INSTRUMENTAL VARIABLE SOLUTION TO AN EXTENDED FRISCH PROBLEM[*]

Petre Stoica ps@syscon.uu.se
Mats Cedervall mc@syscon.uu.se
Joakim Sorelius js@syscon.uu.se
Torsten Söderström ts@syscon.uu.se
Systems and Control Group, Uppsala University
PO Box 27, S-751 03 Uppsala, Sweden
Tel: +46 18 183074; fax: +46 18 503611

ABSTRACT

In signal processing and time series analysis applications we often encounter cases in which a number of (noise-free) variables are linearly related and we want to make inferences on the number and the form of the linear relations among those variables from noisy observations of them. The Frisch problem is concerned with the aforementioned inferences under the assumption that the components of the observation noise vector are mutually uncorrelated. In this paper we extend the Frisch problem by allowing the noise vector components to be correlated in an arbitrary (and unknown) way. The EXtended FRIsch problem of this paper is called EXFRI for short. To make EXFRI solvable we basically assume that the observation noise is temporally white whereas the noise-free signals are temporally correlated. We show that, under the assumptions made, the EXFRI problem has a computationally simple and statistically elegant Instrumental Variable (IV) solution, which is essentially based on a canonical correlation decomposition procedure.

1 INTRODUCTION

The following problem is encountered in several signal processing applications. We observe a number of variables whose noise-free versions are linearly related. From a set of available noisy observations we want to make inferences on the maximum number and the form of the independent linear relations among the variables under study. The Frisch problem, named after the scientist who introduced it more than half a century ago [1], is a special case of the aforementioned general problem, in which the observation noise components are assumed to be mutually uncorrelated. In Section 2 of this paper we argue that such an assumption is fairly restrictive in many applications. In order to cope with such applications we allow the noise vector components to be correlated in an arbitrary (and unknown) way. For convenience of exposition we refer to the EXtended FRIsch problem considered in this paper as to EXFRI. Of course, no meaningful inferences can be made when such a generality of the noise vector, as in EXFRI, is allowed. To make such inferences possible we basically assume that the observation noise is temporally white whereas the noise-free signal vector exhibits sufficient temporal correlation (in a sense to be made precise in the next section). Under these assumptions we show that EXFRI admits a *unique, computationally simple* and *statistically elegant solution* which mainly amounts to the computation of a canonical correlation decomposition (CCD). Note that the set of solutions to the original Frisch problem is infinite, in spite of its rather strong assumptions, and also that there exists no rigorous procedure for choosing a particular solution from that set.

2 PROBLEM FORMULATION

Consider the following data model:

$$x(t) = \tilde{x}(t) + \varepsilon(t) \quad (m \times 1) \; (t = 1, \ldots, N) \qquad (1)$$

where $\tilde{x}(t)$ denotes the noise-free signal vector, $\varepsilon(t)$ is the observation noise vector, and N is the number of available measurements. In this paper we adopt a statistical framework, and therefore assume that $\{\tilde{x}(t)\}$ and $\{\varepsilon(t)\}$ are second-order ergodic stochastic sequences. We should note that sometimes the data model in (1) is considered in a deterministic framework in which qualifications, such as "uncorrelatedness", are assumed to hold in each *finite* "realization" *exactly.* We believe that the statistical framework is more flexible because it does not make such impractical assumptions and, additionally, it allows to assess the effect of the finite-sample deviations from the asymptotic assumptions made on the solution obtained. The noise-free signal vector in (1) is assumed to satisfy the following equation:

$$B^T \tilde{x}(t) \equiv 0 \qquad (2)$$

where B^T is a $(m - n) \times m$ matrix whose rank is equal to $(m - n)$. Here n is the *smallest possible integer* for

[*] This work has been supported by the Swedish Research Council for Engineering Sciences (TFR) under contract 281-95-800, and CEC-HCM/(NFR) under contract CNRX-CT93-0412 (SIMONET)/Ö-AB/HC 03962-300.

which the above equation holds true. Let $R_{\tilde{x}\tilde{x}}$ denote the covariance matrix of $\{\tilde{x}(t)\}$,

$$R_{\tilde{x}\tilde{x}} = E\left[\tilde{x}(t)\tilde{x}^T(t)\right]. \tag{3}$$

We can make use of $R_{\tilde{x}\tilde{x}}$ to reformulate the condition (2) above as follows:

- rank $(R_{\tilde{x}\tilde{x}}) = n,$ (4)
- the columns of B form a basis of the null space of $R_{\tilde{x}\tilde{x}}$. (5)

The problem of interest here is *the estimation of* n *and* B *from* $\{x(t)\}_{t=1}^N$. Evidently B can only be determined modulo a post-multiplication by a nonsingular matrix. In other words, what we can really estimate is *the range space of* B.

The estimation problem stated above finds several applications in the signal processing and time series analysis areas [2], such as:

(a) Time series modeling and factor analysis, with particular emphasis on economic problems.

(b) System identification from noisy input-output measurements. In this application n is the order of the system, and the system parameters can be obtained from B.

(c) Array signal processing for direction-of-arrival estimation. Here n is the number of signals impinging on the array, and the directions-of-arrival of the signals can be derived from the matrix B by using MUSIC or other subspace fitting methodologies.

(d) Blind channel equalization from multiple sensor or fractionally sampled data. In this application, which is related to both (b) and (c) above, n determines the length of the channel memory, and the impulse response of the channel can be obtained from B.

The Frisch problem deals with the estimation of n and B in the data model introduced above, under the following assumptions (translated to our statistical framework):

A The elements of $\varepsilon(t)$ are mutually uncorrelated, so that the covariance matrix of $\varepsilon(t)$ is diagonal

A1 $\tilde{x}(t)$ and $\varepsilon(s)$ are statistically independent of each other, for all t and s.

Assumption A1 appears reasonable, and hence *we also adopt it* in what follows. However, assumption A of the Frisch problem may be impractical in many cases. In the application (a), since there are (linear) relations between the elements of $\tilde{x}(t)$, we might also expect some form of correlation among the observation errors corrupting $\{\tilde{x}(t)\}$. In the other applications (b)-(d), the elements of $\varepsilon(t)$ are more often than not correlated with one another. As a matter of fact, most of the current research in these applications has been driven by the desire to eliminate the "diagonal noise covariance matrix" assumption. In what follows we make *no assumption on the noise covariance matrix*, which therefore is allowed to be arbitrary and unknown. In such a general scenario, the matrix B and the integer-valued parameter n could not be recovered from the data without making some further assumptions. In addition to A1, we assume that:

A2 The vector $\varepsilon(t)$ is temporally white. That is, the sequence $\{\varepsilon(t)\}$ consists of independent and identically distributed random variables with zero mean and covariance matrix denoted by Q:

$$E\left[\varepsilon(t)\right] = 0; \quad E\left[\varepsilon(t)\varepsilon^T(s)\right] = Q\delta_{t,s} \tag{6}$$

A3 The sequence $\{\tilde{x}(t)\}$ is temporally autocorrelated in a sufficiently strong manner, so that the following condition holds true:

$$\text{rank}\left(E\left\{\tilde{x}(t)\left[\tilde{x}^T(t-1)\cdots\tilde{x}^T(t-p)\right]\right\}\right) = n \tag{7}$$

for some $p \geq 1$ (recall that $n \triangleq \text{rank}\,(R_{\tilde{x}\tilde{x}})$).

Note that the rank of the matrix in (7) cannot be larger than n. Hence A3 amounts to assuming that the aforementioned matrix achieves its maximum possible rank. Quite often $\tilde{x}(t)$ comprises time series that are strongly correlated along the time axis, hence making A3 a likely assumption. Regarding A2, the observation errors that appear at different time instants are often mutually uncorrelated, hence making (6) a plausible assumption.

3 ESTIMATION OF B

In this section we assume that n is given. From (1) and (2) we obtain:

$$B^T x(t) = B^T \varepsilon(t). \tag{8}$$

This equation may be viewed as a form of linear regression with (spatially) non-white "residuals". Furthermore, the "residuals" and the "regressors" in (8) are correlated with one another, so that (8) can be regarded as a "pseudo-linear regression". Let

$$z(t) = \left[\begin{array}{ccc} x^T(t-1) & \cdots & x^T(t-p) \end{array}\right]^T \quad (p \geq 1). \tag{9}$$

Under assumptions A1 and A2, the vector $z(t)$ above is uncorrelated with the right-hand-side of (8), and hence it can be used as an *Instrumental Variable* (IV) vector for making inferences about n and B. Define

$$\hat{R}_{xz} \triangleq \frac{1}{N}\sum_{t=p+1}^{N} x(t)z^T(t) \tag{10}$$

and similarly $\hat{R}_{xx}$ and $\hat{R}_{zz}$. Let R_{xz}, R_{xx} and R_{zz} denote the corresponding theoretical covariance matrices. From (8) we get the following IV equations:

$$B^T\hat{R}_{xz} = B^T\frac{1}{N}\sum_{t=p+1}^{N}\varepsilon(t)z^T(t) \tag{11}$$

or, in a vectorized form,

$$\left(\hat{R}_{xz}^T \otimes I\right) b = \mu. \tag{12}$$

Here, $\otimes$ denotes the Kronecker matrix product and $b = \text{vec}\,(B^T)$ where, for a matrix $X = [x_1 \cdots x_n]$ the vector $\text{vec}\,(X)$ is by definition equal to $[x_1^T \cdots x_n^T]^T$. The kth $(m-n) \times 1$ subvector of μ appearing in (12) above is given by

$$\mu_k = \frac{1}{N} \sum_{t=p+1}^{N} B^T \varepsilon(t) z_k(t) \tag{13}$$

where $z_k(t)$ is the kth component of $z(t)$ $(k = 1, \ldots, mp)$.

An *asymptotically best consistent (ABC) estimate* of b in (12) is given by the minimizer of the following criterion:

$$f = b^T \left(\hat{R}_{xz} \otimes I\right) \hat{\Omega}^{-1} \left(\hat{R}_{xz}^T \otimes I\right) b \tag{14}$$

where $\hat{\Omega}$ is a consistent estimate of

$$\Omega = E\left(\mu\mu^T\right). \tag{15}$$

In [2] we have derived an expression for the matrix $\hat{\Omega}$ required in (14) and have also shown that a matrix B minimizing (14) is given by

$$\hat{B} = \hat{R}_{xx}^{-1/2} \hat{G}. \tag{16}$$

Here $\hat{G}$ is defined through the following eigenvalue decomposition (EVD). Let

$$\hat{R} \triangleq \hat{R}_{xx}^{-1/2} \hat{R}_{xz} \hat{R}_{zz}^{-1} \hat{R}_{xz}^T \hat{R}_{xx}^{-1/2} \quad (m \times m). \tag{17}$$

Then

$$\hat{R} = [\ \overset{n}{\hat{S}}\ \ \overset{m-n}{\hat{G}}\ ] \begin{bmatrix} \hat{\lambda}_1 & & 0 \\ & \ddots & \\ 0 & & \hat{\lambda}_m \end{bmatrix} \begin{bmatrix} \hat{S}^T \\ \hat{G}^T \end{bmatrix} \tag{18}$$

where $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_m$. The corresponding minimum of f is

$$f_{\min} = \sum_{k=n+1}^{m} \hat{\lambda}_k. \tag{19}$$

Furthermore, note that the matrix B in (16) satisfies the normalization

$$B^T \hat{R}_{xx} B = I. \tag{20}$$

With this observation the derivation of *the (real-valued) parameter estimation part* of EXFRI is concluded.

It is interesting to relate the EXFRI estimation methodology previously introduced to the canonical correlation decomposition (CCD) approach in multivariate statistical analysis. The problem of estimating (the range space of) B, under the assumptions made, can also be formulated as follows: obtain a matrix B which satisfies the normalization constraint (20) and is such that the sample cross-correlation between the variates $B^T x(t)$ and the IV-vector $z(t)$ is "as small as possible" (recall that $x(t)$ and $z(t)$ are only cross-correlated via $\tilde{x}(t)$ and that the transformation of $x(t)$ via B^T must annihilate $\tilde{x}(t)$). When formulated as above, the estimation of B becomes a CCD problem, the solution of which is given by the $(m-n)$ least significant canonical variates. Hence, the expression for $\hat{B}$ in (16) follows. The drawback of formulating the estimation of B as a CCD problem is that, unlike in the ABC approach, the optimal statistical properties of the resulting estimates are no longer clear. This is particularly true in the present case where the vectors $x(t)$ and $z(t)$ are not temporally white as is usually required in the CCD analysis.

The previous analogy between EXFRI and CCD is however useful. Indeed, this analogy suggests that the estimation of n can be done (like in the CCD) by testing the significance of the $(m-n)$ smallest canonical correlations $\{\hat{\lambda}_k\}_{k=n+1}^m$. More exactly, we can estimate n as the smallest integer for which $f_{\min}$ in (19) is "insignificant". The problem is, of course, to derive a statistical rule which can be used to assess whether or not a given value of $f_{\min}$ is "insignificant". Such a rule is obtained in the next section, where we will once again have to deviate from the traditional CCD analysis owing to the already mentioned fact that the sequences $\{x(t)\}$ and $\{z(t)\}$ are temporally autocorrelated.

4 ESTIMATION OF n

Let H_0 denote the null hypothesis,

H_0 : n is the smallest integer for which (2) holds true.

Then it is possible to prove the following theorem [2].

Theorem 4.1 *Under H_0 and asymptotically in N,*

$$N f_{\min} = N \sum_{k=n+1}^{m} \hat{\lambda}_k \sim \chi^2\left((m-n)(mp-n)\right), \tag{21}$$

i.e., *$N f_{\min}$ is χ^2 distributed with $(m-n)(mp-n)$ degrees of freedom.*

The use of Theorem 4.1 to conceive *a test for inferring the value of n* from the available observations $\{x(t)\}_{t=1}^N$ is straightforward. Below we put together the proposed EXFRI-based approaches for estimating n and B, and summarize them in a step by step fashion for the reader's convenience.

Algorithm: EXFRI

1. Choose a $p \geq 1$, and compute $\hat{R}_{xx}$, $\hat{R}_{xz}$, $\hat{R}_{zz}$ and the EVD (18) of the matrix $\hat{R}$ in (17).
   Set $n = 0$.
2. Check whether

$$N \sum_{k=n+1}^{m} \hat{\lambda}_k \leq \chi_\alpha^2\left((m-n)(mp-n)\right) \tag{22}$$

where the threshold $\chi^2_\alpha(\cdot)$ is defined through the equality $\text{prob}\left(u \geq \chi^2_\alpha(k) \mid u \sim \chi^2(k)\right) = \alpha$ (recommended interval: $\alpha \in [0.01, 0.05]$).

- If (22) is not satisfied and $n < m-1$ then set $n \doteq n+1$ and go to the beginning of Step 2.
- If (22) is not satisfied and $n = m-1$ then set $n = m$ and go to End.
- If (22) is satisfied then the current value of n and the corresponding matrix $\hat{B}$ computed with (16) are the EXFRI estimates of n and B.

End.

5 NUMERICAL EXAMPLE

In this section we will assess the performance of the EXFRI methodology by means of a simulated example. The noise free signal vector is chosen as

$$\begin{aligned}\tilde{x}(t) \quad = \quad & [\tilde{x}_1(t)\ \tilde{x}_2(t) \cdots \tilde{x}_{10}(t) \\ & \tilde{x}_1(t) - \tilde{x}_2(t)\ \tilde{x}_3(t) + \tilde{x}_4(t)\ \tilde{x}_5(t) + 2\tilde{x}_6(t) \\ & \tilde{x}_7(t) - 2\tilde{x}_8(t)\ \tilde{x}_9(t) + 3\tilde{x}_{10}(t)\ ]^T.\end{aligned}$$

where $\{\tilde{x}_i\}_{i=1}^{10}$ are generated as uncorrelated, Gaussian AR-signals having zero mean and unit variance. Hence, $m = 15$ and $n = 10$. The covariance matrix of the observation noise vector is chosen as

$$Q_{ij} = \sigma^2 \left[\rho^{|i-j|}\right] \tag{23}$$

with $\rho = 0.9$. The signal to noise ratio (SNR) is defined by

$$SNR = 10\log(1/\sigma^2) \tag{24}$$

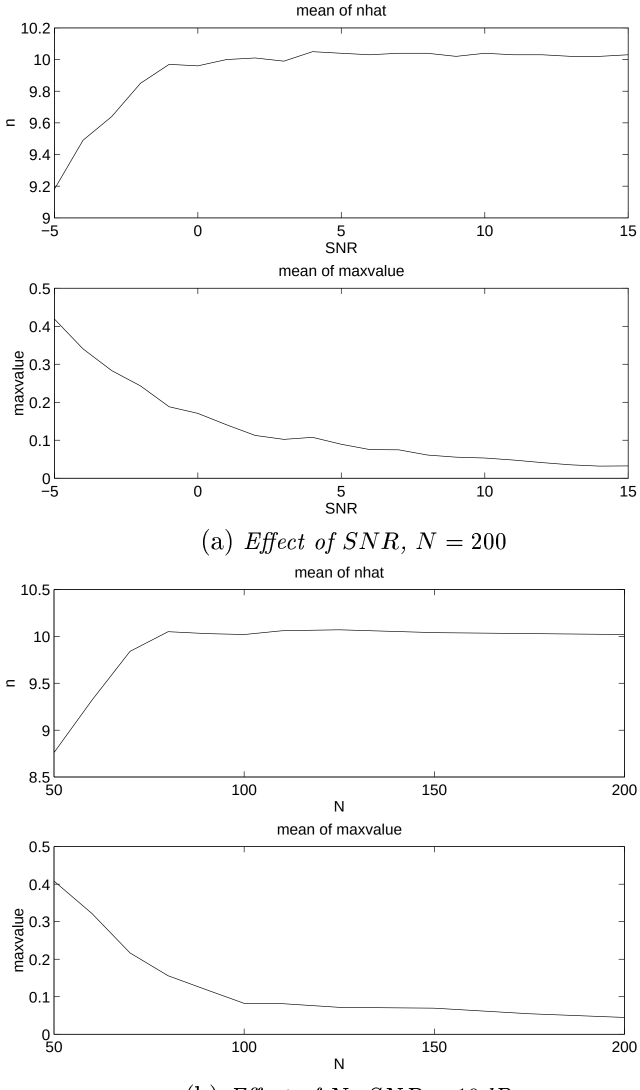
and is expressed in decibels (dB). The probability of false alarm needed in (22) is choosen as $\alpha = 0.05$. We reiterate that we can uniquely estimate only the orthogonal projector onto the range space of B:

$$\Pi \triangleq B(B^TB)^{-1}B^T. \tag{25}$$

The EXFRI methodology is applied to 100 independent realizations of the data (generated with the above parameters) and the following averaged quantities are computed as a function of the SNR and the number of data points N:

- averaged $\hat{n}$
- averaged $\max_{i,j}\left|\left(\widehat{\Pi}_{ij} - \Pi_{ij}\right)\right|$

where the hat denotes an estimated quantity. The latter average is evaluated over all realizations, regardless whether $\hat{n} = n_{\text{true}}$ or not. The results for $p = 1$ (*c.f.* (9)) are shown in Figure 1. The results for larger values of p (not shown here) were similar, as expected.



(a) *Effect of SNR*, $N = 200$

(b) *Effect of* N, $SNR = 10dB$

Figure 1: Average $\hat{n}$ and $\max_{i,j}|(\widehat{\Pi}_{ij} - \Pi_{ij})|$ as a function of SNR (a) and N (b) for $p = 1$.

References

[1] R. Frisch. "Statistical confluence analysis by means of complete regression systems," Technical Report Pub. No. 5, Economic Institute, University of Oslo, 1934.

For the complete reference list, which is omitted here to save space, we refer to:

[2] P. Stoica, J. Sorelius, M. Cedervall and T. Söderström. "Instrumental variable solution to an extended Frisch problem," Technical Report UPTEC 96059R, Department of Technology, Uppsala University, Sweden, 1996.