2024 European Control Conference (ECC)
June 25-28, 2024. Stockholm, Sweden

Relations between prediction error and maximum likelihood methods in an error-in-variables setting

Torsten Söderström[1]

Abstract— **Prediction error (PE) and maximum likelihood (ML) methods are often treated as synonyms when identifying linear dynamic systems from Gaussian data. It is shown how these methods differ when specifically dealing with errors-in-variables problems. These problems can modeled using multivariable times series with a specific internal structure. In such situations the ML estimates have lower variances than the PE estimates. Explicit expressions for the covariance matrices of the estimates are given and analyzed. For the special case when the unperturbed input is white noise it is shown that the system is not identifiable when the PEM estimate is used, while the ML estimates still have quite small variances. In such situations ML is thus much superior to the PE estimates. Another special case concerns non-Gaussian data. In that case a pseudo-ML estimate (using the ML criterion as if the data were Gaussian) will no longer be superior to the PE estimate in terms of error variances.**

Index Terms— **System identification, Errors-in-variables, Maximum Likelihood, Prediction errors**

I. Introduction

The prediction error method (PEM) and the maximum likelihood) ML method are both very well-known in the system identification literature, [1], [7]. In standard situations they coincide and consequently give identical estimates.

In an errors-in-variables (EIV) situation, all standard identification methods yield biased (rather, non-consistent) estimates due to the measured input signal containing additional noise, [5], [4]. PEM and ML can be applied also in an EIV situation after appropriate modifications. However, then typically PEM and ML are no longer equivalent. This paper discusses the difference in behavior between these two estimators in such situations.

The paper is organized as follows. The standard case (non EIV) is reviewed in Section II, and the EIV setting is described in Section III. Formal definitions and algorithms of the estimators are presented in Section IV. Section V reviews the general results on asymptotic distribution of the parameter estimates for Gaussian distributed data. The special case when the unperturbed input is white noise is coped with in Section VI, while the case with non-Gaussian data is treated in Section VII. Finally, conclusions are offered in Section VIII. Due to space limitations, most proofs are omitted, but they can all be found in [6].

II. The standard case

Consider the multivariable system (in time-series form)

$$\begin{aligned} y(t) &= H(q,\theta)e(t) & (1)\\ \mathsf{E}\left\{e(t)e^T(t)\right\} &= \Lambda & (2)\end{aligned}$$

Here $y(t)$ denotes the measured outputs. The transfer function operator H is a function of the shift operator q and is parameterized with the parameter vector θ. The signal $e(t)$ denotes zero mean white noise. The model (1) is assumed to be in innovations form, meaning that $H(q,\theta) = I + \sum_{i=1}^{\infty} H_i(\theta)q^{-i}$, and that $H(q,\theta)$ as well as $H^{-1}(q,\theta)$ are asymptotically stable.

It is possible to include a term $G(q,\theta)u(t)$ in (1), where $u(t)$ denotes the input. To comply with the future treatment in this paper we stick to (1) as a general description. Note that it is possible to let the elements of the vector $y(t)$ in (1) contain both system inputs and system outputs, see Section III.

Under the above assumptions, the one-step prediction error becomes

$$\varepsilon(t,\theta) = H^{-1}(q,\theta)y(t) \tag{3}$$

Two common approaches to identify the system is to apply the prediction error method (PEM) or the maximum likelihood (ML) method.

The PEM estimate is designed to minimize the sample covariance matrix $R_\varepsilon(\theta)$ of the prediction errors

$$R_\varepsilon(\theta) = \frac{1}{N}\sum_{t=1}^{N} \varepsilon(t,\theta)\varepsilon^T(t,\theta) \tag{4}$$

For example, one may consider

$$\hat{\theta} = \arg\min_{\theta} h(R_\varepsilon(\theta)) \tag{5}$$

where $h(R)$ is a positive function. The choice of h will have effect on the asymptotic covariance matrix of $\hat{\theta}$. A typical choice is

$$h(R) = \mathrm{tr}(SR) \tag{6}$$

where S is a user-chosen weighting matrix. It is known, [1], [7], how the covariance matrix of $\hat{\theta}$ depends on S, and that it is minimized for the choice

$$S = \Lambda^{-1} \tag{7}$$

Note that this choice is not practical to use, as Λ is not a priori known. Further, the asymptotic covariance matrix

[1] Torsten Söderström is with Division of Systems and Control, Department of Information Technology, Uppsala University, P O Box 337, SE-751 05 Uppsala, Sweden torsten.soderstrom@it.uu.se

using the criterion (6) with the choice (7) is the same as using the criterion

$$h(R) = \det(R) \tag{8}$$

We therefore consider here the PEM estimate to be

$$\hat{\theta} = \arg \min_{\theta} \det \left(\frac{1}{N} \sum_{t=1}^{N} \varepsilon(t,\theta)\varepsilon^T(t,\theta) \right) \tag{9}$$

$$\hat{\Lambda} = \hat{R}_{\varepsilon}(\hat{\theta}) \tag{10}$$

The ML model is obtained by maximizing the likelihood function $L(\theta, \Lambda)$. It can equivalently be expressed by minimizing the negative logarithm of L. Assuming the data to be Gaussian distributed, it holds (up to a constant)

$$-\log(L(\theta,\Lambda)) = \frac{1}{2}\sum_{t=1}^{N} \varepsilon^T(t,\theta)\Lambda^{-1}\varepsilon(t,\theta) + \frac{N}{2}\log(\det(\Lambda)) \tag{11}$$

The criterion in (11) can be minimized with respect to Λ, and then in a second step with respect to θ. The result happens to lead precisely to the PEM estimate (9), (10), see also [1].

Thus for Gaussian data, PEM and ML give identical estimates.

When data are not Gaussian, the true ML method corresponds to minimization of another criterion than (11). Minimization of (11) can still constitute a meaningful estimator. We will for such cases label it pseudo-maximum likelihood (pML) in this paper.

III. Errors-in-variables models

The errors-in-variables (EIV) identification problem relates to the situation when the input signal cannot be measured without error. For details see, [5].

The EIV situation can be modelled as follows. The system and its measurements are given by

$$y_0(t) = G(q)u_0(t) \tag{12}$$

$$u(t) = u_0(t) + \tilde{u}(t) \tag{13}$$

$$y(t) = y_0(t) + \tilde{y}(t) \tag{14}$$

To apply a PEM or an ML method, we will also need a model for the unperturbed input $u_0(t)$. Here it will be described as an ARMA model

$$u_0(t) = K(q)v(t) \tag{15}$$

where $K(q)$ is a ratio of two monic polynomials, and $v(t)$ is zero mean white noise.

It is assumed that the three noise sources $\tilde{u}(t)$, $\tilde{y}(t)$ and $v(t)$ are white, independent, of zero mean, and with unknown variances λ_u^2, λ_y^2 and λ_v^2, respectively.

Regard now the measured input-output data $z(t) = \begin{pmatrix} y(t) & u(t) \end{pmatrix}^T$ as outputs of a multivariable and structured ARMA process. Then it can be written in the form

$$z(t) = H(q)\varepsilon(t) \tag{16}$$

This can be done by first writing (13)-(15) as

$$z(t) = \begin{pmatrix} y(t) \\ u(t) \end{pmatrix} = \begin{pmatrix} G(q) \\ 1 \end{pmatrix} u_0(t) + \begin{pmatrix} \tilde{y}(t) \\ \tilde{u}(t) \end{pmatrix} \tag{17}$$

To get the description (16) we next apply spectral factorization

$$\Phi_z = \begin{pmatrix} G \\ 1 \end{pmatrix} \Phi_{u_0} \begin{pmatrix} G^* & 1 \end{pmatrix} + \begin{pmatrix} \lambda_y^2 & 0 \\ 0 & \lambda_u^2 \end{pmatrix} = H\Lambda H^* \tag{18}$$

where H and H^{-1} are restricted to be asymptotically stable, and $\lim_{q\to\infty} H(q) = I$. The filter $H(q)$ will depend on all the unknown quantities in $G(q)$, $K(q)$ and the variances. It will though be unaffected if all the three variances are multiplied by the same factor. For these reasons introduce the parameter vector η as

$$\eta = \begin{pmatrix} \theta \\ r \end{pmatrix} \tag{19}$$

where $r = \lambda_v^2$ will correspond the unknown common level of the three variances. Further the vector θ will include all unknown parameters of $G(q)$ and $K(q)$ as well as the variance ratios λ_u^2/λ_v^2 and λ_y^2/λ_v^2.

As a result we have the following innovations form model for the measured data

$$z(t) = H(q,\theta)\varepsilon(t,\theta) \tag{20}$$

$$\mathsf{E}\left\{\varepsilon(t,\theta)\varepsilon^T(t,\theta)\right\} = \Lambda(\theta,r) = rQ(\theta) \tag{21}$$

Note that the previous reasoning shows that $\Lambda(\theta, r)$ is proportional to r, and (21) can be taken as a definition of $Q(\theta)$. In the treatment to follow it is assumed that z and ε are vectors of dimension n (even if $n = 2$ is the case that specifically apply to the EIV situation for a single-input single-output system), and that θ is a vector of dimension n_θ.

We note in passing that how to compute the innovations model (20) from the description (13)-(15) is based on spectral factorization. It can, for example, be carried out by first writing the total system in state space form, and then solve for the associated Kalman filter, see [2] and [4] for details.

IV. Estimation algorithms

Note that the model (20)-(21) differs from the one in Section II, see (1), (2). The essential difference is that in (21) the covariance matrix Λ carries some information about θ. This turns out to be useful. It will lead to a difference between PEM and ML, as explained in what follows.

PEM is still obtained by minimizing the covariance matrix of the prediction errors. This step is complemented with a way to estimate the scalar r.

ML is still obtained by minimizing the negative logarithm of the likelihood function. Now one can no longer minimize it separately with respect to Λ.

A. Prediction error method

The estimate of θ is defined as before, see (9). It needs to be complemented with an estimate of r. It is argued in [3] that an appropriate PEM estimate of η is

$$\hat{\theta} = \arg\min_{\theta} \det\left(\frac{1}{N}\sum_{t=1}^{N}\varepsilon(t,\theta)\varepsilon^{T}(t,\theta)\right) \tag{22}$$

$$\hat{r} = \frac{1}{nN}\sum_{t=1}^{N}\varepsilon(t,\hat{\theta})^{T}Q^{-1}(\hat{\theta})\varepsilon(t,\hat{\theta}) \tag{23}$$

$$\hat{\eta}_{\mathrm{PEM}} = \left(\begin{array}{c}\hat{\theta}\\ \hat{r}\end{array}\right) \tag{24}$$

Assuming $\hat{\theta}$ is consistent ($\hat{\theta} \to \theta$ as $N \to \infty$), it follows from a short calculation that $\hat{r}$ in (23) is consistent as well.

B. Maximum likelihood method

The ML estimate is still defined as the minimizing element of the negative log-likelihood function. This means that

$$\left(\hat{\theta},\hat{r}\right) = \arg\min_{\theta,r}\left[\frac{1}{2}\sum_{t=1}^{N}\varepsilon^{T}(t,\theta)\Lambda^{-1}(\theta,r)\varepsilon(t,\theta) + \frac{N}{2}\log(\det(\Lambda(\theta,r)))\right] \tag{25}$$

$$\hat{\eta}_{\mathrm{ML}} = \left(\begin{array}{c}\hat{\theta}\\ \hat{r}\end{array}\right) \tag{26}$$

The asymptotic accuracies, measured as $\mathrm{cov}\left(\hat{\eta}_{\mathrm{PEM}}\right)$ and $\mathrm{cov}\left(\hat{\eta}_{\mathrm{ML}}\right)$ are analyzed in Sections V, VI and VII.

V. Analysis in case of Gaussian data

The parameter estimates $\hat{\eta}_{\mathrm{PEM}}$ and $\hat{\eta}_{\mathrm{ML}}$ are both asymptotically Gaussian distributed, cf [1].

Set

$$\psi(t) = \frac{\partial\varepsilon(t,\theta)}{\partial\theta} \tag{27}$$

$$M = \mathsf{E}\left\{\psi^{T}(t)\Lambda^{-1}\psi(t)\right\} \tag{28}$$

Note that $\psi(t)$ is an $n \times n_\theta$ matrix, and M is an $n_\theta \times n_\theta$ matrix.

The following lemma describes the asymptotic distributions (as $N \to \infty$). The main steps of the proof are based on the general analysis in [1].

Lemma 1. Assume that the matrix M in (28) is invertible. The estimates are asymptotically Gaussian distributed, as

$$\sqrt{N}\left(\hat{\eta}_{\mathrm{PEM}} - \eta\right) \sim \mathsf{N}(0, C_{\mathrm{PEM}}) \tag{29}$$

$$\sqrt{N}\left(\hat{\eta}_{\mathrm{ML}} - \eta\right) \sim \mathsf{N}(0, C_{\mathrm{ML}}) \tag{30}$$

where $\sim$ denotes convergence in distribution.

The covariance matrix C_{PEM} can be written as

$$C_{\mathrm{PEM}} = P_{\eta} = \left(\begin{array}{cc}P_{\theta} & P_{\theta r}\\ P_{\theta r}^{T} & P_{r}\end{array}\right) \tag{31}$$

$$P_{\theta} \triangleq M^{-1} \tag{32}$$

$$P_{r} \triangleq N\mathrm{var}\left(\hat{r}\right) = \frac{2r^{2}}{n} + b^{T}P_{\theta}b \tag{33}$$

$$P_{\theta,r} \triangleq N\mathrm{cov}\left(\hat{\theta},\hat{r}\right) = -P_{\theta}b \tag{34}$$

where

$$b \triangleq \frac{r}{n}\left(\ \mathrm{tr}\left(\Lambda_{1}\Lambda^{-1}\right)\ \ \ldots\ \ \mathrm{tr}\left(\Lambda_{n_\theta}\Lambda^{-1}\right)\ \right)^{T} \tag{35}$$

$$\Lambda_{i} \triangleq \frac{\partial\Lambda(\theta)}{\partial\theta_{i}}, \quad i = 1,\ldots,n_\theta \tag{36}$$

The asymptotic covariance matrix for the ML estimate can be found to be

$$C_{\mathrm{ML}} = (S+R)^{-1} \tag{37}$$

$$S = \mathsf{E}\left\{\left(\frac{\partial\varepsilon(t)}{\partial\eta}\right)^{T}\Lambda^{-1}\frac{\partial\varepsilon(t)}{\partial\eta}\right\} \tag{38}$$

$$R_{i,j} = \frac{1}{2}\mathrm{tr}\left(\frac{\partial\Lambda(\eta)}{\partial\eta_{i}}\Lambda^{-1}\frac{\partial\Lambda(\eta)}{\partial\eta_{j}}\Lambda^{-1}\right), \quad i,j = 1,\ldots,n_\theta+1 \tag{39}$$

The matrix R can be partitioned as

$$R = \left(\begin{array}{cc}R_{11} & R_{12}\\ R_{12}^{T} & R_{22}\end{array}\right) \tag{40}$$

with R_{22} a scalar, and (for $i,j = 1,\ldots,n_\theta$)

$$\left(R_{11}\right)_{i,j} = \frac{1}{2}\mathrm{tr}\left(Q_{i}Q^{-1}Q_{j}Q^{-1}\right) \tag{41}$$

$$\begin{aligned}\left(R_{12}\right)_{i} &= \frac{1}{2}\mathrm{tr}\left(rQ_{i}\frac{1}{r}Q^{-1}Q\frac{1}{r}Q^{-1}\right)\\ &= \frac{1}{2r}\mathrm{tr}\left(Q_{i}Q^{-1}\right) = \frac{n}{2r^{2}}b_{i}\end{aligned} \tag{42}$$

$$R_{22} = \frac{1}{2}\mathrm{tr}\left(Q\Lambda^{-1}Q\Lambda^{-1}\right) = \frac{n}{2r^{2}} \tag{43}$$

Further, the ML estimate satisfies the Cramér-Rao lower bound, meaning that

$$C_{\mathrm{PEM}} - C_{\mathrm{ML}} \geq 0 \tag{44}$$

that is, the left hand side of (44) is a nonnegative definite matrix. ■

A consequence of (44) is that any linear parameter combination of $\hat{\eta}_{\mathrm{PEM}}$ has at least as large variance as the same combination of $\hat{\eta}_{\mathrm{ML}}$.

Next we will examine the difference in (44) closer. Will the difference between the two covariance matrices sometimes be significantly large? What happens with the accuracies if the estimates are applied to non-Gaussian data? These issues are treated in the coming two sections, respectively.

VI. Special case of white input signal

Assume here that the unperturbed input signal $u_0(t)$ is white noise, and that all data are Gaussian. Set

$$G(q) = B(q)/A(q) \tag{45}$$

and assume that the polynomials $A(q)$ and $B(q)$ are coprime. For this case we will show that the matrix M in (28) is singular. The consequence is that the system is then *not identifiable when PEM is used*, and $P_\theta = M^{-1}$ cannot be computed. In a sense, for this special case therefore PEM has much (not to say infinitely) worse accuracy than ML.

For the special case under study, one can find explicit expressions for the innovations form (20), (21). It holds that

$$\Lambda = \begin{pmatrix} \Lambda_{11} & 0 \\ 0 & \lambda_u^2 + \lambda_v^2 \end{pmatrix} \qquad (46)$$

$$H = \begin{pmatrix} C/A & \frac{\lambda_u^2 \lambda_v^2}{\lambda_u^2+\lambda_v^2} B/A \\ 0 & 1 \end{pmatrix} \qquad (47)$$

$$H^{-1} = \begin{pmatrix} A/C & -\frac{\lambda_u^2 \lambda_v^2}{\lambda_u^2+\lambda_v^2} B/C \\ 0 & 1 \end{pmatrix} \qquad (48)$$

where Λ_{11} and the monic polynomial C are defined as the solution to the spectral factorization equation

$$\Lambda_{11} C(z) C(z^{-1}) = \lambda_y^2 A(z) A(z^{-1}) + \frac{\lambda_u^2 \lambda_v^2}{\lambda_u^2 + \lambda_v^2} B(z) B(z^{-1}) \qquad (49)$$

The parameter vector is set as

$$\theta = \begin{pmatrix} a_1 & \dots & a_{n_a} & b_1 & \dots & b_{n_b} & \lambda_u^2/\lambda_v^2 & \lambda_y^2/\lambda_v^2 \end{pmatrix}^T \qquad (50)$$

meaning also that $r = \lambda_v^2$.

We now have the following result.

Lemma 2. When $u_0(t)$ is white noise, and $G(q)$ is given as in (45), the matrix M is singular. ■

Recall that the lemma implies that the system is not identifiable when PEM is used.

The lack of identifiability when PEM is used for white unperturbed input, can also be analyzed by using the innovations form model (47)-(48), see [6] for details.

Some consequences of the lemma are discussed in Section VIII.

VII. Analysis for the non-Gaussian case

It was shown earlier that $C_{\mathrm{PEM}} - C_{\mathrm{ML}} \geq 0$, assuming Gaussian data. Consider now the same estimates as before, but allow the innovations to have a general distribution. We are then to compare PEM and pML. We will first derive expressions for the covariance matrices C_{PEM} and C_{pML} in the non-Gaussian case. Thereafter we will check the difference of these two matrices and see whether it is still nonnegative definite, or if it has another character.

Let the generic estimate $\hat{\eta}$ be the minimizing element of the criterion $V_N(\eta)$. Its covariance matrix is found as

$$\lim_{N\to\infty} N \mathrm{cov}(\hat{\eta}) = V_\infty^{''}(\eta)^{-1} W V_\infty^{''}(\eta)^{-1} \qquad (51)$$

$$W = \lim_{N\to\infty} N \mathsf{E}\left\{ V_N'(\eta) V_N'(\eta)^T \right\} \qquad (52)$$

Most of the previous analysis (but not all!!) will still apply when the matrices in (51) and (52) are to be evaluated.

A. Some notations and preliminary results

Consider a multivariable generic case, (20), (21). Set here

$$Q_i = \frac{\partial}{\partial \eta_i} Q, \quad \Lambda_i = \frac{\partial}{\partial \eta_i} \Lambda, \quad Q_{i,j} = \frac{\partial^2}{\partial \eta_i \partial \eta_j} Q \qquad (53)$$

As a preparation we have the following result.

Lemma 3. Let x be an n-dimensional random variable, with zero mean and covariance matrix Λ. Then

$$\mathsf{E}\left\{ x^T \Lambda^{-1} x \right\} = n \qquad (54)$$

$$\mathsf{E}\left\{ \left(x^T \Lambda^{-1} x\right)^2 \right\} = (1+\beta) n^2 \qquad (55)$$

for some positive value of β. ■

Introduce the matrix F, of dimension $n_\theta \times n_\theta$ by

$$F_{i,j} = \mathsf{E}\left\{ \varepsilon^T(t) Q^{-1} Q_i Q^{-1} \varepsilon(t) \varepsilon^T(t) Q^{-1} Q_j Q^{-1} \varepsilon(t) \right\} \qquad (56)$$

and the vector g of dimension n_θ:

$$g_i = \mathsf{E}\left\{ \varepsilon^T(t) Q^{-1} Q_i Q^{-1} \varepsilon(t) \varepsilon^T(t) Q^{-1} \varepsilon(t) \right\} \qquad (57)$$

where $\varepsilon(t)$ is a vector-valued white noise of zero mean and covariance matrix Λ.

We have the following result.

Lemma 4. If $\varepsilon(t)$ is Gaussian distributed, then

$$\beta = \frac{2}{n} \qquad (58)$$

$$F = n^2 b b^T + 4 r^2 R_{11} \qquad (59)$$

$$g = (n^2 + 2n) r b \qquad (60)$$

■

B. Analysis of PEM

The PEM estimate of η is defined as in (22). The asymptotic covariance matrix of the PEM estimate $\hat{\eta}_{\mathrm{PEM}}$ is given in (31). The modification here is the first term of P_r, that involves fourth order moments of $\varepsilon(t)$.

Lemma 5. The first term for P_r will for non-Gaussian data be

$$P_r = \beta r^2 \qquad (61)$$

■

C. Analysis of pML

The pML estimate of η is given by (26).

First we should calculate the Hessian of $V_\infty(\eta)$ as well as the asymptotic covariance matrix of the gradient $V_N^{'}(\eta)$.

Differentiation with respect to $\theta_i, i = 1, \dots, n_\theta$ gives

$$\frac{\partial V_N(\theta, r)}{\partial \theta_i} = \frac{1}{Nr} \sum_{t=1}^{N} \varepsilon^T(t,\theta) Q^{-1} \frac{\partial \varepsilon(t,\theta)}{\partial \theta_i}$$
$$- \frac{1}{2Nr} \sum_{t=1}^{N} \varepsilon^T(t,\theta) Q^{-1}(\theta) Q_i(\theta) Q^{-1}(\theta) \varepsilon(t,\theta)$$
$$+ \frac{1}{2} \frac{1}{\det(Q(\theta))} \det(Q(\theta)) \mathrm{tr}\left(Q^{-1}(\theta) Q_i(\theta) \right) \qquad (62)$$

Differentiation with respect to r gives

$$\frac{\partial V_N(\theta, r)}{\partial r} = -\frac{1}{2Nr^2}\sum_{t=1}^{N}\varepsilon^T(t,\theta)Q^{-1}(\theta)\varepsilon(t,\theta) + \frac{n}{2r} \quad (63)$$

Next we proceed to find the Hessian in the asymptotic case (that is, when $N \to \infty$). The result is in the following lemma.

Lemma 6. For pML the Hessian is in the asymptotic case ($N \to \infty$) given by ($i, j = 1, \dots, n_\theta$)

$$\begin{aligned} \frac{\partial^2}{\partial\theta_i\partial\theta_j}V_\infty(\theta, r) &= (M + R_{11})_{i,j} && (64)\\ \frac{\partial^2}{\partial\theta_i\partial r}V_\infty(\theta, r) &= \frac{n}{2r^2}b_i && (65)\\ \frac{\partial^2}{\partial r^2}V_\infty(\theta, r) &= \frac{n}{2r^2} && (66)\end{aligned}$$

■

Note that the expression for the Hessian is, as expected, indeed the same as in the Gaussian case, cf. the lemma to (37).

Next we proceed with finding the asymptotic normalized covariance matrix of the gradient. Set

$$W = \begin{pmatrix} W_{11} & W_{12} \\ W_{21} & W_{22}\end{pmatrix} = \lim_{N\to\infty} N\mathsf{E}\left\{\frac{\partial V_N}{\partial \eta}\frac{\partial V_N}{\partial \eta}\right\} \quad (67)$$

Lemma 7. It holds for $i, j = 1, \dots, n_\theta$

$$\begin{aligned} (W_{11})_{i,j} &= M_{i,j} + \frac{1}{4r^2}F_{i,j} - \frac{n^2}{4r^2}b_i b_j && (68)\\ (W_{12})_i &= \frac{1}{4r^3}\left(g_i - n^2 b_i r\right) && (69)\\ W_{22} &= \frac{\beta n^2}{4r^2} && (70)\end{aligned}$$

■

To summarize the analysis so far, we have (see Lemmas 6 and 7):

$$\frac{\partial^2 V_\infty}{\partial \eta^2} = \begin{pmatrix} M + R_{11} & \frac{n}{2r^2}b \\ \frac{n}{2r^2}b^T & \frac{n}{2r^2}\end{pmatrix} \quad (71)$$

$$\begin{aligned} &\lim_{N\to\infty} N\mathsf{E}\left\{\left(\frac{\partial V_N}{\partial \eta}\right)^T\frac{\partial V_N}{\partial \eta}\right\}\\ &= \begin{pmatrix} M + \frac{1}{4r^2}F - \frac{n^2}{4r^2}bb^T & \frac{1}{4r^3}(g - n^2 rb) \\ \frac{1}{4r^3}(g - n^2 rb)^T & \frac{1}{4r^2}\beta n^2\end{pmatrix} \quad (72)\end{aligned}$$

D. Comparison

We are now finally set to compare the covariance matrices $C_{\rm PEM}$ and $C_{\rm ML}$. To this aim we evaluate the matrix

$$\begin{aligned} S &\triangleq \frac{\partial^2 V_{\rm ML}}{\partial \eta^2}\left[C_{\rm PEM} - C_{\rm ML}\right]\frac{\partial^2 V_{\rm ML}}{\partial \eta^2}\\ &= \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22}\end{pmatrix} && (73)\end{aligned}$$

It will be convenient to examine the matrix S rather than the difference $C_{\rm PEM} - C_{\rm pML}$. These two matrices will have similar properties, in particular concerning the existence of positive and negative eigenvalues.

We have the following result.

Lemma 8

Let A, B and P be symmetric matrices related as

$$A = PBP \quad (74)$$

and let P be positive definite. Then

a) If B has a strictly positive eigenvalue, so has A.
b) If B has a strictly negative eigenvalue, so has A.
c) If B has a zero eigenvalue, so has A.
d) If B is positive definite (all eigenvalues strictly positive), so is A.
e) If B is negative definite (all eigenvalues strictly negative), so is A.
f) If B is positive semidefinite (smallest eigenvalue equal zero), so is A.
g) If B is negative semidefinite (largest eigenvalue equal zero), so is A.
h) If B is indefinite (smallest eigenvalue strictly negative, largest eigenvalue strictly positive), so is A.

■

Concerning the properties of the matrix S we have the following result.

Lemma 9. The blocks of the matrix S satisfy

$$\begin{aligned} S_{22} &= 0 && (75)\\ S_{12} &= \frac{1}{4r^3}\left[(\beta+1)n^2 rb - g\right] && (76)\end{aligned}$$

■

We will make use of the following result.

Lemma 10 Let the symmetric, partitioned matrix S fulfill

$$S = \begin{pmatrix} S_{11} & S_{12} \\ S_{12}^T & 0\end{pmatrix} \quad (77)$$

where S_{12} is a nonzero column vector. Then the matrix S is indefinite. ■

Summing up for the moment, we note

- $S_{22} = 0$ implies that S is indefinite, unless it holds $S_{12} = 0$.
- In the Gaussian case, the quantities fulfill

$$\beta = \frac{2}{n}, \quad g = (n^2 + 2n)rb, \quad F = n^2 bb^T + 4r^2 R_{11} \quad (78)$$

and thus

$$\begin{aligned} S_{12} &= \frac{\beta n^2}{4r^2} b - \frac{1}{4r^3} g + \frac{n^2}{4r^2} b \\ &= \frac{1}{4r^3} \left[2nrb + n^2 rb - (n^2 + 2n) rb \right] \\ &= 0 \end{aligned} \tag{79}$$

- In general, unless

$$g = (\beta + 1) n^2 rb \tag{80}$$

S will be indefinite, cf (76).

What does (80) mean? It is equivalent to

$$\begin{aligned} &\mathsf{E}\left\{ \varepsilon^T Q^{-1} Q_i Q^{-1} \varepsilon \varepsilon^T Q^{-1} \varepsilon \right\} \\ &= \frac{1}{n^2} \mathsf{E}\left\{ \left(\varepsilon^T \Lambda^{-1} \varepsilon \right)^2 \right\} n^2 r \frac{r}{n} \operatorname{tr}\left(Q_i Q^{-1} \right) \end{aligned} \tag{81}$$

or rewritten as

$$\begin{aligned} &n \mathsf{E}\left\{ \varepsilon^T Q^{-1} Q_i Q^{-1} \varepsilon \varepsilon^T Q^{-1} \varepsilon \right\} \\ &= \mathsf{E}\left\{ \left(\varepsilon^T Q^{-1} \varepsilon \right)^2 \right\} \operatorname{tr}\left(Q_i Q^{-1} \right) \end{aligned} \tag{82}$$

Set

$$x = Q^{-1/2} \varepsilon, \quad W = Q^{-1/2} Q_i Q^{-1} \tag{83}$$

The ratio of the two sides in (82) can then be written as

$$\delta = \frac{n \mathsf{E}\left\{ x^T W x x^T x \right\}}{\mathsf{E}\left\{ (x^T x)^2 \right\} \operatorname{tr}(W)} \tag{84}$$

Consider now the special case when x has a point-wise distribution. Then

$$\begin{aligned} \delta &= \frac{n (x^T W x)(x^T x)}{(x^T x)^2 \operatorname{tr}(X)} \\ &= \frac{x^T W x}{x^T x} \frac{n}{\operatorname{tr}(W)} \end{aligned} \tag{85}$$

If x is the eigenvector associated with the smallest eigenvalue of W we have

$$\delta = \frac{n \lambda_{\min}(W)}{\sum_i \lambda_i(W)} \tag{86}$$

while the case of x being the eigenvector associated with the largest eigenvalue leads to

$$\delta = \frac{n \lambda_{\max}(W)}{\sum_i \lambda_i(W)} \tag{87}$$

Unless W is proportional to I, we have $\delta < 1$ in (86) and $\delta > 1$ in (87). We can thus conclude that $S_{12} \neq 0$ in general.

To summarize the analysis so far, we find that in the non-Gaussian case, then normally S as well as the difference $C_{\text{PEM}} - C_{\text{pML}}$ will be indefinite (and thus have both positive and negative eigenvalues). This means that there is then no strict 'order relation' between the two covariance matrices.

One can also comment that for the special case of Gaussian noise, it is already known that (cf Section V) that S is nonnegative definite. We therefore have in that case $S_{22} = 0$ (as in (75)) and $S_{12} = 0$ (as in (79)).

VIII. Conclusions and summarizing discussion

When standard identification methods are applied to input-output data that are noise-corrupted, biased parameter estimates occur due to the presence of input noise.

The prediction error (PE) and the maximum likelihood (ML) estimates have been considered for some multivariable times series models with internal structure. Such models appear for a general set of errors-in-variables problems in system identification. The considered model are characterized using the innovations form. The innovation filter as well as the innovation covariance matrix depend on the unknown parameter vector. That both quantities depend on the parameter vector makes PEM and ML to differ. There is a further parameter that comes in as a scaling factor of the innovation covariance matrix.

Four different cases have been considered.

- For Gaussian data in standard (non-EIV) situations, PEM and ML coincide. This case is well-known. [Section II]
- For Gaussian data in an EIV situation, PEM and ML differ due to the fact that the innovations covariance matrix Λ carries additional information about the parameter vector. This fact is more efficiently exploited in ML than in PEM. The covariance matrix of the parameter estimates is smaller for ML than for PEM. [Section V]
- For Gaussian data in an EIV situation where the unperturbed input signal is white noise, the system is not identifiable when PEM is applied. Thus ML is much superior to PEM for such cases. Numerical experimentation suggest that this also applies to cases where the unperturbed input is close to white ($K(q)$ in (15) has all poles close to the origin). [Section VI]
- Finally, the situation of non-Gaussian data was treated where PEM was compared to a pseudo-ML (the ML algorithm constructed under assumption of Gaussian data). For such cases, there is no strict order relation between the covariance matrices of the parameter estimates. Depending on which linear combination of parameters that is considered, either PEM or pML can give better accuracy. [Section VII]

References

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

[2] T Söderström. *Discrete-time Stochastic Systems - Estimation and Control, 2nd edition.* Springer-Verlag, London, UK, 2002.

[3] T. Söderström. Computing the covariance matrix for PEM estimates and the Cramer-Rao lower bound for linear state space models. Technical Report 2005-019, Department of Information Technology, Uppsala University, Uppsala, Sweden, June 2005.

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

[5] T Söderström. *Errors-in-Variables Methods in System Identification.* Springer-Verlag, London, UK, 2018.

[6] T. Söderström. Relations between prediction error and maximum likelihood methods in an error-in-variables setting. Extended version with full proofs. Technical Report 2023-003, Department of Information Technology, Uppsala University, Uppsala, Sweden, 2023. Available as http://www.it.uu.se/research/publications/ reports/2023-003.

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