodes are ordered such that $v_a(t)$ is a full rank noise process.

The topology detection is a multi-step procedure, that consists of the following main steps:

- Step 1: Obtain noise rank p and ordering of nodes.
- Step 2: Reconstruct the innovation.
- Step 3: Obtain noise correlation structure.

We fist estimate the noise rank p and order the nodes such that we know $\breve{H}$ equals the expression given in (5), where the last $L-p$ columns are known. With this information we consistently reconstruct the innovation in Step 2 according to SLR Step 2. In step 3 we estimate parameter vector η, and employ model selection techniques AIC, BIC and CV, and Glasso to estimate the noise correlation structure.

We avoid measures such as regularization by following the approach of Ljung and Wahlberg (1992) throughout the noise topology detection and parametric identification of the given network, and from here on we consider $n = n(N)$ i.e. the model order n increases as the data length N increases.

We proceed with presenting the three steps of the developed topology detection method.

4.1 Step 1: Noise rank p and ordering of nodes

With the noise model unknown, the first objective is to obtain the noise rank p and the ordering of the nodes such that $v_a(t)$ is a full rank process. The un-ordered covariance of $\breve{\Lambda}^0$ is defined as $\tilde{\Lambda}^0$, and $\tilde{H}^0$ is defined as the un-ordered reduced rank noise model $\breve{H}^0 \in \mathbb{R}^{L\times L}$. According to the reasoning of Weerts et al. (2018a) the data generating $\tilde{\Lambda}^0$ contains the information on rank p and the ordering information for the nodes under appropriate assumptions, i.e. the rank $\tilde{\Lambda}^0 = p$. Moreover, there exists a permutation matrix such that $[I_p\ 0]\,\Pi^\top\tilde{\Lambda}^0\Pi\,[I_p\ 0]^\top = \Lambda^0$, with $\Lambda^0 \in \mathbb{R}^{p\times p}$. The permutation matrix Π orders the nodes of the network such that H_a is monic, and is subject to the full rank noise $e = [e_1\ \dots\ e_p]^\top$.

We obtain the un-ordered covariance matrix $\tilde{\Lambda}$ using a high order ARX model. We extend the predictor definition in (11) to include external excitation $r(t)$

$$\hat{w}(t|t-1) := \bar{\mathbb{E}}\{w(t)|w^{t-1}, r^t\}, \tag{30}$$

with the predictor inputs $w(t)$ and $r(t)$ we can model the dynamic network (2) in MIMO setting according to

$$\hat{w}(t|t-1,\zeta) = \big(I - \tilde{A}(\zeta)\big)\tilde{w}(t) + \tilde{B}(\zeta)\tilde{r}(t), \tag{31}$$

with the predictor filters

$$\begin{aligned}\tilde{A}(\zeta) &= \tilde{H}^{-1}\big(I - \tilde{G}\big)\\ \tilde{B}(\zeta) &= \tilde{H}^{-1}\tilde{R}.\end{aligned} \tag{32}$$

Because the noise rank p is unknown, we choose $\tilde{H}^{-1}$ in the predictor model (32) to be a full matrix. Therefore $\tilde{A}(\zeta) \in \mathbb{R}^{L\times L}$ and $\tilde{B}(\zeta) \in \mathbb{R}^{L\times K}$ are fully parametrized, and can be seen as an unstructured model obtained with an indirect approach. Moreover, the parameter vector ζ contains the polynomial coefficients of the ARX model, where the elements of the $\tilde{A}(\zeta)$ and $\tilde{B}(\zeta)$ matrices are polynomials of model order n. This step is analogous to among others, the first step in SLS (Weerts et al., 2018), the joint direct method (Weerts et al., 2018b), and the initialization procedure of SLR (Dankers, 2019).

Because the ARX model parametrizes the inverse of $\tilde{H}$ in its $\tilde{A}(\zeta)$ and $\tilde{B}(\zeta)$ matrix, it is possible to split the MIMO setting into a MISO one, giving

$$\hat{w}_j(t|t-1,\zeta) = \big(I_j - \tilde{A}_j(\zeta)\big)\tilde{w}(t) + \tilde{B}_j(\zeta)\tilde{r}(t) \tag{33}$$

where $\tilde{A}_j(\zeta)$ and $\tilde{B}_j(\zeta)$ are the fully parametrized rows of matrices $\tilde{A}(\zeta)$ and $\tilde{B}(\zeta)$ belonging to node j.

We identify the ARX model using identification criterion

$$\hat{\zeta}^n_{j_N} = \underset{\zeta}{\operatorname{argmin}} \frac{1}{N}\sum_{t=1}^{N} \varepsilon_j(t,\zeta)^2, \tag{34}$$

where the prediction error is defined by

$$\varepsilon_j(t,\zeta) = w_j(t) - \hat{w}_j(t|t-1,\zeta). \tag{35}$$

The parameter vector ζ is estimated with

$$\hat{\zeta}^n_{j_N} = \left[\frac{1}{N}\sum_{t=1}^{N}\varphi_j(t)\varphi_j^\top(t)\right]^{-1}\frac{1}{N}\sum_{t=1}^{N}\varphi_j(t)w_j(t) \tag{36}$$

that is an analytical solution, and where $\varphi_j(t)$ contains signals $\tilde{w}(t)$ and $\tilde{r}(t)$ that serve as inputs for the predictor (33). With the estimates of ζ available we can derive an estimate of $\tilde{\Lambda}^0$ as

$$\tilde{\Lambda} = \frac{1}{N}\sum_{t=1}^{N}\tilde{\varepsilon}(t,\hat{\zeta}^n_N)\tilde{\varepsilon}^\top(t,\hat{\zeta}^n_N), \tag{37}$$

since the data generating system has a strictly proper $\tilde{G}^0$ matrix and matrix $\tilde{H}^0$ has monicity properties. Consequently the rank $\tilde{\Lambda}$ gives an estimate of the noise rank p.

The permutation matrix Π for ordering of the nodes can be derived from $\tilde{\Lambda}$, assuming the noise variances $\sigma^2_{e_j}$ have unique values for $j = 1,\cdots,p$. Additionally we assume the direct feedthrough matrix Γ^0 contains at most one 1 element per row. We inspect the diagonal elements of $\tilde{\Lambda}$, in which we determine the p unique noise variances $\sigma^2_{e_j}$ and order them by choice. We let permutation matrix Π order $\tilde{\Lambda}$ such that the first p diagonal entries equal vector $[\sigma^2_{e_1},\ldots,\sigma^2_{e_p}]$. Permutation matrix Π is used to order all nodes, and is not necessarily unique. With the nodes ordered $\tilde{\Lambda}$ (37) is an estimate of $\breve{\Lambda}^0$ (6).

To estimate the noise rank p and the permutation matrix Π we do not require the estimates of ζ to be consistent. We derive the noise rank from the covariance matrix (37), regardless of whether the estimates $\hat{\zeta}^n_N$ are of ζ_0 or of ζ_1 obtained with a different non-unique model.

In the next step we use estimates $\hat{\zeta}^n_N$ to reconstruct the innovation signal that will be used as a predictor input following SLR Step 2. To ensure we obtain consistent estimates in the following steps we require the innovation signals to be consistent. Therefore we repeat the estimation procedure and derive the conditions under which estimates $\hat{\zeta}^n_N$ and consequently the reconstructed innovation are consistent.

4.2 Step 2: Reconstruct the innovation

With the noise rank p available and the nodes ordered we gained additional information on the $\breve{H}^0$ (5), namely we know the last $L-p$ columns in $\breve{H}$ are $[0\ I]^\top$. Moreover, taking the inverse of $\breve{H}$ does not affect the last $L-p$ columns

$$\breve{H}^{-1} = \begin{bmatrix} H_a^{-1} & 0\\ -\big(H_b - \Gamma\big)H_a^{-1} & I\end{bmatrix}. \tag{38}$$

To utilize the known zeros in the $\breve{H}^{-1}$ structure, we assume the known R^0 matrix satisfies the following assumption.

Assumption 2: We consider the known external excitation signals $r(t)$ to drive a known R^0 matrix with the following restricted structure

$$R^0 r = \begin{bmatrix} R^0_a & 0\\ 0 & R^0_b\end{bmatrix}\begin{bmatrix} r_a\\ r_b\end{bmatrix}, \text{ with } R^0_a \in \mathbb{R}^{L\times K_a}, R^0_b \in \mathbb{R}^{L\times K_b}, \tag{39}$$

such that $r_a(t)$ drive nodes $w_a(t)$, and $r_b(t)$ drive nodes $w_b(t)$.

Considering dynamic networks that satisfy Assumption 2, we know

$$\begin{aligned}&\breve{A}^0 = (\breve{H}^0)^{-1}(I - G^0),\\ &\breve{B}^0 = (\breve{H}^0)^{-1}R^0 = \begin{bmatrix} (H^0_a)^{-1}R^0_a & 0\\ -\big(H^0_b - \Gamma^0\big)(H^0_a)^{-1}R^0_a & R^0_b\end{bmatrix},\end{aligned} \tag{40}$$

where R^0_b in the lower right corner of $\breve{B}^0$ is known and we therefore do not need to parametrize the last K_b columns of $\breve{B}^0$. Thus we can define the following predictors per node

$$\hat{w}_{a,j}(t|t-1,\zeta) = \big(I_j - \breve{A}_j(\zeta)\big)w + \breve{B}_j(\zeta)r_a, \tag{41}$$

and

$$\hat{w}_{b,j}(t|t-1,\zeta) = \big(I_j - \breve{A}_j(\zeta)\big)w + \breve{B}_j(\zeta)r_a - R_{b,j}r_b, \tag{42}$$

where $\breve{A}_j(\zeta) \in \mathbb{R}^{1\times L}$ is the fully parametrized row j of matrix

$$\breve{A}(\zeta) = \breve{H}^{-1}(I - G), \tag{43}$$

and $\breve{B}_j(\zeta) \in \mathbb{R}^{1\times K_a}$ is the fully parametrized row j of matrix

$$\breve{B}(\zeta) = \breve{H}^{-1}\begin{bmatrix} R_a\\ 0\end{bmatrix}, \tag{44}$$

where the last K_b columns of $\breve{B}^0(q)$ are not included in the parametrization. We repeat the procedure in Step 1 with the newly defined predictors (41) (42), and obtain consistent estimates $\hat{\zeta}^n_{j_N}$ using analytical solution (36), where φ_j contains the predictor inputs $w(t)$ and $r_a(t)$.

The external signals r_b are not used to parametrize the ARX model, they provide excitation to the network. The conditions for consistency are formulated in Proposition 1 and the proof is added in the appendix.

Proposition 1: Consistency $\hat{\zeta}_N^n$
Consider a dynamic network that satisfies Assumption 1 for noise rank $p \leq L$ and Assumption 2 for $p < L$. Additionally, consider the prediction error identification criterion (34) with predictors (41) and (42). Then the transfer function matrices $\breve{A}_j^0(q)$ and $\breve{B}_j^0(q)$ (40) are consistently estimated with the analytical solution (36), if the following conditions hold:

(1) The external excitation $r(t)$ is uncorrelated to the noise $e(t)$,
(2) The spectral density of $\kappa = [r_a(t)\ w(t)]^\top$, $\Phi_\kappa(\omega) > 0$ for a sufficiently high number of frequencies ω,
(3) The data generating system is in the model set, i.e. there exists a ζ_0 such that $\breve{A}_j(q, \zeta_0) = \breve{A}_j^0(q)$, and $\breve{B}_j(q, \zeta_0) = \breve{B}_j^0(q) \in \mathbb{R}^{1 \times K_a}$ where the last K_b columns of $\breve{B}_j^0(q)$ are not parametrized.

Corollary 1 Condition (1) and (2) of Proposition 1 are given for all signals present in the network. These conditions remain unchanged when we convert from MISO predictors to one MIMO predictor. Therefore the proof also holds for a MIMO predictor.

Proof: See appendix.

Weerts et al. (2018b) shows that the ordered prediction error vector $\breve{\varepsilon}(t, \zeta)$ has the same dependencies as the noise, and we can split the vector according to

$$\breve{\varepsilon}(t, \zeta) = \begin{bmatrix} \varepsilon_a(t, \zeta) \\ \varepsilon_b(t, \zeta) \end{bmatrix} = \begin{bmatrix} w_a(t) - \hat{w}_a(t|t-1, \zeta) \\ w_b(t) - \hat{w}_b(t|t-1, \zeta) \end{bmatrix}. \tag{45}$$

For $\zeta = \zeta_0$, assuming ζ_0 exists, we can state

$$\varepsilon_a(t, \zeta_0) = e(t) \quad \text{and} \quad \varepsilon_b(t, \zeta_0) = \Gamma^0 e(t), \tag{46}$$

such that $\Gamma^0 \varepsilon_a(t, \zeta_0) = \varepsilon_b(t, \zeta_0)$. The estimate of Γ^0 can be derived using

$$\hat{\Gamma}_N^n = \Big(\frac{1}{N}\sum_{t=1}^{N} \varepsilon_b(\hat{\zeta}_N^n)\varepsilon_a^\top(\hat{\zeta}_N^n)\Big)\Big(\frac{1}{N}\sum_{t=1}^{N} \varepsilon_a(\hat{\zeta}_N^n)\varepsilon_a^\top(\hat{\zeta}_N^n)\Big)^{-1}. \tag{47}$$

Because $\hat{\zeta}_N^n$ is consistent, the estimate $\hat{\Gamma}_N^n$ will converge to

$$\Gamma^0 \Lambda^0 (\Lambda^0)^{-1} = \Gamma^0. \tag{48}$$

By computing the prediction error with the consistent estimate $\hat{\zeta}_N^n$ per node for each time instance $t = 1, \ldots, N$, the prediction error vector is an estimate of the noise $\breve{e}(t) = \begin{bmatrix} e(t)\ \Gamma^0 e(t) \end{bmatrix}^\top$. From here on we will refer to the prediction error with parametric estimates substituted as the reconstructed innovation, where the consistent $\hat{\zeta}_N^n$ gives

$$\begin{aligned} \varepsilon_a(t, \hat{\zeta}_N^n) &\to e(t) \qquad \text{w.p. 1 as } N \to \infty\, \forall t, \\ \varepsilon_b(t, \hat{\zeta}_N^n) &\to \Gamma^0 e(t) \quad \text{w.p. 1 as } N \to \infty\, \forall t. \end{aligned} \tag{49}$$

At this point we know the direct feedthrough terms of H_a^0 and H_b^0 if ordering of the nodes is in place. We additionally know the sizes of H_a^0 and H_b^0, where the number of columns equals the noise rank p. Therefore the next step will focus on obtaining the correlation structure, i.e. the off diagonal elements present in $\breve{H}^0$ that contain noise dynamics.

In the next step we use the reconstructed innovation signal as an additional predictor input following the SLR step 2. By doing so we are able to convexly estimate parameter vector η that parametrizes the FIR functions of G^0 and H^0. We apply model order selection techniques such as AIC, BIC, CV and GLasso to the parametric noise model $H(\eta)$ to estimate the noise correlation structure.

4.3 Step 3: Noise correlation structure

In this step we apply model order selection techniques to the estimate of parameter vector η to infer the noise correlation structure. We follow the SLR Step 2 to estimate parameter vector η, that uses the reconstructed innovation as an additional input. Due to this additional predictor input the estimate of η is obtained with an analytical solution. Therefore the noise topology detection is a convex procedure.

To estimate η we define the new predictor according to

$$\hat{w}(t|t-1) := \bar{\mathbb{E}}\{w(t)|w^{t-1}, \breve{\varepsilon}(\hat{\zeta}_N^n)^{t-1}\}. \tag{50}$$

With predictor inputs $w(t)$ and $\breve{\varepsilon}(\hat{\zeta}_N^n)$ we can model the network (2), again using a ARX model, according to

$$\hat{w}(t|t-1, \eta) = \big(I - A(\eta)\big)w(t) + \breve{B}(\eta)\breve{\varepsilon}(\hat{\zeta}_N^n) + Rr(t), \tag{51}$$

where

$$\breve{\varepsilon}(t, \hat{\zeta}_N^n) = \begin{bmatrix} \varepsilon_a(t, \hat{\zeta}_N^n) \\ \varepsilon_b(t, \hat{\zeta}_N^n) \end{bmatrix} \tag{52}$$

acts as additional input and

$$\begin{aligned} A(\eta) &= I - G(\eta), \\ \breve{B}(\eta) &= \breve{H}(\eta) - I. \end{aligned} \tag{53}$$

We can rewrite the predictor per node as

$$\begin{aligned} &\hat{w}_j(t|t-1, \eta) = \\ &\sum_{l \in \mathcal{N}_j} G_{jl}(\eta) w_l + \sum_{s \in \mathcal{V}_j} \big(\breve{H}_{js}(\eta) - I_{js}\big)\breve{\varepsilon}_s(\hat{\zeta}_{j_N}) + \sum_{k \in \mathcal{R}_j} R r_k, \end{aligned} \tag{54}$$

where η contains the coefficients of the impulse responses of G_{jl} and the noise model of model order n. The set $\mathcal{V}_j$ defines the set of indices of noise signals $\breve{\varepsilon}_s(\hat{\zeta}_N^n)$ for which $H_{js}^0 \neq 0$, i.e. the set indicates $\mathcal{V}_j$ which elements in the noise topology are 1. However, the complete set $\mathcal{V}_j$ is still unknown.

Important to note is that

$$\begin{aligned} \big(\breve{H}(\eta) - I\big)\breve{\varepsilon}(\hat{\zeta}_N^n) &= \Big(\begin{bmatrix} H_a(\eta) & 0 \\ H_b(\eta) - \Gamma & I \end{bmatrix} - I\Big)\breve{\varepsilon}(\hat{\zeta}_N^n), \\ &= \begin{bmatrix} H_a(\eta) - I \\ H_b(\eta) - \Gamma \end{bmatrix} \varepsilon_a(\hat{\zeta}_N^n), \end{aligned} \tag{55}$$

and because $\varepsilon_a(\hat{\zeta}_N^n)$ converges to the noise $e(t)$, the set $\mathcal{V}_j$ contains at most p indices. Moreover, this indicates the predictor (50) is actually defined as

$$\hat{w}(t|t-1) := \bar{\mathbb{E}}\{w(t)|w^{t-1}, \varepsilon_a(\hat{\zeta}_N^n)^{t-1}\}. \tag{56}$$

We identify the ARX model according to identification criterion

$$\hat{\eta}_{j_N}^n = \underset{\eta}{\operatorname{argmin}} \frac{1}{N}\sum_{t=1}^{N} \varepsilon_j(t, \zeta)^2, \tag{57}$$

where $\varepsilon_j(t,\zeta) = w_j(t) - \hat{w}_j(t|t-1,\eta)$. With the reconstructed innovation as additional predictor input, the dynamics of $G_{jl}(\eta)$ and $\breve{H}_{js}(\eta) - I_{js}$ are parametrized separately. With the known network topology for G^0 we find the noise correlation structure in the estimates $\breve{H}(\eta) - I$.

We obtain estimates of η using analytical solution (36) by replacing ζ with η, and where $\varphi_j(t)$ contains predictor inputs $w(t)$ and $\varepsilon_a(t, \hat{\eta}_N^n)$ of predictor (54).

Based on the estimate $\hat{\eta}_{N_j}^n$ we estimate the noise correlation structure. To this end we test the predictor (54), with possible combinations in set $\mathcal{V}_j$ and employ model selection techniques (Yuan et al., 2011), referred to as structure selection. Because we use convex ARX models to estimate η, model selection techniques such as AIC, BIC and Cross-validation (CV) are also convex. AIC and BIC selects a set $\mathcal{V}_j$ that has the least information loss, where there is a trade off between how well the model fits and the model complexity. CV selects a set $\mathcal{V}_j$ that gives the lowest root mean squared error (RMSE). Because we derive the noise topology per node, we have to test at most 2^L possible sets $\mathcal{V}_j$ for L nodes. This results in a lower computation burden compared to when we detect the topology in a MIMO setting, where we would have to test at most 2^{L^2-L} possible sets $\mathcal{V}$ (Yuan et al., 2011). Utilizing the known direct feedthrough terms reduces the number of sets $\mathcal{V}_j$ to test. However for large networks these model selection techniques can still become computationally heavy.

The Glasso is a convex extension to lasso where the l_1 norm regularization in lasso is replaced by a l_2 norm regularization such that groups of parameters are penalized. If we fully parametrize the noise model, this allows for penalization on the estimated impulse responses of the noise model. With correct penalization only dynamics that is actually present in the data generating network remains, i.e. the estimated FIR transfer function $H_{js}(\eta)$ is either zero or not. The Glasso estimates are computed using

$$\frac{1}{2}||w_j^N - \hat{w}_j^N||_2^2 + \lambda_j \sum_{i=1}^{\mathrm{C}\{\mathcal{N}_j\}+p} \sqrt{\eta_{ji}^\top I_n \eta_{ji}} \qquad (58)$$

where superscript N indicates we use all data samples, such that $w_j^N = [w_j(1), \ldots, w_j(N)]^\top$, and where $\eta_{ji} = \left[\eta_1^{ji}, \ldots, \eta_n^{ji}\right]^\top$, I_n has dimension $(n \times n)$, $\mathrm{C}\{\mathcal{N}_j\}$ is $\mathrm{Cardinal}\{\mathcal{N}_j\}$, and λ_j is the tuning parameter of Glasso. The tuning of λ_j is described in the numerical illustrations in Section 8.

The noise topology consists of the noise rank (dimension), and if ordering is in place we can estimate the noise correlation structure (off-diagonal elements). The noise topology is estimated with an entirely convex method.

The next section describes the developed identification method, where we utilize the estimated noise topology and obtain consistent estimates of parameter vector η. Here we combine the SLR with WNSF such that we are able to estimate the network parameters with reduced variance.

5. DEVELOPED IDENTIFICATION METHOD

The developed identification procedure draws inspiration from several papers of Dankers, Weerts, Galrinho, Ljung and colleagues. The method is best described as a combination of SLR (Dankers, 2019) and WNSF (Galrinho et al., 2019). For a majority of identification methods parametrizing the off-diagonal elements in the noise model leads to a non-convex optimization criterion. The SLR method however, models the FIR functions of G^0 and H^0 separately, where the optimization remains convex even if correlated noise is present. The WNSF method is a convex method that estimates the rational functions belonging to high order models by using the said high order model as an intermediate step. The FIR functions obtained with the SLR method are similar to the intermediate model for WNSF, making the methods easy to combine. By doing so, we are able to parametrize the BJ model structure, and obtain a reduced variance compared to methods that employ other model structures to describe the BJ network, such as ARMAX and FIR models.

The identification method is a step wise procedure that continues where the noise topology detection method stops. We therefore continue the numbering of the steps, where the identification steps are

- Step 4: Update reconstructed innovation
- Step 5: Parametric model for reduced variance.

In Step 4 we repeat Step 3 of the noise topology detection, where we now incorporate the estimated noise topology. We again follow SLR Step 2 where we parametrize the FIR functions $G(\eta)$ and $H(\eta)$. In step 5 we follow the WNSF method and obtain parametric models $G(\theta)$ and $H(\theta)$ that describe the network of BJ model structure.

Note that the Steps 1 to 4 all use high order models, i.e. a large number of parameters is estimated. The final step, Step 5 focuses on obtaining the desired parametric model for reduced variance results.

We proceed with a description of the identification steps, starting with Step 4 that updates the reconstructed innovation. In order for the reconstructed innovation to be consistent we require the estimate of η to be consistent. This updated innovation is then used in the next step to obtain a parametric model where we reduce the variance.

5.1 Step 4: Update reconstructed innovation

By repeating the parametrization in Step 3 of the noise topology detection procedure while fixing the estimated noise topology in predictor (54), we obtain consistent estimate $\hat{\eta}_N^n$ with (36) where ζ is replaced with η and $\varphi_j(t)$ contains predictor inputs $w(t)$ and $\varepsilon_a(\hat{\zeta}_N^n)$. The conditions for the consistency are formulated in Proposition 2 and the proof is given in the appendix.

Proposition 2: Consistency $\hat{\eta}_N^n$
Consider a dynamic network that satisfies Assumption 1, and assume the topology of $H^0(q)$ is estimated correctly. Additionally, consider the prediction error identification criterion (57) with predictor (54) for all j. Then the transfer function matrices of $G^0(q)$ and $H^0(q)$ are consistently estimated with the analytical solution (36), if the following conditions hold:

(1) The external excitation $r(t)$ is uncorrelated to the noise $e(t)$,

(2) The spectral density of $\bar{\kappa} = \left[w_{\{\mathcal{N}_j\}}(t)\ e_{\{\mathcal{V}_j\}}(t)\right]^\top$, $\Phi_{\bar{\kappa}}(\omega) > 0$ for all j and for a sufficiently high number of frequencies ω,
(3) The data generating system is in the model set, i.e. there exists a η_0 such that $G(q,\eta_0) = G^0(q)$ and $H(q,\eta_0) = H^0(q)$.

Proof: See appendix.

With consistent estimate $\hat{\eta}_N^n$ we can update the reconstructed innovation $\breve{\varepsilon}(t,\hat{\eta}_N^n) = [\varepsilon_a(t,\hat{\eta}_N^n)\ \varepsilon_b(t,\hat{\eta}_N^n)]^\top$ consistently for each time step t=1,...,N

$$\breve{\varepsilon}(t,\hat{\eta}_N^n) \to \breve{e}(t) \qquad \text{w.p. 1 as } N \to \infty\, \forall t, \tag{59}$$

where $\breve{e} = \left[e(t)\ \Gamma^0 e(t)\right]^\top$, and the innovation is reconstructed per node using predictor (54) defined according to (56) in

$$\breve{\varepsilon}_j(t,\eta) = w_j(t) - \hat{w}_j(t|t-1,\eta). \tag{60}$$

In the next step we will go from a high order model that parametrizes the FIR functions of G^0 and H^0, to a parametric model that describes the BJ model structure of the data generating network according to (9).

5.2 Step 5: Parametric model for reduced variance

In this step we use the WNSF approach to obtain parameter vector θ that describes the rational transfer functions of G^0_{jl} and H^0_{js} (8). We continue in the MISO setting where we expand the SISO WNSF for OE models, that is similar to the approach used in Fonken et al. (2020). We will follow the WNSF steps to describe the final parametrization steps.

WNSF step 1: High order model We update predictor (54) according to

$$\hat{w}_j(t|t-1,\eta) = \sum_{l\in\mathcal{N}_j} G_{jl}(\eta)w_l + \sum_{k\in\mathcal{R}_j} Rr_k + \sum_{s\in\mathcal{V}_j} \left(\breve{H}_{js}(\eta) - I_{js}\right)\breve{\varepsilon}_s(\hat{\eta}^n_{j_N}), \tag{61}$$

that is used to re-estimate η, where we have added the noise topology information, thus set $\mathcal{V}_j$ is known.

At this point we have a high variance on the estimate of η but negligible bias if model order n is chosen sufficiently large. In the next step we reduce the variance by reducing the number of parameters to estimate.

WNSF step 2: Parametric model For the second step of WNSF we utilize the relation between the high order models and their rational description defined in (9). We rewrite the relation to

$$\begin{aligned} F_{jl}(\theta)G_{jl}(\eta) - L_{jl}(\theta) &= 0\,, \\ D_{js}(\theta)H_{js}(\eta) - C_{js}(\theta) &= 0, \end{aligned} \tag{62}$$

where $H_{js}(\eta)$ is an entry of matrix $H_a(\eta)$ and $H_b(\eta)$, thus including the known direct feedthrough terms. We know the direct feedthrough terms in $H^0 \in \mathbb{R}^{L\times p}$, that are derived by ordering the nodes correctly and estimating Γ^0 (47). Consequently we can rewrite the relation linear in θ analogous to (23) in MISO notation

$$\eta_j - Q_j(\eta)\theta_j = 0, \tag{63}$$

where

$$Q_j(\eta) = \begin{bmatrix} Q_j^g & 0 \\ 0 & Q_j^h \end{bmatrix}, \tag{64}$$

with Q_j^g and Q_j^h diagonal matrices with entries

$$\begin{aligned} Q_j^{g^{jl}}(\eta) &= \left[-\mathcal{T}_{n\times m_f}[G_{jl}(\eta)]\ \bar{I}_{n\times m_l}\right], \\ Q_j^{h^{js}}(\eta) &= \left[-\mathcal{T}_{n\times m_d}[H_{js}(\eta)]\ \bar{I}_{n\times m_c}\right], \end{aligned} \tag{65}$$

the top left corner of $\bar{I}_{n\times m}$ is $I_{m\times m}$ and has zeros otherwise, and $\mathcal{T}_{n\times m}[X_{ji}(q)]$ is a lower triangular matrix where the first column is $\left[x_0^{ji}\ \cdots\ x_{n-1}^{ji}\right]^\top$ with $X_{ji}(q) = \sum_{k=0}^{\infty} x_k^{ji} q^{-k}$. We obtain the consistent estimates of θ analogous to (24) in MISO notation

$$\hat{\theta}^{[0]}_{j_N} = \left(Q_j^\top(\hat{\eta}^n_{j_N})Q_j(\hat{\eta}^n_{j_N})\right)^{-1} Q_j^\top(\hat{\eta}^n_{j_N})\hat{\eta}^n_{j_N}, \tag{66}$$

that is an analytical solution. Consequently we obtain an updated reconstruction of the innovation $\breve{\varepsilon}_j(t,\hat{\theta}^{[0]}_{j_N})$ using predictor (61).

In the next step we reduce the variance further by re-estimating the obtained parametric models $G(\theta)$ and $H(\theta)$.

WNSF step 3: Re-estimation of parametric model In the final parametrization step we correct for (25) given in MISO notation

$$\hat{\eta}^n_{j_N} - Q_j(\eta)\theta_j = T_j(\theta)(\hat{\eta}^n_{j_N} - \eta^n_{j_0}), \tag{67}$$

with $T_j(\theta)$ a diagonal matrix with the denominator polynomials as entries

$$\begin{aligned} T_j^{g^{jl}}(\theta) &= \mathcal{T}_{n\times n}[F_{jl}(\eta)], \\ T_j^{h^{js}}(\theta) &= \mathcal{T}_{n\times n}[D_{js}(\eta)], \end{aligned} \tag{68}$$

where $\mathcal{T}_{n\times n}[X_{ji}(q)]$ is a lower triangular matrix where the first column is $\left[1\ x_1^{ji}\ \cdots\ x_m^{ji}\ 0_{n-m-1}\right]^\top$ with $X_{ji}(q) = 1 + \sum_{k=1}^{\infty} x_k^{ji} q^{-k}$.

This last step is an iterative step, where the estimates of θ are updated analogous to (28)

$$\hat{\theta}^{[k+1]}_{j_N} = \left(Q_j^\top(\hat{\eta}^n_{j_N})W_j(\hat{\theta}^{[k]}_{j_N})Q_j(\hat{\eta}^n_{j_N})\right)^{-1} Q_j^\top(\hat{\eta}^n_{j_N})W_j(\hat{\theta}^{[k]}_{j_N})\hat{\eta}^n_{j_N}, \tag{69}$$

that is an analytical solution. Instead of using the weighting filer in (29) we use

$$W_j(\theta^{[k]}) = T_j^{-\top}(\theta^{[k]})\Big(\frac{1}{N}\sum_{t=1}^{N}\varphi_j^{[k]}(t)(\varphi_j^{[k]}(t))^\top\Big)T_j^{-1}(\theta^{[k]}), \tag{70}$$

where we update the reconstructed innovation $\varepsilon_a(t,\hat{\theta}^{[k]}_N)$ in $\varphi_j^{[k]}$ in the weighting matrix $W_j(\theta^{[k]})$ for each iteration k, combining WNSF Step 3 with SLR Step 3. For consistency of the estimates of parameter vector θ we refer to the proof in Galrinho et al. (2019), with the actual model orders m known.

Algorithm 1 Algorithm for full network identification in dynamic networks, including noise topology detection

Inputs: $w(t), r(t)$
Output: Noise topology, $\hat{\theta}_N$

Noise topology detection

(1) Estimate noise rank p, and order the nodes if $p < L$.
(2) Reconstruct innovation $\breve{\varepsilon}(t, \hat{\zeta}_N^n)$ (45) with consistent estimate $\hat{\zeta}_N^n$ (36) obtained with predictors (41) and (42).
(3) Estimate noise correlation structure with
    (a) Structure selection with AIC, BIC and CV,
    (b) Glasso,
    applied to estimate $\hat{\eta}_N^n$ obtained with predictor (54) defined in (56).

Identification

(4) Update reconstructed innovation $\breve{\varepsilon}(t, \hat{\eta}_N^n)$ (60) with consistent estimate $\hat{\eta}_N^n$ that is obtained with predictor (54), where estimated noise topology is fixed.
(5) Estimate system parameters for reduced variance
    (a) Obtain an initial estimate $\hat{\theta}_N^{[0]}$, (66)
    (b) Re-estimate $\hat{\theta}_N^{[k+1]}$ iteratively with (69), where we update the reconstructed innovation each iteration in $\varphi_j^{[k]}$ that is in the weighting matrix (70).

We continue to iterate until the convergence criterion has been reached

$$\frac{\left\|\hat{\theta}_N^{[k]} - \hat{\theta}_N^{[k-1]}\right\|}{\left\|\hat{\theta}_N^{[k-1]}\right\|} < 0.0001. \tag{71}$$

This convergence criterion is also used in the simulation results in Section 8.

In the next section we convert the spectral conditions in Proposition 1 and 2 to a generic condition, and show how these condition hold for a 2-node example.

6. DATA INFORMATIVITY

Condition (2) of Proposition 1 and 2 is a spectral data informativity condition, but is difficult to interpret for setting up an experimental design. In this section we replace the spectral condition with a generic one, i.e. independent on the numerical values of the network dynamics. By doing so we can evaluate if data informativity is satisfied based on the network and noise topology, and the properties of the external signals. We formulate the conditions in terms of properties and locations analogous to Lemma 1 and Proposition 1 from Van den Hof and Ramaswamy (2020), by means of vertex-disjoint paths. Two paths in a graph are vertex-disjoint if they do not share a node. We use the vertex-disjoint path notation to indicate how many separate paths there are from the external signals $r(t)$ and $e(t)$ to the nodes $w(t)$.

This section proceeds with the formulation of the generic data informativity conditions, followed by a 2-node example.

6.1 Vertex-disjoint paths

For consistently estimating the full network we turn to Proposition 1 and 2, that ensure we obtain consistent parametric estimates of parameter vector ζ and η. The consistent parametric estimates are used to consistently reconstruct the innovation. The reconstructed innovation is then used in the identification method to consistently obtain estimates of parameter vector θ.

The generic equivalent to Condition (2) of Proposition 1 is given in Proposition 3.

Proposition 3: Data informativity
The spectral condition on $\kappa = [r_a(t)\ w(t)]^\top$, $\Phi_\kappa(\omega) > 0$ in Condition (2) of Proposition 1 is generically satisfied if there are at least L vertex-disjoint paths from $[r_b(t)\ e(t)]^\top$ to $w(t)$.

Proof: See appendix

Proposition 3 requires to have external excitation signals at certain locations in the network, combining data informativity conditions with identifiability (Van den Hof and Ramaswamy, 2020). The derivation in the Proof of Proposition 3 states that for consistency for the reduced noise rank case $p < L$, nodes $w_b(t)$ are to be driven via external excitation signals $r_b(t)$, i.e. rank $R_b = L - p$ which is a sufficient condition. For the full noise rank case $p = L$ the derivation in the proof shows we do not require external excitation signals $r(t)$ to be present for the estimate of ζ to be consistent.

For the generic condition for Condition (2) of Proposition 2 we introduce notation $e_{\{\mathcal{X}_j\}}(t)$, where $\mathcal{X}_j$ is the set of indices excluding indices that are already present in set $\mathcal{V}_j$. Thus if element $\{ji\}$ in the noise topology is 1, the noise model element H_{ji}^0 contains dynamics, and index $i \in \mathcal{V}_j$.

Proposition 4: Data informativity
The spectral condition on $\bar{\kappa} = \left[w_{\{\mathcal{N}_j\}}(t)\ e_{\{\mathcal{V}_j\}}(t)\right]^\top$, $\Phi_{\bar{\kappa}}(\omega) > 0$ in Condition (2) of Proposition 2 is generically satisfied if there are at least Cardinal$\{\mathcal{N}_j\}$ vertex-disjoint paths from $\left[r(t)\ e_{\{\mathcal{X}_j\}}(t)\right]^\top$ to $w_{\{\mathcal{N}_j\}}(t)$.

Proof: See appendix

Proposition 4 requires to have external excitation at certain locations in the network, combining data informativity with identifiability according to (Van den Hof and Ramaswamy, 2020).

Remark 1: Proposition 3 and 4 are given separately for the corresponding steps, not taking into account how they affect each other. For consistency to hold both propositions need to be satisfied. Thus for the reduced rank $p < L$ we need to satisfy the sufficient data informativity condition in Proposition 3 such that Condition (2) of Proposition 1 generically holds. The derivation in the proof shows that we need excitation from $r_b(t)$ to excite nodes $w_b(t)$. In addition, we also need to satisfy the necessary path based data informativity condition in Proposition 4 such that Condition (2) in Proposition 2 generically holds, in order to obtain unbiased results. The full rank case $p = L$ only requires external excitation signals $r(t)$ at locations such that Condition (2) of Proposition 2 is satisfied by Proposition 4. Moreover, if Proposition 3 holds and Proposition 4 is not satisfied for all j, the full network identification is not consistent. However, because we use MISO predictors the nodes j that do satisfy Proposition 4 can still be identified consistently in a local identification setting.

We proceed to elaborate the vertex disjoint path conditions by means of an example for the full rank and reduced rank noise.

6.2 Example

Full rank noise We consider the 2-node example shown in Figure (2), that has full rank correlated noise. We assume that the network satisfies Assumption 1 and both the network and noise topology are known.

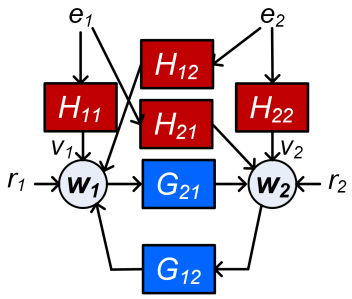


Fig. 2. 2-node network with correlated full rank noise

The goal of this example is to elaborate on the path based data informativity conditions given in Proposition 3 and 4. To be more specific, we show which external excitation signals are required in order to obtain consistent estimates of G^0_{jl} and H^0_{js} (8). We follow the steps of Algorithm 1, and split the MIMO optimization problem in two MISO optimization problems to estimate the full network.

In the case $r_1(t) = r_2(t) = 0$, the estimate of ζ obtained with Algorithm 1 is consistent. For this estimate we need 2 vertex disjoint paths from the external signals to the nodes $w(t)$ according to Proposition 3. Since the noise is full rank we have two noise sources that have 2 vertex disjoint paths to each node $w(t)$. Thus proposition 3 is satisfied, and we can consistently reconstruct the innovation. However, the path based condition in Proposition 4 is not satisfied. To obtain consistent estimates of η we use signals $w_{\{\mathcal{N}_j\}}(t)$ to parametrize the modules $G_{jl}(\eta)$ and use the noise signals $e_1(t)$ and $e_2(t)$ to parametrize the noise dynamics $H_{js}(\eta)$. We therefore cannot use the noise signals to excite the nodes $w_{\{\mathcal{N}_j\}}(t)$ and the set $\mathcal{X}_j$ does not appear there. With no $r(t)$ signals present there is no excitation on nodes $w_{\{\mathcal{N}_j\}}(t)$ and data informativity does not hold.

For the method presented in Van den Hof and Ramaswamy (2020), adding a nonzero $r_1(t)$ or $r_2(t)$ does not satisfy the data informativity condition. In that situation data informativity is satisfied if both $r_1(t)$ and $r_2(t)$ are present and a MIMO predictor model is used to estimate the network. The same holds for the joint-direct method, where the 2-node example is presented in Van den Hof et al. (2017).

For the method presented in this paper data informativity does hold if either $r_1(t)$ or $r_2(t)$ is present, while using MISO predictors. Proposition 4 states that for each node $w_j(t)$ with $j = 1, 2$ we require Cardinal$\{\mathcal{N}_j\} = 1$ vertex-disjoint paths from $\left[r(t)\ e_{\{\mathcal{X}_j\}}(t)\right]^\top$ to $w_{\{\mathcal{N}_j\}}(t)$. As mentioned before the set $\mathcal{X}_j$ does not appear because both noise signals are used to parametrize the full noise model. Therefore the path based condition in Proposition 4 is satisfied for the full network by adding $r_1(t)$ or $r_2(t)$ since only 1 vertex-disjoint path is required per node. To be more specific, if for example $r_1(t)$ is present we can translate the persistence of excitement according to

$$\begin{aligned} w_{\{\mathcal{N}_1\}} &= w_2 \\ &= (1 + G_{12}G_{21})^{-1} r_1 \\ w_{\{\mathcal{N}_2\}} &= w_1 \\ &= (1 + G_{12}G_{21})^{-1} G_{21} r_1, \end{aligned} \tag{72}$$

indicating that for all nodes there is a vertex-disjoint path from $r_1(t)$ to $w_{\{\mathcal{N}_j\}}(t)$, therefore data informativity holds for the full 2-node network.

Next we consider the 2-node network with reduced rank noise.

Reduced rank noise We consider the 2-node example, however now with

$$v_1 = He_1 = \begin{bmatrix} H_{11} \\ H_{21} \end{bmatrix} e_1 \tag{73}$$

present and $v_2(t) = 0$. In addition, the network satisfies Assumption 2. Just like the full rank case we satisfy the path based condition for Proposition 4 if either $r_1(t)$ or $r_2(t)$ is present since for both nodes set $\mathcal{X}_j$ does not appear there and we require 1 vertex disjoint path from $r(t)$ to $w_{\{\mathcal{N}_j\}}$. However, the sufficient path based condition in Proposition 3 needs $r_2(t)$ to be nonzero, where $r_2(t) = r_b(t)$. To estimate the parameter vector ζ Algorithm 1 uses the node signals $w_1(t)$ and $w_2(t)$, and (if present) the external excitation signal $r_1(t)$ to parametrize the ARX model with predictors (41) and (42). Since $r_1(t)$ is used to parametrize a part of the ARX model it cannot be used as excitation to $w(t)$. The noise signal $e_1(t)$ excites the node $w_1(t)$, but due to the rank of the noise node $w_2(t)$ is not excited by a noise signal. Therefore we need signal $r_2(t)$ to excite node $w_2(t)$ to satisfy the sufficient path based condition in Proposition 3. The 2 vertex disjoint paths are from $[r_2(t)\ e_1(t)]^\top$ to $w(t)$.

The path based data informativity conditions for Algorithm 1 indicate that we can estimate networks consistently while using MISO predictors. Compared to the local direct method in Van den Hof and Ramaswamy (2020) and the joint direct method, Algorithm 1 additionally requires less external signals to be allocated in order to consistently identify the 2-node example shown in Figure (2).

7. ANALYTICAL PROPERTIES AND SCALABILITY

7.1 Notes on analytical properties

As mentioned before we follow the approach of Ljung and Wahlberg (1992) by letting model order $n(N)$ increase at a certain rate, specified in Condition D of Ljung and Wahlberg (1992). Due to the stability assumptions on the network the impulse responses of G^0 and H^0 tend to zero at a certain rate making it possible to capture the dynamics with a finite model order n, even though the actual impulse responses can be of infinite length. Thus the consistency for the high order models holds.

The proof of consistency of WNSF also includes maximum likelihood properties for SISO and certain multivariate systems, i.e. minimum variance is reached. For the presented method we do not go beyond the consistency properties. We can say however, that unlike the joint direct method

that does have maximum likelihood properties, we do not take into account the dependencies in the noise. Therefore we know that at least for the reduced rank case additional steps are required to achieve minimum variance.

7.2 Scalability and variance

The scalability of a method is related to the computational complexity. We show what the trade off between computational burden and variance is for the computational choices made in Algorithm 1. We will compare convex with non-convex optimization problems and MIMO versus MISO. Additionally we compare the scalability of the SLS method with Algorithm 1 and the expected trade off in terms of variance.

Convex versus non-convex Both convex and non-convex methods can achieve minimum variance results. When using a non-convex criterion the number of local minima grows with the complexity and size of the network. Additionally each local minima must be evaluated to see if it is the global minimum, and requires proper initialization to prevent getting stuck in a local minimum. Convex optimization problems always converge to the global minimum. To be more specific the SLR, SLS and Algorithm 1 employ analytical solutions. Thus non-convex methods such as the joint direct method will have a higher computational burden compared to convex methods for larger networks in terms of local minima.

MISO versus MIMO The predictor models used in Algorithm 1 allow for a row-wise minimization. By implementing a MISO optimization criterion we decompose a large optimization problem in smaller, more manageable optimization problems. These manageable optimization problems can be computed sequentially or in parallel. We can therefore reduce the computation time significantly by implementing parallel computation compared to a MIMO optimization problem. However, not considering the whole network simultaneously could come at a cost in variance.

Algorithm 1 versus SLS We compare the computational burden and performance of the developed Algorithm 1 to SLS (Weerts et al., 2018). We show how many polynomials need to be estimated for the methods. Moreover, we know as the number of parameters n_p to estimate increases the variance will also increase. We therefore also compare the performance of Algorithm 1 to SLS in terms of variance by considering the relation between number of parameters to estimate n_p and data length N. We do this by comparing the methods for data generating networks that have BJ and ARMAX model structures, assuming the noise is full rank. We focus on the final parametrization step of the SLS compared to Algorithm 1, and indicate separately the intermediate high order models.

The SLS method is originally developed for networks of ARMAX structure with full rank noise. If the model orders are chosen appropriately SLS can capture any model structure due to pole-zero cancellations. The method is closely related to the WNSF method; based on the WNSF proof the SLS method is inferred to be consistent and asymptotically efficient under suitable assumptions.

For a BJ model structure Algorithm 1 parametrizes each denominator and numerator polynomial separately per node. The SLS method first has to transform the BJ model structure to an ARMAX one, by adding poles and zeros. The SLS exploits the fact that there is one common denominator per node, thus has to parametrize one denominator and the different numerator polynomials per node. The number of polynomials to estimate for the different methods increases linearly per node according to Table (1). Although the SLS has to parametrize less polynomials, the number of parameters in these polynomials is increased for a BJ model structure compared to Algorithm 1 due to the added poles and zeros. The total number of parameters n_p needed to estimate is compared to the minimal number of parameters required n_p^0, where both methods are represented in Table (2). For networks of BJ model structure Algorithm 1 has an advantage over SLS, where Algorithm 1 has to parametrize less parameters compared to SLS. Based on the ratio between the number of parameters to estimate n_p and data length N, we expect Algorithm 1 achieves a lower variance compared to SLS for BJ model structures.

For an ARMAX model structure Algorithm 1 over-parametrizes compared to SLS, since Algorithm 1 parametrizes each polynomial separately without taking into account that there are common polynomials. With the sizes of the polynomials to estimate equal for both methods, the SLS has to parametrize less parameters compared to Algorithm 1, shown in Table (2). Based on the ratio between the number of parameters to estimate n_p and data length N, we expect SLS achieves a lower variance compared to Algorithm 1 for ARMAX model structures.

Table 1. Number of final polynomials to estimate per node j

	Final parametrization	Intermediate step
SLS	$1 + Cardinal\{\mathcal{N}_j, \mathcal{V}_j\}$	L+K
Algorithm 1	$2\,Cardinal\{\mathcal{N}_j, \mathcal{V}_j\}$	L+K

Table 2. Total number of parameters n_p that need to be estimated versus the actual number of coefficients in the data generating network n_p^0, focusing on the final parametrization steps of SLS and Algorithm 1

	BJ	ARMAX
SLS	$n_p > n_p^0$	$n_p = n_p^0$
Algorithm 1	$n_p = n_p^0$	$n_p > n_p^0$

Since we consider networks of which the noise model is unknown, the model structure is not known beforehand, Algorithm 1 can be used to detect if there are for example common denominator or numerator polynomials present.

Inferring from table (2) is that when choosing the appropriate method for a certain model structure we restrict the number of parameters to estimate. Reducing the number of parameters contributes to reducing the variance. The variance results are demonstrated in the simulation results in Section 8. Considering full noise rank networks, the SLS is performing best if the data generating system fits in an ARMAX model structure. For more general structures such as BJ, Algorithm 1 is more appropriate.

For scalability we therefore prefer convex optimization problems, where the MIMO optimization problem can be split up in L MISO problems. They reduce the computational burden compared to large non-convex optimization problems, because all solutions are of closed form and the smaller optimization problems can be executed sequentially or in parallel. However, splitting the optimization problem can come at a cost in variance, because we no longer consider the network as a whole. By choosing Algorithm 1 to identify networks of BJ structure we expect to obtain a smaller variance compared to SLS, with a similar computational burden.

8. NUMERICAL ILLUSTRATIONS

In this section we show the results of different steps in Algorithm 1 for both the full noise rank and reduced noise rank case. We also compare the performance of Algorithm 1 with SLS.

For the topology detection and identification simulation study we assume $R^0 = I$ such that data informativity holds, and consider the following system

$$G(\theta) = \begin{bmatrix} 0 & 0 & 0 & G_{14} & 0 & 0 \\ 0 & 0 & 0 & 0 & G_{25} & 0 \\ G_{31} & 0 & 0 & 0 & G_{35} & 0 \\ 0 & G_{42} & 0 & 0 & 0 & 0 \\ G_{51} & 0 & 0 & 0 & 0 & G_{56} \\ 0 & 0 & G_{63} & 0 & 0 & 0 \end{bmatrix}, \tag{74}$$

with the elements of $G_{jl}(\theta)$

$$\begin{aligned} G_{14} &= \tfrac{0.38q^{-1}+0.24q^{-2}}{1-1.35q^{-1}+0.54q^{-2}}, & G_{25} &= \tfrac{0.20q^{-1}}{1-1.30q^{-1}+0.60q^{-2}}, \\ G_{31} &= \tfrac{0.39q^{-1}}{1-0.80q^{-1}+0.20q^{-2}}, & G_{35} &= \tfrac{0.16q^{-1}}{1-1.24q^{-1}+0.51q^{-2}}, \\ G_{42} &= \tfrac{-0.30q^{-1}}{1-0.60q^{-1}+0.20q^{-2}}, & G_{51} &= \tfrac{-0.6q^{-1}}{1+0.45q^{-1}+0.12q^{-2}}, \\ G_{56} &= \tfrac{-0.22q^{-1}}{1-1.22q^{-1}+0.46q^{-2}}, & G_{63} &= \tfrac{-0.11q^{-1}}{1-1.49q^{-1}+0.62q^{-2}}, \end{aligned} \tag{75}$$

for the full rank case $p = L$ we consider noise model

$$H(\theta) = \begin{bmatrix} H_{11} & 0 & 0 & H_{14} & 0 & 0 \\ 0 & H_{22} & 0 & 0 & 0 & H_{26} \\ 0 & 0 & H_{33} & 0 & H_{35} & 0 \\ 0 & H_{42} & 0 & H_{44} & 0 & 0 \\ 0 & 0 & H_{53} & 0 & H_{55} & 0 \\ 0 & H_{62} & 0 & 0 & 0 & H_{66} \end{bmatrix}, \tag{76}$$

with elements

$$\begin{aligned} H_{11} &= \tfrac{1+0.52q^{-1}}{1+0.41q^{-1}}, & H_{14} &= \tfrac{0.41q^{-1}}{1-0.56q^{-1}}, \\ H_{22} &= \tfrac{1+0.44q^{-1}}{1+0.35q^{-1}}, & H_{26} &= \tfrac{0.49q^{-1}}{1-0.49q^{-1}}, \\ H_{33} &= \tfrac{1-0.20q^{-1}}{1+0.43q^{-1}}, & H_{35} &= \tfrac{-0.56q^{-1}}{1-0.40q^{-1}}, \\ H_{42} &= \tfrac{0.26q^{-1}}{1-0.62q^{-1}}, & H_{44} &= \tfrac{1+0.52q^{-1}}{1+0.45q^{-1}}, \\ H_{53} &= \tfrac{0.32q^{-1}}{1-0.65q^{-1}}, & H_{55} &= \tfrac{1-0.20q^{-1}}{1+0.43q^{-1}}, \\ H_{62} &= \tfrac{-0.56q^{-1}}{1-0.56q^{-1}+0.21q^{-2}}, & H_{66} &= \tfrac{1+0.24q^{-1}}{1+0.53q^{-1}}, \end{aligned} \tag{77}$$

and for the reduced rank case $p < L$

$$H(\theta) = \begin{bmatrix} H_{11} & 0 & 0 & H_{14} \\ 0 & H_{22} & 0 & 0 \\ 0 & H_{32} & H_{33} & 0 \\ 0 & H_{42} & 0 & H_{44} \\ 0 & H_{52} & H_{53} & 0 \\ 0 & H_{62} & 0 & H_{64} \end{bmatrix}, \tag{78}$$

with elements

$$\begin{aligned} H_{11} &= \tfrac{1+0.52q^{-1}}{1+0.41q^{-1}}, & H_{14} &= \tfrac{0.41q^{-1}}{1-0.56q^{-1}}, \\ H_{22} &= \tfrac{1+0.44q^{-1}}{1+0.35q^{-1}}, & H_{32} &= \tfrac{-0.56q^{-1}}{1-0.40q^{-1}}, \\ H_{33} &= \tfrac{1-0.20q^{-1}}{1+0.43q^{-1}}, & H_{42} &= \tfrac{0.26q^{-1}}{1-0.62q^{-1}}, \\ H_{44} &= \tfrac{1+0.52q^{-1}}{1+0.45q^{-1}}, & H_{52} &= \tfrac{0.50q^{-1}}{1-0.49q^{-1}}, \\ H_{53} &= \tfrac{1+0.66q^{-1}}{1+0.51q^{-1}}, & H_{62} &= \tfrac{1+0.24q^{-1}}{1+0.53q^{-1}}, \\ H_{64} &= \tfrac{-0.56q^{-1}}{1-0.56q^{-1}+0.21q^{-2}}, \end{aligned} \tag{79}$$

where we rounded the parameters to two decimal places. Although the elements of the G and H matrix seem to have independent coefficients in their polynomials, this is not required.

For the simulation study we use normally distributed zero mean external signals, where $\{r(t)\}$ has a variance of 5 and $\{e(t)\}$ has variances $\{0.1, 0.2, 0.3, 0.4, 0.5, 0.2\}$ for the full rank noise case and variances $\{0.1, 0.2, 0.3, 0.4\}$ for the reduced noise rank case. We simulate the nodes according to $w(t) = (I - G^0)^{-1}(R^0 r(t) + H^0 e(t))$ and perform $M = 100$ Monte Carlo runs over five data lengths logarithmically spaced between 300 and 50000.

8.1 Rank p and ordering of the nodes

From covariance matrix $\tilde{\Lambda}$ (37) we derive the noise rank p and ordering of the nodes. Simulations with Algorithm 1 beyond the current example, where we tested multiple data generating networks varying in size and complexity, indicate that we were able to estimate the noise rank p correctly over all tested data lengths N.

8.2 Topology estimation of the noise model

For the topology detection we are interested in which indices belong in set $\mathcal{V}_j$ for all j, where the indices indicate where the edges are located in the noise model. We evaluate the performance of the topology detection by evaluating the trade-off between overestimating and underestimating the number of edges, that is typically used in receiver operating characteristic (ROC) curves, in among others Hajian-Tilaki (2013) and Shi et al. (2019).

If an edge is present in both the data generating noise and the estimated noise topology, we count this edge as a true positive (TP). If an edge is present in the estimated noise topology but does not exist in the data generating system, we count this edge as a false positive (FP). Additionally we let Pos indicate the total number of existing edges and Neg indicates the total number of non-existing edges in the noise model. The ROC curve plots the true positive rate (TPR) versus the false positive rate (FPR), with

$$TPR = \frac{TP}{Pos}, \quad FPR = \frac{FP}{Neg}, \tag{80}$$

where FPR=0 and TPR=1 represented by the point $(0, 1)$, indicates the topology is perfectly reconstructed. We evaluate the closeness to the point $(0, 1)$ by utilizing the distance function

$$dis = \sqrt{FPR^2 + (1 - TPR)^2}, \tag{81}$$

where a smaller dis indicates a better estimation of the topology.

For the structure selection procedure we test all possible combinations in set $\mathcal{V}_j$ and employ AIC, BIC and CV. For AIC we use

$$\frac{1}{2}\log\Big(V_{j_N}(\hat{\eta}^n_{j_N})\Big) + \frac{n_{p_j}}{N}, \tag{82}$$

with n_{p_j} the number of estimated parameters for node j and

$$V_{j_N}(\hat{\eta}^n_{j_N}) = \frac{1}{N}\sum_{t=1}^{N} \varepsilon_j(t, \hat{\eta}^n_{j_N})^2. \tag{83}$$

For BIC we use

$$N * \log\Big(V_{j_N}(\hat{\eta}^n_{j_N})\Big) + N(\log(2\pi)+1) + n_{p_j}\log(N). \tag{84}$$

From these simulations we select set $\mathcal{V}_j$ that gives the smallest AIC or BIC value. For the CV we split the data $Z^N = Z^{(1)}Z^{(2)}$ in a training set $Z^{(1)}$ of length $\frac{2}{3}(N+1)$ and obtain the estimates for the different combinations in set $\mathcal{V}_j$ according to

$$\hat{\eta}^{(1)}_{j_N} = \underset{\eta}{\operatorname{argmin}} V_{j_N}(\eta_j, Z^{(1)}), \tag{85}$$

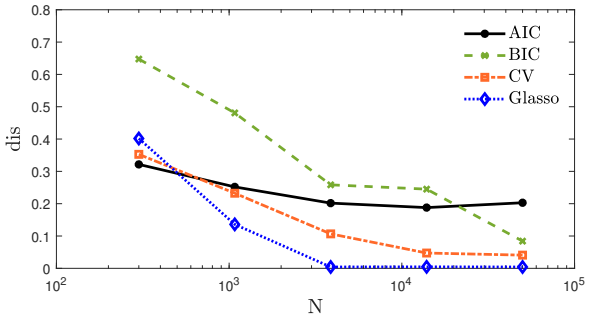
With the validation set $Z^{(2)}$, that contains the remaining data, we minimize objective function

$$V_{j_N}(\hat{\eta}^{(1)}_{j_N}, Z^{(2)}) = \frac{1}{N^{(2)}}\sum_{t=1}^{N^{(2)}} \varepsilon_j(t, \hat{\eta}^{(1)}_{j_N})^2, \tag{86}$$

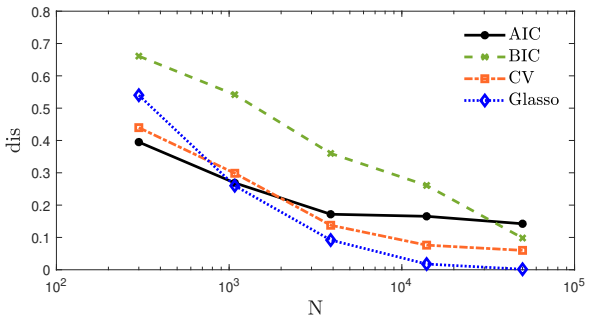
and select the set $\mathcal{V}_j$ that gives the smallest root mean squared error (RMSE)

$$RMSE_j = \sqrt{V_{j_N}(\hat{\eta}^{(1)}_{j_N}, Z^{(2)})}. \tag{87}$$

For Glasso we fully parametrize the noise model, using the known topology of G^0 and fixed $R^0 = I$ the Glasso is implemented according to



(a) full rank noise case $p = L$



(b) reduced rank noise case $p \leq L$

Fig. 3. *dis* as a function of N , averaged over the Monte Carlo runs. In (a) we show the results for the full rank noise case, and (b) shows the reduced rank noise case.

$$\frac{1}{2}||w^N_j - r^N_j - \frac{1}{N}\sum_{t=1}^{N}\varphi_j\eta_j||_2^2 + \lambda_j \sum_{i=1}^{\mathrm{C}\{\mathcal{N}_j\}+p}\sqrt{\eta_{ji}^\top I_n \eta_{ji}}, \tag{88}$$

where $\mathrm{C}\{\mathcal{N}_j\}$ is Cardinal$\{\mathcal{N}_j\}$, and λ_j is the tuning parameter of Glasso. We inspect all elements of the noise model matrix that is parametrized with the Glasso estimates. If an element $H_{ji}(\hat{\eta}_N)$ of the noise model matrix contains nonzero Glasso estimates we say this element contains dynamics, and therefore an edge is present. We include the index number i of the detected edge in the set $\mathcal{V}_j$. To prevent arbitrary small parameters are seen as dynamics we define a tolerance, where the Glasso estimates are nonzero if the l_2 norm of these estimates is larger than 10^{-3}. The choice to include the estimates of $G_{jl}(\eta)$ in the penalization is due to the implementation of Glasso. The Glasso code used is from Boyd et al. (2011) and has not yet been explored fully. For good estimates on the noise topology, we utilize the known topology of G^0 and deal with known $R^0 r(t)$ signals appropriately.

Tuning of λ_j is done via a grid based search similar to the CV structure selection. First we select a grid $\lambda^{grid}_j = \{0, 25, 50, \cdots, 2000\}$ containing λ_j values to test. For each grid point we estimate $\hat{\eta}^{grid}_j$ using Glasso, from where the topology is derived by inspecting the noise model for dynamics as mentioned before, and fix the topology H^{grid}_j per node. Next we apply CV using topology H^{grid}_j and estimate the RMSE$_j$. The grid point with the lowest RMSE$_j$ is selected as the λ_j value. Repeating the tuning procedure over a number of runs gives the minimally required value for λ_j. The tuning procedure is applied to all nodes for the different data lengths N.

Figures (3a) shows the topology detection results for the full rank case, with the distance averages over 100 Monte Carlo runs. Figure (3b) shows the reduced rank case. In both figures we see that the Glasso performs best and detects the topology with little to no error for large data lengths N. The AIC tends to estimate less edges than are present, and the BIC tends to overestimate the number of edges.

8.3 Estimation results

We will now present the results for full network identification, expressed in MSE as a function of data length N. Because Algorithm 1 is consistent we have a negligible bias and the MSE represents the variance. For these results we use the the correct estimated noise topology from the previous step.

In Algorithm 1 the reconstructed innovation in the WNSF approach is computed by

$$\breve{\varepsilon}_j(\hat{\theta}^{[k]}_{j_N}) = w^N_j - \Big[G(q,\hat{\theta}^{[k]}_{j_N})\ R(q)\ \bar{H}(q,\hat{\theta}^{[k]}_{j_N})\Big]\begin{bmatrix} w^N \\ r^N \\ \varepsilon^N_a(\hat{\eta}^n_{j_N})\end{bmatrix}, \tag{89}$$

where $\bar{H}(q,\hat{\theta}^{[k]}_{j_N})$ is $\begin{bmatrix} H_a(q,\hat{\theta}^{[k]}_{j_N}) \\ H_b(q,\hat{\theta}^{[k]}_{j_N})\end{bmatrix} - \begin{bmatrix} I \\ \hat{\Gamma}\end{bmatrix}$, due to the defined relations in (62) where the known direct feedthrough terms are included.

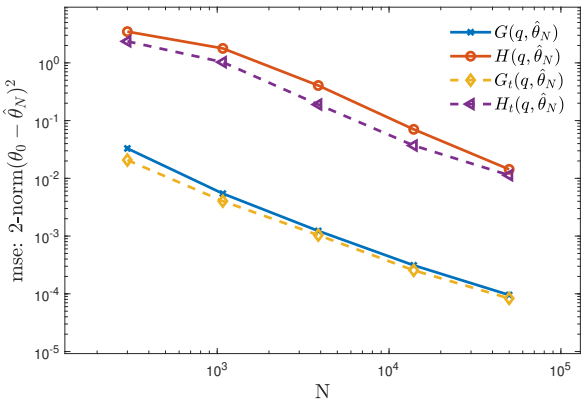
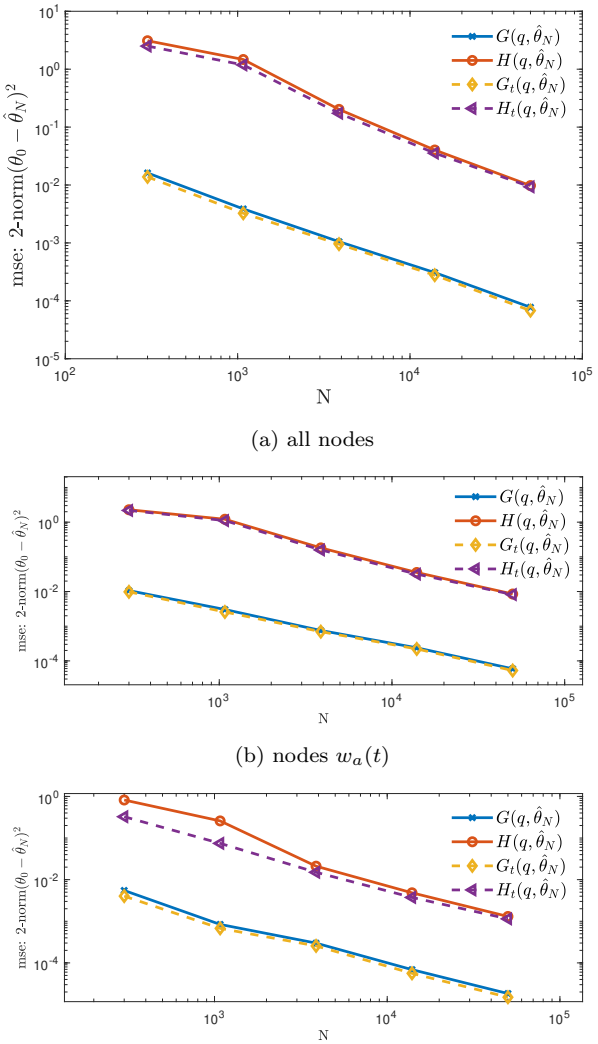


Fig. 4. MSE between $\hat{\theta}_N$ and θ_0 as function of sample size, averaged over the Monte Carlo runs for the full rank case $p = L$ computed with Algorithm 1, where subscript $\{t\}$ indicates Algorithm 1 used the true noise as an input instead of the reconstructed innovation.

For the full rank case Figure (4) presents the sample MSE, i.e. $\text{MSE}(N) = \frac{1}{M}\sum_{c=1}^{M}\left\|\hat{\theta}_{N,c}-\theta_0\right\|^2$, where c indicates the Monte Calro run. The results for the whole network are shown, while using L MISO linear regressions. The solid lines represent Algorithm 1 where the estimates are obtained using the reconstructed innovation as input. The dotted lines represent Algorithm 1 where we use the realization of the actual noise $e(t)$ as input, indicated by subscript $\{t\}$. As the data length N increases we see convergence between the solid and dotted lines. From where we could infer that for large data lengths N we achieve minimum variance results, if Algorithm 1 with the realization of the actual noise as predictor input obtains estimates with maximum likelihood properties. Furthermore all MSE results continue to converge towards zero which is in line with the consistency proof.

Figure (5a) presents the sample MSE for the reduced rank case. At first hand the results seem similar to the full rank case. However, looking at the results split over the nodes $w_a(t)$ and $w_b(t)$ in Figures (5b) (5c), the convergence between the dotted and solid lines, especially for $G(q,\theta)$ is not clear in Figure (5c), where the solid and dotted lines seem to run more parallel or converge at a significantly slower rate even though these lines are close to each other. From the simulation results we could infer that for the nodes $w_b(t)$ we do not achieve minimum variance results. This is probably due to the fact we do not take into account the dependencies in the innovation $\Gamma^0\varepsilon_a = \varepsilon_b$, of which we only use signals ε_a as predictor inputs. The figures in (5) still show that all MSE results converge towards zero that is in line with the consistency claim.

Finally we also show what happens when we replace the last step of Algorithm 1, that is based on the WNSF approach, with the OE prediction error method. Thus instead of employing analytical solutions we estimate parameter vector θ with the non-convex `oe()` command in MATLAB, where we use the reconstructed innovation as an additional predictor input. The OE is initialized using the estimates obtained from the original Algorithm 1. The results for full noise rank are shown in Figure (6) with the



(a) all nodes

(b) nodes $w_a(t)$

(c) nodes $w_b(t)$

Fig. 5. MSE between $\hat{\theta}_N$ and θ_0 as function of sample size, averaged over the Monte Carlo runs for the reduced rank case $p = L$. Computed with Algorithm 1, where subscript $\{t\}$ indicates Algorithm 1 used the true noise as an input instead of the reconstructed innovation. In (a) we show the results over the full network, (b) show the results only for nodes $w_a(t)$, and (c) shows the results only for nodes $w_b(t)$.

original algorithm that has the reconstructed innovation as input, Algorithm 1 with a realization of the actual noise as input indicated with subscript $\{t\}$, and Algorithm 1 with WNSF replaced by OE that is indicated with subscript $\{OE\}$. The OE noise model estimates lie on top of the $H_t(q,\hat{\theta}_N)$ noise model estimates with the true noise as input. For the estimates of $G(q,\theta)$, the OE estimates converges to $G(q,\hat{\theta}_N)$ with reconstructed noise as input. This simulation shows that we are able to initialize the OE model such that it does not get stuck in a local minimum, and we obtain similar results for large N using WNSF or OE in the final parametrization step. The same holds for the reduced rank noise case, that is not shown.

The results of this simulation study indicate that we obtain consistent estimates of the BJ model structure, while applying MISO optimization.

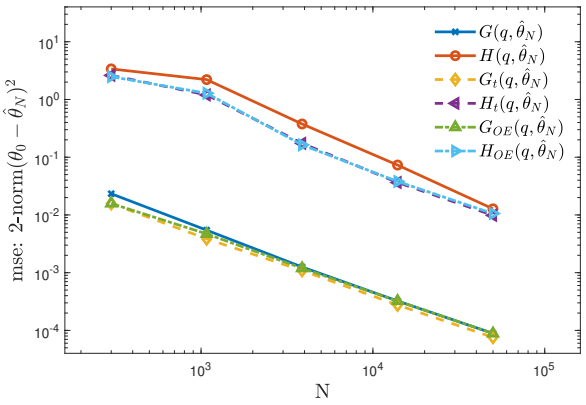


Fig. 6. MSE between $\hat{\theta}_N$ and θ_0 as function of sample size, averaged over the Monte Carlo runs for the full rank case $p = L$, shown for Algorithm 1 with the reconstructed innovation as input, and the actual noise as input indicated with subscript $\{t\}$, and Algorithm 1 where we replaced the WNSF approach in the final step by OE, indicated with subscript $\{OE\}$.

8.4 Algorithm 1 compared to SLS

We compare the performance of Algorithm 1 to SLS (Weerts et al., 2018) through a simulation study, where the noise is full rank. We focus on the MSE as a function of data length. Because both methods are consistent, the MSE represents the variance. Moreover, we perform $M = 100$ Monte Carlo runs.

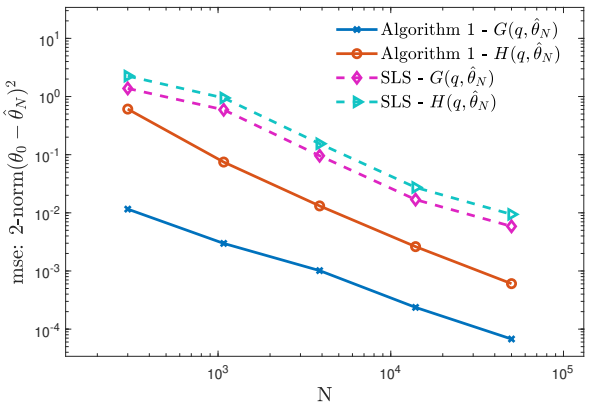
For networks with a BJ model structure, the SLS method has to overparametrize compared to Algorithm 1, due to the added poles and zeros to transform the BJ structure to an ARMAX one. For the simulation we consider a 3-node network transfer functions in BJ model structure

$$\begin{aligned} G_{12} &= \tfrac{0.50q^{-1}}{1+0.20q^{-1}}, \quad H_{11} = \tfrac{1+0.70q^{-1}}{1+0.80q^{-1}}, \\ G_{21} &= \tfrac{0.30q^{-1}}{1+0.40q^{-1}}, \quad H_{22} = \tfrac{1-0.45q^{-1}}{1-0.70q^{-1}}, \\ G_{32} &= \tfrac{-0.45q^{-1}}{1-0.60q^{-1}}, \quad H_{33} = \tfrac{1+0.50q^{-1}}{1-0.40q^{-1}}, \end{aligned} \tag{90}$$

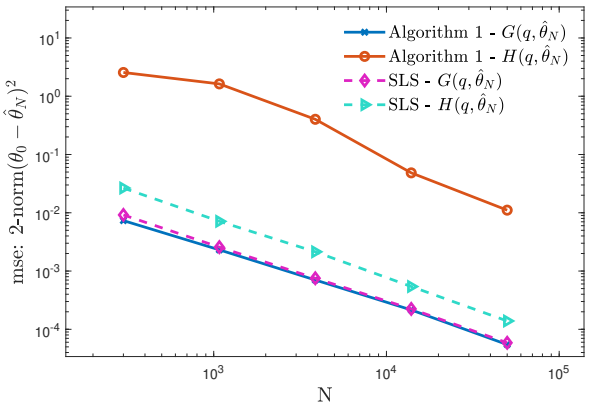
Figure (7a) presents the MSE(N) of the full network for both Algorithm 1 and SLS. Algorithm 1 outperforms SLS on both estimates of $G(\theta)$ and $H(\theta)$ in terms of variance, that is in line with the advantages mentioned before of Algorithm 1 over SLS for BJ model structures.

For an ARMAX model structure the SLS estimates less parameters compared to Algorithm 1, where the SLS method utilizes the fact that the rational transfer functions have a common denominator. For the simulation we consider transfer functions

$$\begin{aligned} G_{12} &= \tfrac{-0.7q^{-1}}{1+-0.7q^{-1}}, \quad H_{11} = \tfrac{1-0.65q^{-1}}{1-0.70q^{-1}}, \\ G_{21} &= \tfrac{0.30q^{-1}}{1+0.40q^{-1}}, \quad H_{22} = \tfrac{1+0.30q^{-1}}{1+0.40q^{-1}}, \\ G_{32} &= \tfrac{-0.30q^{-1}}{1-0.60q^{-1}}, \quad H_{33} = \tfrac{1-0.65q^{-1}}{1-0.60q^{-1}}. \end{aligned} \tag{91}$$



(a) Network with a BJ model structure



(b) Network with an ARMAX model structure

Fig. 7. MSE between $\hat{\theta}_N$ and θ_0 as function of data length N, averaged over the Monte Carlo runs, for Algorithm 1 and SLS method. In (a) the simulation is performed on an ARMAX model structure, and (b) the simulation is performed on a BJ model structure.

Figure (7b) presents the sample MSE(N) for both Algorithm 1 and SLS. The simulations results show that the estimates of $G(\theta)$ have similar MSE results for both algorithms. The difference between the algorithms is more clearly visible in the estimates of $H(\theta)$, where SLS achieves a reduced variance compared to Algorithm 1 Do note that the simulations are performed on a rather simple network.

The computation time of both algorithms for the ARMAX system are given in Table (3), averaged over 100 Monte Carlo runs. The computation time for Algorithm 1 is significantly lower, despite the fact the SLS should have a slight advantage over Algorithm 1 for ARMAX model structures. A possible explanation for the difference of the computation time lies in the implementation of the algorithms. Algorithm 1 employs analytical solutions in the algorithm, whereas the SLS employs the CVX toolbox that increases the computation time. It should be noted however, that simulations beyond the presented examples indicate that the SLS algorithm seems to handle a wider range of transfer functions compared to Algorithm 1 making SLS more robust, i.e. for certain data generating networks Algorithm 1 runs into numerical issues. The root cause of this phenomena is undetermined. The comparison

of Algorithm 1 with SLS also confirms we obtain estimates of θ that are consistent.

Table 3. Average computation time of Algorithm 1 and SLS for a 3 node ARMAX network in seconds over data lengths N

N	300	1078	3973	13916	50000
Alg. 1	2.02	2.61	2.12	3.38	10.34
SLS	7.07	8.00	13.69	38.41	156.35

9. CONCLUSION

In this paper we present a convex Algorithm that can handle reduced rank noise with low computational burden. We follow a step wise procedure where we first extend the SLR identification method to detect the noise topology, and thereafter combine the SLR method with the WNSF method to consistently identify networks of BJ model structure. Simulation results indicate we can identify the noise topology with little to no error if data length N is sufficiently large. We show that the presented method is consistent, and provide path based data informativity conditions, that guides where to allocate external excitation signals in the experimental design. Moreover, we show that the presented method is faster and has a reduced variance with respect to the SLS method for networks of a BJ structure. Considering large networks subject to correlated and or reduced rank noise, the presented method is promising due to its scalability and low variance results.

10. FUTURE WORK

Limitations of the study are the strictly proper conditions on the data generating network. In practice this limitation does not always apply. Additional delay conditions in the network can lead to less restrictive assumptions on the data generating system (Van den Hof et al., 2013; Ramaswamy and Van den Hof, 2021). It is therefore relevant to study how the presented method deals with less strict limitations on the data generating network.

Furthermore, another point of improvement is on the presented path based data informativity conditions. Currently we have to examine each node separately to evaluate if data informativity holds for the full network. A more practical approach would be to have the conditions formulated for the full network, like in Cheng et al. (2019) where they aim for generic identifiability of the whole network with a minimal number excitation signals. This way we can evaluate the conditions directly for the full network by evaluating the topologies and allocations of external excitation in the network.

For future studies it would also be relevant to study how the presented method handles sensor noise, which is not shown in this paper. The sensor noise is of relevance when the method is applied to practical systems, where the measurement data is distorted by the sensor noise. Furthermore the method has not been tested on maximum likelihood properties. When maximum likelihood holds, the method is competitive to state of the art methods such as the joint direct method, due to its scalability.

Appendix A. PROOF OF PROPOSITION 1

The proof is first given for the full rank noise case and thereafter extended to the reduced rank case.

A.1 Full rank noise, $p = L$

Consider the residual of node j

$$\begin{aligned}\varepsilon_j(t,\zeta) &= w_j(t) - \hat{w}_j(t|t-1)\\ &= A_j(\zeta)w(t) - B_j(\zeta)u(t),\end{aligned} \tag{A.1}$$

where $A_j(\zeta)$ and $B_j(\zeta)$ are the fully parametrized rows of matrices $A(\zeta) \in \mathbb{R}^{L\times L}$ and $B(\zeta) \in \mathbb{R}^{L\times K}$ belonging to node j, defined in (32) with accent ($\tilde{}$) removed. Substituting the data generating system written as

$$w(t) = (A^0)^{-1}B^0 r(t) + (A^0)^{-1}e(t), \tag{A.2}$$

in the residual of node j gives

$$\begin{aligned}\varepsilon_j(t,\zeta) &= A_j(\zeta)\Big((A^0)^{-1}B^0 r + (A^0)^{-1}e\Big) - B_j(\zeta)r\\ &= \big(A_j^0 - \Delta A_j(\zeta)\big)\Big((A^0)^{-1}B^0 r + (A^0)^{-1}e\Big) - B_j(\zeta)r\\ &= B_j^0 r - \Delta A_j(\zeta)\Big((A^0)^{-1}B^0 r + (A^0)^{-1}e\Big) + e_j - B_j(\zeta)r\\ &= \Delta B_j(\zeta)r - \Delta A_j(\zeta)\Big((A^0)^{-1}B^0 r + (A^0)^{-1}e\Big) + e_j,\end{aligned} \tag{A.3}$$

where $\Delta A_j(\zeta) = A_j^0 - A_j(\zeta)$ and $\Delta B_j(\zeta) = B_j^0 - B_j(\zeta)$.

The consistency proof consists of two steps

(1) Show the objective function is bounded from below by the noise variance $\bar{V}_j(\zeta) := \bar{\mathbb{E}}\varepsilon_j^2(t,\zeta) \geq \sigma_{e_j}^2$, where the minimum is achieved for $\Delta A_j(\zeta) = 0$ and $\Delta B_j(\zeta) = 0$.
(2) Show the global minimum is unique.

Step (1) By showing the terms

$$\Delta B_j(\zeta) - \Delta A_j(\zeta)\Big((A^0)^{-1}B^0 + (A^0)^{-1}e\Big), \tag{A.4}$$

are uncorrelated to $e_j(t)$ we can prove the minimum is achieved for $\Delta A_j = 0$ and $\Delta B_j = 0$. Using the properties of the system we can show (A.4) is uncorrelated to e_j, since

- condition (1) of Proposition 1 states that the noise $e(t)$ is uncorrelated to external signals $r(t)$. Therefore $\Delta B_j(\zeta)r(t)$ and $\Delta A_j(\zeta)(A^0)^{-1}B^0 r(t)$ are uncorrelated to $e_j(t)$,
- due to the monic property of H^0 and the strictly proper property of G^0 the term A_{jj}^0 is a monic function and the off diagonal terms A_{ji}^0 for $i \neq j$ are strictly proper. Modeling $A_j(\zeta)$ in a similar way as A_j^0 gives a strictly proper $\Delta A_j(\zeta)$. Thus term $\Delta A_j(\zeta)(A^0)^{-1}e(t)$ is a function of $e(t-i)$ for $i > 0$ that is uncorrelated to $e_j(t)$ due to the whiteness of the noise signal.

The objective function, in simplified notation is given by

$$\bar{V}_j(\zeta) = \bar{\mathbb{E}}\Big[\Big(\Delta B_j(\zeta)r - \Delta A_j(\zeta)w\Big)^2\Big] + \sigma_{e_j}^2, \tag{A.5}$$

from which we can infer that the objective function $\bar{V}_j(\zeta)$ is bounded from below by the noise variance $\sigma_{e_j}^2$.

Step (2) For the second step we show the minimum is unique, by showing

$$\bar{V}_j(\zeta) = \sigma^2_{e_j} \implies \zeta_j = \zeta_{j_0} \tag{A.6}$$

holds. We substitute $\bar{V}_j(\zeta) = \sigma^2_{e_j}$ in (A.5), that gives

$$\bar{\mathbb{E}}\Big[\Big(\Delta B_j(\zeta)r - \Delta A_j(\zeta)w\Big)^2\Big] + \sigma^2_{e_j} = \sigma^2_{e_j}, \tag{A.7}$$

and can be rewritten to

$$\bar{\mathbb{E}}\Big[\Big(\begin{bmatrix}\Delta B_j(\zeta) & -\Delta A_j(\zeta)\end{bmatrix}\begin{bmatrix}r\\ w\end{bmatrix}\Big)^2\Big] = 0. \tag{A.8}$$

Using Parseval's theorem gives:

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} \Delta x^\top(e^{j\omega},\zeta)^\top \Phi_\kappa(\omega)\Delta x(e^{-j\omega},\zeta)\mathrm{d}\omega = 0, \tag{A.9}$$

with $\kappa = [r\ w]$. By Condition (2) the spectral density $\Phi_\kappa(\omega)$ is positive definite. Therefore equation (A.9) holds only for $\Delta x^\top = 0$ which is satisfied by Condition (3). The global minimum of $\bar{V}_j(\zeta)$ is thus unique for $A_j(\zeta) = A^0_j$ and $B_j(\zeta) = B^0_j$.

A.2 Reduced rank noise $p < L$

Extending the proof to the reduced rank case, using predictor (41),(42) and R as in (39), we obtain the residual in MIMO notation

$$\begin{aligned}
\breve{\varepsilon}(t,\zeta) &= \breve{A}(\zeta)\Big((\breve{A}^0)^{-1}\breve{B}^0 r + (\breve{A}^0)^{-1}\breve{e}\Big) - \breve{B}(\zeta)r_a - R_b r_b \\
&= \big(\breve{A}^0 - \Delta\breve{A}(\zeta)\big)\Big((\breve{A}^0)^{-1}\breve{B}^0 r + (\breve{A}^0)^{-1}\breve{e}\Big) - \breve{B}(\zeta)r_a - R_b r_b \\
&= \Delta\breve{B}(\zeta)r_a - \Delta\breve{A}(\zeta)\Big((\breve{A}^0)^{-1}\breve{B}^0 r + (\breve{A}^0)^{-1}\breve{e}\Big) + \begin{bmatrix}I\\ \Gamma^0\end{bmatrix} e \\
&= \Delta\breve{B}(\zeta)r_a - \Delta\breve{A}(\zeta)\Big((\breve{A}^0)^{-1}\breve{B}^0 r + (A^0)^{-1}e\Big) + \begin{bmatrix}I\\ \Gamma^0\end{bmatrix} e,
\end{aligned} \tag{A.10}$$

where $\Delta\breve{A}(\zeta) = \breve{A}^0 - \breve{A}(\zeta)$ and $\Delta\breve{B}(\zeta) = \breve{B}^0 - \begin{bmatrix}\breve{B}(\zeta) & \begin{matrix}0\\ R_b\end{matrix}\end{bmatrix}$, with $\breve{B}(\zeta) = \breve{H}^{-1}\begin{bmatrix}R_a & 0\end{bmatrix}^\top \in \mathbb{R}^{L\times K_a}$. Note that notation $(\breve{A})^{-1}$, $\breve{B}$ and $\breve{B}$ utilize the square noise model $\breve{H} \in \mathbb{R}^{L\times L}$ and $(A)^{-1}$ uses the noise model $H \in \mathbb{R}^{L\times p}$.

The prediction error per node is defined by

$$\breve{\varepsilon}_j(t,\zeta) = \quad \Delta\breve{B}_j(\zeta)r_a - \Delta\breve{A}_j(\zeta)w + \breve{e}_j. \tag{A.11}$$

The proof for *Step 1* still holds since $\Delta\breve{A}_j(\zeta)$ contains monic terms $\breve{A}^0_{jj}$ and $\breve{A}_{jj}(\zeta)$, and has strictly proper off-diagonal terms. The objective function $\bar{V}_j$ is therefore bounded from below by $\sigma^2_{\breve{e}_j}$.

For *Step 2* of the proof we rewrite (A.8) to

$$\bar{\mathbb{E}}\Big[\Big(\begin{bmatrix}\Delta\breve{B}_j(\zeta) & -\Delta\breve{A}_j(\zeta)\end{bmatrix}\begin{bmatrix}r_a\\ w\end{bmatrix}\Big)^2\Big] = 0, \tag{A.12}$$

For the reduced rank case Parseval's theorem (A.9), with $\kappa = [r_a\ w]$, still holds, and the global minimum of $\bar{V}_j(\zeta)$ is therefore unique for $\breve{A}_j(\zeta) = \breve{A}^0_j$ and $\breve{B}_j(\zeta) = \breve{B}^0_j \in \mathbb{R}^{1\times K_a}$. □

Remark A.1 In the proof for $p < L$ we only restrict parametrization in the $\breve{B}(\zeta)$ matrix by utilizing the noise rank p, and Assumption 2. We still fully parametrize $\breve{A}(\zeta)$. It is also possible to restrict parametrization in $\breve{A}(\zeta)$ using the known zeros in the last $L-p$ columns of $(\breve{H}^0)^{-1}$ and the known topology of G^0, this however still results in a rather full parametrization of $\breve{A}(\zeta)$ for smaller networks. For large sparse networks it could be beneficial to also reduce the parametrization in matrix $\breve{A}(\zeta)$.

Appendix B. PROOF OF PROPOSITION 2

The proof is first given for the full rank noise case and thereafter extended to the reduced rank case. Also included is the proof in terms of least squares.

B.1 Full rank noise, $p = L$

Consider the MIMO predictor given by

$$\hat{w}(t|t-1) = G(\eta)w + Rr + (H(\eta) - I)\varepsilon(\hat{\zeta}^n_N), \tag{B.1}$$

where $\varepsilon(\hat{\zeta}^n_N)$ from step 1 of the method is a consistent estimate, i.e.

$$\varepsilon(\hat{\zeta}^n_N) \to e(t) \quad \text{w.p. 1 as } N \to \infty\, \forall t. \tag{B.2}$$

Therefore we rewrite the predictor as

$$\hat{w}(t|t-1) = G(\eta)w + Rr + \big(H(\eta) - I\big)e. \tag{B.3}$$

Considering the data generating system in (1) the residual in MISO representation is

$$\begin{aligned}
\varepsilon_j(t,\eta) &= w_j(t) - \hat{w}_j(t|t-1,\eta) \\
&= \sum_{l\in\mathcal{N}_j} \Delta G_{jl}(\eta)w_l + \sum_{s\in\mathcal{V}_j} \Delta H_{js}(\eta)e_s + e_j,
\end{aligned} \tag{B.4}$$

where $\Delta G_{jl}(\eta) = G^0_{jl} - G_{jl}(\eta)$ and $\Delta H_{js}(\eta) = H^0_{js} - H_{js}(\eta)$.

The consistency proof consists of two steps

(1) Show the objective function is bounded from below by the noise variance $\bar{V}_j(\theta) := \bar{\mathbb{E}}\varepsilon^2_j(t,\theta) \geq \sigma^2_{e_j}$, where the minimum is achieved for $\Delta G_{jl} = 0$ and $\Delta H_{js} = 0$.
(2) Show the global minimum is unique.

Step 1 By showing

$$\sum_{l\in\mathcal{N}_j} \Delta G_{jl}(\eta)w_l + \sum_{s\in\mathcal{V}_j} \Delta H_{js}(\eta)e_s \tag{B.5}$$

is uncorrelated to $e_j(t)$ we can prove the minimum is achieved for $\Delta G_{jl} = 0$ and $\Delta H_{js} = 0$. This can be shown using the properties of the system

- The system only contains strictly proper transfer functions in G^0_{jl}, $G_{jl}(\theta)$ is therefore parametrized as a strictly proper function, thus the term $\Delta G_{jl}w_l$ is a function of $w_l(t-i)$, $i > 1$,
- ΔH_{js} is a strictly proper term since H^0_{jj} and $H_{jj}(\theta)$ are both monic and H^0_{js} and $H_{js}(\theta)$ for $s \neq j$ are strictly proper, the term $\Delta H_{js}e_s$ is therefore a function of $e_s(t-i)$, $i > 1$,

and the objective function

$$\bar{V}_j(\eta) = \bar{\mathbb{E}}\Big[\Big(\sum_{l\in\mathcal{N}_j} \Delta G_{jl}(\eta)w_l + \sum_{s\in\mathcal{V}_j} \Delta H_{js}(\eta)e_s\Big)^2\Big] + \sigma^2_{e_j} \tag{B.6}$$

is therefore bounded from below by the noise variance $\bar{V}_j(\theta) \geq \sigma^2_{e_j}$.

Step 2 Showing the minimum is unique is done by showing

$$\bar{V}_j(\eta) = \sigma^2_{e_j} \implies \eta_j = \eta_{j_0} \tag{B.7}$$

holds. Equation (B.6) with $\bar{V}_j(\eta) = \sigma^2_{e_j}$ substituted is equal to

$$\begin{aligned}\bar{\mathbb{E}}\Big[\Big(\sum_{l\in\mathcal{N}_j}\Delta G_{jl}(\eta)w_l+\sum_{s\in\mathcal{V}_j}\Delta H_{js}(\eta)e_s\Big)^2\Big]+\sigma^2_{e_j}=\sigma^2_{e_j}\\ \bar{\mathbb{E}}\Big[\Big(\sum_{l\in\mathcal{N}_j}\Delta G_{jl}(\eta)w_l+\sum_{s\in\mathcal{V}_j}\Delta H_{js}(\eta)e_s\Big)^2\Big]=0,\end{aligned} \tag{B.8}$$

and can be rewritten to

$$\bar{\mathbb{E}}\Big[\Big(\Delta x^\top\begin{bmatrix}w_{\{\mathcal{N}_j\}}\\ e_{\{\mathcal{V}_j\}}\end{bmatrix}\Big)^2\Big]=0, \tag{B.9}$$

where

$$\Delta x^\top=\begin{bmatrix}\Delta G_{jl\in\mathcal{N}_j} & \Delta H_{js\in\mathcal{V}_j}\end{bmatrix}. \tag{B.10}$$

Using Parseval's theorem gives

$$\frac{1}{2\pi}\int_{-\pi}^{\pi}\Delta x^\top(e^{j\omega},\eta)^\top\Phi_{\bar{\kappa}}(\omega)\Delta x(e^{-j\omega},\eta)\mathrm{d}\omega=0, \tag{B.11}$$

with $\bar{\kappa}=\begin{bmatrix}w_{\{\mathcal{N}_j\}}(t) & e_{\{\mathcal{V}_j\}}(t)\end{bmatrix}^\top$. By Condition (2) the spectral density $\Phi_{\bar{\kappa}}$ is positive definite. Therefore equation (B.11) holds only for $\Delta x^\top=0$. The Parseval's theorem shows the the global minimum of $\bar{V}_j(\eta)$ is unique for $G_{jl}(\eta)=G^0_{jl}$ and $H_{js}(\eta)=H^0_{js}$ by Condition (3).

B.2 Reduced rank noise, $p<L$

We will extend the proof to the reduced rank case, where the residuals in MIMO notation are given by

$$\breve{\varepsilon}(t,\eta)=\Delta G(\eta)w+H^0e-(\breve{H}(\eta)-I)\breve{\varepsilon}(\hat{\zeta}^n_N). \tag{B.12}$$

From Proposition 1 we know $\hat{\zeta}^n_N$ is consistent and therefore

$$\breve{\varepsilon}(\hat{\zeta}^n_N)\to\breve{e}\quad\text{w.p. 1 as } N\to\infty\,\forall t. \tag{B.13}$$

such that the residual can be rewritten to

$$\breve{\varepsilon}(t,\eta)=\Delta G(\eta)w+\begin{bmatrix}\Delta H_a(\eta)\\ \Delta\bar{H}_b(\eta)\end{bmatrix}e+\begin{bmatrix}I\\ \Gamma^0\end{bmatrix}e, \tag{B.14}$$

where $\Delta G(\eta)=G^0-G(\eta)$, $\Delta H_a(\eta)=H^0_a-H_a(\eta)$ and $\Delta\bar{H}_b(\eta)=\bar{H}^0_b-\bar{H}_b(\eta)$, with $\bar{H}_b=H_b-\Gamma$.

Deriving the residuals per nodes $w_a(t)$ and $w_b(t)$ gives

$$\begin{aligned}\breve{\varepsilon}_{a,j}(t,\eta)&=\sum_{l\in\mathcal{N}_j}\Delta G_{jl}(\eta)w_l+\sum_{s\in\mathcal{V}_j}\Delta H_{a,js}(\eta)e_s+e_j,\\ \breve{\varepsilon}_{b,j}(t,\eta)&=\sum_{l\in\mathcal{N}_j}\Delta G_{jl}(\eta)w_l+\sum_{s\in\mathcal{V}_j}\Delta H_{b,js}(\eta)e_s+\Gamma^0e_j.\end{aligned} \tag{B.15}$$

Since $H^0_{a,jj}$ and $H_{a,jj}(\eta)$ are monic and the off diagonal elements $H_{a,ji}$ for $i\neq j$ being strictly proper, the term $\Delta H_{a,js}(\eta)$ is strictly proper. $\Delta\bar{H}_{b,js}(\eta)$ is strictly proper because both $\bar{H}^0_{b,js}$ and $\bar{H}_{b,js}(\eta)$ are strictly proper. Therefore the aforementioned proof of Step 1 still holds for the reduced rank case and the objective function $\bar{V}_j$ is bounded from below by $\sigma^2_{\breve{e}_j}$. From here the proof of the reduced rank case is analogous to the full rank case, keeping in mind both matrices $\Delta H(\eta)=\begin{bmatrix}\Delta H_a(\eta)\\ \Delta\bar{H}_b(\eta)\end{bmatrix}$ and H^0 are of dimension $(L\times p)$.

B.3 Least squares approach for consistency

In terms of least squares we can show consistency for

$$\begin{aligned}\hat{\eta}^n_N-\eta_0=&\Big(\frac{1}{N}\sum_{t=1}^N\varphi\varphi^\top\Big)^{-1}\Big(\frac{1}{N}\sum_{t=1}^N\varphi w\Big)-\\ &\Big(\frac{1}{N}\sum_{t=1}^N\varphi\varphi^\top\Big)^{-1}\Big(\frac{1}{N}\sum_{t=1}^N\varphi\varphi^\top\Big)\eta_0\\ =&\Big(\frac{1}{N}\sum_{t=1}^N\varphi\varphi^\top\Big)^{-1}\Big(\frac{1}{N}\sum_{t=1}^N\varphi\varepsilon\Big),\end{aligned} \tag{B.16}$$

where $\hat{\eta}^n_N$ is consistent if

(a) $\bar{\mathbb{E}}[\varphi\varphi^\top]$ is non singular
(b) $\bar{\mathbb{E}}[\varphi\varepsilon]=0$

Condition (a) is satisfied when the system has sufficient excitation, which is satisfied by Condition (1) and (2) of Proposition 2, the whiteness of the noise and Assumption 1g. Condition (b) is satisfied by the white noise condition, thus

$$\bar{\mathbb{E}}\begin{bmatrix}w_{\{\mathcal{N}_j\}}(t-1)\\ \vdots\\ w_{\{\mathcal{N}_j\}}(t-n)\\ \varepsilon_{\{\mathcal{V}_j\}}(t-1,\hat{\zeta}^n_N)\\ \vdots\\ \varepsilon_{\{\mathcal{V}_j\}}(t-n,\hat{\zeta}^n_N)\end{bmatrix}\varepsilon_j(t)=0 \tag{B.17}$$

holds, since both $w_{\{\mathcal{N}_j\}}(t-i)$ and $\varepsilon_{\{\mathcal{V}_j\}}(t-i,\hat{\zeta}^n_N)$ for $i>1$ are uncorrelated to $\varepsilon_j(t)$. □

Appendix C. PROOF OF PROPOSITION 3

Proposition 3 replaces Condition (2) in Proposition 1. The proof is first given for the full rank noise case and thereafter extended to the reduced rank case.

C.1 Full rank noise $p=L$

We replace the spectral condition $\Phi_\kappa(\omega)>0$, with the generic condition given in Proposition 3. To this end we utilize Lemma 1 and Proposition 1 of Van den Hof and Ramaswamy (2020). We can translate the persistence of excitement from the node signals w to external signals $[r\ e]^\top$, giving

$$w=\bar{J}\begin{bmatrix}r\\ e\end{bmatrix},\ \text{with } \bar{J}=\begin{bmatrix}(A^0)^{-1}B^0 & (A^0)^{-1}\end{bmatrix}. \tag{C.1}$$

Substituting $\bar{J}$ in (A.8) results in

$$\bar{\mathbb{E}}\Big[\Big(\Delta B_j(\zeta)r-\Delta A_j(\zeta)\bar{J}\begin{bmatrix}r\\ e\end{bmatrix}\Big)^2\Big]=0 \tag{C.2}$$

that can be rewritten to

$$\begin{aligned}&\bar{\mathbb{E}}\Big[\Big([\Delta B_j(\zeta)\ -\Delta A_j(\zeta)]\begin{bmatrix}I & 0\\ (A^0)^{-1}B^0 & (A^0)^{-1}\end{bmatrix}\begin{bmatrix}r\\ e\end{bmatrix}\Big)^2\Big]=0\\ &\bar{\mathbb{E}}\Big[\Big(\Delta x^\top J\begin{bmatrix}r\\ e\end{bmatrix}\Big)^2\Big]=0,\end{aligned} \tag{C.3}$$

where

$$\Delta x^\top=[\Delta B_j(\zeta)\ -\Delta A_j(\zeta)],\ \text{and } J=\begin{bmatrix}I & 0\\ \bar{J}\end{bmatrix}. \tag{C.4}$$

Using Parseval's theorem gives:

$$\frac{1}{2\pi}\int_{-\pi}^{\pi}\Delta x^\top(e^{j\omega},\zeta)^\top J\Phi_{\left[\begin{smallmatrix}r\\e\end{smallmatrix}\right]}(\omega)J^*\Delta x(e^{-j\omega},\zeta)\mathrm{d}\omega=0 \tag{C.5}$$

By Condition (1) of Proposition 1, the whiteness of the noise and Assumption 1g the spectral density $\Phi_{\left[\begin{smallmatrix}r\\e\end{smallmatrix}\right]}(\omega)$ is positive definite. Moreover, J is required to be full rank, where $\bar{J}$ is full rank due to Proposition 3. In this situation we can also show $\bar{J}$ is full rank due to the monic property of $H^0 \in \mathbb{R}^{L\times L}$ in matrix $(A^0)^{-1}$. This indicates there are L vertex disjoint paths from $e(t)$ to $w(t)$. Therefore equation (C.5) holds only for $\Delta x^\top = 0$ which is satisfied by Condition (3) of Proposition 1. The global minimum of $\bar{V}_j(\zeta)$ is thus unique for $A_j(\zeta) = A_j^0$ and $B_j(\zeta) = B_j^0$.

C.2 Reduced rank noise $p < L$

For the reduced noise rank case we translate the persistence of excitement from the node signals w to external signals $[r_b\ e]^\top$, giving

$$w=\bar{J}\begin{bmatrix}r_b\\e\end{bmatrix},\ \text{with}\ \bar{J}=\begin{bmatrix}(\breve{A}^0)^{-1}\breve{B}^0_{\backslash p\,col}\ (A^0)^{-1}\end{bmatrix}, \tag{C.6}$$

where $\breve{B}^0_{\backslash p\,col} = \begin{bmatrix}0\ R_b^0\end{bmatrix}^\top$. The total expression (A.12) is rewritten to

$$\bar{\mathbb{E}}\Big[\Big(\Delta x^\top J\begin{bmatrix}r_a\\r_b\\e\end{bmatrix}\Big)^2\Big]=0, \tag{C.7}$$

with

$$\Delta x^\top=\begin{bmatrix}\Delta\breve{B}_j(\zeta)\ -\Delta\breve{A}_j(\zeta)\end{bmatrix},\ \text{and}\ J=\begin{bmatrix}I&0&0\\0&&\bar{J}\end{bmatrix}. \tag{C.8}$$

Because matrix $(A^0)^{-1} \in \mathbb{R}^{L\times p}$ in $\bar{J}$, we cannot directly make a statement about the rank of $\bar{J}$. Instead we employ the generic data informativity condition in Proposition 3, i.e. not taking the numerical values of transfer functions into account. We utilize Lemma 1 and Proposition 1 from Van den Hof and Ramaswamy (2020) to make a statement on data informativity, where the row rank of $\bar{J}$ is assessed. For identifiability and data informativity to hold we require the row rank $\bar{J} \geq L$, that is satisfied by Proposition 3. Since $\Phi_{[r_a\ r_b\ e]^\top}(\omega)$ is positive definite due to Condition (1) of Proposition 1, the whiteness of the noise and Assumption 1g, the Parseval's theorem (C.5) shows the minimum is unique for $\Delta\breve{A}_j = 0$ and $\Delta\breve{B}_j = 0$ by Condition (3) of Proposition 1. □

Remark C.1 Since the dimension of the noise vector $e(t) \in \mathbb{R}^p$, Proposition 3 indicates we must have at least $r_b(t) \in \mathbb{R}^{(L-p)}$ signals available in the reduced rank case, i.e. all nodes $w_b(t)$ should be driven by an external excitation signal from vector $r_b(t)$.

Appendix D. PROOF OF PROPOSITION 4

Proposition 4 replaces Condition (2) in Proposition 2. The proof is first given for the full rank noise case and thereafter extended to the reduced rank case.

D.1 Full rank noise $p = L$

We replace the spectral condition $\Phi_{\bar{\kappa}}(\omega) > 0$, with the generic condition given in Proposition 4. To this end we utilize Lemma 1 and Proposition 1 of Van den Hof and Ramaswamy (2020). We translate the persistence of excitement from the node signals $w_{\{\mathcal{N}_j\}}$ to external signals $\begin{bmatrix}r\ e_{\{\mathcal{X}_j\}}\end{bmatrix}^\top$, where set $\{\mathcal{X}_j\}$ contains the noise indices excluding indices that are already in set $\mathcal{V}_j$, giving

$$\begin{aligned}&w_{\{\mathcal{N}_j\}}=\bar{J}\begin{bmatrix}r\\e_{\{\mathcal{X}_j\}}\end{bmatrix},\\&\text{with}\ \bar{J}=\begin{bmatrix}\mathcal{G}^0_{row_{\{\mathcal{N}_j\}}}R^0\ \mathcal{G}^0_{row_{\{\mathcal{N}_j\}}}H^0_{\backslash col_{\{\mathcal{V}_j\}}}\end{bmatrix},\end{aligned} \tag{D.1}$$

where $\mathcal{G}^0_{row_{\{\mathcal{N}_j\}}}$ are the rows of $\mathcal{G}^0 = (I - G^0)^{-1}$ with the row indices in set $\mathcal{N}_j$, and $H^0_{\backslash col_{\{\mathcal{V}_j\}}}$ are the columns of H^0 excluding the columns indices in set $\mathcal{V}_j$.

Equation (B.9) can therefore be rewritten to

$$\bar{\mathbb{E}}\Big[\Big(\Delta x^\top J\begin{bmatrix}r\\e_{\{\mathcal{X}_j\}}\\e_{\{\mathcal{V}_j\}}\end{bmatrix}\Big)^2\Big]=0,\ \text{with}\ J=\begin{bmatrix}\bar{J}&0\\0\ 0\ I\end{bmatrix}. \tag{D.2}$$

Using Parseval's theorem gives

$$\frac{1}{2\pi}\int_{-\pi}^{\pi}\Delta x^\top(e^{j\omega},\eta)^\top J\Phi_{\left[\begin{smallmatrix}r\\e_{\{\mathcal{X}_j\}}\\e_{\{\mathcal{V}_j\}}\end{smallmatrix}\right]}(\omega)J^*\Delta x(e^{-j\omega},\eta)\mathrm{d}\omega=0 \tag{D.3}$$

For the generic condition to hold we require the row rank $\bar{J} \geq$ Cardinal$\{\mathcal{N}_j\}$, that is satisfied by Proposition 4. The spectral density $\Phi_{[r\ e_{\{\mathcal{X}_j\}}\ e_{\{\mathcal{V}_j\}}]^\top}(\omega)$ is positive definite by Condition (1) of Proposition 2, the whiteness of the noise and Assumption 1g. The Parseval's theorem (D.3) shows the the global minimum of $\bar{V}_j(\eta)$ is unique for $G_{jl}(\eta) = G^0_{jl}$ and $H_{js}(\eta) = H^0_{js}$ by Condition (3) of Proposition 2.

D.2 Reduced rank noise $p < L$

The proof of the reduced rank case is analogous to the full rank case, keeping in mind matrix H^0 is of dimension $(L \times p)$. □

REFERENCES

Araki, M. and Saeki, M. (1983). A quantitative condition for the well-posedness of interconnected dynamical systems. *IEEE Transactions on Automatic Control (TAC)*, 28(5), 569–577.

Bolstad, A., Van Veen, B.D., and Nowak, R. (2011). Causal network inference via group sparse regularization. *IEEE Transactions on Signal Processing*, 59(6), 2628–2641.

Boyd, S., Parikh, N., and Chu, E. (2011). *Distributed optimization and statistical learning via the alternating direction method of multipliers*. Now Publishers Inc.

Cheng, X., Shi, S., and Van den Hof, P.M.J. (2019). Allocation of excitation signals for generic identifiability of dynamic networks. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, 5507–5512.

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

Dankers, A.G. (2019). Optimization method for obtaining estimates in a dynamic network. Technical note.

Dankers, A.G., Van den Hof, P.M.J., Bombois, X., and 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., and Bombois, X. (2012). Dynamic network structure identification with prediction error methods - basic examples. *IFAC Proceedings Volumes*, 45(16), 876 – 881. 16th IFAC Symposium on System Identification.

Durbin, J. (1959). Efficient estimation of parameters in moving-average models. *Biometrika*, 46(3/4), 306–316.

Durbin, J. (1960). The fitting of time-series models. *Revue de l'Institut International de Statistique / Review of the International Statistical Institute*, 28(3), 233–244.

Everitt, N., Bottegal, G., Rojas, C., and Hjalmarsson, H. (2015). On the effect of noise correlation in parameter identification of simo systems. *IFAC-PapersOnLine*, 48, 326–331.

Everitt, N., Galrinho, M., and Hjalmarsson, H. (2018). Open-loop asymptotically efficient model reduction with the Steiglitz–Mcbride method. *Automatica*, 89, 221 – 234.

Fonken, S.J.M., Ferizbegovic, M., and Hjalmarsson, H. (2020). Consistent identification of dynamic networks subject to white noise using Weighted Null-Space Fitting. In *21st IFAC World Congress*. To appear.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Applications of the lasso and grouped lasso to the estimation of sparse graphical models. Technical report, Stanford University.

Galrinho, M., Rojas, C., and Hjalmarsson, H. (2019). Parametric identification using Weighted Null-Space Fitting. *IEEE Transactions on Automatic Control (TAC)*, 64(7), 2798–2813.

Gevers, M., Bazanella, A.S., and Pimentel, G.A. (2019). Identifiability of dynamical networks with singular noise spectra. *IEEE Transactions on Automatic Control (TAC)*, 64(6), 2473–2479.

Hajian-Tilaki, K. (2013). Receiver operating characteristic (roc) curve analysis for medical diagnostic test evaluation. *Caspian journal of internal medicine*, 4(2), 627.

Ljung, L. (1999). *System Identification. Theory for the User, 2nd ed.* Prentice-Hall.

Ljung, L. and Wahlberg, B. (1992). Asymptotic properties of the least-squares method for estimating transfer functions and disturbance spectra. *Advances in Applied Probability*, 24(2), 412–440.

Materassi, D., Salapaka, M.V., and Giarrè, L. (2011). Relations between structure and estimators in networks of dynamical systems. *Proceedings of the IEEE Conference on Decision and Control (CDC)*, 162–167.

Materassi, D. and Innocenti, G. (2010). Topological identification in networks of dynamical systems. *IEEE Transactions on Automatic Control (TAC)*, 55(8), 1860–1871.

Materassi, D. and Salapaka, M.V. (2012). On the problem of reconstructing an unknown topology via locality properties of the wiener filter. *IEEE Transactions on Automatic Control (TAC)*, 57(7), 1765–1777.

Ramaswamy, K.R., Bottegal, G., and Van den Hof, P.M.J. (2018). Local module identification in dynamic networks using regularized kernel-based methods. In *2018 IEEE Conference on Decision and Control (CDC)*, 4713–4718.

Ramaswamy, K.R. and Van den Hof, P.M.J. (2021). A local direct method for module identification in dynamic networks with correlated noise. *IEEE Transactions on Automatic Control (TAC)*, 66 (11). To appear.

Shi, S., Bottegal, G., and Van den Hof, P.M.J. (2019). Bayesian topology identification of linear dynamic networks. In *2019 18th European Control Conference (ECC)*, 2814–2819.

Van den Hof, P.M.J., Dankers, A.G., Heuberger, P.S.C., and 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., and Weerts, H.H.M. (2017). From closed-loop identification to dynamic networks: Generalization of the direct method. In *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*, 5845–5850.

Van den Hof, P.M.J. and Ramaswamy, K.R. (2020). Path-based data-informativity conditions for single module identification in dynamic networks. To appear in Proc. of Conference on Decision and Control (CDC) 2020.

Van den Hof, P.M.J. and Schrama, R.J.P. (1993). An indirect method for transfer function estimation from closed loop data. *Automatica*, 29(6), 1523 – 1527.

Van den Hof, P.M.J., Weerts, H.H.M., and Dankers, A.G. (2017). Prediction error identification with rank-reduced output noise. In *2017 American Control Conference (ACC)*, 382–387.

Wasserman, L. (2000). Bayesian model selection and model averaging. *Journal of Mathematical Psychology*, 44(1), 92 – 107.

Weerts, H.H.M., Galrinho, M., Bottegal, G., Hjalmarsson, H., and Van den Hof, P.M.J. (2018). A Sequential Least Squares algorithm for ARMAX dynamic network identification. *IFAC-PapersOnLine*, 51(15), 844–849.

Weerts, H.H.M., Van den Hof, P.M.J., and Dankers, A.G. (2016). Identifiability of dynamic networks with part of the nodes noise-free. *IFAC-PapersOnLine*, 49(13), 19 – 24.

Weerts, H.H.M., Van den Hof, P.M.J., and Dankers, A.G. (2017). Identification of dynamic networks with rank-reduced process noise. *IFAC-PapersOnLine*, 50(1), 10562–10567.

Weerts, H.H.M., Van den Hof, P.M.J., and Dankers, A.G. (2018a). Identifiability of linear dynamic networks. *Automatica*, 89, 247 – 258.

Weerts, H.H.M., Van den Hof, P.M.J., and Dankers, A.G. (2018b). Prediction error identification of linear dynamic networks with rank-reduced noise. *Automatica*, 98, 256 – 268.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B*, 68, 49–67.

Yuan, Y., Stan, G.B., Warnick, S., and Gonçalves, J. (2011). Robust dynamical network structure reconstruction. *Automatica*, 47(6), 1230–1235.