Linear and Nonlinear Regression with Application to Unbalance Estimation

Peter Nauclér* Torsten Söderström*

Abstract

This paper considers estimation of parameters that enters nonlinearly in a regression model. The problem formulation is closely connected to unbalance estimation of rotating machinery. The parameter estimation problem can after approximation be formulated as a linear estimation procedure, while neglecting the effects of the disturbing term. Two such estimators are derived. In addition, a third approach that handles the uncertainty in a statistically sound way is presented. The three methods are compared and analyzed with respect to their statistical accuracy. Using the example of unbalance estimation of a separator, the nonlinear approach is shown to outperform the other two.

1 Introduction

Consider a parameter estimation problem given by

$$\boldsymbol{y}_k = \left(\boldsymbol{A} + \tilde{\boldsymbol{A}}_k\right)(\boldsymbol{x}_0 + \boldsymbol{x}_k), \quad k = 1, \cdots, M, \tag{1}$$

where

$$\begin{aligned} \boldsymbol{y}_k &\in \mathbb{C}^{n\times 1} && \text{measured variable,} \\ \boldsymbol{A} &\in \mathbb{C}^{n\times n} && \text{unknown variable,} \\ \tilde{\boldsymbol{A}}_k &\in \mathbb{C}^{n\times n} && \text{stochastic disturbance,} \\ \boldsymbol{x}_0 &\in \mathbb{C}^{n\times 1} && \text{unknown variable,} \\ \boldsymbol{x}_k &\in \mathbb{C}^{n\times 1} && \text{user chosen variable,} \end{aligned}$$

and where $\mathbb{C}^{n\times m}$ denotes complex valued $n \times m$ matrix. The basis for the system equation (1) will be motivated in the next section. The quantity M is

*Department of Information Technology, Uppsala University. P.O. Box 337, SE-751 05 Uppsala, Sweden. Email: {Peter.Naucler, Torsten.Soderstrom}@it.uu.se

the number of experiments. The number of unknowns is n^2+n and the number of equations is nM. Therefore, the number of experiments must fulfill

$$M \geq n+1. \tag{2}$$

We are primarily interested in estimating $\boldsymbol{x}_0$ and, therefore, $\boldsymbol{A}$ is treated as a nuisance variable. For each experiment, the sought variable $\boldsymbol{x}_0$ is invariant. The way that $\tilde{\boldsymbol{A}}_k$ enters the system makes the estimation problem nontrivial to handle.

2 Background and Motivation

The system equation (1) has its motivation from parameter estimation problems for systems operating in stationarity. Such problems appear in balancing of high speed machinery, where the purpose is to estimate mass unbalances $\boldsymbol{x}_0$ in complex dynamic systems rotating at high speed. Applications that can be mentioned are machining tools, aircraft turbine engines [15], steam turbines, electric generators [2] and balancing of separators.

The matrix $\boldsymbol{A} = \boldsymbol{A}(\mathrm{i}\omega_0)$ is the frequency response function from the current unbalance state $(\boldsymbol{x}_0 + \boldsymbol{x}_k)$ to the measured harmonic vibrational response $\boldsymbol{y}_k$ at the angular frequency ω_0. The user chosen variable $\boldsymbol{x}_k$ is used to excite the system so that the problem becomes solvable. In the present paper, $\boldsymbol{A}$ is assumed to be quadratic. Thus, the number of inputs is equal to the number of outputs. Generally, however, $\boldsymbol{A}$ can be rectangular which is the case if *e.g.* there are more sensors than inputs. In such circumstances, one should employ the pseudo inverse of $\boldsymbol{A}$ instead of $\boldsymbol{A}^{-1}$, whenever it appears.

The frequency response $\boldsymbol{A}$ is in this context often referred to as the *influence coefficient matrix*. The reason for performing experiments with a system that operates in stationarity is that the relation between measured output and applied input becomes simple. Irrespectively of the order of the system (which can be extremely large), the entries of $\boldsymbol{A}$ becomes scalar complex numbers when the frequency response is 'evaluated' at the single frequency $\omega = \omega_0$. The influence coefficient matrix is a function of the structural properties of the underlying system. These properties are here subject to change between experiments, which leads to the model with disturbing term $\tilde{\boldsymbol{A}}_k$.

2.1 Balancing of Separators

This paper is inspired by the problem of separator balancing. It is an important topic in the field of separator technology. The separator bowl is rotating with high speed, which typically is about 5000 revolutions per minute. The appeared centrifugal force is used to separate different substances, *e.g.* liquids. The use of large centrifugal forces is the core of separation technology. Since the bowl is very heavy and rotates with such a high speed, small mass unbalances

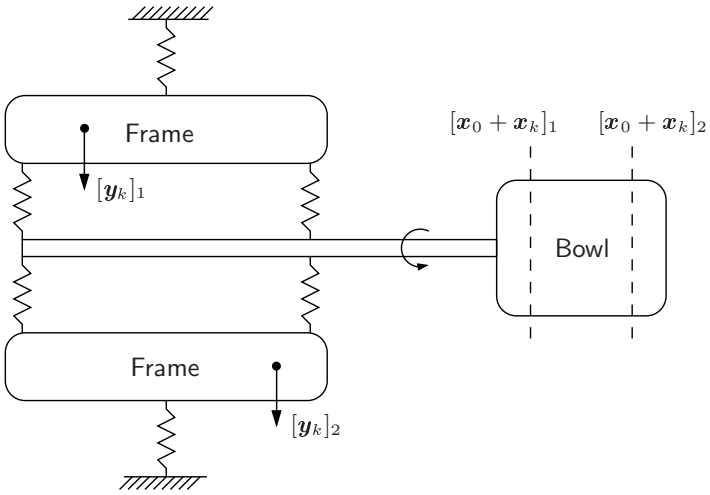

Figure 1: A separator model. The stiffnesses are modeled as complex numbers, which is a way to introduce damping in the system. These stiffnesses are subject to change between experiments.

create large radial bearing forces that may even be harmful. The magnitude of these forces can often be of the magnitude tens of kN, *i.e.* several tonnes. Therefore, the separator must be balanced after manufacturing. When the unbalance estimation is completed, mass corrections are applied to the bowl to counteract the unbalances.

Consider a separator model as shown in Figure 1. The unbalance estimation procedure is to apply trial masses to the rotating bowl in two planes and measure the generated vibration at two sensor positions, as indicated in the figure. Thus, for this system $n = 2$. The measured vibrational frame response is used as an indication of large radial bearing forces. In each new experiment, masses $[\boldsymbol{x}_k]_1$ and $[\boldsymbol{x}_k]_2$ (elements 1 and 2 of $\boldsymbol{x}_k$) are applied and the bowl is driven up to its speed of operation. The influence coefficient matrix $\boldsymbol{A}$ depends on the mechanical properties of the separator, *e.g.* masses, stiffnesses and damping elements. The main source of uncertainty is that the dynamical properties of the system change between experiments. There are several reasons for this. First of all the separator bowl often needs to be opened in order to apply the trial masses. When doing this, some of the structural properties will change due to plays of bearings *etc.*. Also a main source of uncertainty is that different stiffness and damping elements seem to change somewhat between experiments. For example, there are rubber damping elements whose properties depend on temperature and the vibrational amplitude. Such observations are the basis for the model with additive uncertainty to the influence coefficient matrix $\boldsymbol{A}$ that leads to a system equation of the type (1). In the separator model showed in Figure 1, this is modeled with non-invariant stiffnesses. The concept of structural, or hysteretic, damping is employed, which means that the stiffnesses shown in Figure 1 are modeled as complex numbers. This is a way to introduce damping in the system.

In the separator system, equations of the type (1) can be set up for several angular frequencies. The measured quantity $\boldsymbol{y}_k$ and the matrix $\boldsymbol{A}$ then become functions of frequency, while $\boldsymbol{x}_0$ and $\boldsymbol{x}_k$ are frequency independent. Still, the number of experiments must fulfill (2). Thus, the core of the problem is to be able to perform unbalance estimation at a single frequency, which is considered in this paper.

2.2 Existing Estimation Techniques

Equations of the type (1) are commonly found in the literature on balancing of rotating machinery [6, 11, 2, 3]. Commonly, however, the disturbing term $\tilde{\boldsymbol{A}}_k$ is not considered. Even though the equation (1) is nonlinear in the unknowns $\boldsymbol{A}$ and $\boldsymbol{x}_0$, it can be transformed into a linear estimation problem. This is the basis for an approach which is often called the influence coefficient method [1, 2, 15, 8]. It is an experimental method that can be implemented in different ways, but the basis is to use $\boldsymbol{x}_k = \boldsymbol{0}$ in the first experiment. If the disturbance term is negligible, the first measurement become

$$\boldsymbol{y}_1 = \boldsymbol{A}\boldsymbol{x}_0,$$

which can be employed to subtract the effects of $\boldsymbol{x}_0$ in the remaining experiments where $\boldsymbol{x}_k \neq \boldsymbol{0}$. Then the matrix $\boldsymbol{A}$ can be estimated and when it is considered to be known, it is straightforward to compute an estimate of $\boldsymbol{x}_0$. The equations needed to carry out such a procedure can be arranged in different ways, but the basics are as described above.

Commonly, a deterministic viewpoint is adopted and no special treatment is dedicated to the effects of disturbances. In [9] an optimal weighting is introduced under the assumption that the uncertainty is additive on the output side of (1), *i.e.* sensor noise. In addition, a statistical analysis is carried out. However, in the current context it is assumed that sensor noise is negligible compared to the system disturbance $\tilde{\boldsymbol{A}}_k$. To the best of our knowledge there are no statistical analysis associated with estimation of (1) and no algorithms proposed that are devoted to a sound statistical treatment of the disturbing variable.

In this paper three different approaches for unbalance estimation are derived and analyzed with respect to their statistical accuracies in the presence of $\tilde{\boldsymbol{A}}_k$. The first two relies on a deterministic viewpoint in the derivation, which leads to linear estimation procedures that have closed form solutions. The third approach instead makes use of the statistical properties of $\tilde{\boldsymbol{A}}_k$, which leads to a nonlinear optimization problem.

Although unbalance estimation is a motivation for the system equation (1), we do not claim to deliver a complete solution to the problem of balancing. Instead the focus in this paper is on the parameter estimation problem (1) as such, and the statistical analysis associated with the derived algorithms.

3 Preliminaries

The purpose of this section is to introduce some notation and mathematical tools that will be utilized in the sequel of this paper.

The vec operator is the operator that stacks the columns of a matrix, *i.e.* if $\boldsymbol{A} = [\boldsymbol{a}_1 \ \cdots \ \boldsymbol{a}_n]$ where $\boldsymbol{a}_k$ is column k. Using this property, we define

$$\boldsymbol{\alpha} \triangleq \operatorname{vec}(\boldsymbol{A}) = \begin{bmatrix} \boldsymbol{a}_1 \\ \vdots \\ \boldsymbol{a}_n \end{bmatrix}.$$

Similarly, the vectorized uncertainty matrix is defined as

$$\tilde{\boldsymbol{\alpha}}_k = \operatorname{vec}(\tilde{\boldsymbol{A}}_k)$$

and furthermore the uncertainty vector from all M experiments becomes

$$\tilde{\boldsymbol{\alpha}} = \begin{bmatrix} \tilde{\boldsymbol{\alpha}}_1 \\ \vdots \\ \tilde{\boldsymbol{\alpha}}_M \end{bmatrix}.$$

Before proceeding, an assumption regarding the uncertainty matrix is needed:

Assumption 1. The uncertainty matrix is zero mean and $\tilde{\boldsymbol{A}}_k$ is uncorrelated with $\tilde{\boldsymbol{A}}_l$ for $k \neq l$. It has an associated covariance matrix

$$\mathbf{E}\tilde{\boldsymbol{\alpha}}_k \tilde{\boldsymbol{\alpha}}_l^T = \boldsymbol{R}_{\boldsymbol{\alpha}} \delta_{k,l},$$

where $\mathbf{E}$ denotes the expectation operator and $\delta_{k,l}$ is the Kronecker delta function. □

The fact that the uncertainty is independent between experiments implies that

$$\boldsymbol{R}_{\boldsymbol{\alpha},M} \triangleq \operatorname{cov}(\tilde{\boldsymbol{\alpha}}) = \boldsymbol{I}_M \otimes \boldsymbol{R}_{\boldsymbol{\alpha}}, \tag{3}$$

where $\boldsymbol{I}_M$ is the identity matrix of dimension M and $\otimes$ is the Kronecker product.

The vec operator has many useful properties. One that will be extensively employed in this context is

$$\operatorname{vec}(\boldsymbol{ABC}) = \left(\boldsymbol{C}^T \otimes \boldsymbol{A}\right) \operatorname{vec}(\boldsymbol{B}).$$

Application of this result on the system equation (1) yields

$$\operatorname{vec}(\boldsymbol{y}_k) = \boldsymbol{y}_k = \left((\boldsymbol{x}_0 + \boldsymbol{x}_k)^T \otimes \boldsymbol{I}_n\right) \boldsymbol{\alpha} + \left((\boldsymbol{x}_0 + \boldsymbol{x}_k)^T \otimes \boldsymbol{I}_n\right) \tilde{\boldsymbol{\alpha}}_k.$$

The pseudo inverse of a full rank tall matrix $\boldsymbol{\Phi}$ is given by

$$\boldsymbol{\Phi}^\dagger = \left(\boldsymbol{\Phi}^* \boldsymbol{\Phi}\right)^{-1} \boldsymbol{\Phi}^*,$$

where * denotes Hermitian transpose. If $\boldsymbol{\Phi}$ is square, then $\boldsymbol{\Phi}^\dagger = \boldsymbol{\Phi}^{-1}$.

The series expansion

$$\left(\boldsymbol{A} + \tilde{\boldsymbol{A}}\right)^{-1} \approx \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\tilde{\boldsymbol{A}}\boldsymbol{A}^{-1} + \cdots$$

will prove useful. The second order terms can be omitted if $\tilde{\boldsymbol{A}}$ is small compared to $\boldsymbol{A}$.

Let $\boldsymbol{B}(\boldsymbol{x})$ and $\boldsymbol{C}(\boldsymbol{x})$ be matrices whose entries are functions of a real valued vector $\bar{\boldsymbol{x}}$. Furthermore, let $[\bar{\boldsymbol{x}}]_k$ be the k-th element of the vector $\bar{\boldsymbol{x}}$. Then we define

$$\boldsymbol{B}^{(k)} = \frac{\partial \boldsymbol{B}(\bar{\boldsymbol{x}})}{\partial [\bar{\boldsymbol{x}}]_k}$$

and

$$\boldsymbol{B}^{(kl)} = \frac{\partial^2 \boldsymbol{B}(\bar{\boldsymbol{x}})}{\partial [\bar{\boldsymbol{x}}]_k \partial [\bar{\boldsymbol{x}}]_l}.$$

For products of matrices the chain rule applies,

$$(\boldsymbol{BC})^{(k)} = \boldsymbol{B}^{(k)}\boldsymbol{C} + \boldsymbol{B}\boldsymbol{C}^{(k)},$$

where the $\bar{\boldsymbol{x}}$-argument is dropped for notational convenience. For differentiation of matrix inverses it holds that

$$\boldsymbol{B}^{-(k)} \triangleq \left(\boldsymbol{B}^{-1}\right)^{(k)} = -\boldsymbol{B}^{-1}\boldsymbol{B}^{(k)}\boldsymbol{B}^{-1}.$$

4 Linear Deterministic Estimation

One approach to handle the problem of estimating $\boldsymbol{x}_0$ from measured data is to adopt a deterministic viewpoint. Thus, if the effects caused by $\tilde{\boldsymbol{A}}$ is neglected, or considered to be insignificant, (1) simplifies to

$$\boldsymbol{y}_k = \boldsymbol{A}\left(\boldsymbol{x}_0 + \boldsymbol{x}_k\right), \quad k = 1, \cdots, M, \tag{4}$$

where only $\boldsymbol{A}$ and $\boldsymbol{x}$ are unknown. Equation (4) is the basis for two different approaches to estimate $\boldsymbol{x}_0$. The two identification procedures are labeled A1 and A2, respectively. Both these approaches are employed in the balancing industry [7].

4.1 Approach A1

From the relation (4), the unknown variable $\boldsymbol{x}_0$ can be found using a simple procedure. The first step is to subtract the effects of $\boldsymbol{x}_0$ from (4). This is performed by choosing $\boldsymbol{x}_1 = \boldsymbol{0}$ which yields

$$\boldsymbol{y}_1 = \boldsymbol{A}\boldsymbol{x}_0 \tag{5}$$

and for the remaining $M-1$ equations, we form

$$\boldsymbol{z}_k \triangleq \boldsymbol{y}_k - \boldsymbol{y}_1, \quad k = 2, \cdots, M,$$

which yields

$$\boldsymbol{z}_k = \boldsymbol{A}\boldsymbol{x}_k, \quad k = 2, \cdots, M \tag{6}$$

if the disturbance is neglected. Both $\boldsymbol{z}_k$ and $\boldsymbol{x}_k$ are known and therefore it is straightforward to compute an estimate of the nuisance variable $\boldsymbol{A}$. This can be performed in different ways. One option is apply the vec operator to (6), which gives

$$\boldsymbol{z}_k = \left(\boldsymbol{x}_k^T \otimes \boldsymbol{I}_n\right) \boldsymbol{\alpha} \tag{7}$$

and upon stacking the experiments in a tall vector

$$\boldsymbol{z} = \begin{bmatrix} \boldsymbol{z}_1^T & \cdots & \boldsymbol{z}_{M-1}^T \end{bmatrix}^T \tag{8}$$

one obtains

$$\boldsymbol{z} = \boldsymbol{\Phi}_1 \boldsymbol{\alpha}, \tag{9}$$

where

$$\boldsymbol{\Phi}_1 = \begin{bmatrix} \boldsymbol{x}_2^T \otimes \boldsymbol{I}_n \\ \vdots \\ \boldsymbol{x}_M^T \otimes \boldsymbol{I}_n \end{bmatrix}. \tag{10}$$

By use of (9) an estimate of the nuisance variable $\boldsymbol{A}$ can be found. Thereafter it is straightforward to estimate $\boldsymbol{x}_0$ using (5). Thus, the two-step procedure becomes:

Step 1: Let $\boldsymbol{x}_1 = \boldsymbol{0}$ and $\boldsymbol{x}_k \neq \boldsymbol{0}$ for $k \geq 2$. Form (8), (10), and compute

$$\hat{\boldsymbol{\alpha}} = \boldsymbol{\Phi}_1^{\dagger} \boldsymbol{z}. \tag{11}$$

Thereafter, form the estimate $\hat{\boldsymbol{A}}$ from $\hat{\boldsymbol{\alpha}}$.

Step 2: Use the first experiment (5) and $\hat{\boldsymbol{A}}$ to estimate $\boldsymbol{x}_0$:

$$\hat{\boldsymbol{x}}_0 = \hat{\boldsymbol{A}}^{-1} \boldsymbol{y}_1. \tag{12}$$

The procedure to estimate unbalances by using $\boldsymbol{y}_1 = \boldsymbol{A}\boldsymbol{x}_0$ in order to linearize the equations is often referred to as the *influence coefficient method* in the balancing literature [6, 2, 9]. There exist many variants on how to organize the equations [3]. Another alternative that is more computationally efficient is to arrange the equations in the first step as

$$\begin{bmatrix} \boldsymbol{z}_2 & \cdots & \boldsymbol{z}_M \end{bmatrix} = \boldsymbol{A} \begin{bmatrix} \boldsymbol{x}_2 & \cdots & \boldsymbol{x}_M \end{bmatrix}$$
$$\hat{\boldsymbol{A}} = \begin{bmatrix} \boldsymbol{z}_2 & \cdots & \boldsymbol{z}_M \end{bmatrix} \begin{bmatrix} \boldsymbol{x}_2 & \cdots & \boldsymbol{x}_M \end{bmatrix}^{\dagger}.$$

The pseudo inverse of *broad* matrices should then be employed. The two variants yield the same result, but the one chosen for this paper is more tractable from a statistical analysis point of view.

4.2 Statistical Properties of A1

The statistical analysis is carried out under the following conditions:

Assumption 2. The stochastic disturbance $\tilde{\boldsymbol{A}}_k$ is small compared to $\boldsymbol{A}$. This means that the signal to noise ratio, SNR, is large. □

Remark 1. The number of experiments M is *not* assumed to be large. □

These conditions will be employed also for the analysis of the methods A2 and A3 that will be introduced in the sequel. Remark 1 is important since for the underlying application, a very large number of experiments would not be feasible. The first and second order statistics of A1 are summarized in the following lemma.

Lemma 1. *The expected value of the estimate* (12) *is,*

$$\mathbf{E}\hat{\boldsymbol{x}}_0 = \boldsymbol{x}_0 + \mathcal{O}\left(\mathbf{E}\|\tilde{\boldsymbol{A}}_k\|^2\right)$$

and its covariance matrix is for large SNR given by

$$\operatorname{cov}(\hat{\boldsymbol{x}}_0) = \boldsymbol{A}^{-1}\boldsymbol{C}_1\boldsymbol{R}_{\boldsymbol{\alpha},M}\boldsymbol{C}_1^*\boldsymbol{A}^{-*},$$

where

$$\boldsymbol{C}_1 = \left(\boldsymbol{x}_0^T \otimes \boldsymbol{I}_n\right)\left[\left(\boldsymbol{I}_{n^2} + \boldsymbol{\Phi}_1^\dagger \boldsymbol{C}_{1b}\right) \quad -\boldsymbol{\Phi}_1^\dagger \boldsymbol{C}_{1a}\right], \tag{13}$$

$$\boldsymbol{C}_{1a} = \begin{bmatrix} \left(\boldsymbol{x}_0 + \boldsymbol{x}_2\right)^T \otimes \boldsymbol{I}_n & & \boldsymbol{0} \\ & \ddots & \\ \boldsymbol{0} & & \left(\boldsymbol{x}_0 + \boldsymbol{x}_M\right)^T \otimes \boldsymbol{I}_n \end{bmatrix} \quad \in \mathbb{C}^{n(M-1)\times n^2(M-1)}, \tag{14}$$

$$\boldsymbol{C}_{1b} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \otimes \left(\boldsymbol{x}_0^T \otimes \boldsymbol{I}_n\right) \quad \in \mathbb{C}^{n(M-1)\times n^2}. \tag{15}$$

Proof *The proof is given in Appendix A.* □

4.3 Approach A2

In the second step (12) of A1, the unkwnon variable $\boldsymbol{x}_0$ is estimated using the first experiment only. This can be problematic if $\tilde{\boldsymbol{A}}_1$ happens to be large and the approach is not the soundest from a statistical point of view. One way to avoid this problem is to introduce the variable $\boldsymbol{m} = \boldsymbol{A}\boldsymbol{x}_0$. Equation (4) then becomes linear in $\boldsymbol{m}$ and $\boldsymbol{A}$ and all experiments can be used to identify these unknown parameters. Then, $\hat{\boldsymbol{x}}_0$ can be computed using their estimates. This

approach is taken in A2. Still, no nonlinear optimization is needed and the approach will be shown to have better statistical properties than A1.

Again, vectorization is employed and (4) can be rewritten as

$$\boldsymbol{y}_k = \boldsymbol{m} + \left(\boldsymbol{x}_k^T \otimes \boldsymbol{I}_n\right) \boldsymbol{\alpha} \tag{16}$$

and if all experiments are assembled one obtains

$$\boldsymbol{y} = \boldsymbol{\Phi}_2 \boldsymbol{\theta}, \tag{17}$$

where

$$\boldsymbol{y} = \begin{bmatrix} \boldsymbol{y}_1 \\ \vdots \\ \boldsymbol{y}_M \end{bmatrix}, \quad \boldsymbol{\Phi}_2 = \begin{bmatrix} \boldsymbol{I}_n & \boldsymbol{x}_1^T \otimes \boldsymbol{I}_n \\ \vdots & \vdots \\ \boldsymbol{I}_n & \boldsymbol{x}_M^T \otimes \boldsymbol{I}_n \end{bmatrix}, \quad \boldsymbol{\theta} = \begin{bmatrix} \boldsymbol{m} \\ \boldsymbol{\alpha} \end{bmatrix}.$$

The parameter vector $\boldsymbol{\theta}$ is estimated using the least squares estimate

$$\hat{\boldsymbol{\theta}} = \boldsymbol{\Phi}_2^{\dagger} \boldsymbol{y}, \tag{18}$$

which means that $\boldsymbol{m} = \boldsymbol{A}\boldsymbol{x}_0$ and $\boldsymbol{\alpha} = \mathrm{vec}(\boldsymbol{A})$ are estimated by use of *all* M experiments. Thereafter, $\hat{\boldsymbol{A}}$ is formed from $\hat{\boldsymbol{\alpha}}$ and $\boldsymbol{x}_0$ is estimated as

$$\hat{\boldsymbol{x}}_0 = \hat{\boldsymbol{A}}^{-1} \hat{\boldsymbol{m}}. \tag{19}$$

4.4 Statistical Properties of A2

Again the statistical analysis is carried out under Assumption 2 and Remark 1. The results are summarized in Lemma 2.

Lemma 2. *The expected value of the estimate* (19) *is,*

$$\mathbf{E}\hat{\boldsymbol{x}}_0 = \boldsymbol{x}_0 + \mathcal{O}\left(\mathbf{E}\|\tilde{\boldsymbol{A}}_k\|^2\right) \tag{20}$$

and its covariance matrix is for large SNR given by

$$\mathrm{cov}\left(\hat{\boldsymbol{x}}_0\right) = \boldsymbol{A}^{-1} \boldsymbol{C}_{2a} \boldsymbol{\Phi}_2^{\dagger} \boldsymbol{C}_2 \boldsymbol{R}_{\boldsymbol{\alpha},M} \boldsymbol{C}_2^* \boldsymbol{\Phi}_2^{\dagger *} \boldsymbol{C}_{2a}^* \boldsymbol{A}^{-*}, \tag{21}$$

where

$$\boldsymbol{C}_2 = \begin{bmatrix} \left(\boldsymbol{x}_0 + \boldsymbol{x}_1\right)^T \otimes \boldsymbol{I}_n & & \boldsymbol{0} \\ & \ddots & \\ \boldsymbol{0} & & \left(\boldsymbol{x}_0 + \boldsymbol{x}_M\right)^T \otimes \boldsymbol{I}_n \end{bmatrix} \in \mathbb{C}^{nM \times n^2 M}, \tag{22}$$

$$\boldsymbol{C}_{2a} = \begin{bmatrix} \boldsymbol{I}_n & -\boldsymbol{x}_0^T \otimes \boldsymbol{I}_n \end{bmatrix}, \quad \in \mathbb{C}^{n \times n(n+1)}. \tag{23}$$

Proof *See Appendix A.* □

5 Approach 3: Nonlinear Regression

In this section we drive a loss function that handles the stochastic uncertainty $\tilde{\boldsymbol{A}}$ in a more sophisticated fashion. This leads to a problem formulation with a loss function that is nonlinear in $\boldsymbol{x}_0$. Thus, there exists no closed form solution and a numerical search procedure is required. In order to use standard optimization routines, the system equation (1) is reformulated as a real valued problem. This is done by representing the complex valued quantities with their real and imaginary parts separated. This operation is denoted here with $\bar{(\cdot)}$ and we let

$$\bar{\boldsymbol{y}}_k = \begin{bmatrix} \boldsymbol{y}_{kR} \\ \boldsymbol{y}_{kI} \end{bmatrix} = \begin{bmatrix} \mathrm{Re}\,(\boldsymbol{y}_k) \\ \mathrm{Im}\,(\boldsymbol{y}_k) \end{bmatrix}, \quad \bar{\boldsymbol{x}}_0 = \begin{bmatrix} \boldsymbol{x}_{0R} \\ \boldsymbol{x}_{0I} \end{bmatrix}, \quad \bar{\boldsymbol{x}}_k = \begin{bmatrix} \boldsymbol{x}_{kR} \\ \boldsymbol{x}_{kI} \end{bmatrix}, \quad \in \mathbb{R}^{2n\times 1},$$

where $\mathbb{R}^{n\times m}$ denotes real valued $n\times m$ matrix and where $\mathrm{Re}(y_k)$ and $\mathrm{Im}(y_k)$ are the real and imaginary parts of $\boldsymbol{y}_k$, respectively. The corresponding convention with subscripts R and I will be employed in the sequel. Furthermore, the vectorized matrices with separated real and imaginary parts are defined as

$$\bar{\boldsymbol{\alpha}} = \begin{bmatrix} \boldsymbol{\alpha}_R \\ \boldsymbol{\alpha}_I \end{bmatrix} = \begin{bmatrix} \mathrm{Re}\,(\mathrm{vec}\,(\boldsymbol{A})) \\ \mathrm{Im}\,(\mathrm{vec}\,(\boldsymbol{A})) \end{bmatrix} \in \mathbb{R}^{2n^2\times 1},$$

$$\bar{\tilde{\boldsymbol{\alpha}}}_k = \begin{bmatrix} \tilde{\boldsymbol{\alpha}}_{kR} \\ \tilde{\boldsymbol{\alpha}}_{kI} \end{bmatrix} = \begin{bmatrix} \mathrm{Re}\left(\mathrm{vec}\left(\tilde{\boldsymbol{A}}_k\right)\right) \\ \mathrm{Im}\left(\mathrm{vec}\left(\tilde{\boldsymbol{A}}_k\right)\right) \end{bmatrix} \in \mathbb{R}^{2n^2\times 1}, \quad \bar{\tilde{\boldsymbol{\alpha}}} = \begin{bmatrix} \bar{\tilde{\boldsymbol{\alpha}}}_1 \\ \vdots \\ \bar{\tilde{\boldsymbol{\alpha}}}_M \end{bmatrix} \in \mathbb{R}^{2n^2 M\times 1},$$

and the corresponding covariance matrices are defined as

$$\mathrm{cov}\left(\bar{\tilde{\boldsymbol{\alpha}}}_k\right) = \bar{\boldsymbol{R}}_{\boldsymbol{\alpha}}, \quad \mathrm{cov}\left(\bar{\tilde{\boldsymbol{\alpha}}}\right) = \bar{\boldsymbol{R}}_{\boldsymbol{\alpha},M} = \boldsymbol{I}_M \otimes \bar{\boldsymbol{R}}_{\boldsymbol{\alpha}}.$$

A given complex valued equation

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}, \quad \Rightarrow \quad \boldsymbol{y}_R + \mathrm{i}\boldsymbol{y}_I = (\boldsymbol{A}_R + \mathrm{i}\boldsymbol{A}_I)\,(\boldsymbol{x}_R + \mathrm{i}\boldsymbol{x}_I)$$

can be reformulated as a real valued relation

$$\begin{bmatrix} \boldsymbol{y}_R \\ \boldsymbol{y}_I \end{bmatrix} = \begin{bmatrix} \boldsymbol{A}_R & -\boldsymbol{A}_I \\ \boldsymbol{A}_I & \boldsymbol{A}_R \end{bmatrix} \begin{bmatrix} \boldsymbol{x}_R \\ \boldsymbol{x}_I \end{bmatrix}. \tag{24}$$

and by use of the vec operator (on each block row of (24)) one obtains

$$\begin{bmatrix} \boldsymbol{y}_R \\ \boldsymbol{y}_I \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}_R^T \otimes \boldsymbol{I}_n & -\boldsymbol{x}_I^T \otimes \boldsymbol{I}_n \\ \boldsymbol{x}_I^T \otimes \boldsymbol{I}_n & \boldsymbol{x}_R^T \otimes \boldsymbol{I}_n \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha}_R \\ \boldsymbol{\alpha}_I \end{bmatrix}.$$

Using this fact, the system equation (1) can be rewritten as

$$\bar{\boldsymbol{y}}_k = \boldsymbol{D}_k(\bar{\boldsymbol{x}}_0)\bar{\boldsymbol{\alpha}} + \boldsymbol{D}_k(\bar{\boldsymbol{x}}_0)\tilde{\bar{\boldsymbol{\alpha}}}_k, \quad k = 1, \cdots, M,$$

where

$$\boldsymbol{D}_k(\bar{\boldsymbol{x}}_0) = \begin{bmatrix} (\boldsymbol{x}_{0R} + \boldsymbol{x}_{kR})^T \otimes \boldsymbol{I}_n & -(\boldsymbol{x}_{0I} + \boldsymbol{x}_{kI})^T \otimes \boldsymbol{I}_n \\ (\boldsymbol{x}_{0I} + \boldsymbol{x}_{kI})^T \otimes \boldsymbol{I}_n & (\boldsymbol{x}_{0R} + \boldsymbol{x}_{kR})^T \otimes \boldsymbol{I}_n \end{bmatrix} \in \mathbb{R}^{2n\times 2n^2}. \tag{25}$$

If all experiments are stacked in a tall vector

$$\bar{\boldsymbol{y}} = \begin{bmatrix} \bar{\boldsymbol{y}}_1^T & \bar{\boldsymbol{y}}_2^T & \cdots & \bar{\boldsymbol{y}}_M^T \end{bmatrix}^T,$$

one obtains

$$\boxed{\bar{\boldsymbol{y}} - \boldsymbol{B}(\bar{\boldsymbol{x}}_0)\bar{\boldsymbol{\alpha}} = \boldsymbol{C}(\bar{\boldsymbol{x}}_0)\bar{\tilde{\boldsymbol{\alpha}}},} \tag{26}$$

where

$$\boldsymbol{B}(\bar{\boldsymbol{x}}_0) = \begin{bmatrix} \boldsymbol{D}_1(\bar{\boldsymbol{x}}_0) \\ \vdots \\ \boldsymbol{D}_M(\bar{\boldsymbol{x}}_0) \end{bmatrix}, \quad \boldsymbol{C}(\bar{\boldsymbol{x}}_0) = \begin{bmatrix} \boldsymbol{D}_1(\bar{\boldsymbol{x}}_0) & & \boldsymbol{0} \\ & \ddots & \\ \boldsymbol{0} & & \boldsymbol{D}_M(\bar{\boldsymbol{x}}_0) \end{bmatrix}. \tag{27}$$

The covariance matrix of the residual term $\boldsymbol{C}(\bar{\boldsymbol{x}}_0)\bar{\tilde{\boldsymbol{\alpha}}}$ is denoted by

$$\boldsymbol{Q}_{\boldsymbol{\alpha}}(\bar{\boldsymbol{x}}_0) = \boldsymbol{C}(\bar{\boldsymbol{x}}_0)\bar{\boldsymbol{R}}_{\boldsymbol{\alpha},M}\boldsymbol{C}^T(\bar{\boldsymbol{x}}_0) \quad \in \mathbb{R}^{2nM\times 2nM}, \tag{28}$$

which is a function of the unknown variable $\bar{\boldsymbol{x}}_0$. Similarly to the approaches A1 and A2 an estimate of $\bar{\boldsymbol{x}}_0$ is found by minimizing a quadratic criterion. However, in order to make the covariance matrix of the estimation error minimal, the equations should be weighted with the inverse of $\boldsymbol{Q}_{\boldsymbol{\alpha}}$ [13]. The criterion then reads

$$V(\bar{\boldsymbol{x}}, \boldsymbol{\alpha}) = \|\bar{\boldsymbol{y}} - \boldsymbol{B}(\bar{\boldsymbol{x}})\bar{\boldsymbol{\alpha}}\|^2_{\boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1}(\bar{\boldsymbol{x}})}. \tag{29}$$

Minimization of V with respect to $\boldsymbol{\alpha}$ is straightforward. For a fixed value of $\bar{\boldsymbol{x}} = \bar{\boldsymbol{x}}_*$, the minimum is [13]

$$\hat{\bar{\boldsymbol{\alpha}}} = \left[\boldsymbol{B}^T(\bar{\boldsymbol{x}}_*)\boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1}(\bar{\boldsymbol{x}}_*)\boldsymbol{B}(\bar{\boldsymbol{x}}_*)\right]^{-1}\boldsymbol{B}^T(\bar{\boldsymbol{x}}_*)\boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1}(\bar{\boldsymbol{x}}_*)\bar{\boldsymbol{y}} \tag{30}$$

and insertion of (30) into (29) yields a concentrated loss function

$$\begin{aligned} W(\bar{\boldsymbol{x}}) &= \min_{\boldsymbol{\alpha}} V(\bar{\boldsymbol{x}}, \boldsymbol{\alpha}) = \left\| \bar{\boldsymbol{y}} - \boldsymbol{B}\left[\boldsymbol{B}^T\boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1}\boldsymbol{B}\right]^{-1}\boldsymbol{B}^T\boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1}\bar{\boldsymbol{y}} \right\|^2_{\boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1}} \\ &= \left(\bar{\boldsymbol{y}}^T - \bar{\boldsymbol{y}}^T\boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1}\boldsymbol{B}\left[\boldsymbol{B}^T\boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1}\boldsymbol{B}\right]^{-1}\boldsymbol{B}^T\right)\boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1}\left(\bar{\boldsymbol{y}} - \boldsymbol{B}\left[\boldsymbol{B}^T\boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1}\boldsymbol{B}\right]^{-1}\boldsymbol{B}^T\boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1}\bar{\boldsymbol{y}}\right) \\ &= \bar{\boldsymbol{y}}^T\boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1/2}\left(\boldsymbol{I}_{2nM} - \boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1/2}\boldsymbol{B}\left[\boldsymbol{B}^T\boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1}\boldsymbol{B}\right]^{-1}\boldsymbol{B}^T\boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1/2}\right)\boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1/2}\bar{\boldsymbol{y}} \end{aligned} \tag{31}$$

where the dependence on $\bar{\boldsymbol{x}}$ is dropped for brevity. The concentrated loss function (31) can be formulated as

$$W(\bar{\boldsymbol{x}}) = \bar{\boldsymbol{y}}^T\boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1/2}(\bar{\boldsymbol{x}})\boldsymbol{\Pi}^{\perp}(\bar{\boldsymbol{x}})\boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1/2}(\bar{\boldsymbol{x}})\bar{\boldsymbol{y}}, \tag{32}$$

where $\boldsymbol{\Pi}^{\perp}$ is the orthogonal projector onto the null-space of $\boldsymbol{B}^T\boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1/2}$ and it is given by

$$\boldsymbol{\Pi}^{\perp} = \boldsymbol{I}_{2nM} - \boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1/2}\boldsymbol{B}\left[\boldsymbol{B}^T\boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1}\boldsymbol{B}\right]^{-1}\boldsymbol{B}^T\boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1/2}. \tag{33}$$

The parameter estimation problem becomes a two-step procedure:

$$\hat{\bar{\boldsymbol{x}}}_0 = \min_{\bar{\boldsymbol{x}}} W(\bar{\boldsymbol{x}}) \tag{34}$$

$$\hat{\bar{\boldsymbol{\alpha}}} = \left[\boldsymbol{B}^T(\hat{\bar{\boldsymbol{x}}}_0)\boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1}(\hat{\bar{\boldsymbol{x}}}_0)\boldsymbol{B}(\hat{\bar{\boldsymbol{x}}}_0)\right]^{-1}\boldsymbol{B}^T(\hat{\bar{\boldsymbol{x}}}_0)\boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1}(\hat{\bar{\boldsymbol{x}}}_0)\bar{\boldsymbol{y}}. \tag{35}$$

By the separation into two estimation steps the complexity of the optimization problem has been significantly reduced. Minimization of the original loss function (29) would require a nonlinear search over $2(n^2+n)$ unknown parameters. By use of the concentrated loss function (31), the problem is reduced to a nonlinear minimization over $2n$ variables and a simple weighted linear least squares fit to find the remaining $2n^2$ unknown parameters. The second step is only needed if the nuisance variable $\boldsymbol{A}$ is of any importance.

The optimization problem (34) is often referred to as a variable projection problem [4]. Such optimization problems frequently appear in sensor array processing [14] and in many other applications [5]. However, the fact that $\boldsymbol{Q}_{\boldsymbol{\alpha}}$ in (32) is a function of the unknown variable is quite uncommon. Notice that $\boldsymbol{Q}_{\boldsymbol{\alpha}}$ depends on the uncertainty covariance matrix $\bar{\boldsymbol{R}}_{\boldsymbol{\alpha}}$ through (28). Therefore, $\bar{\boldsymbol{R}}_{\boldsymbol{\alpha}}$ needs to be a priori known or estimated.

5.1 Statistical Properties of A3

First notice that the outcome $\hat{\bar{\boldsymbol{x}}}_0$ from the optimization (34) is such that

$$W^{(k)}(\hat{\bar{\boldsymbol{x}}}_0) = 0$$

for a successful minimization. Assume that the estimate $\hat{\bar{\boldsymbol{x}}}_0$ lies in a neighborhood close to the true value $\bar{\boldsymbol{x}} = \bar{\boldsymbol{x}}_0$, *i.e.* $\hat{\bar{\boldsymbol{x}}}_0 = \bar{\boldsymbol{x}}_0 + \bar{\boldsymbol{x}}_\varepsilon$, where $\bar{\boldsymbol{x}}_\varepsilon$ is small. Then (see *e.g.* [10, 13]),

$$\mathbf{0} = \left.\frac{\partial^T W}{\partial \bar{\boldsymbol{x}}}\right|_{\bar{\boldsymbol{x}}=\hat{\bar{\boldsymbol{x}}}_0} = \left.\frac{\partial^T W}{\partial \bar{\boldsymbol{x}}}\right|_{\bar{\boldsymbol{x}}=\bar{\boldsymbol{x}}_0+\bar{\boldsymbol{x}}_\varepsilon} \approx \left.\frac{\partial^T W}{\partial \bar{\boldsymbol{x}}}\right|_{\bar{\boldsymbol{x}}=\bar{\boldsymbol{x}}_0} + \left.\frac{\partial^2 W}{\partial \bar{\boldsymbol{x}}^2}\right|_{\bar{\boldsymbol{x}}=\bar{\boldsymbol{x}}_0} \bar{\boldsymbol{x}}_\varepsilon. \tag{36}$$

Remember that $W^{(k)} = \partial W/\partial[\bar{\boldsymbol{x}}]_k$, where $[\bar{\boldsymbol{x}}]_k$ is the k-th element of $\bar{\boldsymbol{x}}$, see Section 3. Equation (36) implies that the estimation error approximately is

$$\bar{\boldsymbol{x}}_\varepsilon = -\left(\frac{\partial^2 W}{\partial \bar{\boldsymbol{x}}^2}\right)^{-1} \frac{\partial^T W}{\partial \bar{\boldsymbol{x}}}, \tag{37}$$

where the derivatives should be evaluated at $\bar{\boldsymbol{x}} = \bar{\boldsymbol{x}}_0$. The accuracy of the estimate then becomes

$$\operatorname{cov}(\bar{\boldsymbol{x}}_\varepsilon) = \left(\frac{\partial^2 W}{\partial \bar{\boldsymbol{x}}^2}\right)^{-1} \operatorname{cov}\left(\frac{\partial^T W}{\partial \bar{\boldsymbol{x}}}\right) \left(\frac{\partial^2 W}{\partial \bar{\boldsymbol{x}}^2}\right)^{-1}. \tag{38}$$

Thus, we need to evaluate the gradient and the Hessian,

$$\frac{\partial^T W}{\partial \bar{\boldsymbol{x}}} = \begin{bmatrix} W^{(1)} \\ \vdots \\ W^{(2n)} \end{bmatrix}, \quad \frac{\partial^2 W}{\partial \bar{\boldsymbol{x}}^2} = \begin{bmatrix} W^{(11)} & W^{(12)} & \dots & W^{(1(2n))} \\ W^{(21)} & W^{(22)} & \dots & W^{(2(2n))} \\ \vdots & \vdots & \ddots & \vdots \\ W^{((2n)1)} & W^{((2n)2)} & \dots & W^{((2n)(2n))} \end{bmatrix}, \tag{39}$$

of the loss function and evaluate them at $\bar{\boldsymbol{x}} = \bar{\boldsymbol{x}}_0$. In order to accomplish this it is useful to rewrite the criterion function (32) as

$$W(\bar{\boldsymbol{x}}) = \bar{\boldsymbol{y}}^T \boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1}(\bar{\boldsymbol{x}}) \left(\boldsymbol{I}_{2nM} - \boldsymbol{B}(\bar{\boldsymbol{x}})\boldsymbol{P}(\bar{\boldsymbol{x}})\right) \bar{\boldsymbol{y}}, \tag{40}$$

where

$$\boldsymbol{P}(\bar{\boldsymbol{x}}) = \left[\boldsymbol{B}^T(\bar{\boldsymbol{x}})\boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1}(\bar{\boldsymbol{x}})\boldsymbol{B}(\bar{\boldsymbol{x}})\right]^{-1}\boldsymbol{B}^T(\bar{\boldsymbol{x}})\boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1}(\bar{\boldsymbol{x}}). \tag{41}$$

The matrix $\boldsymbol{P}$ has some useful properties that are summarized in what follows. For notational convenience the dependence on $\bar{\boldsymbol{x}}$ is dropped. All proofs are given in Appendix B.

Proposition 1.

$$\boldsymbol{P}\boldsymbol{B} = \boldsymbol{I}_{2nM} \tag{42}$$

□

Proposition 2.

$$\boldsymbol{P}^{(k)} = \left[\boldsymbol{B}^T\boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1}\boldsymbol{B}\right]^{-1}\left(\boldsymbol{B}^{T(k)}\boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1} + \boldsymbol{B}^T\boldsymbol{Q}_{\boldsymbol{\alpha}}^{-(k)}\right)(\boldsymbol{I}_{2nM} - \boldsymbol{B}\boldsymbol{P}) - \boldsymbol{P}\boldsymbol{B}^{(k)}\boldsymbol{P} \tag{43}$$

□

A very useful consequence of the first proposition is

Proposition 3.

$$\boldsymbol{P}^{(k)}\boldsymbol{B} = -\boldsymbol{P}\boldsymbol{B}^{(k)}. \tag{44}$$

□

The final proposition is related to the second derivatives of $\boldsymbol{P}$:

Proposition 4.

$$\begin{aligned}\left[\boldsymbol{B}^T\boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1}\boldsymbol{B}\right]\boldsymbol{P}^{(kl)}\boldsymbol{B} =& \boldsymbol{B}^T\boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1}\left(\boldsymbol{B}^{(l)}\boldsymbol{P}\boldsymbol{B}^{(k)} + \boldsymbol{B}^{(k)}\boldsymbol{P}\boldsymbol{B}^{(l)}\right)\\ &- \left(\boldsymbol{B}^{T(k)}\boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1} + \boldsymbol{B}^T\boldsymbol{Q}_{\boldsymbol{\alpha}}^{-(k)}\right)(\boldsymbol{I}_{2nM} - \boldsymbol{B}\boldsymbol{P})\,\boldsymbol{B}^{(l)}\\ &- \left(\boldsymbol{B}^{T(l)}\boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1} + \boldsymbol{B}^T\boldsymbol{Q}_{\boldsymbol{\alpha}}^{-(l)}\right)(\boldsymbol{I}_{2nM} - \boldsymbol{B}\boldsymbol{P})\,\boldsymbol{B}^{(k)}\end{aligned} \tag{45}$$

□

After these technical results we present one lemma needed in order to compute the gradient and Hessian (39) of the concentrated loss function $W(\bar{\boldsymbol{x}})$, (32).

Lemma 3. *Under Assumption 2* ($\|\tilde{\boldsymbol{A}}_k\| \ll \|\boldsymbol{A}\|$) *it holds that*

$$W^{(k)}(\bar{\boldsymbol{x}}_0) \approx -2\bar{\boldsymbol{\alpha}}^T\boldsymbol{B}^{T(k)}\boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1/2}\boldsymbol{\Pi}^{\perp}\boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1/2}\boldsymbol{C}\bar{\tilde{\boldsymbol{\alpha}}} \tag{46}$$

$$W^{(kl)}(\bar{\boldsymbol{x}}_0) \approx 2\bar{\boldsymbol{\alpha}}^T\boldsymbol{B}^{T(k)}\boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1/2}\boldsymbol{\Pi}^{\perp}\boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1/2}\boldsymbol{B}^{(l)}\bar{\boldsymbol{\alpha}}. \tag{47}$$

□

We are now ready to give the main result of this section:

Lemma 4. *The estimation procedure A3 yields*

$$\mathbf{E}\hat{\boldsymbol{x}}_0 = \boldsymbol{x}_0 + \mathcal{O}\left(\mathbf{E}\|\tilde{\boldsymbol{A}}_k\|^2\right) \tag{48}$$

and the accuracy is for large SNR given by

$$\operatorname{cov}\left(\hat{\bar{\boldsymbol{x}}}_0\right) = \boldsymbol{\mathcal{H}}^{-1}\boldsymbol{G}\bar{\boldsymbol{R}}_{\boldsymbol{\alpha},M}\boldsymbol{G}^T\boldsymbol{\mathcal{H}}^{-1}, \tag{49}$$

where

$$[\boldsymbol{\mathcal{H}}]_{kl} = 2\bar{\boldsymbol{\alpha}}^T\boldsymbol{B}^{T(k)}\boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1/2}\boldsymbol{\Pi}^{\perp}\boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1/2}\boldsymbol{B}^{(l)}\bar{\boldsymbol{\alpha}}, \quad (k,l) = 1,\cdots,2n, \tag{50}$$

$$[\boldsymbol{G}]_{k,:} = -2\bar{\boldsymbol{\alpha}}^T\boldsymbol{B}^{T(k)}\boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1/2}\boldsymbol{\Pi}^{\perp}\boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1/2}\boldsymbol{C}, \quad k = 1,\cdots,2n, \tag{51}$$

where $[\boldsymbol{G}]_{k,:}$ *means row* k *of the matrix* $\boldsymbol{G}$. □

Again, notice that all proofs are given in Appendix B.

Remark 2: The results presented apply for any variable projection problem of the type (32). The only assumption made is that $\boldsymbol{B}$ is linear in $\bar{\boldsymbol{x}}$, so that $\boldsymbol{B}^{(kl)} = \boldsymbol{0}$. If $\boldsymbol{B}$ would be a nonlinear function of $\bar{\boldsymbol{x}}$, terms that involve $\boldsymbol{B}^{(kl)}$ appear in the results above. The details needed in order to carry out the final computations for the specific problem at hand are given in Appendix C.

5.2 Computational Aspects

The loss function (32) is a nonlinear function of the unknown variable $\bar{\boldsymbol{x}}_0$. Therefore, numerical optimization is needed in order to compute the estimate $\hat{\bar{\boldsymbol{x}}}_0$. For this purpose, there are some computational issues that need to be addressed.

5.2.1 Initialization

Any optimization routine need to be started with an initial guess of the minimizing variable. Instead of just choosing *e.g.* $\hat{\bar{\boldsymbol{x}}}_0 = \boldsymbol{0}$, the optimization is initialized with the outcome from the procedure A2.

5.2.2 Existence of Local Minima

It is not easily seen if there exists local minima from the expression (32). So far, no problems with convergence to inaccurate estimates have been experienced. If $n = 1$, it is possible to visually depict the level curves of the concentrated loss function. Such an example is shown in Figure 2. Here, the number of experiments is $M = 7$ and

$$\boldsymbol{A} = 1 + 0.78\mathrm{i}, \quad \boldsymbol{x}_0 = 0.55, \quad \bar{\boldsymbol{R}}_{\boldsymbol{\alpha}} = \operatorname{cov}\left(\bar{\tilde{\boldsymbol{\alpha}}}_k\right) = 10^{-3}\boldsymbol{I}_2.$$

The figure shows that at least in this case the loss function is well behaved.

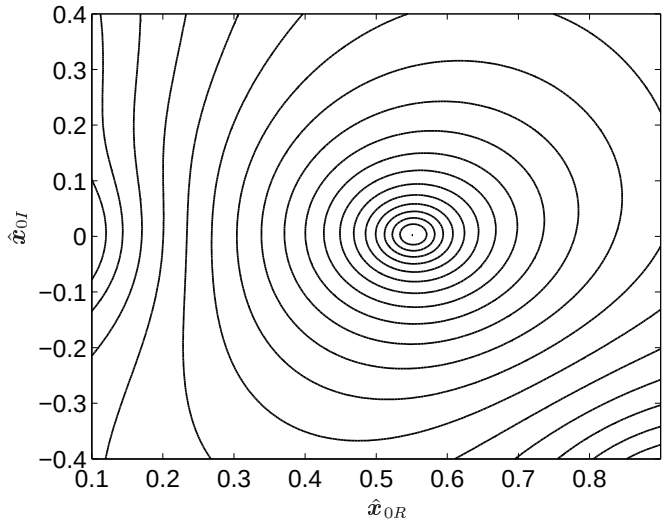

Figure 2: Level curves of the loss function. The true parameter value is $\boldsymbol{x}_0 = 0.55$.

5.2.3 The Covariance Matrix $\bar{\boldsymbol{R}}_{\boldsymbol{\alpha}}$

In most applications, the covariance matrix $\boldsymbol{Q}_{\boldsymbol{\alpha}}$ should be positive definite. However, situations where it is ill conditioned, or rank deficient may occur. Such situations need to be taken care of. It can be done using regularization,

$$\boldsymbol{Q}_{\boldsymbol{\alpha}} = \boldsymbol{C}\bar{\boldsymbol{R}}_{\boldsymbol{\alpha},M}\boldsymbol{C}^T + \mu\boldsymbol{I}_{2nM},$$

where μ is a small real number.

In order to use approach A3, the statistics of the uncertainty must be known or estimated beforehand. The good news is that only the structure of $\bar{\boldsymbol{R}}_{\boldsymbol{\alpha}}$ and not its absolute value is of importance. A scaling of the covariance matrix will only scale the loss function (32). Thus, the value of $\hat{\hat{\boldsymbol{x}}}_0$ that minimizes value (32) will remain the same.

5.2.4 Computation of the Loss Function

When the projection matrix $\boldsymbol{\Pi}^{\perp}$ is computed, the effects of rounding errors may become significant. Therefore, it should be computed in a numerically sound way. First, rewrite (33) as [12]

$$\boldsymbol{\Pi}^{\perp} = \boldsymbol{I}_{2nM} - \boldsymbol{M}\boldsymbol{M}^{\dagger}, \quad \boldsymbol{M} = \boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1/2}\boldsymbol{B}$$

and perform the QR factorization

$$\boldsymbol{M} = \boldsymbol{Q}\boldsymbol{R} = \begin{bmatrix}\boldsymbol{Q}_1 & \boldsymbol{Q}_2\end{bmatrix}\begin{bmatrix}\boldsymbol{R}_1 \\ \boldsymbol{0}\end{bmatrix}, \tag{52}$$

where $\boldsymbol{Q}$ is an orthogonal matrix and $\boldsymbol{R}$ is upper triangular. Equation (52) implies

$$\boldsymbol{M} = \boldsymbol{Q}_1 \boldsymbol{R}_1,$$

which gives

$$\boldsymbol{M}^\dagger = \left(\boldsymbol{R}_1^T \boldsymbol{Q}_1^T \boldsymbol{Q}_1 \boldsymbol{R}_1\right)^{-1} \boldsymbol{R}_1^T \boldsymbol{Q}_1^T = \boldsymbol{R}_1^{-1} \boldsymbol{Q}_1^T,$$

where the last equality follows from the orthogonality of $\boldsymbol{Q}_1$. Using this result, the projection matrix can be written as

$$\boldsymbol{\Pi}^\perp = \boldsymbol{I}_{2nM} - \boldsymbol{Q}_1 \boldsymbol{R}_1 \boldsymbol{R}_1^{-1} \boldsymbol{Q}_1^T = \boldsymbol{I}_{2nM} - \boldsymbol{Q}_1 \boldsymbol{Q}_1^T = \boldsymbol{Q}_2 \boldsymbol{Q}_2^T. \tag{53}$$

Equation (53) less sensitive to rounding errors compared to direct computation of (33). In addition, the use of $\boldsymbol{Q}_2$ forces (53) to be positive semidefinite. Therefore, the QR decomposition approach should be used for the numerical computations.

5.2.5 Speeding up Convergence

Many optimization routines converge in fewer iterations if in each step the analytical value of the gradient of the loss function is supplied. Such expressions are given in Appendix D, for any $\bar{\boldsymbol{x}} \neq \bar{\boldsymbol{x}}_0$.

6 Numerical Examples

In this section the performance of the different estimators are compared by means of two numerical examples. The first example is ad hoc and used to compare the analytical accuracy results with Monte Carlo simulations. The second example is more realistic and deals with unbalance estimation of rotating machinery.

6.1 Example 1

The accuracy results for the three different approaches described in Section 4 (A1, A2) and Section 5 (A3), are not really transparent and therefore they are evaluated in a simple example.

Consider a system given by

$$\boldsymbol{y}_k = \left(\boldsymbol{A} + \tilde{\boldsymbol{A}}_k\right)\left(\boldsymbol{x}_0 + \boldsymbol{x}_k\right), \quad k = 1, \cdots, M,$$

where $\boldsymbol{y}_k$ is the measurement of experiment k and $\boldsymbol{x}_k$ is a user chosen variable. The unknown system parameters are given by

$$\boldsymbol{A} = \begin{bmatrix} 0.3 + 4\mathrm{i} & 0.2 + \mathrm{i} \\ 0.2 - \mathrm{i} & 0.5 - 0.5\mathrm{i} \end{bmatrix}, \quad \boldsymbol{x}_0 = \begin{bmatrix} 1 \\ 0.9 \end{bmatrix}$$

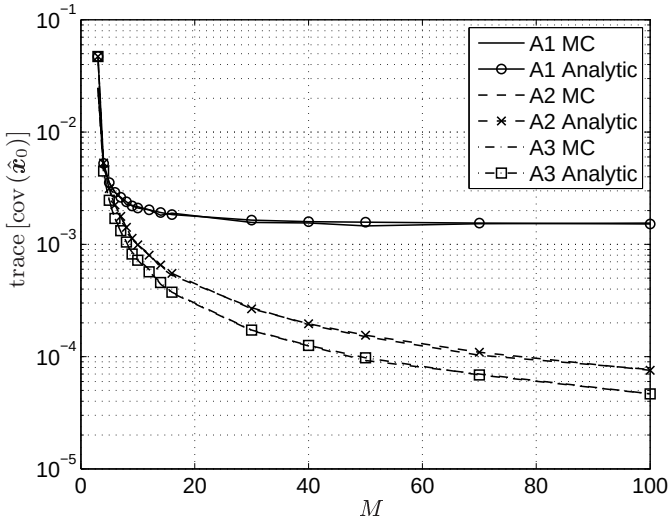

Figure 3: Results of Monte Carlo simulations (MC) and analytical results for Example 1.

and the disturbance has zero mean and covariance matrix

$$\bar{\boldsymbol{R}}_{\boldsymbol{\alpha}} = 10^{-4}\boldsymbol{I}_8.$$

For a given experiment, the user chosen variable is either zero or drawn from a uniform distribution,

$$\boldsymbol{x}_k = \begin{cases} \boldsymbol{0} & \text{if } k = 1 \\ \in \mathcal{CU}(-2,2) & \text{if } k = 2, \cdots, M, \end{cases}$$

where $\mathcal{CU}(-2,2)$ means that the real and imaginary parts both are uniformly distributed in the interval $(2,2)$. The variables $\boldsymbol{x}_2, \cdots, \boldsymbol{x}_M$ are independent. The effect of the number of experiments M used to estimate $\boldsymbol{x}_0$ is evaluated using Monte Carlo simulations. For each value of M, the covariance matrix of the estimates from algorithms A1, A2 and A3 are computed from 1000 realizations. The same data set is employed for all algorithms and the results are compared with the corresponding analytical covariance matrices.

Figure 3 shows the trace of $\operatorname{cov}(\hat{\boldsymbol{x}}_0)$ for the different algorithms. Notice that $\operatorname{trace}[\operatorname{cov}(\hat{\boldsymbol{x}}_0)] = \operatorname{trace}\left[\operatorname{cov}\left(\hat{\tilde{\boldsymbol{x}}}_0\right)\right]$. The smallest number of experiments required in order to estimate $\boldsymbol{x}_0$ is $M = n + 1 = 3$. Notice also that if $M = 3$, all algorithms have precisely the same performance. It depends on that the number of equations are equal to the number of unknowns and there exist only one *unique* solution, which is independent on how the equations are weighted.

The curves in Figure 3 show that the analytic results coincide very well with the results from Monte Carlo simulations. It is also seen that even though A1 may compete with A2 and A3 when the estimation is performed with a small

number of experiments (*e.g.* $M < 6$), it performs significantly worse than A2 and A3 when M is increased. The reason for this is that A1 relies so much on the first experiment when $\hat{\boldsymbol{x}}_0$ is computed. Therefore, A1 does not benefit from statistical averaging when the number of experiments is increased.

In this simple example, A2 and A3 performs almost similar. The algorithm A3 does benefit from its optimal weighting and always performs slightly better than A2. However, one may argue that the improvement is dearly bought. In the next example, we show that A3 sometimes actually pays off.

6.2 Example 2

Consider a model of a separator as described in Section 2.1 and shown in Figure 1. It is a 2-dimensional model with 12 degrees of freedom. The beam at which the separator bowl is attached is however modeled with the Euler-Bernoulli partial differential equation. The masses of the bowl and the frames are in the order hundreds of kilograms. The stiffnesses are modeled using the concept of hysteretic damping. It means they are modeled as a complex valued stiffness, which is a is a way to introduce damping in the system. The damping does not change with frequency, in contrast to viscous damping. The complex valued stiffnesses are subject to change between experiments, which leads to the uncertainty term. Between each experiment, each stiffness varies uniformly ± 1 percent around its nominal value. The modeling is quite extensive and the details are by purpose left out in order to make the presentation compact. The system model becomes

$$\boldsymbol{y}_k = \left(\boldsymbol{A} + \tilde{\boldsymbol{A}}_k\right)(\boldsymbol{x}_0 + \boldsymbol{x}_k), \quad k = 1, \cdots, M,$$

where

$$A = 10^{-4}\begin{bmatrix} -0.0095 - 0.5335\mathrm{i} & 0.0036 + 0.1743\mathrm{i} \\ -0.0089 - 0.4344\mathrm{i} & 0.0017 + 0.1932\mathrm{i} \end{bmatrix}, \quad [\mathrm{m}/(\mathrm{sg})],$$

$$\boldsymbol{x}_0 = \begin{bmatrix} 21\mathrm{e}^{37\frac{\pi}{180}\mathrm{i}} \\ 17\mathrm{e}^{111\frac{\pi}{180}\mathrm{i}} \end{bmatrix}, \quad [\mathrm{g}].$$

The unit of $\boldsymbol{A}$ depends on the fact that the measured quantity is in [m/s] and the applied masses are in grams [g]. The quantities are complex valued since they are associated with a magnitude and an angular position. The structure of the covariance matrix $\bar{\boldsymbol{R}}_{\boldsymbol{\alpha}}$ is depicted in Figure 4.

In order to use A3, the statistics of the uncertainty must known or estimated somehow. Two scenarios here are evaluated. The first is that the statistics of the uncertainty is fully known. The other scenario is that it is completely unknown and therefore $\hat{\bar{\boldsymbol{R}}}_{\boldsymbol{\alpha}} = \boldsymbol{I}_8$ is employed. The latter choice clearly deviates from the true covariance matrix as depicted in Figure 4. Still, the algorithm A3 can be used, but the weighting is no longer optimal. Therefore, it is not necessarily so that A3 should perform better than the other two approaches in this case.

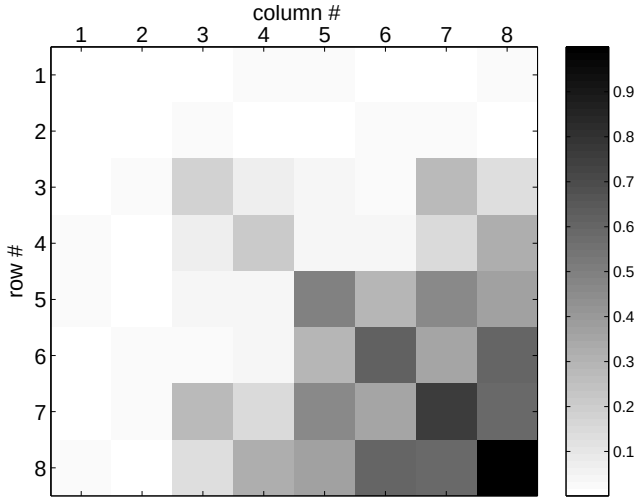

Figure 4: The structure of $\bar{\boldsymbol{R}}_{\boldsymbol{\alpha}}$. Each square shows the magnitude of the corresponding element in $\bar{\boldsymbol{R}}_{\boldsymbol{\alpha}}$. The matrix is scaled so that the greatest element have unit magnitude.

Each trial weight $[\boldsymbol{x}_k]_i$ has certain mass m_i and angular position ϕ_i, relative to a reference position in the bowl. Typically, $\boldsymbol{x}_1 = \boldsymbol{0}$, since in the first experiment it is decided if balancing is at all needed. Thus, if balancing is needed, the first experiment is for 'free'. In this example $M \geq 3$ is required and it is chosen to use

$$\begin{bmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \boldsymbol{x}_3 \end{bmatrix} = \begin{bmatrix} 0 & 30 & -30 \\ 0 & 30 & 30 \end{bmatrix}$$

as the trial masses (in grams) for the first three experiments. This is done to ensure that the trial masses do not become 'too' linearly dependent. If $M > 3$, the further experiments are drawn from a statistical distribution

$$\boldsymbol{x}_k = \begin{bmatrix} [\boldsymbol{x}_k]_1 \\ [\boldsymbol{x}_k]_2 \end{bmatrix} = \begin{bmatrix} m_1 \mathrm{e}^{\phi_1 \mathrm{i}} \\ m_2 \mathrm{e}^{\phi_2 \mathrm{i}} \end{bmatrix}, \quad k \geq 4,$$

where

$$m_i \in \{30, 40, 50, 60\}\,[\mathrm{g}], \quad \phi_i \in \mathcal{U}(0, 2\pi)\,[\mathrm{rad}]. \tag{54}$$

All values of m_i are equally probable and $\mathcal{U}(0, 2\pi)$ is a discrete uniform distribution with resolution 1 degree. Not too much effort is put on choosing 'good' candidates for trial masses. Instead, the masses are changed according to (54) for each new Monte Carlo realization. The purpose with this procedure is to diminish the effect of specific choices of $\boldsymbol{x}_k$ and instead put the focus on the performance of the estimators.

Again, Monte Carlo simulations are used to evaluate the performance of the three estimation algorithms. The covariance matrix of the estimates are computed using 300 realizations for each value of M. The result is shown in Figure

5. The figure shows that if the true covariance matrix of the uncertainty is known, the nonlinear estimation method A3 outperforms A1 and A2. Even with the ad hoc choice $\hat{\bar{\boldsymbol{R}}}_{\boldsymbol{\alpha}} = \boldsymbol{I}$, A3 gives better performance compared to A1 and A2. Such a choice is probably natural if the statistics of the uncertainty is completely unknown. In reality, user choices of $\hat{\bar{\boldsymbol{R}}}_{\boldsymbol{\alpha}}$ would probably lead to a performance of A3 that lies somewhere in between the curves marked with squares. Thus, better knowledge about the system at hand is expected to yield better estimates.

Finally, we show a histogram plot of the estimation error for $M = 14$. The error of $[\hat{\bar{\boldsymbol{x}}}_0]_1 = \mathrm{Re}([\hat{\boldsymbol{x}}_0]_1)$ is shown. It can be seen that the estimation error is centered around zero and the distribution is by far most narrow when A3 with $\hat{\bar{\boldsymbol{R}}}_{\boldsymbol{\alpha}} = \bar{\boldsymbol{R}}_{\boldsymbol{\alpha}}$ is employed.

7 Conclusions

An estimation problem that is motivated by the application of unbalance estimation of rotating machinery have been considered. Three different estimation techniques (A1, A2 and A3) are derived and analyzed with respect to their respective statistical properties.The estimation problem is special in the way that the disturbance is entering the system equations. Instead of noisy measurements (ordinary least squares problems) or noisy inputs (errors in variables problems), the main source of uncertainty is here considered to act on the system parameters in a stochastic fashion.

For a simple example, the derived analytical results are compared with Monte Carlo simulations which show very good agreement. The analytical accuracy expressions could be employed as a basis for experiment design, *i.e.* the problem of finding a sequence of $\boldsymbol{x}_k$ that minimizes the estimation error.

An example of unbalance estimation of a separator is also considered. Here, it is shown that the accuracy can be significantly improved if the nonlinear estimation approach A3 is employed. This is particularly so if the number of experiments is increased. In such circumstances, it matters very much how the estimation is performed. The nonlinear approach A3 may then perform considerably much better than the linear estimators A1 and A2.

Acknowledgment

We are grateful to Dr. Lars Hillström at Alfa Laval Machine Dynamics for fruitful discussions and for letting us use the separator model.

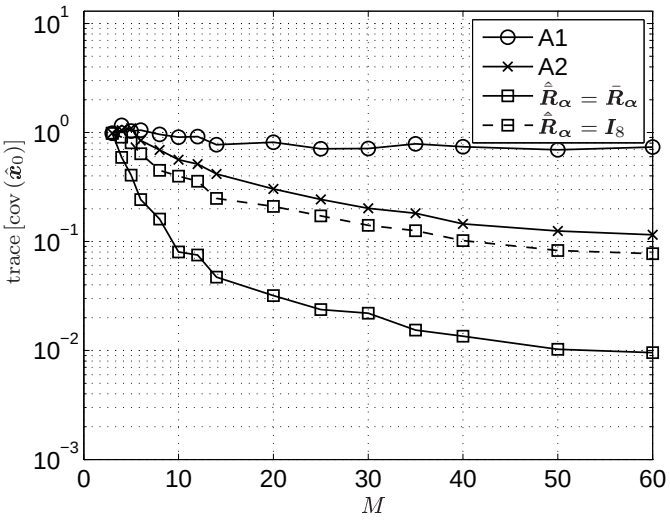

Figure 5: Performance of the different estimators for the separator example.

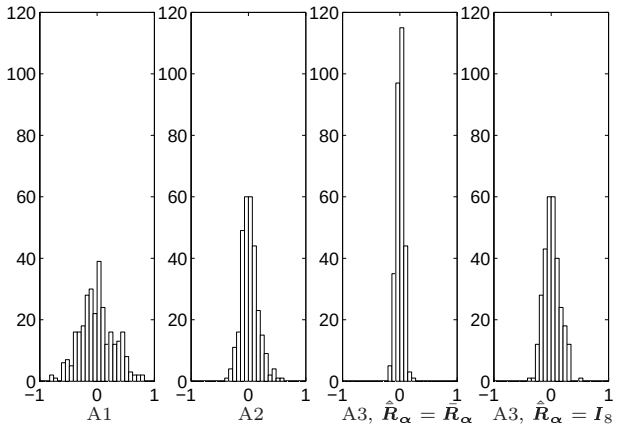

Figure 6: Histogram plot of the estimation error of the real part of $[\boldsymbol{x}_0]_1$. The number of realizations is 300.

References

[1] M. S. Darlow. Balancing of high speed machinery: Theory, methods and experimental results. *Mechanical Systems and Signal Processing*, 1(1):105–134, 1987.

[2] M. S. Darlow. *Balancing of High Speed Machinery*. Springer-Verlag, New York, NY, 1989.

[3] W. C Foiles, P. E. Allaire, and E. J. Gunter. Review: Rotor balancing. *Shock and Vibration*, 5:325–336, 1998.

[4] G. H. Golub and V. Pereyra. The differentiation of pseudo-inverses and nonlinear least squares problems whose variables separate. *Siam Journal of Numerical Analysis*, 10(2):413–432, 1973.

[5] G. H. Golub and V. Pereyra. Separable nonlinear least squares: the variable projection method and its applications. *Inverse Problems*, 19(2):R1–R26, 2003.

[6] T. P. Goodman. A least-squares method for computing balance corrections. *Journal of Engineering for Industry*, 86(3):273–279, 1964.

[7] L. Hillström. *Personal communication*, 2008.

[8] Y. Kang, Y. P. Chang, M. H. Tseng, P. H. Tang, and Y. F. Chang. A modified approach based on influence coefficient method for balancing crankshafts. *Journal of Sound and Vibration*, 234(2):277–296, 200.

[9] L. O. Larsson. On the determination of the influence coefficients in rotor balancing, using linear regression analysis. In *Vibrations in Rotating Machinery*, pages 93–97, Cambridge, UK, September 1976.

[10] L. Ljung. *System Identification*. Prentice–Hall, Upper Saddle River, NJ, USA, 2nd edition, 1999.

[11] J. W. Lund and J. Tonnesen. Analysis and experiments on multi-plane balancing of a flexible rotor. *Journal of Engineering for Industry*, 94(1):233–242, 1972.

[12] K. Mahata. *Estimation Using Low Rank Signal Models*. PhD thesis, Department of Information Technology, Uppsala University, Uppsala, Sweden, 2003.

[13] T. Söderström and P. Stoica. *System Identification*. Prentice Hall International, Hemel Hempstead, United Kingdom, 1989.

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

[15] S. Zhou and J. Shi. Active balancing and vibration control of rotating machinery: A survey. *The Shock and Vibration Digest*, 33(5):361–371, 2001.

A Proofs of Section 4

Proof of Lemma 1

The identification procedure is derived while neglecting the effects of $\tilde{\boldsymbol{A}}_k$. In the presence of this disturbance (6) and (7) modify to

$$\begin{aligned}\boldsymbol{z}_k &= \boldsymbol{A}\boldsymbol{x}_k - \tilde{\boldsymbol{A}}_1\boldsymbol{x}_0 + \tilde{\boldsymbol{A}}_k\left(\boldsymbol{x}_0+\boldsymbol{x}_k\right)\\ &= \left(\boldsymbol{x}_k^T\otimes\boldsymbol{I}_n\right)\boldsymbol{\alpha} - \left(\boldsymbol{x}_0^T\otimes\boldsymbol{I}_n\right)\tilde{\boldsymbol{\alpha}}_1 + \left(\left(\boldsymbol{x}_0+\boldsymbol{x}_k\right)^T\otimes\boldsymbol{I}_n\right)\tilde{\boldsymbol{\alpha}}_k, \quad k=2,\cdots,M.\end{aligned}$$

The $\boldsymbol{z}$ vector (9) then becomes

$$\begin{aligned}\boldsymbol{z} &= \boldsymbol{\Phi}_1\boldsymbol{\alpha} - \begin{bmatrix}1\\ \vdots\\ 1\end{bmatrix}\otimes\left(\boldsymbol{x}_0^T\otimes\boldsymbol{I}_n\right)\tilde{\boldsymbol{\alpha}}_1 + \begin{bmatrix}\left(\boldsymbol{x}_0+\boldsymbol{x}_2\right)^T\otimes\boldsymbol{I}_n & & \boldsymbol{0}\\ & \ddots & \\ \boldsymbol{0} & & \left(\boldsymbol{x}_0+\boldsymbol{x}_M\right)^T\otimes\boldsymbol{I}_n\end{bmatrix}\begin{bmatrix}\tilde{\boldsymbol{\alpha}}_2\\ \vdots\\ \tilde{\boldsymbol{\alpha}}_M\end{bmatrix}\\ &= \boldsymbol{\Phi}_1\boldsymbol{\alpha} + \begin{bmatrix}-\boldsymbol{C}_{1b} & \boldsymbol{C}_{1a}\end{bmatrix}\tilde{\boldsymbol{\alpha}},\end{aligned} \tag{A.1}$$

with $\boldsymbol{C}_{1a}$ and $\boldsymbol{C}_{1b}$ as defined in (14) and (15), respectively.

The first step of the estimation procedure is to compute an estimate of $\boldsymbol{\alpha}$, as in (11)

$$\begin{aligned}\hat{\boldsymbol{\alpha}} &= \boldsymbol{\Phi}_1^{\dagger}\boldsymbol{z}\\ &= \boldsymbol{\alpha} + \boldsymbol{\Phi}_1^{\dagger}\begin{bmatrix}-\boldsymbol{C}_{1b} & \boldsymbol{C}_{1a}\end{bmatrix}\tilde{\boldsymbol{\alpha}}\\ &\triangleq \boldsymbol{\alpha} + \boldsymbol{\alpha}_{\varepsilon},\end{aligned}$$

where

$$\boldsymbol{\alpha}_{\varepsilon} = \boldsymbol{\Phi}_1^{\dagger}\begin{bmatrix}-\boldsymbol{C}_{1b} & \boldsymbol{C}_{1a}\end{bmatrix}\tilde{\boldsymbol{\alpha}}$$

Thus, the estimate of $\boldsymbol{A}$ can be written as

$$\hat{\boldsymbol{A}} = \boldsymbol{A} + \boldsymbol{A}_{\varepsilon}, \tag{A.2}$$

where $\boldsymbol{A}_{\varepsilon}$ is formed from $\boldsymbol{\alpha}_{\varepsilon}$, *i.e.* $\mathrm{vec}(\boldsymbol{A}_{\varepsilon}) = \boldsymbol{\alpha}_{\varepsilon}$.

Let

$$\boldsymbol{m} = \boldsymbol{A}\boldsymbol{x}_0,$$

which means that

$$\hat{\boldsymbol{m}} = \boldsymbol{y}_1 = \boldsymbol{m} + \tilde{\boldsymbol{A}}_1\boldsymbol{x}_0, \tag{A.3}$$

that follows from (1) and (5). Equation (A.3) can be rewritten as

$$\hat{\boldsymbol{m}} = \boldsymbol{m} + \boldsymbol{m}_{\varepsilon}, \tag{A.4}$$

where

$$\boldsymbol{m}_\varepsilon = \left(\boldsymbol{x}_0^T \otimes \boldsymbol{I}_n\right) \tilde{\boldsymbol{\alpha}}_1.$$

Using (12), (A.2) and (A.4), the estimate of $\boldsymbol{x}_0$ can be written as

$$\begin{aligned}
\hat{\boldsymbol{x}}_0 &= \left(\boldsymbol{A} + \boldsymbol{A}_\varepsilon\right)^{-1} \left(\boldsymbol{m} + \boldsymbol{m}_\varepsilon\right) \\
&= \left(\boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{A}_\varepsilon\boldsymbol{A}^{-1} + \cdots\right) \left(\boldsymbol{m} + \boldsymbol{m}_\varepsilon\right) \\
&= \boldsymbol{x}_0 + \boldsymbol{A}^{-1} \left(\boldsymbol{m}_\varepsilon - \boldsymbol{A}_\varepsilon \boldsymbol{x}_0\right) + \cdots \\
&= \boldsymbol{x}_0 + \boldsymbol{A}^{-1} \left(\left(\boldsymbol{x}_0^T \otimes \boldsymbol{I}_n\right) \tilde{\boldsymbol{\alpha}}_1 - \left(\boldsymbol{x}_0^T \otimes \boldsymbol{I}_n\right) \boldsymbol{\alpha}_\varepsilon\right) + \cdots \\
&= \boldsymbol{x}_0 + \boldsymbol{A}^{-1} \left(\boldsymbol{x}_0^T \otimes \boldsymbol{I}_n\right) \left(\tilde{\boldsymbol{\alpha}}_1 - \boldsymbol{\Phi}_1^\dagger \begin{bmatrix} -\boldsymbol{C}_{1b} & \boldsymbol{C}_{1a} \end{bmatrix} \tilde{\boldsymbol{\alpha}}\right) + \cdots \qquad \text{(A.5)} \\
&\approx \boldsymbol{x}_0 + \boldsymbol{A}^{-1} \left(\boldsymbol{x}_0^T \otimes \boldsymbol{I}_n\right) \begin{bmatrix} \left(\boldsymbol{I}_{n^2} + \boldsymbol{\Phi}^\dagger \boldsymbol{C}_{1b}\right) & -\boldsymbol{\Phi}_1^\dagger \boldsymbol{C}_{1a} \end{bmatrix} \tilde{\boldsymbol{\alpha}}, \qquad \text{(A.6)}
\end{aligned}$$

where the approximation in (A.6) follows from the fact that $\|\tilde{\boldsymbol{A}}_k\|$ is assumed to be much smaller than $\|\boldsymbol{A}\|$. Therefore, also $\|\boldsymbol{A}_\varepsilon\|$ is much smaller than $\|\boldsymbol{A}\|$. From (A.5) it is concluded that

$$\mathbf{E}\hat{\boldsymbol{x}}_0 = \boldsymbol{x}_0 + \mathcal{O}\left(\mathbf{E}\|\tilde{\boldsymbol{A}}_k\|^2\right)$$

since the error term in (A.6) is linear in $\tilde{\boldsymbol{\alpha}}$, which has zero mean. For large SNR, (A.6) is a valid approximation. Then, covariance matrix of $\hat{\boldsymbol{x}}_0$ becomes

$$\mathrm{cov}\left(\hat{\boldsymbol{x}}_0\right) = \boldsymbol{A}^{-1}\boldsymbol{C}_1 \mathrm{cov}\left(\tilde{\boldsymbol{\alpha}}\right) \boldsymbol{C}_1^* \boldsymbol{A}^{-*},$$

with $\boldsymbol{C}_1$ given by (13). Furthermore, the covariance matrix of $\tilde{\boldsymbol{\alpha}}$ is given by (3), which concludes the proof. □

Proof of Lemma 2

In the presence of $\tilde{\boldsymbol{A}}_k$, (16) modifies to

$$\boldsymbol{y}_k = \boldsymbol{m} + \left(\boldsymbol{x}_k^T \otimes \boldsymbol{I}_n\right) \boldsymbol{\alpha} + \left(\left(\boldsymbol{x}_0 + \boldsymbol{x}_k\right)^T \otimes \boldsymbol{I}_n\right) \tilde{\boldsymbol{\alpha}}_k.$$

and (17) becomes

$$\boldsymbol{y} = \boldsymbol{\Phi}_2 \boldsymbol{\theta} + \boldsymbol{C}_2 \tilde{\boldsymbol{\alpha}}, \qquad \text{(A.7)}$$

with $\boldsymbol{\Phi}_2$ and $\boldsymbol{C}_2$ as defined in (17) and (22), respectively.

Application of (A.7) on (18) yields

$$\begin{aligned}
\hat{\boldsymbol{\theta}} &= \boldsymbol{\Phi}_2^\dagger \left(\boldsymbol{\Phi}_2 \boldsymbol{\theta} + \boldsymbol{C}_2 \tilde{\boldsymbol{\alpha}}\right) \\
&= \boldsymbol{\theta} + \boldsymbol{\theta}_\varepsilon, \qquad \text{(A.8)}
\end{aligned}$$

where

$$\boldsymbol{\theta}_\varepsilon = \boldsymbol{\Phi}_2^\dagger \boldsymbol{C}_2 \tilde{\boldsymbol{\alpha}}. \qquad \text{(A.9)}$$

Equation (A.8) shows that $\hat{\boldsymbol{\theta}}$ is an unbiased estimate of $\boldsymbol{\theta} = \begin{bmatrix} \boldsymbol{m}^T & \boldsymbol{\alpha}^T \end{bmatrix}^T$ since the error term is linear in $\tilde{\boldsymbol{\alpha}}$. The covariance matrix of the estimate then reads

$$\operatorname{cov}\left(\hat{\boldsymbol{\theta}}\right) = \boldsymbol{\Phi}_2^{\dagger} \boldsymbol{C}_2 \operatorname{cov}\left(\tilde{\boldsymbol{\alpha}}\right) \boldsymbol{C}_2^{*} \boldsymbol{\Phi}_2^{\dagger *},$$

where $\operatorname{cov}\left(\tilde{\boldsymbol{\alpha}}\right) = \boldsymbol{R}_{\boldsymbol{\alpha},M}$.

The partition of $\hat{\boldsymbol{\theta}}$ in (A.8) implies that

$$\begin{aligned} \hat{\boldsymbol{m}} &= \boldsymbol{m} + \boldsymbol{m}_{\varepsilon}, \\ \hat{\boldsymbol{\alpha}} &= \boldsymbol{\alpha} + \boldsymbol{\alpha}_{\varepsilon} \quad \Rightarrow \hat{\boldsymbol{A}} = \boldsymbol{A} + \boldsymbol{A}_{\varepsilon}, \end{aligned}$$

for some $\boldsymbol{m}_{\varepsilon}$ and $\boldsymbol{A}_{\varepsilon}$. Therefore, (19) can be written as

$$\begin{aligned} \hat{\boldsymbol{x}}_0 &= \left(\boldsymbol{A} + \boldsymbol{A}_{\varepsilon}\right)^{-1} \left(\boldsymbol{m} + \boldsymbol{m}_{\varepsilon}\right) \\ &= \left(\boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{A}_{\varepsilon}\boldsymbol{A}^{-1} + \cdots\right) \left(\boldsymbol{m} + \boldsymbol{m}_{\varepsilon}\right) \\ &= \boldsymbol{x}_0 + \boldsymbol{A}^{-1} \left(\boldsymbol{m}_{\varepsilon} - \boldsymbol{A}_{\varepsilon}\boldsymbol{x}_0\right) + \cdots \\ &= \boldsymbol{x}_0 + \boldsymbol{A}^{-1} \left(\boldsymbol{m}_{\varepsilon} - \left(\boldsymbol{x}_0^T \otimes \boldsymbol{I}_n\right) \boldsymbol{\alpha}_{\varepsilon}\right) + \cdots \\ &= \boldsymbol{x}_0 + \boldsymbol{A}^{-1} \left(\begin{bmatrix} \boldsymbol{I}_n & -\boldsymbol{x}_0^T \otimes \boldsymbol{I}_n \end{bmatrix} \begin{bmatrix} \boldsymbol{m}_{\varepsilon} \\ \boldsymbol{\alpha}_{\varepsilon} \end{bmatrix} \right) + \cdots \\ &= \boldsymbol{x}_0 + \boldsymbol{A}^{-1} \boldsymbol{C}_{2a} \boldsymbol{\theta}_{\varepsilon} + \cdots \\ &\approx \boldsymbol{x}_0 + \boldsymbol{A}^{-1} \boldsymbol{C}_{2a} \boldsymbol{\Phi}_2^{\dagger} \boldsymbol{C}_2 \tilde{\boldsymbol{\alpha}}, \end{aligned} \tag{A.10}$$

where in the last equality (A.9) has been used and the approximation follows from that the second order terms are assumed to be negligible. Therefore, the expectation of the estimate (19) can be written as (20). From (A.10) it directly follows that for large SNR, the covariance matrix of $\hat{\boldsymbol{x}}_0$ is given by (21). □

B Proofs of Section 5

Proof of Proposition 1

The result directly follows from the definition (41) of $\mathbf{P}$. □

Proof of Proposition 2

Application of the chain rule and the rule for differentiation of matrix inverses yields

$$\begin{aligned}
\boldsymbol{P}^{(k)} &= \left\{ \left[\boldsymbol{B}^T \boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1} \boldsymbol{B}\right]^{-1} \boldsymbol{B}^T \boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1} \right\}^{(k)} \\
&= \left[\boldsymbol{B}^T \boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1} \boldsymbol{B}\right]^{-(k)} \boldsymbol{B}^T \boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1} + \left[\boldsymbol{B}^T \boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1} \boldsymbol{B}\right]^{-1} \left(\boldsymbol{B}^{T(k)} \boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1} + \boldsymbol{B}^T \boldsymbol{Q}_{\boldsymbol{\alpha}}^{-(k)} \right) \\
&= - \left[\boldsymbol{B}^T \boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1} \boldsymbol{B}\right]^{-1} \left(\boldsymbol{B}^{T(k)} \boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1} \boldsymbol{B} + \boldsymbol{B}^T \boldsymbol{Q}_{\boldsymbol{\alpha}}^{-(k)} \boldsymbol{B} + \boldsymbol{B}^T \boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1} \boldsymbol{B}^{(k)} \right) \\
&\quad \times \underbrace{\left[\boldsymbol{B}^T \boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1} \boldsymbol{B}\right]^{-1} \boldsymbol{B}^T \boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1}}_{\boldsymbol{P}} + \left[\boldsymbol{B}^T \boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1} \boldsymbol{B}\right]^{-1} \left(\boldsymbol{B}^{T(k)} \boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1} + \boldsymbol{B}^T \boldsymbol{Q}_{\boldsymbol{\alpha}}^{-(k)} \right) \\
&= \left[\boldsymbol{B}^T \boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1} \boldsymbol{B}\right]^{-1} \left(\boldsymbol{B}^{T(k)} \boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1} + \boldsymbol{B}^T \boldsymbol{Q}_{\boldsymbol{\alpha}}^{-(k)} \right) (\boldsymbol{I}_{2nM} - \boldsymbol{B}\boldsymbol{P}) - \boldsymbol{P}\boldsymbol{B}^{(k)}\boldsymbol{P}.
\end{aligned}$$

□

Proof of Proposition 3

Application of the chain rule on (42) yields

$$\boldsymbol{P}^{(k)}\boldsymbol{B} + \boldsymbol{P}\boldsymbol{B}^{(k)} = \boldsymbol{0} \quad \Leftrightarrow \quad \boldsymbol{P}^{(k)}\boldsymbol{B} = -\boldsymbol{P}\boldsymbol{B}^{(k)},$$

which is the desired result. □

Proof of Proposition 4

First notice that $\boldsymbol{B}^{(kl)} = \boldsymbol{0}$. Next, differentiate (43) with respect to $[\bar{\boldsymbol{x}}]_l$, and make use of the chain rule

$$\begin{aligned}\boldsymbol{P}^{(kl)} = &\left\{ \left[\boldsymbol{B}^T\boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1}\boldsymbol{B}\right]^{-1} \left(\boldsymbol{B}^{T(k)}\boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1} + \boldsymbol{B}^T\boldsymbol{Q}_{\boldsymbol{\alpha}}^{-(k)}\right)\right\}^{(l)} (\boldsymbol{I}_{2nM} - \boldsymbol{B}\boldsymbol{P}) \\ &+ \left[\boldsymbol{B}^T\boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1}\boldsymbol{B}\right]^{-1} \left(\boldsymbol{B}^{T(k)}\boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1} + \boldsymbol{B}^T\boldsymbol{Q}_{\boldsymbol{\alpha}}^{-(k)}\right) \left(-\boldsymbol{B}^{(l)}\boldsymbol{P} - \boldsymbol{B}\boldsymbol{P}^{(l)}\right) \\ &- \boldsymbol{P}^{(l)}\boldsymbol{B}^{(k)}\boldsymbol{P} - \boldsymbol{P}\boldsymbol{B}^{(k)}\boldsymbol{P}^{(l)}.\end{aligned}$$

Using Proposition 1 and Proposition 3 we obtain

$$\begin{aligned}\left[\boldsymbol{B}^T\boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1}\boldsymbol{B}\right]\boldsymbol{P}^{(kl)}\boldsymbol{B} &= -\left(\boldsymbol{B}^{T(k)}\boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1} + \boldsymbol{B}^T\boldsymbol{Q}_{\boldsymbol{\alpha}}^{-(k)}\right)\left(\boldsymbol{B}^{(l)}\boldsymbol{P} + \boldsymbol{B}\boldsymbol{P}^{(l)}\right)\boldsymbol{B} \\ &\quad - \boldsymbol{B}^T\boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1}\boldsymbol{B}\boldsymbol{P}^{(l)}\boldsymbol{B}^{(k)}\boldsymbol{P}\boldsymbol{B} - \underbrace{\boldsymbol{B}^T\boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1}\boldsymbol{B}\boldsymbol{P}}_{\boldsymbol{B}^T\boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1}}\boldsymbol{B}^{(k)}\boldsymbol{P}^{(l)}\boldsymbol{B} \\ &= -\left(\boldsymbol{B}^{T(k)}\boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1} + \boldsymbol{B}^T\boldsymbol{Q}_{\boldsymbol{\alpha}}^{-(k)}\right)(\boldsymbol{I}_{2nM} - \boldsymbol{B}\boldsymbol{P})\,\boldsymbol{B}^{(l)} \\ &\quad - \boldsymbol{B}^T\boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1}\boldsymbol{B}\boldsymbol{P}^{(l)}\boldsymbol{B}^{(k)} + \boldsymbol{B}^T\boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1}\boldsymbol{B}^{(k)}\boldsymbol{P}\boldsymbol{B}^{(l)} \qquad \text{(B.1)}\end{aligned}$$

and using Proposition 2, Proposition 3 and some algebraic manipulations, the term that involves $\boldsymbol{P}^{(l)}$ is expanded

$$\begin{aligned}&\boldsymbol{B}^T\boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1}\boldsymbol{B}\boldsymbol{P}^{(l)}\boldsymbol{B}^{(k)} = \cdots = \\ &= \left(\boldsymbol{B}^{T(l)}\boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1} + \boldsymbol{B}^T\boldsymbol{Q}_{\boldsymbol{\alpha}}^{-(l)}\right)(\boldsymbol{I}_{2nM} - \boldsymbol{B}\boldsymbol{P})\,\boldsymbol{B}^{(k)} - \boldsymbol{B}^T\boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1}\boldsymbol{B}^{(l)}\boldsymbol{P}\boldsymbol{B}^{(k)}. \qquad \text{(B.2)}\end{aligned}$$

Finally, combining (B.1) and (B.2) gives the desired result. □

Proof of Lemma 3

Differentiation of (40) yields

$$W^{(k)} = \bar{\boldsymbol{y}}^T\boldsymbol{Q}^{(k)}\bar{\boldsymbol{y}}, \qquad \text{(B.3)}$$

where

$$\boldsymbol{Q}^{(k)} = \boldsymbol{Q}_{\boldsymbol{\alpha}}^{-(k)}(\boldsymbol{I}_{2nM} - \boldsymbol{B}\boldsymbol{P}) - \boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1}\left(\boldsymbol{B}^{(k)}\boldsymbol{P} + \boldsymbol{B}\boldsymbol{P}^{(k)}\right). \qquad \text{(B.4)}$$

Let $\bar{\boldsymbol{y}} = \boldsymbol{B}(\bar{\boldsymbol{x}}_0)\bar{\boldsymbol{\alpha}} + \boldsymbol{C}(\bar{\boldsymbol{x}}_0)\bar{\tilde{\boldsymbol{\alpha}}}$ as in (26) and evaluate (B.3) at $\bar{\boldsymbol{x}} = \bar{\boldsymbol{x}}_0$. This gives

$$W^{(k)}\Big|_{\bar{\boldsymbol{x}}=\bar{\boldsymbol{x}}_0} = \bar{\boldsymbol{\alpha}}^T\boldsymbol{B}^T\boldsymbol{Q}^{(k)}\boldsymbol{B}\bar{\boldsymbol{\alpha}} + 2\bar{\boldsymbol{\alpha}}^T\boldsymbol{B}^T\boldsymbol{Q}^{(k)}\boldsymbol{C}\bar{\tilde{\boldsymbol{\alpha}}} + \bar{\tilde{\boldsymbol{\alpha}}}^T\boldsymbol{C}^T\boldsymbol{Q}^{(k)}\boldsymbol{C}\bar{\tilde{\boldsymbol{\alpha}}}. \qquad \text{(B.5)}$$

By use of Proposition 1 and Proposition 3 it follows that $\boldsymbol{B}^T\boldsymbol{Q}^{(k)}\boldsymbol{B} = \boldsymbol{0}$, so the first term vanishes. Next, it is argued that if $\|\tilde{\boldsymbol{A}}_k\| \ll \|\boldsymbol{A}\|$, then the term $\bar{\tilde{\boldsymbol{\alpha}}}^T\boldsymbol{C}^T\boldsymbol{Q}^{(k)}\boldsymbol{C}\bar{\tilde{\boldsymbol{\alpha}}}$ is negligible compared to the middle term of (B.5). It remains to compute

$$2\bar{\boldsymbol{\alpha}}^T\boldsymbol{B}^T\boldsymbol{Q}^{(k)}\boldsymbol{C}\bar{\tilde{\boldsymbol{\alpha}}} = 2\bar{\tilde{\boldsymbol{\alpha}}}^T\boldsymbol{C}^T\boldsymbol{Q}^{(k)}\boldsymbol{B}\bar{\boldsymbol{\alpha}}.$$

Again, Proposition 1 and Proposition 3 give

$$W^{(k)}(\bar{\boldsymbol{x}}_0) \approx -2\bar{\tilde{\boldsymbol{\alpha}}}^T \boldsymbol{C}^T \boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1} \left(\boldsymbol{I}_{2nM} - \boldsymbol{B}\boldsymbol{P}\right) \boldsymbol{B}^{(k)} \bar{\boldsymbol{\alpha}}, \tag{B.6}$$

which can be equivalently written as (46).

Next, we want to find an expression for $W^{(kl)}$. Differentiation of (B.3) with respect to $[\bar{\boldsymbol{x}}]_l$ yields

$$W^{(kl)} = \bar{\boldsymbol{y}}^T \boldsymbol{Q}^{(kl)} \bar{\boldsymbol{y}},$$

where $\boldsymbol{Q}^{(k)}$ is given by (B.4). If the model (26) for $\bar{\boldsymbol{y}}$ is inserted, one obtains

$$\begin{aligned} W^{(kl)} &= \bar{\boldsymbol{\alpha}}^T \boldsymbol{B}^T \boldsymbol{Q}^{(kl)} \boldsymbol{B} \bar{\boldsymbol{\alpha}} + 2\bar{\boldsymbol{\alpha}}^T \boldsymbol{B}^T \boldsymbol{Q}^{(kl)} \boldsymbol{C} \bar{\tilde{\boldsymbol{\alpha}}} + \bar{\tilde{\boldsymbol{\alpha}}}^T \boldsymbol{C}^T \boldsymbol{Q}^{(kl)} \boldsymbol{C} \bar{\tilde{\boldsymbol{\alpha}}} \\ &\approx \bar{\boldsymbol{\alpha}}^T \boldsymbol{B}^T \boldsymbol{Q}^{(kl)} \boldsymbol{B} \bar{\boldsymbol{\alpha}}. \end{aligned} \tag{B.7}$$

The approximation follows from that the term that is quadratic in $\bar{\boldsymbol{\alpha}}$ is nonzero, and the assumption $\|\tilde{\boldsymbol{A}}_k\| \ll \|\boldsymbol{A}\|$.

Differentiation of (B.4)with respect to $[\bar{\boldsymbol{x}}]_l$ yields

$$\begin{aligned} \boldsymbol{Q}^{(kl)} = \boldsymbol{Q}_{\boldsymbol{\alpha}}^{-(kl)}(\boldsymbol{I}_{2nM} - \boldsymbol{B}\boldsymbol{P}) - \boldsymbol{Q}_{\boldsymbol{\alpha}}^{-(k)} \left(\boldsymbol{B}^{(l)}\boldsymbol{P} + \boldsymbol{B}\boldsymbol{P}^{(l)}\right) - \boldsymbol{Q}_{\boldsymbol{\alpha}}^{-(l)} \left(\boldsymbol{B}^{(k)}\boldsymbol{P} + \boldsymbol{B}\boldsymbol{P}^{(k)}\right) \\ - \boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1} \left(\boldsymbol{B}^{(k)}\boldsymbol{P}^{(l)} + \boldsymbol{B}^{(l)}\boldsymbol{P}^{(k)} + \boldsymbol{B}\boldsymbol{P}^{(kl)}\right). \end{aligned}$$

Computation of (B.7) with application of Propositions 1–4, $\boldsymbol{B}^{(kl)} = \boldsymbol{0}$, and evaluation at $\bar{\boldsymbol{x}} = \bar{\boldsymbol{x}}_0$ yields

$$\begin{aligned} W^{(kl)} &\approx \bar{\boldsymbol{\alpha}}^T \boldsymbol{B}^T \boldsymbol{Q}^{(kl)} \boldsymbol{B} \bar{\boldsymbol{\alpha}} \\ &= \bar{\boldsymbol{\alpha}}^T \Big(\boldsymbol{0} - \boldsymbol{B}^T \boldsymbol{Q}_{\boldsymbol{\alpha}}^{-(k)} \left(\boldsymbol{B}^{(l)} - \boldsymbol{B}\boldsymbol{P}\boldsymbol{B}^{(l)}\right) - \boldsymbol{B}^T \boldsymbol{Q}_{\boldsymbol{\alpha}}^{-(l)} \left(\boldsymbol{B}^{(k)} - \boldsymbol{B}\boldsymbol{P}\boldsymbol{B}^{(k)}\right) \\ &\quad - \boldsymbol{B}^T \boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1} \left(-\boldsymbol{B}^{(k)}\boldsymbol{P}\boldsymbol{B}^{(l)} - \boldsymbol{B}^{(l)}\boldsymbol{P}\boldsymbol{B}^{(k)}\right) - \boldsymbol{B}^T \boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1} \boldsymbol{B}\boldsymbol{P}^{(kl)}\boldsymbol{B} \Big) \bar{\boldsymbol{\alpha}} \\ &= \bar{\boldsymbol{\alpha}}^T \Big(\boldsymbol{B}^{T(k)} \boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1} (\boldsymbol{I}_{2nM} - \boldsymbol{B}\boldsymbol{P}) \boldsymbol{B}^{(l)} + \boldsymbol{B}^{T(l)} \boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1} (\boldsymbol{I}_{2nM} - \boldsymbol{B}\boldsymbol{P}) \boldsymbol{B}^{(k)} \Big) \bar{\boldsymbol{\alpha}} \qquad \text{(B.8)} \\ &= 2\bar{\boldsymbol{\alpha}}^T \boldsymbol{B}^{T(k)} \boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1} \left(\boldsymbol{I}_{2nM} - \boldsymbol{B}\boldsymbol{P}\right) \boldsymbol{B}^{(l)} \bar{\boldsymbol{\alpha}}, \qquad \text{(B.9)} \end{aligned}$$

where (B.8) follows from Proposition 4 and some algebra. Equation (B.9) follows since $\boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1} (\boldsymbol{I}_{2nM} - \boldsymbol{B}\boldsymbol{P})$ is a symmetric matrix. The expression (B.9) can be equivalently written as (47). □

Proof of Lemma 4

The estimation error is given by (37), which by use of Lemma 3 can be written as

$$\bar{\boldsymbol{x}}_\varepsilon \approx -\boldsymbol{\mathcal{H}}^{-1} \boldsymbol{G} \bar{\tilde{\boldsymbol{\alpha}}}, \tag{B.10}$$

which has zero expectation for large SNR, since $\boldsymbol{\mathcal{H}}$ and $\boldsymbol{G}$ are constant matrices. Using (B.5) and (B.6), the expecation of $\hat{\bar{\boldsymbol{x}}}_0$ can be written as (48). Furthermore, using (B.10) the covariance matrix of $\bar{\boldsymbol{x}}_\varepsilon$ satisfy $\operatorname{cov}(\bar{\boldsymbol{x}}_\varepsilon) = \operatorname{cov}\left(\hat{\bar{\boldsymbol{x}}}_0\right)$. Then, (49) immediately follows from (38) and Lemma 3. □

C Expressions for the Derivatives of $\boldsymbol{B}$ and $\boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1}$

Differentiation of $\boldsymbol{B}(\bar{\boldsymbol{x}})$ and $\boldsymbol{C}(\bar{\boldsymbol{x}})$

The derivatives with respect to $[\bar{\boldsymbol{x}}]_k$ of the matrices $\boldsymbol{B}$ and $\boldsymbol{C}$ as defined in (27) become

$$\boldsymbol{B}^{(k)} = \begin{bmatrix} \frac{\partial \boldsymbol{D}_1(\bar{\boldsymbol{x}})}{\partial [\bar{\boldsymbol{x}}]_k} \\ \vdots \\ \frac{\partial \boldsymbol{D}_M(\bar{\boldsymbol{x}})}{\partial [\bar{\boldsymbol{x}}]_k} \end{bmatrix}, \quad \boldsymbol{C}^{(k)} = \begin{bmatrix} \frac{\partial \boldsymbol{D}_1(\bar{\boldsymbol{x}})}{\partial [\bar{\boldsymbol{x}}]_k} & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \frac{\partial \boldsymbol{D}_2(\bar{\boldsymbol{x}})}{\partial [\bar{\boldsymbol{x}}]_k} & \cdots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \frac{\partial \boldsymbol{D}_M(\bar{\boldsymbol{x}})}{\partial [\bar{\boldsymbol{x}}]_k} \end{bmatrix},$$

where $\boldsymbol{D}_i(\bar{\boldsymbol{x}})$ is given by (25). For brevity it is repeated

$$\boldsymbol{D}_i(\bar{\boldsymbol{x}}) = \begin{bmatrix} (\boldsymbol{x}_R + \boldsymbol{x}_{iR})^T \otimes \boldsymbol{I}_n & -(\boldsymbol{x}_I + \boldsymbol{x}_{iI})^T \otimes \boldsymbol{I}_n \\ (\boldsymbol{x}_I + \boldsymbol{x}_{iI})^T \otimes \boldsymbol{I}_n & (\boldsymbol{x}_R + \boldsymbol{x}_{iR})^T \otimes \boldsymbol{I}_n \end{bmatrix}, \quad \bar{\boldsymbol{x}} = \begin{bmatrix} \boldsymbol{x}_R \\ \boldsymbol{x}_I \end{bmatrix} \begin{matrix} \}n \\ \}n \end{matrix} \tag{C.1}$$

and we see that its derivative differs somewhat depending on whether $[\bar{\boldsymbol{x}}]_k$ belongs to $\boldsymbol{x}_R$ or $\boldsymbol{x}_I$. Thus if $k \leq n$, then $[\bar{\boldsymbol{x}}]_k \in \boldsymbol{x}_R$; and if $k > n$, then $[\bar{\boldsymbol{x}}]_k \in \boldsymbol{x}_I$. The derivative of $\boldsymbol{D}_i$ then becomes

$$\begin{aligned} \boldsymbol{D}_i^{(k)} &= \frac{\partial \boldsymbol{D}_i(\bar{\boldsymbol{x}})}{\partial [\bar{\boldsymbol{x}}]_k} \\ &= \begin{cases} \begin{bmatrix} 0 & \cdots & 1\,(\text{pos. } k) & \cdots \\ 0 & \cdots & 0 & \cdots \end{bmatrix} \begin{bmatrix} \cdots & 0 & \cdots & 0 \\ \cdots & 1(\text{pos. } n+k) & \cdots & 0 \end{bmatrix} \otimes \boldsymbol{I}_n & \text{if } k \leq n, \\ \begin{bmatrix} 0 & \cdots & 0 & \cdots \\ 0 & \cdots & 1(\text{pos. } k) & \cdots \end{bmatrix} -\begin{bmatrix} \cdots & 1(\text{pos. } n+k) & \cdots & 0 \\ \cdots & 0 & \cdots & 0 \end{bmatrix} \otimes \boldsymbol{I}_n & \text{if } k > n, \end{cases} \end{aligned} \tag{C.2}$$

which is a constant matrix that is independent on the input variable $\boldsymbol{x}_i$, for $i = 1, \cdots, M$. The final expressions of the derivatives of $\boldsymbol{B}$ and $\boldsymbol{C}$ become

$$\boldsymbol{B}^{(k)} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \otimes \boldsymbol{D}^{(k)}, \quad \boldsymbol{C}^{(k)} = \boldsymbol{I}_M \otimes \boldsymbol{D}^{(k)}. \tag{C.3}$$

□

Differentiation of $\boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1}$

The weighting matrix $\boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1}$ was defined in (28). It is here repeated

$$\boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1}(\bar{\boldsymbol{x}}) = \left(\boldsymbol{C}(\bar{\boldsymbol{x}}) \bar{\boldsymbol{R}}_{\boldsymbol{\alpha},M} \boldsymbol{C}^T(\bar{\boldsymbol{x}})\right)^{-1}$$

and differentiation with respect to $[\bar{\boldsymbol{x}}]_k$ yields

$$\begin{aligned}\boldsymbol{Q}_{\boldsymbol{\alpha}}^{-(k)} &= -\boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1}\boldsymbol{Q}_{\boldsymbol{\alpha}}^{(k)}\boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1} \\ &= -\boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1}\left(\boldsymbol{C}^{(k)}\bar{\boldsymbol{R}}_{\boldsymbol{\alpha},M}\boldsymbol{C}^{T} + \boldsymbol{C}\bar{\boldsymbol{R}}_{\boldsymbol{\alpha},M}\boldsymbol{C}^{(k)T}\right)\boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1} \\ &= -\boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1}\left(\left(\boldsymbol{I}_M \otimes \boldsymbol{D}^{(k)}\right)\bar{\boldsymbol{R}}_{\boldsymbol{\alpha},M}\boldsymbol{C}^{T} + \boldsymbol{C}\bar{\boldsymbol{R}}_{\boldsymbol{\alpha},M}\left(\boldsymbol{I}_M \otimes \boldsymbol{D}^{(k)}\right)^{T}\right)\boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1},\end{aligned}$$

where $\boldsymbol{D}^{(k)}$ is given by (C.2). □

D The Gradient of the Loss Function

The gradient of the loss function (32) is given by

$$\frac{\partial^T W}{\partial \bar{\boldsymbol{x}}} = \begin{bmatrix} \bar{\boldsymbol{y}}^T \boldsymbol{Q}^{(1)} \bar{\boldsymbol{y}} \\ \vdots \\ \bar{\boldsymbol{y}}^T \boldsymbol{Q}^{(2n)} \bar{\boldsymbol{y}} \end{bmatrix},$$

where

$$\begin{aligned} \boldsymbol{Q}^{(k)} = &- \left(\boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1} \boldsymbol{Q}_{\boldsymbol{\alpha}}^{(k)} + \boldsymbol{P}^T \left(\boldsymbol{B}^{(k)T} - \boldsymbol{B}^T \boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1} \boldsymbol{Q}_{\boldsymbol{\alpha}}^{(k)} \right) \right) \boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1/2} \boldsymbol{\Pi}^{\perp} \boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1/2} \\ &- \boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1/2} \boldsymbol{\Pi}^{\perp} \boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1/2} \boldsymbol{B}^{(k)} \boldsymbol{P}, \end{aligned} \tag{D.1}$$

$$\boldsymbol{Q}_{\boldsymbol{\alpha}}^{(k)} = \boldsymbol{C}^{(k)} \bar{\boldsymbol{R}}_{\boldsymbol{\alpha},M} \boldsymbol{C}^T + \boldsymbol{C} \bar{\boldsymbol{R}}_{\boldsymbol{\alpha},M} \boldsymbol{C}^{(k)T} \tag{D.2}$$

and where $\boldsymbol{B}^{(k)}$ and $\boldsymbol{C}^{(k)}$ are given by (C.3).

Proof First, notice that

$$W = \bar{\boldsymbol{y}}^T \boldsymbol{Q} \bar{\boldsymbol{y}} = \bar{\boldsymbol{y}}^T \boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1} \left(\boldsymbol{I}_{2nM} - \boldsymbol{B}\boldsymbol{P} \right) \bar{\boldsymbol{y}}$$

and differentiate $\boldsymbol{Q}$ with respect to $[\bar{\boldsymbol{x}}]_k$,

$$\boldsymbol{Q}^{(k)} = \boldsymbol{Q}_{\boldsymbol{\alpha}}^{-(k)} \left(\boldsymbol{I}_{2nM} - \boldsymbol{B}\boldsymbol{P} \right) - \boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1} \left(\boldsymbol{B}^{(k)} \boldsymbol{P} + \boldsymbol{B} \boldsymbol{P}^{(k)} \right).$$

By use of Proposition 2 and $\boldsymbol{Q}_{\boldsymbol{\alpha}}^{-(k)} = -\boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1} \boldsymbol{Q}_{\boldsymbol{\alpha}}^{(k)} \boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1}$ one obtains

$$\begin{aligned} \boldsymbol{Q}^{(k)} = &- \left(\boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1} \boldsymbol{Q}_{\boldsymbol{\alpha}}^{(k)} + \boldsymbol{P}^T \left(\boldsymbol{B}^{(k)T} - \boldsymbol{B}^T \boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1} \boldsymbol{Q}_{\boldsymbol{\alpha}}^{(k)} \right) \right) \boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1} \left(\boldsymbol{I}_{2nM} - \boldsymbol{B}\boldsymbol{P} \right) \\ &- \boldsymbol{Q}_{\boldsymbol{\alpha}}^{-1} \left(\boldsymbol{I}_{2nM} - \boldsymbol{B}\boldsymbol{P} \right) \boldsymbol{B}^{(k)} \boldsymbol{P}, \end{aligned}$$

which is equal to (D.1). Equation (D.2) directly follows from differentiation of $\boldsymbol{Q}_{\boldsymbol{\alpha}} = \boldsymbol{C} \bar{\boldsymbol{R}}_{\boldsymbol{\alpha},M} \boldsymbol{C}^T$. □