TWMCC ⋆ Texas-Wisconsin Modeling and Control Consortium

Technical report number 2007-02

Estimation of the Disturbance Structure from Data using Semidefinite Programming and Optimal Weighting*

Murali R. Rajamani† and James B. Rawlings‡
Department of Chemical and Biological Engineering
University of Wisconsin-Madison
Madison, WI 53706

29 June 2007

*A version of this report was submitted for publication in *Automatica*
†rmurali@wisc.edu
‡rawlings@engr.wisc.edu

Abstract

Tuning a state estimator for a linear state space model requires knowledge of the characteristics of the independent disturbances entering the states and the measurements. In Odelson, Rajamani, and Rawlings (2006), the correlations between the innovations data were used to form a least-squares problem to determine the covariances for the disturbances. In this paper we present new and simpler necessary and sufficient conditions for the uniqueness of the covariance estimates. We also formulate the optimal weighting to be used in the least-squares objective in the covariance estimation problem to ensure minimum variance in the estimates. A modification to the above technique is then presented to estimate the stochastic disturbance structure that affects the states. The disturbance structure also provides information about the minimum number of disturbances affecting the state. This minimum number is usually unknown and must be determined from data. A semidefinite optimization problem is solved to estimate the disturbance structure and the covariances of the noises entering the system.

Keywords

State estimation; Kalman filter; covariance estimation; disturbance structure; optimal weighting; minimum variance estimation; semidefinite programming

1 Introduction

We start with the linear time-invariant state-space model in discrete time:

$$x_{k+1} = Ax_k + Bu_k + Gw_k \tag{1a}$$

$$y_k = Cx_k + v_k \tag{1b}$$

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 dimensions of the system matrices are $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $G \in \mathbb{R}^{n\times g}$ and $C \in \mathbb{R}^{p\times n}$. The noises corrupting the state and the output ($w_k \in \mathbb{R}^g$ and $v_k \in \mathbb{R}^p$) are modelled as zero-mean Gaussian noise sequences with covariances Q_w and R_v respectively. The noises w_k and v_k are assumed to be statistically independent for simplicity. The case where w_k and v_k are dependent can be handled as shown in Åkesson, Jørgensen, Poulsen, and Jørgensen (2007). The optimal filtering or state estimation for the model given in Equations 1a, 1b when there are no constraints on the input and the state is given by the classical Kalman filter (Kalman and

Bucy, 1961). If the Gaussian assumption is relaxed, the Kalman filter is the still the optimal filter among the class of all linear filters (Goodwin and Sin, 1984; Anderson and Moore, 1979).

If complete knowledge about the deterministic part of the model i.e. A, B, C is assumed, then the Kalman filter or for that matter any state estimator requires 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. Physical systems often have only a few independent disturbances which affect the states. This implies a tall G matrix with more rows than columns. In Odelson et al. (2006), an autocovariance least-squares method for estimating the covariances Q_w, R_v was presented. The estimation technique was based on the correlations between the measurements at different times. The correlation based method was largely pioneered by Mehra (1970, 1971, 1972) and adapted by many others (Neethling and Young, 1974; Isaksson, 1987; Carew and Bélanger, 1973; Bélanger, 1974; Noriega and Pasupathy, 1997). 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. This type of independence of the disturbances is unlikely for physical reasons. To the best of our knowledge, there exists no technique in the literature to estimate the structure of the disturbances entering the state, which we do in this paper. We also give the formula for a linear unbiased minimum variance estimation of the covariances. Throughout we assume complete knowledge about A, B, C and treat the stochastic part of the model as the only unknowns.

The rest of the paper is organized as follows: In Section 2 we give some mathematical preliminaries that are required to understand the rest of the paper. Section 3 gives the formulation of the Autocovariance Least-Squares (ALS) technique simplified from Odelson et al. (2006). The main contributions of this paper are then presented in Sections 4, 5 and 6. Simple mathematical conditions to check for uniqueness of the covariance estimates are proved in Section 4 and the results used in the remaining sections. In Section 5, we find the optimal weighting matrix to calculate the linear unbiased minimum variance estimates of the covariances. In Section 6 we estimate the noise shaping matrix G from data using Semidefinite Programming (SDP). The G matrix contains information about the disturbance structure and the number of independent disturbances affecting the state equals to the number of columns in G.

2 Background

Assumption 1. *We assume that the pair (A, C) is observable*

We use the notation $\hat{x}_k$ to denote any estimate of the state x_k. If $L \in \mathbb{R}^{n \times p}$ is any

arbitrary, stable filter gain, then the state estimates are calculated recursively as:

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

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:

$$\begin{aligned} P_o &= AP_oA^T - AP_oC^T(CP_oC^T + R_v)^{-1}CP_oA^T + GQ_wG^T \\ L_o &= P_oC^T(CP_oC^T + R_v)^{-1} \end{aligned} \tag{3}$$

The optimal estimate error covariance is $P_o = E[(x_k - \hat{x}_k)(x_k - \hat{x}_k)^T]$ calculated as above. As seen in Equation 3, tuning the optimal state estimator (the Kalman filter for the linear unconstrained case), requires information about the covariances Q_w and R_v. In the absence of this information 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.

Given some arbitrary (stable, perhaps suboptimal) initial estimator L, we can write the evolution of the state estimate error $\varepsilon_k = x_k - \hat{x}_k$ by subtracting Equation 2 from 1a and substituting 1b:

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

in which $\mathscr{Y}_k$ are the L-innovations defined as $\mathscr{Y}_k \triangleq y_k - C\hat{x}_k$. Note that the L-innovations are uncorrelated in time if the initial state estimator L is optimal (i.e. $L = L_o$) (Anderson and Moore, 1979). We use the term L-innovations to distinguish them from the optimal innovations obtained by using the optimal state estimates.

Assumption 2. *The L-innovations data $\{\mathscr{Y}_1, \cdots \mathscr{Y}_{N_d}\}$ used in the techniques described in this paper are obtained after the system has reached steady state and any initial transience can be neglected when $\bar{A}$ is stable*

Given a set of steady state L-innovations data $\{\mathscr{Y}_1, \cdots \mathscr{Y}_{N_d}\}$, we want to form a weighted least-squares problem in the unknown disturbance covariances, GQ_wG^T, R_v. One of the motivations behind using a least-squares approach is to avoid a complicated nonlinear approach required for techniques involving maximum likelihood estimation eg. Shumway and Stoffer (1982).

In the subspace ID literature (Gevers, 2006; Van Overschee and De Moor, 1994, 1995; Viberg, 1995; Juang and Phan, 1994; Qin, Lin, and Ljung, 2005), the identification procedures estimate the model and the stochastic parameters starting with the model in the innovations form, which is Equation 2 rewritten as:

$$\hat{x}_{k+1} = A\hat{x}_k + Bu_k + AL_oe_k \tag{5a}$$

$$y_k = C\hat{x}_k + e_k \tag{5b}$$

Here e_k are the optimal innovations (as opposed to the L-innovations) and hence uncorrelated in time. The estimation of the system matrices $\hat{A}, \hat{B}, \hat{C}$ is carried out along with the optimal Kalman filter gain $\hat{L}_o$, where the $\hat{\cdot}$ symbol denotes an estimate.

Notice the difference between Equations 5a,5b and Equations 1a, 1b. If the subspace ID techniques are used to identify only the stochastic parameters then the disturbance covariances as identified as $A\hat{L}_o S\hat{L}_o^T A^T$ instead of GQ_wG^T for the state noise and S instead of R_v for the measurements, where S is the covariance of e_k given by:

$$S = CP_oC^T + R_v$$

where, P_o is defined in Equation 3.

Remark 1. *As shown above, subspace ID techniques estimate a different set of covariances than G, Q_w, R_v. The aims of subspace ID are different and the estimates of the stochastic parameters are simply used to compute the optimal estimator gain. Finding the covariance parameters affecting the system (G, Q_w, R_v) on the other hand provides more flexibility in the choice of the state estimator. For example we may wish to employ a constrained, nonlinear moving horizon estimator (Rao, Rawlings, and Lee, 2001). In addition estimating G, Q_w, R_v gives a more informative handle to monitor the disturbances than monitoring changes in the optimal estimator gain.*

Also see Remark 2 for requirements about exciting inputs in subspace ID techniques.

3 The Autocovariance Least-Squares (ALS) Technique

Following the derivation along the lines of Odelson et al. (2006), we use Equation 4 to write the following expectation of covariances:

$$E(\mathscr{Y}_k\mathscr{Y}_k^T) = CPC^T + R_v \tag{6}$$

$$E(\mathscr{Y}_{k+j}\mathscr{Y}_k^T) = C\bar{A}^jPC^T - C\bar{A}^{j-1}ALR_v \quad j \geq 1 \tag{7}$$

which are independent of k because of our steady state assumption. Again using Equation 4 we note that P satisfies the Lyapunov equation:

$$P = \bar{A}P\bar{A}^T + \underbrace{\begin{bmatrix} G & -AL \end{bmatrix}}_{\bar{G}} \underbrace{\begin{bmatrix} Q_w & 0 \\ 0 & R_v \end{bmatrix}}_{\bar{Q}_w} \bar{G}^T \tag{8}$$

In Odelson et al. (2006) the autocovariance matrix was defined as:

$$\mathscr{R}(N) = E\begin{bmatrix} \mathscr{Y}_k\mathscr{Y}_k^T & \cdots & \mathscr{Y}_k\mathscr{Y}_{k+N-1}^T \\ \vdots & \ddots & \vdots \\ \mathscr{Y}_{k+N-1}\mathscr{Y}_k^T & \cdots & \mathscr{Y}_k\mathscr{Y}_k^T \end{bmatrix} \tag{9}$$

where N is the number of lags. To avoid redundant definition of the lagged covariances, here we use only the first block column of the autocovariance matrix $\mathscr{R}_1(N)$:

$$\mathscr{R}_1(N) = E\begin{bmatrix} \mathscr{Y}_k\mathscr{Y}_k^T \\ \vdots \\ \mathscr{Y}_{k+N-1}\mathscr{Y}_k^T \end{bmatrix} \tag{10}$$

Using Equations 6, 7 and 8, we can write the $\mathscr{R}_1(N)$ as:

$$\mathscr{R}_1(N) = \mathscr{O}PC^T + \Gamma R_v \tag{11}$$

in which

$$\mathscr{O} = \begin{bmatrix} C \\ C\bar{A} \\ \vdots \\ C\bar{A}^{N-1} \end{bmatrix} \quad \Gamma = \begin{bmatrix} I_p \\ -CAL \\ \vdots \\ -C\bar{A}^{N-2}AL \end{bmatrix} \tag{12}$$

The single column block development of the ALS technique as above is preferred over the use of the full $\mathscr{R}(N)$ matrix as in Odelson et al. (2006) due to the simpler formulation when using only $\mathscr{R}_1(N)$.

In this result and those to follow, we employ the standard definitions and properties of the Kronecker product, Kronecker sum and the direct sum (Steeb, 1991; Graham, 1981; Van Loan, 2000). If use the s subscript to denote the column-wise stacking of the matrix into a vector, a useful Kronecker product result is $(AXB)_s = (B^T \otimes A)X_s$ (here $\otimes$ is the standard symbol for the Kronecker product).

We then stack Equation 11 and use the stacked form of Equation 8 to substitute out P:

$$\begin{aligned} b = (\mathscr{R}_1(N))_s = &[(C \otimes \mathscr{O})(I_{n^2} - \bar{A} \otimes \bar{A})^{-1}](GQ_wG^T)_s \\ &+ [(C \otimes \mathscr{O})(I_{n^2} - \bar{A} \otimes \bar{A})^{-1}(AL \otimes AL) \\ &+ (I_p \otimes \Gamma)](R_v)_s \end{aligned} \tag{13}$$

Now that we have Equation 13, we use the ergodic property of the L-innovations to estimate the autocovariance matrix $\mathscr{R}_1(N)$ from the given set of data (Jenkins and Watts, 1968):

$$\widehat{E[\mathscr{Y}_k\mathscr{Y}_{k+j}^T]} = \frac{1}{N_d - j}\sum_{i=1}^{N_d-j} \mathscr{Y}_i\mathscr{Y}_{i+j}^T \tag{14}$$

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 then we define the matrix $\mathbb{Y}$ as follows:

$$\mathbb{Y} \triangleq \begin{bmatrix} \mathscr{Y}_1 & \mathscr{Y}_2 & \cdots & \mathscr{Y}_{N_d-N+1} \\ \mathscr{Y}_2 & \mathscr{Y}_3 & \cdots & \mathscr{Y}_{N_d-N+2} \\ \vdots & \vdots & \vdots & \vdots \\ \mathscr{Y}_N & \mathscr{Y}_{N+1} & \vdots & \mathscr{Y}_{N_d} \end{bmatrix} \tag{15}$$

$\mathbb{Y} \in \mathbb{R}^{\tilde{p}\times\tilde{n}}$ where, $\tilde{n} \triangleq N_d - N + 1$ and $\tilde{p} \triangleq Np$. Using Equation 14, the estimate $\widehat{\mathscr{R}_1}(N)$ is then given by:

$$\widehat{\mathscr{R}_1}(N) = \frac{1}{N_d - N + 1}\mathbb{Y}\mathbb{Y}^T_{(1:p,:)} \tag{16}$$

and $\hat{b} = (\widehat{\mathscr{R}_1}(N))_s$. Here, $\mathbb{Y}_{(1:p,:)}$ is the first row block of $\mathbb{Y}$ also given by:

$$\mathbb{Y}_{(1:p,:)} = \underbrace{\begin{bmatrix} I_p & 0 & \cdots & 0 \end{bmatrix}}_{\mathbb{E}} \mathbb{Y} \tag{17}$$

Given the linear relation in Equation 13 and the estimate $\hat{b}$ from Equation 16, we can formulate the following positive definite constrained least-squares problem in the symmetric elements of the covariances GQ_wG^T, R_v:

$$\begin{aligned} \Phi = \min_{GQ_wG, R_v} \quad & \left\| \mathscr{A} \begin{bmatrix} \mathscr{D}_n(GQ_wG^T)_{ss} \\ (R_v)_{ss} \end{bmatrix} - \hat{b} \right\|^2_W \\ \text{subject to,} \quad & GQ_wG^T, R_v \geq 0, \quad R_v = R_v^T \end{aligned} \tag{18}$$

Here we introduce the notation of $(R_v)_{ss}$ to denote the column-wise stacking of only the symmetric $p(p+1)/2$ elements of the matrix R_v (eliminating the supra-diagonal elements). More explicitly there exists an unique matrix $\mathscr{D}_p \in \mathbb{R}^{p^2 \times \frac{p(p+1)}{2}}$ called the *duplication matrix* (Magnus and Neudecker, 1999, p. 49) containing ones and zeros that gives the relation $(R_v)_s = \mathscr{D}_p(R_v)_{ss}$.

Using Equation 13, we can then write $\mathscr{A}$ explicitly as:

$$\begin{aligned} \mathscr{A} =& [\mathscr{A}_1 \quad \mathscr{A}_2] \\ \mathscr{A}_1 =& [(C \otimes \mathscr{O})(I_{n^2} - \bar{A} \otimes \bar{A})^{-1}] \\ \mathscr{A}_2 =& [(C \otimes \mathscr{O})(I_{n^2} - \bar{A} \otimes \bar{A})^{-1}(AL \otimes AL) \\ & + (I_p \otimes \Gamma)]\mathscr{D}_p \end{aligned} \tag{19}$$

where, the *duplication matrix* $\mathscr{D}_p$ is included to ensure symmetry in the covariance estimates.

The estimation method in Equation 18 is referred to as the Autocovariance Least-Squares (ALS) technique in the sequel. A recent application of the ALS technique was presented in Zhuang, Rajamani, Rawlings, and Stoustrup (2007a,b). The ALS technique can also be used to estimate the optimal filter gain when there are integrating disturbance models in model predictive control (Rajamani, Rawlings, Qin, and Downs, 2006).

Remark 2. *A significant advantage of using the ALS technique and the modifications presented in the rest of this paper over other identification techniques is the use of only steady state data in the calculations. This means that unlike other identification techniques there is no requirement for exciting inputs to be applied to the system.*

4 Conditions for Uniqueness

In this section, we assume that the G matrix is a known $\in \mathbb{R}^{n\times g}$ matrix. Without loss of generality we can also assume G to be of full column rank. If G is not full column rank then a new matrix $\tilde{G}$ can be defined with its columns independent and such that $\tilde{G}\tilde{G}^T = GG^T$.

We next derive simple conditions for uniqueness for the ALS problem with Q_w, R_v as unknowns and a known G. In the rest of this section we also assume that the weighting for the norm in the objective function is $W = I$.

$$\begin{aligned} \Phi = \min_{Q_w,R_v} \quad & \left\| \tilde{\mathscr{A}} \begin{bmatrix} (Q_w)_{ss} \\ (R_v)_{ss} \end{bmatrix} - \hat{b} \right\|^2 \\ \text{s.t.} \quad & Q_w, R_v \geq 0, \quad R_v = R_v^T, \quad Q_w = Q_w^T \end{aligned} \tag{20}$$

where, $\tilde{\mathscr{A}} = \begin{bmatrix} \mathscr{A}_1(G\otimes G)\mathscr{D}_g & \mathscr{A}_2 \end{bmatrix}$

Lemma 1. *The optimization in Equation 20 has a unique solution if and only if $\tilde{\mathscr{A}}$ in Equation 20 has full column rank.*

Proof. Existence of a feasible solution is proved by observing that $Q_w = I_g$ and $R_v = I_p$ are valid solutions satisfying the constraints. To prove uniqueness, we see that $\tilde{\mathscr{A}}$ having full column rank guarantees the objective function in Equation 20 to be strictly convex. The constraints are on the covariance matrices being positive definite and hence also convex (Vandenberghe and Boyd, 1996; Boyd, Ghaoui, Feron, and Balakrishnan, 1994). Uniqueness then follows for a strictly convex objective function subject to convex constraints (Boyd and Vandenberghe, 2004, p. 137). □

Assumption 3. *We assume that the state transition matrix A in non-singular. If the original A is singular, then a similarity transformation can be used to eliminate the states with zero eigenvalues and the noise covariances redefined.*

Lemma 2. *If (A, C) is observable and A is non-singular, then the matrix $\tilde{\mathscr{A}}$ in Equation 20 has a null space if and only if the matrix M defined by, $M = (C \otimes I_n)(I_{n^2} - \bar{A} \otimes \bar{A})^{-1}(G \otimes G)\mathscr{D}_g$ also has a null space, and the null space of $\mathscr{A}_1(G \otimes G)\mathscr{D}_g$ which multiplies $(Q_w)_{ss}$ in Equation 20 is equal to the null space of M.*

The derivation is given in Appendix B.

Theorem 1. *If (A, C) is observable and A is non-singular, the optimization in Equation 20 has a unique solution if and only if $dim[Null(M)] = 0$, where:*

$$M = (C \otimes I_n)(I_{n^2} - \bar{A} \otimes \bar{A})^{-1}(G \otimes G)\mathscr{D}_g$$

Proof. The proof follows from Lemmas 1 and 2. □

Corollary 1. *If C is full column rank (i.e. the number of sensors equal the number of states), then the optimization in Equation 20 is unique.*

Proof. C having full column rank implies M in Theorem 1 has full rank and hence an empty null space. The optimization in Equation 20 then gives a unique solution according to Theorem 1. □

5 Minimum Variance and Optimal Weighting

Theorem 2. *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.

This statement is a classical generalized least squares result for the linear regression model first considered by Aitken (1935). A more recent proof can be found for example in Magnus and Neudecker (1999, p. 259).

The weighted least-squares estimation of the covariances is given by the ALS technique as shown by Equation 18. In Odelson et al. (2006) 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 next derive the formula for the minimum variance weighting matrix W.

Following the analogy of Theorem 2 for Equation 18, if $\hat{b}$ is an unbiased estimator of b, then $b = E[\hat{b}]$. Define $S \triangleq E[(\hat{b} - b)(\hat{b} - b)^T] = \text{cov}(\hat{b})$ as the covariance of $\hat{b}$. Then $W = S^{-1}$ is the weighting that gives minimum variance for the ALS problem. It is shown in Odelson et al. (2006) that $\hat{b}$ in Equation 16 is an unbiased estimator.

Lemma 3. *Given the L-innovations from Equation 4 and the definition of* $\mathbb{Y}$ *from Equation 15, 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 30). The random matrix* $\mathbb{Y}$ *is distributed normally with* $\mathbb{Y} \sim N(0, \Omega)$.

Proof of Lemma 3 is given in Appendix A.

Note that the formula for Ω as given by Equation 30 depends on the unknown disturbance covariances Q_w, R_v and G.

Theorem 3. *The minimum variance weight to use in the the ALS objective in Equation 18 is given by* $W = S^{\dagger}$*, where,*

$$S = \frac{T(I_{\tilde{n}^2\tilde{p}^2} + K_{(\tilde{n}\tilde{p})(\tilde{n}\tilde{p})})((K_{\tilde{p}\tilde{n}}\Omega K_{\tilde{n}\tilde{p}}) \otimes (K_{\tilde{p}\tilde{n}}\Omega K_{\tilde{n}\tilde{p}}))T^T}{(N_d - N + 1)^2} \tag{21}$$

and K_{ij} *is the commutation matrix defined in Magnus and Neudecker (1979).* T *is defined as:*

$$T = (\mathbb{E} \otimes I_p)(I_{\tilde{p}^2} \otimes (I_{\tilde{n}})_s)^T(I_{\tilde{p}} \otimes K_{\tilde{p}\tilde{n}} \otimes I_{\tilde{n}})$$

and $\mathbb{E} = [I_p, 0 \cdots 0]$. $\tilde{n} = N_d - N + 1$ *and* $\tilde{p} = Np$

Proof. Since $\mathbb{Y} \in \mathbb{R}^{\tilde{p}\times\tilde{n}}$ is a matrix as defined in Equation 15 which is normally distributed with mean 0 and covariance Ω as defined in Lemma 3, the fourth moment of $\mathbb{Y}$ is defined as follows:

$$\text{cov } [\mathbb{Y}\mathbb{Y}^T] \triangleq \text{cov}((\mathbb{Y}\mathbb{Y}^T)_s)$$

The formula for the fourth moment of $\mathbb{Y}$ i.e cov$(\mathbb{Y}\mathbb{Y}^T)$ for a normal distribution is given by:

$$\begin{aligned} &\text{cov } (\mathbb{Y}\mathbb{Y}^T) = \\ &\quad T_1(I_{\tilde{n}^2\tilde{p}^2} + K_{(\tilde{n}\tilde{p})(\tilde{n}\tilde{p})})((K_{\tilde{p}\tilde{n}}\Omega K_{\tilde{n}\tilde{p}}) \otimes (K_{\tilde{p}\tilde{n}}\Omega K_{\tilde{n}\tilde{p}}))T_1^T \end{aligned} \tag{22}$$

where, $T_1 = (I_{\tilde{p}^2} \otimes (I_{\tilde{n}})_s)^T(I_{\tilde{p}} \otimes K_{\tilde{p}\tilde{n}} \otimes I_{\tilde{n}})$. The formula follows from the the results in Ghazal and Neudecker (2000). See Ghazal and Neudecker (2000) for more details on the derivation. The commutation matrix K_{ij} is a $\in \mathbb{R}^{ij\times ij}$ matrix containing only 1's and 0's and gives the following relationship between $(A)_s$ and $(A^T)_s$ when A is a $\in \mathbb{R}^{i\times j}$ matrix: $(A)_s = K_{ij}(A^T)_s$ and $(A^T)_s = K_{ji}(A)_s$.

We also have from Equations 16 and 17:

$$\begin{aligned}\hat{b} &= (\mathbb{Y}\mathbb{Y}^T\mathbb{E}^T)_s \\ &= (\mathbb{E}\otimes I_p)(\mathbb{Y}\mathbb{Y}^T)_s\end{aligned}$$

From Equation 22 we can then calculate the covariance of $\hat{b}$:

$$\begin{aligned}S &= \mathrm{cov}(\hat{b}) \\ &= \frac{(\mathbb{E}\otimes I_p)\mathrm{cov}(\mathbb{Y}\mathbb{Y}^T)(\mathbb{E}^T\otimes I_p)}{(N_d - N + 1)^2}\end{aligned}$$

Thus we get Equation 21 as the covariance of $\hat{b}$ where $T = (\mathbb{E}\otimes I_p)T_1$.

The optimal weight is then $W = S^{-1}$ following Theorem 2. 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 depending 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 $\hat{Q}, \hat{R}_v$, where $Q = GQ_wG^T$ and calculate Ω and $W = S^{-1}$ using Equations 30 and 21.
2. Use the estimated weight in the ALS technique to estimate $\hat{Q}, \hat{R}_v$ using Equation 18
3. Use estimates in previous step to recalculate W
4. Iterate until convergence

The convergence of the above iterative scheme has not been tested because of the computational burden (see Remark 4).

Remark 3. *If the initial estimator gain L was optimal, the L-innovations (or just innovations) would be white. The formula for S (Equation 21) would then be much simpler and would be the second moment of the Wishart distribution (Anderson, 2003). 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 Dee, Cohn, Dalcher, and Ghil (1985).*

Remark 4. *The computation of S from Equation 21 becomes prohibitively large even for a small dimensional problem with large data sets. This is a drawback for any practical application until efficient means for the computation are found.*

Remark 5. *Although the weight may be estimated from data, a large data set is required before getting reliable estimates for the weights. An attractive alternative to circumvent the need for large data sets is to use Bootstrapping, for example Stoffer and Wall (1991).*

5.1 Example of Lower Variance

Consider the following model for the system:

$$x_{k+1} = \begin{bmatrix} 0.732 & -0.086 \\ 0.172 & 0.990 \end{bmatrix} x_k + \begin{bmatrix} 1 & 0 \\ 0 & 0.2 \end{bmatrix} w_k$$
$$y_k = x_k + v_k$$

Data is generated by drawing the noises from the following distributions:

$$w_k \sim N\left(0, \begin{bmatrix} 0.5 & 0 \\ 0 & 0.2 \end{bmatrix}\right), \qquad v_k \sim N\left(0, \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}\right)$$

The ALS estimation of the covariances Q_w, R_v for a set of data simulated using $W = I$ and using the minimum variance weight (iterative scheme) from the above section is compared. The covariance estimation is repeated 100 times and the results are plotted to check for the variance in the estimates. The diagonal elements of the estimated Q_w, R_v are plotted.

As seen in Figures 1 and 2 using the optimal weight gives estimates having much lower variance than using $W = I$.

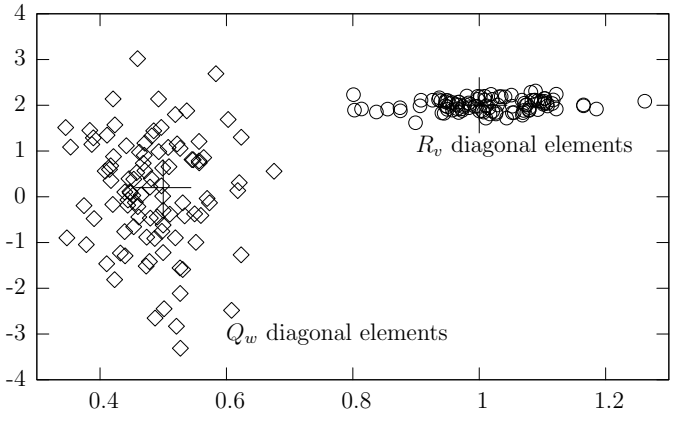

Figure 1: Covariance estimates using $W = I$ in ALS

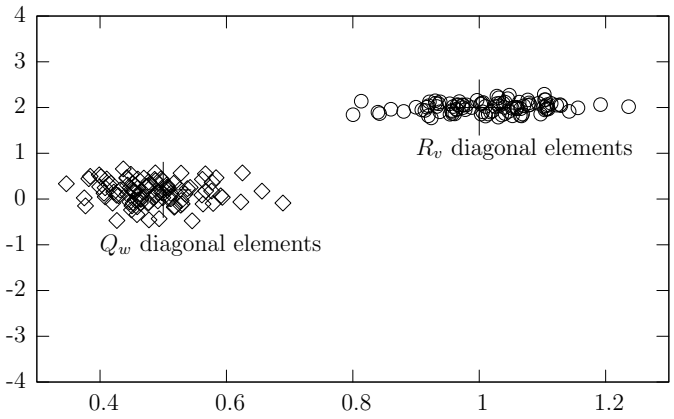

Figure 2: Covariance estimates using a minimum variance weight in ALS

6 ALS-SDP method

In this section the G matrix is also assumed to be unknown in addition to the Q_w, R_v matrices. An estimation technique is presented that estimates the structure of the G matrix modelling the minimum number of independent disturbances affecting the state.

Generally a linear model of a system has many states and only a few independent disturbances corrupting these states. 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} (Equations 1a and 1b). It is unlikely to have information about the G matrix in most applications. Information contained in the measurements is also usually not enough to estimate a full GQ_wG^T matrix uniquely (this can be checked using Theorem 1). If there are fewer sensors than the states, there can be multiple covariances that generate the state noises making up the same output data (Corollary 1).

When G is unknown, our aim is to find the minimum rank Q (where, $Q = GQ_wG^T$). A minimum rank Q can be decomposed as follows:

$$Q = \tilde{G}\tilde{G}^T, \qquad \tilde{Q}_w = I \tag{23}$$

It should be noted that the choice $\tilde{Q}_w = I$ is not a binding choice for the covariance because any other choice of $\tilde{Q}_w$ can be easily absorbed into $\tilde{G}$ by redefining $\tilde{G}_1 = \tilde{G}\sqrt{Q_w^{-1}}$ so that $Q = \tilde{G}\tilde{G}^T = \tilde{G}_1 Q_w \tilde{G}_1^T$.

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 get information about the minimum number of independent disturbances entering the data in addition to the disturbance structure.

Remark 6. *With reference to Equation 23, one might think that a more natural procedure would be to solve the ALS optimization in Equation 18 directly with G as the optimization variable and constraining Q_w instead of solving with Q and then following with 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 and the extra flexibility in allowing for minimization of the rank of Q.*

In the development of the remaining results in this section , we take the weight $W = I$. The more general case is addressed in Remark 7 at the end of this section. The rank can be explicitly added to the objective in Equation 18 through a tradeoff parameter ρ multiplying the rank:

$$\begin{aligned} \Phi_* = &\min_{Q,R_v} \underbrace{\left\| \mathscr{A} \begin{bmatrix} (Q)_s \\ (R_v)_s \end{bmatrix} - \hat{b} \right\|^2}_{\Phi} + \rho \text{Rank } (Q) \\ &Q, R_v \geq 0, \quad Q = Q^T, \quad R_v = R_v^T \end{aligned} \tag{24}$$

The constraints are in the form of convex Linear Matrix Inequalities (LMI) (Boyd et al., 1994; VanAntwerp and Braatz, 2000). The norm part of the objective is also convex. The rank however can only take integer values making the problem NP hard. The solution of minimizing the rank subject to LMI constraints is an open research question and current techniques are largely based on heuristics (Vandenberghe and Boyd, 1996).

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 (Fazel, 2002).

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

The trace of a matrix is a convex function of Q. The optimization in Equation 24 can be rewritten with the trace replacing the rank:

$$\begin{aligned} \Phi_1 = &\min_{Q,R_v} \underbrace{\left\| \mathscr{A} \begin{bmatrix} (Q)_s \\ (R_v)_s \end{bmatrix} - b \right\|^2}_{\Phi} + \rho \text{Tr } (Q) \\ &Q, R_v \geq 0, \quad Q = Q^T, \quad R_v = R_v^T \end{aligned} \tag{25}$$

Lemma 4. *Given an optimization problem in the matrix variable $X \in \mathbb{R}^{n \times n}$ with the following form:*

$$\begin{aligned} &\min_X (AX_s - b)^T (AX_s - b) + \rho Tr\ (X) \\ \textit{subject to} \quad &X \geq 0, \quad X = X^T \end{aligned}$$

with the matrices A and b appropriately dimensioned, the optimization can be rewritten in the following standard primal Semidefinite Programming problem:

$$\begin{aligned} \min_{x} \quad & c^T x \\ \textit{subject to} \quad & F(x) \geq 0 \\ \textit{where} \quad & F(x) \triangleq F_0 + \sum_{i=1}^{m} x_i F_i \end{aligned}$$

with the symmetric matrices $F_0, \cdots, F_m \in \mathbb{R}^{n \times n}$ and the vector $c \in \mathbb{R}^m$ chosen appropriately.

The derivation is given in Appendix C.

Given the above Lemma 4, if we define $X = \text{diag}(Q, R_v)$ then Equation 25 is in the form of a Semidefinite Programming (SDP) problem with A defined accordingly. We refer to this problem as the ALS-SDP (Autocovariance Least-Squares with Semidefinite Programming) in the sequel.

Lemma 5. *If $p < n$ (i.e. number of measurements is fewer than the number of states), then the following holds for Equation 25:*

$$dim[Null(\mathscr{A})] \geq (n-p)(n-p+1)/2$$

Proof. The dimension condition follows by substituting $G = I$ in Lemma 2, noting that $(I_{n^2} - \bar{A} \otimes \bar{A})$ is full rank and using the rank condition in Hua (1990). □

Theorem 4. *A solution $(\hat{Q}, \hat{R}_v)$ to the ALS-SDP in Equation 25 is unique if $dim[Null(M)] = 0$ where,*

$$M = (C \otimes I_n)(I_{n^2} - \bar{A} \otimes \bar{A})^{-1}(G \otimes G)\mathscr{D}_g$$

and G is any full column rank decomposition of $\hat{Q} = GG^T$ (G is an unique decomposition within an orthogonal matrix multiplication).

Proof. The function:

$$\Phi = \left\| \mathscr{A} \begin{bmatrix} (G \otimes G)(Q_w)_s \\ (R_v)_s \end{bmatrix} - b \right\|^2$$

is the first part of the objective in Equation 25 and also the same as the objective in Equation 20. Following Theorem 1 and Lemma 1, $\dim[\text{Null}(M)] = 0$ then implies that Φ is strictly convex at the solution $Q_w = I_g, R_v = \hat{R}_v$.

The other part of the objective in Equation 25 i.e. Tr (Q) is linear in the variable Q and hence is also convex. The overall objective in Equation 25 is then strictly convex at the solution $\hat{Q}, \hat{R}_v$ when $\dim[\text{Null}(M)] = 0$. Uniqueness of the solution thus follows from minimization of a strictly convex objective subject to convex constraints (Boyd and Vandenberghe, 2004). □

The ALS-SDP method gives a feasible solution for each value of the tradeoff parameter ρ by using simple Newton-like algorithms. The choice of ρ is made from a tradeoff plot of Tr (Q) versus Φ from Equation 25. A good value of ρ is such that Tr (Q) is small and any further decrease in value of Tr (Q) causes significant increase in the value of Φ. This ensures that the rank of Q is minimized without significant compromise on the original objective Φ (Rajamani and Rawlings, 2006).

The matrix inequalities $Q_w \geq 0, R_v \geq 0$ can be handled by an optimization algorithm adding a logarithmic barrier function to the objective. The optimization algorithm then minimizes:

$$\Phi_1 = \min_{Q_w, R_v} \Phi + \rho \text{Tr } (Q) - \mu \log \begin{vmatrix} Q & 0 \\ 0 & R_v \end{vmatrix} \tag{26}$$

in which, μ is the *barrier parameter* and $|\cdot|$ denotes the determinant of the matrix (Nocedal and Wright, 1999). The log-determinant barrier is an attractive choice because it has analytical first and second derivatives. Appendix D lists some useful matrix derivatives arising in the optimization in Equation 26. As with other barrier techniques, with $\mu \to 0$, the solution to the SDP tends to the optimum solution. The following approach was used to solve the barrier augmented SDP.

1. Choose a value for the tradeoff parameter ρ
2. Iteration k=0
3. Choose a starting value of μ (say $\mu = 100$)
4. Solve the SDP and let the solution be Q_k, R_k
5. Decrease value of μ (say choose the new value as $\mu/2$)
6. Increase value of k by 1 and repeat step 4 till $\mu < 10^{-7}$
7. Check conditions in Theorem 4 for uniqueness. If the solution is not unique then repeat with higher value for ρ.

Other path following type of algorithms can be found in Wolkowicz, Saigal, and Vandenberghe (2000, chap. 10). The convexity of Equation 26 ensures a unique termination of the minimization algorithm. The algorithm scales efficiently for large dimensional problems.

Remark 7. *The inclusion of the weight W as derived in Section 5 would give an estimate of Q which has minimum variance among all constrained unbiased linear estimators. However since W itself is a highly nonlinear function of the unknown covariances, the whole purpose of the convex optimization in Equation 25 is defeated if W is included as a part of the optimization. Apart from that, the computational challenge of calculating W justifies its exclusion from the objective.*

6.1 Example

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

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

with noises drawn from the distributions:

$$w_k \sim N(0, 0.5), \qquad v_k \sim N(0, 1)$$

Although the data is generated by a single column G matrix, we assume G is unknown and estimate it using the ALS-SDP procedure.

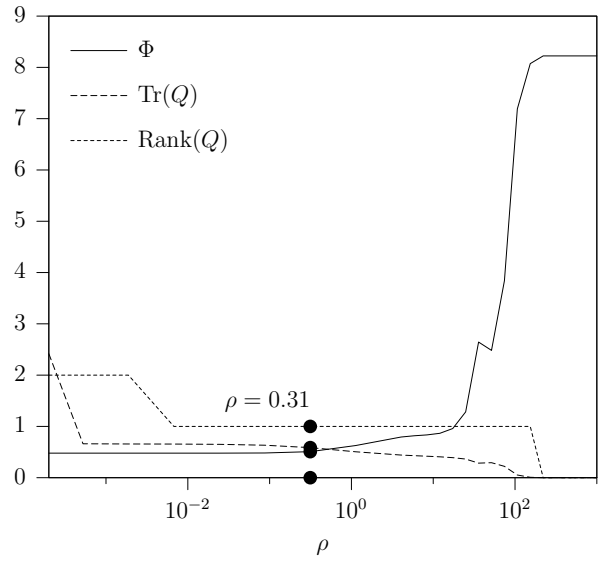

Figure 3: Values of competing parts of the objective function in Equation 25 for different values of ρ and the rank of Q

The results from the new ALS-SDP are shown in Figures 3 and 4. 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 $\rho = 0.31$ is 1, which is the number of independent disturbances entering the state in the simulated data (columns of G).

Also the estimated disturbance structure and covariances using $\rho = 0.31$ is:

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

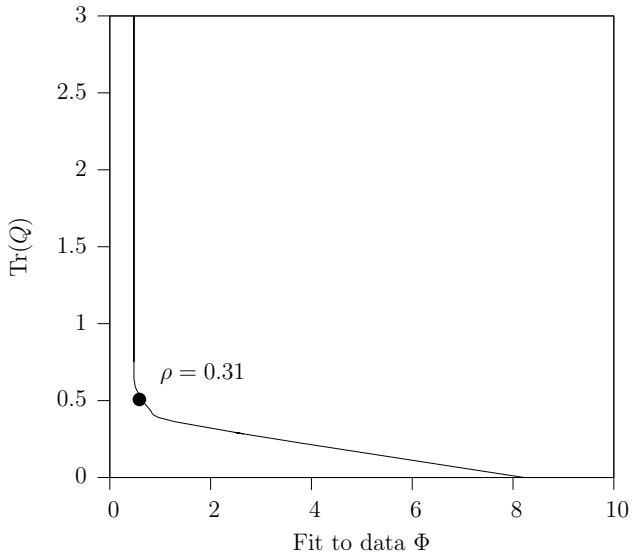

Figure 4: Tradeoff plot between Φ and $Tr(Q)$ from Equation 25 to choose the tradeoff parameter ρ

After decomposition according to Equation 23 we get, $\hat{G} = [0.670, 0.372]^T, \hat{Q}_w = 1$. Again if $\hat{Q}_w$ were chosen to be 0.5, then the decomposition of $\hat{Q}$ gives $\hat{G} = [0.95, 0.52]^T$, which is close to the actual G simulating the data.

The estimated positive semidefinite $\hat{Q}$ and a positive definite $\hat{R}_v$ can then be used to tune any state estimator chosen by the user. With the above estimated covariances for $\rho = 0.31$, the Kalman filter tuning $\hat{L}$ is compared with the optimal L_o:

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

7 Conclusions

Given a set of system matrices A, C, a known noise shaping matrix G and an initial arbitrary stable filter gain L, uniqueness of the estimates of Q_w and R_v using the ALS technique can be checked using the simple conditions in Theorem 1. The computational burden in checking these conditions is minimal even for large dimension systems. Estimates of the noise covariances from data are minimum variance only if the least-squares is weighted with the optimal weight. This weight was shown to depend on the fourth moment of data and a formula was derived (Theorem 3). An example was presented to show the reduced variance in the covariance estimates when using the minimum variance weight. The complicated nature of the formulae

do not make them practical using current computational techniques. The weight however puts to rest the issue of the existence of the best linear unbiased estimator for the covariances. One of the major uncertainties in the process industries is the disturbance structure affecting the significant variables of the plant. For linear models we showed that the disturbance structure is captured accurately by the matrix G in Equation 1a, which shapes the noises entering the states. Estimation of the minimum number of disturbances affecting the states is equivalent to minimizing the rank of G. An estimation procedure using SDP and a rank heuristic was shown to give a tradeoff between fit to 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 and the disturbance structure.

8 Acknowledgments

The authors thank Professor Stephen P. Boyd for helpful discussions. The authors acknowledge financial support from NSF through grant #CNS–0540147 and PRF through grant #43321–AC9.

References

A. C. Aitken. On least squares and linear combinations of observations. *Proc. R. Soc. Edingburgh*, 55:42–48, 1935.

B. M. Åkesson, J. B. Jørgensen, N. K. Poulsen, and S. B. Jørgensen. A generalized autocovariance least-squares method for Kalman filter tuning. Submitted for publication in *Journal of Process Control*, June 2007.

B. D. O. Anderson and J. B. Moore. *Optimal Filtering*. Prentice-Hall, Englewood Cliffs, N.J., 1979.

T. W. Anderson. *An Introduction to Multivariate Statistical Analysis*. John Wiley & Sons, New York, third edition, 2003.

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

S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

S. Boyd, L. E. Ghaoui, E. Feron, and V. Balakrishnan. *Linear Matrix Inequalities in Control Theory*. SIAM, Philadelphia, 1994.

B. Carew and P. Bélanger. Identification of optimum filter steady-state gain for systems with unknown noise covariances. *IEEE Trans. Auto. Cont.*, 18(6):582–587, 1973.

D. Dee, S. Cohn, A. Dalcher, and M. Ghil. An efficient algorithm for estimating noise covariances in distributed systems. *IEEE Trans. Auto. Cont.*, 30(11):1057–1065, 1985.

M. Fazel. *Matrix Rank Minimization with Applications*. PhD thesis, Dept. of Elec. Eng., Stanford University, 2002.

M. Gevers. A personal view of the development of system identification. *IEEE Control Systems Magazine*, 26(6):93–105, December 2006.

G. A. Ghazal and H. Neudecker. On second-order and fourth-order moments of jointly distributed random matrices: a survey. *Linear Algebra Appl.*, 321:61–93, 2000.

G. C. Goodwin and K. S. Sin. *Adaptive Filtering Prediction and Control*. Prentice-Hall, Englewood Cliffs, New Jersey, 1984.

A. Graham. *Kronecker products and matrix calculus with applications*. Ellis Horwood Limited, West Sussex, England, 1981.

D. Hua. On the symmetric solutions of linear matrix equations. *Linear Algebra Appl.*, 131:1–7, 1990.

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.

G. Jenkins and D. Watts. *Spectral Analysis and its Applications*. Holden-Day, 500 Sansome Street, San Fransisco, California, 1968.

J. Juang and M. Phan. Identification of system, observer, and controller from closed-loop experimental data. *J. Guid. Control Dynam.*, 17(1):91–96, January-February 1994.

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.

J. R. Magnus and H. Neudecker. The commutation matrix: Some properties and applications. *Ann. Stat.*, 7(2):381–394, March 1979.

J. R. Magnus and H. Neudecker. *Matrix Differential Calculus with Applications in Statistics and Econometrics*. John Wiley, New York, 1999.

R. Mehra. On the identification of variances and adaptive Kalman filtering. *IEEE Trans. Auto. Cont.*, 15(12):175–184, 1970.

R. Mehra. On-line identification of linear dynamic system with application to Kalman filtering. *IEEE Trans. Auto. Cont.*, 16(1):12–21, 1971.

R. Mehra. Approaches to adaptive filtering. *IEEE Trans. Auto. Cont.*, 17:903–908, 1972.

C. Neethling and P. Young. Comments on "Identification of optimum filter steady-state gain for systems with unknown noise covariances". *IEEE Trans. Auto. Cont.*, 19(5):623–625, 1974.

J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer-Verlag, New York, 1999.

G. Noriega and S. Pasupathy. Adaptive estimation of noise covariance matrices in real-time preprocessing of geophysical data. *IEEE Trans. Geosci. Remote Sensing*, 35(5):1146–1159, 1997.

B. J. Odelson, M. R. Rajamani, and J. B. Rawlings. A new autocovariance least-squares method for estimating noise covariances. *Automatica*, 42(2):303–308, February 2006. URL `http://www.elsevier.com/locate/automatica`.

S. J. Qin, W. Lin, and L. Ljung. A novel subspace identification approach with enforced causal models. *Automatica*, 41(12):2043–2053, December 2005.

M. R. Rajamani and J. B. Rawlings. Estimation of noise covariances and disturbance structure from data using least squares with optimal weighting. In *Proceedings of AIChE Annual Meeting*, San Francisco, California, November 2006.

M. R. Rajamani, J. B. Rawlings, S. J. Qin, and J. J. Downs. Equivalence of MPC disturbance models identified from data. In *Proceedings of Chemical Process Control—CPC 7*, Lake Louise, Alberta, Canada, January 2006.

C. V. Rao, J. B. Rawlings, and J. H. Lee. Constrained linear state estimation – a moving horizon approach. *Automatica*, 37(10):1619–1628, 2001.

R. H. Shumway and D. S. Stoffer. An approach to time series smoothing and forecasting using the em algorithm. *J. Time Series Anal.*, 3:253–264, 1982.

W. Steeb. *Kronecker product of matrices and applications*. Mannheim, 1991.

D. S. Stoffer and K. Wall. Bootstrapping state space models: Gaussian maximum likelihood estimation and the Kalman filter. *J. Am. Stat. Assoc.*, 86:1024–1033, 1991.

C. F. Van Loan. The ubiquitous Kronecker product. *J. Comput. Appl. Math.*, 123: 85–100, 2000.

P. Van Overschee and B. De Moor. N4SID: subspace algorithms for the identification of combined deterministic-stochastic systems. *Automatica*, 30(1):75–93, 1994.

P. Van Overschee and B. De Moor. A unifying theorem for three subspace system identification algorithms. *Automatica*, 31(12):1853–1864, December 1995.

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

L. Vandenberghe and S. Boyd. Semidefinite programming. *SIAM Rev.*, 38(1):49–95, March 1996.

M. Viberg. Subspace methods for the identification of linear time-invariant systems. *Automatica*, 31(12):1835–1851, 1995.

H. Wolkowicz, R. Saigal, and L. Vandenberghe, editors. *Handbook of Semidefinite Programming: Theory, Algorithms and Applications*. Kluwer Academic Publishers, Massachusetts, 2000.

W. Zhuang, M. R. Rajamani, J. B. Rawlings, and J. Stoustrup. Application of auto-covariance least-squares method for model predictive control of hybrid ventilation in livestock stable. In *Proceedings of the American Control Conference*, New York, USA, July 11-13 2007a.

W. Zhuang, M. R. Rajamani, J. B. Rawlings, and J. Stoustrup. Model predictive control of thermal comfort and indoor air quality in livestock stable. In *Proceedings of the European Control Conference*, Kos, Greece, July 2-5 2007b.

A Proof of Lemma 3

Let $Y_p = [\mathscr{Y}_1^T \cdots \mathscr{Y}_{N_d}^T]^T$. Then we have,

$$\begin{bmatrix} \mathscr{Y}_1 \\ \vdots \\ \mathscr{Y}_N \\ \hline \mathscr{Y}_2 \\ \vdots \\ \mathscr{Y}_{N+1} \\ \hline \vdots \\ \hline \mathscr{Y}_{N_d-N+1} \\ \vdots \\ \mathscr{Y}_{N_d} \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & \cdots & 0 & \cdots & & \\ \vdots & \ddots & \vdots & \vdots & \cdots & \cdots \\ 0 & \cdots & 1 & \cdots & & \\ \vdots & 1 & \cdots & 0 & & \\ 0 & \vdots & \ddots & \vdots & \cdots & \cdots \\ \vdots & 0 & \cdots & 1 & \ddots & \\ & & \cdots & \cdots & \cdots & \ddots \\ & & \cdots & \cdots & \cdots & I_N \end{bmatrix}}_{\mathbb{E}_1} \begin{bmatrix} \mathscr{Y}_1 \\ \vdots \\ \mathscr{Y}_N \\ \hline \mathscr{Y}_{N+1} \\ \vdots \\ \mathscr{Y}_{N_d} \end{bmatrix} \tag{27}$$

Thus, $\mathbb{Y}_s = \mathbb{E}_1 Y_p$. Now we look at the distribution of Y_p.
From Equation 4, we have

$$\varepsilon_k = \bar{A}^k \varepsilon_0 + \sum_{j=0}^{k-1} \bar{A}^{k-j-1} \bar{G} \begin{bmatrix} w_j \\ v_j \end{bmatrix} \tag{28}$$

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

$$E[Y_p] = E \begin{bmatrix} \mathscr{Y}_1 \\ \vdots \\ \mathscr{Y}_{N_d} \end{bmatrix} = 0$$

Since $E(Y_p) = 0$, the covariance of Y_p is also its the second moment.
Now, calculate Ω_p the second moment of Y_p as follows:

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

Using Equations 6, 7 and 8, we get:

$$\begin{aligned} \Omega_p = & \underbrace{\begin{bmatrix} C \\ C\bar{A} \\ \vdots \\ C\bar{A}^{N_d-1} \end{bmatrix}}_{\mathcal{O}} P\mathcal{O}^T + \begin{bmatrix} R_v & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & R_v \end{bmatrix} + \Psi \begin{bmatrix} R_v & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & R_v \end{bmatrix} \\ & + \underbrace{\begin{bmatrix} 0 & 0 & 0 & 0 \\ C\bar{G} & 0 & 0 & 0 \\ \vdots & \ddots & & \vdots \\ C\bar{A}^{N_d-2}\bar{G} & \cdots & C\bar{G} & 0 \end{bmatrix}}_{\Gamma_f} \begin{bmatrix} \bar{Q}_w & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \bar{Q}_w \end{bmatrix} \Gamma_f^T + \begin{bmatrix} R_v & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & R_v \end{bmatrix} \Psi^T \end{aligned} \tag{29}$$

where,

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

Following Equation 28, 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:

$$Y_p \sim N(0, \Omega_p)$$

Next we use Equation 27 and the above result to get the distribution of $\mathbb{Y}$. Since $\mathbb{Y}$ is a matrix, its mean and covariance are defined for the stacked version of the matrix i.e. $\mathbb{Y}_s$. Given the linear relationship between Y_p and $\mathbb{Y}_s$, we get,

$$\mathbb{Y}_s \sim N(0, \mathbb{E}_1 \Omega_p \mathbb{E}_1^T)$$

Thus, the covariance of $\mathbb{Y}$ is:

$$\Omega = \mathbb{E}_1 \Omega_p \mathbb{E}_1^T \tag{30}$$

where Ω_p is given by Equation 29 and $\mathbb{E}_1$ is defined in Equation 27.

□

B Proof of Lemma 2

Let $[q_N, r_N]^T$ be an element in the null space of $\tilde{\mathscr{A}}$ in Equation 20, where the dimensions are $q_N \in \mathbb{R}^{\frac{g(g+1)}{2}\times 1}$ and $r_N \in \mathbb{R}^{\frac{p(p+1)}{2}\times 1}$. This implies:

$$\tilde{\mathscr{A}} \begin{bmatrix} q_N \\ r_N \end{bmatrix} = 0 \qquad \text{or,}$$
$$\begin{bmatrix} \mathscr{A}_1(G\otimes G)\mathscr{D}_g & \mathscr{A}_2 \end{bmatrix} \begin{bmatrix} q_N \\ r_N \end{bmatrix} = 0$$

where, $\mathscr{A}_1$ and $\mathscr{A}_2$ from Equation 19 are:

$$\begin{aligned} \mathscr{A}_1 =& (C\otimes \mathscr{O})A^\dagger \\ \mathscr{A}_2 =& [(C\otimes \mathscr{O})A^\dagger(AL\otimes AL) + (I_p \otimes \Gamma)]\mathscr{D}_p \quad \text{and} \\ A^\dagger =& (I_{n^2} - \bar{A}\otimes\bar{A})^{-1} \end{aligned}$$

We then have:

$$\begin{aligned} (C\otimes\mathscr{O})A^\dagger(G\otimes G)\mathscr{D}_g q_N + \\ [(C\otimes\mathscr{O})A^\dagger(AL\otimes AL) + (I_p\otimes\Gamma)]\mathscr{D}_p r_N = 0 \end{aligned} \tag{31}$$

We can rewrite $\mathscr{O}$ and Γ as:

$$\mathscr{O} = \begin{bmatrix} C \\ \mathscr{O}_1\bar{A} \end{bmatrix}, \qquad \Gamma = \begin{bmatrix} I_p \\ \mathscr{O}_1(-AL) \end{bmatrix}$$

where,

$$\mathscr{O}_1 = \begin{bmatrix} C \\ C\bar{A} \\ \vdots \\ C\bar{A}^{N-2} \end{bmatrix}$$

If (A, C) is observable (Assumption 1), then $\mathscr{O}_1$ has full column rank for $N \geq (n+1)$.

Partitioning $\mathscr{O}$ and Γ as above, we can write Equation 31 as the following Equations:

$$\begin{aligned} (C\otimes C)A^\dagger[(G\otimes G)\mathscr{D}_g q_N + (AL\otimes AL)\mathscr{D}_p r_N] \\ +\mathscr{D}_p r_N = 0 \end{aligned} \tag{32a}$$

$$\begin{aligned} (C\otimes\mathscr{O}_1\bar{A})A^\dagger[(G\otimes G)\mathscr{D}_g q_N + (AL\otimes AL)\mathscr{D}_p r_N] \\ +(I_p\otimes\mathscr{O}_1(-AL))\mathscr{D}_p r_N = 0 \end{aligned} \tag{32b}$$

By expanding $\bar{A} = A - ALC$ and using Equation 32a, Equation 32b simplifies to:

$$(I_p \otimes \mathscr{O}_1 A)(C \otimes I_n)A^\dagger[(G \otimes G)\mathscr{D}_g q_N \\ +(AL \otimes AL)\mathscr{D}_p r_N] = 0$$

Since $\mathscr{O}_1$ is full column rank (Assumption 1) and A is non singular (Assumption 3), $(I_p \otimes \mathscr{O}_1 A)$ is also full column rank. This implies:

$$(C \otimes I_n)A^\dagger[(G \otimes G)\mathscr{D}_g q_N + (AL \otimes AL)\mathscr{D}_p r_N] = 0 \tag{33}$$

Substituting Equation 33 in 32a and noting that $(C \otimes C)$ can be written as $(I_n \otimes C)(C \otimes I_n)$, we get:

$$\mathscr{D}_p r_N = 0$$

Equation 33 then simplifies to:

$$(C \otimes I_n)A^\dagger(G \otimes G)\mathscr{D}_g q_N = 0$$

Thus, q_N is an element in the null space of $M = (C \otimes I_n)(I_{n^2} - \bar{A} \otimes \bar{A})^{-1}(G \otimes G)\mathscr{D}_g$ and $r_N = 0$.

Proving the second part of the lemma is straightforward by starting with Equation 33 and multiplying with $(I_n \otimes \mathscr{O})$, which is full column rank. □

C Proof of Lemma 4

Since the X matrix is constrained to be symmetric, we only need to consider the symmetric $p = \frac{n(n+1)}{2}$ elements of X. Let these symmetric elements of X be stacked in the vector $z \in \mathbb{R}^p$.

The original optimization in Lemma 4 can then be written as:

$$\min_z \quad (\tilde{A}z - b)^T(\tilde{A}z - b) + d^T z$$
$$\text{subject to} \quad \sum_{i=1}^{i=p} z_i B_i \geq 0$$

where, $\tilde{A}$ is the A modified to operate only on the symmetric elements of X, $B_i \in \mathbb{R}^{n \times n}, i = 1, \cdots, p$ are the basis matrices for a symmetric $\in \mathbb{R}^{n \times n}$ matrix and d is chosen appropriately such that $\rho \text{Tr}\,(X) = d^T z$.

Using a standard trick for converting quadratic objective functions into Linear Matrix Inequalities (see for example Vandenberghe and Boyd (1996)), we get:

$$\begin{aligned} \min_{z,t} \quad & t + d^T z \\ \text{subject to} \quad & \sum_{i=1}^{i=p} z_i B_i \geq 0 \\ & \begin{bmatrix} t & (\tilde{A}z - b)^T \\ (\tilde{A}z - b) & I \end{bmatrix} \geq 0 \end{aligned}$$

Define $m = p + 1$, $x = [z^T, t]^T \in \mathbb{R}^m$, $c = [d^T, 1]^T \in \mathbb{R}^m$ and the matrices

$$F_0 = \begin{bmatrix} 0 & -b^T & \\ -b & I & \\ & & 0 \end{bmatrix}, \quad F_m = \begin{bmatrix} 1 & 0 & \\ 0 & 0 & \\ & & 0 \end{bmatrix}, \quad F_i = \begin{bmatrix} 0 & a_i^T & \\ a_i & 0 & \\ & & B_i \end{bmatrix}$$
$$i = 1, \cdots, p$$

here, the 0 represents zero matrices with appropriate dimensions. We then get the final form of the optimization as a standard primal Semidefinite Programming(SDP) problem:

$$\begin{aligned} \min_{x} \quad & c^T x \\ \text{subject to} \quad & F(x) \geq 0 \\ \text{where} \quad & F(x) \triangleq F_0 + \sum_{i=1}^{m} x_i F_i \end{aligned}$$

□

D Some Useful Derivatives of Matrix Functions

The results below follow from Magnus and Neudecker (1999); Graham (1981).

Given $Q \in \mathbb{R}^{n \times n}$ is a symmetric matrix and $A \in \mathbb{R}^{p \times \frac{n(n+1)}{2}}$ and $b \in \mathbb{R}^p$ are some arbitrary constant matrices with $p \geq n$.

$$F = (AQ_s - b)^T (AQ_s - b) + \rho \text{Tr}\ (Q) - \mu \log|Q|$$

The first and second derivatives for the above function with respect to the matrix Q

are given by:

$$\left\{\left[\frac{\partial F}{\partial Q}\right]_{Q_k}\right\}_s = 2A^T A(Q_k)_s - 2A^T b + \rho(I_n)_s - \mu(Q_k^{-1})_s \tag{34}$$

$$\left\{\left[\frac{\partial^2 F}{\partial Q^2}\right]_{Q_k}\right\}_s = 2A^T A - \mu(Q_k^{-1} \otimes Q_k^{-1}) \tag{35}$$