Why are errors-in-variables problems often tricky?

Torsten Söderström

Systems and Control, Department of Information Technology,
Uppsala University, P O Box 337, SE-75105 Uppsala, Sweden
email: ts@syscon.uu.se

version November 22, 2002

Abstract

There are several identification methods designed for the errors-in-variables problem. In this paper we focus on discussing the underlying assumptions for several of these methods. Some assumptions are shown to have far reaching consequences. For example, if the noise-free input happens to be periodic, simple estimators that give consistent parameter estimates of the system parameters can easily be designed. If the variances of the input and output noises are unknown, some structural assumption must be added for the system dynamics to be identifiable. On the other hand, should the ratio between output noise variance and input noise variance be known, it is possible not only to estimate the system parameters consistently, but also to combine this with a reasonable estimate of the unperturbed input.

1 Introduction

Many different solutions have been presented for system identification of linear dynamic systems from noise–corrupted output measurements see, for example, [16], [22]. On the other hand, the estimation of the parameters for linear dynamic systems when also the input is affected by noise is recognized as a more difficult problem. Representations where errors or measurement noises are present on both the inputs and outputs are usually called "errors–in–variables" models and play an important role when the identification purpose is the determination of the inner laws that describe the process, rather than the prediction of its future behavior.

The class of scientific disciplines which make use of such a kind of representations is very broad, as proved by the several applications collected in [28], [29], such as time series modelling, array signal processing for direction–of–arrival estimation, blind channel equalization, multivariate calibration in analytical chemistry, image processing, astronomical data reduction, etc. In case of static systems, errors–in–variables representations are closely related to other well–known topics such as *latent variables* models and *factor* models [9], [31].

With reference to these systems, the assumptions (*prejudices*) which lie behind the identification procedure have been analyzed in [12], [13] with particular attention to the Frisch scheme [8]. The extension to the dynamic case has been considered in the literature by several authors with different approaches; in fact the problem can be examined from different point of views. For example, we can distinguish between time–domain [16], [22], and frequency-domain methods [18] approaches, exploiting deterministic signals (e.g. periodic inputs [7]) or stochastic processes, methods that make use of second–order statistics only or based on higher order cumulant statistics [26], [27]. When only second–order statistics are exploited the identification of errors–in–variables models cannot, in general, admit a single solution [1], [2], [3].

Some comparison between different approaches for errors-in-variables modelling are given in [20], [21] and [23].

The focus here is on the role of the assumptions. In 'classical' identification problems where the input signal is measured without errors, some assumptions on the noise properties and the input properties have often not any impact on the consistency on the parameter estimates, but on their accuracy. On the other hand, in the errors-in-variables context, the assumptions can often have far reaching consequences.

The paper is organized as follows. The next section presents the basic setup, while the principal identifiability problem is dealt with in Section 3. There it is shown that it is not possible to get identifiability from one single experiment without imposing some (structural) assumptions, that cannot be verified from the data. In Section 4 we describe bias-compensated versions of the least-squares method, while other estimators based on repeated and correlated experiments are treated in Section 5. It can be appealing to let an identification scheme include estimation of the unknown unperturbed input. This idea is discussed in Section 6, where we show that it is feasible only if we have some appropriate *a priori* information about the noise variances. The relation between the total least-squares estimator and a maximum likelihood estimator, both treating all values of the unperturbed input as additional unknowns are described in Section 7. Section 8 presents an attempt to model the noise effects by fitting an ARMAX model to the noisy input-output data, and offers an explanation why this is not fruitful. We show in Section 9, how the system can be identified when a parametric model of the unperturbed input is applied, and summarizing conclusions are given in Section 10. Several of the sections are ended by summarizing different results, where the role of the key assumptions is highlighted.

2 Basic setup and assumptions

As a typical model example, consider the following system depicted in Figure 1 with noise-corrupted input and output measurements.

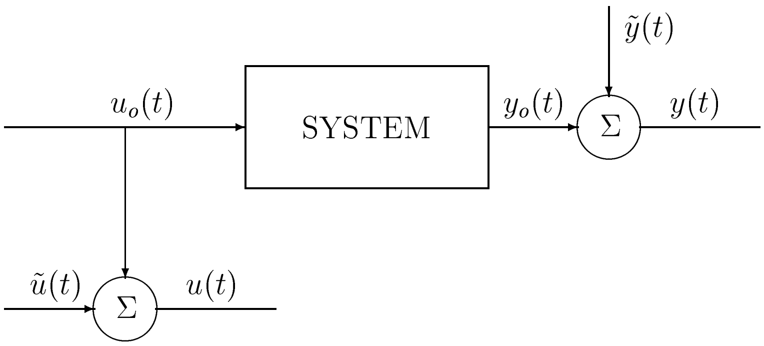

Figure 1: The basic setup for a dynamic error-in-variables problem.

The noise–free input is denoted by $u_o(t)$ and the undisturbed output by $y_o(t)$. They are linked through the linear difference equation

$$A(q^{-1})\, y_o(t) = B(q^{-1})\, u_o(t), \tag{2.1}$$

where $A(q^{-1})$ and $B(q^{-1})$ are polynomials of the type

$$\begin{aligned} A(q^{-1}) &= 1 + a_1\, q^{-1} + \ldots + a_n\, q^{-n} \\ B(q^{-1}) &= b_1\, q^{-1} + \ldots + b_n\, q^{-n} \end{aligned} \tag{2.2}$$

and q^{-1} is the backward shift operator, *i.e.* $q^{-1}\, x(t) = x(t-1)$. It is not restrictive to assume that the polynomials $A(q^{-1})$, $B(q^{-1})$ have equal degree n, which represents the order of the system.

In an errors–in–variables environment we assume that the observations are corrupted by additive measurement noises $\tilde{u}(t)$ and $\tilde{y}(t)$, at the input and output respectively. Therefore, the available signals are of the form

$$\begin{aligned} u(t) &= u_o(t) + \tilde{u}(t) \\ y(t) &= y_o(t) + \tilde{y}(t) \end{aligned} \; . \tag{2.3}$$

The general problem to be treated here, is the following.

Problem. Let the available measurements $u(t)$ and $y(t)$, for $t = 1, \ldots, N$, be generated in accordance with the previous assumptions. Determine the system characteristics, *i.e.* the transfer function

$$G(q^{-1}) = \frac{B(q^{-1})}{A(q^{-1})}. \tag{2.4}$$

■

In order to proceed, a number of assumptions must be introduced. To some degree they can be made more or less restrictive. In the coming subsections we therefore introduce a number of assumptions, that are partly alternative ones. We will then discuss what assumptions that are necessary for different results, and what the consequences will be if the assumptions are relaxed.

2.1 Assumptions on the system

A1a. The dynamic system (1) is asymptotically stable, *i.e.* $A(z)$ has all zeros outside the unit circle.

A1b. All the system modes are observable and controllable, *i.e.* $A(z)$ and $B(z)$ have no common factors.

A1c. The order n of the system is *a priori* known.

Out of the above assumptions, **A1c** is the hardest in practice. It implies that we do not treat the undermodelling case.

2.2 Assumptions on the measurement noise

A2a. The processes $\tilde{u}(t)$ and $\tilde{y}(t)$ are mutually uncorrelated, and uncorrelated with the noise–free signals $u_o(t)$ and $y_o(t)$.

A2b. The sequences $\tilde{u}(t)$ and $\tilde{y}(t)$ are zero–mean white noises with variances λ_u and λ_y, respectively.

The assumption **A2b** may be regarded as a bit restrictive. If $\tilde{y}(t)$ should model not only measurement noise, but also account for process noise and modelling errors, then it is likely correlated and *not* white.

We can have different situations concerning to what degree the noise variances are known.

A2c The noise variances λ_u and λ_y are unknown.

A2d The noise variances λ_u and λ_y are both known.

A2e The ratio λ_y/λ_u is known.

Obviously, the condition **A2c** is more general (and more realistic) than **A2e**. It will turn out in the coming analysis, that when **A2e** is implied, much stronger results will hold than under **A2c**. Of course, assumption **A2d** is still more restrictive than **A2e**.

It is sometimes possible to allow for weaker assumptions. The measurement noises may be somewhat correlated. Some estimation methods can be extended to handle that $\tilde{u}(t)$ is a moving average process, $\tilde{y}(t)$ is an ARMA process, or that $\tilde{u}(t)$ and $\tilde{y}(t)$ are correlated.

2.3 Assumptions on the true input

A3a. The true input $u_o(t)$ is a zero–mean stationary ergodic random signal.

We may also have

A3b. The spectral density of $u_o(t)$ is a rational function. Note that this implies that $u_o(t)$ can be modelled as an ARMA process.

For some identification methods, see Section 9, but certainly not all, the assumptions **A3a** and **A3b** have to be imposed.

2.4 Assumptions on the experiment

A4a. One experimental data set is available.

A4b. More than one experimental data set is available. The unperturbed input data $u_o(t)$ from the different data sets are correlated.

When **A4b** holds, a simple estimation algorithm can be applied, see Section 5. However, it is not obvious when the assumption can be realistically applied.

2.5 Notations

The following notations will be convenient. The parameter vector to be estimated is

$$\theta = (a_1 \dots a_n \; b_1 \dots b_n)^T. \tag{2.5}$$

Similarly we introduce the regressor vector

$$\varphi(t) = (-y(t-1) \dots - y(t-n) \; u(t-1) \dots u(t-n))^T. \tag{2.6}$$

Further, we will use the conventions:

- θ_o denotes the true parameter vector, and $\hat{\theta}$ denotes its estimate.
- $\varphi_o(t)$ denotes the noisefree part of the regressor vector. Hence

$$\varphi_o(t) = (-y_o(t-1) \dots - y_o(t-n) \; u_o(t-1) \dots u_o(t-n))^T. \tag{2.7}$$

- $\tilde{\varphi}(t)$ denotes the noise-contribution to the regressor vector. This means that

$$\tilde{\varphi}(t) = (-\tilde{y}(t-1) \ldots - \tilde{y}(t-n)\ \tilde{u}(t-1) \ldots \tilde{u}(t-n))^T. \tag{2.8}$$

3 Identifiability and verifiability

Consider the case of Gaussian distributed data. This means that only (first and) second order moments carry information about the distribution, and that higher-order moments do not bring further information. We may alternatively say that we limit our study for the time being to infer information from second-order statistics.

It turns out that in such cases, without introducing more explicit assumptions, such as **A2e**, it is not possible to uniquely identify the system. As illustration, consider the following example.

Example 3.1. Let the measurement noises be auto-correlated, with spectral densities $\phi_{\tilde{u}}$, $\phi_{\tilde{y}}$, respectively. (Note that this is weaker than **A2b**.) Set

$$z(t) = \begin{pmatrix} u(t) \\ y(t) \end{pmatrix} . \tag{3.1}$$

In the Gaussian case all information in the data is represented by the second order statistics, notably by the spectrum

$$\phi_z = \begin{pmatrix} 1 & G^* \\ G & GG^* \end{pmatrix} \phi_{u_o} + \begin{pmatrix} \phi_{\tilde{u}} & 0 \\ 0 & \phi_{\tilde{y}} \end{pmatrix} . \tag{3.2}$$

With ϕ_z given and G, ϕ_{u_o}, $\phi_{\tilde{u}}$, $\phi_{\tilde{y}}$ as unknowns, there is no unique solution to (3.2). Recall that due to symmetry, the nondiagonal elements of ϕ_z brings identical information.

To be more explicit, let the estimates of the aforementioned variables be denoted by $\hat{G}$, $\hat{\phi}_{u_o}$, $\hat{\phi}_{\tilde{u}}$, $\hat{\phi}_{\tilde{y}}$. Equating the elements of (3.2) leads to

$$\begin{aligned} \hat{\phi}_{u_o} + \hat{\phi}_{\tilde{u}} &= \phi_{u_o} + \phi_{\tilde{u}} & (3.3)\\ \hat{G}\hat{\phi}_{u_o} &= G\phi_{u_o} & (3.4)\\ \hat{G}\hat{G}^*\hat{\phi}_{u_o} + \hat{\phi}_{\tilde{y}} &= GG^*\phi_{u_o} + \phi_{\tilde{y}} \ . & (3.5) \end{aligned}$$

Choose a spectrum $\hat{\phi}_{u_o}$ (so far arbitrarily). One can with straightforward calculations derive

$$\begin{aligned} \hat{G} &= G\frac{\phi_{u_o}}{\hat{\phi}_{u_o}} & (3.6)\\ \hat{\phi}_{\tilde{u}} &= \phi_{\tilde{u}} + \phi_{u_o} - \hat{\phi}_{u_o} & (3.7)\\ \hat{\phi}_{\tilde{y}} &= \phi_{\tilde{y}} + GG^*\left(\phi_{u_o} - \frac{\phi_{u_o}^2}{\hat{\phi}_{u_o}}\right) \ . & (3.8) \end{aligned}$$

Requiring that $\hat{\phi}_{\tilde{u}}$ and $\hat{\phi}_{\tilde{y}}$ are positive definite function will give the possible values of $\hat{\phi}_{u_o}$. Some straightforward calculations give

$$-\frac{\phi_{u_o}\phi_{\tilde{y}}}{\phi_{u_o}GG^* + \phi_{\tilde{y}}} \leq \hat{\phi}_{u_o} - \phi_{u_o} \leq \phi_{\tilde{u}} \ . \tag{3.9}$$

This shows clearly that except for the trivial and true solution $\hat{\phi}_{u_o} = \phi_{u_o}$, many more solutions are possible. ■

In the paper [17], the problem of identifying multi-variable EIV models from the second-order moments of the observations is considered. Starting from the fact that this knowledge is not sufficient for uniquely identifying the underlying system, a class of observationally equivalent systems is derived and the mapping relating this class to the spectral densities of the observations is investigated. A particular attention is given to the models characterized by the maximum number of outputs.

Results:

1. Without assumptions on the structures (assumptions on parameterizations), there will *not* be a unique solution to the identification problem.
2. It is *not* possible to verify from the data, whether or not these structural assumptions are verified.

■

4 Bias-compensation of the least-squares estimate

When estimating the parameter vector θ from the linear regression model

$$y(t) = \varphi(t)\theta + \varepsilon(t), \tag{4.1}$$

the normal equations will be (in the asymptotic case of an infinite number of data points)

$$[E\varphi(t)\varphi^T(t)]\hat{\theta}_{LS} = E\varphi(t)y(t). \tag{4.2}$$

As the data contain one contribution due to the unperturbed input $u_o(t)$ and another due to the noise, the normal equations can be written as

$$[E\varphi_o(t)\varphi_o^T(t) + E\tilde{\varphi}(t)\tilde{\varphi}^T(t)]\hat{\theta}_{LS} = E\varphi_o(t)y_o(t) + E\tilde{\varphi}(t)\tilde{y}(t). \tag{4.3}$$

Under assumption **A2a**, the output noise is white, and hence it holds

$$E\tilde{\varphi}(t)\tilde{y}(t) = 0. \tag{4.4}$$

As the true parameter vector θ_o satisfies

$$[E\varphi_o(t)\varphi_o^T(t)]\theta_o = E\varphi_o(t)y_o(t) \tag{4.5}$$

it follows that there will be a bias,

$$\hat{\theta}_{LS} \neq \theta_o. \tag{4.6}$$

The bias is caused by the matrix

$$E\tilde{\varphi}(t)\tilde{\varphi}^T(t) = R_{\tilde{\varphi}\tilde{\varphi}} = \begin{pmatrix} \lambda_y I & 0 \\ 0 & \lambda_u I \end{pmatrix} \tag{4.7}$$

There are many identification schemes available that aim to compensate for the bias effect that is induced by the noise sources. Assuming the noise variances are known or estimated, a bias-compensated least-squares scheme can be constructed as

$$\hat{\theta}_{BCLS} = \left[\hat{R}_{\varphi\varphi} - \begin{pmatrix} \hat{\lambda}_y I & 0 \\ 0 & \hat{\lambda}_u I \end{pmatrix} \right]^{-1} \hat{r}_{\varphi y} \ . \tag{4.8}$$

where

$$\hat{R}_{\varphi\varphi} = \frac{1}{N} \sum_{t=1}^{N} \varphi(t)\varphi^T(t), \qquad \hat{r}_{\varphi y} = \frac{1}{N} \sum_{t=1}^{N} \varphi(t)y(t) \tag{4.9}$$

In the case of no input noise, and only output noise, it is fairly straightforward to devise an estimator for λ_y, using a total least-squares (TLS) scheme, see [24].

One of the first solutions for identification of dynamic errors–in–variables models using a bias-compensated scheme was presented in the works of Koopmans [14] and Levin [15] (see also [4]). In more recent years the Koopmans–Levin method has been re-proposed by several authors with slightly different variants [5], [6], [30]. One of the main drawbacks of the Koopmans–Levin method is constituted by the fact that the input and output noises are assumed white and the ratio of their variances is *a priori* known, (*i.e.* **A2e** is assumed to hold).

The compensated-bias least-squares techniques [24], based on suitable pre-filtering elaborations of the available data, are even more flexible since they can work under more general noise conditions than those of assumption **A2a, A2b**. The appealing properties of these approaches have been investigated in recent years and several recursive and off–line algorithms have been proposed [32], [33], [34].

Note that it is crucial in all these schemes that the measurement noise sequences are white (assumption **A2b**). Let us also remark, that a wide class of bias-compensated least squares methods can in fact be interpreted as specific variants of the instrumental variable method, see [25].

5 Use of repeated and correlated experiments

Consider the linear regression model

$$y(t) = \varphi^T(t)\theta + \varepsilon(t) \tag{5.1}$$

where $\varepsilon(t)$ denotes the equation error. Assume that more than one data set is available, so that there are, in fact, a number of models (at least two data sets will be needed)

$$y^{(i)}(t) = \varphi^{(i)T}(t)\theta + \varepsilon^{(i)}(t), \quad i = 1, 2, \ldots \tag{5.2}$$

The true parameter vector fits perfectly the models when undisturbed data are used:

$$y_o^{(i)}(t) = \varphi_o^{(i)T}(t)\theta_o. \tag{5.3}$$

Assume now that

1. the noise is independent in the different data sets
2. the unperturbed regressor vector is (well) correlated in the different data sets.

Using two data sets, we may then form

$$\begin{aligned}
& E[\varphi^{(1)}(t)\varphi^{(2)T}(t)]\theta_o - E[\varphi^{(1)}(t)y^{(2}(t))] \\
= \; & E[\varphi_o^{(1)}(t) + \tilde{\varphi}^{(1)}(t)][\varphi_o^{(2)T}(t) + \tilde{\varphi}^{(2)T}(t)]\theta_o - E[\varphi_o^{(1)}(t) + \tilde{\varphi}^{(1)}(t)][y_o^{(2)}(t) + \tilde{y}^{(2)}(t)] \\
= \; & E[\varphi_o^{(1)}(t)\varphi_o^{(2)T}(t)]\theta_o - E[\varphi_o^{(1)}(t)y_o^{(2)}(t)] \\
= \; & E\varphi_o^{(1)}(t)[\varphi_o^{(2)T}(t)\theta_o - y_o^{(2)}(t)] = 0 \qquad (5.4)
\end{aligned}$$

Assume that the matrix $E[\varphi_o^{(1)}(t)\varphi_o^{(2)T}(t)]$ is nonsingular. This is partly a condition on the inputs being persistently exciting. It is also a condition on sufficient correlation between the data sets. The consequence is that from *two* data sets, it is possible to derive a consistent parameter estimator as

$$\left[\frac{1}{N}\sum_{t=1}^{N}\varphi^{(1)}(t)\varphi^{(2)T}(t)\right]\hat{\theta} = \left[\frac{1}{N}\sum_{t=1}^{N}\varphi^{(1)}(t)y^{(2)}(t)\right]. \qquad (5.5)$$

This is indeed an instrumental variable estimator, [22].

A special case is mentioned in [7], [18], where some specific measurement situations allow for repeated experiments where the noise-free regressor vector remain the same, that is $\varphi_o^{(i)}(t)$ does not vary from one experiment to another (it does not vary with i). It is though also possible to apply the estimator (5.5) in other situations, and the 'experiments' can be allowed to be overlapping, as long as the basic assumptions are satisfied. In particular two subsets of a single experiments may be used, as illustrated in Figure 2.

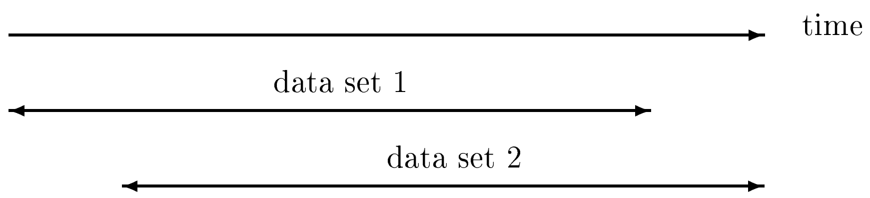

Figure 2: Split of a total data set into two overlapping ones.

Results:

1. If the total data set can be split into parts, where the measurement noise $\tilde{y}(t)$, $\tilde{\varphi}(t)$ in the different parts are uncorrelated, but the noisefree regressors $\varphi_o(t)$ are correlated, it is possible to derive simple and consistent estimates of the parameter vector θ.
2. The estimates are IV variants, with delayed input-output data as instruments.

6 Nonparametric modeling of the input

A tempting approach is to treat the signal values of the undisturbed input, $u_o(t),\ t = 1, \ldots, N$ as auxiliary unknowns to be estimated. Even if this can be easily done in practice, it seems though not to be a pertinent approach in general as shown in the following. When assumption **A2e** holds it is though fully feasible.

Consider a case as simple as possible, namely a purely static system:

$$\begin{array}{rcl} y(t) & = & bu_o(t) + \tilde{y}(t) \\ u(t) & = & u_o(t) + \tilde{u}(t) \end{array} . \tag{6.1}$$

We regard

$$\theta = (b\ \lambda_y\ \lambda_u)^T, \qquad U_o = (u_o(1) \ldots u_o(N))^T \tag{6.2}$$

as the unknowns. Assume the data to be Gaussian distributed. The negative loglikehood function will then be (neglecting a constant term and skipping a factor of $N/2$)

$$\begin{array}{rcl} L(\theta, U_o) & = & \dfrac{1}{N\lambda_y} \sum_{t=1}^{N} [y(t) - bu_o(t)]^2 \\ & & + \dfrac{1}{N\lambda_u} \sum_{t=1}^{N} [u(t) - u_o(t)]^2 + \log \lambda_y + \log \lambda_u \ . \end{array} \tag{6.3}$$

The loss function $L(\theta, U_o)$ is easily first minimized with respect to U_o giving

$$u_o(t) = \frac{b\lambda_u y(t) + \lambda_y u(t)}{b^2 \lambda_u + \lambda_y},\ t = 1, \ldots, N \ . \tag{6.4}$$

The concentrated loss function becomes, after inserting (6.4) into (6.3) and some straightforward calculations

$$\begin{array}{rcl} V(\theta) & = & \min\limits_{U_o} L(\theta, U_o) \\ & = & \dfrac{1}{(b^2 \lambda_u + \lambda_y)} \left(\hat{R}_y + b^2 \hat{R}_u - 2b\hat{R}_{yu} \right) + \log \lambda_y + \log \lambda_u, \end{array} \tag{6.5}$$

where

$$\hat{R}_y = \frac{1}{N} \sum_{t=1}^{N} y^2(t), \qquad \hat{R}_u = \frac{1}{N} \sum_{t=1}^{N} u^2(t), \qquad \hat{R}_{yu} = \frac{1}{N} \sum_{t=1}^{N} y(t)u(t) \ . \tag{6.6}$$

One can now consider some different cases, based on what assumptions on the noise variances that apply. Three cases are presented here as lemmas.

Lemma 6.1. Assume that the noise variances λ_u and λ_y both are known, (**A2d**). Let N tend to infinity. Then the loss function V is minimized with respect to b for the true value b_o.

Proof. See appendix A. ∎

Lemma 6.2. Assume that the ratio λ_y/λ_u between the noise variances is known, (**A2e**). Let N tend to infinity. Then the loss function V is minimized with respect to b for the true value b_o.

Proof. See appendix A. ■

Lemma 6.3 Assume that both noise variances λ_y and λ_y are unknown, (**A2c**). Let N tend to infinity. Then the loss function is *not* minimized for $b = b_o$.

Proof. See appendix A. (This result was also shown in [21].) ■

Results:

1. The estimate $\hat{b}$ will not be consistent in the cases
   - λ_u and λ_y are both unknown, (**A2c**).
   - One of the variances λ_u and λ_y is known and the other is unknown. (This is proved similarly as Lemma 6.3.)
2. The estimate $\hat{b}$ will be consistent, if either
   - both the noise variances λ_y and λ_u are known, (**A2d**).
   - their ratio λ_y/λ_u is known, (**A2e**).
3. Under any of the assumptions **A2d** or **A2e** it is in fact possible to consistently estimate the gain b_o, even if there are $2N$ data values, and $N+2$ (or $N+1$) unknowns. It is hence not the fact that the number of unknowns grows as fast as the number of data points, that causes the consistency problem in other cases.
4. Extension of the above results to the dynamic case are possible, and will be reported elsewhere.

7 Relation between total least-squares and maximum likelihood

7.1 Total least-squares

Consider the overdetermined linear system of equations

$$Ax \approx b. \tag{7.1}$$

The least-squares solution is well-known to be

$$x_{LS} = (A^T A)^{-1} A^T b \tag{7.2}$$

assuming A has full rank. The solution corresponds to attributing all uncertainties in (7.1) to b, while A is treated as exactly known.

In contrast, for the total least-squares solution, [30], [10], [11], it is assumed that both A and b are subject to uncertainties. The TLS solution, x_{TLS}, satisfies

$$(A + \Delta A \quad b + \Delta b) \begin{pmatrix} x_{TLS} \\ -1 \end{pmatrix} = 0 \tag{7.3}$$

where the perturbation is made as small as possible

$$\| \, (\Delta A \quad \Delta b) \, \|_F \rightarrow \min \tag{7.4}$$

where $\| \cdot \|_F$ denotes the Frobenius norm. The TLS solution is computed from a singular value decomposition of the augmented matrix $[A \; b]$. To be specific, let

$$(A \quad b) = U_1 \Sigma_1 V_1^T + u_{n+1} \sigma_{n+1} v_{n+1}^T \tag{7.5}$$

where σ_{n+1} is the smallest singular value with associated left singular vector u_{n+1} and right singular vector v_{n+1}. Then

$$x_{TLS} = -\frac{v_{n+1}(1:n)}{v_{n+1}(n+1)} \tag{7.6}$$

The denominator $v_{n+1}(n+1)$ is nonzero generically. Further, the corresponding perturbation of A and b can also be determined:

$$\begin{aligned} (\Delta A \quad \Delta b) &= -u_{n+1} \sigma_{n+1} v_{n+1}^T \\ &= (A \quad b) \, v_{n+1} v_{n+1}^T \\ &= (A \quad b) \begin{pmatrix} x_{TLS} \\ -1 \end{pmatrix} \begin{pmatrix} x_{TLS}^T & -1 \end{pmatrix} \frac{1}{1 + \| \, x_{TLS} \, \|^2} \end{aligned} \tag{7.7}$$

The LS and TLS solutions can be illustrated as in Figure 3, where a straight line, $y \sim kx$, is fitted to some data points. For the LS solution all uncertainty is attributed to the y_i values. When all uncertainties are set to the x_i values, a so called data least squares solution is obtained. The TLS solution corresponds geometrically to minimization of the sum of squared perpendicular distances from the data points to the line. Clearly, the three solutions will lead to different lines (different values of k).

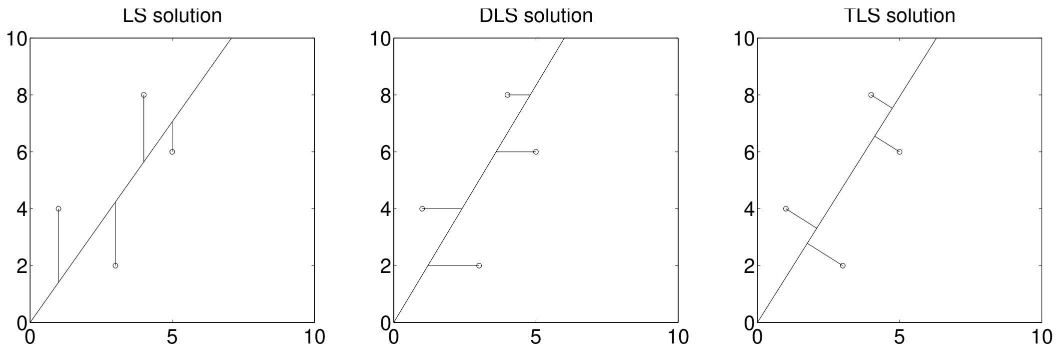

Figure 3: Illustration of the least squares (LS) estimate, the data least squares (DLS) estimate and the total least squares (TLS) estimate.

Further, it can be seen that for the TLS solution to make full sense, the uncertainties in x and y directions should be the same. If this is not the case, some scaling of the variables should first be exercised.

7.2 The TLS solution to the errors-in-variables problem

In this subsection we consider the scalar static case treated in Section 6. We assume that the ratio

$$r = \frac{\lambda_y^o}{\lambda_u^o} \tag{7.8}$$

is known. This allows for an appropriate scaling. To this aim introduce

$$v(t) \triangleq y(t)/\sqrt{r}. \tag{7.9}$$

Then $\tilde{v}(t) = \tilde{y}(t)/\sqrt{r}$ and $\tilde{u}(t)$ have both variance λ_u^o. The linear regression problem

$$\begin{pmatrix} y(1) \\ y(2) \\ \vdots \\ y(N) \end{pmatrix} \approx \begin{pmatrix} u(1) \\ u(2) \\ \vdots \\ u(N) \end{pmatrix} b \tag{7.10}$$

becomes after scaling

$$c = b/\sqrt{r}, \tag{7.11}$$

$$\begin{pmatrix} v(1) \\ \vdots \\ v(N) \end{pmatrix} \approx \begin{pmatrix} u(1) \\ \vdots \\ u(N) \end{pmatrix} c. \tag{7.12}$$

The TLS estimate is according to (7.11)

$$\hat{b}_{TLS} = \hat{c}_{TLS}\sqrt{r}. \tag{7.13}$$

Further,

$$\begin{pmatrix} 1 \\ -\hat{c}_{TLS} \end{pmatrix}$$

is the eigenvector associated to the smallest eigenvalue of

$$\begin{pmatrix} v(1) & \dots & v(N) \\ u(1) & \dots & u(N) \end{pmatrix} \begin{pmatrix} v(1) & u(1) \\ \vdots & \vdots \\ v(N) & u(N) \end{pmatrix} \triangleq \begin{pmatrix} \hat{R}_v & \hat{R}_{vu} \\ \hat{R}_{vu} & \hat{R}_u \end{pmatrix}.$$

The result after some calculations are as follows.

Lemma 7.1. For the present example, the TLS solution reads

$$\hat{b}_{TLS} = \frac{\hat{R}_y - r\hat{R}_u}{2\hat{R}_{yu}} + \frac{1}{2\hat{R}_{yu}} \left[(r\hat{R}_u - \hat{R}_y)^2 + 4r\hat{R}_{yu}^2 \right], \tag{7.14}$$

$$\widehat{u_o}(t)_{TLS} = \frac{\hat{b}_{TLS} y(t) + r u(t)}{\hat{b}_{TLS}^2 + r}. \tag{7.15}$$

Proof. See Appendix B. ■

7.3 Relation between the TLS solution and the maximum likelihood estimate

It is of interest to compare the TLS results to what a maximum likelihood (ML) estimation would give. The ML estimates are obtained by maximizing the loss function L (6.3) of Section 6.

Lemma 7.2. For the present example, the ML solution reads

$$\hat{b}_{ML} = -\frac{r\hat{R}_u - \hat{R}_y}{2\hat{R}_{yu}} + \frac{1}{2\hat{R}_{yu}} \left[(r\hat{R}_u - \hat{R}_y)^2 + 4r\hat{R}_{yu}^2 \right]^{1/2}, \tag{7.16}$$

$$\widehat{u_o}(t)_{ML} = \frac{\hat{b}_{ML} y(t) + r u(t)}{\hat{b}_{ML}^2 + r}. \tag{7.17}$$

Proof. See Appendix B. ■

Results:

1. Under the given assumptions (static deterministic model), the total least-squares estimate $\hat{b}_{TLS}$ coincides with the maximum likelihood estimate $\hat{b}_{ML}$. See also [30], where such a relation is described, under the assumption that the noise in all elements of A and b are white and of equal variance.

2. The same statement is true for the estimate of the unperturbed input, that is $\widehat{u_o}(t)_{TLS} \equiv \widehat{u_o}(t)_{ML}$.

8 ARMAX modelling of the system

The idea here is to rewrite the total dynamics into an ARMAX form as follows. It is easily seen that

$$\begin{aligned} A(q^{-1})y(t) - B(q^{-1})u(t) &= A(q^{-1})[y_o(t) + \tilde{y}(t)] - B(q^{-1})[u_o(t) + \tilde{u}(t)] \\ &= A(q^{-1})\tilde{y}(t) - B(q^{-1})\tilde{u}(t) \ . \end{aligned} \tag{8.1}$$

As the right hand side is the difference of two moving average processes, one of order n and the other of order $n-1$, one can apply a spectral factorization approach to get an equivalent description, in the sense that the statistical properties of the total noise term will remain the same. Therefore, introduce a monic polynomial $C(q^{-1})$ and a positive scalar σ such that

$$C(z)C(z^{-1})\sigma = A(z)A(z^{-1})\lambda_y + B(z)B(z^{-1})\lambda_u \ . \tag{8.2}$$

Then we can replace the noise term in (8.1) by *one* moving average process as

$$C(q^{-1})e(t) = A(q^{-1})\tilde{y}(t) - B(q^{-1})\tilde{u}(t) \ . \tag{8.3}$$

where the noise source $e(t)$ will be a white noise sequence of variance σ.

Thus it is tempting to try a prediction error approach, [16], [22], where one is estimating the (extended) parameter vector

$$\theta = \left(\hat{a}_1 \dots \hat{a}_n \ \hat{b}_1 \dots \hat{b}_n \ \hat{c}_1 \dots \hat{c}_n\right)^T \tag{8.4}$$

in the straightforward way by minimizing the loss function

$$V(\theta) = \frac{1}{N}\sum_{t=1}^{N} \varepsilon^2(t,\theta) \tag{8.5}$$

and the prediction errors $\varepsilon(t,\theta)$ are computed from the model structure

$$\hat{C}(q^{-1})\varepsilon(t,\theta) = \hat{A}(q^{-1})y(t) - \hat{B}(q^{-1})u(t) \ . \tag{8.6}$$

One can show that the gradient of the criterion *does not* vanish for the true parameter vector, that is

$$V'(\theta_o) \neq 0 \tag{8.7}$$

unless the input is measured exactly (so that $\lambda_u = 0$). Details are shown in [21]. This means that the estimator will not be consistent (and hence not close to the true value θ_o for large N).

A way of interpreting that the approach will not give consistent estimates, is to examine the cross-covariance function between the input $u(t)$ and the prediction error $\varepsilon(t)$:

$$r_{\varepsilon u}(\tau) = E\varepsilon(t+\tau)u(t) \ . \tag{8.8}$$

If a system is operating in open loop, then for an identifiable parameterization the cross-covariance function will be identically zero. The system may be identifiable also under causal feedback, in which case it is required

$$r_{\varepsilon u}(\tau) = 0 \quad \text{for } \tau > 0 \ . \tag{8.9}$$

The cross-covariance function is to be evaluated for the true parameter vector θ_o. In the case under study, the cross-covariance function becomes

$$\begin{aligned}
r_{\varepsilon u}(\tau) &= E\varepsilon(t+\tau)u(t) \\
&= E\left[\frac{\hat{A}(q^{-1})}{\hat{C}(q^{-1})}\left(\frac{B(q^{-1})}{A(q^{-1})}u_o(t+\tau) + \tilde{y}(t+\tau)\right)\right. \\
&\quad \left. - \frac{\hat{B}(q^{-1})}{\hat{C}(q^{-1})}\Big(u_o(t+\tau) + \tilde{u}(t+\tau)\Big)\right]_{\hat{\theta}=\theta_o} \Big[u_o(t) + \tilde{u}(t)\Big] \\
&= E\left[\frac{A(q^{-1})}{C(q^{-1})}\tilde{y}(t+\tau) - \frac{B(q^{-1})}{C(q^{-1})}\tilde{u}(t+\tau)\right]\Big[u_o(t) + \tilde{u}(t)\Big] \\
&= -E\frac{B(q^{-1})}{C(q^{-1})}\tilde{u}(t+\tau)\tilde{u}(t) \neq 0 \ .
\end{aligned} \tag{8.10}$$

The expected value becomes nonzero, as $\lambda_u > 0$, and by the definition (8.2) the polynomial $C(q^{-1})$ is never identical to 1. Thus the noise term $\frac{B(q^{-1})}{C(q^{-1})}\tilde{u}(t+\tau)$ will be correlated with all past values of $\tilde{u}(t)$. Note that the relation (8.10) means that the process has a *non-causal* feedback, which is the reason why identifiability is lost.

Results:

1. ARMAX modelling of the system is in a sense possible. However, it is not practical for estimation purposes, as it would imply some non-causality, through a feedback mechanism from disturbances to the input.

2. The PEM estimate of the ARMAX parameters will not be consistent.

9 Parametric modelling of the input

In this method the errors–in–variable model (2.1), (2.3) is regarded as a multivariable system with both $u(t)$ and $y(t)$ as outputs [20]. Of crucial importance for this approach is the assumption **A3b**, *i.e.* the noise–free input $u_o(t)$ is characterized by a rational spectrum. The signal $u_o(t)$ can thus be represented by its innovations form, described as an ARMA process of the type

$$u_o(t) = \frac{C(q^{-1})}{D(q^{-1})}\; e(t) \;, \tag{9.1}$$

where $e(t)$ is a white noise with variance λ_e and the polynomials $C(q^{-1})$, $D(q^{-1})$ are of the form

$$\begin{aligned} C(q^{-1}) &= 1 + c_1\, q^{-1} + \ldots + c_p\, q^{-p} \\ D(q^{-1}) &= 1 + d_1\, q^{-1} + \ldots + d_p\, q^{-p} \end{aligned} \;, \tag{9.2}$$

with known degree p; moreover, $C(z)$ and $D(z)$ are relatively prime and asymptotically stable polynomials. In this way, the whole errors–in–variables model

$$\begin{aligned} u(t) &= u_o(t) + \tilde{u}(t) \\ y(t) &= G(q^{-1})\, u_o(t) + \tilde{y}(t) \\ u_o(t) &= C(q^{-1})/D(q^{-1})\;\, e(t) \end{aligned} \tag{9.3}$$

can be considered as a system with a two–dimensional output vector $z(t) = [u(t)\, y(t)]^T$ and three mutually uncorrelated white noise sources $e(t)$, $\tilde{u}(t)$ and $\tilde{y}(t)$:

$$\begin{pmatrix} u(t) \\ y(t) \end{pmatrix} = \begin{pmatrix} \frac{C(q^{-1})}{D(q^{-1})} & 1 & 0 \\ \frac{B(q^{-1})C(q^{-1})}{A(q^{-1})D(q^{-1})} & 0 & 1 \end{pmatrix} \begin{pmatrix} e(t) \\ \tilde{u}(t) \\ \tilde{y}(t) \end{pmatrix} . \tag{9.4}$$

In this way the model $C(q^{-1})/D(q^{-1})$ of the undisturbed input is a part of the errors–in–variables representation and its coefficients must be estimated together with the parameters of $A(q^{-1})$ and $B(q^{-1})$.

Using well–known techniques the model (9.4) can be converted into the innovations form

$$z(t) = S(q^{-1})\, \varepsilon(t) \;, \tag{9.5}$$

where $S(z)$ is the square (2×2) canonical (stable and minimum–phase) spectral factor of $\phi_z(\omega)$ and $\varepsilon(t)$ is the two–dimensional innovations vector with identity covariance matrix. One way to compute (9.5) from (9.4) is to convert the model (9.4) into a state space form. Then, the innovations form (9.5) can be computed from the Kalman filter after solving a Riccati equation. Since the innovations $\varepsilon(t)$ of model (9.5) can also be interpreted as the one step ahead prediction errors, a prediction error method can be applied for the identification of the system parameters. The parameter vector θ includes the coefficients of polynomials $A(q^{-1})$, $B(q^{-1})$, $C(q^{-1})$, $D(q^{-1})$ and the variances λ_e, λ_y, λ_u. The solution of the identification problem in Section 2 is then determined by minimizing the loss function

$$J(\theta) = \det\left(\frac{1}{N}\sum_{t=1}^{N} \varepsilon(t)\,\varepsilon^T(t)\right) , \tag{9.6}$$

where N is the number of input–output samples.

Results:

1. This method yields consistent parameter estimates under quite mild conditions. This was proved in [19], [20].
2. The procedure may fail to give good results when only poor initial parameter estimates are available.
3. There are (many) alternative ways to introduce parametric models for the unperturbed input $u_o(t)$, that leads to consistent parameter estimators.

10 Conclusions

For errors-in-variables identification, it is often crucial what general assumptions on the noise properties and on the unperturbed input that are imposed.

It is shown that some assumptions of this kind needs to be made in order to get consistent estimates of the system parameters. Furthermore, it is not possible to verify the validity of these assumptions from the experimental data used for the estimation.

In contrast, in the case of no input errors (which applies in 'classical identification' situations) these assumptions are not so critical. Then they normally do not influence the consistency properties, but the accuracy of the estimates.

Concerning the assumptions on the experimental conditions, the consequences can be summarized as follows:

1. If the unperturbed input is known to be *periodic*, it is easy to derive simple and consistent estimators. These estimators can also be applied to more general inputs, if the recorded data sets can be splitted into two (possibly overlapping) data sets, where the unperturbed inputs are correlated, and the noise contributions to the data are uncorrelated.

2. Another situation occurs when the unperturbed input is regarded as *completely unknown*, and estimated along with the system parameters. In the general case, it is not possible to get consistent parameter estimates. However, if the ratio between the output noise variance and the input noise variance is known, this approach is fully feasible. It leads to consistent estimates of the system parameters, despite that simultaneously all values of the unperturbed input are estimated.

3. A third possibility is to model the unperturbed input in a *parametric* way, for example that it is generated as a stationary process with a rational spectrum. This means that it can be modelled as an ARMA process, and a form of joint input-output estimation process can be applied. This will be a prediction error method tied to a particular parameterization of the measured data, and gives consistent parameter estimates under weak assumptions.

4. A TLS-based estimator is feasible to use only in cases when the ratio between the noise variances is known.

A Proofs for Section 6

As a preparation, we first repeat the concentrated loss function, see (6.5), (6.6):

$$\begin{aligned} V(\theta) &= \min_{U_o} L(\theta, U_o) \\ &= \frac{1}{(b^2\lambda_u + \lambda_y)} \left(\hat{R}_y + b^2 \hat{R}_u - 2b\hat{R}_{yu} \right) + \log \lambda_y + \log \lambda_u, \end{aligned} \tag{A.1}$$

where

$$\hat{R}_y = \frac{1}{N}\sum_{t=1}^{N} y^2(t), \qquad \hat{R}_u = \frac{1}{N}\sum_{t=1}^{N} u^2(t), \qquad \hat{R}_{yu} = \frac{1}{N}\sum_{t=1}^{N} y(t)u(t) \ . \tag{A.2}$$

The minimum point of $V(\theta)$ will obviously be some function of the covariance elements in (A.2). This is 'alarming' from a consistency point of view, as for the asymptotic case ($N \to \infty$) we have

$$\begin{aligned} R_y &= Ey^2(t) = b_o^2 R_{u_o} + \lambda_y^o, \\ R_{yu} &= b_o R_{u_o}, \\ R_u &= R_{u_o} + \lambda_u^o \ . \end{aligned} \tag{A.3}$$

These covariance elements depend not only on the three true values of the θ elements: b_o, λ_u^o, λ_y^o, but also on R_{u_o}. Expressed differently, regarding $dV/d\theta = 0$ as a system of equations with b_o, λ_u^o, λ_y^o, R_{u_o} as unknowns will lead to an underdetermined system.

More specifically, the gradient of $V(\theta)$ becomes

$$\begin{aligned} V_b &= \frac{1}{(\lambda_y + b^2\lambda_u)^2}\left[(b^2\lambda_u + \lambda_y)(2b\hat{R}_u - 2\hat{R}_{yu}) - 2b\lambda_u(\hat{R}_y + b^2\hat{R}_u - 2b\hat{R}_{yu})\right] \\ &= \frac{2}{(\lambda_y + b^2\lambda_u)^2}\left[b^2\lambda_u\hat{R}_{yu} + b(\lambda_y\hat{R}_u - \lambda_u\hat{R}_y) - \lambda_y\hat{R}_{yu}\right]. \end{aligned} \tag{A.4}$$

$$V_{\lambda_u} = -\frac{b^2}{(b^2\lambda_u + \lambda_y)^2}(\hat{R}_y + b^2\hat{R}_u - 2b\hat{R}_{yu}) + \frac{1}{\lambda_u}. \tag{A.5}$$

$$V_{\lambda_y} = -\frac{1}{(b^2\lambda_u + \lambda_y)^2}(\hat{R}_y + b^2\hat{R}_u - 2b\hat{R}_{yu}) + \frac{1}{\lambda_y}. \tag{A.6}$$

A.1 Proof of Lemma 6.1

When λ_u and λ_y are known, the equation determining the limiting value of b is given by $V_b = 0$. This leads to the the second order equation

$$b^2\lambda_u\hat{R}_{yu} + b(\lambda_y\hat{R}_u - \lambda\hat{R}_y) - \lambda_y\hat{R}_{yu} = 0 \tag{A.7}$$

and there will be two solutions. Substituting the estimated covariances $\hat{R}_y$, $\hat{R}_u$ and $\hat{R}_{yu}$ by their asymptotic values according to (A.3), and setting $\lambda_y = \lambda_y^o$, $\lambda_u = \lambda_u^o$, we transform (A.7) into

$$b^2\lambda_u^o b_o R_{u_o} + b(\lambda_y^o[R_{u_o} + \lambda_u^o] - \lambda_u^o[b_o^2 R_{u_o} + \lambda_y^o]) - \lambda_y^o b_o R_{u_o} = 0$$

or

$$R_{u_o}(b-b_o)(\lambda_u^o b b_o+\lambda_y^o)=0. \tag{A.8}$$

The first solution gives the true value

$$b=b_o \tag{A.9}$$

leading to

$$V=1. \tag{A.10}$$

The second solution

$$b=-\frac{\lambda_y^o}{\lambda_u^o b_o} \tag{A.11}$$

leads after some calculations to

$$V=1+R_{u_o}\frac{b_o^2\lambda_u^o+\lambda_y^o}{\lambda_u^o\lambda_y^o}>1. \tag{A.12}$$

Obviously, (A.9) is the global minimum point of the loss function. ■

.

A.2 Proof of Lemma 6.2

In this case set

$$r=\frac{\lambda_y}{\lambda_u}=\frac{\lambda_y^o}{\lambda_u^o} \tag{A.13}$$

which is a known ratio, due to the assumption. We can then parameterize V with the unknowns

$$b,\ \lambda_u=\lambda,\ \lambda_y=r\lambda. \tag{A.14}$$

Setting the gradient to zero, still leads to (A.7). Using (A.13) and (A.14) gives

$$b^2\lambda\hat{R}_{yu}+b(r\lambda\hat{R}_u-\lambda\hat{R}_y)+r\lambda\hat{R}_{yu}=0. \tag{A.15}$$

After shortening a factor λ and introducing the asymptotic values (A.3) one gets

$$b^2 b_o R_{u_o}+b(\lambda_y^o/\lambda[R_{u_o}+\lambda_u^o]-[b_o^2R_{u_o}+\lambda_y^o])+\lambda_y^o/\lambda_u^o b_o R_{u_o}=0$$

and then

$$R_{u_o}(b-b_o)(bb_o+\lambda_y^o/\lambda_u^o)=0 \tag{A.16}$$

leading again to two solutions. In order to investigate the corresponding values of the loss function, we need also to determine the corresponding values of λ. The parameter λ is determined by

$$\begin{aligned}
0 &= V_\lambda=V_{\lambda_u}+V_{\lambda_y}r \\
&= -\frac{b^2+r}{(b^2\lambda_u+\lambda_y)^2}[\hat{R}_y+b^2\hat{R}_u-2b\hat{R}_{yu}]+\frac{1}{\lambda_u}+\frac{r}{\lambda_y} \\
&= -\frac{1}{\lambda^2}\frac{1}{b^2+r}[\hat{R}_y+b^2\hat{R}_u-2b\hat{R}_{yu}]+\frac{2}{\lambda}.
\end{aligned} \tag{A.17}$$

Hence we get

$$\lambda = \frac{1}{2(b^2+r)}[\hat{R}_y + b^2\hat{R}_u - 2b\hat{R}_{yu}]. \tag{A.18}$$

The corresponding values of V hence become

$$\begin{aligned} V &= \frac{1}{\lambda(b^2+r)}[\hat{R}_y + b^2\hat{R}_u - 2b\hat{R}_{yu}] + \log(r\lambda) + \log\lambda \\ &= 2 + \log r + 2\log\left[\frac{\hat{R}_y + b^2\hat{R}_u - 2b\hat{R}_{yu}}{2(b^2+r)}\right]. \end{aligned} \tag{A.19}$$

We evaluate the stationary points of V for the asymptotic case, where (A.3) applies. The first stationary point is given by

$$b = b_o, \quad \lambda = \frac{\lambda_o}{2} \tag{A.20}$$

leading to

$$\begin{aligned} V &= 2 + \log\frac{\lambda_y^o}{\lambda_u^o} + 2\log\frac{\lambda_u^o}{2} \\ &= 2 + \log\lambda_y^o + \log\lambda_u^o - 2\log 2. \end{aligned} \tag{A.21}$$

The second stationary point is found after some calculations to be

$$b = -\frac{\lambda_y^o}{b_o\lambda_u^o}, \qquad \lambda = \frac{\lambda_u^o}{2} + \frac{R_{u_o}}{2r}(b_o^2 + r) \tag{A.22}$$

which, again after some calculations, leads to

$$V = 2 + \log\frac{\lambda_y^o}{\lambda_u^o} + 2\log\left[\frac{\lambda_u^o}{2} + \frac{R_{u_o}(b_o^2+r)}{2r}\right]. \tag{A.23}$$

It is clear that the V value of (A.21) is smaller than (A.23). Hence (A.20) is the (global) minimum point. It gives the true value (b_o) of b, although there is a bias of the estimate λ. ∎

A.3 Proof of Lemma 6.3

When b, λ_u and λ_y are unknowns, the stationary points are solved by setting (A.4) - (A.6) to zero. When the asymptotic expressions (A.3) of the covariance elements are substituted it can be seen that (A.4) is satisfied for the true parameter values, but (A.5) and (A.6) are not. Hence, the $b = b_o$, $\lambda_u = \lambda_u^0$, $\lambda_y = \lambda_y^o$ is *not* a stationary point. A more detailed analysis carried out in [21] shows that in fact *no* stationary point satisfies $b = b_o$. (Note that this is in contrast to the situation treated in Lemma 6.2).

B Proofs for Section 7

B.1 Proof of Lemma 7.1

In order to apply TLS in a decent way, the problem has first to be rescaled so that the dependent and independent variables have equal variances.

The eigenvalue problem in question is easily formulated as

$$\begin{pmatrix} \hat{R}_v - \alpha & \hat{R}_{vu} \\ \hat{R}_{vu} & \hat{R}_u - \alpha \end{pmatrix} \begin{pmatrix} 1 \\ -\hat{c}_{TLS} \end{pmatrix} = 0. \tag{B.1}$$

The first equation of (B.1) gives

$$\hat{c}_{TLS} = \frac{\hat{R}_v - \alpha}{\hat{R}_{vu}} \tag{B.2}$$

which after insertion into the second equation gives

$$\hat{R}_{vu} - (\hat{R}_u - \alpha)\frac{\hat{R}_v - \alpha}{\hat{R}_{vu}} = 0.$$

This characteristic equation is easily rewritten as

$$\alpha^2 - \alpha(\hat{R}_u + \hat{R}_v) + (\hat{R}_u\hat{R}_v - \hat{R}_{uv}^2) = 0 \tag{B.3}$$

and the smallest eigenvalue (which is also the smallest singular value of

$$(\mathbf{v}\ \mathbf{u}) \triangleq \begin{pmatrix} v(1) & u(1) \\ \vdots & \vdots \\ v(N) & u(N) \end{pmatrix} \tag{B.4}$$

is obviously

$$\alpha = \frac{1}{2}(\hat{R}_u + \hat{R}_v) - \frac{1}{2}\left[(\hat{R}_u - \hat{R}_v)^2 + 4\hat{R}_{uv}^2\right]^{1/2}. \tag{B.5}$$

Inserting the expression (B.5) into (B.2) and using (7.13) and (7.9) gives

$$\begin{aligned} \hat{b}_{TLS} &= \frac{\sqrt{r}}{\hat{R}_{vu}}\left[\frac{\hat{R}_v - \hat{R}_u}{2} + \frac{1}{2}\left[(\hat{R}_v - \hat{R}_u)^2 + 4\hat{R}_{uv}^2\right]^{1/2}\right] \\ &= \frac{r}{\hat{R}_{vu}}\left[\frac{\hat{R}_y/r - \hat{R}_u}{2} + \frac{1}{2}\left[(\hat{R}_u - \hat{R}_y/r)^2 + \frac{4}{r}\hat{R}_{yu}^2\right]^{1/2}\right] \\ &= \frac{\hat{R}_y - r\hat{R}_u}{2\hat{R}_{yu}} + \frac{1}{2\hat{R}_{yu}}\left[(r\hat{R}_u - \hat{R}_y)^2 + 4r\hat{R}_{yu}^2\right]^{1/2} \end{aligned}$$

which is precisely (7.14).

The TLS estimate of $u_o(t)$ can be found from (7.7):

$$\begin{aligned}
\mathbf{u}_{TLS} &= \mathbf{u} - \Delta\mathbf{u} \\
&= \mathbf{u} - (\mathbf{u} \;\; \mathbf{v}) \begin{pmatrix} c_{TLS} \\ -1 \end{pmatrix} \begin{pmatrix} c_{TLS} & -1 \end{pmatrix} \frac{1}{1 + c_{TLS}^2} \begin{pmatrix} 1 \\ 0 \end{pmatrix} \\
&= \mathbf{u} - \left[\frac{b_{TLS}}{\sqrt{r}} \mathbf{u} - \frac{1}{\sqrt{r}} \mathbf{y} \right] \frac{b_{TLS}/\sqrt{r}}{1 + b_{TLS}^2/r} \\
&= \mathbf{u} - [b_{TLS}\mathbf{u} - \mathbf{y}] \frac{b_{TLS}}{r + b_{TLS}^2} \\
&= \frac{1}{r + b_{TLS}^2} [r\mathbf{u} + b_{TLS}\mathbf{y}]. && \text{(B.6)}
\end{aligned}$$

which is the same as (7.17).

■

B.2 Proof of Lemma 7.2

The estimate $\hat{b}_{ML}$ is the solution of $V_b = 0$, (A.4). Inserting $\lambda_y = r\lambda_u$, this gives

$$b^2 \hat{R}_{yu} + b(r\hat{R}_u - \hat{R}_y) - r\hat{R}_{yu} = 0 \tag{B.7}$$

with has solutions

$$\hat{b}_{ML} = -\frac{r\hat{R}_u - \hat{R}_y}{2\hat{R}_{yu}} \pm \frac{1}{2\hat{R}_{yu}} \left[(r\hat{R}_u - \hat{R}_y)^2 + 4r\hat{R}_{yu}^2 \right]^{1/2}. \tag{B.8}$$

A closer analysis, for example using the asymptotic case (A.3), reveals that the positive sign applies in (B.8). The corresponding estimate (7.17) of $u_o(t)$ follows directly from (6.4). ■

References

[1] B. D. O. Anderson. Identification of scalar errors-in-variables models with dynamics. *Automatica*, 21:709–716, 1985.

[2] B. D. O. Anderson and M. Deistler. Identifiability of dynamic errors-in-variables models. *J. Time Series Analysis*, 5:1–13, 1984.

[3] B. D. O. Anderson and M. Deistler. Dynamic errors-in-variables systems in three variables. *Automatica*, 23:611–616, 1987.

[4] M. Aoki and P. C. Yue. On *a priori* error estimate of some identification methods. *IEEE Transactions on Automatic Control*, AC-15:541–548, October 1970.

[5] J. A. Cadzow and O. M. Solomon. Algebraic approach to system identification. *IEEE Transactions on Signal Processing*, ASSP-34(3):462–469, June 1986.

[6] K. V. Fernando and H. Nicholson. Identification of linear systems with input and output noise: the Koopmans–Levin method. *IEE Proceedings, Part D*, 132(1):30–36, January 1985.

[7] U. Forsell, F. Gustafsson, and T. McKelvey. Time-domain identification of dynamic errors-in-variables systems using periodic excitation signals. In *Proc. IFAC 14th World Congress*, Beijing, P.R.China, 1999.

[8] R. Frisch. Statistical confluence analysis by means of complete regression systems. Technical Report 5, University of Oslo, Economics Institute, Oslo, Norway, 1934.

[9] W. A. Fuller. *Measurement Error Models.* Wiley, New York, NY, 1987.

[10] G. H. Golub and C. F. Van Loan. An analysis of the total least squares problem. *SIAM J. Numerical Analysis*, 17:883–893, 1980.

[11] G. H. Golub and C. F. Van Loan. *Matrix Computations, 3rd ed.* John Hopkins University Press, Baltimore, MD, USA, 1989.

[12] R. E. Kalman. Identification from real data. In M. Hazewinkel and A. H. G. Rinnoy Kan, editors, *Current Developments in the Interface: Economics, Econometrics, Mathematics*, Dordrecht, The Netherlands, 1982. D. Reidel.

[13] R. E. Kalman. System identification from noisy data. In A. R. Bednarek and L. Cesari, editors, *Dynamical Systems II*, New York, N.Y., 1982. Academic Press.

[14] T. J. Koopmans. *Linear Regression Analysis of Economic Time Series.* N. V. Haarlem, The Netherlands, 1937.

[15] M. J. Levin. Estimation of a system pulse transfer function in the presence of noise. *IEEE Transactions on Automatic. Control*, AC–9:229–235, July 1964.

[16] L. Ljung. *System Identification - Theory for the User, 2nd edition.* Prentice Hall, Upper Saddle River, NJ, USA, 1999.

[17] W. Scherrer and M. Deistler. A structure theory for linear dynamic errors-in-variables models. *SIAM Journal on Control and Optimization*, 36(6):2148–2175, November 1998.

[18] J. Schoukens, R. Pintelon, G. Vandersteen, and P. Guillaume. Frequency domain system identification using non-parametric noise models estimated from a small number of data sets. *Automatica*, 33:1073–1086, 1997.

[19] T Söderström. Some methods for identification of linear systems with noisy input/output data. In *5th IFAC/IFORS Symposium on Identification and System Parameter Estimation*, Darmstadt, Germany, September 1979.

[20] T. Söderström. Identification of stochastic linear systems in presence of input noise. *Automatica*, 17:713–725, 1981.

[21] T. Söderström, U. Soverini, and K. Mahata. Perspectives on errors-in-variables estimation for dynamic systems. *Signal Processing*, 82(8):1139–1154, August 2002.

[22] T. Söderström and P. Stoica. *System Identification.* Prentice Hall International, Hemel Hempstead, UK, 1989.

[23] U. Soverini and T. Söderström. Identification methods of dynamic systems in presence of input noise. In *Proc. SYSID 2000, IFAC 12th Symposium on System Identification*, Santa Barbara, California, June 21–23 2000.

[24] P. Stoica and T. Söderström. Bias correction in least–squares identification. *International Journal of Control*, 35(3):449–457, 1982.

[25] P. Stoica, T. Söderström, and V. Šimonytė. Study of a bias-free least squares parameter estimator. *IEE Proc. Control Theory Appl.*, 142:1–6, 1995.

[26] J. K. Tugnait. Stochastic system identification with noisy input using cumulant statistics. *IEEE Transactions on Automatic Control*, AC-37:476–485, 1992.

[27] J. K. Tugnait and Y. Ye. Stochastic system identification with noisy input-output measurement using polyspectra. *IEEE Transactions on Automatic Control*, AC-40:670–683, 1995.

[28] S. Van Huffel, editor. *Recent Advances in Total Least Squares Techniques and Errors-in-Variables Modelling.* SIAM, Philadelphia, USA, 1997.

[29] S. Van Huffel and Ph. Lemmerling, editors. *Total Least Squares and Errors-in-Variables Modelling. Analysis, Algorithms and Applications.* Kluwer, Dordrecht, The Netherlands, 2002.

[30] S. Van Huffel and J. Vandewalle. *The Total Least Squares Problem: Computation Aspects and Analysis.* SIAM, Philadelphia, USA, 1991.

[31] J. H. Van Schuppen. Stochastic realization problems. In N. Nijmeijer and J. M. Schumacher, editors, *Three Decades of Mathematical System Theory*, Berlin, 1989. Springer-Verlag.

[32] W. X. Zheng. On least–squares identification of stochastic linear systems with noisy input–output data. *International Journal of Adaptive Control and Signal Processing*, 13:11–143, 1999.

[33] W. X. Zheng and C. B. Feng. Unbiased parameter estimation of linear systems in presence of input and output noise. *International Journal of Adaptive Control and Signal Processing*, 3:231–251, 1989.

[34] W. X. Zheng and C. B. Feng. Identification of a class of dynamic errors–in–variables models. *International Journal of Adaptive Control and Signal Processing*, 6:41–440, 1992.