Perspectives on errors-in-variables estimation for dynamic systems

Torsten Söderström[†*], Umberto Soverini[‡], and Kaushik Mahata[†]

† Department of Systems and Control, Information Technology,
Uppsala University, P O Box 27, SE-75103 Uppsala, Sweden
* corresponding author; email: ts@syscon.uu.se

‡ DEIS - Department of Electronics, Information Science and Systems,
University of Bologna, Viale Risorgimento 2, IT–40136 Bologna, Italy
email: usoverini@deis.unibo.it

version April 17, 2001

Abstract

The paper gives an overview of various methods for identifying dynamic errors-in-variables systems. Several approaches are classified by how the original information in time-series data of the noisy input and output measurements is condensed before further processing. For some methods, such as instrumental variable estimators, the information is condensed into a nonsymmetric covariance matrix as a first step before further processing. In a second class of methods, where a symmetric covariance matrix is used instead, the Frisch scheme and other bias-compensation approaches appear. When dealing with the estimation problem in the frequency domain, a milder data reduction typically takes place by first computing spectral estimators of the noisy input-output data. Finally, it is also possible to apply maximum likelihood and prediction error approaches using the original time-domain data in a direct fashion. This alternative will often require quite high computational complexity but yield good statistical efficiency.

The paper is also presenting various properties of parameter estimators for the errors-in-variables problem, and a few conjectures are included, as well as some perspectives and experiences by the authors.

1 Introduction

System identification of linear dynamic systems from noise–corrupted output measurements is a fundamental research problem which has been investigated in the past decades. Many different solutions have been presented, see, for example, [20, 21, 30]. 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 which only recently has received increasing attention. 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 behaviour.

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 [36], 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 [14], [24].

With reference to these systems, the assumptions (*prejudices*) which lie behind the identification procedure have been thoroughly analyzed in [16], [17] with particular attention to the Frisch scheme [13]. This scheme, in particular, assumes that each variable is affected by an unknown amount of additive noise and each noise component is independent of every other noise component and of every variable. As a consequence of this fact, in this case the solution is constituted by a whole family of models compatible with the set of noisy data, unlike other traditional approaches (e.g. least squares, total least squares, maximum likelihood, etc.) where the solution is characterized by a single model.

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 [20], [30], and frequency-domain methods [23] approaches, exploiting deterministic signals (e.g. periodic inputs [12]) or stochastic processes, methods that make use of second–order statistics only or based on higher order cumulant statistics [34], [35]. When only second–order statistics are exploited it is a well–known result that, in general, the identification of errors–in–variables models cannot admit a single solution [1], [2], [3].

Also the stochastic input–output noises can be represented in different ways by assuming different model structures. For example, in [6] the analysis of the Frisch scheme, originally related to the static case, has been extended to linear dynamic SISO systems under the assumption of white and uncorrelated additive noises with unknown variances. In [5] the identification of SISO errors–in–variables models has been analyzed by means of frequency–domain techniques. In [8] the identifiability conditions have been studied in case of SISO errors–in–variables model with coloured input–output noises with unknown rational spectra.

Some comparison between different approaches are given in [26] and [31]. In this paper we will extend the overviews and give further perspectives and comparisons. The paper is organized as follows. The next section presents a model example and gives the general setup used in the paper. The possibility to weaken the assumptions is also discussed. We also give a way to classify different estimators based on how the information in the original time series are condensed before the estimates are computed. Section 3 is devoted to presenting some difficulties that are inherent in the problem. Some basic classification schemes of estimators are introduced in Section 4. Section 5 is devoted to instrumental variable estimators, and other schemes based on a nonsymmetric covariance matrix, while Section 6 presents the Frisch scheme and other bias-compensation methods, which are all based on a symmetric covariance matrix. Some frequency domain estimators are presented in Section 7, while Section 8 is devoted to a prediction error method using directly the originally time-series data. The paper is finished by some numerical examples and a concluding discussion in Section 9.

2 Basic setup and example

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

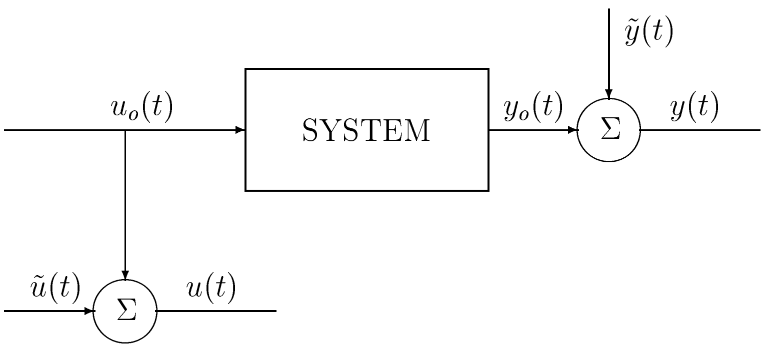

Figure 1: The basic setup for a dynamic error-in-variables problem.

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

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

In this work the following assumptions are introduced.

A1. The dynamic system (1) is asymptotically stable, i.e. $A(z)$ has all zeros outside the unit circle; all the system modes are observable and controllable, i.e. $A(z)$ and $B(z)$ have no common factors. Moreover, the order n of the system is *a priori* known.

A2. The true input $u_o(t)$ is a zero–mean stationary ergodic random signal.

We may also have

A2'. The spectral density of $u_o(t)$ is a rational function.

A3. The sequences $\tilde{u}(t)$ and $\tilde{y}(t)$ are zero–mean white noises with unknown variances λ_u and λ_y, respectively. These processes are mutually uncorrelated and uncorrelated with the noise–free signals $u_o(t)$ and $y_o(t)$.

The problem under investigation can thus be stated as follows.

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

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

and the variances λ_u and λ_y of the noise sequences. ■

The assumptions can sometimes be relaxed. In a few cases additional assumptions need to be imposed.

- 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.
- Most of the methods can be extended to multivariable systems.
- The problem will be significantly simplified if the noise variances λ_u and λ_y are known, or even only the ratio λ_y/λ_u.

- For a maximum likelihood or prediction error approach, there is a need to assume also a model parameterization of the unperturbed input signal $u_o(t)$, see Section 8. Note that Assumption **A2'** implies that $u_o(t)$ can be modelled as an ARMA process.

In order to keep the description reasonably simple, we will stick to the assumptions stated earlier. The essential ingredients in the problem are still kept with this formulation.

3 What difficulties are inherent in the problem?

In this section we present two ways of coping with the estimation problem, that both are fairly natural. As we will see, both fail to solve the problem in a successful way, and we will offer some explanations. We will also give some aspects on the identifiability properties.

3.1 Rewriting the dynamics into an ARMAX model structure

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

$$\begin{aligned} A(q^{-1})y(t) - B(q^{-1})u(t) &= A(q^{-1})[y_o(t) + \tilde{y}(t)] - B(q^{-1})[u_o(t) + \tilde{u}(t)] \\ &= A(q^{-1})\tilde{y}(t) - B(q^{-1})\tilde{u}(t) \ . \end{aligned} \tag{3.1}$$

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

$$C(z)C(z^{-1})\sigma = A(z)A(z^{-1})\lambda_y + B(z)B(z^{-1})\lambda_u \ . \tag{3.2}$$

Then we can replace the noise term in (3.1) by *one* moving average process as

$$C(q^{-1})e(t) = A(q^{-1})\tilde{y}(t) - B(q^{-1})\tilde{u}(t) \ . \tag{3.3}$$

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, [20], [30], where one is estimating the (extended) parameter vector

$$\theta = \left(\hat{a}_1 \dots \hat{a}_n \ \hat{b}_1 \dots \hat{b}_n \ \hat{c}_1 \dots \hat{c}_n\right)^T \tag{3.4}$$

in the straightforward way by minimizing the loss function

$$V(\theta) = \frac{1}{N}\sum_{t=1}^{N} \varepsilon^2(t, \theta) \tag{3.5}$$

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

$$\hat{C}(q^{-1})\varepsilon(t,\theta) = \hat{A}(q^{-1})y(t) - \hat{B}(q^{-1})u(t) \ . \tag{3.6}$$

The asymptotic estimate, as the number of data points N tends to infinity, is given by the minimizing vector of the asymptotic loss function

$$\begin{aligned} V(\theta) &= E\varepsilon^2(t,\theta) \\ &= E\left[\frac{\hat{A}(q^{-1})}{\hat{C}(q^{-1})}y(t) - \frac{\hat{B}(q^{-1})}{\hat{C}(q^{-1})}u(t)\right]^2 \\ &= E\left[\frac{\hat{A}(q^{-1})}{\hat{C}(q^{-1})}\left(\frac{B(q^{-1})}{A(q^{-1})}u_o(t) + \tilde{y}(t)\right) - \frac{\hat{B}(q^{-1})}{\hat{C}(q^{-1})}\left(u_o(t) + \tilde{u}(t)\right)\right]^2 \\ &= E\left[\frac{\hat{A}(q^{-1})B(q^{-1}) - A(q^{-1})\hat{B}(q^{-1})}{A(q^{-1})\hat{C}(q^{-1})}u_o(t)\right]^2 \\ &\quad + E\left[\frac{\hat{A}(q^{-1})}{\hat{C}(q^{-1})}\tilde{y}(t)\right]^2 + E\left[\frac{\hat{B}(q^{-1})}{\hat{C}(q^{-1})}\tilde{u}(t)\right]^2 \end{aligned} \tag{3.7}$$

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

$$V'(\theta_o) \neq 0 \tag{3.8}$$

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

An alternative 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 $\varepsilon(t)$:

$$r_{\varepsilon u}(\tau) = E\varepsilon(t+\tau)u(t) \ . \tag{3.9}$$

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

$$r_{\varepsilon u}(\tau) = 0 \quad \text{for } \tau > 0 \ . \tag{3.10}$$

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

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

$$= \quad -E\frac{B(q^{-1})}{C(q^{-1})}\tilde{u}(t+\tau)\tilde{u}(t) \neq 0 \; . \tag{3.11}$$

The expected value becomes nonzero, as $\lambda_u > 0$, and by the definition (3.2) 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 (3.11) means that the process has a *non-causal* feedback, which is the reason why identifiability is lost.

3.2 Nonparametric estimation of the true input signal

Another 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 as shown in the following.

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

$$\begin{aligned} y(t) &= bu_o(t) + \tilde{y}(t) \\ u(t) &= u_o(t) + \tilde{u}(t) \end{aligned} \tag{3.12}$$

We regard

$$\theta = (b \ \lambda_y \ \lambda_u)^T, \qquad U_o = (u_o(1) \ldots u_o(N))^T \tag{3.13}$$

as the unknowns. Assume the data to be Gaussian distributed. The negative loglikehood function will then be (neglecting a constant term)

$$\begin{aligned} L(\theta, U_o) &= \frac{1}{2\lambda_y}\sum_{t=1}^{N}[y(t) - bu_o(t)]^2 \\ &+ \frac{1}{2\lambda_u}\sum_{t=1}^{N}[u(t) - u_o(t)]^2 + \frac{N}{2}\log\lambda_y + \frac{N}{2}\log\lambda_u \; . \end{aligned} \tag{3.14}$$

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

$$u_o(t) = \frac{b\lambda_u y(t) + \lambda_y u(t)}{b^2\lambda_u + \lambda_y}, \ t = 1, \ldots, N \; . \tag{3.15}$$

It is shown in the appendix B that the concentrated loss function

$$V(\theta) = \min_{U_o} L(\theta, U_o) \tag{3.16}$$

is *not* minimized for the true parameter values.

3.3 Some comments on identifiability

Based on the analysis of Section 3.2, we make the following claim.

Conjecture 3.1. Including all the values of the undisturbed input as additional parameter estimates, will not make it possible to estimate the parameter vector of concern, θ, in a consistent way. ■

Note that the conjecture refers to a situation of nonparametric treatment of the undisturbed input. It does not matter if the estimation is carried out in the time domain or in the frequency domain. However, in case $u_o(t)$ is described by a parametric model (say with a fixed finite number of parameters), then the situation is different. See Section 8 for more details.

Conjecture 3.2. Some structural assumptions (such as model order, correlation structure of measurement noise) are needed for the estimation to be uniquely solved, but these assumptions cannot be verified from data. Expressed differently, the given data cannot be expressed in a unique way as due to measurement noise and the unknown input. ■

The following example supports and illustrates Conjecture 3.2.

Example 3.1. Let the measurement noises be auto-correlated, with spectral densities $\phi_{\tilde{u}}$, $\phi_{\tilde{y}}$, respectively. Set

$$z(t) = \begin{pmatrix} u(t) \\ y(t) \end{pmatrix} . \tag{3.17}$$

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

$$\phi_z = \begin{pmatrix} 1 & G^* \\ G & GG^* \end{pmatrix} \phi_{u_o} + \begin{pmatrix} \phi_{\tilde{u}} & 0 \\ 0 & \phi_{\tilde{y}} \end{pmatrix} . \tag{3.18}$$

With ϕ_z given and G, ϕ_{u_o}, $\phi_{\tilde{u}}$, $\phi_{\tilde{y}}$ as unknowns, there is no unique solution to (3.18). 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 (3.18) leads to

$$\begin{aligned} \hat{\phi}_{u_o} + \hat{\phi}_{\tilde{u}} &= \phi_{u_o} + \phi_{\tilde{u}} \\ \hat{G}\hat{\phi}_{u_o} &= G\phi_{u_o} && (3.19) \\ \hat{G}\hat{G}^*\hat{\phi}_{u_o} + \hat{\phi}_{\tilde{y}} &= GG^*\phi_{u_o} + \phi_{\tilde{y}} \; . && (3.20) \end{aligned}$$

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

$$\begin{aligned} \hat{G} &= G\frac{\phi_{u_o}}{\hat{\phi}_{u_o}} && (3.21) \\ \hat{\phi}_{\tilde{u}} &= \phi_{\tilde{u}} + \phi_{u_o} - \hat{\phi}_{u_o} && (3.22) \\ \hat{\phi}_{\tilde{y}} &= \phi_{\tilde{y}} + GG^*\left(\phi_{u_o} - \frac{\phi_{u_o}^2}{\hat{\phi}_{u_o}}\right) . && (3.23) \end{aligned}$$

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

$$-\frac{\phi_{u_o}\phi_{\tilde{y}}}{\phi_{u_o}GG^* + \phi_{\tilde{y}}} \leq \hat{\phi}_{u_o} - \phi_{u_o} \leq \phi_{\tilde{u}} \; . \tag{3.24}$$

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 frequency domain an essential assumption for identifiability is that the noisy input–output signals $u(t)$, $y(t)$ have a rational spectrum, [8]. In this case the identifiability of the EIV system is ensured even if the orders of the processes are not *a priori* known, provided that no zero/pole cancellation occurs between the transfer function $G(q^{-1})$ and the ARMA model of the noise–free input $u_0(t)$, and all the ARMA processes involved in the EIV representation do not share common poles.

In the paper [22], the problem of identifying multivariable EIV models from the population second 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 continuity property of mapping relating this class to the spectral densities of the observations is investigated. A particular attention is turned to the models characterized by the maximum number of outputs.

4 Classification of estimators

Estimation methods can be classified and organised in different ways. Here, we have found it useful to group methods together, based on an initial data compression step. After a first 'pre-processing' of the data, some reduced information is set up and used for the final computation of the parameter estimates. In case the condensed information is really a sufficient statistics, one would even be able to achieve statistical efficiency in the final step. Also when this is not the case, such an estimation scheme can still be useful, for example due to low computational complexity.

The two steps of the estimators, with an initial data compression, are illustrated in the Figure 2 below.

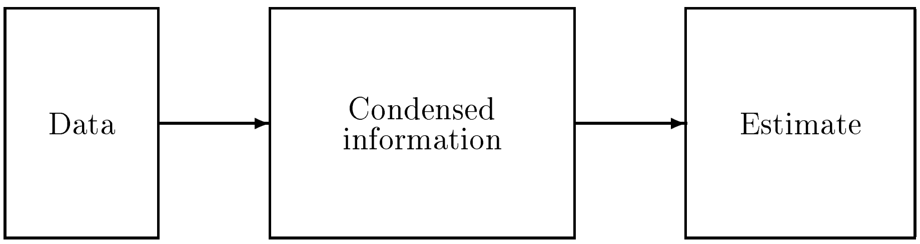

Figure 2: Classification of the data compression prior to the estimation step.

The different groups of methods that will be discussed in the sections to follow, differ in the way the data compression is carried out. The following cases will be treated.

a) Using a nonsymmetric covariance matrix

b) Using a symmetric covariance matrix

c) Using the input-output spectrum

d) Using the time-series data

5 Estimation methods based on a covariance matrix

In this section we will use the parameter vector

$$\theta = (a_1 \dots a_n \; b_1 \dots b_n)^T \triangleq (a^T \; b^T)^T \tag{5.1}$$

and the extended parameter vector

$$\overline{\theta} = \begin{pmatrix} 1 \\ \theta \end{pmatrix} . \tag{5.2}$$

Similarly we introduce the regressor vector

$$\varphi(t) = (-y(t-1) \dots - y(t-n) \; u(t-1) \dots u(t-n))^T \tag{5.3}$$

and the extended regressor vector

$$\begin{aligned} \overline{\varphi}(t) &= (-y(t) \dots - y(t-n) \; u(t-1) \dots u(t-n))^T \\ &= (-y(t) \;\; \varphi^T(t))^T . \end{aligned} \tag{5.4}$$

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

Using the above notations, and the system description (2.1) it follows that

$$\overline{\varphi}_o^T(t)\overline{\theta}_o = (-y_o(t) \;\; \varphi_o^T(t)) \begin{pmatrix} 1 \\ \theta_o \end{pmatrix} = -A(q^{-1})y_o(t) + B(q^{-1})u_o(t) = 0 . \tag{5.5}$$

Further introduce a vector $z(t)$ of dimension $2n$ or larger, satisfying

$$Ez(t)\tilde{\overline{\varphi}}^T(t) = 0 , \tag{5.6}$$

where

$$\tilde{\overline{\varphi}}^T(t) = (-\tilde{y}(t) \dots . - \tilde{y}(t-n) \; \tilde{u}(t-1) \dots \tilde{u}(t-n)). \tag{5.7}$$

As the noise sequences $\{\tilde{y}(t)\}$ and $\{\tilde{u}(t)\}$ are assumed white, one can for example let the vector $z(t)$ in (5.6) be formed from delayed inputs $u(t-L)$ and outputs $y(t-L)$ with the delay $L > n$ in order to satisfy (5.6).

We can now write, using (5.5) and (5.6),

$$Ez(t)\overline{\varphi}^T(t)\overline{\theta}_o = Ez(t)\left(\overline{\varphi}_o^T(t) + \tilde{\overline{\varphi}}^T(t)\right)\overline{\theta}_o = 0 \tag{5.8}$$

which we write for short as

$$R_{z\overline{\varphi}}\overline{\theta}_o = 0 \ . \tag{5.9}$$

The matrix in (5.9) can easily be estimated from the data as

$$\widehat{R_{z\overline{\varphi}}} = \frac{1}{N}\sum_{t=1}^{N} z(t)\overline{\varphi}^T(t) \tag{5.10}$$

and due to the relation (5.9) one can derive several estimators from the approximate relation

$$\widehat{R_{z\overline{\varphi}}}\hat{\overline{\theta}} \approx 0 \ . \tag{5.11}$$

One possible approach for defining the estimate, leading to an instrumental variable (IV) estimator (cf [26], [30]) is to partitionate the matrix in (5.11) as

$$\begin{pmatrix} \hat{r} & \widehat{R_{z\varphi}} \end{pmatrix} \begin{pmatrix} 1 \\ \hat{\theta} \end{pmatrix} = \hat{r} + \widehat{R_{z\varphi}}\hat{\theta} \approx 0 \ . \tag{5.12}$$

In case $\dim z(t) = 2n$, the relation (5.12) gives a linear system of equation with an exact solution, namely the basic IV estimator

$$\hat{\theta}_{IV} = -\widehat{R_{z\varphi}}^{-1}\hat{r} \ . \tag{5.13}$$

When the vector $z(t)$ has higher dimension than $2n$, (5.12) gives an overdetermined system, and has in general no exact solution. A weighted least squares estimator may be taken, which gives the extended IV estimator, see [29], [30],

$$\hat{\theta}_{EIV} = -(\widehat{R_{z\varphi}}^T W \widehat{R_{z\varphi}})^{-1}(\widehat{R_{z\varphi}}^T W\hat{r}) \tag{5.14}$$

where W is a positive definite weighting matrix (no weighting, that is $W = I$, is one possible choice).

An alternative approach for exploiting (5.11) is to note that the relation means that the matrix $\widehat{R_{z\overline{\varphi}}}$ is close to singular. Let $\hat{w}$ denote its right singular vector, associated with its smallest singular value. Then, except for a normalization, $\hat{w}$ should coincide with $\hat{\overline{\theta}}$. Using Matlab notations, this leads to the estimate

$$\hat{\theta}_{SVD} = \hat{w}_{2:2n+1}/\hat{w}_1 \ . \tag{5.15}$$

For the case when $\dim z(t) = 2n$ the estimator $\hat{\theta}_{IV}$ and $\hat{\theta}_{SVD}$ coincide. This is easily seen as follows. First note that in this case $\widehat{R_{z\overline{\varphi}}}$ is of dimension $2n \times (2n+1)$

and is hence singular. The smallest singular value is zero, and the associated singular vector turns out to be $(1 \;\; \hat{\theta}_{IV}^T)^T$:

$$\widehat{R_{z\overline{\varphi}}} \begin{pmatrix} 1 \\ \hat{\theta}_{IV} \end{pmatrix} = \begin{pmatrix} \hat{r} & \widehat{R_{z\varphi}} \end{pmatrix} \begin{pmatrix} 1 \\ -\widehat{R_{z\varphi}}^{-1} \hat{r} \end{pmatrix} = 0 \; . \tag{5.16}$$

When $\dim z(t) > 2n$ holds, the two estimators differ. However, it can be shown, see [27], that they have the same asymptotic covariance matrix in the following sense. It holds that for large N

$$\sqrt{N}(\hat{\theta}_{IV} - \theta_o) \;\; \overset{N\to\infty}{\longrightarrow} \;\; \mathcal{N}(0, P_{IV}) \tag{5.17}$$

$$\sqrt{N}(\hat{\theta}_{SVD} - \theta_o) \;\; \overset{N\to\infty}{\longrightarrow} \;\; \mathcal{N}(0, P_{SVD}) \; , \tag{5.18}$$

where

$$P_{IV} = \sigma(R_{z\varphi}^T W R_{z\varphi})^{-1} R_{z\varphi}^T W \mathrm{cov}\left(C(q^{-1})z(t)\right) W R_{z\varphi}(R_{z\varphi}^T W R_{z\varphi})^{-1} \tag{5.19}$$

$$P_{SVD} = \sigma(R_{z\varphi}^T R_{z\varphi})^{-1} R_{z\varphi}^T \mathrm{cov}\left(C(q^{-1})z(t)\right) R_{z\varphi}(R_{z\varphi}^T R_{z\varphi})^{-1} \; . \tag{5.20}$$

The covariance matrix P_{IV}, (5.19) apparently depends on the weighting matrix W. It can be shown in the same way as for other instrumental variable estimators, see [29], [30], that there is an optimal choice of the weighting matrix. The covariance matrix P_{IV} is minimized with the choice

$$W = \left(\mathrm{cov}\left(C(q^{-1})z(t)\right)\right)^{-1} \; . \tag{5.21}$$

Comments:

- One of the principal advantages of IV the method is its applicability under fairly general noise conditions.
- It is rather inexpensive from a computational point of view.
- However, with this procedure a poor accuracy of the parameter estimates is often obtained.
- The main idea of the IV technique, based on well–known properties of the correlation functions, has been subsequently developed and generalized in several ways, e.g. by combining it with a weighted subspace fitting approach [32]. This combined instrumental variable weighted subspace fitting, or IV-WSF, can be viewed as an extension to the approach discussed above. Here the regressor vector $\varphi(t)$ in (5.3) corresponds to a model order p such that $p \geq n$. Let the null vector of dimension $p - n$ be denoted by η and $\theta_p = [1 \; a^T \; \eta^T \; b^T \; \eta^T]^T$. Denote the first column of $p - n + 1$ dimensional identity matrix by ζ. Consider the Toeplitz matrix $G(\theta)$ having θ_p as its first column and ζ^T as its first row. Then it can be shown that (5.9) gets modified to

  $$R_{z\overline{\varphi}} G(\theta) = 0. \tag{5.22}$$

  An approach to exploit (5.22) is to fit $G(\theta)$ to span the null space of $R_{z\overline{\varphi}}$. Let the columns of $\tilde{G}$ be the right singular vectors corresponding to the least

$p-n+1$ singular values of $\widehat{R_{z\overline{\varphi}}}$. Then one way to determine $\hat{\theta}$ is to minimise the distance between $G(\hat{\theta})$ and the span of $\tilde{G}$, i.e.

$$\hat{\theta} = \arg\min_{\theta}\left(\min_{C}\frac{1}{2}\left[\mathrm{vec}(G(\theta)-\tilde{G}C)\right]^T W\left[\mathrm{vec}(G(\theta)-\tilde{G}C)\right]\right) \tag{5.23}$$

where C is a nonsingular $(p-n+1)\times(p-n+1)$ matrix, W is a positive definite weighting matrix and may be data dependent, and $\mathrm{vec}(\cdot)$ denotes the vectorization operation. This can be shown to be equivalent to (see [32] for details)

$$\hat{\theta} = \arg\min_{\theta}\frac{1}{2}\left[\mathrm{vec}(S^T G(\theta))\right]^T Q\left[\mathrm{vec}(S^T G(\theta))\right]. \tag{5.24}$$

where Q is a positive definite weighting matrix and may be data dependent, and the columns of S are the right singular vectors corresponding to the largest $p+n$ singular values of $\widehat{R_{z\overline{\varphi}}}$. For optimal selection of the weighting matrix Q and the corresponding implementational aspects we refer to [32]. The optimal weight determination is computationally expensive. The details of the computationally economical sub-optimal IV-WSF can be found in [9].

- $R_{z\overline{\varphi}}$ has to be full rank for the IV based estimates to exist. This is a persistence-of-excitation like condition to be satisfied by the noise-free input signal.

- IV based subspace algorithms employing the state space models are proposed in [10]. These algorithms can be applied to MIMO systems operating in open or closed loop, where one has to account for the process noise also.

6 Bias-compensating schemes

Now return to (5.11) but substitute $z(t)$ by $\varphi(t)$. Set further

$$r = -E\varphi(t)y(t)\ . \tag{6.1}$$

Now, (5.9) will no longer hold, but instead

$$\begin{aligned} R_{\varphi\overline{\varphi}}\overline{\theta}_o &= r + R_{\varphi\varphi}\theta_o \\ &= r + [R_{\varphi_o\varphi_o} + R_{\tilde{\varphi}\tilde{\varphi}}]\theta_o \\ &= R_{\tilde{\varphi}\tilde{\varphi}}\theta_o\ . \end{aligned} \tag{6.2}$$

The matrix

$$R_{\tilde{\varphi}\tilde{\varphi}} = \begin{pmatrix} \lambda_y I & 0 \\ 0 & \lambda_u I \end{pmatrix} \tag{6.3}$$

in these normal equations will cause a bias in the least squares estimator

$$\hat{\theta}_{LS} = -\hat{R}_{\varphi\varphi}^{-1}\hat{r}\ . \tag{6.4}$$

Assuming the noise variances are known or estimated, a bias-compensated least squares scheme can be constructed as

$$\hat{\theta}_{BCLS} = \left[\hat{R}_{\varphi\varphi} - \begin{pmatrix} \hat{\lambda}_y I & 0 \\ 0 & \hat{\lambda}_u I \end{pmatrix} \right]^{-1} \hat{r} \ . \tag{6.5}$$

In the case of no input noise, it is fairly straightforward to devise an estimator for λ_y, using a TLS scheme, see [33].

The Frisch scheme can be viewed as a more advanced bias-compensating scheme. Note that due to the property (5.5) the matrix

$$R_{\overline{\varphi}_o\overline{\varphi}_o} = E\overline{\varphi}_o(t)\overline{\varphi}_o^T(t) \tag{6.6}$$

is singular (more specifically, positive semidefinite with one eigenvalue equal to zero). Now let $\hat{\lambda}_u$ be some estimate of λ_u. The estimated output noise variance $\hat{\lambda}_y$ is then constructed so that

$$\hat{R}_{\overline{\varphi}\overline{\varphi}} - \begin{pmatrix} \hat{\lambda}_y I & 0 \\ 0 & \hat{\lambda}_u I \end{pmatrix} \tag{6.7}$$

is singular. Hence $\hat{\lambda}_y$ will be an implicit function of $\hat{\lambda}_u$, [6]. The procedure can be repeated with the model order n increased by one unit. In case $\hat{R}_{\overline{\varphi}\overline{\varphi}}$ would be replaced by its true value $R_{\overline{\varphi}\overline{\varphi}}$ we would then get a situation as displayed in Figure 3.

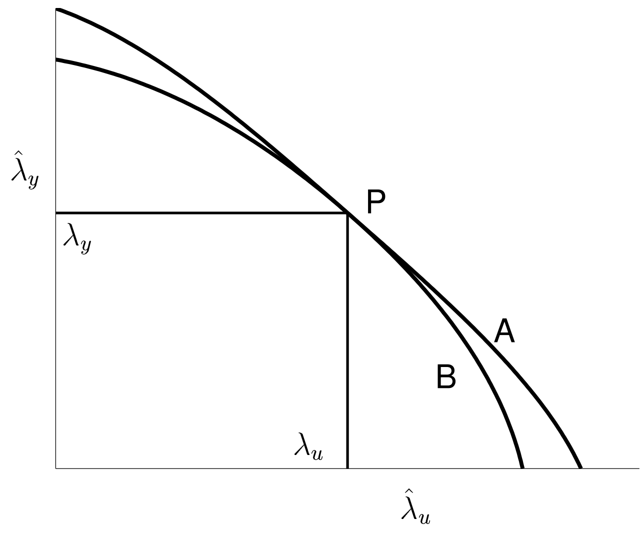

Figure 3: Illustration of the principle for Frisch estimation.

In Figure 3 curve B corresponds to the true model order, while curve A applies for the increased model order. The coordinates of the common point P give precisely the true noise variances λ_u, λ_y.

For a finite data set the situation is less ideal, and there is not a distinct point P where the curves A and B share a common point. We refer to [6], [31] for more detailed aspects on how the Frisch scheme can be implemented.

Comments on early work in the field:

- 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 [18] and Levin [19] (see also [4]).

  In more recent years the Koopmans–Levin method has been re-proposed by several authors with slightly different variants [7], [11], [37].

  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. These restrictive assumptions drastically reduce the practical use of the method.

- The Frisch scheme approach takes its name from a classical algebraic estimation problem, see [13], which was extended to the dynamic case in in [6]. In this work it was developed a procedure for identifying linear dynamic systems in presence of additive and uncorrelated white noises with unknown (and possibly different) variances.

- The compensated-bias least-squares techniques [33], based on suitable prefiltering elaborations of the available data, are even more flexible since they can work under more general noise conditions than those of assumption **A2**. The appealing properties of these approaches have been deeply investigated in recent years and several recursive and off–line algorithms have been proposed [40], [41], [42]. The BELS method is such a bias-compensation approach, where the noise variances and the parameter vector are estimated iteratively. Let us consider the model with over-parameterised numerator polynomial. Define

$$\theta^* = [a^{*T} \; b^{*T}]^T, \tag{6.8}$$

  where $a^* = (a_1 \dots a_n \; 0)^T$ and $b^* = (b_1 \dots b_n)^T$, the corresponding regressor vector being

$$\varphi^*(t) = (-y(t-1) \dots - y(t-n-1) \; u(t-1) \dots u(t-n))^T. \tag{6.9}$$

  It is possible to form the biased least squares estimate $\hat{\theta}^*_{LS}$. It follows from (6.1) - (6.4), with necessary modifications in notations,

$$\lim_{N\to\infty} \hat{\theta}^*_{LS} = \theta^*_o - R^{-1}_{\varphi^*\varphi^*} R_{\tilde{\varphi}^*\tilde{\varphi}^*} \theta^*_o. \tag{6.10}$$

  One can also establish that (refer [38] for details),

$$\lim_{N\to\infty} \frac{1}{N} \sum_{t=1}^{N} \left[y(t) - \varphi^{*T}(t) \hat{\theta}^*_{LS} \right]^2 = \lim_{N\to\infty} \left[\lambda_y \left(1 + a^{*T} \hat{a}^*_{LS} \right) + \lambda_u b^{*T} \hat{b}^*_{LS} \right] \tag{6.11}$$

$$-\lim_{N\to\infty} h^T\hat{a}^*_{LS} = \lambda_y h^T R_{11} a^* + \lambda_u h^T R_{12} b^* \tag{6.12}$$

where we introduced the $(n+1)$ dimensional vector $h = [0\ldots 0\ 1]^T$. The matrix R_{11} and R_{12} are $(n+1)\times(n+1)$ and $(n+1)\times n$, respectively, and formed by the partition of the matrix $R^{-1}_{\varphi^*\varphi^*}$ as follows,

$$R^{-1}_{\varphi^*\varphi^*} = \begin{pmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{pmatrix}. \tag{6.13}$$

The iteration is initialized using $\hat{\theta}^*_o = \hat{\theta}^*_{LS}$ and one can solve approximately for λ_u and λ_y using (6.11) and (6.12). These values of λ_u and λ_y are used in (6.10) to get a new estimate of $\hat{\theta}^*_o$. The steps are repeated until convergence. Similar procedure can be derived by overparameterising the numerator polynomial also. The method can be extended for coloured output noise. We refer to [38], [39] for a detailed analysis.

7 Frequency domain estimators

The idea here is similar to that of the Frisch scheme, with the important difference that we use the full spectral information in the data and not only the covariance matrix (5.10).

To be more specific, introduce the notation

$$z(t) = \begin{pmatrix} u(t) \\ y(t) \end{pmatrix}, \qquad z_o(t) = \begin{pmatrix} u_o(t) \\ y_o(t) \end{pmatrix} \tag{7.1}$$

for the input-output data. The spectral density of the input-output data satisfies

$$\phi_z(\omega) = \begin{pmatrix} 1 \\ G(e^{i\omega}) \end{pmatrix} \begin{pmatrix} 1 & G^*(e^{i\omega}) \end{pmatrix} \phi_{u_o}(\omega) + \begin{pmatrix} \lambda_y & 0 \\ 0 & \lambda_u \end{pmatrix}, \qquad \forall\omega \tag{7.2}$$

If the spectral density matrix ϕ_z is known, and an appropriate diagonal matrix is subtracted, then one would get a rank 1 matrix, corresponding to the first term of (7.2), for all frequencies ω. In case the decomposition as in (7.2) can be carried out, the first term would easily lead to estimates of the transfer function $G(e^{i\omega})$ and the true input spectrum $\phi_{u_o}(\omega)$.

The above idea is developed in [5] using an FFT-based spectral estimator. The quality of the estimates is moderate. A more promising approach, based on a black-box modelling of the spectral density ϕ_z is described in [28].

Another frequency domain identification technique which is employing periodic noise-free input signal is proposed in [15]. It is essential for the method to work that the signal $u_o(t)$ is periodic. The discrete Fourier Transform based technique uses the fact that the DFT coefficients of the time domain measurement noise, are complex-normally distributed and independent over frequencies. First, for each period of the observed periodic data, a maximum likelihood estimate of the parameter vector is obtained. This is followed by the averaging of the corresponding estimates of the frequency response functions. This method gives biased parameter estimates because the ML estimation step uses estimated variances of the noises. To improve upon the performance of the estimators a 'logarithmic averaging' is used over the periods. This approach yields practically consistent estimates.

8 Maximum likelihood and prediction error approach

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 [26]. Of crucial importance for this approach is the assumption **A2**, 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

$$u_o(t) = \frac{C(q^{-1})}{D(q^{-1})}\; e(t) \ , \tag{8.1}$$

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

$$\begin{aligned} C(q^{-1}) &= 1 + c_1\, q^{-1} + \ldots + c_p\, q^{-p} \\ D(q^{-1}) &= 1 + d_1\, q^{-1} + \ldots + d_p\, q^{-p} \end{aligned} \ , \tag{8.2}$$

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

$$\begin{aligned} u(t) &= u_o(t) + \tilde{u}(t) \\ y(t) &= G(q^{-1})\, u_o(t) + \tilde{y}(t) \\ u_o(t) &= C(q^{-1})/D(q^{-1})\ e(t) \end{aligned} \tag{8.3}$$

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 $\tilde{y}(t)$:

$$\begin{pmatrix} u(t) \\ y(t) \end{pmatrix} = \begin{pmatrix} \frac{C(q^{-1})}{D(q^{-1})} & 1 & 0 \\ \frac{B(q^{-1})C(q^{-1})}{A(q^{-1})D(q^{-1})} & 0 & 1 \end{pmatrix} \begin{pmatrix} e(t) \\ \tilde{u}(t) \\ \tilde{y}(t) \end{pmatrix} \ . \tag{8.4}$$

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 (8.4) can be converted into the innovations form

$$z(t) = S(q^{-1})\, \varepsilon(t) \ , \tag{8.5}$$

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 (8.5) from (8.4) is to convert the model (8.4) into a state space form. Then, the innovations form (8.5) can be computed from the Kalman filter after solving a Riccati equation. Since the innovations $\varepsilon(t)$

of model (8.5) 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

$$J(\theta) = \det\left(\frac{1}{N}\sum_{t=1}^{N} \varepsilon(t)\,\varepsilon^T(t)\right) , \tag{8.6}$$

where N is the number of input–output samples.

Comments and experiences:

- As it always happens for identification of multivariable systems, the choice of a suitable parameterization of the model is a crucial point of the approach.
- In [25, 26] it was proved that this method yields consistent parameter estimates under quite mild conditions.
- Its main disadvantage is that the numerical optimization procedure is, in general, quite involved since at every step a Riccati equation must be solved in order to find the innovations $\varepsilon(t)$ used in (8.6).
- The procedure may fail to give good results when only poor initial parameter estimates are available.
- Even if this technique can work with rather arbitrarily correlated noises, an extended system representation by including also models for the noises is needed.

9 Numerical illustration and concluding discussion

Some comparisons of different methods for identifying errors-in-variables models have appeared in the literature, for example in [26], [31]. Here we present some additional illustrations of three methods:

- The instrumental variable estimator (5.13), with delayed inputs in the instrumental vector $z(t)$.
- The Frisch scheme presented in Section 6.
- The prediction error method presented in Section 8.

The focus will be on comparing the computational load and the achieved model accuracy.

For this purpose two systems will be used for numerical simulations, namely

$$
\begin{aligned}
\mathcal{S}_1 \quad & G(q^{-1}) = \tfrac{1.0q^{-1}}{1-0.8q^{-1}}, \qquad u_o(t) = \tfrac{1}{1-0.5q^{-1}} e(t) \ , \\
& \lambda_u = 1, \quad \lambda_y = 1, \quad \lambda_e = 1 \ , \\
\mathcal{S}_2 \quad & G(q^{-1}) = \tfrac{1.0q^{-1}+0.5q^{-2}}{1-1.5q^{-1}+0.7q^{-2}}, \qquad u_o(t) = \tfrac{1}{1-0.9q^{-1}} e(t) \ , \\
& \lambda_u = 1, \quad \lambda_y = 1, \quad \lambda_e = 1 \ .
\end{aligned}
\tag{9.1}
$$

Example 9.1. We compare the computational load by measuring the number of Matlab flops. This value should not be regarded as a very exact measure of the computational complexity, as it is influenced by the implementation and coding. Nevertheless, it will give the correct order of magnitude. It can be noted that all three methods estimate the coefficients of $A(q^{-1})$ and $B(q^{-1})$. In addition, the Frisch scheme also gives estimate two more parameters (λ_u and λ_y), and PEM another two (for modelling $u_o(t)$). Hence the total number of parameters in the models will be as follows.

	IV	**Frisch**	**PEM**
$\mathcal{S}_1$	2	4	6
$\mathcal{S}_2$	4	6	8

The mean values of the number of Matlab flops from 10 realisations, each with $N = 200$ data points, are presented in Figure 4.

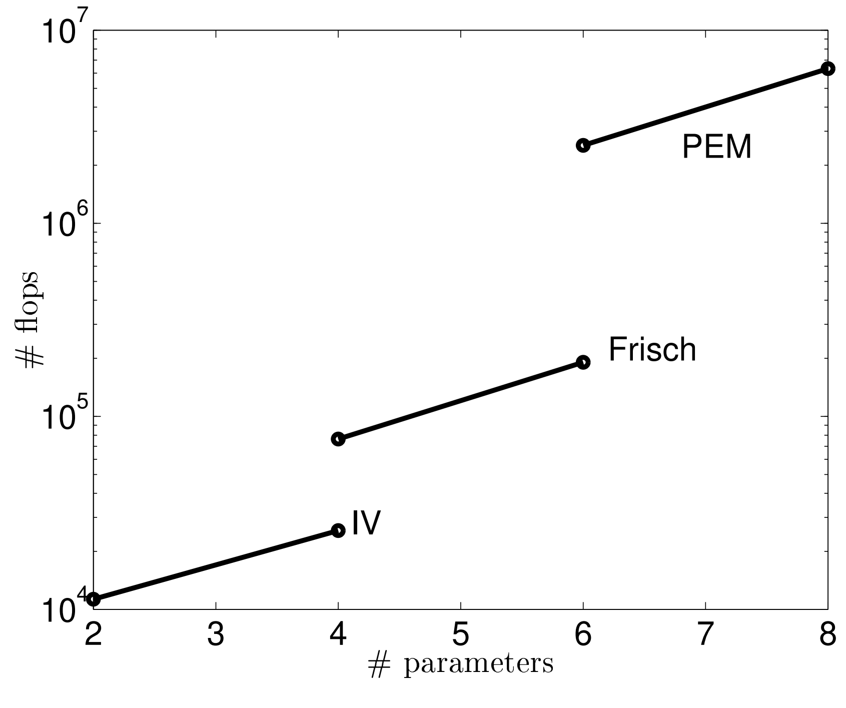

Figure 4: Illustration of the computational loads.

It is apparent that PEM is considerably more computationally demanding than the other two methods. This is fairly natural, as it involves a nonlinear optimisation problem, where each function evaluation requires a considerable amount of calculations. ■

Example 9.2. In this example the achieved model accuracy is studied. For this purpose the system $\mathcal{S}_2$, (9.1), is used with different values of the input noise variance λ_u. The model accuracy is expressed using the scalar measure

$$\text{NRMSE} = \frac{1}{\| \hat{\theta} \|} \sqrt{\frac{1}{M} \sum_{i=1}^{M} \| \hat{\theta}_i - \theta \|^2} \tag{9.2}$$

Here, $\hat{\theta}_i$ is parameter estimate in realisation i, and M is the number of realisations. In the numerical study, $M = 20,\ N = 200$. The findings are given in Figure 5.

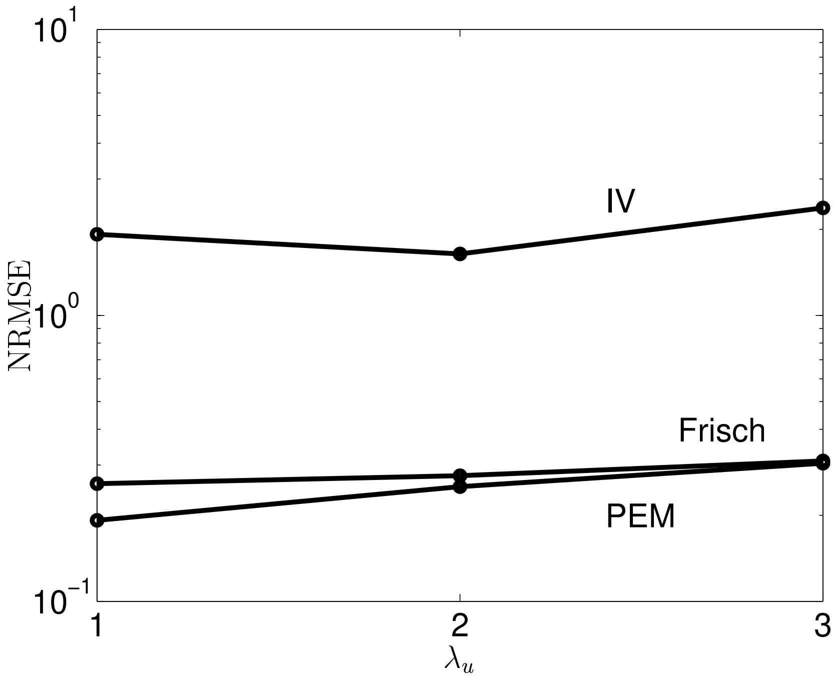

Figure 5: Illustration of the estimation accuracy.

One can see that PEM gives best accuracy, as expected (it should achieve the Cramér-Rao lower bound when N goes to infinity). Further, one finds that Frisch gives somewhat weaker accuracy, while the instrumental variable approach has the worst performance. ■

Appendix

A Gradient of the loss function (3.5)

We write the loss function (3.7) as

$$V(\theta) = V_1(\theta) + V_2(\theta) \tag{A.1}$$

where

$$\begin{aligned} V_1(\theta) &= E\left[\frac{\hat{A}(q^{-1})B(q^{-1}) - A(q^{-1})\hat{B}(q^{-1})}{A(q^{-1})\hat{C}(q^{-1})}u_o(t)\right]^2 && \text{(A.2)} \\ V_2(\theta) &= E\left[\frac{\hat{A}(q^{-1})}{\hat{C}(q^{-1})}\tilde{y}(t)\right]^2 + E\left[\frac{\hat{B}(q^{-1})}{\hat{C}(q^{-1})}\tilde{u}(t)\right]^2 && \text{(A.3)} \end{aligned}$$

As $V_1(\theta) \geq V_1(\theta_o) = 0$, it follows that $V_1(\theta)$ has a minimum for the true parameter vector θ_o, and hence

$$V_1'(\theta_o) = 0 \tag{A.4}$$

We then focus on evaluating the gradient of $V_2(\theta)$ at the true parameter vector. Let j be an integer satisfying $1 \leq j \leq n$. Then

$$\begin{aligned} \frac{\partial V_2(\theta)}{\partial a_j} &= 2E\left[\frac{A(q^{-1})}{C(q^{-1})}\tilde{y}(t)\right]\left[\frac{q^{-j}}{C(q^{-1})}\tilde{y}(t)\right] && \text{(A.5)} \\ \frac{\partial V_2(\theta)}{\partial b_j} &= 2E\left[\frac{B(q^{-1})}{C(q^{-1})}\tilde{u}(t)\right]\left[\frac{q^{-j}}{C(q^{-1})}\tilde{u}(t)\right] && \text{(A.6)} \\ \frac{\partial V_2(\theta)}{\partial c_j} &= -2E\left[\frac{A(q^{-1})}{C(q^{-1})}\tilde{y}(t)\right]\left[\frac{q^{-j}A(q^{-1})}{C^2(q^{-1})}\tilde{y}(t)\right] - 2E\left[\frac{B(q^{-1})}{C(q^{-1})}\tilde{u}(t)\right]\left[\frac{q^{-j}B(q^{-1})}{C^2(q^{-1})}\tilde{u}(t)\right] \\ &= -2\frac{1}{2\pi i}\oint\left[\frac{A(z)}{C(z)}\frac{z^{-j}A(z^{-1})}{C^2(z^{-1})}\lambda_y + \frac{B(z)}{C(z)}\frac{z^{-j}B(z^{-1})}{C^2(z^{-1})}\lambda_u\right]\frac{dz}{z} \\ &= -2\frac{1}{2\pi i}\oint\frac{z^{-j}}{C(z^{-1})}\frac{A(z)A(z^{-1})\lambda_y + B(z)B(z^{-1})\lambda_u}{C(z)C(z^{-1})}\frac{dz}{z} \\ &= -2\frac{1}{2\pi i}\oint\frac{z^{-j}}{C(z^{-1})}\sigma\frac{dz}{z} = \frac{2\sigma}{2\pi i}\oint\frac{z^{j-1}}{C(z)}dz = 0 && \text{(A.7)} \end{aligned}$$

The last equality in (A.7) follows as the integral has no poles inside the unit circle. On the other hand, the elements in (A.5), (A.6) are in general nonzero. As an illustration, particularly in the case of $n = 1$ we get

$$\begin{aligned} \frac{\partial V_2(\theta)}{\partial a} &= 2E\left[\frac{1+aq^{-1}}{1+cq^{-1}}\tilde{y}(t)\right]\left[\frac{q^{-1}}{1+cq^{-1}}\tilde{y}(t)\right] = \frac{2\lambda_y}{1-c^2}(a-c) \neq 0\ , \\ \frac{\partial V_2(\theta)}{\partial b} &= 2E\left[\frac{bq^{-1}}{1+cq^{-1}}\tilde{u}(t)\right]\left[\frac{q^{-1}}{1+cq^{-1}}\tilde{u}(t)\right] = \frac{2\lambda_u}{1-c^2}b \neq 0\ . \end{aligned}$$

B The concentrated loss function

The concentrated loss function (3.16) becomes, after inserting (3.15) into (3.14)

$$\begin{aligned}
V(\theta) &= \min_{U_o} L(\theta, U_o) \\
&= \frac{1}{2\lambda_y}\sum_{t=1}^{N}\left[\frac{\lambda_y(y(t)-bu(t))}{b^2\lambda_u+\lambda_y}\right]^2 \\
&\quad + \frac{1}{2\lambda_u}\sum_{t=1}^{N}\left[\frac{b\lambda_u(y(t)-bu(t))}{b^2\lambda_u+\lambda_y}\right]^2 + \frac{N}{2}\log\lambda_y + \frac{N}{2}\log\lambda_u \\
&= \frac{N}{2}\left[\frac{1}{(b^2\lambda_u+\lambda_y)}\left(\hat{R}_y + b^2\hat{R}_u - 2b\hat{R}_{yu}\right) + \log\lambda_y + \log\lambda_u\right] \qquad \text{(B.1)}
\end{aligned}$$

where

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

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

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

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, (B.3) 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, setting the gradient of $V(\theta)$ to zero gives:

$$\begin{aligned}
V_b: \quad & (b^2\lambda_u+\lambda_y)(2b\hat{R}_u - 2\hat{R}_{yu}) - 2b\lambda_u(\hat{R}_y + b^2\hat{R}_u - 2b\hat{R}_{yu}) = 0 \\
& b^2\lambda_u\hat{R}_{yu} + b(\lambda_y\hat{R}_u - \lambda_u\hat{R}_y) - \lambda_y\hat{R}_{yu} = 0\ . \qquad \text{(B.4)} \\
V_{\lambda_u}: \quad & -\frac{b^2}{(b^2\lambda_u+\lambda_y)^2}(\hat{R}_y + b^2\hat{R}_u - 2b\hat{R}_{yu}) + \frac{1}{\lambda_u} = 0\ . \qquad \text{(B.5)} \\
V_{\lambda_y}: \quad & -\frac{1}{(b^2\lambda_u+\lambda_y)^2}(\hat{R}_y + b^2\hat{R}_u - 2b\hat{R}_{yu}) + \frac{1}{\lambda_y} = 0\ . \qquad \text{(B.6)}
\end{aligned}$$

Dividing (B.5) by (B.6) gives

$$\lambda_y = b^2\lambda_u\ . \qquad \text{(B.7)}$$

Inserting (B.7) into (B.4) gives

$$b^2\lambda_u\hat{R}_{yu} + b\lambda_u(b^2\hat{R}_u - \hat{R}_y) - b^2\lambda_u\hat{R}_{yu} = 0$$

leading to (except for the degenerate solutions $b = 0$ and $\lambda_u = 0$)

$$b = \pm\sqrt{\hat{R}_y/\hat{R}_u}\ . \qquad \text{(B.8)}$$

This clearly shows the lack of uniqueness, and hence of loss of identifiability. In case we would like to characterize the complete solution anyway, inserting (B.7) and (B.8) into (B.5) gives

$$-\frac{b^2}{4b^4\lambda_u^2}(\hat{R}_y + b^2\hat{R}_u - 2b\hat{R}_{yu}) + \frac{1}{\lambda_u} = 0 \ ,$$

leading to

$$\lambda_u = \frac{1}{2b^2}(\hat{R}_y - b\hat{R}_{yu}) = \frac{1}{2}\hat{R}_u \mp \frac{1}{2}\sqrt{\frac{\hat{R}_u}{\hat{R}_y}}\hat{R}_{yu} \ , \tag{B.9}$$

and finally

$$\lambda_y = \frac{\hat{R}_y}{\hat{R}_u}\lambda_u = \frac{1}{2}\hat{R}_y \mp \frac{1}{2}\sqrt{\frac{\hat{R}_y}{\hat{R}_u}}\hat{R}_{yu} \ . \tag{B.10}$$

It is obvious that the estimates (B.8), (B.8), (B.10) differ from the true values.

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] S. Beghelli, P. Castaldi, and U. Soverini. A frequential approach for errors-in-variables models. In *Proc. European Control Conference, ECC '97*, Brussels, Belgium, July 1–4 1997.

[6] S. Beghelli, R.P. Guidorzi, and U. Soverini. The Frisch scheme in dynamic system identification. *Automatica*, 26:171–176, 1990.

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

[8] P. Castaldi and U. Soverini. Identification of dynamic errors-in-variables models. *Automatica*, 32(4):631–636, April 1996.

[9] M Cedervall and P Stoica. System identification from noisy measurements by using instrumental variables and subspace fitting. *Circuits, Systems and Signal Processing*, 15:2:275–290, 1996.

[10] C. T. Chou and M. Verhaegen. Subspace algorithms for the identification of multivariable dynamic errors-in-variables models. *Automatica*, 33(10):1857–1869, October 1997.

[11] 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.

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

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

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

[15] P Guillaume, R Pintelon, and J Schoukens. Robust parametric transfer function estimaton using complex logarithmic frequency response data. *IEEE Transactions on Automatic Control*, 40:1180–1190, 1995.

[16] 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.

[17] 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.

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

[19] 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.

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

[21] L. Ljung and T. Söderström. *Theory and Practice of Recursive Identification.* MIT Press, Cambridge, MA, USA, 1983.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

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

[27] T. Söderström and K. Mahata. On subspace fitting and instrumental variable approaches for identification of error-in-variables models. 2001. Submitted for publication.

[28] T. Söderström, K. Mahata, and U. Soverini. A frequency domain Frisch scheme for dynamic errors-in-variables identification. 2001. In preparation.

[29] T. Söderström and P. Stoica. *Instrumental Variable Methods for System Identification.* Springer-Verlag, Berlin, 1983.

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

[31] 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.

[32] P Stoica, M Cedervall, and A Eriksson. Combined instrumental variable and subspace fitting approach to parameter estimation of noisy input-output systems. *IEEE Transactions on Signal Processing*, 43:2386–2397, 1995.

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

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

[35] 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.

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

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

[38] W. X. Zheng. Transfer function estimation form noisy input and output data. *International Journal of Adaptive Control and Signal Processing*, 12:365–380, 1998.

[39] W. X. Zheng. An efficient algorithm for stochastic system identification with noisy input. In *Proc. 38th IEEE Conference on Decision and Control*, pages 3657–3662, Phoenix, AZ, December 1999.

[40] 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.

[41] 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.

[42] 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.