Variance Analysis of Linear SIMO Models with Spatially Correlated Noise [⋆]

Niklas Everitt [a], Giulio Bottegal [a], Cristian R. Rojas [a], Håkan Hjalmarsson [a]

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

Abstract

Substantial improvement in accuracy of identified linear time-invariant single-input multi-output (SIMO) dynamical models is possible when the disturbances affecting the output measurements are spatially correlated. Using an orthogonal representation for the modules composing the SIMO structure, in this paper we show that the variance of a parameter estimate of a module is dependent on the model structure of the other modules, and the correlation structure of the disturbances. In addition, we quantify the variance-error for the parameter estimates for finite model orders, where the effect of noise correlation structure, model structure and signal spectra are visible. From these results, we derive the noise correlation structure under which the mentioned model parameterization gives the lowest variance, when one module is identified using less parameters than the other modules.

Key words: System identification, Asymptotic variance, Linear SIMO models, Least-squares.

1 Introduction

Recently, system identification in dynamic networks has gained popularity, see e.g., Van den Hof *et al.* (2013); Dankers *et al.* (2013*b*,*a*, 2014); Ali *et al.* (2011); Materassi *et al.* (2011); Torres *et al.* (2014); Haber and Verhaegen (2014); Gunes *et al.* (2014); Chiuso and Pillonetto (2012). In this framework, signals are modeled as nodes in a graph and edges model transfer functions. To estimate a transfer function in the network, a large number of methods have been proposed. Some have been shown to give consistent estimates, provided that a certain subset of signals is included in the identification process (Dankers *et al.*, 2013*b*,*a*). In these methods, the user has the freedom to include additional signals. However, little is known on how these signals should be chosen, and how large the potential is for variance reduction. From a theoretical point of view there are only a few results regarding specific network structures e.g., Hägg *et al.* (2011); Ramazi *et al.* (2014); Wahlberg *et al.* (2009). To get a better understanding of the potential of adding

available information to the identification process, we ask the fundamental questions: will, and how much, an added sensor improve the accuracy of an estimate of a certain target transfer function in the network? We shall attempt to give some answers by focusing on a special case of dynamic networks, namely single-input multi-output (SIMO) systems, and in a wider context, dynamic networks.

SIMO models are interesting in themselves. They find applications in various disciplines, such as signal processing and speech enhancement (Benesty *et al.*, 2005), (Doclo and Moonen, 2002), communications (Bertaux *et al.*, 1999), (Schmidt, 1986), (Trudnowski *et al.*, 1998), biomedical sciences (McCombie *et al.*, 2005) and structural engineering (Ulusoy *et al.*, 2011). Some of these applications are concerned with spatio-temporal models, in the sense that the measured output can be strictly related to the location at which the sensor is placed (Stoica *et al.*, 1994), (Viberg *et al.*, 1997), (Viberg and Ottersten, 1991). In these cases, it is reasonable to expect that measurements collected at locations close to each other are affected by disturbances of the same nature. In other words, noise on the outputs can be correlated; understanding how this noise correlation affects the accuracy of the estimated model is a key issue in data-driven modeling of SIMO systems.

For SIMO systems, our aim in this contribution is to

quantify the model error induced by stochastic disturbances and noise in prediction error identification. We consider the situation where the true system can accurately be described by the model, i.e., the true system lies within the set of models used, and thus the bias (systematic) error is zero. Then, the model error mainly consists of the variance-error, which is caused by disturbances and noise when the model is estimated using a finite number of input-output samples. In particular, we shall quantify the variance error in terms of the noise covariance matrix, input spectrum and model structure. These quantities are also crucial in answering the questions we have posed above, namely, when and how much, adding a sensor pays off in terms of accuracy of the estimated target transfer function.

There are expressions for the model error, in terms of the asymptotic (in sample size) (co-) variance of the estimated parameters, for a variety of identification methods for multi-output systems (Ljung, 1999),(Ljung and Caines, 1980). Even though these expressions correspond to the Cramér-Rao lower bound, they are typically rather opaque, in that it is difficult to discern how the model accuracy is influenced by the aforementioned quantities. There is a well-known expression for the variance of an estimated frequency response function that lends itself to the kind of analysis we wish to facilitate (Ljung, 1985; Yuan and Ljung, 1984; Zhu, 1989). This formula is given in its SIMO version by (see e.g., Zhu (2001))

$$\text{Cov}\,\hat{G}(e^{j\omega}) \approx \frac{n}{N} \Phi_u(\omega)^{-1} \Phi_v(\omega), \qquad (1)$$

where Φ_u is the input spectrum and Φ_v is the spectrum of the noise affecting the outputs. Notice that the expression is valid for large number of samples N and large model order n. For finite model order there are mainly results for SISO models (Hjalmarsson and Ninness, 2006; Ninness and Hjalmarsson, 2004, 2005a,b) and recently, multi-input-single-output (MISO) models. For MISO models, the concept of connectedness (Gevers et $al.$, 2006) gives conditions on when one input can help reduce the variance of an identified transfer function. These results were refined in Mårtensson (2007). For white inputs, it was recently shown in Ramazi et $al.$ (2014) that an increment in the model order of one transfer function leads to an increment in the variance of another estimated transfer function only up to a point, after which the variance levels off. It was also quantified how correlation between the inputs may reduce the accuracy. The results presented here are similar in nature to those in Ramazi et $al.$ (2014), while they regard another special type of multi-variable models, namely multi-input single-output (MISO) models. Note that variance expressions for the special case of SIMO cascade structures are found in Wahlberg et $al.$ (2009), Everitt et $al.$ (2013).

1.1 Contribution of this paper

As a motivation, let us first introduce the following two output example. Consider the model:

$$y_1(t) = \theta_{1,1}u(t-1) + e_1(t),$$
$$y_2(t) = \theta_{2,2}u(t-2) + e_2(t),$$

where the input $u(t)$ is white noise and $e_k, k = 1, 2$ is measurement noise. We consider two different types of measurement noise (uncorrelated with the input). In the first case, the noise is perfectly correlated, let us for simplicity assume that $e_1(t) = e_2(t)$. For the second case, $e_1(t)$ and $e_2(t)$ are independent. It turns out that in the first case we can perfectly recover the parameters $\theta_{1,1}$ and $\theta_{2,2}$, while, in the second case we do not improve the accuracy of the estimate of $\theta_{1,1}$ by also using the measurement $y_2(t)$. The reason for this difference is that, in the first case, we can construct the noise free equation

$$y_1(t) - y_2(t) = \theta_{1,1}u(t-1) - \theta_{2,2}u(t-2)$$

and we can perfectly recover $\theta_{1,1}$ and $\theta_{2,2}$, while in the second case neither $y_2(t)$ nor $e_2(t)$ contain information about $e_1(t)$.

Also the model structure plays an important role for the benefit of the second sensor. To this end, we consider a third case, where again $e_1(t) = e_2(t)$. This time, the model structure is slightly different:

$$y_1(t) = \theta_{1,1}u(t-1) + e_1(t),$$
$$y_2(t) = \theta_{2,1}u(t-1) + e_2(t).$$

In this case, we can construct the noise free equation

$$y_1(t) - y_2(t) = (\theta_{1,1} - \theta_{2,2})u(t-1).$$

The fundamental difference is that now only the difference $(\theta_{1,1} - \theta_{2,1})$ can be recovered exactly, but not the parameters $\theta_{1,1}$ and $\theta_{2,1}$ themselves. They can be identified from $y_1(t)$ and $y_2(t)$ separately, as long as y_1 and y_2 are measured. A similar consideration is made in Ljung et $al.$ (2011), where SIMO cascade systems are considered.

This paper will generalize these observations in the following contributions:

- We provide novel expressions for the variance-error of an estimated frequency response function of a SIMO linear model in orthonormal basis form, or equivalently, in fixed denominator form (Ninness et $al.$, 1999). The expression reveals how the noise correlation structure, model orders and input variance affect the variance-error of the estimated frequency response function.

- For a non-white input spectrum, we show where in the frequency spectrum the benefit of the correlation structure is focused.
- When one module, i.e., the relationship between the input and one of the outputs, is identified using less parameters, we derive the noise correlation structure under which the mentioned model parameterization gives the lowest total variance.

The paper is organized as follows: in Section 2 we define the SIMO model structure under study and provide an expression for the covariance matrix of the parameter estimates. Section 3 contains the main results, namely a novel variance expression for LTI SIMO orthonormal basis function models. The connection with MISO models is explored in Section 4. In Section 5, the main results are applied to a non–white input spectrum. In section 6 we derive the correlation structure that gives the minimum total variance, when one block has less parameters than the other blocks. Numerical experiments illustrating the application of the derived results are presented in Section 7. A final discussion ends the paper in Section 8.
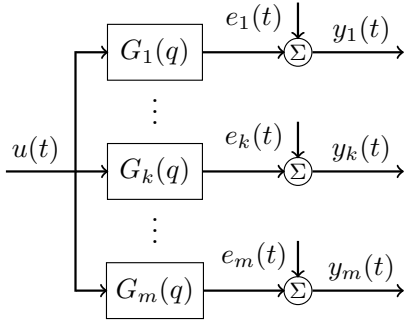
2 Problem Statement

Fig. 1. *Block scheme of the linear SIMO system.*

We consider linear time-invariant dynamic systems with one input and m outputs (see Fig. 1). The model is described as follows:

$$\begin{bmatrix} y_1(t) \\ y_2(t) \\ \vdots \\ y_m(t) \end{bmatrix} = \begin{bmatrix} G_1(q) \\ G_2(q) \\ \vdots \\ G_m(q) \end{bmatrix} u(t) + \begin{bmatrix} e_1(t) \\ e_2(t) \\ \vdots \\ e_m(t) \end{bmatrix}, \qquad (2)$$

where q denotes the forward shift operator, i.e., $qu(t) = u(t+1)$ and the $G_i(q)$ are causal stable rational transfer functions. The G_i are modeled as

$$G_i(q, \theta_i) = \Gamma_i(q)\theta_i, \quad \theta_i \in \mathbb{R}^{n_i}, \quad i = 1, \ldots, m, \quad (3)$$

where $n_1 \leq \ldots \leq n_m$ and $\Gamma_i(q) = \begin{bmatrix} \mathcal{B}_1(q), \ldots, \mathcal{B}_{n_i}(q) \end{bmatrix}$, for some orthonormal basis functions $\{\mathcal{B}_k(q)\}_{k=1}^{n_m}$. Orthonormal with respect to the scalar product defined for complex functions $f(z), g(z) : \mathbb{C} \to \mathbb{C}^{1 \times m}$ as $\langle f, g \rangle := \frac{1}{2\pi} \int_{-\pi}^{\pi} f(e^{i\omega})g^*(e^{i\omega}) \, d\omega$. Let us introduce the vector notation

$$y(t) := \begin{bmatrix} y_1(t) \\ y_2(t) \\ \vdots \\ y_m(t) \end{bmatrix}, \quad e(t) := \begin{bmatrix} e_1(t) \\ e_2(t) \\ \vdots \\ e_m(t) \end{bmatrix}.$$

The noise sequence $\{e(t)\}$ is zero mean and temporally white, but may be correlated in the spatial domain:

$$\begin{aligned} \mathrm{E}\left[e(t)\right] &= 0 \\ \mathrm{E}\left[e(t)e(s)^T\right] &= \delta_{t-s}\Lambda \end{aligned} \qquad (4)$$

for some positive definite matrix covariance matrix Λ, and where $\mathrm{E}\left[\cdot\right]$ is the expected value operator. We express Λ in terms of its Cholesky factorization

$$\Lambda = \Lambda_{CH}\Lambda_{CH}^T, \qquad (5)$$

where $\Lambda_{CH} \in \mathbb{R}^{m \times m}$ is lower triangular, i.e.,

$$\Lambda_{CH} = \begin{bmatrix} \gamma_{11} & 0 & \ldots & 0 \\ \gamma_{21} & \gamma_{22} & \ldots & 0 \\ \vdots & \ldots & \ddots & 0 \\ \gamma_{m1} & \gamma_{m2} & \cdots & \gamma_{mm} \end{bmatrix} \qquad (6)$$

for some $\{\gamma_{ij}\}$. Also notice that since $\Lambda > 0$,

$$\Lambda^{-1} = \Lambda_{CH}^{-T}\Lambda_{CH}^{-1}. \qquad (7)$$

We summarize the assumptions on input, noise and model as follows:

Assumption 1 *The input $\{u(t)\}$ is zero mean stationary white noise with finite moments of all orders, and variance $\sigma^2 > 0$. The noise $\{e(t)\}$ is zero mean and temporally white, i.e, (4) holds with $\Lambda > 0$, a positive definite matrix. It is assumed that $\mathrm{E}\left[|e(t)|^{4+\rho}\right] < \infty$ for some $\rho > 0$. The data is generated in open loop, that is, the input $\{u(t)\}$ is independent of the noise $\{e(t)\}$. The true input-output behavior of the data generating system can be captured by our model, i.e the true system can be described by (2) and (3) for some parameters $\theta_i^o \in \mathbb{R}^{n_i}$, $i = 1, \ldots, m$, where $n_1 \leq \ldots \leq n_m$. The orthonormal basis functions $\{\mathcal{B}_k(q)\}$ are assumed stable.* $\square$

Remark 2 *Generalization of the model structure:*

3

- *The assumption that the modules have the same orthogonal parameterization in (3) is less restrictive than it might appear at first glance, and made for clarity and ease of presentation. A model consisting of non-orthonormal basis functions can be transformed into this format by a linear transformation, which can be computed by the Gram-Schmidt procedure (Trefethen and Bau, 1997). Notice that fixed denominator models, with the same denominator in all transfer functions, also fit this framework (Ninness et al., 1999). However, it is essential for our analysis that the modules share poles.*
- *It is not necessary to restrict the input to be white noise. A colored input introduces a weighting by its spectrum Φ_u, which means that the basis functions need to be orthogonal with respect to the inner product $\langle f, g \rangle_{\Phi_u} := \langle f\Phi_u, g \rangle$. If $\Phi_u(z) = \sigma^2 R(z)R^*(z)$, where $R(z)$ is a monic stable minimum phase spectral factor; the transformation $\tilde{\Gamma}_i(q) = \sigma^{-1}R(q)^{-1}\Gamma_i(q)$ is a procedure that gives a set of orthogonal basis functions in the weighted space. If we use this parameterization, all the main results of the paper carry over naturally. However, in general, the new parameterization does not contain the same set of models as the original parameterization. Another way is to use the Gram-Schmidt method, which maintains the same model set and the main results are still valid. If we would like to keep the original parameterization, we may, in some cases, proceed as in Section 5 where the input signal is generated by an AR-spectrum.*
- *Note that the assumption that $n_1 \leq \ldots \leq n_m$ is not restrictive as it only represents an ordering of the modules in the system.*

2.1 Weighted least-squares estimate

By introducing $\theta = \left[\theta_1^T, \ldots, \theta_m^T\right]^T \in \mathbb{R}^n$, $n := \sum_{i=1}^m n_i$ and the $n \times m$ transfer function matrix

$$
\tilde{\Psi}(q) = \begin{bmatrix} \Gamma_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \Gamma_m \end{bmatrix},
$$

we can write the model (2) as a linear regression model

$$
y(t) = \varphi^T(t)\theta + e(t), \qquad (8)
$$

where

$$
\varphi^T(t) = \tilde{\Psi}(q)^T u(t).
$$

An unbiased and consistent estimate of the parameter vector θ can be obtained from weighted least-squares, with optimal weighting matrix Λ^{-1} (see, e.g., Ljung

(1999); Söderström and Stoica (1989)). Λ is assumed known, however, this assumption is not restrictive since Λ can be estimated from data. The estimate of θ is given by

$$
\hat{\theta}_N = \left(\sum_{t=1}^N \varphi(t)\Lambda^{-1}\varphi^T(t)\right)^{-1} \sum_{t=1}^N \varphi(t)\Lambda^{-1}y(t). \quad (9)
$$

Inserting (8) in (9) gives

$$
\hat{\theta}_N = \theta + \left(\sum_{t=1}^N \varphi(t)\Lambda^{-1}\varphi^T(t)\right)^{-1} \sum_{t=1}^N \varphi(t)\Lambda^{-1}e(t).
$$

Under Assumption 1, the noise sequence is zero mean, hence $\hat{\theta}_N$ is unbiased. It can be noted that this is the same estimate as the one obtained by the prediction error method and, if the noise is Gaussian, by the maximum likelihood method (Ljung, 1999). It also follows that the asymptotic covariance matrix of the parameter estimates is given by

$$
\text{AsCov} \, \hat{\theta}_N = \left(\text{E}\left[\varphi(t)\Lambda^{-1}\varphi^T(t)\right]\right)^{-1}. \quad (10)
$$

Here $\text{AsCov} \, \hat{\theta}_N$ is the asymptotic covariance matrix of the parameter estimates, in the sense that the asymptotic covariance matrix of a stochastic sequence $\{f_N\}_{N=1}^\infty, f_N \in \mathbb{C}^{1 \times q}$ is defined as [1]

$$
\text{AsCov} \, f_N := \lim_{N \to \infty} N \cdot \text{E}\left[(f_N - \text{E}\left[f_N\right])^* (f_N - \text{E}\left[f_N\right])\right].
$$

In the problem we consider, using Parseval's formula and (7), the asymptotic covariance matrix, (10), can be written as [2]

$$
\text{AsCov} \, \hat{\theta}_N = \left[\frac{1}{2\pi}\int_{-\pi}^{\pi} \Psi(e^{j\omega})\Psi^*(e^{j\omega}) \, d\omega\right]^{-1}
$$
$$
= \langle \Psi, \Psi \rangle^{-1}, \quad (11)
$$

where

$$
\Psi(q) = \frac{1}{\sigma}\tilde{\Psi}(q)\Lambda_{CH}^{-T}. \quad (12)
$$

Note that $\Psi(q)$ is block upper triangular since $\tilde{\Psi}(q)$ is block diagonal and Λ_{CH}^{-T} is upper triangular.

---

[1] This definition is slightly non-standard in that the second term is usually conjugated. For the standard definition, in general, all results have to be transposed, however, all results in this paper are symmetric.

[2] Non-singularity of $\langle \Psi, \Psi \rangle$ usually requires parameter identifiability and persistence of excitation (Ljung, 1999).

2.2 The introductory example

With formal assumptions in place, we now consider the introductory example in greater detail. Consider the model

$$y_1(t) = \theta_{1,1} q^{-1} u(t) + e_1(t), \tag{13}$$
$$y_2(t) = \theta_{2,1} q^{-1} u(t) + \theta_{2,2} q^{-2} u(t) + e_2(t) \tag{14}$$

which uses the delays q^{-1} and q^{-2} as orthonormal basis functions. With $\theta = [\theta_{1,1} \, \theta_{2,1} \, \theta_{2,2}]^T$; the corresponding regression matrix is

$$\varphi(t)^T = \begin{bmatrix} u(t-1) & 0 & 0 \\ 0 & u(t-1) & u(t-2) \end{bmatrix}.$$

The noise vector is generated by

$$\begin{bmatrix} e_1(t) \\ e_2(t) \end{bmatrix} = L w(t) = \begin{bmatrix} 1 & 0 \\ \sqrt{1-\beta^2} & \beta \end{bmatrix} \begin{bmatrix} w_1(t) \\ w_2(t) \end{bmatrix}, \tag{15}$$

where $w_1(t)$ and $w_2(t)$ are uncorrelated white processes with unit variance. The parameter $\beta \in [0, 1]$ tunes the correlation between $e_1(t)$ and $e_2(t)$. When $\beta = 0$, the two processes are perfectly correlated (i.e., identical); conversely, when $\beta = 1$, they are completely uncorrelated. Note that, for every $\beta \in [0, 1]$, one has $\mathrm{E}\left[e_1(t)^2\right] = \mathrm{E}\left[e_2(t)^2\right] = 1$. In fact, the covariance matrix of $e(t)$ becomes

$$\Lambda = L L^T = \begin{bmatrix} 1 & \sqrt{1-\beta^2} \\ \sqrt{1-\beta^2} & 1 \end{bmatrix}.$$

Then, when computing (11) in this specific case gives

$$\mathrm{AsCov}\, \hat{\theta}_N = \frac{1}{\sigma^2} \begin{bmatrix} 1 & \sqrt{1-\beta^2} & 0 \\ \sqrt{1-\beta^2} & 1 & 0 \\ 0 & 0 & \beta^2 \end{bmatrix}. \tag{16}$$

We note that:

$$\mathrm{AsCov} \left[\hat{\theta}_{1,1} \, \hat{\theta}_{2,1}\right]^T = \frac{1}{\sigma^2} \Lambda,$$
$$\mathrm{AsCov}\, \hat{\theta}_{2,2} = \frac{1}{\sigma^2} \beta^2.$$

The above expressions reveals two interesting facts:

(1) The (scalar) variances of $\hat{\theta}_{1,1}$ and $\hat{\theta}_{2,1}$, namely the estimates of parameters of the two modules related to the same time lag, are not affected by possible correlation of the noise processes, i.e., they are independent of the value of β. However, note that the cross correlation between $\hat{\theta}_{1,1}$ and $\hat{\theta}_{2,1}$ in (16):

$$\mathrm{Var}\,(\hat{\theta}_{1,1} - \sqrt{1-\beta^2}\hat{\theta}_{2,1})$$
$$= \begin{bmatrix} 1 \\ -\sqrt{1-\beta^2} \end{bmatrix}^T \frac{1}{\sigma^2} \Lambda \begin{bmatrix} 1 \\ -\sqrt{1-\beta^2} \end{bmatrix}$$
$$= \frac{1}{\sigma^2} \beta^2. \tag{17}$$

This cross correlation will induce a cross correlation in the transfer function estimates as well.

(2) As seen in (16), the variance of $\hat{\theta}_{1,2}$ strongly depends on β. In particular, when β tends to 0, one is ideally able to estimate $\hat{\theta}_{1,2}$ perfectly. Note that in the limit case $\beta = 0$ one has $e_1(t) = e_2(t)$, so that (2) can be rearranged to obtain the noise-free equation

$$y_1(t) - y_2(t) = (\theta_{1,1} - \theta_{2,1})u(t-1) + \theta_{1,2}u(t-2),$$

which shows that both $\theta_{1,2}$ and the difference $\theta_{1,1} - \theta_{2,1}$ can be estimated perfectly. This can of course also be seen from (16), cf. (17).

The example shows that correlated measurements can be favorable for estimating for estimating certain parameters, but not necessarily for all. The main focus of this article is to generalize these observations to arbitrary basis functions, number of systems and number of estimated parameters. Additionally, the results are used to derive the optimal correlation structure of the noise. But first, we need some technical preliminaries.

2.3 Review of the geometry of the asymptotic variance

The following lemma is instrumental in deriving the results that follow.

Lemma 3 (*Lemma II.9 in Hjalmarsson and Mårtensson (2011)*) *Let $J : \mathbb{R}^n \to \mathbb{C}^{1 \times q}$ be differentiable with respect to θ, and $\Psi \in \mathcal{L}_2^{n \times m}$; let $\mathcal{S}_\Psi$ be the subspace of $\mathcal{L}^{1 \times m}$ spanned by the rows of Ψ and $\{\mathcal{B}_k^{\mathcal{S}}\}_{k=1}^r, r \leq n$ be an orthonormal basis for $\mathcal{S}_\Psi$. Suppose that $J'(\theta^o) \in \mathbb{C}^{n \times q}$ is the gradient of J with respect to θ and $J'(\theta^o) = \Psi(z_o)L$ for some $z_0 \in \mathbb{C}$ and $L \in \mathbb{C}^{m \times q}$. Then*

$$\mathrm{AsCov}\, J(\hat{\theta}_N) = L^* \sum_{k=1}^r \mathcal{B}_k^{\mathcal{S}}(z_o)^* \mathcal{B}_k^{\mathcal{S}}(z_o)\, L. \tag{18}$$

2.4 Non-estimable part of the noise

As seen in the example of Section 2.2, strong noise correlation may be helpful in the estimation. In fact, the

variance error will depend on the non-estimable part of the noise, i.e., the part that cannot be linearly estimated from other noise sources. To be more specific, define the signal vector $e_{j\setminus i}(t)$ to include the noise sources from module 1 to module j, with the one from module i excluded, i.e.,

$$
e_{j\setminus i}(t) :=
\begin{cases}
\begin{bmatrix} e_1(t), \ldots, e_j(t) \end{bmatrix}^T & j < i, \\
\begin{bmatrix} e_1(t), \ldots, e_{i-1}(t) \end{bmatrix}^T & j = i, \\
\begin{bmatrix} e_1(t), \ldots, e_{i-1}(t), e_{i+1}(t), \ldots, e_j(t) \end{bmatrix}^T & j > i.
\end{cases}
$$

Now, the linear minimum variance estimate of e_i given $e_{j\setminus i}(t)$, is given by

$$\hat{e}_{i|j}(t) := \varrho_{ij}^T e_{j\setminus i}(t). \tag{19}$$

Introduce the notation

$$\lambda_{i|j} := \mathrm{Var}\left[e_i(t) - \hat{e}_{i|j}(t) \right], \tag{20}$$

with the convention that $\lambda_{i|0} := \lambda_i$. The vector ϱ_{ij} in (19) is given by

$$\varrho_{ij} = \left[\mathrm{Cov}\, e_{j\setminus i}(t) \right]^{-1} \mathrm{E}\left[e_{j\setminus i}(t) e_i(t) \right].$$

We call

$$e_i(t) - \hat{e}_{i|j}(t)$$

the non-estimable part of $e_i(t)$ given $e_{j\setminus i}(t)$.

Definition 4 *When $\hat{e}_{i|j}(t)$ does not depend on $e_k(t)$, where $1 \leq k \leq j$, $k \neq i$, we say that $e_i(t)$ is orthogonal to $e_k(t)$ conditionally to $e_{j\setminus i}(t)$.*

The variance of the non-estimable part of the noise is closely related to the Cholesky factor of the covariance matrix Λ. We have the following lemma.

Lemma 5 *Let $e(t) \in \mathbb{R}^m$ have zero mean and covariance matrix $\Lambda > 0$. Let Λ_{CH} be the lower triangular Cholesky factor of Λ, i.e., Λ_{CH} satisfies (5), with $\{\gamma_{ik}\}$ as its entries as defined by (6). Then for $j < i$,*

$$\lambda_{i|j} = \sum_{k=j+1}^{i} \gamma_{ik}^2.$$

Furthermore, $\gamma_{ij} = 0$ is equivalent to that $e_i(t)$ is orthogonal to $e_j(t)$ conditionally to $e_{j\setminus i}(t)$.

PROOF. See Appendix A. ■

Similar to the above, for $i \leq m$, we also define

$$e_{i:m}(t) := \begin{bmatrix} e_i(t) & \ldots & e_m(t) \end{bmatrix}^T,$$

and for $j < i$ we define $\hat{e}_{i:m|j}(t)$ as the linear minimum variance estimate of $e_{i:m}(t)$ based on the other signals $e_{j\setminus i}(t)$, and

$$\Lambda_{i:m|j} := \mathrm{Cov}\left[e_{i:m}(t) - \hat{e}_{i:m|j}(t) \right].$$

As a small example of why this formulation is useful, consider the covariance matrix below, where there is correlation between any pair $(e_i(t), e_j(t))$:

$$
\Lambda = \begin{bmatrix} 1 & 0.6 & 0.9 \\ 0.6 & 1 & 0.54 \\ 0.9 & 0.54 & 1 \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & 0 & 0 \\ 0.6 & 0.8 & 0 \\ 0.9 & 0 & 0.44 \end{bmatrix}}_{\Lambda_{CH}} \begin{bmatrix} 1 & 0.6 & 0.9 \\ 0 & 0.8 & 0 \\ 0 & 0 & 0.44 \end{bmatrix}.
$$

From the Cholesky factorization above we see that, since γ_{32} is zero, Lemma 5 gives that $e_3(t)$ is orthogonal to $e_2(t)$ given $e_{2\setminus 3}(t)$, i.e., there is no information about $e_3(t)$ in $e_2(t)$ if we already know $e_1(t)$. This is not apparent from Λ where every entry is non-zero. If we know $e_1(t)$ a considerable part of $e_2(t)$ and $e_3(t)$ can be estimated. Without knowing $e_1(t)$, $\lambda_1 = \lambda_2 = \lambda_3 = 1$, while if we know $e_1(t)$, $\lambda_{2|1} = 0.64$ and $\lambda_{3|1} = 0.19$.

3 Main results

In this section, we present novel expressions for the variance-error of an estimated frequency response function. The expression reveals how the noise correlation structure, model orders and input variance affect the variance-error of the estimated frequency response function. We will analyze the effect of the correlation structure of the noise on the transfer function estimates. To this end, collect all m transfer functions into

$$G := \begin{bmatrix} G_1 & G_2 & \ldots & G_m \end{bmatrix}.$$

For convenience, we will simplify notation according to the following definition:

Definition 6 *The asymptotic covariance of $\hat{G}(e^{j\omega_0}) := G(e^{j\omega_0}, \hat{\theta}^N)$ for the fixed frequency ω_0 is denoted by*

$$\mathrm{AsCov}\, \hat{G}.$$

In particular, the variance of $\hat{G}_i(e^{j\omega_0}) := G_i(e^{j\omega_0}, \hat{\theta}_i^N)$ for the fixed frequency ω_0 is denoted by

$$\mathrm{AsVar}\, \hat{G}_i.$$

6

Define χ_k as the index of the first system that contains the basis function $\mathcal{B}_k(e^{j\omega_0})$. Notice that $\chi_k - 1$ is the number of systems that do not contain the basis function.

Theorem 7 *Let Assumption 1 hold. Suppose that the parameters $\theta_i \in \mathbb{R}^{n_i}$, $i = 1, \ldots, m$, are estimated using weighted least-squares (9). Let the entries of θ be arranged as follows:*

$$\bar{\theta} = [\theta_{1,1} \ldots \theta_{m,1} \; \theta_{1,2} \ldots \theta_{m,2} \ldots \theta_{1,n_1} \ldots$$
$$\ldots \theta_{m,n_1} \theta_{2,n_1+1} \ldots \theta_{m,n_1+1} \ldots \theta_{m,n_m}]^T. \quad (21)$$

and the corresponding weighted least-squares estimate be denoted by $\hat{\bar{\theta}}$. Then, the covariance of $\hat{\bar{\theta}}$ is

$$\mathrm{AsCov}\, \hat{\bar{\theta}} = \frac{1}{\sigma^2}\, \mathrm{diag}(\Lambda_{1:m}, \Lambda_{\chi_2:m|\chi_2-1}, \ldots,$$
$$\ldots, \Lambda_{\chi_{n_m}:m|\chi_{n_m}-1}). \quad (22)$$

In particular, the covariance of the parameters related to basis function number k is given by

$$\mathrm{AsCov}\, \hat{\bar{\theta}}_k = \frac{1}{\sigma^2}\Lambda_{\chi_k:m|\chi_k-1}, \quad (23)$$

where

$$\hat{\bar{\theta}}_k = \left[\hat{\theta}_{\chi_k,k} \ldots \hat{\theta}_{m,k}\right]^T,$$

and where, for $\chi_k \le i \le m$,

$$\mathrm{AsVar}\, \hat{\theta}_{i,k} = \frac{\lambda_{i|\chi_k-1}}{\sigma^2}. \quad (24)$$

It also holds that

$$\mathrm{AsCov}\, \hat{G} = \sum_{k=1}^{n_m} \begin{bmatrix} \mathbf{0}_{\chi_k-1} & \mathbf{0} \\ \mathbf{0} & \mathrm{AsCov}\,\hat{\bar{\theta}}_k \end{bmatrix} |\mathcal{B}_k(e^{j\omega_0})|^2, \quad (25)$$

where $\mathrm{AsCov}\, \hat{\bar{\theta}}_k$ is given by (23) and $\mathbf{0}_{\chi_k-1}$ is a $\chi_k - 1 \times \chi_k - 1$ matrix with all entries equal to zero. For $\chi_k = 1$, $\mathbf{0}_{\chi_k-1}$ is an empty matrix. In (25), $\mathbf{0}$ denotes zero matrices of dimensions compatible to the diagonal blocks.

PROOF. See Appendix B. ∎

Remark 8 *The covariance of $\hat{\bar{\theta}}_k$, which contain the parameters related to basis function k, is determined by which other models share the basis function $\mathcal{B}_k$; cf. (17) of the introductory example. The asymptotic covariance of $\hat{G}$ can be understood as a sum of the contributions from each of the n_m basis functions. The covariance contribution from a basis function $\mathcal{B}_k$ is weighted by $|\mathcal{B}_k(e^{j\omega_0})|^2$ and only affects the covariance between systems that contain that basis function, as visualized in Figure 2.*
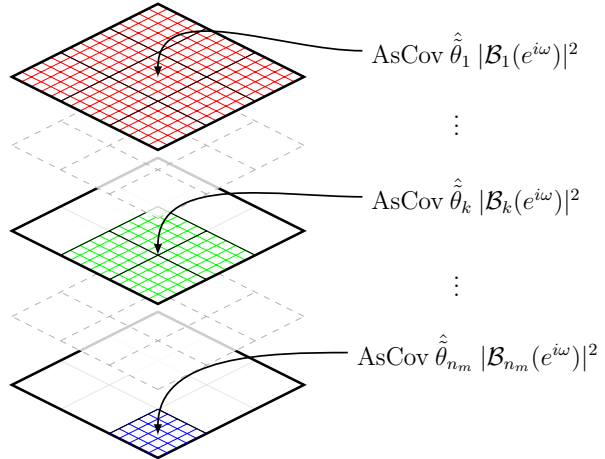

Fig. 2. A graphical representation of AsCov $\hat{G}$ where each term of the sum in (25) is represented by a layer. A basis function only affects the covariance between modules that also contain that basis function. Thus, the first basis function affects the complete covariance matrix while the last basis function n_m only affects modules $\chi_{n_m}, \ldots, m$.

Remark 9 *The orthogonal basis functions correspond to a decomposition of the output signals into orthogonal components and the problem in a sense becomes decoupled. As an example, consider the system described by*

$$y_1(t) = \theta_{1,1}\mathcal{B}_1(q)u(t) + e_1(t),$$
$$y_2(t) = \theta_{2,1}\mathcal{B}_1(q)u(t) + e_2(t),$$
$$y_3(t) = \theta_{3,1}\mathcal{B}_1(q)u(t) + \theta_{3,2}\mathcal{B}_2(q)u(t) + e_3(t), \quad (26)$$

Suppose that we are interested in estimating $\theta_{3,2}$. For this parameter, (24) becomes

$$\mathrm{AsVar}\, \hat{\theta}_{3,2} = \frac{\lambda_{3|2}}{\sigma^2} \quad (27)$$

To understand the mechanisms behind this expression, let $u_1(t) = \mathcal{B}_1(q)u(t)$, and $u_2(t) = \mathcal{B}_2(q)u(t)$ so that the system can be visualized as in Figure 3, i.e., we can consider u_1 and u_2 as separate inputs.

First we observe that it is only y_3 that contains information about $\theta_{3,2}$, and the term $\theta_{3,1}u_1$ contributing to y_3 is a nuisance from the perspective of estimating $\theta_{3,2}$. This term vanishes when $u_1 = 0$ and we will not be able to achieve better accuracy than the optimal estimate of $\theta_{3,2}$ for this idealized case. So let us study this setting first. Straightforward application of the least-squares method, using u_2 and y_3, gives an estimate of $\theta_{3,2}$ with variance λ/σ^2, which is larger than (27) when e_3 depends on e_1 and e_2. However, in this idealized case, $y_1 = e_1$ and $y_2 = e_2$, and these signals can thus be used to estimate e_3. This estimate can then be subtracted from y_3 before the least-squares method is applied. The remaining noise in y_3 will have variance $\lambda_{3|2}$, if e_3 is optimally estimated (see (19)–(20)), and hence the least-squares estimate will now have variance $\lambda_{3|2}/\sigma^2$, i.e., the same as (27).

To understand why it is possible to achieve the same accuracy as this idealized case when u_1 is non-zero, we need to observe that our new inputs $u_1(t)$ and $u_2(t)$ are orthogonal (uncorrelated)[3]. Returning to the case when only the output y_3 is used for estimating $\theta_{3,2}$, this implies that we pay no price for including the term $\theta_{3,1}u_1$ in our model, and then estimating $\theta_{3,1}$ and $\theta_{3,2}$ jointly, i.e., the variance of $\hat{\theta}_{3,2}$ will still be λ/σ^2[4]. The question now is if we can use y_1 and y_2 as before to estimate e_3. Perhaps surprisingly, we can use the same estimate as when u_1 was zero. The reader may object that this estimate will now, in addition to the previous optimal estimate of e_3, contain a term which is a multiple of u_1. However, due to the orthogonality between u_1 and u_2, this term will only affect the estimate of $\theta_{3,1}$ (which we anyway were not interested in, in this example), and the accuracy of the estimate of $\theta_{3,2}$ will be $\lambda_{3|2}/\sigma^2$, i.e. (27). Figure 4 illustrates the setting with $\tilde{y}_3$ denoting y_3 subtracted by the optimal estimate of e_3. In the figure, the new parameter $\tilde{\theta}_{3,1}$ reflects that the relation between u_1 and $\tilde{y}_3$ is different from $\theta_{3,1}$ as discussed above. A key insight from this discussion is that for the estimate of a parameter in the path from input i to output j, it is only outputs that are not affected by input i that can be used to estimate the noise in output j; when this particular parameter is estimated, using outputs influenced by input i will introduce a bias, since the noise estimate will then contain a term that is not orthogonal to this input. In (24), this manifests itself in that the numerator is $\lambda_{i|\chi_k-1}$, only the $\chi_k - 1$ first systems do not contain u_i.
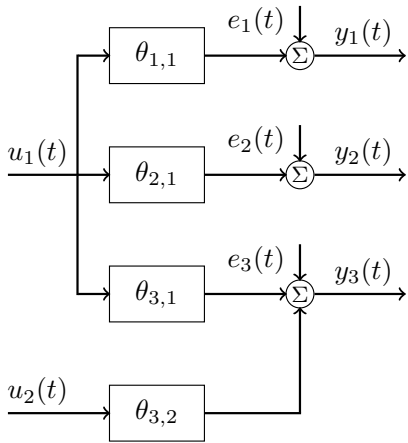

Fig. 3. The SIMO system of Remark 9, described by (26).

We now turn our attention to the variance of the individual transfer function estimates.

---

[3] This since $u(t)$ is white and $\mathcal{B}_1$ and $\mathcal{B}_2$ are orthonormal.
[4] With u_1 and u_2 correlated, the variance will be higher, see Section 4 for a further discussion of this topic.
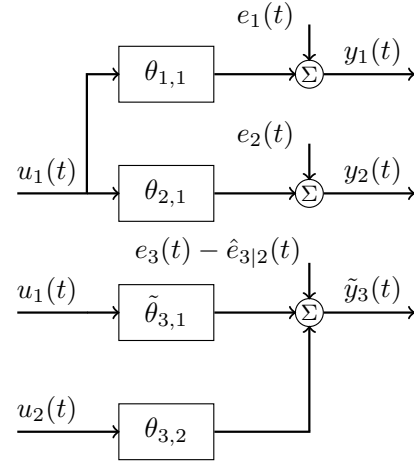

Fig. 4. The SIMO system of Remark 9, described by (26), but with $\tilde{y}_3$ denoting y_3 subtracted by the optimal estimate of e_3 and $\tilde{\theta}_{3,1}$ reflects that the relation between u_1 and $\tilde{y}_3$ is different from $\theta_{3,1}$.

Corollary 10 *Let the same assumptions as in Theorem 7 hold. Then, for any frequency ω_0, it holds that*

$$\text{AsVar}\,\hat{G}_i = \sum_{k=1}^{n_i} |\mathcal{B}_k(e^{j\omega_0})|^2 \text{AsVar}\,\hat{\theta}_{i,k}, \qquad (28)$$

where

$$\text{AsVar}\,\hat{\theta}_{i,k} = \frac{\lambda_{i|\chi_k-1}}{\sigma^2}, \qquad (29)$$

and $\lambda_{i|j}$ is defined in (20).

PROOF. Follows from Theorem 7, since (28) is a diagonal element of (25). ∎

From Corollary 10, we can tell when increasing the model order of G_j will increase the asymptotic variance of $\hat{G}_i$.

Corollary 11 *Under the same conditions as in Theorem 7, if we increase the number of estimated parameters of G_j from n_j to n_j+1, the asymptotic variance of G_i will increase, if and only if all the following conditions hold:*

(1) $n_j < n_i$,
(2) $e_i(t)$ is not orthogonal to $e_j(t)$ conditioned on $e_{j\setminus i}(t)$,
(3) $|\mathcal{B}_{n_j+1}(e^{j\omega_0})|^2 \neq 0$.

PROOF. See Appendix C. ∎

Remark 12 *Corollary 11 explicitly tells when an increase in the model order of G_j from n_j to $n_j + 1$ will increase the variance of G_i. If $n_j \geq n_i$, there will be no increase in the variance of G_i, no matter how many additional parameters we introduce to the model G_j, which was also seen the introductory example in Section 2.2. Naturally, if $e_i(t)$ is orthogonal to $e_j(t)$ conditioned on $e_{j \setminus i}(t)$, $\hat{e}_{i|j}(t)$ does not depend on $e_j(t)$ and there is no increase in variance of $\hat{G}_i$, cf. Remark 9.*

3.1 A graphical representation of Corollary 11

Following the notation in Bayesian Networks (Koski and Noble, 2012), Conditions 1) and 2) in Corollary 11 can be interpreted graphically. Each module is represented by a vertex in a weighted directed acyclic graph $\mathcal{G}$. Let the vertices be ordered by model order, i.e., let the first vertex correspond to $\hat{G}_1$. Under Assumption 1, with the additional assumption that module i is the first module with order n_i, let there be an edge, denoted by $j \rightarrow i$, from vertex j to i, $j < i$, if $e_i(t)$ is not orthogonal to $e_j(t)$ conditioned on $e_{j \setminus i}(t)$. Notice that this is equivalent to $\gamma_{ij} \neq 0$. Let the weight of the edge be γ_{ij} and define the parents of vertex i to be all nodes with a link to vertex i, i.e., $pa_{\mathcal{G}}(i) := \{j : j \rightarrow i\}$. Then, (29), together with Lemma 5, shows that only outputs corresponding to parents of node i affect the asymptotic variance. Indeed, a vertex without parents has variance

$$\text{AsVar}\,\hat{G}_i = \frac{\lambda_i}{\sigma^2} \sum_{k=1}^{n_i} |\mathcal{B}_k(e^{j\omega_0})|^2, \qquad (30)$$

which corresponds to (28) with

$$\lambda_{i|0} = \ldots = \lambda_{i|i-1} = \lambda_i.$$

Thus, AsVar $\hat{G}_i$ is independent of the model order of the other modules.

As an example, consider four systems with the lower Cholesky factor of the covariance of $e(t)$ given by:

$$
\Lambda =
\begin{bmatrix}
1 & 0.4 & 0.2 & 0 \\
0.4 & 1.16 & 0.08 & 0.3 \\
0.2 & 0.08 & 1.04 & 0 \\
0 & 0.3 & 0 & 1.09
\end{bmatrix}
$$
$$
= \underbrace{\begin{bmatrix}
1 & 0 & 0 & 0 \\
0.4 & 1 & 0 & 0 \\
0.2 & 0 & 1 & 0 \\
0 & 0.3 & 0 & 1
\end{bmatrix}}_{\Lambda_{CH}}
\begin{bmatrix}
1 & 0 & 0 & 0 \\
0.4 & 1 & 0 & 0 \\
0.2 & 0 & 1 & 0 \\
0 & 0.3 & 0 & 1
\end{bmatrix}^T \qquad (31)
$$

If the model orders are distinct, the corresponding graph is given in Figure 5, where one can see that AsVar $\hat{G}_4$ depends on y_2 (and on y_4 of course), but depends neither on y_3 nor y_1 since $\gamma_{43} = \gamma_{41} = 0$, AsVar $\hat{G}_3$ depends on y_1, but not on y_2 since $\gamma_{32} = 0$ and AsVar $\hat{G}_2$ depends on y_1, while the variance of $\hat{G}_1$ is given by (30). If $n_2 = n_4$, the first condition of Corollary 11 is not satisfied and we have to cut the edge between $\hat{G}_4$ and $\hat{G}_2$. Similarly, if $n_1 = n_2$, we have to cut the edge between $\hat{G}_2$ and $\hat{G}_1$, and if additionally $n_1 = n_2 = n_3$, we have to cut the edge between $\hat{G}_3$ and $\hat{G}_1$.
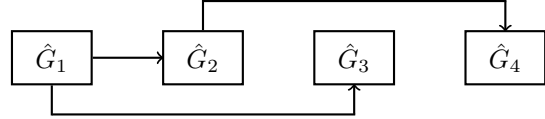

Fig. 5. Graphical representation of Conditions 1) and 2) in Corollary 11 for the Cholesky factor given in (31).

4 Connection between MISO and SIMO

There is a strong connection between the results presented here and those regarding MISO systems presented in Ramazi *et al.* (2014). We briefly restate the problem formulation and some results from Ramazi *et al.* (2014) to show the connection. The MISO data generating system is in some sense the dual of the SIMO case. With m spatially correlated inputs and one output, a MISO system is described by

$$
y(t) =
\begin{bmatrix}
G_1(q) & G_2(q) & \ldots & G_m(q)
\end{bmatrix}
\begin{bmatrix}
u_1(q) \\
u_2(q) \\
\vdots \\
u_m(q)
\end{bmatrix}
+ e(t).
$$

The input sequence $\{u(t)\}$ is zero mean and temporally white, but may be correlated in the spatial domain,

$$\text{E}\left[u(t)\right] = 0$$
$$\text{E}\left[u(t)u(s)^T\right] = \delta_{t-s}\Sigma_u,$$

for some positive definite matrix covariance matrix $\Sigma_u = \Sigma_{CH}\Sigma_{CH}^T$, where Σ_{CH} is the lower triangular Cholesky factor of Σ. The noise $e(t)$ is zero mean and has variance λ. The asymptotic covariance of the estimated parameters can be expressed using (11) with

$$\Psi = \Psi^{\text{MISO}} := \tilde{\Psi}\Sigma_{CH}. \qquad (32)$$

We make the convention that $\sum_{k=k_1}^{k_2} x_k = 0$ whenever $k_1 > k_2$.

9

Theorem 13 (Theorem 4 in Ramazi *et al.* (2014))
Under Assumption 1, but with $n_1 \geq n_2 \geq \ldots, \geq n_m$, for any frequency ω_0 it holds that

$$\text{AsVar } \hat{G}_i = \sum_{j=i}^{m} \frac{\lambda}{\sigma_{i|j}^2} \sum_{k=n_{j+1}+1}^{n_j} |\mathcal{B}_k(e^{j\omega_0})|^2, \qquad (33)$$

where $n_{m+1} := 0$ and $\sigma_{i|j}^2$ is the variance of the non-estimable part of $u_i(t)$ given $u_{j\backslash i}(t)$.

Corollary 14 (Corollary 6 in Ramazi *et al.* (2014))
Under Assumption 1, but with $n_1 \geq n_2 \geq \ldots, \geq n_m$. Suppose that the order of block j is increased from n_j to n_{j+1}. Then there is an increase in the asymptotic variance of $\hat{G}_i$ if and only if all the following conditions hold:

(1) $n_j < n_i$,
(2) $u_i(t)$ is not orthogonal to $u_j(t)$ conditioned on $u_{j\backslash i}(t)$,
(3) $|\mathcal{B}_{n_{j+1}}(e^{j\omega_0})|^2 \neq 0$.

Remark 15 *The similarities between Corollary 10 and Theorem 13, and between Corollary 11 and Corollary 14 are striking. In both cases it is the non-estimable part of the input and noise, respectively, along with the estimated basis functions that are the key determinants for the resulting accuracy. Just as in Corollary 10, Theorem 13 can be expressed with respect to the basis functions:*

$$\text{AsVar } \hat{G}_i = \sum_{k=1}^{n_i} \text{AsVar } \hat{\theta}_{i,k} \, |\mathcal{B}_k(e^{j\omega_0})|^2. \qquad (34)$$

However, now

$$\text{AsVar } \hat{\theta}_{i,k} = \frac{\lambda}{\sigma_{i|\chi_k}^2} \qquad (35)$$

where $\sigma_{i|l}^2$ is determined by the correlation structure of the inputs $u_i(t)$ to the systems $G_i(q, \theta_i)$ that do share basis function $\mathcal{B}_k(q)$ ($i = 1, \ldots, \chi_k$). Note that in the SIMO case we had

$$\text{AsVar } \hat{\theta}_{i,k} = \frac{\lambda_{i|\chi_k}}{\sigma^2}$$

where $\lambda_{i|\chi_k}$ is determined by the correlation structure of the noise sources $e_i(t)$ affecting systems $G_i(q, \theta_i)$ that do not share basis function $\mathcal{B}_k(q)$ ($i = 1, \ldots, \chi_k$). Note that (33) found in Ramazi et al. (2014) does not draw the connection to the variance of the parameters. This is made explicit in the alternate expressions (35) and (34).

The correlation between parameters related to the same basis functions is not explored in Ramazi *et al.* (2014). In fact, it is possible to follow the same line of reasoning leading to Theorem 7 and arrive at the counter-part for

MISO systems. Let the first χ_k systems contain basis function k, so

$$\text{AsVar } \hat{\hat{\theta}}_k^{MISO} = \lambda \, \Sigma_{1:\chi_k}^{-1}$$

where $\Sigma_{1:\chi_k}$ denotes the covariance matrix of the first χ_k inputs. Hence

$$\text{AsCov } \hat{G} = \lambda \sum_{k=1}^{n_1} \begin{bmatrix} \Sigma_{1:\chi_k}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0}_{m-\chi_k} \end{bmatrix} |\mathcal{B}_k(e^{j\omega_0})|^2,$$

and

$$\text{AsCov } \hat{\hat{\theta}}^{MISO} = \lambda \, \text{diag}(\Sigma_{1:\chi_1}^{-1}, \Sigma_{1:\chi_2}^{-1}, \ldots, \Sigma_{1:\chi_{n_m}}^{-1}). \quad (36)$$

Note that, while the correlation between the noise sources is beneficial, the correlation in the input is detrimental for the estimation accuracy. Intuitively, if we use the same input to parallel systems, and only observe the sum of the outputs, there is no way to determine the contribution from the individual systems. On the other hand, as observed in the example in Section 2.2, if the noise is correlated, we can construct equations with reduced noise and improve the accuracy of our estimates.

This difference may also be understood from the structure of Ψ, which through (11) determines the variance properties of any estimate. Consider a single SISO system G_1 as the basic case. For the SIMO structure considered in this paper, as noted before, Ψ^{SIMO} of (12) is block upper triangular with m columns (the number of outputs), while Ψ^{MISO} is block lower triangular with as many columns as inputs. Ψ^{MISO} is block lower triangular since $\tilde{\Psi}$ is block diagonal and Σ_{CH} is lower triangular in (32). Adding an output y_j to the SIMO structure corresponds to extending Ψ^{SIMO} with one column (and n_j rows):

$$\Psi_e^{\text{SIMO}} = \begin{bmatrix} \Psi^{\text{SIMO}} & \star \\ 0 & \star \end{bmatrix}, \qquad (37)$$

where the zero comes from that Ψ_e^{SIMO} also is block upper triangular. Since Ψ^{MISO} is block lower triangular, adding an input u_j to the MISO structure extends Ψ^{MISO} with n_j rows (and one column):

$$\Psi_e^{\text{MISO}} = \begin{bmatrix} \Psi^{\text{MISO}} & 0 \\ \star & \star \end{bmatrix}, \qquad (38)$$

where $\star$ denotes the added column and added row respectively. Addition of columns to Ψ decreases the variance of G_1, while addition of rows increases the variance. First, a short motivation of this will be given. Second, we will discuss the implication for the variance analysis.

Addition of one more column to Ψ in (37) decreases the variance of G_1. With $\Psi = \begin{bmatrix} \psi_1 & \dots & \psi_m \end{bmatrix}$, $\langle \Psi, \Psi \rangle = \sum_k^m \langle \psi_k, \psi_k \rangle$, where $\langle \psi_k, \psi_k \rangle \geq 0$ for every k. The variance of the parameter estimate $\hat{\theta}_N$ decreases with the addition of a column, since

$$\langle \Psi_e, \Psi_e \rangle^{-1} \leq [\langle \psi_{m+1}, \psi_{m+1} \rangle + \langle \Psi, \Psi \rangle]^{-1}.$$

On the other hand, addition of rows leads to an increase in variance of $\hat{G}_1$, e.g., consider (18) in Lemma 3,

$$\text{AsCov } G_1(\hat{\theta}_N) = L^T \sum_{k=1}^{r} \mathcal{B}_k^{\mathcal{S}}(z_o)^* \mathcal{B}_k^{\mathcal{S}}(z_o) \, L,$$

where $L = \Sigma_{CH}^{-T} \begin{bmatrix} 1 & 0 & \dots & 0 \end{bmatrix}^T$ for any number of inputs, and $\{\mathcal{B}_k^{\mathcal{S}}\}_{k=1}^{r}$ is a basis for the linear span of the rows of Ψ^{MISO}. As can be seen from (38), the first rows of Ψ_e^{MISO} are the same as for Ψ^{MISO} and the first r basis functions can therefore be taken the same (with a zero in the last column). To accommodate for the extra rows, n_e extra basis functions $\{\mathcal{B}_k^{\mathcal{S}}\}_{k=r+1}^{r+n_e}$ are needed. Thus, $\{\mathcal{B}_k^{\mathcal{S}}\}_{k=1}^{r+n_e}$ is a basis for the linear span of Ψ_e^{MISO}. We see that the variance of $G_1(\hat{\theta}_N^e)$ is larger than $G_1(\hat{\theta}_N)$ since

$$\text{AsCov } G_1(\hat{\theta}_N^e) = \text{AsCov } G_1(\hat{\theta}_N)$$
$$+ L^T \sum_{k=r+1}^{r+n_e} \mathcal{B}_k^{\mathcal{S}}(z_o)^* \mathcal{B}_k^{\mathcal{S}}(z_o) \, L,$$

and $L^T \mathcal{B}_k^{\mathcal{S}}(z_o)^* \mathcal{B}_k^{\mathcal{S}}(z_o) L \geq 0$ is positive semidefinite for every k.

Every additional input of the MISO system corresponds to addition of rows to Ψ. The increase is strictly positive provided that the explicit conditions in Corollary 14 hold.

Every additional output of the SIMO system corresponds to the addition of one more column to Ψ. However, the benefit of the additional columns is reduced by the additional rows arising from the additional parameters that need to be estimated, cf. Corollary 11 and the preceding discussion. When the number of additional parameters has reached n_1 or if $e_1(t)$ is orthogonal to $e_j(t)$ conditioned on $u_{j \setminus 1}(t)$ the benefit vanishes completely. The following examples clarify this point.

Example 16 *We consider the same example as in Section 2.2 for three cases of model orders of the second model, $n_2 = 0, 1, 2$. These cases correspond to Ψ given by the first $n_2 + 1$ rows of*

$$\Psi_e^{SIMO}(q) = \begin{bmatrix} q^{-1} & -\sqrt{1-\beta^2}/\beta q^{-1} \\ 0 & 1/\beta q^{-1} \\ 0 & 1/\beta q^{-2} \end{bmatrix},$$

respectively. When only y_1 is used ($\Psi = q^{-1}$) :

$$\text{AsVar } \hat{\theta}_{1,1} = \langle \Psi, \Psi \rangle^{-1} = 1.$$

When $n_2 = 0$, the second measurement gives a benefit determined by how strong the correlation is between the two noise sources:

$$\text{AsVar } \hat{\theta}_{1,1} = \langle \Psi, \Psi \rangle^{-1} = (1 + (1-\beta^2)/\beta^2)^{-1} = \beta^2.$$

However, already if we have to estimate one parameter in G_2 the benefit vanishes completely, i.e., for $n_2 = 1$:

$$\text{AsCov } \hat{\theta} = \langle \Psi, \Psi \rangle^{-1} = \begin{bmatrix} 1 & \sqrt{1-\beta^2} \\ \sqrt{1-\beta^2} & 1 \end{bmatrix}.$$

The third case, $n_2 = 2$, corresponds to the example in Section 2.2, which shows that the first measurement y_1 improves the estimate of $\theta_{2,2}$ (compared to only estimating $\hat{G}_2$ using y_2):

$$\text{AsVar } \hat{\theta}_{2,2} = \lambda_{2|1} = \beta^2.$$

Example 17 *We consider the corresponding MISO system with unit variance noise $e(t)$ and $u(t)$ instead having the same spectral factor*

$$\Sigma_{CH} = \begin{bmatrix} 1 & 0 \\ \sqrt{1-\beta^2} & \beta \end{bmatrix}.$$

for $\beta \in (0, 1]$. We now study the impact of the second input $u_2(t)$ when

$$G_1(q) = \theta_{1,1} q^{-1} + \theta_{1,2} q^{-2}, \quad G_2(q) = \sum_{k=1}^{n_2} \theta_{2,k} q^{-k}$$

and $n_2 = 0, 1$. These two cases correspond to Ψ given by the first $n_2 + 2$ rows of

$$\Psi_e^{MISO}(q) = \begin{bmatrix} q^{-1} & 0 \\ q^{-2} & 0 \\ \sqrt{1-\beta^2} q^{-1} & \beta q^{-1} \end{bmatrix},$$

respectively. When only u_1 is used or G_2 is known ($\Psi = \begin{bmatrix} q^{-1} & q^{-2} \end{bmatrix}^T$):

$$\text{AsCov } \hat{\theta}_1 = \langle \Psi, \Psi \rangle^{-1} = I.$$

When $n_2 = 1$, the variance of $\hat{\theta}_{1,1}$ is increased depending on the correlation between the two inputs:

$$\text{AsVar } \hat{\theta} = \frac{1}{\beta^2} \begin{bmatrix} 1 & 0 & -\sqrt{1-\beta^2} \\ 0 & \beta^2 & 0 \\ -\sqrt{1-\beta^2} & 0 & 1 \end{bmatrix}.$$

Also notice that the asymptotic covariance of $\left[\hat\theta_{1,1}\ \hat\theta_{2,1}\right]^T$ is given by Σ^{-1}, the inverse of the covariance matrix of $u(t)$ and that AsVar $\hat\theta_{2,2} = \sigma_1^{-1}$. As β goes to zero the variance of $\left[\hat\theta_{1,1}\ \hat\theta_{2,1}\right]^T$ increases and at $\beta = 0$ the two inputs are perfectly correlated and we loose identifiability.

5 Effect of input spectrum

In this section we will see how a non white input spectrum changes the results of Section 3. Using white noise filtered through an AR-filter as input, we may use the developed results to show where in the frequency range the benefit of the correlation structure is focused. An alternative approach, when a non-white input spectrum is used, is to change the basis functions as discussed in Remark 2. However, the effect of the input filter is internalized in the basis functions and it is hard to distinguish the effect of the input filter. We will instead use FIR basis functions for the SIMO system, which are not orthogonal with respect to the inner product induced by the input spectrum, cf. Remark 2. We let the input filter be given by

$$u(t) = \frac{1}{A(q)} w(t) \tag{39}$$

where $w(t)$ is a white noise sequence with variance σ_w^2 and the order n_a of A is less than the order of G_1, i.e., $n_a \leq n_1$. In this case, following the derivations of Theorem 7, it can be shown that

$$\text{AsVar } \hat G_i = \sum_{k=1}^{n_i} \text{AsVar } \hat\theta_{i,k} |\mathcal{B}_k(e^{j\omega_0})|^2 \tag{40}$$

where

$$\text{AsVar } \hat\theta_{i,k} = \frac{\lambda_{i|\chi_k-1}}{\Phi_u(\omega_0)}$$

and the basis functions $\mathcal{B}_k$ have changed due to the input filter. The solutions boil down to finding explicit basis functions $\mathcal{B}_k$ for the well known case (Ninness and Gustafsson, 1997) when

$$\text{Span}\left\{\frac{\Gamma_n}{A(q)}\right\} = \text{Span}\left\{\frac{q^{-1}}{A(q)}, \frac{q^{-2}}{A(q)}, \ldots, \frac{q^{-n}}{A(q)}\right\}$$

where $A(q) = \prod_{k=1}^{n_a}(1 - \xi_k q^{-1})$, $|\xi_k| < 1$ for some set of specified poles $\{\xi_1, \ldots, \xi_{n_a}\}$ and where $n \geq n_a$. Then, it holds (Ninness and Gustafsson, 1997) that

$$\text{Span}\left\{\frac{\Gamma_n}{A(q)}\right\} = \text{Span}\{\mathcal{B}_1, \ldots, \mathcal{B}_n\}$$

where $\{\mathcal{B}_k\}$ are the Takenaka-Malmquist functions given by

$$\mathcal{B}_k(q) := \frac{\sqrt{1 - |\xi_k|^2}}{q - \xi_k}\, \phi_{k-1}(q),\ k = 1, \ldots, n$$

$$\phi_k(q) := \prod_{i=1}^{k} \frac{1 - \overline{\xi_i}q}{q - \xi_i},\ \ \phi_0(q) := 1$$

and with $\xi_k = 0$ for $k = n_a + 1, \ldots, n$. We summarize the result in the following theorem:

Theorem 18 *Let the same assumptions as in Theorem 7 hold. Additionally the input $u(t)$ is generated by an AR-filter as in (39). Then for any frequency ω_0 it holds that*

$$\text{AsVar } \hat G_i = \frac{1}{\Phi_u(\omega_0)}\left(\lambda_i \sum_{k=1}^{n_a} \frac{1 - |\xi_k|^2}{|e^{j\omega_0} - \xi_k|^2} + \lambda_i(n_1 - n_a)\right.$$
$$\left. + \sum_{j=2}^{i} \lambda_{i|j-1}(n_j - n_{j-1})\right) \tag{41}$$

PROOF. The proof follows from (40) with the basis functions given by the Takenaka-Malmquist functions and using that $\phi_{k-1}(q)$ is all-pass, and for $k > n_a$, also $\mathcal{B}_k(q)$ is all-pass. This means that $|\mathcal{B}_k(e^{j\omega})|^2 = 1$ for all $k > n_a$. ∎

Remark 19 *The second sum in (41) is where the benefit from the correlation structure of the noise at the other sensors enters through $\lambda_{i|j-1}$. This contribution is weighted by $1/\Phi_u(\omega)$. The benefit thus enters mainly where $\Phi_u(\omega)$ is small. The first sum gives a variance contribution that is less focused around frequencies close to the poles ($|e^{j\omega_0} - \xi_k|^2$ is canceled by $\Phi_u(\omega)$). This contribution is not reduced by correlation between noise sources. Shaping the input thus gives a weaker effect on the asymptotic variance than what is suggested by the asymptotic in model order result (1), and what would be the case if there would be no correlation between noise sources (replacing $\lambda_{i|j-1}$ by λ_i in (41)).*

For the example in Section 2.2 for filtered input with $n_1 = 2$, $n_2 = 3$ and $n_a = 1$, (41) simplifies to

$$\text{AsVar } \hat G_2 = \frac{\lambda_2}{\sigma^2} + \frac{\lambda_2}{\Phi_u(\omega_0)} + \frac{\lambda_{2|1}}{\Phi_u(\omega_0)}. \tag{42}$$

In Figure 6 the variance of $\hat G_2$ for an input filtered with $A(q) = 1 - 0.8q^{-1}$ is presented. The filter is of low-pass type and thus gives high input power at low frequencies which results in low variance at those frequencies. Correlation between noise sources decreases the variance mainly where $\Phi_u(\omega)$ is small, i.e., at higher frequencies.
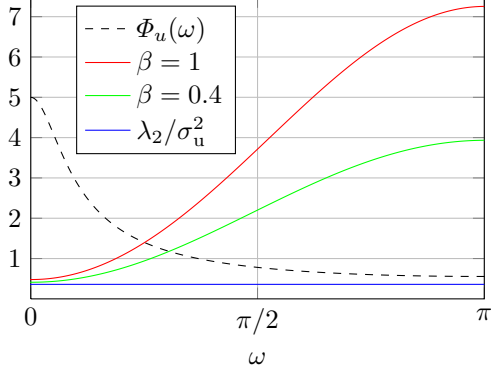
Fig. 6. Asymptotic variance of $\hat{G}_3(e^{j\omega}, \hat{\theta}_2)$ for $\beta = 1$ and $\beta = 0.4$. Also shown is λ_2/σ^2, the first term of AsVar $\hat{G}_3(e^{j\omega}, \hat{\theta}_2)$ in (42).

6 Optimal correlation structure

In this section we characterize the optimal correlation structure, in order to minimize the total variance. In the previous section, we have seen that not estimating a basis function k leads to a reduction in variance of the other parameters related to basis function k. In this section, we will find the optimal correlation structure when the parameterization of the system is such that $n_1 + 1 = n_2 = \ldots = n_m$, i.e., the first module has one parameter less than the others. Let $\tilde{\theta} := [\theta_{2,n_2}, \theta_{3,n_2}, \ldots, \theta_{m,n_2}]^T$, i.e., the sub vector of parameters related to basis function $\mathcal{B}_{n_2}(q)$. Assume the SIMO structure given by (2) and let the input be white noise with unit variance. Recalling Theorem 7, the covariance matrix of $\hat{\theta}$ is given by $\Lambda_{2:m|1}$. In particular, the variance of the entries of $\tilde{\theta}$ is

$$\text{AsVar } \hat{\theta}_{k,n_2} = \frac{\lambda_{k|1}}{\sigma^2}, \quad k = 2, \ldots, m. \tag{43}$$

As before, $\lambda_{k|1}$ is the non-estimable part of $e_k(t)$ given $e_1(t)$.

Recalling Theorem 7, the covariance matrix of $\tilde{\theta}$ is given by $\frac{1}{\sigma^2}\Lambda_{2:m|2}$. The total variance of $\hat{\tilde{\theta}}$ is defined as

$$\text{tvar } \tilde{\theta} := \sum_{i=2}^{m} \text{AsVar } \hat{\theta}_{i,n_2} = \text{Tr } \frac{1}{\sigma^2}\Lambda_{2:m|1}.$$

We are interested in understanding how the correlation of $e_1(t)$ with $e_2(t), \ldots, e_m(t)$, i.e., $\Lambda_{12} \ldots, \Lambda_{1m}$, should be in order to obtain the minimum value of the above total variance. This problem can be expressed as follows:

$$\begin{aligned}
\underset{\Lambda_{12} \ldots, \Lambda_{1m}}{\text{minimize}} \quad & \text{Tr } \Lambda_{2:m|1} \\
\text{s.t.} \quad & \Lambda \geq 0,
\end{aligned} \tag{44}$$

where the constraint on Λ implies that not all choices of the entries Λ_{1i} are allowed. Directly characterizing the

noise structure using this formulation of the problem appears to be hard. Therefore, it turns out convenient to introduce an upper triangular Cholesky factorization of Λ, namely define B *upper triangular* such that $\Lambda = BB^T$. Note that

(1) the rows of B, b_i^T, $i = 1, \ldots, m$, are such that $\|b_i\|^2 = \lambda_i$;
(2) $\mathbf{E}\{e_1 e_i\} = \Lambda_{1i} = b_1^T b_i$;
(3) there always exists an orthonormal matrix Q such that $B = \Lambda_{CH}Q$, where Λ_{CH} is the previously defined lower triangular Cholesky factor.

Lemma 20 *Let*

$$B = \begin{bmatrix} \eta & p^T \\ 0 & M \end{bmatrix}, \, M \in \mathbb{R}^{m-1 \times m-1}, \, p \in \mathbb{R}^{m-1 \times 1}, \, n \in \mathbb{R};$$

then

$$\Lambda_{2:m|1} = M(I - \frac{1}{\lambda_1})pp^T M^T. \tag{45}$$

PROOF. See Appendix D. ∎

Using the previous lemma we reformulate (44); keeping M fixed and letting p vary, we have

$$\begin{aligned}
\underset{b_1}{\text{maximize}} \quad & \text{Tr } \frac{1}{\lambda_1}Mpp^T M^T \\
\text{s.t.} \quad & \|b_1\|^2 = \lambda_1,
\end{aligned} \tag{46}$$

with $b_1^T = [\eta \quad p^T]$. Note that the constraint $\Lambda \geq 0$ is automatically satisfied. Let us define $v_1, \ldots, v_{m-1}$ as the right-singular vectors of M, namely the columns of the matrix V in the singular value decomposition (SVD) $M = USV^T$. The following result provides the structure of B that solves (44).

Theorem 21 *Let the input be white noise. Let $n_1 + 1 = n_2 = \ldots = n_m$. Then (44) is solved by an upper triangular Cholesky factor B such that its first row is*

$$b_1^{*T} = \lambda_1[0 \quad v_1^T]. \tag{47}$$

PROOF. Observe the following facts:

(1) $\text{Tr } \frac{1}{\lambda_1}Mpp^T M^T = \frac{1}{\lambda_1}p^T M^T Mp = \frac{1}{\lambda_1}\|Mp\|^2$;
(2) since $b_1^T = [\eta \quad p^T]$, it is clear that a candidate solution to (46) is of the form $b_1^{*T} = [0 \quad p^{*T}]$.

It follows that Problem (46) can be written as

$$\begin{aligned}
\underset{b_1}{\text{maximize}} \quad & \|Mp\|^2 \\
\text{s.t.} \quad & \|p\|^2 = \lambda_1,
\end{aligned} \tag{48}$$

13

whose solution is known to be (a rescaling of) the first right singular vector of M, namely v_1. Hence $b_1^{*T} = \lambda_1 \begin{bmatrix} 0 & v_1^T \end{bmatrix}$. ∎

Remark 22 *As has been pointed out in Section 2, Λ is required to be positive definite. Thus, the ideal solution provided by Theorem 21 is not applicable in practice, where one should expect that η, the first entry of b_1, is always nonzero. In this case, the result of Theorem 21 can be easily adapted, leading to $b_1^{*T} = \begin{bmatrix} \eta & \sqrt{1 - \frac{\eta^2}{\lambda_1}} v_1^T \end{bmatrix}$.*

7 Numerical examples

In this section, we illustrate the results derived in the previous sections through three sets of numerical Monte Carlo simulations.

7.1 Effect of model order

Corollary 11 is illustrated in Figure 7, where the following systems are identified using $N = 500$ input-output measurements:

$$G_i = \tilde{\Gamma}_i \theta_i, \quad i = 1, 2, 3 \qquad \tilde{\Gamma}_i(q) = F(q)^{-1} \Gamma_i(q),$$
$$\Gamma_i(q) = \begin{bmatrix} \mathcal{B}_1(q), \dots, \mathcal{B}_{n_i}(q) \end{bmatrix}, \qquad \mathcal{B}_k(q) = q^{-k}, \qquad (49)$$

with

$$F(q) = \frac{1}{1 - 0.8q^{-1}}, \qquad \theta_1^0 = \begin{bmatrix} 1 & 0.5 & 0.7 \end{bmatrix}^T,$$
$$\theta_2^0 = \begin{bmatrix} 1 & -1 & 2 \end{bmatrix}^T, \qquad \theta_3^0 = \begin{bmatrix} 1 & 1 & 2 & 1 & 1 \end{bmatrix}^T.$$

The input $u(t)$ is drawn from a Gaussian distribution with variance $\sigma^2 = 1$, filtered by $F(q)$. The measurement noise is normally distributed with covariance matrix $\Lambda = \Lambda_{CH} \Lambda_{CH}^T$, where

$$\Lambda_{CH} = \begin{bmatrix} 1 & 0 & 0 \\ 0.6 & 0.8 & 0 \\ 0.7 & 0.7 & 0.1 \end{bmatrix},$$

thus $\lambda_1 = \lambda_2 = \lambda_3 = 1$. The sample variance is computed using

$$\text{Cov } \hat{\theta}_s = \frac{1}{MC} \sum_{k=1}^{MC} |G_3(e^{j\omega_0}, \theta_3^0) - G_3(e^{j\omega_0}, \hat{\theta}_3)|^2,$$

where $MC = 2000$ is the number of realizations of the input and noise. The same realizations of the input and noise are used for all model orders.

The variance of $G_3(e^{j\omega}, \hat{\theta}_3)$ increases with increasing n_i, $i = 1, 2$, but only up to the point where $n_i = n_3 = 5$.
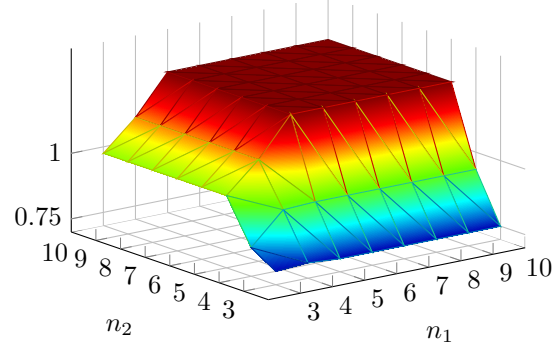

Fig. 7. Sample variance of $G_3(e^{j\omega}, \hat{\theta}_3)$ as a function of the number of estimated parameters of G_1 and G_2.

After that, any increase in n_1 or n_2 does not increase the variance of $G_3(e^{j\omega}, \hat{\theta}_3)$, as can be seen in Figure 7. The behavior can be explained by Corollary 11: when $n_3 \geq n_1, n_2$, G_3 is the last block, having the highest number of parameters, and any increase in n_1, n_2 increases the variance of G_3. When for example $n_1 \geq n_3$, the blocks should be reordered so that G_3 comes before G_1. In this case, when n_1 increases the first condition of Corollary 11 does not hold and hence the variance of $G_3(e^{j\omega}, \hat{\theta}_3)$ does not increase further.

7.2 Effect of removing one parameter

In the second set of numerical experiments, we simulate the same type of system as in Section 2.2. We let β vary in the interval [0.001, 1]. For each β, we generate $MC = 2000$ Monte Carlo experiments, where in each of them we collect $N = 500$ input-output samples. At the i-th Monte Carlo run, we generate new trajectories for the input and the noise and we compute $\hat{\theta}_i$ as in (9). The sample covariance matrix, for each β, is computed as

$$\text{Cov } \hat{\theta}_s = \frac{1}{MC} \sum_{i=1}^{MC} (\hat{\theta}_i - \theta)(\hat{\theta}_i - \theta)^T.$$

Figure 8 shows that the variance of $\hat{\theta}_{21}$ is always close to one, no matter what the value of β is. It also shows that the variance of the estimate $\hat{\theta}_{22}$ behaves as β^2. In particular, when β approaches 0 (i.e., almost perfectly correlated noise processes), the variance of such estimate tends to 0. All of the observations are as predicted by Corollary 7 and the example in Section 2.2.

7.3 Optimal correlation structure

A third simulation experiment is performed in order to illustrate Theorem 21. We consider a system with $m = 3$ outputs; the modules are

$$G_1(q) = 0.3q^{-1}, \qquad G_2(q) = 0.8q^{-1} - 0.4q^{-2},$$
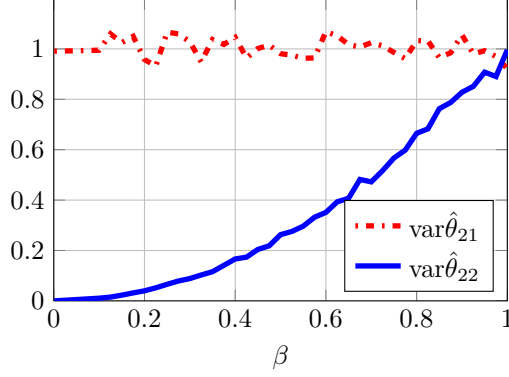$$G_3(q) = 0.1q^{-1} + 0.2q^{-2},$$

Fig. 8. Sample variance of the parameters of the first module (as functions of β) when the second module has one parameter.

so that

$$\theta = [\theta_{11}\,\theta_{21}\,\theta_{31}\,\theta_{22}\,\theta_{32}]^T = [0.3\ 0.8\ 0.1\ -0.4\ 0.2]^T .$$

The noise process is generated by the following upper triangular Cholesky factor:

$$R = \begin{bmatrix} \varepsilon & \sqrt{1-\varepsilon^2}\cos\alpha & \sqrt{1-\varepsilon^2}\sin\alpha \\ 0 & 0.8 & 0.6 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} \varepsilon & p^T \\ 0 & M \end{bmatrix},$$

where $\varepsilon = 0.1$ and $\alpha \in [0, \pi]$ is a parameter tuning the correlation of $e_1(t)$ with $e_2(t)$ and $e_3(t)$. The purpose of this experiment is to show that, when α is such that $p = [\sqrt{1-\varepsilon^2}\cos\alpha \quad \sqrt{1-\varepsilon^2}\sin\alpha]^T$ is aligned with the first right-singular vector v_1 of M, then the total variance of the estimate of the sub-vector $\tilde{\theta} = [\theta_{12}\,\theta_{22}]^T$ is minimized. In the case under analysis, $v_1 = [0.447\ 0.894]^T = [\cos\alpha_0\ \sin\alpha_0]^T$, with $\alpha_0 = 1.11$. We let α take values in $[0, \pi]$ and for each α we generate $MC = 2000$ Monte Carlo runs of $N = 500$ input-output samples each. We compute the sample total variance of $\tilde{\theta}$ as

$$\text{tvar } \hat{\tilde{\theta}}_{\text{s}} = \frac{1}{MC} \sum_{i=1}^{MC} \left((\hat{\theta}_{22,i} - \theta_{22})^2 + (\hat{\theta}_{32,i} - \theta_{32})^2 \right),$$

where $\hat{\theta}_{22,i}$ and $\hat{\theta}_{32,i}$ are the estimates obtained at the i-th Monte Carlo run.

The results of the experiment are reported in Figure 9. It can be seen that the minimum total variance of the estimate of $\hat{\tilde{\theta}}$ is attained for values close to α_0 (approximations are due to low resolution of the grid of values of α). An interesting observation regards the value of α for which the total variance is maximized: this happens when $\alpha = 2.68$, which yields the second right-singular vector of M, namely $v_2 = [-0.894\ 0.447]^T$.
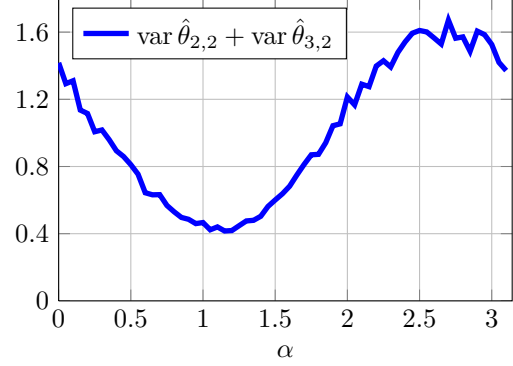

Fig. 9. Total sample variance of the parameter vector $\tilde{\theta}$ as function of α.

8   Conclusions

The purpose of this paper has been to examine how the estimation accuracy of a linear SIMO model depends on the correlation structure of the noise and model structure and model order. A formula for the asymptotic covariance of the frequency response function estimate and the model parameters has been developed for the case of temporally white, but possibly spatially correlated additive noise. It has been shown that the variance decreases when parts of the noise can be linearly estimated from measurements of other blocks with less estimated parameters. The expressions reveal how the order of the different blocks and the correlation of the noise affect the variance of one block. In particular, it has been shown that the variance of the block of interest levels off when the number of estimated parameters in another block reaches the number of estimated parameters of the block of interest. The optimal correlation structure for the noise was determined for the case when one block has one parameter less than the other blocks.

References

Ali, M., A. Ali, S.S. Chughtai and H. Werner (2011). Consistent identification of spatially interconnected systems. In 'Proceedings of the 2011 American Control Conference'. pp. 3583–3588.

Bartlett, M. (1951). 'An inverse matrix adjustment arising in discriminant analysis'. *The Annals of Mathematical Statistics* **22**(1), pp. 107–111.

Benesty, J., S. Makino and J. Chen (2005). *Speech enhancement*. Springer.

Bertaux, N., P. Larzabal, C. Adnet and E. Chaumette (1999). A parameterized maximum likelihood method for multipaths channels estimation. In 'Proceedings of the 2nd IEEE Workshop on Signal Processing Advances in Wireless Communications'. pp. 391–394.

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

Dankers, A., P.M.J. Van den Hof and P.S.C. Heuberger (2013a). Predictor input selection for direct identification in dynamic networks. In 'Proceeding of the 52nd IEEE Conference on Decision and Control'. pp. 4541–4546.

Dankers, A., P.M.J. Van den Hof, X. Bombois and P.S.C. Heuberger (2013b). Predictor input selection for two stage identification in dynamic networks. In 'Proceeding of the 2013 European Control Conference'. pp. 1422–1427.

Dankers, A., P.M.J. Van den Hof, X. Bombois and P.S.C. Heuberger (2014). Errors-in-variables identification in dynamic networks. In 'Proceedings of the 19th IFAC World Congress'.

Doclo, S. and M. Moonen (2002). 'GSVD-based optimal filtering for single and multimicrophone speech enhancement'. *IEEE Transactions on Signal Processing* **50**(9), 2230–2244.

Everitt, N., C.R. Rojas and H. Hjalmarsson (2013). A geometric approach to variance analysis of cascaded systems. In 'Proceedings of the 52nd IEEE Conference on Decision and Control'.

Gevers, M., L. Mišković, D. Bonvin and A. Karimi (2006). 'Identification of multi-input systems: variance analysis and input design issues'. *Automatica* **42**(4), 559 – 572.

Gunes, B., A. Dankers and P.M.J. Van den Hof (2014). A variance reduction technique for identification in dynamic networks. In 'Proceedings of the 19th IFAC World Congress'.

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

Hägg, P., B. Wahlberg and H. Sandberg (2011). On identification of parallel cascade serial systems. In 'Proceedings of the 18th IFAC World Congress'.

Hjalmarsson, H. and B. Ninness (2006). 'Least-squares estimation of a class of frequency functions: A finite sample variance expression'. *Automatica* **42**(4), 589 – 600.

Hjalmarsson, H. and J. Mårtensson (2011). 'A geometric approach to variance analysis in system identification'. *IEEE Transactions on Automatic Control* **56**(5), 983 –997.

Horn, R. and C.R. Johnson (1990). *Matrix Analysis.* Cambridge University Press.

Koski, T. and J. Noble (2012). 'A review of Bayesian networks and structure learning'. *Mathematica Applicanda* **40**(1), 51–103.

Ljung, L. (1985). 'Asymptotic variance expressions for identified black-box transfer function models'. *IEEE Transactions on Automatic Control* **30**(9), 834 – 844.

Ljung, L. (1999). *System Identification: Theory for the User.* 2 edn. Prentice Hall.

Ljung, L. and P.E. Caines (1980). 'Asymptotic normality of prediction error estimators for approximate system models'. *Stochastics* **3**(1-4), 29–46.

Ljung, L., H. Hjalmarsson and H. Ohlsson (2011).

'Four encounters with system identification'. *European Journal of Control* **17**(5), 449–471.

Mårtensson, J. (2007). Geometric analysis of stochastic model errors in system identification. Ph.D. thesis. KTH, Automatic Control.

Materassi, D., M.V. Salapaka and L. Giarre (2011). Relations between structure and estimators in networks of dynamical systems. In 'Proceedings of the 50th IEEE Conference on Decision and Control and European Control Conference'. pp. 162–167.

McCombie, D., A.T. Reisner and H.H. Asada (2005). 'Laguerre-model blind system identification: Cardiovascular dynamics estimated from multiple peripheral circulatory signals'. *IEEE Transactions on Biomedical Engineering* **52**(11), 1889–1901.

Ninness, B. and F. Gustafsson (1997). 'A unifying construction of orthonormal bases for system identification'. *IEEE Transactions on Automatic Control* **42**, 515–521.

Ninness, B. and H. Hjalmarsson (2004). 'Variance error quantifications that are exact for finite model order'. *IEEE Transactions on Automatic Control* **49**(8), 1275–1291.

Ninness, B. and H. Hjalmarsson (2005a). 'Analysis of the variability of joint input–output estimation methods'. *Automatica* **41**(7), 1123 – 1132.

Ninness, B. and H. Hjalmarsson (2005b). 'On the frequency domain accuracy of closed-loop estimates'. *Automatica* **41**, 1109–1122.

Ninness, B., H. Hjalmarsson and F. Gustafsson (1999). 'The fundamental role of general orthonormal bases in system identification'. *IEEE Transactions on Automatic Control* **44**(7), 1384–1406.

Ramazi, P., H. Hjalmarsson and J. Mårtensson (2014). 'Variance analysis of identified linear MISO models having spatially correlated inputs, with application to parallel Hammerstein models'. *Automatica* **50**(6), 1675 – 1683.

Schmidt, R. (1986). 'Multiple emitter location and signal parameter estimation'. *IEEE Transactions on Antennas and Propagation* **34**(3), 276–280.

Söderström, T. and P. Stoica (1989). *System identification.* Prentice-Hall. Englewood Cliffs.

Stoica, P., M. Viberg and B. Ottersten (1994). 'Instrumental variable approach to array processing in spatially correlated noise fields'. *IEEE Transactions on Signal Processing* **42**(1), 121–133.

Torres, P., J.W. van Wingerden and M. Verhaegen (2014). 'Output-error identification of large scale 1d-spatially varying interconnected systems'. *IEEE Transactions on Automatic Control* **60**(1), 130–142.

Trefethen, L. and D. Bau (1997). *Numerical Linear Algebra.* Society for Industrial and Applied Mathematics.

Trudnowski, D., J.M. Johnson and J.F. Hauer (1998). SIMO system identification from measured ringdowns. In 'Proceedings of the 1998 American Control Conference'. Vol. 5. pp. 2968–2972.

Ulusoy, H., M.Q. Feng and P.J. Fanning (2011). 'System identification of a building from multiple seismic

records'. *Earthquake Engineering & Structural Dynamics* **40**(6), 661–674.

Van den Hof, P. M. J., Arne Dankers, P. S. C. Heuberger and X Bombois (2013). 'Identification of dynamic models in complex networks with prediction error methods - basic methods for consistent module estimates'. *Automatica* **49**(10), 2994–3006.

Viberg, M. and B. Ottersten (1991). 'Sensor array processing based on subspace fitting'. *IEEE Transactions on Signal Processing* **39**(5), 1110–1121.

Viberg, M., P. Stoica and B. Ottersten (1997). 'Maximum likelihood array processing in spatially correlated noise fields using parameterized signals'. *IEEE Transactions on Signal Processing* **45**(4), 996–1004.

Wahlberg, B., H. Hjalmarsson and J. Mårtensson (2009). 'Variance results for identification of cascade systems'. *Automatica* **45**(6), 1443–1448.

Yuan, Z. D. and L. Ljung (1984). 'Black-box identification of multivariable transfer functions–asymptotic properties and optimal input design'. *International Journal of Control* **40**(2), 233–256.

Zhu, Y. (1989). 'Black-box identification of MIMO transfer functions: Asymptotic properties of prediction error models'. *International Journal of Adaptive Control and Signal Processing* **3**(4), 357–373.

Zhu, Y. (2001). *Multivariable system identification for process control.* Elsevier.

A   Proof of Lemma 5

Let $v(t) = \Lambda_{CH}^{-1} e(t)$ for some real valued random variable $e(t)$ (Λ_{CH}^{-1} exists and is unique for $\Lambda > 0$ (Horn and Johnson, 1990)). Then Cov $v(t) = I$. Similarly $e(t) = \Lambda_{CH} v(t)$. The set $\{v_1(t), \ldots, v_j(t)\}$ is a function of $e_1(t), \ldots, e_j(t)$ only and vice versa, for all $1 \leq j \leq m$. Thus, if $e_1(t), \ldots, e_j(t)$ are known, then also $\{v_1(t), \ldots, v_j(t)\}$ are known, but nothing is known about $\{v_{j+1}(t), \ldots, v_m(t)\}$. Thus, for $j < i$ the best linear estimator of $e_i(t)$ given $e_1(t), \ldots, e_j(t)$, is

$$\hat{e}_{i|j}(t) = \sum_{k=1}^{j} \gamma_{ik} v_k(t), \qquad (A.1)$$

and

$$e_i(t) - \hat{e}_{i|j}(t)$$

has variance

$$\lambda_{i|j} = \sum_{k=j+1}^{i} \gamma_{ik}^2.$$

For the last part of the lemma, we realize that the dependence of $\hat{e}_{i|j}(t)$ on $e_j(t)$ in Equation (A.1) is given by γ_{ij}/γ_{jj} (since $v_1(t), \ldots, v_{j-1}(t)$ do not depend on $e_j(t)$). Hence $\hat{e}_{i|j}(t)$ depends on $e_j(t)$ if and only if $\gamma_{ij} \neq 0$.

B   Proof of Theorem 10

Before giving the proof of Theorem 10 we need the following auxiliary lemma.

Lemma 23 *Let $\Lambda > 0$ and real and its Cholesky factor Λ_{CH} be partitioned according to $e_{1:\chi_k-1}$ and $e_{\chi_k:m}$,*

$$\Lambda = \begin{bmatrix} \Lambda_1 & \Lambda_{12} \\ \Lambda_{21} & \Lambda_2 \end{bmatrix}, \qquad \Lambda_{CH} = \begin{bmatrix} (\Lambda_{CH})_1 & 0 \\ (\Lambda_{CH})_{21} & (\Lambda_{CH})_2 \end{bmatrix}.$$

Then

$$\Lambda_{\chi_k:m|\chi_k-1} = (\Lambda_{CH})_2 (\Lambda_{CH})_2^T.$$

PROOF. By the derivations of Lemma A, for some $v(t)$ with Cov $v(t) = I$, $e(t) = \Lambda_{CH} v(t)$ and $v_{1:\chi-1}(t)$ are known since $e_{1:\chi-1}(t)$ are known. Furthermore $\hat{e}_{\chi_k:m|\chi_k-1}(t) = (\Lambda_{CH})_{21} v_{1:\chi-1}(t)$, which implies $e_{\chi_k:m|\chi_k-1}(t) - \hat{e}_{\chi_k:m|\chi_k-1}(t) = (\Lambda_{CH})_2 v_{\chi_k:m}(t)$ and the results follows since Cov $v_{\chi_k:m}(t) = I$. ∎

The asymptotic variance is given by (11) with

$$\Psi(q) = \tilde{\Psi}(q) \Lambda_{CH}^{-T}.$$

Let $n = n_1 + \cdots + n_m$. From the upper triangular structure of Λ_{CH}^{-T} and $n_1 \leq n_2 \leq \ldots \leq n_m$, an orthonormal basis for $\mathcal{S}_\Psi$, the subspace spanned by the rows of Ψ, is given by

$$\begin{aligned} \mathcal{B}_k^{\mathcal{S}}(e^{j\omega}) &:= \begin{bmatrix} \mathcal{B}_k & 0 & \ldots & 0 \end{bmatrix}, & k&=1, \ldots, n_1 \\ \mathcal{B}_k^{\mathcal{S}}(e^{j\omega}) &:= \begin{bmatrix} 0 & \mathcal{B}_{k-n_1} & 0 & \ldots \end{bmatrix}, & k&=n_1+1, \ldots, n_2 \\ &\vdots & & \qquad\qquad (B.1) \\ \mathcal{B}_k^{\mathcal{S}}(e^{j\omega}) &:= \begin{bmatrix} \ldots & 0 & \mathcal{B}_{k-n+n_m} \end{bmatrix}, & k&=n-n_m+1, \ldots, n. \end{aligned}$$

First note that

$$\frac{\partial G}{\partial \theta} = \Psi \Lambda_{CH}^T.$$

Then, using Theorem 3,

$$\sigma^2 \text{ AsVar } \hat{G} = \Lambda_{CH} \sum_{k=1}^{n} \mathcal{B}_k^{\mathcal{S}}(e^{j\omega_0})^* \mathcal{B}_k^{\mathcal{S}}(e^{j\omega_0}) \Lambda_{CH}^T.$$

Sorting the sum with respect to the basis functions $\mathcal{B}_k(e^{j\omega_0})$, we get

$$\sigma^2 \text{ AsCov } \hat{G}_i = \Lambda_{CH} \sum_{k=1}^{n_m} |\mathcal{B}_k(e^{j\omega_0})|^2 \begin{bmatrix} 0_{\chi_k-1} & 0 \\ 0 & I \end{bmatrix} \Lambda_{CH}^T.$$

Using Lemma 23

$$\text{AsCov } \hat{G} = \frac{1}{\sigma^2} \sum_{k=1}^{n_m} \begin{bmatrix} 0_{\chi_k - 1} & 0 \\ 0 & \Lambda_{\chi_k : m | \chi_k - 1} \end{bmatrix} |\mathcal{B}_k(e^{j\omega_0})|^2.$$

Thus (25) follows. We now show the first part of the theorem, that

$$\text{AsCov } \hat{\bar{\theta}} = \frac{1}{\sigma^2} \text{diag}(\Lambda_{1:m}, \Lambda_{\chi_2 : m | \chi_2 - 1}, \dots, \Lambda_{\chi_{n_m} : m | \chi_{n_m} - 1}).$$

The covariance of $\hat{G}$ can be expressed as

$$\text{AsCov } \hat{G} = T \text{AsCov } \hat{\bar{\theta}} \, T^* \qquad \text{(B.2)}$$

where

$$T = \begin{bmatrix} \mathcal{B}_1 I(1) & \mathcal{B}_2 I(2) \dots \mathcal{B}_{n_m} I(n_m) \end{bmatrix},$$

$$I(k) = \begin{bmatrix} \mathbf{0} \\ I_{m - \chi_k + 1} \end{bmatrix} \in \mathbb{R}^{m \times (m - \chi_k + 1)}.$$

However, (B.2) equals (25) for all ω and the theorem follows.

C   Proof of Corollary 11

To prove Corollary 11 we will use (28). First, we make the assumption that j is the last module that has n_j parameters. This assumption is made for convenience since reordering all modules with n_j estimated parameters does not change (28). First of all, we see that if

$$n_j \geq n_i,$$

then (28) does not increase when n_j increases. If instead

$$n_j < n_i,$$

the increase in variance is given by

$$\gamma_{ij}^2 |\mathcal{B}_{n_j+1}(e^{j\omega_0})|^2,$$

which is non-zero iff $\gamma_{ij} \neq 0$ and $|\mathcal{B}_{n_j+1}(e^{j\omega_0})|^2 \neq 0$. From Lemma 5 the theorem follows.

D   Proof of Lemma 20

The inverse of B is

$$B^{-1} = \begin{bmatrix} \eta^{-1} & -q^T \\ 0 & M^{-1} \end{bmatrix}, \qquad q^T := \eta^{-1} p^T M^{-1},$$

so that

$$\Lambda^{-1} = B^{-T} B^{-1} = \begin{bmatrix} \eta^{-2} & -q^T \eta^{-1} \\ -q\eta^{-1} & M^{-T} M^{-1} + qq^T \end{bmatrix}.$$

Hence, using the Sherman–Morrison formula (Bartlett, 1951)

$$\begin{aligned} \Lambda_{2:m|1} &= \left(M^{-T} M^{-1} + qq^T \right)^{-1} \\ &= MM^T - \frac{1}{k} MM^T qq^T MM^T \\ &= MM^T - \frac{1}{k} \frac{M p p^T M^T}{\eta^2}, \end{aligned}$$

where

$$\begin{aligned} k &= 1 + q^T MM^T q = 1 + \frac{p^T p}{\eta^2} \\ &= \frac{\eta^2 + p^T p}{\eta^2} = \frac{\|b_1\|_2^2}{\eta^2} \\ &= \frac{\lambda_1}{\eta^2}, \end{aligned}$$

so (45) follows.