Proceedings of the 17th World Congress
The International Federation of Automatic Control
Seoul, Korea, July 6-11, 2008

Identification of errors-in-variables models using the EM algorithm ⋆

Jaafar ALMutawa *

* *Department of Mathematics and Statistics, King Fahd University of Petroleum and Minerals, (e-mail:jaafarm@kfupm.edu.sa).*

Abstract: This paper advocates a new subspace system identification algorithm for the errors-in-variables (EIV) state space model via the EM algorithm. To initialize the EM algorithm an initial estimate is obtained by the errors-in-variables subspace system identification method: EIV-MOESP (Chou et al. [1997]) and EIV-N4SID (Gustafsson [2001]). The EM algorithm is an algorithm to compute the maximum value for the likelihood function that is consists of two steps; namely the E- and M-steps. The E- and M-steps in the EM algorithm are calculated by computing the conditional expectation under the assumption that the input-output data is completely observed. Numerical example shows that the EM algorithm can monotonically improve the initial estimates obtained by subspace identification methods.
Keywords: Subspace system identification, errors-in-variables model, EM-algorithm.

1. INTRODUCTION

The identification problem for the linear state space models based on a data corrupted by a noise on both inputs and outputs are called errors-in-variables (EIV) models. Recently the errors-in-variables models have received more attention (Chou et al. [1997], Gustafsson [2001], Li et al. [2001], Diversi et al. [2005]) due to its wide application in engineering and economics. Most approaches (Chou et al. [1997], Gustafsson [2001], Li et al. [2001]) are based on statistical frameworks for example the instrumental variables and principal components analysis. The maximum likelihood estimates has been applied by Diversi et al. [2005] to identify the unknown parameters in ARMA model. In particular, in this paper, we shall present a solution to the problem of identifying the unknown parameters in the EIV state space models by applying the EM algorithm.

The maximum likelihood estimation (MLE) of parameters appearing in the state space model has been approached, for most part, using steepest ascent and Newton-Raphson corrections to iteratively solve the non-linear equations (Hamilton [1994]). The steepest ascent method may require a very large number of iterations to close in on the local maximum. The Newton-Raphson procedure is also computationally expensive since it involves a set of recursions for the second order derivatives of the log-likelihood function and require a matrix inversion of second order partial derivatives at each step.

To circumvent these difficulties, Dempster et al. [1977], have introduced the expectation maximization (EM) algorithm, which is an iterative algorithm for computing the MLE in incomplete data problems. The EM algorithm can be broken down into the expectation (E) step and maximization (M) step, with the basic idea being to maximize the incomplete data log-likelihood by maximizing the current conditional expectation of complete data log-likelihood given the incomplete data (Little et al. [2002], Mclachlan et al. [1997]). There are several advantages of using EM algorithm, for example it does not require the second order derivatives to be calculated or approximated and the EM algorithm always increases the likelihood function, converging to at least a stationary point of the log-likelihood function. Stationary values may, of course, be either a local or global maximum or a point on ridge (Shumway et al. [1982]). A disadvantage is that the rate of convergence of the EM algorithm is somewhat slower than the Newton-Raphson procedure, which has a quadratic convergence in the neighborhood of the maximum.

A system identification method based on EM algorithm has been proposed by Shumway et al. [1982], where the stationary stochastic process without exogenous input has been considered for two cases with missing data and without missing data. Identification problems in the ARX model setting subject to missing data have been studied in Isaksson et al. [1993]. The exogenous input for the state space model has been considered by Gibson et al. [2005], in which the output data are purely observed.

In this paper the conditional expectation is computed by the Kalman filter and smoother. In order to apply the EM procedure, initial values are required for the unknown parameters. Therefore we use EIV subspace system identification methods to obtain initial estimates of the EIV state-space model. We will examine two different sets of starting values, since the EM algorithm may converge to different stationary values corresponding to a local rather than global maxima.

The remainder of this article is arranged as follows. In section 2, the problem is stated along with underlying assumptions. In section 3, we briefly review the EM algorithm and derives a EIV state space identification method based on the EM algorithm, where we initialize the EM algorithm by subspace system identification methods. Section 4 gives a simulation result and section 6 concludes the paper. Appendix A presents the Kalman filter and

⋆ This work was supported by King Fahd University of Petroleum and Minerals.

smoother to be used in the E-step, and Appendices B and C gives the proof of Lemma 1 and Lemma 2; respectively.

2. PROBLEM STATEMENT

As depicted in Fig. 1, consider the linear time invariant (LTI) errors-in-variables state space model described by

$$\begin{bmatrix} x_{t+1} \\ y_t \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} x_t \\ \hat{u}_t \end{bmatrix} + \begin{bmatrix} w_t \\ v_t \end{bmatrix}, \tag{1}$$

where $x_t \in \mathbb{R}^n$, $\hat{u}_t \in \mathbb{R}^m$ and $y_t \in \mathbb{R}^p$ are unknown state, true input and measured output vectors respectively. Furthermore, the noises w_t and v_t are Gaussian white

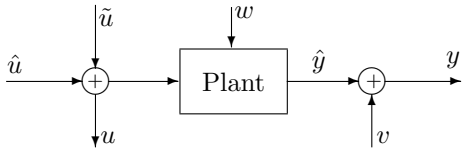

Fig. 1. Errors-in-variables model

noises with zero mean and finite covariance matrices

$$\mathbb{E}\left\{ \begin{bmatrix} w_t \\ v_t \end{bmatrix} \begin{bmatrix} w_t^{\mathrm{T}} & v_t^{\mathrm{T}} \end{bmatrix} \right\} = \begin{bmatrix} Q & S \\ S^{\mathrm{T}} & R \end{bmatrix}. \tag{2}$$

The measured input signals u_t is modeled as

$$u_t = \hat{u}_t + \tilde{u}_t, \tag{3}$$

where $\tilde{u}_t \in \mathbb{R}^m$ is Gaussian white noises with zero mean and finite positive definite covariance matrix $\Sigma_{\tilde{u}}$. Furthermore, we assume in the sequel, that $\tilde{u}_t$ is uncorrelated with $\{\hat{u}_t, w_t, v_t\}$.

It should be noted that the EM-algorithm estimate for the state space model (Shumway et al. [1982], Gibson et al. [2005]) is biased estimate if it is applied to estimate $\begin{bmatrix} A & B \\ C & D \end{bmatrix}$ in the EIV state space model(see Appendix B).

The following assumptions are now introduced.

- **(A1)** System (1) is asymptotically stable, i.e. all eigenvalues of the matrix A are inside the unit circle.
- **(A2)** All system modes are observable and reachable.
- **(A3)** The true input $\hat{u}_t$ is an i.i.d. random variable with $N(\bar{\hat{u}}_t, \Sigma_{\hat{u}})$ where the variance $\Sigma_{\hat{u}} > 0$ and $\hat{u}_t$ is independent of $(x_t, v_t, w_t, \tilde{u}_t)$.
- **(A4)** The noise variances $\Sigma_{\tilde{u}}$ is known.

If assumption **(A3)** has been omitted then our problem setting is not identifiable (see Solary [1969], Bekker et al. [1984]). Furthermore, if the problem settings has been changed to the well known Berkson model, i.e.

$$\hat{u}_t = u_t + \tilde{u}_t,$$

Then it trivial to conclude that the subspace system identification (Verhaegen et al. [1992], Picci et al. [1996], Overschee et al. [1994]), and maximum likelihood estimate (Shumway et al. [1982], Gibson et al. [2005]) still give unbiased estimate.

The problem under investigation can be stated as follows.

Problem: Assume a time sequence of data $\{(u_t, y_t),\ t = 1, \cdots, N\}$ is given. Then, the problem of interest is to estimate the true input data $\{\hat{u}_t : \ t = 1, \cdots, N\}$ and the parameter (A, B, C, D, Q, S, R) within the freedom of equivalent transformation. The fact that we account for the possibility that the input signal is not exactly known, makes the problem difficult, and is often referred to as an errors-in-variables (EIV) problem (Gustafsson [2001]).

3. EM ALGORITHM

In this section, the EM algorithm will be reviewed based on Dempster et al. [1977], Little et al. [2002], Mclachlan et al. [1997], and we will present an algorithm for the estimation of LTI EIV state space models represented by (1) based on the EM algorithm.

3.1 EM algorithm: review

Assume that the observed data Y are generated according to some distribution $f(Y)$. We write $Y = (Y_{obs}, Y_{mis})$ where Y_{obs} represents the observed part of Y and Y_{mis} the missing part. We call Y_{obs} the incomplete data and Y the complete data. The objective is to maximize the incomplete data likelihood

$$\mathcal{L}(Y_{obs}; \Theta) = \int f(Y_{obs}, Y_{mis} \mid \Theta) dY_{mis}, \tag{4}$$

with respect to Θ. To compute the MLE in the presence of missing data, the EM algorithm has been used in two steps, i.e. E-step and M-step.

The E-step finds the conditional expectation of the missing data given the observed data and current estimated parameters, and then substitutes these expectations for the missing data. Specifically, let $\Theta^{(j-1)}$ be the current estimate of the parameter Θ. Then E-step finds the conditional expectation of the complete-data log-likelihood given $\Theta^{(j-1)}$:

$$\mathcal{Q}(\Theta \mid \Theta^{(j-1)}) = \mathrm{E}\{\log \mathcal{L}(Y; \Theta) \mid Y_{obs}, \Theta^{(j-1)}\}. \tag{5}$$

The M-step is particularly simple to describe: compute the MLE of Θ just as if there were no missing data, that is, as if they had been filled in. Hence, the M-step determines $\Theta^{(j)}$ by maximizing the expected complete-data log-likelihood:

$$\mathcal{Q}(\Theta^{(j)} \mid \Theta^{(j-1)}) \geq \mathcal{Q}(\Theta \mid \Theta^{(j-1)}), \quad \forall \Theta.$$

It should be noted that for any EM algorithm, the change from $\Theta^{(j-1)}$ to $\Theta^{(j)}$ does not decrease the log-likelihood.

3.2 EM algorithm for the errors-in-variables model

Define $(Y_N, \ U_N)$ as the incomplete data and $(X_N, Y_N, U_N, \hat{U}_N)$ the complete data where $X_N = \{x_0, x_1, \cdots, x_N\}$ in order to apply the EM algorithm. For simplicity let $\mathcal{F}_N^{(j-1)} = \{Y_N, U_N, \Theta^{(j-1)}\}$, where $(j-1)$ means the previous iteration if j denotes the current iteration. Hence to facilitate the EM algorithm, we need to derive the joint probability density function of Θ based on the complete data.

Let $\alpha_t = \begin{bmatrix} w_t \\ v_t \end{bmatrix}$ and by applying the Bayes theorem, the relationship of the observations to the input, output data and the state is written as:

$$\begin{aligned} p_\Theta\left(\begin{bmatrix} x_{t+1} \\ y_t \end{bmatrix} \middle| X_t, \hat{U}_t \right) &= p_\Theta\left(\begin{bmatrix} x_{t+1} \\ y_t \end{bmatrix} \mid x_t, \hat{u}_t \right) \\ &= f_\alpha\left(\begin{bmatrix} x_{t+1} \\ y_t \end{bmatrix} - \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} x_t \\ \hat{u}_t \end{bmatrix} \right) \end{aligned} \tag{6}$$

where $f_{\alpha_t}(\cdot)$ represent the $(p+n)$-variate normal density function with zero mean and covariance matrix $\begin{bmatrix} Q & S \\ S^{\mathrm{T}} & R \end{bmatrix}$. We see that the observations are conditionally independent given the present input and state, and that the observations are linear and Gaussian.

Assume that

$$x_0 \sim f_0(x_0 - \mu_0)$$

where $f_0(\cdot)$ represents the n-variate normal density of the initial state x_0 with zero mean and covariance matrix Σ_0. Since f_0 and f_{α_t} completely specify the likelihood function, we have

$$\begin{aligned}\mathcal{L}(X,Y,\hat{U};\Theta) &= p_\Theta(X_N, Y_N, \hat{U}_N) \\ &= f_0(x_0-\mu_0)\prod_{t=1}^{N} f_{\alpha_t}\left(\begin{bmatrix} x_{t+1} \\ y_t \end{bmatrix} - \begin{bmatrix} A & B \\ C & D \end{bmatrix}\begin{bmatrix} x_t \\ \hat{u}_t \end{bmatrix}\right) \\ &\times f_{\tilde{u}_t}(u_t - \hat{u}_t) \end{aligned} \tag{7}$$

Under the Gaussian assumption, the complete data log-likelihood of (7) can be written as:

$$\begin{aligned}\log\mathcal{L}(X,Y,\hat{U};\Theta) &= -\frac{1}{2}\log \mid \Sigma_0 \mid -\frac{N}{2}\log \mid \Sigma_{\tilde{u}} \mid \\ &- \frac{N}{2}\log\left|\begin{matrix} Q & S \\ S^{\mathrm{T}} & R \end{matrix}\right| - \frac{1}{2}(x_0-\mu)^{\mathrm{T}}\Sigma_0^{-1}(x_0-\mu) \\ &- \frac{1}{2}\sum_{t=1}^{N}(u_t-\hat{u}_t)^{\mathrm{T}}\Sigma_{\tilde{u}}^{-1}(u_t-\hat{u}_t) \\ &- \frac{1}{2}\sum_{t=1}^{N}\left(\begin{bmatrix} x_{t+1} \\ y_t \end{bmatrix} - \begin{bmatrix} A & B \\ C & D \end{bmatrix}\begin{bmatrix} x_t \\ \hat{u}_t \end{bmatrix}\right)^{\mathrm{T}}\begin{bmatrix} Q & S \\ S^{\mathrm{T}} & R \end{bmatrix}^{-1} \\ &\times \left(\begin{bmatrix} x_{t+1} \\ y_t \end{bmatrix} - \begin{bmatrix} A & B \\ C & D \end{bmatrix}\begin{bmatrix} x_t \\ \hat{u}_t \end{bmatrix}\right) \end{aligned} \tag{8}$$

Thus in view of (8) if the true inputs data $\hat{u}_t$ and states x_t were observed, then the MLE can be computed easily. However, since the true input $\hat{u}_t$ and states x_t are not observed, we compute the MLE by iteratively calculating the E- and M-step of the EM algorithm.

To compute the E-step in the EM algorithm, the conditional expectation of the incomplete data with respect to the observed data are calculated in the following lemma.

Lemma 1. For the EIV state-space model structure (1) under the Gaussian assumption, the function $\mathcal{Q}(\Theta \mid \Theta^{(j-1)})$ defined in (5) can be computed as:

$$\begin{aligned}\mathcal{Q}(\Theta \mid \Theta^{(j-1)}) &= -\frac{1}{2}\log \mid \Sigma_0 \mid -\frac{N}{2}\log \mid \Sigma_{\tilde{u}} \mid -\frac{N}{2}\log\left|\begin{matrix} Q & S \\ S^{\mathrm{T}} & R \end{matrix}\right| \\ &- \frac{1}{2}\mathrm{Tr}\{\Sigma_0^{-1}[P_{0|N} + (x_{0|N}-\mu_0)(x_{0|N}-\mu_0)^{\mathrm{T}}]\} \\ &- \frac{1}{2}\mathrm{Tr}\{\Sigma_{\tilde{u}}^{-1}[\sum_{t=1}^{N}(u_t - \mathrm{E}\{\hat{u}_t \mid \mathcal{F}_N^{(j-1)}\})(u_t - \mathrm{E}\{\hat{u}_t \mid \mathcal{F}_N^{(j-1)}\})^{\mathrm{T}}]\} \\ &- \frac{1}{2}\mathrm{Tr}\left\{\begin{bmatrix} Q & S \\ S^{\mathrm{T}} & R \end{bmatrix}^{-1}\left[\Phi - \Psi\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{\mathrm{T}} - \begin{bmatrix} A & B \\ C & D \end{bmatrix}\Psi^{\mathrm{T}}\right.\right. \\ &\left.\left. + \begin{bmatrix} A & B \\ C & D \end{bmatrix}\Gamma\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{\mathrm{T}}\right]\right\}, \end{aligned} \tag{9}$$

where

$$\Phi = \sum_{t=1}^{N}\begin{bmatrix} \mathrm{E}\{x_{t+1}x_{t+1}^{\mathrm{T}} \mid \mathcal{F}_N^{(j-1)}\} & x_{t+1|N}y_t^{\mathrm{T}} \\ y_t x_{t+1|N}^{\mathrm{T}} & y_t y_t^{\mathrm{T}} \end{bmatrix}, \tag{10}$$

$$\Psi = \sum_{t=1}^{N}\begin{bmatrix} \mathrm{E}\{x_{t+1}x_t^{\mathrm{T}} \mid \mathcal{F}_N^{(j-1)}\} & \mathrm{E}\{x_{t+1}\hat{u}_t^{\mathrm{T}} \mid \mathcal{F}_N^{(j-1)}\} \\ y_t x_{t|N}^{\mathrm{T}} & y_t u_{t|N}^{\mathrm{T}} \end{bmatrix}, \tag{11}$$

$$\Gamma = \sum_{t=1}^{N}\begin{bmatrix} \mathrm{E}\{x_t x_t^{\mathrm{T}} \mid \mathcal{F}_N^{(j-1)}\} & \mathrm{E}\{x_t\hat{u}_t^{\mathrm{T}} \mid \mathcal{F}_N^{(j-1)}\} \\ \mathrm{E}\{\hat{u}_t x_t^{\mathrm{T}} \mid \mathcal{F}_N^{(j-1)}\} & u_t u_t^{\mathrm{T}} + \Sigma_{\tilde{u}} \end{bmatrix}. \tag{12}$$

The conditional expectations in the above lemma can be computed by using the Kalman filter and smoother.

Lemma 2.

$$\begin{aligned}\mathrm{E}\{\hat{u}_t x_t^{\mathrm{T}} \mid \mathcal{F}_N^{(j-1)}\} &= u_{t|N}x_{t|N}^{\mathrm{T}} \\ \mathrm{E}\{x_{t+1}\hat{u}_t^{\mathrm{T}} \mid \mathcal{F}_N^{(j-1)}\} &= B(\Sigma_u - \Sigma_{\tilde{u}})B^{\mathrm{T}} + x_{t+1|N}u_{t|N}\end{aligned}$$

The M-step can be easily achieved via the following lemma.

Lemma 3. The function $\mathcal{Q}(\Theta \mid \Theta^{(j-1)})$ of Lemma 1 is maximized by

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} = \Psi\Gamma^{-1} \tag{13}$$

$$\begin{bmatrix} Q & S \\ S^{\mathrm{T}} & R \end{bmatrix} = N^{-1}(\Phi - \Psi\Gamma^{-1}\Psi^{\mathrm{T}}) \tag{14}$$

The iterative procedure is easy to apply since the initial estimate $\Theta^{(0)}$ can be used to produce the initial $x_{t|N}$, $P_{t|N}$ and $M_{t|N}$ by the recursions in Appendix A. Then, simple calculations in (10)-(12) give the updated $\Theta^{(1)}$, and the recursions are used to generate new $x_{t|N}$, $P_{t|N}$ and $M_{t|N}$. The iterations are stopped when the log-likelihood and parameter estimates converge.

The EM algorithm for EIV state space models is summarized as follows

EM algorithm estimate for EIV state space model

Step 1: Initialize $\Theta^{(0)} = \{A, B, C, D, Q, R\}$ using EIV subspace identification methods MOESP or N4SID.
Step 2: By using $\Theta^{(0)}$, run the Kalman filter and smoother shown in Appendix A.
Step 3: Calculate the E-step using (10)-(12).
Step 4: Calculate the M-step using (15)-(16).
Step 6: Repeat steps 2-4 until we get a satisfactory convergence.

4. NUMERICAL EXAMPLE

The following example is a slightly modified version of the one used in Diversi et al. [2005]. Where the numerical simulation is performed on two inputs two outputs time-invariant system with $N = 500$ described by the following matrices:

$$A = \begin{bmatrix} 0 & 1 & 0 \\ -0.3 & 0.4 & -0.2 \\ -0.1 & 0.2 & 0.4 \end{bmatrix}, \quad B = \begin{bmatrix} 0.8 & -0.8 \\ 0.17 & -0.37 \\ 1.09 & 1.1 \end{bmatrix},$$
$$C = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad D = \begin{bmatrix} 1.7 & 1.5 \\ 0.51 & -1 \end{bmatrix}.$$

The noise free input sequence $\hat{u}_t$ is a zero mean with a unit variance Gaussian process, and a sample of the unmeasurable output data $\hat{y}_t$ is shown in Fig. 2. Furthermore, the

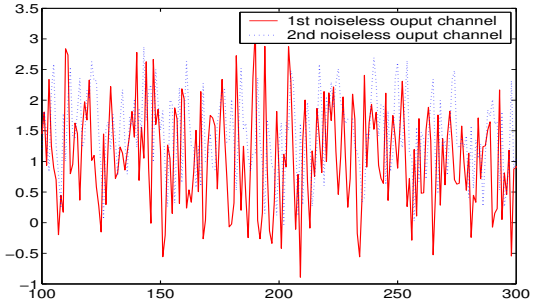

Fig. 2. Un-observed output data $\hat{y}$

noise sequences $w_t, \tilde{y}_t$ and $\tilde{u}_t$ are characterized as follows

$$w_t \sim \mathrm{N}(0_3, 0.1 \times I_3),$$
$$\tilde{u}_t \sim \mathrm{N}(0_2, 0.1 \times I_2),$$
$$\tilde{y}_t \sim \mathrm{N}(0_2, 0.1 \times I_2),$$

A sample for realizations of the noises $\tilde{y}_t$ and $\tilde{u}_t$ are shown in Fig. 3. Moreover, the initial state x_0 is a random vector

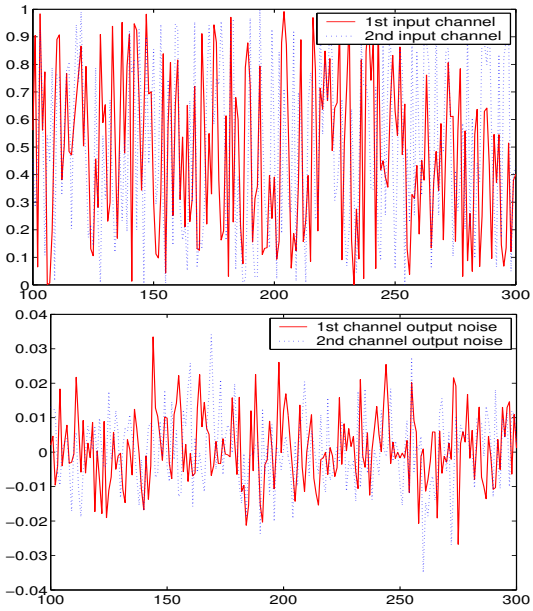

Fig. 3. A sample process noise $\tilde{u}$ and $\tilde{y}$

and has been initialized as $x_{0|-1} = 0$ and $P_{0|-1} = I_n$.

Fig. 4 shows that the prediction error defined by

$$E_N = \frac{1}{N}\sum_{t=1}^{N}(y_t - Cx_{t|t-1})^2$$

decreases by iteration and hence we infer that the estimate converges toward the true system. Fig. 4 also shows that about 40 iterations lead to considerable change in the performance, whereas further iterations do not change it significantly.

5. CONCLUSION

In this paper, we have considered the EM algorithm applied to the errors-in-variables state space models initialized by the classical errors-in-variables subspace system identification algorithms. Then, EM algorithm is derived by taking the conditional expectation of the log-likelihood function under the assumption that the states can be

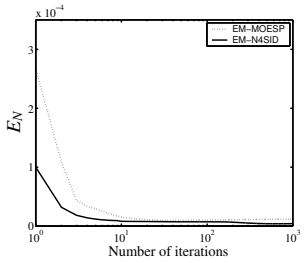

Fig. 4. Error functions for EM initialized by MOESP and N4SID

considered as incompletely observed data. Numerical examples have shown the effectiveness of state space identification method based on the proposed EM algorithm.

REFERENCES

C. Chou and M. Verhaegen: Subspace algorithms for the identification of multivariable dynamic errors-in-variables models; *Automatica*, Vol. 33, No. 10, pp. 1857-1869 (1997)

T. Gustafsson: Subspace identification using instrumental variable techniques; *Automatica*, Vol. 37, No. 1, pp. 2005-2010 (2001)

W. Li and S.J. Qin: Consistent dynamic PCA based on errors-in-variables subspace identification; *J. Process Control*, Vol. 11, pp. 661-678 (2001)

R. Diversi, R. Guidorzi and U. Soverini: Maximum likelihood identification of noisy inputoutput models; *Automatica*, Vol. 43, pp. 464-472 (2007)

J. ALMutawa and H. Tanaka and T. Katayama: EM algorithm for system identification in the presence of outliers; *Trans. Inst. Sys. Contr. and Inform. Eng.*, Vol. 18, No. 5, pp. 146-151 (2005)

M. Verhaegen and P. Dewilde: Subspace model identification, Part 1. The output error state-space model identification class of algorithms; *Int. J. Control*, Vol. 56, No. 5, pp. 1187-1210 (1992)

G. Picci and T. Katayama: Stochastic realization with exogenous inputs and subspace methods identification; *Signal Processing*, Vol. 52, No. 2, pp. 145-160 (1996)

P. Van Overschee and B. De Moore: Subspace algorithms for the identification of combined deterministic-stochastic systems; *Automatica*, Vol. 30, No. 1, pp. 75-93 (1994)

J.D. Hamilton: *Time Series Analysis*, Princeton University Press (1994)

A. Dempster, N.M. Laird and D.B. Rubin: Maximum likelihood from incomplete data via the EM algorithm; *J. Royal Statistical Society*, Series B, Vol. 39, pp. 1-38 (1977)

R. Little and D. Rubin: *Statistical Analysis with Missing Data* (2nd ed.), John Wiley & Sons (2002)

G.J. Mclachlan and T. Krishnan: *EM Algorithm and Extensions*, John Wiley & Sons (1997)

R.H. Shumway and D.S. Stoffer: An approach to time series smoothing and forecasting using the EM algorithm; *J. Time Series Analy.*, Vol. 3, pp. 253-264 (1982)

M. Tanaka and T. Katayama: Robust identification and smoothing for linear system with outliers and missing data; *Proc. 11th IFAC Congress*, Vol. 3, pp. 160-165 (1990)

A. Isaksson: Identification of ARX-models subject to missing data; *IEEE Trans. Automatic Control*, Vol. 38, No. 5, pp. 813-819 (1993)
S. Gibson and B. Ninness: Robust maximum-likelihood estimation of multivariable dynamic; *Automatica*, Vol. 41, pp. 1667-1682 (2005)
R. Diversi, R. Guidorzi and U. Soverini: Kalman filtering in extended noise environments; *IEEE Trans. Automatic Control*, Vol. 50, No. 9, pp. 1396-1402 (2005)
M.E. Solary: The maximum likelihood solution of the problem of estimating a linear functional relationship; *J. Royal Statistics Soc.*, Vol. 31, pp. 372–375 (1969)
J. AlMutawa: Robust Kalman filter and smoother for errors-in-variables model with observation outliers; *International Journal of Control*, Vol. xxx, pp. xx–xx (2008)
P.A. Bekker, A. Kapteyn and T.J. Wansbeek: Measurement error and endogeneity in regression: Bounds for ML and 2SLS estimates; *Dijkstra, T.K., Editor, Misspecification Analysis, Springer*, pp. 85-103, (1984)

Appendix A. KALMAN FILTER AND SMOOTHER FOR EM ITERATION

The Kalman filter is given by (Diversi et al. [2005])

$$y_{t+1|t} = Cx_{t|t} + Du_t, \tag{A.1}$$
$$x_{t+1|t} = Ax_{t|t-1} + Bu_t + K_t\epsilon_t, \tag{A.2}$$
$$K_t = [AP_{t|t-1}C^{\mathrm{T}} + S_t]\Sigma_{\epsilon_t}^{-1}, \tag{A.3}$$
$$P_{t+1|t} = AP_{t|t-1}A^{\mathrm{T}} + Q_t - [AP_{t|t-1}C^{\mathrm{T}} + S_t]\Sigma_{\epsilon_t}^{-1} \times [AP_{t|t-1}C^{\mathrm{T}} + S_t]^{\mathrm{T}}. \tag{A.4}$$

and the Kalman smoother for $t = N, N-1, \cdots, 1$,

$$x_{t-1|N} = x_{t-1|t-1} + S_{t-1}[x_{t|N} - x_{t|t-1}], \tag{A.5}$$
$$P_{t-1|N} = P_{t-1|t-1} + S_{t-1}[P_{t|N} - P_{t|t-1}]S_{t-1}^{\mathrm{T}}, \tag{A.6}$$
$$S_{t-1} = P_{t-1|t-1}AP_{t|t-1}^{-1}, \tag{A.7}$$
$$M_{t|N} = \mathrm{E}\{(x_{t|N} - x_t)(x_{t-1|N} - x_{t-1})^{\mathrm{T}} \mid \mathcal{F}_N^{(j-1)}\}, \tag{A.8}$$
$$M_{N|N} = (I - K_NC)AP_{N-1|N-1}, \tag{A.9}$$
$$M_{t|N} = P_{t|t}S_{t-1}^{\mathrm{T}} + S_t(M_{t+1|N} - AP_{t|t})S_{t-1}^{\mathrm{T}}, \tag{A.10}$$

$$\tilde{u}(t \mid t) = [\Sigma_{\widetilde{uy}}(t) - \Sigma_{\tilde{u}}D^{\mathrm{T}}]\Sigma_\epsilon(t)^{-1}\epsilon(t), \tag{A.11}$$
$$\tilde{y}(t \mid t) = [\Sigma_{\tilde{y}} - \Sigma_{\widetilde{uy}}^{\mathrm{T}}D^{\mathrm{T}}]\Sigma_\epsilon(t)^{-1}\epsilon(t). \tag{A.12}$$

By using (A.11) and (A.12), the minimal variance estimates of $\hat{y}(t)$ and $\hat{u}(t)$ can be written in the form

$$\hat{u}(t \mid t) = u(t) - [\Sigma_{\widetilde{uy}} - \Sigma_{\tilde{u}}D^{\mathrm{T}}]\Sigma_\epsilon(t)^{-1}\epsilon(t), \tag{A.13}$$
$$\hat{y}(t \mid t) = y(t) - [\Sigma_{\tilde{y}} - \Sigma_{\widetilde{uy}}^{\mathrm{T}}D^{\mathrm{T}}]\Sigma_\epsilon(t)^{-1}\epsilon(t), \tag{A.14}$$

which are initialized by $x_{0|0} = \mu$, and $P_{0|0} = P_0$. Then,

$$\mathrm{E}\{x_tx_t^{\mathrm{T}} \mid \mathcal{F}_N^{(j-1)}\} = P_{t|N} + x_{t|N}x_{t|N}^{\mathrm{T}}, \tag{A.15}$$
$$\mathrm{E}\{x_tx_{t-1}^{\mathrm{T}} \mid \mathcal{F}_N^{(j-1)}\} = M_{t|N} + x_{t|N}x_{t-1|N}^{\mathrm{T}}, \tag{A.16}$$

we replace $z(t)$ by its innovation (AlMutawa [2008])

$$\tilde{u}(t \mid N) = \tilde{u}(t \mid t) + \sum_{s=t+1}^{N} \mathrm{cov}\{\tilde{u}(t), \epsilon(s)\}\Sigma_\epsilon(s)^{-1}\epsilon(s), \tag{A.17}$$

$$\tilde{y}(t \mid N) = \tilde{y}(t \mid t) + \sum_{s=t+1}^{N} \mathrm{cov}\{\tilde{y}(t), \epsilon(s)\}\Sigma_\epsilon(s)^{-1}\epsilon(s), \tag{A.18}$$

The covariances can be found as follows

$$\mathrm{cov}\{\tilde{u}(t), \epsilon(s)\} = [\Sigma_{\tilde{u}}(K(t)D - B)^{\mathrm{T}} - \Sigma_{\widetilde{uy}}K^{\mathrm{T}}(t)]L(s-1,t)^{\mathrm{T}}C^{\mathrm{T}}, \tag{A.19}$$

$$\mathrm{cov}\{\tilde{y}(t), \epsilon(s)\} = [\Sigma_{\widetilde{uy}}^{\mathrm{T}}(K(t)D - B)^{\mathrm{T}} - \Sigma_{\tilde{y}}K^{\mathrm{T}}(t)]L(s-1,t)^{\mathrm{T}}C^{\mathrm{T}}, \tag{A.20}$$

and where where $L(s-1,t) = [A - K(s-1)C][A - K(s-2)C]\cdots[A - K(t+1)C]$ and $L(t,t) = I_n$.

Proof 1. The equality follows from

$$\begin{aligned}\mathrm{E}\{\hat{u}_tx_t^{\mathrm{T}} \mid \mathcal{F}_N^{(j-1)}\} &= \mathrm{cov}\{\hat{u}_t, x_t\} + \hat{u}_{t|N}x_{t|N},\\ &= \mathrm{cov}\{\hat{u}_t, Ax_{t-1} + B\hat{u}_{t-1} + w_{t-1}\} + \hat{u}_{t|N}x_{t|N},\\ &= \hat{u}_{t|N}x_{t|N},\end{aligned}$$

and last equality follows from

$$\begin{aligned}\mathrm{E}\{x_{t+1}\hat{u}_t^{\mathrm{T}} \mid \mathcal{F}_N^{(j-1)}\} &= \mathrm{cov}\{x_{t+1}, \hat{u}_t\} + x_{t+1|N}\hat{u}_{t|N},\\ &= \mathrm{cov}\{Ax_t + B\hat{u}_t + w_t, \hat{u}_t\} + x_{t+1|N}\hat{u}_{t|N},\\ &= B\Sigma_{\hat{u}}B^{\mathrm{T}} + x_{t+1|N}\hat{u}_{t|N},\\ &= B(\Sigma_u - \Sigma_{\tilde{u}})B^{\mathrm{T}} + x_{t+1|N}\hat{u}_{t|N}.\end{aligned}$$

Appendix B. BIAS

Even if the state x_t is known, the classical EM-algorithm Gibson et al. [2005], Shumway et al. [1982] estimate of $\begin{bmatrix} A & B \\ C & D \end{bmatrix}$ is biased estimate. To be precise the EM algorithm estimate for Θ if x_t is measured given by

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}_{\tilde{U}_N} = \sum_{t=1}^{N} \begin{bmatrix} x_{t+1}x_t^{\mathrm{T}} & x_{t+1}u_t^{\mathrm{T}} \\ y_tx_t^{\mathrm{T}} & y_tu_t^{\mathrm{T}} \end{bmatrix} \sum_{t=1}^{N} \begin{bmatrix} x_tx_t^{\mathrm{T}} & x_tu_t^{\mathrm{T}} \\ u_tx_t^{\mathrm{T}} & u_tu_t^{\mathrm{T}} \end{bmatrix}^{-1}$$

To be precise consider

$$\begin{aligned}&\mathrm{E}\left\{\begin{bmatrix} A & B \\ C & D \end{bmatrix}_{\tilde{U}_N}\right\} = \mathrm{E}_{\tilde{U}_N}\mathrm{E}\left\{\begin{bmatrix} A & B \\ C & D \end{bmatrix}_{\tilde{U}_N} \mid \tilde{U}_N\right\}\\ &= \mathrm{E}_{\tilde{U}_N}\left\{\sum_{t=1}^{N} \begin{bmatrix} x_{t+1}x_t^{\mathrm{T}} & x_{t+1}u_t^{\mathrm{T}} \\ \hat{y}_tx_t^{\mathrm{T}} & \hat{y}_tu_t^{\mathrm{T}} \end{bmatrix} \sum_{t=1}^{N} \begin{bmatrix} x_tx_t^{\mathrm{T}} & x_tu_t^{\mathrm{T}} \\ u_tx_t^{\mathrm{T}} & u_tu_t^{\mathrm{T}} \end{bmatrix}^{-1}\right\}\\ &= \mathrm{E}_{\tilde{U}}\left\{\sum_{t=1}^{N} \begin{bmatrix} x_{t+1}x_t^{\mathrm{T}} & x_{t+1}u_t^{\mathrm{T}} \\ (y_t - \tilde{y}_t)x_t^{\mathrm{T}} & (y_t - \tilde{y}_t)u_t^{\mathrm{T}} \end{bmatrix} \times \sum_{t=1}^{N} \begin{bmatrix} x_tx_t^{\mathrm{T}} & x_tu_t^{\mathrm{T}} \\ u_tx_t^{\mathrm{T}} & u_tu_t^{\mathrm{T}} \end{bmatrix}^{-1}\right\}\\ &= \mathrm{E}_{\tilde{U}}\left\{\sum_{t=1}^{N} \left(\begin{bmatrix} x_{t+1}x_t^{\mathrm{T}} & x_{t+1}u_t^{\mathrm{T}} \\ y_tx_t^{\mathrm{T}} & y_tu_t^{\mathrm{T}} \end{bmatrix} - \begin{bmatrix} 0 & 0 \\ \tilde{y}_tx_t^{\mathrm{T}} & \tilde{y}_tu_t^{\mathrm{T}} \end{bmatrix}\right) \sum_{t=1}^{N} \begin{bmatrix} x_tx_t^{\mathrm{T}} & x_tu_t^{\mathrm{T}} \\ u_tx_t^{\mathrm{T}} & u_tu_t^{\mathrm{T}} \end{bmatrix}^{-1}\right\}\\ &= \begin{bmatrix} A & B \\ C & D \end{bmatrix} - \text{Bias}\end{aligned}$$

where

$$\text{Bias} = \mathrm{E}_{\tilde{U}} \left\{ \sum_{t=1}^{N} \begin{bmatrix} 0 & 0 \\ \tilde{y}_t x_t^{\mathrm{T}} & \tilde{y}_t u_t^{\mathrm{T}} \end{bmatrix} \sum_{t=1}^{N} \begin{bmatrix} x_t x_t^{\mathrm{T}} & x_t u_t^{\mathrm{T}} \\ u_t x_t^{\mathrm{T}} & u_t u_t^{\mathrm{T}} \end{bmatrix}^{-1} \right\}$$

Appendix C. PROOF OF LEMMA 1

Thus

$$\begin{aligned}
&\mathrm{E}\{\log\mathcal{L}(X,\hat{Y},\hat{U};\Theta) \mid \hat{Y},\hat{U},\Theta^{(j-1)}\} = \mathrm{E}\left\{ -\frac{1}{2}\log \mid \Sigma_0 \mid \right. \\
&-\frac{N}{2}\log \mid \Sigma_{\tilde{u}} \mid -\frac{N}{2}\log \left| \begin{matrix} Q & S \\ S^{\mathrm{T}} & R \end{matrix} \right| - \frac{1}{2}(x_0-\mu)^{\mathrm{T}}\Sigma_0^{-1}(x_0-\mu) \\
&-\frac{1}{2}\sum_{t=1}^{N}(u_t-\hat{u}_t)^{\mathrm{T}}\Sigma_{\tilde{u}}^{-1}(u_t-\hat{u}_t) \\
&-\frac{1}{2}\sum_{t=1}^{N}\left(\begin{bmatrix} x_{t+1} \\ y_t \end{bmatrix} - \begin{bmatrix} A & B \\ C & D \end{bmatrix}\begin{bmatrix} x_t \\ \hat{u}_t \end{bmatrix}\right)^{\mathrm{T}} \begin{bmatrix} Q & S \\ S^{\mathrm{T}} & R \end{bmatrix}^{-1} \\
&\qquad \left.\times\left(\begin{bmatrix} x_{t+1} \\ y_t \end{bmatrix} - \begin{bmatrix} A & B \\ C & D \end{bmatrix}\begin{bmatrix} x_t \\ \hat{u}_t \end{bmatrix}\right)\right\} \qquad \text{(C.1)}
\end{aligned}$$

last equation implies

$$\begin{aligned}
&\mathrm{E}\{\log\mathcal{L}(X,\hat{Y},\hat{U};\Theta) \mid \hat{Y},\hat{U},\Theta^{(j-1)}\} = -\frac{1}{2}\log \mid \Sigma_0 \mid \\
&-\frac{N}{2}\log \left| \begin{matrix} Q & S \\ S^{\mathrm{T}} & R \end{matrix} \right| - \frac{N}{2}\log \mid \Sigma_{\tilde{u}} \mid \\
&-\frac{1}{2}\mathrm{Tr}\{\Sigma_0^{-1}[P_{0|N} + (x_{0|N}-\mu_0)(x_{0|N}-\mu_0)^{\mathrm{T}}]\} \\
&-\frac{1}{2}\mathrm{Tr}\{\Sigma_{\tilde{u}}^{-1}[(u_t-\hat{u}_{t|N})(u_t-\hat{u}_{t|N})^{\mathrm{T}}]\} \\
&-\frac{1}{2}\mathrm{Tr}\left\{\begin{bmatrix} Q & S \\ S^{\mathrm{T}} & R \end{bmatrix}^{-1} \mathrm{E}\left\{\sum_{t=1}^{N}\left(\begin{bmatrix} x_{t+1} \\ y_t \end{bmatrix} - \begin{bmatrix} A & B \\ C & D \end{bmatrix}\begin{bmatrix} x_t \\ \hat{u}_t \end{bmatrix}\right)\right. \\
&\qquad \left.\left.\times\left(\begin{bmatrix} x_{t+1} \\ y_t \end{bmatrix} - \begin{bmatrix} A & B \\ C & D \end{bmatrix}\begin{bmatrix} x_t \\ \hat{u}_t \end{bmatrix}\right)^{\mathrm{T}}\right\}\right\} \qquad \text{(C.2)}
\end{aligned}$$

next step we will expand the last term of the right hand side of (C.2), i.e.

$$\begin{aligned}
&= \sum_{t=1}^{N}\mathrm{E}\left\{\begin{bmatrix} x_{t+1} \\ y_t \end{bmatrix}\begin{bmatrix} x_{t+1} \\ y_t \end{bmatrix}^{\mathrm{T}} - \begin{bmatrix} x_{t+1} \\ y_t \end{bmatrix}\begin{bmatrix} x_t \\ \hat{u}_t \end{bmatrix}^{\mathrm{T}}\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{\mathrm{T}}\right. \\
&-\begin{bmatrix} A & B \\ C & D \end{bmatrix}\begin{bmatrix} x_t \\ \hat{u}_t \end{bmatrix}\begin{bmatrix} x_{t+1} \\ y_t \end{bmatrix}^{\mathrm{T}} \\
&\left.+\begin{bmatrix} A & B \\ C & D \end{bmatrix}\begin{bmatrix} x_t \\ \hat{u}_t \end{bmatrix}\begin{bmatrix} x_t \\ \hat{u}_t \end{bmatrix}^{\mathrm{T}}\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{\mathrm{T}}\right\} \\
&= \sum_{t=1}^{N}\left\{\begin{bmatrix} \mathrm{E}\{x_{t+1}x_{t+1}^{\mathrm{T}}\} & x_{t+1|N}y_t^{\mathrm{T}} \\ y_t x_{t+1|N}^{\mathrm{T}} & y_t y_t^{\mathrm{T}} \end{bmatrix}\right. \\
&-\begin{bmatrix} \mathrm{E}\{x_{t+1}x_t^{\mathrm{T}}\} & \mathrm{E}\{x_{t+1}\hat{u}_t^{\mathrm{T}}\} \\ y_t x_{t|N}^{\mathrm{T}} & y_t u_{t|N}^{\mathrm{T}} \end{bmatrix}\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{\mathrm{T}} \\
&-\begin{bmatrix} A & B \\ C & D \end{bmatrix}\begin{bmatrix} \mathrm{E}\{x_t x_{t+1}^{\mathrm{T}}\} & x_{t|N}y_t^{\mathrm{T}} \\ \mathrm{E}\{\hat{u}_t x_{t+1}^{\mathrm{T}}\} & u_{t|N}y_t^{\mathrm{T}} \end{bmatrix} \\
&\left.+\begin{bmatrix} A & B \\ C & D \end{bmatrix}\begin{bmatrix} \mathrm{E}\{x_t x_t^{\mathrm{T}}\} & \mathrm{E}\{x_t\hat{u}_t^{\mathrm{T}}\} \\ \mathrm{E}\{\hat{u}_t x_t^{\mathrm{T}}\} & u_t u_t^{\mathrm{T}} + \mathrm{E}\{\tilde{u}_t\tilde{u}_t^{\mathrm{T}}\} \end{bmatrix}\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{\mathrm{T}}\right\}
\end{aligned}$$

$$\begin{aligned}
&= \sum_{t=1}^{N}\left\{\begin{bmatrix} \mathrm{E}\{x_{t+1}x_{t+1}^{\mathrm{T}}\} & x_{t+1|N}y_t^{\mathrm{T}} \\ y_t x_{t+1|N}^{\mathrm{T}} & y_t y_t^{\mathrm{T}} \end{bmatrix}\right. \\
&-\begin{bmatrix} \mathrm{E}\{x_{t+1}x_t^{\mathrm{T}}\} & \mathrm{E}\{x_{t+1}\hat{u}_t^{\mathrm{T}}\} \\ y_t x_{t|N}^{\mathrm{T}} & y_t u_{t|N}^{\mathrm{T}} \end{bmatrix}\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{\mathrm{T}} \\
&-\begin{bmatrix} A & B \\ C & D \end{bmatrix}\begin{bmatrix} \mathrm{E}\{x_t x_{t+1}^{\mathrm{T}}\} & x_{t|N}y_t^{\mathrm{T}} \\ \mathrm{E}\{\hat{u}_t x_{t+1}^{\mathrm{T}}\} & u_{t|N}y_t^{\mathrm{T}} \end{bmatrix} \\
&\left.+\begin{bmatrix} A & B \\ C & D \end{bmatrix}\begin{bmatrix} \mathrm{E}\{x_t x_t^{\mathrm{T}}\} & \mathrm{E}\{x_t\hat{u}_t^{\mathrm{T}}\} \\ \mathrm{E}\{\hat{u}_t x_t^{\mathrm{T}}\} & u_t u_t^{\mathrm{T}} + \Sigma_{\tilde{u}} \end{bmatrix}\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{\mathrm{T}}\right\}
\end{aligned}$$

which proves lemma 1.

Appendix D. PROOF OF LEMMA 2

Let $\Theta = \begin{bmatrix} A & B \\ C & D \end{bmatrix}$, then take the partial derivative of (9) with respect to Θ gives

$$\begin{aligned}
&\frac{\partial}{\partial\Theta}\mathcal{Q}(\Theta \mid \Theta^{(j-1)}) = -\frac{1}{2}\frac{\partial}{\partial\Theta}\mathrm{Tr}\left\{\begin{bmatrix} Q & S \\ S^{\mathrm{T}} & R \end{bmatrix}^{-1}\left[\Phi - \Psi\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{\mathrm{T}}\right.\right. \\
&\left.\left.-\begin{bmatrix} A & B \\ C & D \end{bmatrix}\Psi^{\mathrm{T}} + \begin{bmatrix} A & B \\ C & D \end{bmatrix}\Gamma\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{\mathrm{T}}\right]\right\} = 0. \qquad \text{(D.1)}
\end{aligned}$$

this implies that

$$-\Psi - \Psi + \begin{bmatrix} A & B \\ C & D \end{bmatrix}\Gamma + \begin{bmatrix} A & B \\ C & D \end{bmatrix}\Gamma = 0$$

so that

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} = \Psi\Gamma^{-1}$$

and let $X = \begin{bmatrix} Q & S \\ S^{\mathrm{T}} & R \end{bmatrix}$, gives

$$\begin{aligned}
&\frac{\partial}{\partial X}\mathcal{Q}(\Theta \mid \Theta^{(j-1)}) = \frac{\partial}{\partial X}\left\{\log\left|\begin{matrix} Q & S \\ S^{\mathrm{T}} & R \end{matrix}\right|\right. \\
&-\frac{1}{2}\mathrm{Tr}\left\{\begin{bmatrix} Q & S \\ S^{\mathrm{T}} & R \end{bmatrix}^{-1}\left[\Phi - \Psi\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{\mathrm{T}} - \begin{bmatrix} A & B \\ C & D \end{bmatrix}\Psi^{\mathrm{T}}\right.\right. \\
&\left.\left.\left.+\begin{bmatrix} A & B \\ C & D \end{bmatrix}\Gamma\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{\mathrm{T}}\right]\right\}\right\}
\end{aligned}$$

the last equation gives

$$\begin{aligned}
N\begin{bmatrix} Q & S \\ S^{\mathrm{T}} & R \end{bmatrix}^{-1} &= \begin{bmatrix} Q & S \\ S^{\mathrm{T}} & R \end{bmatrix}^{-1}\left\{\Phi - \Psi\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{\mathrm{T}}\right. \\
&\left.-\begin{bmatrix} A & B \\ C & D \end{bmatrix}\Psi^{\mathrm{T}} + \begin{bmatrix} A & B \\ C & D \end{bmatrix}\Gamma\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{\mathrm{T}}\right\}\begin{bmatrix} Q & S \\ S^{\mathrm{T}} & R \end{bmatrix}^{-1}
\end{aligned}$$

using the values of $\begin{bmatrix} A & B \\ C & D \end{bmatrix}$ proves the lemma.