A generalized instrumental variable framework for EIV identification methods when the measurement noises are mutually correlated

Torsten Söderström * **Roberto Diversi** [b] **Umberto Soverini** **

* *Department of Information Technology, Uppsala University, Sweden*
** *Department of Electrical, Electronic and Information Engineering, University of Bologna, Italy.*

Abstract: In this paper, the previously introduced Generalized Instrumental Variable Estimator (GIVE) is extended to the case of errors-in-variables models where the additive input and output noises are mutually correlated white processes. It is shown how many estimators proposed in the literature can be described as various special cases of a generalized instrumental variable framework. It is also investigated how to analyze the common situation where some of the equations that define the estimator are to hold exactly, and others to hold approximately in a least squares sense, providing a detailed study of the accuracy analysis.

Keywords: Identification; errors–in–variables models; mutually correlated noises.

1. INTRODUCTION

In this paper the problem of identifying a linear dynamic system from noisy input–output measurements is addressed. System representations where both the input and output are affected by additive errors are called errors–in–variables (EIV) models and play an important role in several engineering applications Van Huffel (1997), Van Huffel and Lemmerling (2002). The identification of EIV models has been deeply investigated in the literature, see Söderström (2007), Guidorzi et al. (2008) Söderström (2012) and the references therein.

Many recently proposed methods belong to the class of bias-compensated least squares (BCLS) methods or can be interpreted as BCLS methods. Among these approaches it is worth recalling the bias-eliminating least squares (BELS) Zheng (1998), Zheng (2002), the extended compensated least squares Ekman (2005), Ekman et al. (2006) and the dynamic Frisch scheme Beghelli et al. (1993), Diversi et al. (2003), Diversi et al. (2004), Diversi et al. (2006), Diversi et al. (2012).

The BCLS methods rely on the so-called compensated normal equations whose unknowns are the plant parameters and the noise variances. Since the number of such equations coincides with the number of system parameters, more equations are needed for solving the EIV identification problem. Note that certain methods (e.g. the Frisch scheme) require that some of the defining equations are exactly satisfied.

The relations between the BCLS methods have been analyzed in Hong and Söderström (2009) whereas in Söderström (2011) it is shown how these methods can be put into a general framework, resulting into a Generalized Instrumental Variable Estimator (GIVE). The GIVE framework provides also a general accuracy analysis leading to the asymptotic covariance matrix of the parameter estimates Söderström (2011). This framework has been extended to MIMO EIV systems in Söderström (2012).

In this paper, the GIVE approach is extended to EIV models with mutually correlated input and output measurement errors. The paper shows also how the methods that require some BCLS equations to hold exactly can be embedded into the GIVE framework as a limiting case, providing a detailed study of the accuracy analysis. Note that, the case of mutually correlated noises has been rarely treated in the literature and only with reference to specific approaches, see e.g. Beghelli et al. (1997), Diversi (2013), Diversi et al. (2012).

The organization of the paper is as follows. In the next section we provide the setup and problem formulation and introduce notations. The bias-compensation principle is reviewed in Section 3, while Section 4 provides a general framework that can describe many bias-compensating estimation schemes. Section 5 is devoted to illustrate how several identification methods in the literature correspond to various special cases of the general approach. The asymptotic distribution of the parameter estimates is reviewed in Section 6. The above parts appeared in a preliminary form also in the conference paper Söderström et al. (2014). In addition in Section 7 and the associated appendix we provide a detailed study of how to analyse the accuracy when some of the defining equations are to hold exactly, and others to hold approximately. In fact, such a situation applies to many of the estimators proposed in the literature. Concluding remarks are provided in Section 8.

2. SETUP AND PROBLEM FORMULATION

Consider the linear time–invariant SISO system described in Figure 1. The noise–free input and output $u_0(t)$, $y_0(t)$ are linked by the difference equation

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

where $A(z^{-1})$ and $B(z^{-1})$ are polynomials in the backward shift operator z^{-1}

$$\begin{aligned} A(z^{-1}) &= 1 + a_1 z^{-1} + \cdots + a_{n_a} z^{-n_a} \\ B(z^{-1}) &= b_0 + b_1 z^{-1} + \cdots + b_{n_b} z^{-n_b}. \end{aligned} \tag{2}$$

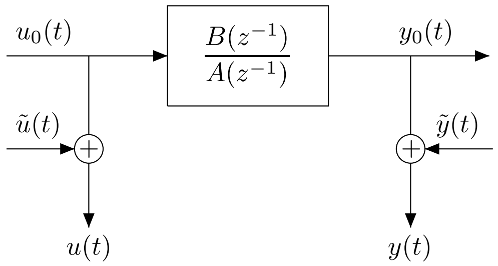

Fig. 1. Errors–in–variables model

In the EIV environment the input and output measurements are assumed as corrupted by additive noise so that the available observations are

$$u(t) = u_0(t) + \tilde{u}(t) \tag{3}$$
$$y(t) = y_0(t) + \tilde{y}(t). \tag{4}$$

In the sequel, the following assumptions will be considered as satisfied.

A1. System (1) is asymptotically stable.
A2. $A(z^{-1})$ and $B(z^{-1})$ do not share any common factor.
A3. The polynomial degrees n_a and n_b are assumed to be *a priori* known.
A4. The noiseless input $u_0(t)$ is either a zero–mean ergodic process or a quasi-stationary determinstic signal, and is persistently exciting of sufficiently high order.
A5. $\tilde{u}(t)$ and $\tilde{y}(t)$ are mutually correlated zero–mean ergodic white processes with covariances

$$\mathsf{E}\left[\tilde{u}(t)\,\tilde{u}(t-\tau)\right] = \lambda_u\,\delta(\tau) \tag{5}$$
$$\mathsf{E}\left[\tilde{y}(t)\,\tilde{y}(t-\tau)\right] = \lambda_y\,\delta(\tau) \tag{6}$$
$$\mathsf{E}\left[\tilde{u}(t)\,\tilde{y}(t-\tau)\right] = \lambda_{yu}\,\delta(\tau), \tag{7}$$

where $\delta(\tau)$ denotes the Kronecker delta function.
A6. $\tilde{u}(t)$ and $\tilde{y}(t)$ are uncorrelated with the noise–free input $u_0(t)$.

We will discuss later, see Section 5, how Assumption A5 may be either extended or simplified.

The problem under investigation is the following.

Problem. Given a set of observations $u(1),\dots,u(N)$, $y(1),\dots,$ $y(N)$, estimate the coefficients a_k $(k = 1,\dots,n_a)$, b_k $(k = 0,\dots,n_b)$ and possibly also the noise covariances λ_u, λ_y, λ_{yu}.

Remark 1. Maximum likelihood (ML) solutions often possess strong properties of high accuracy. However, in the present context no ML solution exists. For an ML solution to exist one need modified assumptions. One possibility would be to assume $\lambda_{yu} = 0$, and *known* noise ratio λ_y/λ_u, see for example Soverini and Söderström (2014). Another possibility is to impose a parametric model of the noisefree input $u_0(t)$. How to use an ARMA model for $u_0(t)$ and to obtain the ML esitmates of all parameters is considered in Söderström (1981), Söderström (2006).

2.1 Some notations

For the subsequent analysis it is useful to define the vectors

$$\varphi_0(t) = [\,-y_0(t-1)\,\dots\,-y_0(t-n_a)\,u_0(t)\,\dots\,u_0(t-n_b)\,]^T \tag{8}$$
$$\varphi(t) = [\,-y(t-1)\,\dots\,-y(t-n_a)\,u(t)\,\dots\,u(t-n_b)\,]^T \tag{9}$$
$$\tilde{\varphi}(t) = [\,-\tilde{y}(t-1)\,\dots\,-\tilde{y}(t-n_a)\,\tilde{u}(t)\,\dots\,\tilde{u}(t-n_b)\,]^T \tag{10}$$

and the parameter vectors

$$\theta = \left[a_1\,\cdots\,a_{n_a}\;b_0\,\cdots\,b_{n_b}\right]^T \tag{11}$$
$$\rho = \left[\lambda_y\;\lambda_u\;\lambda_{yu}\right]^T. \tag{12}$$

It is also convenient to define the extended vectors

$$\phi_0(t) = [\,-y_0(t)\;\varphi_0^T(t)\,]^T \tag{13}$$
$$\phi(t) = [\,-y(t)\;\varphi^T(t)\,]^T \tag{14}$$
$$\tilde{\phi}(t) = [\,-\tilde{y}(t)\;\tilde{\varphi}^T(t)\,]^T \tag{15}$$

and the extended parameter vector

$$\Theta = \left[1\;\;\theta^T\right]^T. \tag{16}$$

In the following, for a stationary random process $x(t)$ we define its covariance function $r_x(\tau)$ as

$$r_x(\tau) = \mathsf{E}\left[x(\tau)\,x(t-\tau)\right] \qquad \tau = 0, \pm 1, \pm 2 \dots \tag{17}$$

where E denotes the expectation operator. Further, the cross–covariance matrix between two random vectors $x(t)$ and $y(t)$ and the cross-covariance vector between a random vector $x(t)$ and a scalar random variable $z(t)$ are denoted as

$$R_{xy} = \mathsf{E}\left[x(t)\,y^T(t)\right] \qquad r_{xz} = \mathsf{E}\left[x(t)\,z(t)\right]. \tag{18}$$

The estimates of these covariances from measured data are denoted as

$$\hat{R}_{xy} = \frac{1}{N}\sum_{t=1}^{N} x(t)y^T(t) \qquad \hat{r}_{xz} = \frac{1}{N}\sum_{t=1}^{N} x(t)z(t)\,. \tag{19}$$

For the parameter vectors θ and ρ, a subscript 0, as in θ_0 and ρ_0, is included when it is emphasized that they are evaluated for the ‘true’ parameter values.

The notation

$$\|x\|_W^2 = x^T W x \tag{20}$$

is used for a weighted squared norm of a vector x, where W is a positive definite weighting matrix.

3. BIAS-COMPENSATED LEAST SQUARES

The EIV model (1)–(4) can be rewritten in the form

$$y(t) = \varphi(t)^T\,\theta + \varepsilon(t) \tag{21}$$
$$\varepsilon(t) = \tilde{y}(t) - \tilde{\varphi}(t)^T\theta\,. \tag{22}$$

When the least squares method is applied to the linear regression (21) the estimate of θ_0 will be biased and non-consistent due to the presence of the measurement noises. In fact for $N \to \infty$, we have

$$\mathsf{E}\left[\varphi(t)\,\varphi^T(t)\right]\theta_{LS} = \mathsf{E}\left[\varphi(t)\,y(t)\right] \tag{23}$$

i.e.

$$R_{\varphi\varphi}\theta_{LS} = r_{\varphi y}\,. \tag{24}$$

Since

$$\mathsf{E}\left[\varphi(t)\,y(t)\right] = \mathsf{E}\left[\varphi_0(t)\,y_0(t)\right] + \mathsf{E}\left[\tilde{\varphi}(t)\,\tilde{y}(t)\right] \tag{25}$$

and $y_0(t) = \varphi_0^T(t)\,\theta_0$, according to Assumptions A5–A6 it results

$$R_{\varphi\varphi}\theta_{LS} = \left(R_{\varphi\varphi} - R_{\tilde{\varphi}\tilde{\varphi}}\right)\theta_0 + r_{\tilde{\varphi}\tilde{y}}\,. \tag{26}$$

To get an unbiased estimate, a Bias Compensated Least Squares (BCLS) scheme can be considered. The basic idea is to remove the noise contributions by estimating in some way the noisy terms $R_{\tilde{\varphi}\tilde{\varphi}}$ and $r_{\tilde{\varphi}\tilde{y}}$, i.e

$$\theta_{BCLS} = \big(R_{\varphi\varphi} - R_{\tilde{\varphi}\tilde{\varphi}}(\rho)\big)^{-1}\big(r_{\varphi y} - r_{\tilde{\varphi}\tilde{y}}(\rho)\big) \, , \tag{27}$$

where ρ is the noise parameter vector introduced in (12). In order to estimate the noise parameters ρ, at least three more equations are needed in addition to the $n_a + n_b$ relations (27).

Under Assumption A5 the terms $R_{\tilde{\varphi}\tilde{\varphi}}(\rho)$ and $r_{\tilde{\varphi}\tilde{y}}(\rho)$ have the following structure

$$R_{\tilde{\varphi}\tilde{\varphi}}(\rho) = \begin{bmatrix} \lambda_y \, I_{n_a} & -\lambda_{yu} \, E_{n_a,n_b+1} \\ -\lambda_{yu} \, E^T_{n_a,n_b+1} & \lambda_u \, I_{n_b+1} \end{bmatrix} \tag{28}$$

$$r_{\tilde{\varphi}\tilde{y}}(\rho) = \left[0_{1,n_a} \mid \lambda_{yu} \mid 0_{1,n_b}\right]^T \, . \tag{29}$$

If $n_a \geq n_b$ matrix E_{n_a,n_b+1} is

$$E_{n_a,n_b+1} = \begin{bmatrix} 0_{n_b,1} & I_{n_b} \\ 0_{n_a-n_b,1} & 0_{n_a-n_b,n_b} \end{bmatrix} \; ; \tag{30}$$

if $n_a < n_b$ matrix E_{n_a,n_b+1} is

$$E_{n_a,n_b+1} = [0_{n_a,1} \;\; I_{n_a} \;\; 0_{n_a,n_b-n_a}] \, . \tag{31}$$

There are several methods proposed and analyzed in the literature that fall into the category (27). Traditionally, the case when the noises are uncorrelated is considered, so that $\lambda_{yu} = 0$. For such cases two more equations are needed. Here, we briefly describe the main ideas for such additional equations, without giving all mathematical details:

- For the bias-eliminating least squares approach, BELS, see e. g. Zheng (1998), Zheng (2002), one of the equations is derived by evaluating the minimal loss
$$V_{LS} = \mathsf{E}\left[(y(t) - \varphi^T(t)\theta_{LS})^2\right] , \tag{32}$$
where $\theta_{LS} = R^{-1}_{\varphi\varphi} \, r_{\varphi y}$, see (24). More equations can be obtained by considering the least squares estimation of an "augmented" model where additional coefficients (equal to zero) are introduced in $A(z^{-1})$ and/or $B(z^{-1})$.
- For the Frisch scheme for EIV identification, Beghelli et al. (1997), Guidorzi et al. (2008), Diversi et al. (2012), the idea is to consider the relation
$$R_{\phi_0\phi_0}\Theta_0 = 0, \tag{33}$$
that can be expressed as
$$\left(R_{\phi\phi} - R_{\tilde{\phi}\tilde{\phi}}(\rho_0)\right)\Theta_0 = 0. \tag{34}$$
Note that one more equation is used as compared to (26). Other relations can be introduced by considering augmented vectors of $\phi(t)$ of different dimensions, corresponding to different model orders.

4. GENERAL FRAMEWORK

The Generalized Instrumental Variable Estimator (GIVE) has been introduced in Söderström (2011) as a class of estimators based on the bias–eliminating principle (27) containing many previously known methods as special cases.

Introduce the total parameter vector η as

$$\eta = \begin{bmatrix} \theta \\ \rho \end{bmatrix} . \tag{35}$$

Introduce a generalized instrumental vector (IV) $z(t)$, composed of delayed values of $y(t)$ and $u(t)$, and of dimension n_z, where

$$n_z \geq \dim(\eta) = n_a + n_b + 4 \tag{36}$$

Correlating $z(t)$ with the equation error $\varepsilon(t)$ in (21) it is possible to write the following over–determined set of equations

$$\big(R_{z\varphi} - R_{\tilde{z}\tilde{\varphi}}(\rho)\big)\theta = \big(r_{zy} - r_{\tilde{z}\tilde{y}}(\rho)\big) \, , \tag{37}$$

where the choice of the instrumental variable $z(t)$ determines the structure of $R_{\tilde{z}\tilde{\varphi}}(\rho)$ and $r_{\tilde{z}\tilde{y}}(\rho)$. Cf. (24) and (26). In order to determine the parameter vectors θ and ρ, some of the entries in $z(t)$ must be correlated with $\varepsilon(t)$.

When n_z is chosen so that inequality applies in (36), the system of equations in (37) is over-determined. A very common choice for the vector $z(t)$ is

$$z(t) = [\, y(t) \; \ldots \; y(t - n_a - p_y) \;\; u(t) \; \ldots \; u(t - n_b - p_u) \,]^T \tag{38}$$

where p_u and p_y are user chosen variables, with $p_y \geq 0$, $p_u \geq 0$ and $p_u + p_y \geq 2$.

In the general case, for the the GIVE method the parameter estimate $\hat{\eta}_{GIVE}$ is defined as the solution to an optimization problem. The GIVE estimate of η is

$$\hat{\eta}_{GIVE} = \arg\min_{\eta} V_{GIVE}(\eta) \tag{39}$$

$$\begin{aligned} V_{GIVE}(\eta) &= \|\hat{r}_{zy} - r_{\tilde{z}\tilde{y}}(\rho) - \big(\hat{R}_{z\varphi} - R_{\tilde{z}\tilde{\varphi}}(\rho)\big)\theta\|^2_{W(\theta)} \\ &= \|\hat{r}_{z\varepsilon} - r_{\tilde{z}\varepsilon} \, (\theta, \rho)\|^2_{W(\theta)} \\ &\triangleq \|\bar{r}_{z\varepsilon}(\theta, \rho)\|^2_{W(\theta)} \end{aligned} \tag{40}$$

In its most general form, one uses a θ-dependent weighting matrix $W(\theta)$, but often $W(\theta)$ is chosen as a constant matrix.

When the weighting matrix W does not depend on θ, the minimization in (39) with respect to θ is easy, as the criterion V_{GIVE} is quadratic in θ. The problem is then indeed a separable nonlinear least squares problem. This means that the estimate (39) can be obtained as the solution of an associated problem of *lower dimension*, cf Golub and Pereyra (1973), Golub and Pereyra (2003):

$$\hat{\theta}_{GIVE} = \left[\bar{R}^T_{z\varphi} W \bar{R}_{z\varphi}\right]^{-1} \bar{R}^T_{z\varphi} W \bar{r}_{zy}|_{\rho=\hat{\rho}_{GIVE}} \tag{41}$$

$$\bar{R}_{z\varphi}(\rho) = \hat{R}_{z\varphi} - R_{\tilde{z}\tilde{\varphi}}(\rho) \tag{42}$$

$$\bar{r}_{z\varphi}(\rho) = \hat{r}_{zy} - r_{\tilde{z}\tilde{y}}(\rho) \tag{43}$$

$$\hat{\rho}_{GIVE} = \arg\min_{\rho} \bar{V}_{GIVE}(\rho) \tag{44}$$

$$\bar{V}_{GIVE}(\rho) = \bar{r}^T W \bar{r} - \bar{r}^T W \bar{R}\left[\bar{R}^T W \bar{R}\right]^{-1} \bar{R}^T W \bar{r} \tag{45}$$

The function $\bar{V}_{GIVE}(\rho)$ is called a concentrated loss function.

There is one special situation that is worth discussing for the case when the system of equations (37) is over-determined. It is not uncommon for such cases that one chooses to require some of the equations to hold exactly, while for the others the difference between the left and the right hand sides is minimized in a weighted least squares sense.

To formulate such a case, split the vector $z(t)$ as

$$z(t) = \begin{bmatrix} z_1(t) \\ z_2(t) \end{bmatrix} \begin{matrix} \}n_1 \text{ elements} \\ \}n_2 \text{ elements} \end{matrix} \tag{46}$$

where we require $\bar{r}_{z_1\varepsilon} = 0$ to hold exactly. For this to be meaningful, we must have

$$n_1 < \dim(\eta) = n_a + n_b + 4 \tag{47}$$

$$n_2 = n_z - n_1 \tag{48}$$

The optimization problem for finding $\hat{\eta}_{GIVE}$ is then

$$\hat{\eta}_{GIVE} = \arg\min_{\rho,\theta} \|\bar{r}_{z_2\varepsilon}(\rho,\theta)\|^2_{W_2} \quad \text{such that } \bar{r}_{z_1\varepsilon}(\rho,\theta) = 0 \tag{49}$$

This estimate can be seen as an extreme case of the general formulation (41)-(45), by choosing

$$W = \begin{bmatrix} \alpha I_{n_1} & 0 \\ 0 & W_2 \end{bmatrix} \tag{50}$$

and letting α tend to infinity.

We have found, cf Hong et al. (2007), Hong and Söderström (2009), Söderström (2011) that many bias-compensating schemes can be formulated as the general estimator (41)-(45), by appropriate choices of $z_1(t)$, $z_2(t)$ and W. We exemplify such choices in the next section. This also means that *these formally different estimators are equivalent.*

What does equivalence mean in this context?

(1) First, one has to distinguish between *the equations* (41)-(45) defining the estimates on one hand, and *the choice of numerical algorithm* employed to solve them on the other. The properties of the estimates (the solution to the equations), such as statistical properties of the estimation error $\hat{\eta} - \eta_0$, do not depend on the way the equations are solved. That is, which particular algorithm that is used for finding the solution has no importance on the properties of the solution itself.

The choice of the algorithm may still be important from a practical point of view. The amount of computations needed, the robustness to rounding errors and to the initial guesses for the nonlinear optimization in (39) or (44) can differ considerably between different algorithms. For the particular problem treated in this paper, see Söderström et al. (2005b), Söderström et al. (2005a) for some examples. According to our experiences, the use of the concentrated loss function formulation (41)-(45) is a preferable and robust way to solve the optimization problem.

(2) The parameter estimates for different estimators defined by solving the same set of equations can differ somewhat for various reasons:
- One aspect is that different weights $W(\theta)$ may have been chosen.
- A more subtle difference, that should be of minor importance is the precise way the covariance estimates (18) are formed from the data. For example, are the time points where all elements of $x(t)$ and $y(t)$ are not available completely discarded in (18), or are not available data replaced by zeros?
- Another reason for minor differences is that all estimator algorithms include some iterative computations, and that different stopping rules may be applied.

5. EXAMPLES AND SPECIAL CASES

We first discuss some special cases as well as extensions of the noise assumption A5. In the second subsection we give several specific examples to show how well-known methods from the literature fit into the GIVE framework of Section 4.

5.1 Special cases and extensions

Here we discuss various modification of the Assumption A5, that $\tilde{y}(t), \tilde{u}(t)$ are mutually correlated white noise processes.

- The case of uncorrelated noises is simply obtained by setting $\lambda_{yu} = 0$, and omitting the corresponding element of the noise parameter vector ρ, (12). This situation is also the most commonly treated one in the literature. This situation also implies that the dimensions of ρ and η are decreased by one unit, and so does the minimal number of equations in (36). The modifications of $r_{\tilde{z}\tilde{y}}$ and $R_{\tilde{z}\tilde{\varphi}}$ in (37) are straightforward.
- One may consider the case when $\tilde{y}(t)$ is arbitrarily auto-correlated, but uncorrelated with the white input noise $\tilde{u}(t)$. For such a case, one might add covariance elements $r_{\tilde{y}}(\tau)$ for a number of $\tau-$values in the ρ vector. However, it then turns out to be infeasible to use any delayed values of the output $y(t)$ in the vector $z(t)$. For any new such vector element added, one has also to include a further unknown $r_{\tilde{y}}(\tau)$. Therefore, it only makes sense in such scenarios to use delayed input variables in the $z(t)$ vector, so for example
$$z(t) = [\, u(t) \ \dots \ u(t - n_b - p_u) \,]^T \tag{51}$$
where, according to (36), $p_u \geq n_a + 1$, as $\rho = \lambda_u$ in this case.
- For the situation above, it is also possible to use only further delayed inputs in the $z(t)$ vector
$$z(t) = [\, u(t - n_b - 1) \ \dots \ u(t - n_b - p_u) \,]^T \tag{52}$$
with $p_u \geq n_a + n_b + 1$. In this case there is indeed no noise parameter vector needed, as $r_{\tilde{z}\tilde{y}}$ and $R_{\tilde{z}\tilde{\varphi}}$ both become zero. The estimate is the instrumental variable estimate described, e.g., in Söderström (1981). It can also be interpreted as a Yule-Walker estimate.
- A more general situation occurs when $\tilde{y}(t)$ consists of two independent components, $\tilde{y}(t) = \tilde{y}_1(t) + \tilde{y}_2(t)$. Then we can use $\tilde{y}_1(t)$ to model an arbitrarily auto-correlated process noise, and let $\tilde{y}_2(t)$ describe white measurement noise. Further assume that $[\ \tilde{y}_2(t) \ \tilde{u}(t) \,]^T$ is a vector-valued white noise. It is then possible to proceed as above with $\rho = [\, \lambda_{yu} \ \lambda_u \,]^T$ and using only delayed input values in the vector $z(t)$. Such a case is treated in Diversi et al. (2010).

5.2 Various examples

We now illustrate how several methods earlier proposed in the literature fit into the general GIVE framework. For each method, we specify how $z_1(t)$, $z_2(t)$ and W_2 are selected. Very often the originally introduced methods are based on a model with $b_0 = 0$, and we make here the straightforward adjustments in the description to treat the general case (1) - (2).

Example 5.1. The bias-eliminating least squares method was introduced in Zheng (1998) and Zheng (2002) and proposed in a number of variants.

The algorithm BELS-1 of Zheng (1998), corresponds to $z_1(t) = \left[\, y(t) \ \varphi^T(t) \,\right]^T$, $z_2(t) = y(t - n_a - 1)$. As $z_2(t)$ is a scalar, there is no need for any weighting matrix W_2.

Simililarly, The algorithm BELS-2 of Zheng (1998), corresponds to $z_1(t) = \left[\, y(t) \ \varphi^T(t) \,\right]^T$, $z_2(t) = u(t - n_b - 1)$. As $z_2(t)$ also in this case is a scalar, there is no need for any weighting matrix W_2.

Remark 2. Note that the signs of the elements in the vector $z(t)$ are not significant, and do not change the equations defining the estimators. For example, one may change a negative delayed output to its positive value, that is, to replace $-y(t-i)$ by $y(t-i)$, without affecting the estimator. This property will sometimes be used below, when deemed convenient. □

Example 5.2. Another variant of the bias-eliminating least squares method is the algorithm BELS-II of Zheng (2002), which copes with the case of arbitrarily correlated output noise. The equations obtained by correlating past outputs, say $y(t-j)$, with the equation errors $\varepsilon(t) = y(t) - \varphi^T(t)\theta$ are then not 'useful', in the sense that for each further equations used, the number of unknowns also increases by one. After eliminating all the equations involving the unknown correlation function $r_{\tilde{y}}(\tau)$ of the output disturbances, the algorithm leads to the use of

$$z_1(t) = \begin{bmatrix} u(t) \\ \vdots \\ u(t-n_b) \end{bmatrix}, \quad z_2(t) = \begin{bmatrix} u(t-n_b-1) \\ \vdots \\ u(t-n_b-n_a-1) \end{bmatrix}$$

for estimating the unknowns θ and λ_u. (No weighting is needed.)

Example 5.3. The Frisch scheme for EIV identification has been proposed in several forms. One of the first appeared in Beghelli et al. (1993). A common aspect for all these methods is that the adjusted normal equations are used. This means precisely that

$$z_1(t) = [\, y(t) \;\dots\; y(t-n_a) \;\; u(t) \;\dots\; u(t-n_b)\,]^T \tag{53}$$

The shifted relation criterion described in Diversi et al. (2004) is based on the following equation

$$\left(R_{\bar{\phi}\bar{\phi}} - R_{\bar{\bar{\phi}}\bar{\bar{\phi}}}(\rho_0)\right)[\, v_1 \;\; v_2\,] = 0. \tag{54}$$

where

$$\bar{\phi}(t) = [\, -y(t) \;\dots\; -y(t-n_a-1) \;\; u(t) \;\dots\; u(t-n_b-1)\,]^T \tag{55}$$

$$\bar{\bar{\phi}}(t) = [\, -\tilde{y}(t) \;\dots\; -\tilde{y}(t-n_a-1) \;\; \tilde{u}(t) \;\dots\; \tilde{u}(t-n_b-1)\,]^T \tag{56}$$

and

$$v_1 = [\, 1 \;\; a_1 \;\dots\; a_{n_a} \;\; 0 \;\; b_0 \;\dots\; b_{n_b} \;\; 0\,]^T \tag{57}$$

$$v_2 = [\, 0 \;\; 1 \;\; a_1 \;\dots\; a_{n_a} \;\; 0 \;\; b_0 \;\dots\; b_{n_b}\,]^T \tag{58}$$

so that four additional relations are used besides the standard Frisch equations. The use of v_1 leads to

$$z_2(t) = [\, y(t-n_a-1) \;\; u(t-n_b-1)\,]^T$$

while the use of v_2 leads to

$$z_2(t) = [\, y(t+1) \;\; u(t+1)\,]^T .$$

Therefore, the use of both v_1 and v_2 corresponds to

$$z_2(t) = [\, y(t-n_a-1) \;\; u(t-n_b-1) \;\; y(t+1) \;\; u(t+1)\,]^T . \tag{59}$$

Finally, in Beghelli et al. (1993) the choice $W_2 = I$ is made.

Example 5.4. Another variant of the Frisch scheme is to use additional Yule-Walker equations, Diversi et al. (2006)

This corresponds to

$$z_1(t) = [\, y(t) \;\dots\; y(t-n_a) \;\; u(t) \;\dots\; u(t-n_b)\,]^T ,$$
$$z_2(t) = [\, u(t-n_b-1) \;\dots\; u(t-n_a-m)\,]^T$$

Equal weighting, $W_2 = I$ is proposed in Diversi et al. (2006). The size m of the vector $z_2(t)$ is normally chosen so that an overdetermined system is obtained.

Adaption of this approach to the general case with cross-correlated noise, $\lambda_{yu} \neq 0$, is treated in Diversi et al. (2012).

Example 5.5. A third variant of the Frisch scheme is based on comparing the correlation functions of the equations errors, using the model on one hand and using the measured data on the other. Details are provided in Diversi et al. (2003), where this approach was first proposed. It is shown in Söderström (2011) that it corresponds to

$$z_1(t) = [\, y(t) \;\dots\; y(t-n_a) \;\; u(t) \;\dots\; u(t-n_b)\,]^T$$

$$\begin{aligned} z_2(t) &= \begin{bmatrix} \varepsilon(t-1,\theta) \\ \vdots \\ \varepsilon(t-k,\theta) \end{bmatrix} \\ &= \begin{bmatrix} 1 & a_1 & \dots & a_{n_a} & & & b_0 & \dots & b_{n_b} & & \\ & \ddots & & & \ddots & & & \ddots & & & \\ & & 1 & a_1 & \dots & a_{n_a} & & & b_0 & \dots & b_{n_b} \end{bmatrix} \\ &\times \begin{bmatrix} y(t) \\ \vdots \\ y(t-n_a-k) \\ u(t) \\ \vdots \\ u(t-n_b-k) \end{bmatrix} \\ &\triangleq M(\theta) \begin{bmatrix} z_1(t) \\ \bar{z}(t) \end{bmatrix} = M_1(\theta) z_1(t) + M_2(\theta)\bar{z}(t) \end{aligned} \tag{60}$$

$$\bar{z}(t) = [\, y(t-n_a-1) \;\dots\; y(t-n_a-k) \;\; u(t-n_b-1) \;\dots\; u(t-n_b-k)\,]^T \tag{61}$$

Further, as $\bar{r}_{z_1\varepsilon} = 0$, the criterion to be minimized can also be written as

$$\begin{aligned} \| \bar{r}_{z_2\varepsilon}(\theta) \|^2 &= \| M_1(\theta)\bar{r}_{z_1\varepsilon}(\theta) + M_2(\theta)\bar{r}_{\bar{z}\varepsilon}(\theta) \|^2 \\ &= \| M_2(\theta)\bar{r}_{\bar{z}\varepsilon}(\theta) \|^2 \end{aligned} \tag{62}$$

We may therefore also identify the vector $z_2(t)$ in the general algorithm with $\bar{z}(t)$ in (61), and let the weighting matrix depend on the parameter vector θ as

$$W_2(\theta) = M_2^T(\theta) M_2(\theta) \tag{63}$$

An extension of this algorithm to handle the general case with correlated noise, $\lambda_{yu} \neq 0$, is presented in Diversi (2013).

Example 5.6. The extended compensated least squares (ECLS) method was proposed in Ekman (2005) and analysed in Ekman et al. (2006). It corresponds to $z_1(t)$ being absent, $z_2(t)$ as in (38), with the weighting matrix $W_2 = I$.

6. ASYMPTOTIC DISTRIBUTION

The asymptotic distribution of $\hat{\eta}_{GIVE}$ was considered in Söderström (2011) for the case $\lambda_{yu} = 0$. The modification to include also a parameter λ_{yu} in the noise parameter vector ρ is straightforward. The main steps in the analysis remain the same, and are repeated here in short for convenience.

It follows from (40) that

$$0 = \bar{r}_{z\varepsilon}^T W(\theta) F \tag{64}$$

where

$$F = \frac{\partial \bar{r}_{z\varepsilon}}{\partial \eta} \tag{65}$$

As for large values of N, using linearlization around the true parameter vector

$$\bar{r}_{z\varepsilon}(\hat{\eta}) \approx \bar{r}_{z\varepsilon}(\eta_0) + F\left(\hat{\eta} - \eta_0\right) \tag{66}$$

it follows that

$$\hat{\eta} - \eta_0 \approx -\left(F^T W F\right)^{-1} F^T W \bar{r}_{z\varepsilon}(\eta_0) \tag{67}$$

We may then invoke the central limit theorem, see for example Söderström and Stoica (1989), to conclude that asymptotically in N,

$$\sqrt{N}\left(\hat{\eta} - \eta_0\right) \xrightarrow{\text{dist}} \mathcal{N}(0, P_{\text{GIVE}}), \tag{68}$$

where the covariance matrix P_{GIVE} is given by

$$P_{\text{GIVE}} \triangleq \left(F^T W F\right)^{-1} F^T W Q W F \left(F^T W F\right)^{-1}, \tag{69}$$

and

$$Q \triangleq \lim_{N\to\infty} N \operatorname{cov}\left(\tilde{r}_{z\varepsilon}\right) \tag{70}$$

$$\tilde{r}_{z\varepsilon} = \frac{1}{N}\sum_{t=1}^{N} z(t,\theta_0)\varepsilon(t,\theta_0) - \mathsf{E}\left\{z(t,\theta_0)\varepsilon(t,\theta_0)\right\} \tag{71}$$

For the case of Gaussian distributed data, it was shown in Söderström (2011) that

$$Q = \sum_{\tau=-\infty}^{\infty} \left[r_z(\tau) r_\varepsilon(\tau) + r_{z\varepsilon}(\tau) r_{z\varepsilon}^T(-\tau)\right]. \tag{72}$$

7. USING CONSTRAINTS

We have claimed above that the case of using constraints, such as requiring that some of the GIVE equations must hold exactly, can be handled as a limiting case. More specifically, one then applies the weighting as in (50) and then let $\alpha \to \infty$.

In this section we will further analyse this idea. We introduce the notations

$$\begin{aligned} F_j &= \frac{\partial \bar{r}_{z_j\varepsilon}(\eta)}{\partial \eta} \quad j = 1,2 \\ f_j &= \bar{r}_{z_j\varepsilon}(\eta_0) \end{aligned} \tag{73}$$

Then we have

$$F = \begin{bmatrix} F_1 \\ F_2 \end{bmatrix} \tag{74}$$

and close to the true parameter values (that is when $\tilde{\eta} \triangleq \hat{\eta} - \eta_0$ is small)

$$\bar{r}_{z_j\varepsilon}(\hat{\eta}) = f_j + F_j \tilde{\eta}, \ \ j = 1,2 \tag{75}$$

One can see from (67), (69) that what matters for the performance is the factor

$$G \triangleq \left(F^T W F\right)^{-1} F^T W \tag{76}$$

We consider first the case when the 'exact' formulation is used. Then the constraint (49) is imposed in the optimization. We can for $\hat{\eta} \approx \eta_0$ write the criterion and the constraint as

$$V = \| \bar{r}_{z_2\varepsilon}(\hat{\eta}) \|_{W_2}^2 = \| f_2 + F_2\tilde{\eta} \|_{W_2}^2 \tag{77}$$

$$0 = \bar{r}_{z_1\varepsilon}(\hat{\eta}) = f_1 + F_1\tilde{\eta} \tag{78}$$

The necessary conditions for a minimum of the quadratic criterion (77) under the linear constraint (78) can be derived by using the Lagrangian function

$$L(\tilde{\eta}) = \frac{1}{2}\left(f_2 + F_2\tilde{\eta}\right)^T W_2 \left(f_2 + F_2\tilde{\eta}\right) + \lambda^T\left(f_1 + F_1\tilde{\eta}\right) \tag{79}$$

Setting the derivatives to zero leads to

$$f_1 + F_1\tilde{\eta} = 0 \tag{80}$$

$$F_2^T W_2\left(f_2 + F_2\tilde{\eta}\right) + F_1^T\lambda = 0 \tag{81}$$

which immediately leads to

$$\tilde{\eta} = -\left[\, I \ \ 0 \,\right] \begin{bmatrix} F_1 & 0 \\ F_2^T W_2 F_2 & F_1^T \end{bmatrix}^{-1} \begin{bmatrix} I & 0 \\ 0 & F_2^T W_2 \end{bmatrix} \begin{bmatrix} f_1 \\ f_2 \end{bmatrix} \tag{82}$$

Next we consider the alternative and general formalism with weighting as in (50)

$$W_\alpha = \begin{bmatrix} \alpha W_1 & 0 \\ 0 & W_2 \end{bmatrix} \tag{83}$$

We get from (76) the factor

$$\begin{aligned} G(\alpha) &= \left(F^T W_\alpha F\right)^{-1} F^T W_\alpha \\ &= \left(\alpha F_1^T W_1 F_1 + F_2^T W_2 F_2\right)^{-1} \left[\, \alpha F_1^T W_1 \ \ F_2^T W_2 \,\right] \end{aligned} \tag{84}$$

We will examine how $G(\alpha)$ behaves when α grows to infinity. To proceed, we introduce assumptions and notations for the dimensions of the matrices as follows:

Matrix	Dimension	rank
F_1	$n_1 \times n$	$\text{rank}(F_1) = n_1$
F_2	$n_2 \times n$	$\text{rank}(F_2) = n_2$
F	$(n_1 + n_2) \times n$	$\text{rank}(F) = n$
W_1	$n_1 \times n_1$	W_1 pos. definite
W_2	$n_2 \times n_2$	W_2 pos. definite

The notations correspond to $n_j = \dim(\hat{r}_{z_j\varepsilon}), \ \ j = 1,2$ and $n = \dim(\theta)$. It must hold that

$$n_1 < n \tag{85}$$

as the constraint $\hat{r}_{z_1\varepsilon}$ should not completely determine the estimate $\hat{\eta}$. Further, the rank conditions on $F_j, j = 1,2$, correspond to require that $\hat{r}_{z_j\varepsilon} = 0$ does not contain linearly dependent equations. There is no condition imposed on the relation between n_2 and n.

Note that in particular the assumptions imply that the matrix $F_1^T W_1 F_1$ is singular (as it is of dimension $n \times n$ and of rank n_1). This fact makes the analysis more complicated compared to a case when the matrix is invertible.

We have first a result for writing $G(\alpha)$:

Lemma 1. Under the dimension and rank assumptions of F_1 and F_2 it holds for any fixed α that

$$\begin{aligned} G(\alpha) &= \left(\alpha F_1^T W_1 F_1 + F_2^T W_2 F_2\right)^{-1} \left[\, \alpha F_1^T W_1 \ \ F_2^T W_2 \,\right] \\ &= \left[\, I \ \ 0 \,\right] \begin{bmatrix} F_1 & -W_1^{-1}/\alpha \\ F_2^T W_2 F_2 & F_1^T \end{bmatrix}^{-1} \begin{bmatrix} I & 0 \\ 0 & F_2^T W_2 \end{bmatrix}. \end{aligned} \tag{86}$$

Proof. The right hand side of (86) can be evaluated as

$$
\begin{aligned}
\text{RHS} &= [\, I \;\; 0 \,] \left(\begin{bmatrix} F_1 & -W_1^{-1}/\alpha \\ F_2^T W_2 F_2 & F_1^T \end{bmatrix} \begin{bmatrix} 0 & I \\ I & 0 \end{bmatrix} \begin{bmatrix} 0 & I \\ I & 0 \end{bmatrix} \right)^{-1} \\
&\quad \times \begin{bmatrix} I & 0 \\ 0 & F_2^T W_2 \end{bmatrix} \\
&= [\, I \;\; 0 \,] \begin{bmatrix} 0 & I \\ I & 0 \end{bmatrix} \begin{bmatrix} -W_1^{-1}/\alpha & F_1 \\ F_1^T & F_2^T W_2 F_2 \end{bmatrix}^{-1} \begin{bmatrix} I & 0 \\ 0 & F_2^T W_2 \end{bmatrix} \\
&= [\, 0 \;\; I \,] \left(\begin{bmatrix} -\alpha W_1 & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} \alpha W_1 F_1 \\ I \end{bmatrix} \right. \\
&\quad \left. \times \left(\alpha F_1^T W_1 F_1 + F_2^T W_2 F_2 \right)^{-1} [\, \alpha F_1^T W_1 \;\; I \,] \right) \\
&\quad \times \begin{bmatrix} I & 0 \\ 0 & F_2^T W_2 \end{bmatrix} \\
&= \left(\alpha F_1^T W_1 F_1 + F_2^T W_2 F_2 \right)^{-1} [\, \alpha F_1^T W_1 \;\; F_2^T W_2 \,] \\
&= \text{LHS}.
\end{aligned}
$$

□

The lemma can now be used to verify that the solution to the approximate problem with weighting indeed does converge to the solution to the 'exact' problem, when the weighting α grows to infinity.

Remark 3. When $\alpha \to \infty$ we have

$$
\begin{aligned}
\lim_{\alpha \to \infty} G(\alpha) &= [\, I \;\; 0 \,] \begin{bmatrix} F_1 & 0 \\ F_2^T W_2 F_2 & F_1^T \end{bmatrix}^{-1} \begin{bmatrix} I & 0 \\ 0 & F_2^T W_2 \end{bmatrix}, \\
&= G
\end{aligned} \tag{87}
$$

which obviously does not depend on W_1. □

Next we attempt to rewrite the expression (87). To express the inverse is relatively complicated in the general case. We first rewrite G by making some block permutations,

$$
\begin{aligned}
G &= [\, I \;\; 0 \,] \left(\begin{bmatrix} 0 & I \\ I & 0 \end{bmatrix} \begin{bmatrix} 0 & I \\ I & 0 \end{bmatrix} \begin{bmatrix} F_1 & 0 \\ F_2^T W_2 F_2 & F_1^T \end{bmatrix} \right. \\
&\quad \left. \times \begin{bmatrix} 0 & I \\ I & 0 \end{bmatrix} \begin{bmatrix} 0 & I \\ I & 0 \end{bmatrix} \right)^{-1} \begin{bmatrix} I & 0 \\ 0 & F_2^T W_2 \end{bmatrix} \\
&= [\, 0 \;\; I \,] \begin{bmatrix} F_1^T & F_2^T W_2 F_2 \\ 0 & F_1 \end{bmatrix}^{-1} \begin{bmatrix} 0 & F_2^T W_2 \\ I & 0 \end{bmatrix}.
\end{aligned} \tag{88}
$$

In what follows we examine alternative ways of writing the matrix $G(\alpha)$.

To proceed, we need the inverse in (88). For that aim we use the following lemma.

Lemma 2. Consider the matrix

$$
B = \begin{bmatrix} K^T & Q \\ 0 & K \end{bmatrix}, \tag{89}
$$

where K is a $m \times n$ matrix of rank m, and Q is an $n \times n$ symmetric and nonnegative definite matrix. Set

$$
P = K^T \left(K K^T \right)^{-1} K = K^T K^{T^\dagger} = K^\dagger K, \tag{90}
$$

$$
P_\perp = I - P, \tag{91}
$$

$$
D_0 = P + P_\perp Q P_\perp, \tag{92}
$$

where $K^\dagger$ denotes the pseudoinverse of K. Assume that D_0 is invertible. Then B is invertible and its inverse is given by

$$
B^{-1} = \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix}, \tag{93}
$$

where

$$
H_{11} = \left(K K^T \right)^{-1} K (I - Q D), \tag{94}
$$

$$
H_{12} = -\left(K K^T \right)^{-1} K Q (I - D Q) K^T \left(K K^T \right)^{-1}, \tag{95}
$$

$$
H_{21} = D, \tag{96}
$$

$$
H_{22} = (I - D Q) K^T \left(K K^T \right)^{-1}, \tag{97}
$$

$$
D = D_0^{-1} P_\perp. \tag{98}
$$

Proof. See the appendix. □

In order to apply Lemma 2 to the inverse in (88), we need to verify that the matrix

$$
D_0 \triangleq P + P_\perp F_2^T W_2 F_2 P_\perp \tag{99}
$$

where

$$
P = F_1^\dagger F_1, \quad P_\perp = I - P \tag{100}
$$

is nonsingular. Clearly, by construction D_0 is symmetric and nonnegative definite. Further,

$$
\begin{aligned}
x^T D_0 x = 0, &\Rightarrow x^T \left(P + P_\perp F_2^T W_2 F_2 P_\perp \right) x = 0, \\
&\Rightarrow x^T P x = 0, \;\; F_2 P_\perp x = 0, \\
&\Rightarrow F_1 x = 0, \;\; F_2 \left(I - F_1^T \left(F_1 F_1^T \right)^{-1} F_1 \right) x = 0, \\
&\Rightarrow F_1 x = 0, \;\; F_2 x = 0, \;\; \Rightarrow F x = 0.
\end{aligned}
$$

as F has rank n. Thus D_0 is nonsingular. Then according to Lemma 2, (88) can be expressed as

$$
G = [\, 0 \;\; I \,] \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix} \begin{bmatrix} 0 & F_2^T W_2 \\ I & 0 \end{bmatrix}, \tag{101}
$$

where

$$
\begin{aligned}
H_{11} &= \left(F_1 F_1^T \right)^{-1} F_1 (I - F_2^T W_2 F_2 D), \\
H_{12} &= -\left(F_1 F_1^T \right)^{-1} F_1 F_2^T W_2 F_2 (I - D F_2^T W_2 F_2) F_1^T \\
&\quad \times \left(F_1 F_1^T \right)^{-1}, \\
H_{21} &= D, \\
H_{22} &= (I - D F_2^T W_2 F_2) F_1^T \left(F_1 F_1^T \right)^{-1}. \\
D &= D_0^{-1} P_\perp
\end{aligned}
$$

Straightforward multiplications in (101) then leads to

$$
G = \left[\, (I - D F_2^T W_2 F_2) F_1^T \left(F_1 F_1^T \right)^{-1} \quad D F_2^T W_2 \, \right]. \tag{102}
$$

For the particular case when rank $F_2 = n$ (which requires dim $\hat{r}_{z_2 \varepsilon} \geq n$), a simpler expression is possible:

Lemma 3. Consider the expression (86) for $G(\alpha)$. Assume

$$
n_2 \geq n = \operatorname{rank}(F_2) \tag{103}
$$

and set

$$
Q = F_2^T W_2 F_2 \tag{104}
$$

Then it holds that the following limit exists and

$$\lim_{\alpha\to\infty} G(\alpha) \triangleq [\, G_1 \;\; G_2 \,] \tag{105}$$

$$G_1 = Q^{-1}F_1^T \left(F_1 Q^{-1} F_1^T\right)^{-1} \tag{106}$$

$$G_2 = \left(I - Q^{-1}F_1^T \left(F_1 Q^{-1}F_1^T\right)^{-1} F_1\right) Q^{-1}F_2^T W_2 \tag{107}$$

Proof. Note that that matrix Q by construction is invertible. Using the matrix inversion lemma,

$$\begin{aligned}&\left(\alpha F_1^T W_1 F_1 + F_2^T W_2 F_2\right)^{-1} = Q^{-1} - Q^{-1}F_1^T \\ &\times \left(\frac{1}{\alpha}W_1^{-1} + F_1 Q^{-1} F_1^T\right)^{-1} F_1 Q^{-1}\end{aligned} \tag{108}$$

Thus from (86)

$$\begin{aligned}G_2 &= \lim_{\alpha\to\infty} \left(Q^{-1} - Q^{-1}F_1^T \left(\frac{1}{\alpha}W_1^{-1} + F_1 Q^{-1}F_1^T\right)^{-1} \right.\\ &\left.\quad \times\, F_1 Q^{-1}\right) F_2^T W_2 \\ &= Q^{-1}F_2^T W_2 - Q^{-1}F_1^T \left(F_1 Q^{-1}F_1^T\right)^{-1} F_1 Q^{-1} F_2^T W_2 \\ &= \left(I - Q^{-1}F_1^T \left(F_1 Q^{-1}F_1^T\right)^{-1} F_1\right) Q^{-1}F_2^T W_2\end{aligned} \tag{109}$$

Similarly,

$$\begin{aligned}G_1 &= \lim_{\alpha\to\infty} \left(\alpha Q^{-1}F_1^T W_1 - \alpha Q^{-1}F_1^T \right.\\ &\quad \left. \left(\frac{1}{\alpha}W_1^{-1} + F_1 Q^{-1}F_1^T\right)^{-1} F_1 Q^{-1}F_1^T W_1\right) \\ &= \lim_{\alpha\to\infty} \left(\alpha Q^{-1}F_1^T \left(\frac{1}{\alpha}W_1^{-1} + F_1 Q^{-1}F_1^T\right)^{-1} \right.\\ &\quad \left. \times \left(\frac{1}{\alpha}W_1^{-1} + F_1 Q^{-1}F_1^T - F_1 Q^{-1}F_1^T\right) W_1\right) \\ &= \lim_{\alpha\to\infty} Q^{-1}F_1^T \left(\frac{1}{\alpha}W_1^{-1} + F_1 Q^{-1}F_1^T\right)^{-1} \\ &= Q^{-1}F_1^T \left(F_1 Q^{-1}F_1^T\right)^{-1}\end{aligned} \tag{110}$$

This completes the proof. □

The results (105) - (107) give an alternative characterization of the matrix G.

8. CONCLUSIONS

It has been shown how many different estimators for the errors-in-variables problem can all be casted in a generalized instrumental variables framework. The setup used allows input and output measurement noises to be mutually correlated. Various estimators known from the literature are shown explicitly to appear as special cases of the provided framework. The paper contains also a detailed analysis of how to analyse the effect of constraints in an approximate way by replacing them with a weighted term in the criterion.

REFERENCES

S. Beghelli, P. Castaldi, R. Guidorzi, and U. Soverini. A robust criterion for model selection in identification from noisy data. In *Proc. 9th International Conference on Systems Engineering*, pages 480–484, Las Vegas, Nevada, USA, 1993.

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.

R. Diversi. Bias-eliminating least-squares identification of errors-in-variables models with mutually correlated noises. *International Journal of Adaptive Control and Signal Processing*, 27:915–924, October 2013.

R. Diversi, R. Guidorzi, and U. Soverini. A new criterion in EIV identification and filtering applications. In *Proc. 13th IFAC Symposium on System Identification*, pages 1993–1998, Rotterdam, The Netherlands, August 27-29 2003.

R. Diversi, R. Guidorzi, and U. Soverini. Frisch scheme–based algorithms for EIV identification. In *Proc. 12th IEEE Mediterranean Conference on Control and Automation*, Kusadasi, Turkey, June 2004.

R. Diversi, R. Guidorzi, and U. Soverini. Yule-Walker equations in the Frisch scheme solution of errors-in-variables identification problems. In *Proc. 17th International Symposium on Mathematical Theory of Networks and Systems (MTNS)*, pages 391–395, Kyoto, Japan, July 2006.

R. Diversi, R. Guidorzi, and U. Soverini. Identification of ARX and ARARX models in the presence of input and output noises. *European Journal of Control*, 16(3):242–255, 2010.

R. Diversi, R. Guidorzi, and U. Soverini. Identification of errors-in-variables models with mutually correlated input and output noises. In *16th IFAC Symposium on System Identification*, pages 1353–1358, Brussels, Belgium, July 2012.

M. Ekman. Identification of linear systems with errors in variables using separable nonlinear least squares. In *16th IFAC World Congress*, Prague, Czech Republic, July 04-08 2005.

M. Ekman, M. Hong, and T. Söderström. A separable nonlinear least squares approach for identification of linear systems with errors in variables. In *Proc. 14th IFAC Symposium on System Identification*, Newcastle, Australia, March 29-31 2006.

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

G. H. Golub and V. Pereyra. The differentiation of pseudoinverses and nonlinear least squares problems whose variables separate. *SIAM J. Numerical Analysis*, 10(2):413–432, 1973.

R. Guidorzi, R. Diversi, and U. Soverini. The Frisch scheme in algebraic and dynamic identification problems. *Kybernetika*, 44(5):585–616, 2008.

M. Hong and T. Söderström. Relations between bias-eliminating least squares, the Frisch scheme and extended compensated least squares methods for identifying errors-in-variables systems. *Automatica*, 45(1):277–282, January 2009.

M. Hong, T. Söderström, and W. X. Zheng. A simplified form of the bias-eliminating least squares method for errors-in-variables identification. *IEEE Transactions on Automatic Control*, 52(9):1754–1756, September 2007.

T. Söderström. On computing the Cramér-Rao bound and covariance matrices for PEM estimates in linear state space models. In *Proc. 14th IFAC Symposium on System Identification*, Newcastle, Australia, March 29-31 2006.

T. Söderström. Errors-in-variables methods in system identification. *Automatica*, 43(6):939–958, June 2007. Survey

paper.
T. Söderström. A generalized instrumental variable estimator for multivariable errors-in-variables problems. *International Journal of Control*, 85(3):287–303, March 2012.
T. Söderström. Identification of stochastic linear systems in presence of input noise. *Automatica*, 17:713–725, 1981.
T. Söderström and P. Stoica. *System Identification*. Prentice Hall International, Hemel Hempstead, UK, 1989.
T. Söderström, M. Hong, and W. X. Zheng. Convergence properties of bias-eliminating algorithms for errors-in-variables identification. *International Journal of Adaptive Control and Signal Processing*, 19(9):703–722, November 2005a.
T. Söderström, M. Hong, and W. X. Zheng. Convergence of bias-eliminating least squares methods for identification of errors-in-variables dynamic systems. In *44th IEEE CDC/ECC 2005*, Seville, Spain, December 12-15 2005b.
T. Söderström, R. Diversi, and U. Soverini. A unified framework for eiv identification methods in the presence of mutually correlated noises. In *IFAC 19th World Congress*, Cape Town, South Africa, August 24-29 2014.
U. Soverini and T. Söderström. Frequency domain maximum likelihood identification of noisy input-output models. In *IFAC 19th World Congress*, Cape Town, South Africa, August 24-29 2014.
T. Söderström. A generalized instrumental variable method for errors-in-variables. *Automatica*, 47(8):1656–1666, August 2011.
T. Söderström. System identification for the errors-in-variables problem. *Transactions of the Institute of Measurement and Control*, 34(7):780–792, October 2012.
S. Van Huffel, editor. *Recent Advances in Total Least Squares Techniques and Errors-in-Variables Modelling*. SIAM, Philadelphia, USA, 1997.
S. Van Huffel and Ph. Lemmerling, editors. *Total Least Squares and Errors-in-Variables Modelling. Analysis, Algorithms and Applications*. Kluwer, Dordrecht, The Netherlands, 2002.
W. X. Zheng. A bias correction method for identification of linear dynamic errors-in-variables models. *IEEE Transactions on Automatic Control*, 47(7):1142–1147, July 2002.
W. X. Zheng. Transfer function estimation form noisy input and output data. *International Journal of Adaptive Control and Signal Processing*, 12:365–380, 1998.

Appendix A. PROOF OF LEMMA 2

The proof is based on direct multiplication. Equation (93) holds precisely when

$$\begin{bmatrix} K^{T^\dagger}(I-QD) & -K^{T^\dagger}Q(I-DQ)K^\dagger \\ D & (I-DQ)K^\dagger \end{bmatrix} \begin{bmatrix} K^T & Q \\ 0 & K \end{bmatrix} = I$$

or equivalently in block form

$$K^{T^\dagger}(I-QD)K^T = I, \quad \text{(A.1)}$$
$$DK^T = 0, \quad \text{(A.2)}$$
$$K^{T^\dagger}(I-QD)Q - K^{T^\dagger}Q(I-DQ)K^\dagger K = 0, \quad \text{(A.3)}$$
$$DQ + (I-DQ)K^\dagger K = I. \quad \text{(A.4)}$$

The left hand side of (A.2) is easily evaluated:

$$DK^T = D_0^{-1}P_\perp K^T = D_0^{-1}\left(I - K^T K^{T^\dagger}\right)K^T = 0,$$

and hence (A.2) is proved. Using (A.2) the left hand side of (A.1) becomes

$$K^{T^\dagger}(I-QD)K^T = K^{T^\dagger}K^T = I,$$

and (A.1) is proved. Next consider the left hand side of (A.4):

$$\begin{aligned} & DQ + (I-DQ)K^\dagger K \\ &= D_0^{-1}\left(P_\perp Q + (D_0 - P_\perp Q)K^\dagger K\right) \\ &= D_0^{-1}\left(P_\perp Q + (P + P_\perp Q P_\perp - P_\perp Q)P\right) \\ &= D_0^{-1}\left(P_\perp Q + P - P_\perp Q P\right) \\ &= D_0^{-1}\left(P + P_\perp Q(I-P)\right) = D_0^{-1}D_0 = I, \end{aligned}$$

and (A.4) is proved. Finally, using (A.4), consider the left hand side of (A.3):

$$\begin{aligned} & K^{T^\dagger}(I-QD)Q - K^{T^\dagger}Q(I-DQ)K^\dagger K \\ &= K^{T^\dagger}(Q - QDQ)\left(I - K^\dagger K\right) \\ &= K^{T^\dagger}Q(I-DQ)\left(I - K^\dagger K\right) \\ &= K^{T^\dagger}Q(I-DQ) - K^{T^\dagger}Q(I-DQ) = 0, \end{aligned}$$

and the lemma is proven. □