Estimation of Noise Covariances and Disturbance Structure from Data Using Least Squares with Optimal Weighting

Murali R Rajamani,* and James B. Rawlings*

Department of Chemical and Biological Engineering,
University of Wisconsin-Madison, 1415 Engineering Drive, Madison, WI 53706

Abstract

For the linear time-invariant state space model, it is well known that the minimum variance state estimator is the Kalman filter when there are no constraints on the states. In [20], the correlations between the innovations was used to form a least-squares problem to determine the optimal tuning for the Kalman filter. In this paper, we formulate the optimal weighting to be used in the above least-squares problem. In addition to that the stochastic disturbance structure that affects the states and the measurements is also usually unknown. We present a semidefinite programming to estimate the disturbance structure and the covariances of the noises entering the system. Once the noise covariances are estimated using the correctly weighted least-squares technique, the state estimator can be tuned optimally. The disturbance structure provides information about the minimum number of disturbances affecting the state.

Keywords: ALP, optimal weighting, semidefinite programming, covariance estimation

* Authors for correspondence. (M.R.R.) Telephone: (608) 265-8607. E-mail: rmurali@wisc.edu. (J.B.R.) Telephone: (608) 263-5859. Fax: (608) 265-8794. E-mail: jbraw@bevo.che.wisc.edu.

I. INTRODUCTION

Consider the Linear Time-Invariant (LTI) state-space model in discrete time:

$$\begin{aligned} x_{k+1} &= Ax_k + Bu_k + Gw_k \\ y_k &= Cx_k + v_k \end{aligned} \tag{1}$$

in which $x_k \in \mathbb{R}^n, u_k \in \mathbb{R}^m, y_k \in \mathbb{R}^p$ are the state, input and output of the system at time t_k. The noises corrupting the state and the output ($w_k \in \mathbb{R}^g$ and $v_k \in \mathbb{R}^p$) are modelled to be zero-mean independent Gaussian noise sequences with covariances Q_w and R_v respectively. With no constraints the optimal state estimation is achieved by the classical Kalman filter [12]. If the Gaussian assumption is relaxed, the Kalman filter is the still the optimal filter among the class of linear filters [10].

If knowledge of the deterministic part of the model i.e. A, B, C is known from ID experiments, then the Kalman filter or any state estimator would require the knowledge of stochastic part of the model i.e. G, Q_w, R_v. The G matrix shapes the disturbance w_k entering the state. Techniques from the ID literature estimate the the statistics of the output residuals along with A, B, C, but not the disturbance structure G or the covariances Q_w and R_v. In [20], the estimation of the covariances Q_w, R_v using the correlations between the measurements at different times was presented. The correlation based method was largely pioneered by Mehra [15–17] and adapted by many others [3, 5, 11, 18, 19]. In technique in [20] result in a least-squares problem to be solved in Q_w and R_v.

The correct weighting is needed for the least-squares estimate of Q_w, R_v to have the minimum variance. The weighted least squares estimate has the lowest variance among the class of all linear estimators. Here we present the theoretical weighting for the minimum variance ALS technique to estimate Q_w, R_v from data.

All of these techniques assume that the disturbance structure as given by the G matrix is known. In the absence of any knowledge about G an assumption that $G = I$ is often made, which implies that an independent disturbance enters each of the states. To best of our knowledge there exists no technique in the literature to estimate the structure of the disturbances as given by the G matrix.

II. BACKGROUND

If $\hat{x}_k$ denotes the estimate of the state and $L \in \mathbb{R}^{n \times p}$ is some arbitrary stable filter gain (not necessarily optimal), then the state estimates are given by:

$$\hat{x}_{k+1} = A\hat{x}_k + Bu_k + AL(y_k - C\hat{x}_k) \tag{2}$$

When the system is unconstrained, the optimal state estimator is the Kalman filter. For the Kalman filter the filer gain L_o is calculated by solving the Riccati equation:

$$\begin{aligned} P &= APA^T - APC^T(CPC^T + R_v)^{-1}CPA^T + GQ_wG^T \\ L_o &= PC^T(CPC^T + R_v)^{-1} \end{aligned} \tag{3}$$

The estimate error covariance is $P = E[(x_k - \hat{x}_k)(x_k - \hat{x}_k)^T]$. In the absence of the knowledge of the noise statistics the covariances are set heuristically and the filter gain L is changed in

an ad-hoc way to get reasonable performance from the closed-loop controller. We can write the evolution of the state estimate error $\varepsilon_k = x_k - \hat{x}_k$ by subtracting Equations 2 from 1:

$$\begin{aligned} \varepsilon_{k+1} &= \underbrace{(A - ALC)}_{\bar{A}} \varepsilon_k + \underbrace{\begin{bmatrix} G & -AL \end{bmatrix}}_{\bar{G}} \begin{bmatrix} w_k \\ v_k \end{bmatrix} \\ \mathscr{Y}_k &= C\varepsilon_k + v_k \end{aligned} \tag{4}$$

in which the L-innovations are defined as $\mathscr{Y}_k \triangleq y_k - C\hat{x}_k$. Note that the L-innovations are not white if the initial state estimator L is not optimal [2].

Skipping the lengthy algebra for brevity and referring the reader to [20] for detailed derivations, we only note here that given past data $\{\mathscr{Y}_1, \mathscr{Y}_2, \cdots, \mathscr{Y}_{N_d}\}$, we can form the following least-squares problem in a vector of unknowns, GQ_wG^T, R_v. The subscript 's' denotes the elements of the matrix stacked in a vector.

$$\begin{aligned} \Phi = \min_{GQ_wG^T, R_v} \quad & \left\| \mathscr{A} \begin{bmatrix} (GQ_wG^T)_s \\ (R_v)_s \end{bmatrix} - \hat{b} \right\|^2 \\ \text{subject to,} \quad & GQ_wG^T, R_v \geq 0, \quad R_v = R_v^T \end{aligned} \tag{5}$$

The $\mathscr{A}$ matrix is calculated using the system matrices and the initial estimator tuning L. The vector $\hat{b}$ is estimated from the autocovariances of the L-innovations $\{\mathscr{Y}_1, \mathscr{Y}_2, \cdots, \mathscr{Y}_{N_d}\}$. The estimation method in Equation 5 is referred to as the **Autocovariance Least-Squares (ALS)** technique in the sequel.

III. MINIMUM VARIANCE AND OPTIMAL WEIGHTING

It is well known that the Bayesian estimation procedure gives the minimum variance estimates of the parameters. The minimum variance estimator for the covariances Q_w, R_v would involve finding the posterior expectation of the covariances conditional on the data. In the limit of having enough data, the a priori distribution of the covariances becomes negligible and Bayesian estimation becomes similar to the Maximum Likelihood (ML) estimation. ML estimation procedures aim to maximize the likelihood function with respect to the parameters. The innovations form of the likelihood function for the Gaussian state-space model described by Equation 1 is given as (ignoring a constant) [21]:

$$-2\ln L_Y(\theta) = \sum_{k=1}^{N_d} \log|\Sigma(\theta)| + \sum_{k=1}^{N_d} \mathscr{Y}_k(\theta)^T \Sigma(\theta)^{-1} \mathscr{Y}_k(\theta) \tag{6}$$

in which $L_Y(\theta)$ is the likelihood and $\Sigma = CPC^T + R_v$.

The estimation of the unknown covariances then follows an iterative scheme:

- Guess GQ_wG^T, R_v and stack them in a vector $\theta = [(GQ_wG^T)_s, (R_v)_s]$.
- Calculate $\Sigma = CPC^T + R_v$ and $\mathscr{Y}_k$ using the guessed values for the covariances.
- Compute the innovations form of the likelihood function from Equation 6 and find a gradient direction to maximize $L_Y(\theta)$.

- Iterate until the likelihood function is maximized.

As emphasized by Shumway and Stoffer [23, chap. 4], the L-innovations $\mathscr{Y}_k$ are dependent on the parameter θ and the likelihood function is a highly nonlinear and complicated function of θ. Newton-Raphson methods can be used recursively to update the parameters and to maximize the objective function in Equation 6. A more robust algorithm was given by Shumway and Stoffer [22] based on the EM (expectation-maximization) algorithm originally used by Dempster [7]. The EM algorithm again is an iterative procedure that like other algorithms struggles to find a global maximum starting from a bad initial guess.

To avoid the complicated nonlinear approach detailed till now in this section, we can instead try to find a linear unbiased minimum variance estimator for the covariances. Given a linear model the linear unbiased minimum variance estimator is the weighted least-squares estimator as given by Theorem 1 [1].

Theorem 1 *For a linear model of the form $y = Ax + e$ with $E[e] = 0$ and $E[ee^T] = R$, the weighted least-squares estimator for x is formulated as:*

$$\min_x \|Ax - y\|^2_{R^{-1}}$$

The weighted least-squares estimator given by

$$\hat{x} = (A^T R^{-1} A)^{-1} A^T R^{-1} y$$

then has the minimum variance among all linear unbiased estimators.

Proof 1 *The statement a classical generalized least squares result for the linear regression model first considered by Aitken [1]. A more recent proof can be found for example in [14, p. 259].*

The weighted least-squares estimation of the covariances is given by the ALS technique as shown by Equation 5 . In [20] however, the weighting matrix W in the ALS problem is taken to be the identity matrix. The minimum variance property for the estimates then does not hold. We will now derive the formula for the weighting matrix W.

Following the analogy of Lemma 1 for Equation 5, if $\hat{b}$ is an unbiased estimator of $\bar{b}$, then $\bar{b} = E[\hat{b}]$. Define $S \triangleq E[(\hat{b} - \bar{b})(\hat{b} - \bar{b})^T]$ as the covariance of $\hat{b}$. Then $W = S^{-1}$ is the weighting for the ALS problem. In is shown in [20] that $\hat{b}$ is indeed an unbiased estimator.

If $\{\mathscr{Y}_1, \mathscr{Y}_2, \cdots \mathscr{Y}_{N_d}\}$ are the set of L-innovations calculated from data as given by Equation 4, and N is the window size used for the autocovariances (see [20]) then we define the matrix $\mathbb{Y}$ as follows (assuming N_d to be divisible by N):

$$\mathbb{Y} \triangleq \begin{bmatrix} \mathscr{Y}_1 & \mathscr{Y}_{N+1} & \cdots & \mathscr{Y}_{N_d-N+1} \\ \mathscr{Y}_2 & \mathscr{Y}_{N+2} & \cdots & \mathscr{Y}_{N_d-N+2} \\ \vdots & \vdots & \vdots & \vdots \\ \mathscr{Y}_N & \mathscr{Y}_{2N} & \vdots & \mathscr{Y}_{N_d} \end{bmatrix} \tag{7}$$

$\mathbb{Y} \in \mathbb{R}^{Np \times \frac{N_d}{N}}$. Define $\tilde{n} \triangleq \frac{N_d}{N}$ and $\tilde{p} \triangleq Np$.

The estimate $\hat{b}$ is given by:

$$\hat{b} = \left(\frac{1}{N_d - N} \mathbb{Y}\mathbb{Y}^T \right)_s \tag{8}$$

Lemma 1 *Given the L-innovations from Equation 4 and the definition of* $\mathbb{Y}$ *from Equation 7, we have*

$$\begin{aligned} E[\mathbb{Y}] &= 0 \\ E[\mathbb{Y}\mathbb{Y}^T] &\triangleq E[\mathbb{Y}_s\mathbb{Y}_s^T] \\ &= \Omega \end{aligned}$$

with Ω *as defined in Appendix A (Equation A2). The random matrix* $\mathbb{Y}$ *is distributed normally with* $\mathbb{Y} \sim N(0, \Omega)$.

Proof of Lemma 1 is given in Appendix A.

Note that the formula for Ω as given by Equation A2 is dependent on the unknown disturbance covariances Q_w, R_v and G. Following the derivation in [9] for calculating higher order moments for normally distributed matrices, we can calculate the covariance of $\hat{b}$. From Equation 8 we have:

$$S = \mathrm{cov}(\hat{b}) \tag{9}$$

$$= \frac{1}{(N_d - N)^2}\mathrm{cov}(\mathbb{Y}\mathbb{Y}^T) \tag{10}$$

Using Lemma 1 and [9], the formula for S is given in Appendix B (Equation B1). The optimal weight is $W = S^{-1}$. If S is singular, then without loss of generality we can take $W = S^{\dagger}$, the Moore-Penrose pseudoinverse of S.

The weight W it a complicated function dependent on the values of the unknown covariances. A recursive calculation may be carried out for calculating W and the covariances.

1. Guess a value for $\widehat{GQ_wG^T}, \hat{R}_v$ and calculate Ω and $W = S^{-1}$ using Equations A2 and B1.
2. Use the ALS technique to estimate $\widehat{GQ_wG^T}, \hat{R}_v$ using Equation 5
3. Use estimates in previous step to recalculate W
4. Iterate until convergence

The convergence of the above iterative scheme has not tested because of the computational burden (see Remark 2).

Remark 1 *If the initial estimator gain L was optimal, the L-innovations (or just innovations) would be white. The formula for S (Equation B1) would then be much simpler and would be the second moment of the Wishart distribution given in literature. White innovations would also imply optimality of the filter and there would be no need to calculate the covariances. In the more practical situation when the L-innovations are not white, the assumption of 'whiteness' would lead to an incorrect weighting. This incorrect weighting was used in [6].*

Remark 2 *The computation of S from Equation B1 becomes prohibitively large even for a small dimensional problem with large data sets. This will be a drawback for any practical application until efficient means for the computation are found.*

IV. ESTIMATION OF DISTURBANCE STRUCTURE

Generally a linear model of a chemical plant has many states and only a few independent disturbances. Any noise w_k that enters the state x_{k+1} is first scaled by the G matrix and then by the C matrix before it is measured in the output y_{k+1} (Equation 1). It is unusual to have model information about the G matrix in applications. Also if there are fewer sensors than the number of states, then the information contained in the measurements is also usually not enough to estimate a full rank GQ_wG^T matrix using Equation 5. There can then be multiple covariance matrices that generate the same statistics in the output data.

Our aim is to find the minimum rank Q (we use Q as a notation to denote the estimate GQ_wG^T in the rest of this section). A minimum rank Q can be decomposed as follows:

$$\begin{aligned} Q &= GQ_wG^T \\ Q = \tilde{G}\tilde{G}^T \qquad & \tilde{Q}_w = I \end{aligned} \tag{11}$$

Having Q with minimum rank would ensure that $\tilde{G}$ has the minimum number of columns. The number of columns in the matrix G is equal to the number of independent disturbances entering the state and equal to the rank of Q. Hence, by estimating $\tilde{G}$, we will get information about the minimum number of independent disturbances entering the data along with the disturbance structure and covariances. With reference to Equation 11, one might think that the natural step would be to solve the optimization directly with G as the optimization variable instead of solving with Q and then following the decomposition. The reason for solving with Q as the optimization variable is to avoid the nonlinearity that would be introduced if the elements of G are used as optimization variables.

The rank can be explicitly added to the objective in Equation 5 through a weighting parameter ρ multiplying the rank:

$$\begin{aligned} \Phi_1 = \min_{Q,R_v} & \underbrace{\left\| \mathscr{A} \begin{bmatrix} (Q)_s \\ (R_v)_s \end{bmatrix} - \hat{b} \right\|^2}_{\Phi} + \rho \text{Rank } (Q) \\ & Q, R_v \geq 0, \quad Q = Q^T, \quad R_v = R_v^T \end{aligned} \tag{12}$$

The constraints are in the form of convex Linear Matrix Inequalities (LMI) [4, 24]. The norm part of the objective is also convex. The rank however can only take integer values and makes the problem NP hard [25]. The solution of minimizing the rank subject to LMI constraints is an open research question and current techniques are largely based on heuristics.

Since the rank is the number of nonzero eigenvalues of a matrix, a good heuristic substitute for the rank is the sum of its eigenvalues or the trace of the matrix. The trace of a matrix is also the largest convex envelope over the rank of the matrix [8].

$$\text{Rank } (Q)_{\min} \geq \frac{1}{\lambda_{\max}(Q)} \text{Tr } (Q)$$

The trace of a matrix is a convex function of Q. The optimization in Equation 12 can be

rewritten with the trace replacing the rank:

$$\Phi_1 = \min_{Q,R_v} \underbrace{\left\| \mathscr{A} \begin{bmatrix} (Q)_s \\ (R_v)_s \end{bmatrix} - b \right\|^2}_{\Phi} + \rho \mathrm{Tr}\ (Q) \qquad (13)$$
$$Q, R_v \geq 0, \quad Q = Q^T, \quad R_v = R_v^T$$

Equation 13 is in the form of a Semidefinite Programming (SDP) problem. We will refer to this problem as the **ALS-SDP (Autocovariance Least-Squares with Semidefinite Programming)**.

The ALS-SDP method gives a feasible solution for each value of the parameter ρ by using simple Newton-like algorithms. The choice of ρ is made from a tradeoff plot of Tr (Q) versus Φ. The choice of ρ is made such that Tr (Q) is small and any further decrease in value of Tr (Q) will cause significant increase in the value of Φ.

V. EXAMPLE

Let the plant be simulated using the following state-space matrices.

$$A = \begin{bmatrix} 0.733 & -0.086 \\ 0.172 & 0.991 \end{bmatrix} \quad C = \begin{bmatrix} 1 & 2 \end{bmatrix} \quad G = \begin{bmatrix} 1 \\ 0.5 \end{bmatrix}$$
$$Q_w = 0.5 \quad R_v = 1$$

Assume G is unknown. Since C is not full column rank, the estimate of $Q = GQ_wG^T$ using ALS is not unique. The ALS-SDP technique then estimates a minimum rank $\hat{Q}$ and $\hat{R}_v$ that represent the statistics in the output data.

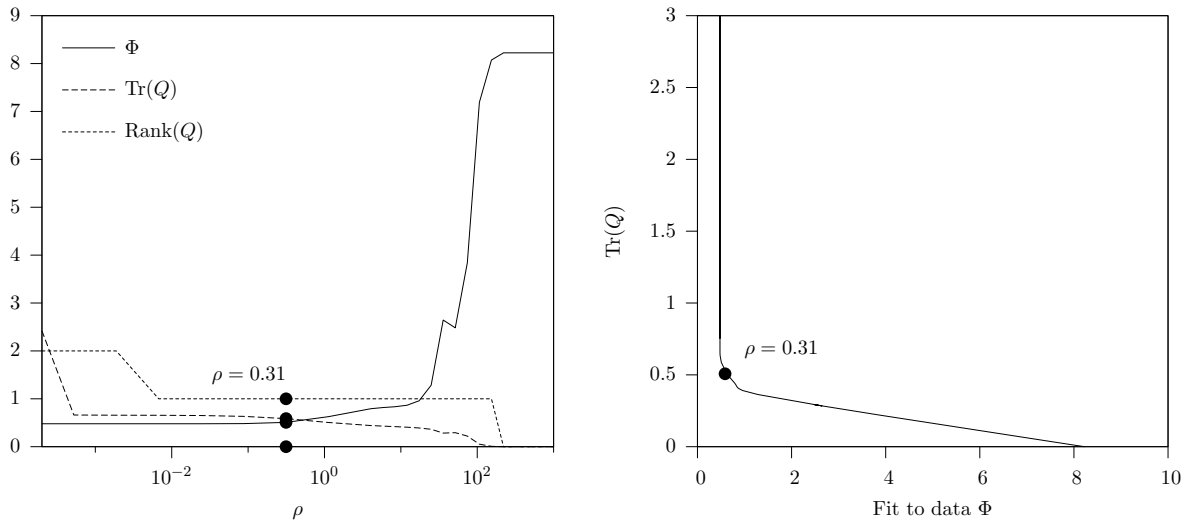

FIG. 1: Tradeoff between fit to data and trace

The results from the new ALS-SDP are shown in Figure 1. The plots show that choice of $\rho = 0.31$ is where the Tr (Q) is the minimum with no significant change in Φ. Also, the

rank(Q) at $\rho = 0.31$ can be seen to be 1, which is the number of independent disturbances entering the state of the simulated data. Also the estimated disturbance structure and covariances were (using $\rho = 0.31$):

$$\hat{Q} = \widehat{GQ_wG^T} = \begin{bmatrix} 0.449 & 0.249 \\ 0.249 & 0.138 \end{bmatrix}, \quad \hat{R}_v = 0.99$$

After decomposition according to Equation 11 we get, $\hat{G} = [0.670, 0.372]^T, \hat{Q}_w = 1$.

Once a positive semidefinite $\hat{Q}$ and a positive definite $\hat{R}_v$ are estimated from data, the Kalman filter gain (optimal state estimator for unconstrained linear models) can be obtained by solving the Riccati equation 3. The comparison between the estimated $\hat{L}$ and the optimal L_o is given below:

$$\hat{L} = \begin{bmatrix} 0.312 \\ 0.211 \end{bmatrix} \qquad L_o = \begin{bmatrix} 0.328 \\ 0.202 \end{bmatrix}$$

VI. CONCLUSIONS

For linear models we showed an estimation procedure for the disturbance structure given by the matrix G in Equation 1. Estimation of the minimum number of disturbances affecting the states is equivalent to minimizing the rank of GG^T. An estimation procedure using semidefinite programming and a rank heuristic was shown to give a tradeoff between fit the data and the minimization of the rank. The 'knee' of the tradeoff curve was shown to give good estimates for the minimum number of disturbances.

[1] A. C. Aitken. On least squares and linear combinations of observations. *Proceedings of the Royal Society of Edinburgh*, 55:42–48, 1935.

[2] Brian D. O. Anderson and John B. Moore. *Optimal Filtering*. Prentice-Hall, Englewood Cliffs, N.J., 1979.

[3] P.R. Bélanger. Estimation of noise covariance matrices for a linear time-varying stochastic process. *Automatica*, 10:267–275, 1974.

[4] Stephen Boyd, Laurent El Ghaoui, Eric Feron, and Venkataramanan Balakrishnan. *Linear Matrix Inequalities in Control Theory*. SIAM, Philadelphia, 1994.

[5] B. Carew and P.R. Bélanger. Identification of optimum filter steady-state gain for systems with unknown noise covariances. *IEEE Transactions on Automatic Control*, 18(6):582–587, 1973.

[6] D.P. Dee, S.E. Cohn, A. Dalcher, and M. Ghil. An efficient algorithm for estimating noise covariances in distributed systems. *IEEE Transactions on Automatic Control*, 30(11):1057–1065, 1985.

[7] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B*, 39:1–38, 1977.

[8] Maryam Fazel. *Matrix Rank Minimization with Applications*. PhD thesis, Dept. of Elec. Eng., Stanford University, 2002.

[9] Ghazal A. Ghazal and Heinz Neudecker. On second-order and fourth-order moments of jointly distributed random matrices: a survey. *Linear Algebra and its Applications*, 321:61–93, 2000.

[10] Graham C. Goodwin and Kwai Sang Sin. *Adaptive Filtering Prediction and Control*. Prentice-Hall, Englewood Cliffs, New Jersey, 1984.

[11] A. Isaksson. Identification of time varying systems through adaptive Kalman filtering. In *Proceedings of the 10th IFAC World Congress, Munich, Germany*, pages 305–310. Pergamon Press, Oxford, UK, 1987.

[12] R. E. Kalman and R. S. Bucy. New results in linear filtering and prediction theory. *Trans. ASME, J. Basic Engineering*, pages 95–108, March 1961.

[13] Jan R. Magnus and H. Neudecker. The commutation matrix: Some properties and applications. *Annals of Statistics*, 7(2):381–394, March 1979.

[14] Jan R. Magnus and Heinz Neudecker. *Matrix Differential Calculus with Applications in Statistics and Econometrics*. John Wiley, New York, 1999.

[15] R.K. Mehra. On the identification of variances and adaptive Kalman filtering. *IEEE Transactions on Automatic Control*, 15(12):175–184, 1970.

[16] R.K. Mehra. On-line identification of linear dynamic system with application to Kalman filtering. *IEEE Transactions on Automatic Control*, 16(1):12–21, 1971.

[17] R.K. Mehra. Approaches to adaptive filtering. *IEEE Transactions on Automatic Control*, 17:903–908, 1972.

[18] C. Neethling and P. Young. Comments on "Identification of optimum filter steady-state gain for systems with unknown noise covariances". *IEEE Transactions on Automatic Control*, 19(5):623–625, 1974.

[19] Gerardo Noriega and Subbarayan Pasupathy. Adaptive estimation of noise covariance matrices in real-time preprocessing of geophysical data. *IEEE Transactions on Geoscience and Remote Sensing*, 35(5):1146–1159, 1997.

[20] Brian J. Odelson, Murali R. Rajamani, and James B. Rawlings. A new autocovariance least-squares method for estimating noise covariances. *Automatica*, 42(2):303–308, February 2006.

[21] Fred C. Schweppe. Evaluation of likelihood functions for gaussian signals. *IEEE Transactions on Information Theory*, 11(1):61–70, January 1965.

[22] Robert H. Shumway and David S. Stoffer. An approach to time series smoothing and forecasting using the em algorithm. *Journal of Time Series Analysis*, 3:253–264, 1982.

[23] Robert H. Shumway and David S. Stoffer. *Time Series Analysis and Its Applications*. Springer-Verlag, New York, 2000.

[24] J.G. VanAntwerp and R.D. Braatz. A tutorial on linear and bilinear matrix inequalities. *Journal of Process Control*, 10(4):363–385, 2000.

[25] Lieven Vandenberghe and Stephen Boyd. Semidefinite programming. *SIAM Review*, 38(1):49–95, March 1996.

APPENDIX A: PROOF OF LEMMA 1

From Equation 4, we have

$$\varepsilon_k = \bar{A}^k \varepsilon_0 + \sum_{j=0}^{k-1} \bar{A}^{k-j-1} \bar{G} \begin{bmatrix} w_j \\ v_j \end{bmatrix} \tag{A1}$$

Taking the expectation of the above expression and noting that $E[v_k] = E[w_k] = 0$, we get,

$$E[\varepsilon_k] = \bar{A}^k E[\varepsilon_0] = 0$$

The equality follows from the stability of the initial filter gain L since for k large enough, we have $\bar{A}^k = (A - ALC)^k \approx 0$.

Taking the expectation of the L-innovations in Equation 4, we get:

$$E[\mathscr{Y}_j] = CE[\varepsilon_j] + E[v_k] = 0$$

holding for all $j \geq k$ (k is the initial period of transience, when for $i < k$, $E[\varepsilon_i]$ cannot be approximated as 0). Thus, we have

$$E[\mathbb{Y}] = E\left[\begin{pmatrix} \mathscr{Y}_1 \\ \mathscr{Y}_2 \\ \vdots \\ \mathscr{Y}_{N_d} \end{pmatrix}\right] = 0$$

Now, calculate Ω the second moment of $\mathbb{Y}$ as follows:

$$\Omega = E\left[\begin{pmatrix} \mathscr{Y}_1 \\ \mathscr{Y}_2 \\ \vdots \\ \mathscr{Y}_{N_d} \end{pmatrix} \begin{pmatrix} \mathscr{Y}_1^T & \mathscr{Y}_2^T & \cdots & \mathscr{Y}_{N_d}^T \end{pmatrix}\right]$$

Following the derivation along lines similar to [20], we get:

$$\begin{aligned} \Omega = & \underbrace{\begin{bmatrix} C \\ C\bar{A} \\ \vdots \\ C\bar{A}^{N_d-1} \end{bmatrix}}_{\mathcal{O}} P\mathcal{O}^T + \begin{bmatrix} R_v & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & R_v \end{bmatrix} + \Psi \begin{bmatrix} R_v & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & R_v \end{bmatrix} \\ & + \underbrace{\begin{bmatrix} 0 & 0 & 0 & 0 \\ C\bar{G} & 0 & 0 & 0 \\ \vdots & \ddots & & \vdots \\ C\bar{A}^{N_d-2}\bar{G} & \cdots & C\bar{G} & 0 \end{bmatrix}}_{\Gamma} \begin{bmatrix} \bar{Q}_w & 0 & 0 & 0 \\ 0 & \bar{Q}_w & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \bar{Q}_w \end{bmatrix} \Gamma^T + \begin{bmatrix} R_v & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & R_v \end{bmatrix} \Psi^T \end{aligned} \tag{A2}$$

where,

$$\Psi = \Gamma \begin{bmatrix} -AL & 0 & 0 & 0 \\ 0 & -AL & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & -AL \end{bmatrix}, \qquad \bar{Q}_w = \begin{bmatrix} Q_w & 0 \\ 0 & R_v \end{bmatrix}$$

Following Equation A1, we see that ε_k is a linear combination of normally distributed noises given $\bar{A}^k \approx 0$ and hence is normal. This implies $\mathscr{Y}_k$ is also normally distributed. We then have:

$$\mathbb{Y} \sim N(0, \Omega)$$

APPENDIX B: WEIGHTING MATRIX

Given Kij is the commutation matrix as defined in [13], we define T as:

$$T \triangleq (I_{\tilde{p}^2} \otimes (I_{\tilde{n}^2})_s)^T (I_{\tilde{p}} \otimes K_{\tilde{p}\tilde{n}} \otimes I_{\tilde{n}})$$

If $\mathbb{Y} \sim N(0, \Omega)$ then the formula for $S = \text{cov}(\mathbb{Y}\mathbb{Y}^T)$ is given by [9]:

$$S = T(I_{\tilde{n}^2\tilde{p}^2} + K_{(\tilde{n}\tilde{p})(\tilde{n}\tilde{p})})((K_{\tilde{p}\tilde{n}}\Omega K_{\tilde{n}\tilde{p}}) \otimes (K_{\tilde{p}\tilde{n}}\Omega K_{\tilde{n}\tilde{p}}))T^T \tag{B1}$$