periodica polytechnica

Electrical Engineering
52/3-4 (2008) 127–135
doi: 10.3311/pp.ee.2008-3-4.01
web: http://www.pp.bme.hu/ee

RESEARCH ARTICLE

An identification approach to dynamic errors-in-variables systems with a preliminary clustering of observations

Levente Hunyadi / *István* Vajk

Received 2008-07-22

Abstract

Errors-in-variables models are statistical models in which not only dependent but also independent variables are observed with error, i.e. they exhibit a symmetrical model structure in terms of noise. The application field for these models is diverse including computer vision, image reconstruction, speech and audio processing, signal processing, modal and spectral analysis, system identification, econometrics and time series analysis. This paper explores applying the errors-in-variables approach to parameter estimation of discrete-time dynamic linear systems. In particular, a framework is introduced in which a preliminary separation step is applied to group observations prior to parameter estimation. As a result, instead of one, two sets of estimates are derived simultaneously, comparing which can yield estimates for noise parameters. The proposed approach is compared to other schemes with simulation examples.

Keywords

errors-in-variables · model and noise parameter estimation · data separation · principal components analysis

Acknowledgement

This work has been supported by the fund of the Hungarian Academy of Sciences for control research and the Hungarian National Research Fund (grant number T68370).

Levente Hunyadi

Department of Automation and Applied Informatics, BME, H-1111 Budapest Goldmann Gy. t. 3., Hungary
e-mail: hunyadi@aut.bme.hu

István Vajk

Department of Electronics Technology, BME, H-1111 Budapest Goldmann Gy. t. 3., Hungary

1 Introduction

The task of system identification is to build mathematical models of a system based on available experimental data. A widely adopted assumption is that the dependent (or *output*) variables are observed with errors, whereas noise-free independent (or *input*) variables are available for modeling. However, this assumption may be violated in practical applications like computer vision, image reconstruction, control systems, speech, audio or signal processing, communications, econometrics and time series analysis where not only the system output but also the input is a measured set or series of quantities, hence observed with error. In fact, these applications put the focus on discovering, understanding or parameterizing the internal relationship between observed quantities rather than on predicting future outcome.

Errors-in-variables (EIV) systems may be *static*, in which case there is no coupling between observed variables, or *dynamic*, where a quantity at time t may depend on a finite number of past quantities. Fig. 1 depicts a dynamic single-input single-output (SISO) EIV configuration. Observe that only the noise-corrupted input and output sequences $u(t)$ and $y(t)$ are observable, the original noise-free sequences $u_0(t)$ and $y_0(t)$ are not, $t = 1, 2, \ldots N$, N denoting the number of observations. As far as the additive noise sequences $\tilde{u}(t)$ and $\tilde{y}(t)$ are concerned, in most cases, a white noise model is assumed, which corresponds to noise due to measurement error. Unlike in control theory, the noise sequence $\tilde{u}(t)$ is not fed through the system. Given this system model, the goal of system identification is to estimate *model* (in usual system identification terminology, *process*) as well as *noise parameters* using the observable noise-contaminated input and output data.

Provided that the ratio of input and output noise variances is a priori known, the task of deriving model and noise parameters is a classical system identification problem. In contrast, a situation where no such information is available is recognized as a more difficult one. In fact, it turns out that under general assumptions, the system is not identifiable, or put alternatively, it produces many equivalent results. In other words, restrictions are necessary for the identification to produce a unique result [1].

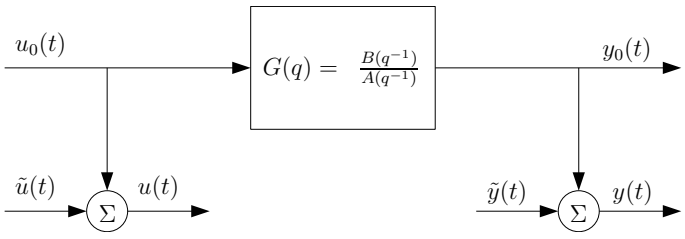

Fig. 1. The basic setup for a discrete-time dynamic errors-in-variables system.

The restriction we explore is that it is possible to *partition observations* into two not necessarily contiguous sets based on some varied noise parameter. The goal is to produce two dissimilar sets from the perspective of an estimation algorithm. In particular, if the initial assumption for the noise parameter is incorrect, the estimates over the two sets will differ significantly. On the other hand, with a correct noise parameter assumption, the estimates will be close to one another. As a result, it is possible to arrive at a correct noise parameter estimate by minimizing the difference between parameter estimates.

The rest of the paper is structured as follows. The general setup of a discrete-time dynamic linear errors-in-variables system is outlined in Section 2, which also introduces the notations used throughout this paper. Section 3 explores some inherent constraints of EIV systems, while Section 4 surveys related work. Next, Section 5 describes the generalized Koopmans–Levin algorithm, which we use for separated observations to derive parameter estimates. Section 6 discusses the main idea of the paper, that is, the data separation methods and the metrics using which estimates are compared. Finally, Section 7 illustrates the feasibility of the outlined approach with some comparative simulation results before concluding with Section 8, which summarizes the key points of the paper.

2 Setup and notations

Consider the SISO errors-in-variables system in Fig. 1. As the system $G(q^{-1})$ is linear, it is described by the linear autoregressive moving average (ARMA) difference equation

$$A(q^{-1})y_0(t) = B(q^{-1})u_0(t) \tag{1}$$

where q^{-1} denotes the backward shift operator such that $q^{-1}u(t) = u(t-1)$ and

$$\begin{aligned} A(q^{-1}) &= 1 + a_1 q^{-1} + \cdots + a_{ma} q^{-ma} \\ B(q^{-1}) &= b_1 q^{-1} + \cdots + b_{mb} q^{-mb}. \end{aligned}$$

Given the aforementioned system description, we may now introduce the model parameter vector θ as well as its autoregressive and moving average components, θ^a and θ^b, respectively:

$$\begin{aligned} \theta &= [\ a_1 \ \ldots \ a_{ma} \ -b_1 \ \ldots \ -b_{mb}\ ]^\top \\ \theta^a &= [\ a_1 \ \ldots \ a_{ma}\ ]^\top \\ \theta^b &= [\ b_1 \ \ldots \ b_{mb}\ ]^\top \end{aligned}$$

whose estimates are denoted by $\hat{\theta}$ and whose true values by θ_0. In general, the notation $\hat{p}$ and p_0 will also be applied to other parameters to indicate the estimated and the true value, respectively.

Similarly, the regressor vector $\varphi(t)$ may be introduced as

$$\begin{aligned} \varphi(t) &= [\ \varphi_y^\top(t) \ \ \varphi_u^\top(t)\ ]^\top \\ \varphi_y(t) &= [\ y(t-1) \ \ldots \ y(t-ma)\ ]^\top \\ \varphi_u(t) &= [\ u(t-1) \ \ldots \ u(t-mb)\ ]^\top \end{aligned}$$

hence the system description in (1) can be recast in the compact linear regression form

$$y(t) = \varphi^\top(t)\theta + \varepsilon(t) \tag{2}$$

where ε is a stochastic disturbance term $\varepsilon(t) = \tilde{y}(t) - \tilde{\varphi}^\top(t)\theta_0$ in which $\tilde{\varphi}$ is the noise contribution of the regressor vector.

Without loss of generality, we may assume that $m = ma = mb$ (or, $m = \max(ma, mb)$), which allows us to use a symmetric model in terms of parameters.

For some approaches, it is preferable to exploit the symmetry of EIV models and use an implicit formula rather than the explicit formula (2). For this end, supplement the model parameters in θ with additional elements such that

$$g = \begin{bmatrix} a_0 & a_1 & \ldots & a_m & -b_0 & -b_1 & \ldots & -b_m \end{bmatrix}^\top$$

and write

$$x^\top(t)\, g = 0 \tag{3}$$

where

$$\begin{aligned} x(t) &= [\ x_y^\top(t) \ \ x_u^\top(t)\ ]^\top \\ x_y(t) &= [\ y(t) \ \ldots \ y(t-m)\ ]^\top \\ x_u(t) &= [\ u(t) \ \ldots \ u(t-m)\ ]^\top \end{aligned}$$

for $t = 0, \ldots, N-m$ where the implicit assumptions $a_0 = 1$, $b_0 = 0$ have been made to make (3) conform to (2).

In many cases, it is more practical to use matrix notation by collecting multiple observations into a large vector or matrix. Notations such as u or y refer to these N-row vectors, while Φ and X collect $N-m+1$ and $N-m$ observations of $\varphi(t)$ and $x(t)$, respectively:

$$\begin{aligned} u &= [\ u_1 \ \ u_2 \ \ldots \ u_N\ ]^\top \\ y &= [\ y_1 \ \ y_2 \ \ldots \ y_N\ ]^\top \\ \Phi &= \begin{bmatrix} y_m & \ldots & y_1 & u_m & \ldots & u_1 \\ y_{m+1} & \ldots & y_2 & u_{m+1} & \ldots & u_2 \\ \vdots & & \vdots & \vdots & & \vdots \\ y_N & \ldots & y_{N-m+1} & u_N & \ldots & u_{N-m+1} \end{bmatrix} \\ X &= \begin{bmatrix} y_{m+1} & \ldots & y_1 & u_{m+1} & \ldots & u_1 \\ y_{m+2} & \ldots & y_2 & u_{m+2} & \ldots & u_2 \\ \vdots & & \vdots & \vdots & & \vdots \\ y_N & \ldots & y_{N-m} & u_N & \ldots & u_{N-m} \end{bmatrix}. \end{aligned}$$

With matrix notation, (2) can be concisely written as an overdetermined system of equations

$$y = \Phi\theta + \varepsilon.$$

As far as noise assumptions are concerned, we will assume white noise for most identification algorithms. The covariance matrix of white input–output noise is parameterizable with two scalars: μ corresponding to noise magnitude, and ρ to noise "direction", such that

$$C = \begin{bmatrix} \sigma_y^2 & 0 \\ 0 & \sigma_u^2 \end{bmatrix} = \mu C_\rho = \mu \begin{bmatrix} \sin^2\rho & 0 \\ 0 & \cos^2\rho \end{bmatrix}. \quad (4)$$

Likewise, observations can be characterized with their sample covariance matrices. Define the sample covariance matrix and vector R_φ and $r_{\varphi y}$, as well as their estimates $\hat{R}_\varphi$ and $\hat{r}_{\varphi y}$, in a way that

$$\begin{aligned} R_\varphi &= \mathbb{E}\left(\varphi(t)\varphi^\top(t)\right) \\ r_{\varphi y} &= \mathbb{E}\left(\varphi(t)y(t)\right) \\ \hat{R}_\varphi &= \frac{1}{N}\sum_{t=1}^{N}\varphi(t)\varphi^\top(t) = \frac{1}{N}\Phi^\top\Phi \\ \hat{r}_{\varphi y} &= \frac{1}{N}\sum_{t=1}^{N}\varphi(t)y(t) = \frac{1}{N}\Phi^\top Y \end{aligned}$$

where $\hat{R}_\varphi$ and $\hat{r}_{\varphi y}$ are estimates for R_φ and $r_{\varphi y}$ from N samples. A similar covariance matrix may be introduced for the observation vector $x(t)$, given in the implicit form (3), which incorporates the covariance matrix for both $\varphi(t)$ and $y(t)$.

3 Identifiability aspects

Identification of errors-in-variables systems where no a priori knowledge of the noise ratio is available is not possible with the assumption of

- Gaussian white input sequence $u_0(t)$ and
- Gaussian white input and output noises $\tilde{u}(t)$ and $\tilde{y}(t)$,

or, in other words, by being constrained to using at most *second-order* characteristics, such as the autocorrelation function in the time domain or the power spectrum in the frequency domain. In fact, such problems lead to many indistinguishable solutions under these conditions. In order to make such systems uniquely identifiable, additional restrictions have to be imposed either on the noise-free input signal or the noise characteristics [11]:

- One option is to make *distributional* assumptions where the input (or noise) signal is supposed to satisfy some non-Gaussian (skewed) distribution. Higher-order statistics methods [13] exploit that either the noise-free input signal (or the noise) is non-Gaussian distributed and use third- or fourth-order statistics to identify the system.
- A second, equally feasible option is to make *structural* assumptions on the systems, e.g. to assume more detailed models for the noise-free input and the measurement noises, in particular, modeling them as ARMA processes. For instance, let the noise-free input signal be generated by an ARMA process

  $$D(q^{-1})u_0(t) = C(q^{-1})e_u(t)$$

  where $e_u(t)$ is a white noise sequence with unknown variance. This approach may lead to a unique decomposition of the observation spectrum into frequencies partly attributable to noise-free input and partly to measurement noise. An in-depth analysis of this approach is given in [1].
- A third option is to use *repeated experiments* in which either the input signal can be controlled or it changes characteristics at some point in time while noise properties remain the same throughout the experiment. Such a setup enables data to be arranged into disjoint but contiguous sets. With as many sets as unknown noise parameters, the system can, in principle, be identified.

4 Related work

There is extensive literature on the identification of parametric errors-in-variables systems, see [11] for a comprehensive survey. Methods aiming at simultaneously deriving process and noise parameters include instrumental variables [4, 12], bias-compensating least squares [7, 16], the Frisch scheme [3, 6], structured total least squares [9], frequency-domain [10], prediction error and efficient maximum likelihood [14] methods, which differ in the noise and experimental conditions they assume, the computational complexity they demand as well as the statistical accuracy they provide. Below we summarize the key points of some of these algorithms, which we subsequently compare to our proposed algorithm in Section 7.

4.1 Least squares

The least squares (LS) estimate known from statistical literature can then be formulated as

$$\hat{\theta}_{LS} = \Phi^\dagger y = (\Phi^\top\Phi)^{-1}\Phi^\top y \quad (5)$$

where $M^\dagger$ denotes the Moore–Penrose generalized inverse of M.

However, this identification method gives consistent estimates (i.e. the solution converges to the true parameter vector as $N \to \infty$) only under restrictive conditions, notably, when only the output observation is corrupted with noise. Reformulating (5) using covariance matrices yields

$$\hat{\theta}_{LS} = \hat{R}_\varphi^{-1}\hat{r}_{\varphi y}. \quad (6)$$

Assuming white noise on both input and output sequences, the covariance matrices may be decomposed into a model part and

a noise part such that

$$\begin{aligned} R_\varphi &= R_{\varphi 0} + R_{\tilde{\varphi}} \\ r_{\varphi y} &= r_{\varphi_0 y_0} + r_{\tilde{\varphi}\tilde{y}} = R_{\varphi_0 y_0}\theta_0 \end{aligned}$$

where $r_{\tilde{\varphi}\tilde{y}} = 0$ (the two sequences are not correlated) and $r_{\varphi_0 y_0} = R_{\varphi_0 y_0}\theta_0$ from (6), in which case,

$$R_\varphi \hat{\theta}_{LS} = (R_\varphi - R_{\tilde{\varphi}})\theta_0$$

thus $\hat{\theta}_{LS}$ is biased due to the term $R_{\tilde{\varphi}}$.

4.2 Bias-compensating least squares

The principle of bias compensated least squares (BCLS) methods is to adjust the LS estimate to eliminate the bias due to $R_{\tilde{\varphi}}$. Consequently,

$$\hat{\theta}_{BCLS} = (\hat{R}_\varphi - \hat{R}_{\tilde{\varphi}})^{-1}\hat{r}_{\varphi y} \tag{7}$$

in which the unknown $\hat{R}_{\tilde{\varphi}}$, which depends on the noise parameters σ_y^2 and σ_u^2, has to be estimated in some way.

It is clear that if the ratio of noise variances is unknown, (7) contains $2m+2$ unknowns but comprises of only $2m$ equations, one for each of the model parameters. Consequently, additional equations have to supplement the above set of equations. One relation can be obtained by using the minimum error of the least squares estimate:

$$V_{LS} = \min_\theta \mathbb{E}\left(y(t) - \varphi^\top(t)\theta\right)^2 = \sigma_y^2 + \hat{\theta}_{LS}^\top R_{\tilde{\varphi}}\theta_0. \tag{8}$$

In a practical scenario, the expected value is not known but is computed using the available samples as well as the current estimates for θ. This suggests that (unlike LS estimation) the compensated LS procedure is iterative.

In order to get a second extra equation, an extended model structure should be considered. A possible extension is appending an additional $-y(t-na-1)$ to the regressor vector $\varphi(t)$ and a corresponding a_{na+1} parameter to θ (whose true value is 0) and using the extended versions in the formulae of the original model in (7):

$$\begin{aligned} \bar{\theta} &= [\ -a_1 \quad \dots \quad -a_{na} \quad -a_{na+1} \quad b_1 \quad \dots \quad b_{nb}\ ] \\ \bar{\varphi}(t) &= [\ \bar{\varphi}_y^\top(t) \quad \bar{\varphi}_u^\top(t)\ ]^\top \\ \bar{\varphi}_y(t) &= [\ y(t-1) \quad \dots \quad y(t-ma) \quad y(t-ma-1)\ ]^\top \\ \bar{\varphi}_u(t) &= [\ u(t-1) \quad \dots \quad u(t-mb)\ ]^\top. \end{aligned}$$

These additional equations allow us to infer estimates for σ_u^2 and σ_y^2. Once these have been estimated, the bias of the least squares estimate is eliminated to achieve consistent estimates. In practice, these estimates are often rather crude, which can be significantly improved by augmenting multiple input or output parameters. As the number of equations in this case exceeds the number of unknowns, an overdetermined system of equations has to be solved in a least squares sense.

The iterative bias-compensating estimation algorithm is therefore as follows [16]:

1 Set the initial value of $\hat{\theta}_0$ to $\hat{\theta}_{LS}$ according to (5).

2 Solve (8) and the equation(s) corresponding to the extended model using the current parameter estimates $\hat{\theta}_k$ to get estimates for the noise elements $\hat{v}_{k+1} = [\ \hat{\sigma}_y^2 \quad \hat{\sigma}_u^2\ ]$.

3 Using (7), compute new parameter estimates $\hat{\theta}_{k+1}$ using $\hat{\theta}_k$ and $\hat{v}_{k+1}$, and repeat from step 2.

4.3 The Frisch scheme

The Frisch scheme provides a recursive algorithm strikingly similar to the BCLS approach so that many of its variants may be interpreted as a special form of BCLS, operating on similar extended models with comparable performance results, see [6]. It is based on the idea that the sample covariance matrix R_{x_0} of the true values of observations yields the zero vector when multiplied by the true parameter values. In other words, for the estimated quantities, it holds that

$$\hat{R}_{x_0}\hat{g} = (\hat{R}_x - \hat{R}_{\tilde{x}})\hat{g} = 0 \tag{9}$$

where we have used that $R_x = R_{x_0} + R_{\tilde{x}}$ where $\hat{R}_{\tilde{x}} = \hat{R}_{\tilde{x}}(\sigma_y^2,\ \sigma_u^2)$ is a(n estimated) covariance matrix corresponding to white noise on both y and u. Similarly to the BCLS case, we have more unknowns than equations in (9). However, assuming an estimate of σ_u^2 is available, σ_y^2 may be computed such that the difference matrix $\hat{R}_x - \hat{R}_{\tilde{x}}$ is singular. This is achieved by

$$\sigma_y^2 = \lambda_{min}\left(R_y - R_{yu}(R_u - \sigma_u^2 I_m)^{-1}R_{uy}\right) \tag{10}$$

where R_y and R_u denote the sample covariance matrices belonging to output- and input-related entries in x, respectively, and the λ_{min} operator denotes the minimum eigenvalue of the operand matrix.

In order to determine σ_u^2, one of the more robust approaches is to compute so-called residuals and compare their statistical properties to what can be predicted from the model [3]. A residual is defined as

$$\varepsilon(t) = A(q^{-1})y(t) - B(q^{-1})u(t).$$

Additionally, introduce the covariance vector belonging to shift $k \geqq 1$ as

$$r(k) = \mathbb{E}(w(t)w(t+k))$$

and its estimate from finite samples as

$$\hat{r}(k) = \frac{1}{N-k}\sum_1^{N-k} w(t)w(t+k)$$

The idea is to compute the sample covariance vector using $\varepsilon(t)$ where

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

in which $\hat{A}(q^{-1})$ and $\hat{B}(q^{-1})$ encapsulate current model parameter estimates, and compare it to a theoretical covariance vector, in which

$$\varepsilon_0(t) = \hat{A}(q^{-1})\hat{\tilde{y}}(t) - \hat{B}(q^{-1})\hat{\tilde{u}}(t)$$

and $\hat{\tilde{y}}(t)$ as well as $\hat{\tilde{u}}(t)$ are independent white noise sequences with variance as determined by the current estimates $\hat{\sigma}_y^2$ and $\hat{\sigma}_u^2$. The dimension of the covariance vector r (i.e. the maximum shift k) is a user-supplied parameter.

The entire algorithm runs as follows:

1 Assume an initial value for $\hat{\sigma}_u^2$.

2 Compute an estimate $\hat{\sigma}_y^2$ using (10).

3 Compute model parameters based on (9).

4 Determine the residuals $\varepsilon(t, \hat{\theta})$ using estimated model parameters in $\hat{A}(q^{-1})$ and $\hat{B}(q^{-1})$ as well as observed output and input sequences $y(t)$ and $u(t)$, and compute the related sample covariance vector $\hat{r}$.

5 Determine the theoretical reference covariance vector using residuals $\varepsilon_0(t)$ generated by estimated process parameters and white noise sequences $\hat{\tilde{y}}(t)$ and $\hat{\tilde{u}}(t)$ where

$$\begin{aligned} \mathbb{E}\left(\hat{\tilde{y}}(t)\right)^2 &= \hat{\sigma}_y^2 \\ \mathbb{E}\left(\hat{\tilde{u}}(t)\right)^2 &= \hat{\sigma}_u^2. \end{aligned}$$

Compare the sample and the reference covariance vectors by setting $V = \delta^\top W \delta$ where δ is a difference vector of covariances and W is a (user-chosen) weighing matrix.

6 Repeat from step 1 minimizing V.

4.4 Instrumental variables

Instrumental variables are a family of methods to give a quick estimate of model parameters without requiring an iterative approach. The idea is to choose an instrument vector $z(t)$, which is uncorrelated with the noise term $\varepsilon(t)$ in (2) and as correlated as possible with $\varphi(t)$. Which elements the vector $z(t)$ is to contain depends on the specific approach. The estimates are then computed as

$$\hat{\theta}_{IV} = \left(R_{z\varphi}^\top W R_{z\varphi}\right)^{-1} R_{z\varphi}^\top W r_{zy}$$

where W is a user-selected weighing matrix, often chosen as $W = I$. A possible arrangement [12] for the vector $z(t)$ is

$$\begin{aligned} z_y^\top(t) &= [\ y(t-1-dy) \quad \ldots \quad y(t-dy-my)\ ] \\ z_u^\top(t) &= [\ u(t-1-du) \quad \ldots \quad u(t-du-mu)\ ] \\ z^\top(t) &= [\ -z_y^\top(t) \quad z_u^\top(t)\ ] \end{aligned}$$

where $dy \geqq ma$ and $du \geqq mb$, while my and mu determine how many extra shifted y and u components to take.

4.5 Structured total least squares with data splitting

An estimation scheme described in [9] makes use of the repeated experiment approach to employ a structured total least squares method for system identification. They assume that the input sequence u_0 changes characteristics at a time instant t, leading to two sequences of data points that have a different mean and dispersion. The idea is to use the two sequences to determine the ratio of input and output noise, i.e. λ in the noise covariance matrix

$$C = \begin{bmatrix} \sigma_u^2 & 0 \\ 0 & \sigma_y^2 \end{bmatrix} = \mu \begin{bmatrix} \lambda & 0 \\ 0 & 1 \end{bmatrix}.$$

Once λ has been determined, the noise covariance matrix is known up to a scaling factor, and hence a structured version of the total least squares approach [5] can be used to uniquely identify the system.

Their algorithm thus proceeds as follows:

1 Determine the time instant t at which input characteristics change and thereby create two disjoint data sequences.

2 Solve a univariate optimization problem to estimate the noise covariance ratio λ. The optimization problem entails solving weighted total least squares problems simultaneously for both sequences and iteratively arriving at a solution by minimizing an appropriate cost function involving the identified parameters belonging to each respective sequence.

3 Identify the entire system with the estimated λ by means of the standard generalized total least squares method.

The cost function used in their paper is

$$\hat{\lambda} = \arg\min_{\lambda} \left((\mu_1 - \mu_2)^2 + c \sin^2\left(\angle(\theta_1, \theta_2)\right)\right)$$

where c is a scaling constant usually chosen as $c = 1$, $\angle$ denotes the angle enclosed by the parameter vectors, and μ_i and θ_i come from a structured total least squares problem. (Notice that both μ and θ are functions of λ.)

5 The Generalized Koopmans–Levin estimator

The Koopmans method for static systems, and its extension, the Koopmans–Levin (KL) method [8] for dynamic systems are classical methods that provide a simple non-iterative way to estimate model or process parameters but the estimation variance is fairly large. Meanwhile, the maximum likelihood (ML) estimation approach, the "best possible" estimator, is much more robust but is inherently iterative and hence entails a larger computational complexity. In this section, a generalized Koopmans–Levin (GKL) approach that unifies the KL and ML algorithms will be developed following [15]. The unified algorithm incorporates a scaling parameter q that allows us to freely trade estimation accuracy for efficiency.

Let us first introduce a generalized version of the observation matrix that has $N - q + 1$ rows and $2q$ columns as opposed to the original observation matrix that had $N - m + 1$ rows and $2m$ columns:

$$X_q = \begin{bmatrix} y_q & \cdots & y_1 & u_q & \cdots & u_1 \\ y_{q+1} & \cdots & y_2 & u_{q+1} & \cdots & u_2 \\ \vdots & & \vdots & \vdots & & \vdots \\ y_N & \cdots & y_{N-q+1} & u_N & \cdots & u_{N-q+1} \end{bmatrix} \tag{11}$$

such that $X_{m+1} = X$ (the latter in terms of notation introduced in Section 2).

Using the notations introduced in Section 2, the process parameters g of a system can be obtained using the KL algorithm by minimizing the loss function

$$J_{KL} = \frac{1}{2} \frac{g^\top \left(\sum_{t=m+1}^{N} x(t) x(t)^\top \right) g}{g^\top C_{KL} g}$$

where $C_{KL} = C_\rho \otimes I_{m+1}$. This can be rewritten in a more compact form as

$$J_{KL} = \frac{1}{2} \frac{g^\top X_{KL}^\top X_{KL} g}{g^\top C_{KL} g} \tag{12}$$

where $X_{KL} = X_{m+1}$.

A practical way to solve the minimization problem above is to consider the generalized eigenvalue-eigenvector problem

$$(X_{KL}^\top X_{KL} - \lambda C_{KL}) g = 0$$

where the optimal value g_{opt} will be equal to the eigenvector corresponding to the smallest eigenvalue. If C_{KL} is factorized as

$$C_{KL} = \bar{C}_{KL}^\top \bar{C}_{KL}$$

the optimization problem can be solved by means of generalized singular value decomposition (GSVD).

A similar loss function as in (12) may be derived for ML estimation. Define x_{ML} as a 1-by-$2N$ vector such that $x_{ML} = X_N$ and

$$G = \begin{bmatrix} G_a \\ -G_b \end{bmatrix}$$

in which G_a and G_b are banded Toeplitz matrices of parameters a_i and b_i such that $x_0 G = 0$ (assuming x_0 is the noise-free equivalent of x_{ML}):

$$G_a = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 & 0 \\ a_1 & 1 & 0 & \dots & 0 & 0 \\ a_2 & a_1 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ a_m & a_{m-1} & a_{m-2} & \dots & 0 & 0 \\ 0 & a_m & a_{m-1} & \dots & 0 & 0 \\ 0 & 0 & a_m & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & a_m & a_{m-1} \\ 0 & 0 & 0 & \dots & 0 & a_m \end{bmatrix}_{N, N-m}$$

and G_b can be constructed in a similar manner.

The likelihood function can then be formulated as

$$p(x_{ML} \mid g) \propto \exp\left(-\frac{1}{2} (x_{ML} - x_0)(C_{ML})^{-1} (x_{ML} - x_0) \right)$$

where C_{ML} is a large matrix $C_{ML} = C_\rho \otimes I_N$.

Taking the constraint $x_0 G = 0$ into account, maximizing the likelihood function is equivalent to minimizing the loss function

$$J_{ML} = \frac{1}{2} x_{ML} G (G^\top C_{ML} G)^{-1} G^\top x_{ML}^\top. \tag{13}$$

Comparing (12) and (13), a common form for these loss functions seems likely. Introducing $D_{KL} = X_{KL}^\top X_{KL}$, (12) can be reformulated as

$$J_{KL} = \frac{1}{2} \text{trace} \left((g^\top C_{KL} g)^{-1} g^\top D_{KL} g \right).$$

On the other hand,

$$g^\top D_{KL} g = g^\top X_{KL}^\top X_{KL} g = e_{KL}^\top e_{KL}$$

in which $e_{KL} = X_{KL} g$ (a column vector) represents the error.

Meanwhile, (13) can be similarly transformed using $D_{ML} = x_{ML}^\top x_{ML}$ to give

$$J_{ML} = \frac{1}{2} \text{trace} \left((G^\top C_{ML} G)^{-1} G^\top D_{ML} G \right)$$

where again

$$G^\top D_{ML} G = G^\top x_{ML}^\top x_{ML} G = e_{ML}^\top e_{ML}.$$

in which $e_{ML} = x_{ML} G$ (a row vector) is the error.

The striking similarity between the two loss functions leads us to a joint loss function that includes both the KL and the ML approach as a special case. Let us introduce the error matrix E_q such that $E_q = X_q G_q$ and $D_q = X_q^\top X_q$, where X_q is a $2q$-by-$N - q + 1$ matrix in which q is a scaling parameter between $m + 1$ (yielding Koopmans–Levin) and N (yielding maximum likelihood), and let G_q be a parameter matrix and C_q a noise covariance matrix of matching dimensions. The joint loss function thus has the form

$$J = \frac{1}{2} \text{trace} \left((G_q^\top C_q G_q)^{-1} G_q^\top D_q G_q \right). \tag{14}$$

6 Estimation with preliminary clustering

We have previously mentioned that it is not possible to uniquely identify an errors-in-variables system in the absence of a priori information on the noise ratio without imposing restrictions on the system or the experimental setup. Indeed, the outlined methods have made such implicit assumptions by incorporating covariance matrices that have to be invertible (thereby supposing a sufficient excitation) or by requiring that noise properties remain the same while the input changes at some point. The restriction we explore here is that observations (possibly after subject to some transformation) are *separable* into groups for which an estimator yields substantially different parameter estimates. The idea is to vary some parameter of the noise model, typically the noise "direction" ρ, thereby traversing the noise space, and compare parameter estimates using some distance metrics. When the distance is minimum, we may conclude that the "true" noise model has been found. Once the noise model is known, the problem reverts to the classical identification case, and a maximum likelihood estimator can be applied over the entire set of observations to get "true" model parameters. Our primary task is therefore to identify efficient separation mechanisms and appropriate distance metrics.

The fundamental idea behind the aforementioned approach is that observations have a "hidden knowledge" of the true noise covariance structure. The aim of the separation step is to partition the set of observations so that they are as far as possible from the perspective of the noise structure, i.e. they react differently to various assumptions of noise structure. Consequently, when the noise "direction" ρ is varied and parameter estimates are derived for each value of ρ, they are likely to differ substantially when an incorrect "direction" has been assumed. On the other hand, if the assumption for ρ matches the true value, the two sets of observations are likely to behave similarly when subject to parameter estimation. Notice the underlying assumption that separation should produce sets with sufficiently different characteristics. Otherwise, estimates may be close to each other even if the noise structure is not appropriate, yielding a false value for noise variances and, in turn, model parameter values.

6.1 The estimation process

In order to get an insight into the estimation process, suppose that an appropriate separation mechanism has been selected. The estimation process then runs as follows:

1 A noise model is assumed. As estimator methods (including the Koopmans–Levin and the maximum likelihood methods) often automatically compute noise magnitude given a noise covariance structure, it is sufficient to parameterize a covariance matrix C in (4) corresponding to white noise with a single scalar ρ that represents noise "direction".

2 Using the noise-polluted observations $y(t)$ and $u(t)$, the observation matrix X_q in (11) is constructed. q is a parameter of the user's choice such that $q \gg m$, with higher values (to a limit) yielding more accurate results at the expense of computational cost.

3 Rows of X_q, each of which represents an observation at time t, are grouped into two sets by means of a clustering algorithm.

4 The generalized Koopmans–Levin estimator derives parameter estimates for each of the sets independently by minimizing the joint loss function J in (14) given the chosen noise model. While we have selected this particular estimator, it is equally possible to use other types of estimators that need a noise structure model.

5 Parameter estimates for the two sets are compared using some distance metrics.

As the value of ρ is within the range $[0; \frac{\pi}{2}]$ (0 corresponding to input noise, $\frac{\pi}{2}$ to output noise only), minimizing d yields the "true" value for $\hat{\rho}$. Once an estimate for ρ is at our disposal, we may apply an efficient maximum likelihood estimator [14] to compute "true" model parameters estimates $\hat{\theta}$ as well as the noise magnitude $\hat{\mu}$, and hence $\hat{\sigma}_y$ and $\hat{\sigma}_u$.

6.2 Clustering based on principal component analysis

The goal of data clustering is to devise an unsupervised analysis to partition observations into disjoint sets such that points belonging to the same set are similar, while those belonging to different sets are dissimilar. Principal component analysis (PCA) is a widely used statistical method for dimension reduction. The basis for dimension reduction is that PCA picks up the dimensions with the largest variances. The idea of PCA-based separation is to compute the singular value decomposition (SVD), which is the basis for PCA, and inspect one or more of the principal singular vectors. More specifically, decompose the data matrix $\bar{D}$ such that

$$\bar{D} = \bar{U}\,\Sigma\,\bar{V}^{\top}$$

and denote the columns of $\bar{U}$ as $\bar{u}_i$ so that the first principal vector is $\bar{u}_1$. A set indicator may then be introduced so that

$$\begin{aligned} S_1 &= \left\{ i \mid f(\bar{u}_{p_i}(i), \ldots, \bar{u}_{p_f}(i)) \geqq 0 \right\} \\ S_2 &= \left\{ i \mid g(\bar{u}_{p_i}(i), \ldots, \bar{u}_{p_f}(i)) < 0 \right\} \end{aligned}$$

where $f,\ g\ :\ \mathbb{R}^{p_f - p_i} \to \mathbb{R}$, $i = 1 \ldots N$ and $p_f - p_i$ determines how many principal components to take into consideration. Conveniently, g is chosen to be the complement of f, such that $S_1 \cap S_2 = \emptyset$.

The most natural way to assess the performance of the functions f and g is to compare the covariance matrices R_1 and R_2 the separated observations they bring forth would produce. The aim is to produce characteristically different elements in the lower (or equivalently, upper) triangle of R_1 and R_2 calculated by taking the observations that belong to each of the two respective sets. The notion characteristically different may be measured by computing the distance

$$d = 1 - \frac{\langle R_1,\ R_2 \rangle}{||R_1||_F\ ||R_2||_F} = 1 - \frac{\operatorname{trace}(R_1 R_2)}{||R_1||_F\ ||R_2||_F} \qquad (15)$$

where $\langle M_1,\ M_2 \rangle$ is the inner product of matrices M_1 and M_2, and $||M||_F$ denotes the Frobenius norm of the matrix M. As $0 \leqq \langle R_1,\ R_2 \rangle \leqq ||R_1||_F\ ||R_2||_F$, d is in the range $[0; 1]$.

Directly maximizing d in (15) can be a cumbersome endeavor. Consequently, computationally simpler alternatives have to be considered. Choices to f include:

- $\operatorname{sgn} \bar{u}_1(i) \geqq 0$ where $\operatorname{sgn} x$ is the sign function. This is essentially equivalent to performing a k-means clustering on the data with $k = 2$ [2].
- $\prod_{k=p_i}^{p_f} \operatorname{sgn} \bar{u}_k(i) \geqq 0$. If corresponding elements in the covariance matrices have opposite signs, it is likely that the estimation algorithm produces similar estimates for the two sets only in case of correct noise assumption. A natural combination is to choose $p_i = 1$ and $p_f = 2$.
- $|\bar{u}_1|(i) > m_1$ where m_1 is the median of the values in the first principal vector $\bar{u}_1$.

- $||\bar{u}_{p_i \ldots p_f}|| > m$, which is a generalization of the above, where $\bar{u}_{p_i \ldots p_f}$ stands for the indexed principal components and m is the median value of the norm. For a 2-dimensional case with $p_i = 1$ and $p_f = 2$, this corresponds to a circle in the $\bar{u}_1$ vs $\bar{u}_2$ plane where observations are grouped whether they fall inside or outside the circle.

What remains to discuss is the exact matrix to use in place of the data matrix $\bar{D}$ that is subject to decomposition. The following are viable alternatives:

- the extended observation matrix X_q as defined by (11); or
- the components of X_q that correspond to input observations, which we denote by U_q.

Notice that the observation matrix X_q consists of both input and output observations, each of which is contaminated with noise with a different variance σ_u^2 and σ_y^2, respectively. Consequently, it is better to replace the Euclidean distance with the Mahalanobis distance that takes the different scalings into account by incorporating the noise matrix $\bar{C} = C_q$ in (14) into separation mechanisms. In accordance, the generalized version of singular value decomposition (GSVD) has to be employed instead of SVD such that

$$\begin{aligned} \bar{D} &= \bar{U}\,\Sigma_1 \bar{X}^\top \\ \bar{C} &= \bar{V}\,\Sigma_2 \bar{X}^\top \\ I &= \Sigma_1^\top \Sigma_1 + \Sigma_2^\top \Sigma_2 \end{aligned}$$

where $\bar{U}$ and $\bar{V}$ are unitary matrices and I is the unit matrix.

6.3 Comparing parameter estimates

There are various ways parameter estimates over the separated sets can be compared. The most straightforward is to use the relative distance

$$d = \frac{||\hat{\theta}_1 - \hat{\theta}_2||}{||\hat{\theta}_1||\,||\hat{\theta}_2||}$$

where $\hat{\theta}_k$ represents the estimated parameter vector on set k. It is, however, often more practical to compare autoregressive (AR) components $\hat{\theta}_k^a$ of the model only, i.e. parameters a_i, which often produces more accurate results, especially for sequences with low moving average excitation. Accordingly,

$$d = \frac{||\hat{\theta}_1^a - \hat{\theta}_2^a||}{||\hat{\theta}_1^a||\,||\hat{\theta}_2^a||}.$$

As a third option, the angle enclosed by the estimated parameter vectors may be compared, such that

$$d = \angle(\theta_1, \theta_2).$$

These metrics do not take noise magnitude into account. The combined distance metrics

$$d = \left((\mu_1 - \mu_2)^2 + c \sin^2\left(\angle(\theta_1, \theta_2)\right)\right)$$

proposed in [9] can be utilized for a possibly more accurate noise direction estimate where c is a scaling constant, often chosen as $c = 1$.

7 Simulation results

Finally, we show some simulation results to illustrate the feasibility of the outlined approach and compare its performance to that of related work.

Consider the discrete linear model described by the relationship

$$y_0(t) = \frac{B(q^{-1})}{A(q^{-1})} u_0(t) = \frac{0.1q^{-1} + 0.05q^{-2}}{1 - 1.5q^{-1} + 0.7q^{-2}} u_0(t) \quad (16)$$

and let $N = 1000$, $\rho = 20^\circ$, $\mu = 0.1$, and define an ARMA input sequence that is described by the relationship

$$u_0(t) = \frac{1}{1 - 0.2q^{-1} + 0.5q^{-2}} e_u(t) \quad (17)$$

where $e_u(t)$ is a white random sequence with variance 1. The parameters for the input and output sequences have been chosen to produce a signal-to-noise ratio of approximately 10dB. As far as the parameters of the identification algorithms are concerned, set them to $q = 6$, $p_i = 1$ and $p_f = 2$ in the separation function $\prod_{k=p_i}^{p_f} \operatorname{sgn} \bar{u}_k(i) \geqq 0$, $\bar{D} = X_q$ (for the GKL algorithm), the maximum lag to m (for the Frisch algorithm), 4 extra equations to augment the BELS estimator, $dy = 3$, $du = 3$, $my = 4$ and $mu = 4$ (for the IV estimator). Next, perform a Monte–Carlo simulation of 100 runs. The means and variances of the estimates $\hat{\theta}$ are summarized in Table 1. For the sake of comparison, the special entry "ML with ρ" denotes the theoretical configuration where the identification is performed using a maximum likelihood estimator with a *known* noise direction ρ, where variance asymptotically approaches the Cramér–Rao lower bound.

As seen from Table 1, the performance of the proposed approach is comparable to other approaches, even if the variances are somewhat in favor of related work. However, the ARMA input sequence in (17) was an idealistic assumption. Next, consider a symmetric square signal with a duty cycle of $T = 75$ (large enough for the model parameters to appear in the output) and an amplitude of $A = 1$ as an input sequence, which is a less benign input as it provides little excitation for determining moving average components in (16). The results are summarized in Table 2. Apparently, the proposed approach exhibits much more favorable variances than compared related work. In fact, some parameters cannot be reliably estimated whatsoever with the other methods shown.

8 Conclusion

We have investigated an approach to identifying linear dynamic errors-in-variables systems with a preliminary clustering step. We have seen that the aim of the clustering step is to produce two separate sets which are distant from each other in a certain sense. In other words, when parameter estimates are derived for each of the two sets, they are likely to be close to one another only if an initial noise assumption was correct. In fact, assuming an incorrect noise covariance structure leads to easily identifiable groups of observations, whereas a correct assumption makes no such distinction of observations possible. As a

Tab. 1. Comparative performance of estimator algorithms with ARMA input sequence.

$\hat{\theta}$	true	ML with ρ			Frisch			BELS			IV			proposed		
a_1	-1.5	-1.5152	$\pm$	0.0168	-1.4987	$\pm$	0.0286	-1.4641	$\pm$	0.0572	-1.4871	$\pm$	0.0686	-1.4942	$\pm$	0.0640
a_2	0.7	0.7087	$\pm$	0.0158	0.6981	$\pm$	0.0269	0.6694	$\pm$	0.0474	0.6903	$\pm$	0.0573	0.6901	$\pm$	0.0568
b_1	0.1	0.1059	$\pm$	0.0041	0.1000	$\pm$	0.0066	0.0910	$\pm$	0.0175	0.0925	$\pm$	0.0251	0.1040	$\pm$	0.0062
b_2	0.05	0.0564	$\pm$	0.0064	0.0506	$\pm$	0.0095	0.0491	$\pm$	0.0082	-0.1151	$\pm$	0.0226	0.0523	$\pm$	0.0183

Tab. 2. Comparative performance of estimator algorithms with a square signal input sequence.

$\hat{\theta}$	true	Frisch			BELS			IV			proposed		
a_1	-1.5	-1.4455	$\pm$	0.2919	-1.4762	$\pm$	0.1162	-1.4514	$\pm$	0.1146	-1.5146	$\pm$	0.0570
a_2	0.7	0.6739	$\pm$	0.1707	0.6591	$\pm$	0.0609	0.6793	$\pm$	0.0617	0.7030	$\pm$	0.0366
b_1	0.1	0.0481	$\pm$	0.1737	0.0393	$\pm$	0.2035	-0.0746	$\pm$	0.4220	0.1032	$\pm$	0.0297
b_2	0.05	0.1053	$\pm$	0.1697	0.0506	$\pm$	0.1497	-0.0987	$\pm$	0.4282	0.0483	$\pm$	0.0469

result, traversing a noise space, the "true" noise model can be discovered by minimizing the distance between parameter estimates over the two sets.

References

1 **Agüero J C, Goodwin G C**, *Identifiability of errors in variables dynamic systems*, Automatica, **44**(2), (2008), 371–382, DOI 10.1016/j.automatica.2007.06.011.

2 **Ding C, He X**, *K-means clustering via principal component analysis*, Proceedings of the 21st international conference on machine learning, ACM, New York, NY, USA, 2004, 29, DOI 10.1145/1015330.1015408, (to appear in print).

3 **Diversi R, Guidorzi R, Soverini U**, *A new criterion in EIV identification and filtering applications*, Proc. of 13th IFAC symposium on system identification, 2003, 1993–1998.

4 **Ekman M, Hong M, Söderström T**, *A separable nonlinear least-squares approach for identification of linear systems with errors in variables*, 14th IFAC symposium on system identification, 2006 March, 178–183.

5 **de Groen P**, *An introduction to total least squares*, Nieuw Archief voor Wiskunde, **4**(14), (1996), 237–253.

6 **Hong M, Söderström T, Soverini U, Diversi R**, *Comparison of three Frisch methods for errors-in-variables identification*, Proc. of 17th IFAC world congress, Seoul, Korea, 2008 July, 420–425.

7 **Hong M, Söderström T, Zheng W X**, *A simplified form of the bias-eliminating least squares method for errors-in-variables identification*, Department of Information Technology, Uppsala University, 2006. report number: 2006-040.

8 **Levin M J**, *Estimation of a system pulse transfer function in the presence of noise*, Proc. of joint automatic control conference, 1963, 452–458.

9 **Markovsky I, Kukush A, Huffel S V**, *On errors-in-variables with unknown noise variance ratio*, Proc. of 14th IFAC symposium on system identification, 2006 March, 172–177.

10 **Schoukens J, Pintelon R, Vandersteen G, Guillaume P**, *Frequency domain system identification using non-parametric noise models estimated from a small number of data sets*, Automatica, **33**, (1997), 1073–1086, DOI 10.1016/S0005-1098(97)00002-2.

11 **Söderström T**, *Errors-in-variables methods in system identification*, Automatica, **43**, (2007), 939–958, DOI 10.1016/j.automatica.2006.11.025.

12 **Thil S, Gilson M, Garnier H**, *On instrumental variable-based methods for errors-in-variables model identification*, Proc. of 17th IFAC world congress, Seoul, Korea, 2008 July, 420–425.

13 **Thil S, Hong M, Söderström T, Gilson M, Garnier H**, *Statistical analysis of a third-order cumulants based algorithm for discrete-time errors-in-variables identification*, Proc. of 17th IFAC world congress, Seoul, Korea, 2008 July, 420–425.

14 **Vajk I, Hetthéssy J**, *Efficient estimation of errors-in-variables models*, 17th IFAC world congress, Seoul, Korea, 2008 July.

15 **Vajk I**, *Identification methods in a unified framework*, Automatica, **41**(8), (2005), 1385–1393, DOI 10.1016/j.automatica.2005.03.012.

16 **Zheng W X**, *A bias correction method for identification of linear dynamic errors-in-variables models*, IEEE Transactions on Automatic Control, **47**(7), (2002 July), 1142–1147, DOI 10.1109/TAC.2002.800661.