TWCCC ⋆ Texas – Wisconsin – California Control Consortium

Technical report number 2014-02

Identification of Disturbance Covariances Using Maximum Likelihood Estimation [*]

Megan A. Zagrobelny[†]and James B. Rawlings[‡]
Department of Chemical and Biological Engineering
University of Wisconsin-Madison

December 15, 2014

Abstract

Disturbance model identification is necessary both for estimator design and controller performance monitoring. Here we present a maximum likelihood estimation (MLE) method to identify process and measurement noise covariances. By writing the outputs in terms of the process and measurement noises, we form a normal distribution for the sequence of measurements. The variance of this distribution is a function of the unknown noise covariances, and the likelihood is optimized with respect to these covariances. We show that a solution to this problem exists. By comparing the first order conditions to those of the autocovariance-least squares (ALS) method, we derive necessary conditions for uniqueness. Several numerical methods, including utilizing the sparsity of the solution, are presented and demonstrated to decrease the computational time for the problem. Simulations are used to compare the MLE method to several existing methods: the ALS method, an alternate MLE method based on the innovations, and an expectation maximization method. Although the solving the MLE problem is considerably slower than solving the ALS problem, the MLE solution is shown to maximize the likelihood compared to the ALS problem.

[*] The authors gratefully acknowledge the financial support of the industrial members of the Texas-Wisconsin Modeling and Control Consortium and NSF through grant #CTS-1159088

[†] zagrobelny@wisc.edu

[‡] rawlings@engr.wisc.edu

1 Background

Knowledge of the disturbances affecting the system is a key aspect both for controller performance monitoring and for estimator design. Methods for identifying noise covariances are often divided into four categories: Bayesian estimation, maximum likelihood estimation (MLE), covariance matching, and correlation techniques, such as the autocovariance least-squares (ALS) method [12]. Subspace ID techniques, which are primarily designed for system identification, also provide a noise model, but they identify the optimal Kalman filter gain and innovation variance rather than the process and measurement noise [13].

Several early maximum likelihood methods also focus on finding filter parameters and then estimate the process and measurement noise covariances from these results [10, 4, 11]. More stringent conditions must be satisfied to have a unique estimate for the covariances, compared to estimating only the filter parameters [4, 11].

More recently, [3] and [5] use maximum likelihood or Bayesian estimation to estimate parameters in a grey-box model. The general grey-box model has a known structure but some parameters are unknown, which may include the noise covariances. In [9], a Bayesian method is presented to estimate the covariances when the deterministic system parameters are completely known. First a grid of possible covariances is created; then, at each point in the grid, state estimation is performed and the likelihood and posterior probability are calculated.

Since direct maximum likelihood methods require solving a nonlinear optimization problem, [16] proposed an iterative method using the expectation maximization (EM) technique. In the EM method, an initial guess of the unknown parameters are chosen, and the states are estimated from these unknown parameters via the Kalman smoother. Then the unknown parameters are found by maximum likelihood estimation assuming that the states are equal to the smoother estimates. Since the states are known, this maximization step simplifies to simple algebraic equations. This process of estimating the states using the Kalman smoother and optimizing the parameters using MLE is repeated until the parameters converge, which is guaranteed for the EM method.

[2] developed both a direct maximum likelihood method and an EM method for nonlinear systems, based on the extended Kalman filter. The direct maximum likelihood method is written in terms of the innovations, which are calculated at each iteration of the optimizer. Since the innovations are white under the optimal estimator, the likelihood for the entire data set is written as a product of the likelihoods for each innovation. This MLE method assumes that the deterministic system parameters are known. Like the direct MLE method, the EM method also only estimates Q_w and R_v, whereas [16] estimated the state transition matrix, A, as well as the noise covariances. Both the maximum likelihood and expectation maximization methods accurately identified Q_w and R_v in simulation; the methods also led to improved estimation for laboratory data. They applied both methods to systems with measurements sampled at multiple rates. [6] applied the EM method of [2] to linear systems and expanded this method to cases in which the noise-shaping matrix G is known. Several examples demonstrated that this method reduces the variance as compared to the ALS method.

Here we directly formulate the maximum likelihood problem as estimation of parameters affecting the variance of a normally distributed signal. Estimating the covariance of a normal distribution is discussed in detail in [1] when the entire covariance is unknown. [7] derived first and second order conditions for the maximum likelihood estimator of the mean and covariance for a normal distribution, where the covariance is a function of a finite number of parameters. We simplify the maximum likelihood problem by assuming an LTI system with known deterministic system matrices and unknown noise covariances.

The paper is organized as follows. First, we develop the maximum likelihood problem starting from the state space model. Then we discuss existence of the solution. We compare this method to the existing ALS technique and give necessary conditions for the solution to be unique. Finally, we give recommendations to improve the numerical optimization and illustrate the method on three examples.

2 Setting up the problem

We begin with the state space model

$$\begin{aligned} x^+ &= Ax + w \\ y &= Cx + v \\ \begin{bmatrix} w \\ v \end{bmatrix} &\sim N\left(0, \begin{bmatrix} Q_w & 0 \\ 0 & R_v \end{bmatrix}\right) \end{aligned}$$

in which $x, w \in \mathbb{R}^n$, $y, v \in \mathbb{R}^p$, and w and v are uncorrelated in time. We seek maximum likelihood estimates of the unknown covariance matrices Q_w and R_v given the system matrices A and C and a sequence of measurements $y(0), \dots, y(N-1)$:

$$\begin{gathered} \max_{Q_w, R_v} \ln p_y(y(0) \dots y(N-1) | Q_w, R_v) \\ \text{subject to } Q_w, R_v \geq 0 \end{gathered}$$

To derive an expression for the likelihood, we write all the measurements in a single vector and relate them to the noises entering the system and an initial state $x(0)$:

$$\begin{aligned} \begin{bmatrix} y(0) \\ y(1) \\ \vdots \\ y(\tilde{N}-1) \end{bmatrix} &= \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{\tilde{N}-1} \end{bmatrix} x(0) + \begin{bmatrix} v(0) \\ v(1) \\ \vdots \\ v(\tilde{N}-1) \end{bmatrix} \\ &+ \begin{bmatrix} 0 & 0 & \dots & 0 \\ C & 0 & \dots & 0 \\ \dots & \ddots & & \vdots \\ CA^{\tilde{N}-2} & CA^{\tilde{N}-3} & \dots & C \end{bmatrix} \begin{bmatrix} w(0) \\ w(1) \\ \vdots \\ w(\tilde{N}-2) \end{bmatrix} \end{aligned} \tag{1}$$

in which $\tilde{N} = N + K$.

For simplicity of presentation, we assume that A is stable[1] and choose K such that $|A^i| \leq \delta,\ \forall i \geq K$ for some small scalar threshold $\delta > 0$. Then all the measurements $y(K+i)$ (for $i > 0$) are approximately independent of the initial state, as well as many of the past noises. Considering only the measurements at time K or later, (1) simplifies to

$$\begin{bmatrix} y(K) \\ y(K+1) \\ \vdots \\ y(\tilde{N}-1) \end{bmatrix} \approx \begin{bmatrix} v(K) \\ v(K+1) \\ \vdots \\ v(\tilde{N}-1) \end{bmatrix} + \underbrace{\begin{bmatrix} CA^{K-1} & \dots & C & 0 & \dots & 0 \\ 0 & CA^{K-1} & \dots & C & \dots & 0 \\ \vdots & \ddots & & & & \vdots \\ 0 & \dots & 0 & CA^{K-1} & \dots & C \end{bmatrix}}_{\mathbb{O}} \begin{bmatrix} w(0) \\ w(1) \\ \vdots \\ w(\tilde{N}-2) \end{bmatrix} \tag{2}$$

Since all of the noises are normally distributed, $\begin{bmatrix} y(K)' & \dots & y(\tilde{N}-1)' \end{bmatrix}'$ also has a normal distribution. As the indices in (2) are arbitrary, we have the distribution

$$Y := \begin{bmatrix} y(0) \\ y(1) \\ \vdots \\ y(N-1) \end{bmatrix} \sim N(0, P)$$

$$P = \mathbb{O} \begin{bmatrix} Q_w & & \\ & \ddots & \\ & & Q_w \end{bmatrix} \mathbb{O}' + \begin{bmatrix} R_v & & \\ & \ddots & \\ & & R_v \end{bmatrix} \tag{3}$$

Note that we can also write P as

$$P = \sum_{i=1}^{N+K-1} \mathbb{O}_i Q_w \mathbb{O}_i' + \sum_{j=1}^{N} \mathbb{I}_j R_v \mathbb{I}_j' \tag{4}$$

in which $\mathbb{O}_i$ is the i^{th} $pN \times n$ block column of $\mathbb{O}$ and $\mathbb{I}_i$ is the i^{th} $pN \times p$ block column of I_{Np}. Finally, we write the maximum likelihood problem as

$$\min_{Q_w, R_v} \phi(Q_w, R_v) = \ln \det P + Y' P^{-1} Y$$

$$\text{subject to } Q_w, R_v \geq 0 \tag{5}$$

in which P is defined in (3) and (4). Note that $\phi(Q_w, R_v)$ is equal to $-2 \ln p_Y(Y|Q_w, R_v)$ without the constant term.

[1] In the case that A is unstable, the MLE problem for an observable system can be formulated by choosing a stable estimator with gain L and posing the problem in terms of the matrix $A - ALC$ and the L-innovations, $y(k) - C\hat{x}(k|k)$, rather than the outputs. The covariance of the vector of L-innovations is slightly more complicated than that of the outputs as the past measurement noises affect the current innovation, but the MLE problem is analogous to the problem presented here.

3 Existence

We next consider under what conditions a solution to the maximum likelihood optimization problem in (5) exists. To better motivate the results that follow, we first consider a more standard case in which we have N independent samples of a normally distributed variable with an unknown covariance. In the following two propositions, we show that the maximum likelihood estimate for this covariance exists with probability one.

Proposition 1. *Let $R \in \mathbb{R}^{p\times p}$ be positive definite and matrix $\mathbb{Y} \in \mathbb{R}^{p\times N}$ have rank p with its column partitioning denoted by*

$$\mathbb{Y} = \begin{bmatrix} y_1 & y_2 & \cdots & y_N \end{bmatrix}$$

with $y_i \in \mathbb{R}^p$ and $N \geq p$. Define $f(R)$ as

$$f(R) := N \ln\det R + \sum_{i=1}^{N} y_i' R^{-1} y_i$$

Then $f(R) \to \infty$ if either $\lambda_i(R) \to 0^+$ for any eigenvalue or $R \to \infty$.

Proof. Since R is positive definite, it has eigenvalue decomposition $R = W\Lambda W'$ in which $W \in \mathbb{R}^{p\times p}$ is orthogonal and $\Lambda \in \mathbb{R}^{p\times p}$ is diagonal with positive diagonal elements, $\lambda_i > 0$, $i = 1, 2, \ldots p$. Evaluating f gives

$$f(R) = N\sum_{j=1}^{p} \ln(\lambda_i) + \sum_{i=1}^{N} y_i' W\Lambda^{-1}W' y_i$$

Partitioning W by its columns, $W = \begin{bmatrix} w_1 & w_2 & \cdots & w_p \end{bmatrix}$, we express the second term as

$$\begin{aligned}
\sum_{i=1}^{N} y_i' W\Lambda^{-1}W' y_i &= \sum_{i=1}^{N} y_i' \Bigg(\sum_{j=1}^{p} \frac{1}{\lambda_j} w_j w_j' \Bigg) y_i = \sum_j \frac{1}{\lambda_j} \sum_i y_i' w_j w_j' y_i \\
&= \sum_j \frac{1}{\lambda_j} \sum_i w_j' y_i y_i' w_j = \sum_j \frac{1}{\lambda_j} w_j' \mathbb{Y}\mathbb{Y}' w_j \\
&= \sum_{j=1}^{p} \frac{1}{\lambda_j} r_j' r_j
\end{aligned}$$

in which $r_j := \mathbb{Y}' w_j$. Since $\mathbb{Y}$ has full row rank and $w_j \neq 0$ for $j = 1, 2, \ldots, p$, we must have $r_j \neq 0$. Therefore, $a_j^2 := r_j' r_j$ are positive scalars for $j = 0, 1, \ldots, p$. Substituting this result into f gives

$$f(R) = \sum_{j=1}^{p} b_j \qquad\qquad b_j := N\ln(\lambda_j) + \frac{a_j^2}{\lambda_j}$$

With this decomposition of $f(R)$, we consider its behavior as $\lambda_i(R) \to 0^+$ and $R \to \infty$:

1. $\lambda_i(R) \to 0^+$. Note that for any $a_j^2 > 0$, $\lim_{\lambda_j \to 0^+} \ln(\lambda_j) + a_j^2/\lambda_j \to \infty$, *i.e.*, $1/\lambda_j$ goes to ∞ faster than $\ln \lambda_j$ goes to $-\infty$. Therefore, as any $\lambda_j \to 0^+$, $b_j \to \infty$. For the eigenvalues that remain positive, b_j has a finite value. Therefore we conclude that $\lim_{R \to 0^+} f(R) \to \infty$ and the first limit is established.

2. $R \to \infty$. Let λ_1 be the largest eigenvalue of R. The condition $R \to \infty$ implies that $\lambda_1 \to \infty$, although some eigenvalues may tend to zero as well. As any eigenvalue goes to infinity, the corresponding $b_i \to \infty$, due to the log term. As we just showed, if any $\lambda_j \to 0$, $b_j \to \infty$. The remaining b_i terms, which correspond to strictly positive and finite eigenvalues, remain finite. Since at least $\lambda_1 \to \infty$, then at least one $b_i \to \infty$. Since no $b_i \to -\infty$, $f(R) \to \infty$ as $R \to \infty$.

■

Proposition 2. *Given $\mathbb{Y}$ and $f(R)$ as defined in Proposition 1, a solution to the maximum likelihood problem* $\min_{R>0} f(R)$ *exists.*

Proof. Choose some $R_1 > 0$ such that $f(R) = \alpha$ is finite. Then define the set

$$L := \{R \mid R \geq 0, f(R) \leq \alpha\}$$

L is a non-empty subset of the feasible region. Since $f(R) > \alpha$ for any feasible R which is not in L, the solution to the MLE problem, if it exists, lies in L. $f(R)$ is continuous on L and the set L is closed and bounded. Therefore, by the Weierstrass theorem, the problem $\min_{R \in L} f(R)$ has a solution. This solution also solves $\min_{R>0} f(R)$ ■

Next we return to the maximum likelihood problem defined in (5). The propositions above do not directly apply because we have only one sample of the Np-vector Y. As each y_i is correlated, we must treat Y as a single vector. In addition P has a known structure in terms of Q_w and R_v, whereas R in Proposition 1 is entirely unknown. First we consider the behavior of $\phi(Q_w, R_v)$ on the boundary as P becomes semi-definite or $P \to \infty$.

Proposition 3. *Let the data $Y \in \mathbb{R}^{Np}$ be generated from a normal distribution with mean zero and covariance $P^* > 0$ (strictly positive definite), so that $\begin{bmatrix} y_1 & \dots & y_N \end{bmatrix}$ is rank p with probability one. Assume also (A, C) observable and $N \geq n$. Then $\phi(Q_w, R_v) \to \infty$ if either any eigenvalue $\lambda_i(P) \to 0^+$ or $P \to \infty$.*

Proof. Since P is symmetric, it has eigendecomposition $P = W \Lambda W'$. Then

$$Y' P^{-1} Y = Y' W \Lambda^{-1} W' Y = a' \Lambda^{-1} a = \sum_{i=1} \frac{a_i^2}{\lambda_i}$$

in which the scalar a_i is the i^{th} element of the vector $a := W'Y$.

Then we write the objective function as

$$\phi(Q_w, R_v) = \sum_{i=1}^{Np} b_i \qquad\qquad b_i := \ln(\lambda_i) + \frac{a_i^2}{\lambda_i}$$

If λ_i is finite, then b_i is finite as well. As $\lambda_i \to \infty$, $b_i \to \infty$ because the first term goes to infinity and the second to zero. As $\lambda_i \to 0$, $\ln(\lambda_i) \to -\infty$. When $a_i \neq 0$, then $\frac{a_i^2}{\lambda_i} \to \infty$ faster than $\ln(\lambda_i) \to -\infty$, so $b_i \to \infty$.

In this case we are no longer guaranteed that $a_i \neq 0$. However, due to the structure of P, as one eigenvalue of P tends to zero, then N eigenvalues of P tend to zero at the same rate, as explained below. Let $\lambda_1 \ldots \lambda_N$ be the eigenvalues of P which go to zero. Then $\phi \to \infty$ as long as at least one of $a_1 \ldots a_N$ is non-zero. In other words, $\phi \to \infty$ as long as $W_0'Y \neq 0$, where W_0 is the null space of P.

Next we show that $W_0'Y \neq 0$ with probability one. We write P as

$$P = P_Q + P_R \qquad P_Q = \mathbb{O} \begin{bmatrix} Q_w & & \\ & \ddots & \\ & & Q_w \end{bmatrix} \mathbb{O}' \qquad P_R = \begin{bmatrix} R_v & & \\ & \ddots & \\ & & R_v \end{bmatrix} \tag{6}$$

Since $P \geq 0$, $W_i'PW_i = 0$ implies that W_i is in the null space of P. As P_Q and P_R are both positive semidefinite, then we must have $W_i'P_QW_i = W_i'P_RW_i = 0$ for any W_i in the null space of P. In other words, W_i is in the null space of P if and only if it is in the null space of *both* P_Q and P_R.

Consider the block-diagonal structure of P_R. Let one eigenvalue of R_v go to zero and let v_1 be the corresponding eigenvector. We write the null space of P_R as

$$W_{R0} = \begin{bmatrix} v_1 & & 0 \\ & \ddots & \\ 0 & & v_1 \end{bmatrix}$$

Due to the structure of P_Q, either W_{R0} lies in the null space of P_Q, in which case $W_0 = W_{R0}$, or else no non-zero vector lies in both null spaces, in which case P is non-singular (see Appendix A).

Since $W_0 = W_{R0}$, then $(W_0'Y)_j = v_1'y_j$ and $W_0'Y = v_i' \begin{bmatrix} y_1 \ldots y_n \end{bmatrix}$. Since $\begin{bmatrix} y_1 \ldots y_n \end{bmatrix}$ is full row rank with probability one, we are guaranteed that $W_0'Y \neq 0$. Further, since the dimension of W_0 is either zero or N, then if one eigenvalue of P tends to zero, N eigenvalues of P approach zero at the same rate.

Next we consider the case in which multiple eigenvalues of R_v tend to zero. Let R_m denote a matrix in which the first m eigenvalues of R_v tend to zero. Then we perturb R_m slightly:

$$R_{mr} = R_m + rW_R \mathrm{diag}\left(\begin{bmatrix} 0 & 1 & \frac{1}{2} & \cdots & \frac{1}{m-1} & 0 & \ldots 0 \end{bmatrix}\right) W_R' \tag{7}$$

in which W_R contains the eigenvectors of R and r is a positive scalar. The perturbed matrix R_{mr} has only one zero eigenvalue.

Let Q_r denote Q with a zero eigenvalue such that P_Q and P_R have the same null space. As shown above, as $(Q_w, R_v) \to (Q_r, R_{mr})$, then $\phi(Q_r, R_{mr}) \to \infty$. Since we can choose any positive r for the perturbation in (7), R_m is arbitrarily close to R_{mr}. Since ϕ is continuous in Q and R and R_{mr} is continuous in r, $\phi(Q_r, R_{mr})$ is also continuous in r. Thus $\phi(Q_r, R_m) \to \infty$ as well.

Therefore, as any eigenvalue of P goes to infinity or zero, $\phi(Q_w, R_v) \to \infty$. ■

Proposition 4. *Given that the assumptions in Proposition 3 are satisfied, a solution exists to the maximum likelihood problem defined in* (5)*.*

Proof. As Q_w or $R_v \to \infty$, $P \to \infty$ (see Appendix B) and $\phi \to \infty$ (by Proposition 3). As $Q_w \to 0$ or $R_v \to 0$, either P is positive definite and $\phi(Q_w, R_v)$ is finite, or else $P \to 0$ and $\phi(Q_w, R_v) \to \infty$ (see Appendix B and Proposition 3). Let $\Omega := \{(Q_w, R_v) \mid Q_w \geq 0, R_v \geq 0\}$ be the feasible region of (Q_w, R_v). Choose a feasible point $(Q_1, R_1) \in \Omega$ such that $P(Q_1, R_1)$ is non-singular and let $\phi_1 = \phi(Q_1, R_1)$. Then define

$$L := \{(Q_w, R_v) \mid Q_w \geq 0, R_v \geq 0, \phi(Q_w, R_v) \leq \phi_1\}$$

L is a non-empty subset of Ω. Since any (Q_w, R_v) that lies in Ω but not in L must have $\phi > \phi_1$, the solution to (5), if it exists, lies in L.

Since $\phi(Q_w, R_v)$ is continuous on L and the set L is closed and bounded the problem

$$\min_{Q_w, R_v} \phi(Q_w, R_v) \quad \text{subject to } (Q_w, R_v) \in L$$

has a solution by the Weierstrass theorem. Therefore, a solution to (5) exists. ■

Note that Propositions 1-4 rely on the assumption that $\begin{bmatrix} y_1 \dots y_N \end{bmatrix}$ is full row rank. As shown in Appendix C, this condition is satisfied with probability one when Y is generated from a normal distribution with a positive definite covariance matrix.

4 Uniqueness of the Solution

We find first and second differentials of $\phi(P) = \phi(Q_w, R_v)$ defined in (5). Appendix D contains several matrix differentials used in the following derivation. From (4), we write dP as

$$dP = \sum_{i=1}^{N+K-1} \mathbb{O}_i \, (dQ_w) \, \mathbb{O}_i' + \sum_{i=1}^{N} \mathbb{I}_i \, (dR_v) \, \mathbb{I}_i' \tag{8}$$

We then write $d\phi$ as

$$d\phi = \operatorname{tr}\left((dP)\, P^{-1} \left(P - YY'\right) P^{-1}\right) \tag{9}$$

Using dP as defined in (8), we write $d\phi$ as:

$$\begin{aligned} d\phi = \operatorname{tr}&\left((dQ_w) \sum_i \mathbb{O}_i' P^{-1}(P - YY')P^{-1}\mathbb{O}_i\right) \\ + \operatorname{tr}&\left((dR_v) \sum_i \mathbb{I}_i' P^{-1}(P - YY')P^{-1}\mathbb{I}_i\right) \end{aligned}$$

Any solution to (5) on the interior of the region $Q_w, R_v \geq 0$ must satisfy $d\phi = 0$ for all dQ_w and dR_v and therefore satisfies the equations

$$\sum_i \mathbb{O}_i' P^{-1}(P - YY')P^{-1}\mathbb{O}_i = 0$$

$$\sum_i \mathbb{I}_i' P^{-1}(P - YY')P^{-1}\mathbb{I}_i = 0$$

Note that we cannot choose $\hat{P} = YY'$, as that choice of $\hat{P}$ would exceed our degrees of freedom and result in P singular.

We have for the second differential

$$\begin{aligned} d^2\phi = &\operatorname{tr}((dP)\, P^{-1}\, (dP)\, P^{-1}) \\ &- 2\operatorname{tr}\left((dP)\, P^{-1}\, (dP)\, P^{-1}(P - YY')P^{-1}\right) \end{aligned} \tag{10}$$

We can write $d^2\phi$ in terms of dQ_w and dR_v, but the equation quickly becomes very complicated.

Any minimum on the interior satisfies $d\phi = 0$ and $d^2\phi > 0$. For any $P > 0$ and $dP \neq 0$, the first term in (10) is strictly positive. However, the sign of the second term remains unknown, even at a stationary point. The number of stationary points is also unknown. Therefore, we cannot easily establish when the MLE problem has a unique solution from looking at the differentials. In addition, although we cannot have a solution on the boundary $P \to 0^+$, we may still have solutions on the boundary $Q_w \to 0^+$ or $R_v \to 0^+$. In the next section, we gain further insight on the conditions for uniqueness by comparing this problem with the ALS problem.

5 Connection to Autocovariance Least-Squares (ALS) Method

Here we follow the derivation as in [14]. For simplicity, we assume that $L = 0$ in both the MLE and ALS problems.

We rewrite the MLE first order condition in (9) as

$$\operatorname{tr}\left(P^{-1}\,(dP)\left(I_{pN} - P^{-1}YY'\right)\right) = 0 \tag{11}$$

Let $\mathcal{I}_n := \operatorname{vec}(I_n)$. Noting that for any $n \times n$ matrix A, $\operatorname{tr}(A) = \mathcal{I}_n'\operatorname{vec}(A)$, we rewrite (11) as

$$\mathcal{I}_{Np}'\left(\left(I_{pN} - YY'P^{-1}\right) \otimes P^{-1}\right)\operatorname{vec}(dP) = 0 \tag{12}$$

Note that there was an error in [14, p. 128], which is fixed here. Next we write (12) in terms of dQ_w and dR_v. Starting with Y in terms of $x(0)$ as in (1) and defining

$$\mathscr{O} = \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{N-1} \end{bmatrix} \qquad \Gamma_f = \begin{bmatrix} 0 & 0 & \dots & 0 & 0 \\ C & 0 & \dots & 0 & 0 \\ \dots & \ddots & & & \vdots \\ CA^{N-1} & CA^{N-2} & \dots & C & 0 \end{bmatrix}$$

we write P as

$$P = \mathscr{O}P_x\mathscr{O}' + \Gamma_f \bigoplus_{i=1}^{N} Q_w \Gamma_f' + \bigoplus_{i=1}^{N} R_v \tag{13}$$

in which $P_x = \text{cov}(x) = AP_xA + Q_w$ and $\bigoplus_{i=1}^{N} A$ indicates the direct sum. Using the Lyapunov equation for P_x, we vectorize (13) to obtain

$$\begin{aligned}\text{vec}(P) &= \begin{bmatrix}\mathcal{A}_1 & \mathcal{A}_2\end{bmatrix}\begin{bmatrix}\text{vec}(Q_w)\\ \text{vec}(R_v)\end{bmatrix} \\ \mathcal{A}_1 &= (\mathscr{O}\otimes\mathscr{O})(I_{n^2} - A\otimes A)^{-1} + (\Gamma_f\otimes\Gamma_f)\mathcal{J}_{n,N} \\ \mathcal{A}_2 &= \mathcal{J}_{p,N}\end{aligned} \tag{14}$$

in which the permutation matrix $\mathcal{J}_{m,N}$ satisfies the relationship $\text{vec}\left(\bigoplus_{i=1}^{N} A\right) = \mathcal{J}_{m,N}\text{vec}(A)$ (for an $m\times m$ matrix A). Apart from the approximation $A^K \approx 0$, the formula for P here is equivalent to that in the previous sections. We simply choose to write P in terms of $x(0)$ rather than including additional past noise terms.

Letting $\mathbb{A}_0 = \mathcal{I}_{Np}'\left(\left(I_{pN} - YY'P^{-1}\right)\otimes P^{-1}\right)$ and using (14) in (12), we write the first order condition as

$$\mathbb{A}_0\begin{bmatrix}\mathcal{A}_1 & \mathcal{A}_2\end{bmatrix} = \begin{bmatrix}0 & \dots 0\end{bmatrix} \tag{15}$$

We rewrite $\mathbb{A}_0$ as

$$\mathbb{A}_0 = \text{vec}(P^{-1})' - \frac{1}{2}\text{vec}(YY')'\left(I_{(pN)^2} + K_{(pN)(pN)}\right)\left(P^{-1}\otimes P^{-1}\right)$$

in which the commutation matrix K_{ij} is such that $\text{vec}(A) = K_{ij}\text{vec}(A')$ where A has dimensions $i\times j$. Then, taking the transpose of (15), we write the first order condition for the maximum likelihood problem as

$$0 = \begin{bmatrix}\mathcal{A}_1'\\ \mathcal{A}_2'\end{bmatrix}\text{vec}(P^{-1}) - \frac{1}{2}\begin{bmatrix}\mathcal{A}_1'\\ \mathcal{A}_2'\end{bmatrix}\left(P^{-1}\otimes P^{-1}\right)\left(I_{(pN)^2} + K_{(pN)(pN)}\right)\text{vec}(YY') \tag{16}$$

We compare this condition to the first order condition for the ALS problem, which forms a least-squares optimization problem for the elements of Q_w and R_v. For the full matrix, unconstrained, weighted ALS problem, when $N = N_d$ (*i.e.* the number of autocovariances is equal to the number of data points), the ALS solution satisfies the first order condition

$$\begin{bmatrix}\mathcal{A}_1'\\ \mathcal{A}_2'\end{bmatrix}W^\dagger\begin{bmatrix}\mathcal{A}_1 & \mathcal{A}_2\end{bmatrix}\begin{bmatrix}\text{vec}(Q_w)\\ \text{vec}(R_v)\end{bmatrix} - \begin{bmatrix}\mathcal{A}_1'\\ \mathcal{A}_2'\end{bmatrix}W^\dagger\text{vec}(YY') \tag{17}$$

We define W and its psuedoinverse as

$$\begin{aligned}W &= \frac{1}{2}\left(I_{(pN)^2} + K_{(pN)(pN)}\right)(P\otimes P) \\ W^\dagger &= \frac{1}{2}\left(P^{-1}\otimes P^{-1}\right)\left(I_{(pN)^2} + K_{(pN)(pN)}\right)\end{aligned}$$

Using this value of $W^\dagger$ in (17) and utilizing the fact that

$$\text{vec}(P^{-1}) = \frac{1}{2}\left(P^{-1}\otimes P^{-1}\right)\left(I_{(pN)^2} + K_{(pN)(pN)}\right)\text{vec}(P)$$

then the ALS first order condition is identical to (16).

From equation (18) in [15], W is the covariance of $\text{vec}(YY')$ when $N = N_d$, and therefore it is the minimum variance weighting for the ALS problem. Thus, the MLE method is equivalent to the optimally-weighted full matrix ALS method with $N = N_d$ (neglecting the semidefinite constraints). This conclusion allows us to make several observations:

1. Since W depends on the unknown Q_w and R_v, solving the optimally-weighted ALS problem requires either nonlinear optimization or an iterative procedure as suggested in [15].

2. From (14), when $\begin{bmatrix}\mathcal{A}_1 & \mathcal{A}_2\end{bmatrix}$ is not full rank, more than one (Q_w, R_v) maps to any given P. Since the likelihood depends on Q_w and R_v only through P, there is not a unique solution to the MLE problem.

3. As this rank condition is necessary for the unweighted ALS problem to have a unique solution, when there is not a unique ALS solution, there cannot be a unique MLE solution.

4. It does not necessarily follow that there *is* a unique MLE solution when there is a unique ALS solution.

5. It is particularly worthwhile to note that, in the case when the noise-shaping matrix G is unknown, the following conditions are necessary for the ALS or MLE problem to have a unique solution:

   (a) (A, C) observable

   (b) $\text{rank}(C) = n$

   (c) $\text{rank}(A) = n$

6 Solving the Problem

One limitation of the maximum likelihood method is that it requires the computation, storage, and manipulation of very large matrices used in the likelihood. Here we suggest several methods to reduce the computation time:

1. **Sparsity**: P and the matrices from which it is composed are sparse, as seen in (2) and (3). By treating these matrices as sparse, we reduce both the storage requirements and the computation time.

2. **Cholesky Decomposition**: Computing $\ln \det(P)$ for large P presents challenges in both numerical accuracy and computation time. If P has many eigenvalues that are less than one, computing the log determinant directly may return an answer of minus infinity, while in reality this term has a finite value. Calculating the log determinant via the eigenvalues produces a more accurate numerical result in Octave and Matlab. However, finding the eigenvalues may be computationally expensive, and Octave and Matlab do not utilize sparsity in this step. A faster method is to compute the log

determinant via Cholesky factorization. The positive definite matrix P is decomposed uniquely into $P = LL'$ in which L is lower-triangular. The log determinant of P is computed as $\log\det(P) = 2\sum_i \log(L_{ii})$ in which L_{ii} are the diagonal entries of L.

3. **Solving Linear Systems of Equations**: Directly inverting P to calculate $Y'P^{-1}Y$ is computationally expensive, and the computation time is not reduced for sparse matrices. To avoid computing the inverse directly, we first find the vector X which solves the equation $PX = Y$ and then calculate $Y'P^{-1}Y = Y'X$. In Octave and Matlab, the "mldivide" function (abbreviated by the \ symbol) uses efficient algorithms, based on the structure of P, to solve $PX = Y$.

We also recommend optimizing over $\tilde{Q}$ and $\tilde{R}$, in which $Q_w = \tilde{Q}\tilde{Q}'$ and $R_v = \tilde{R}\tilde{R}'$ rather than optimizing directly over Q_w and R_v, as this decomposition enforces both the positive definite and the symmetry constraints of Q_w and R_v.

6.1 Optimal Innovation MLE Method

The MLE method proposed by [2] utilizes the fact that the innovations, $y(k)-\hat{y}(k|k-1)$, are white under an optimal estimator. This method reduces the computational time because the objective function is written in terms of the independent innovations rather than the correlated outputs. The optimal innovations MLE problem is written as

$$\begin{aligned} \min_{Q_w,R_v} & N\ln(\det(\Sigma_e)) + \sum_{i=0}^{N-1} (y(k) - \hat{y}(k|k-1))' \,\Sigma_e^{-1}\, (y(k) - \hat{y}(k|k-1)) \\ & \text{subject to: Kalman filter equations} \\ & \qquad Q_w, R_v \geq 0 \end{aligned}$$

in which Σ_e is the variance of the innovation. This method was designed for nonlinear systems using the extended Kalman filter. We apply it to a linear time invariant system using the following steps in each iteration of the optimizer:

1. Calculate the steady-state predictor gain and innovation variance (Σ_e) from the estimator Riccati equation, using the current values of Q_w and R_v.

2. Calculate the innovations using the Kalman filter equations.

3. Calculate the block diagonal matrix $P = I_N \otimes \Sigma_e$; use sparsity to reduce the storage space of P.

4. Calculate the objective function as $\phi = N\log\det(\Sigma_e) + Y'_{\text{inn}}(P\backslash Y_{\text{inn}})$ in which the Np-vector Y_{inn} contains all the innovations[2].

[2] In the examples studied, it is faster to compute the term $Y'_{\text{inn}}(P\backslash Y_{\text{inn}})$ than to calculate and add the individual terms $(y(k) - \hat{y}(k|k-1))' \,\Sigma_e^{-1}\, (y(k) - \hat{y}(k|k-1))$

Since calculating the innovations requires a value for $\hat{x}(0)$, we also optimize over this parameter.

For this method, P is block diagonal, so $Y_{\mathrm{inn}}' P^{-1} Y_{\mathrm{inn}}$ is computed more quickly. Computing the log determinant is also significantly faster, as only the determinant of the $p \times p$ matrix Σ_e is calculated, rather than the determinant of the $Np \times Np$ matrix P. These advantages come at the cost of computing the innovations within the optimizer at each iteration, since the estimator gain changes as Q_w and R_v are updated. However, for larger systems, the optimal innovation MLE method significantly reduces the computational time. Both formulations of the MLE problem lead to the same estimates of Q_w and R_v.

7 Examples

7.1 Scalar Example

Consider the example

$$A = 0.600 \qquad C = 0.483 \qquad Q_w = 7 \qquad R_v = 3$$

We used $N = 1000$ data points and $K = 23$ (placing a threshold of 10^{-5} on the norm of A). We solved the MLE problem in Octave using the built-in function sqp. We also solved the ALS problem for comparison, in which the optimal weighting is approximated from the data and the window is fixed at $N_{\mathrm{ALS}} = 15$. The results are summarized in Table 1 and are compared to the sample variances of the process and measurement noises used in the simulation. These sample variances would be the best estimate for Q_w and R_v if the sequence of noises were known. Both the MLE and ALS method achieve similar results, but the MLE solution produces the lowest objective value compared to the ALS solution and the sample variances. Figure 1 plots the objective function vs. Q_w and R_v; we see that the objective function does indeed have a unique minimum and tends to infinity on the boundaries of P. For $N = 1000$, the computation times for the MLE and ALS methods are comparable. However, when $N = 10000$, the ALS technique is faster by two orders of magnitude. Unlike in the MLE method, the computation time in the ALS method has little dependence on the number of data points, since the size of the optimization problem is unchanged.

7.2 Comparison to the Expectation Maximization Approach

Using the same scalar example, we compare the MLE and ALS methods to the expectation maximization (EM) approach described in [6] and [2]. We simulated 50 instances of the problem and calculated Q_w and R_v using all three approaches. Figure 2 plots $\hat{Q}_w$ and $\hat{R}_v$ for each approach. The estimates for all the methods are centered around the true mean values, and the variances of the estimates are similar. The MLE and EM methods produce nearly identical results.

We summarize the results in Table 2. The estimates from all three methods have similar means and variances, although the MLE and EM methods lead to slightly lower variances than does the ALS technique. Since the MLE and EM methods produce approximately

Table 1: Comparison of results for scalar example.

$N = 1000$

	Q_w	R_v	ϕ	Time (s)
MLE	8.69	2.74	2656.85	2.88
ALS	8.66	2.65	2657.10	1.29
Sample Var.	6.79	3.07	2660.04	

$N = 10000$

	Q_w	R_v	ϕ	Time (s)
MLE	6.83	3.10	26 366.66	113.81
ALS	6.60	3.15	26 367.13	1.72
Sample Var.	6.90	3.09	26 366.74	

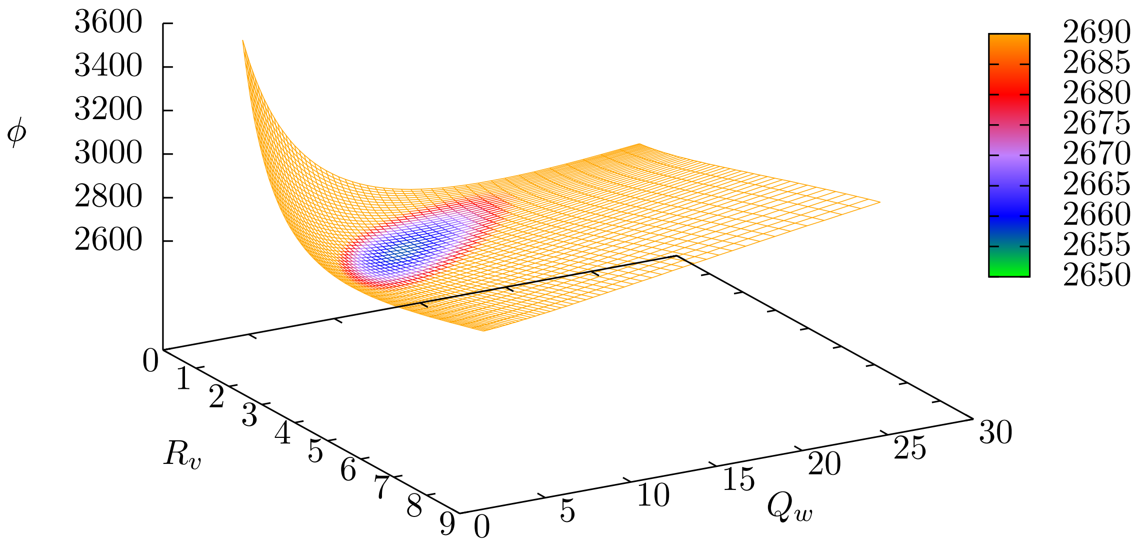

Figure 1: Likelihood objective value vs. Q_w and R_v for scalar example with $N = 1000$ data points. The objective has a unique minimum and goes to infinity on the boundaries of P.

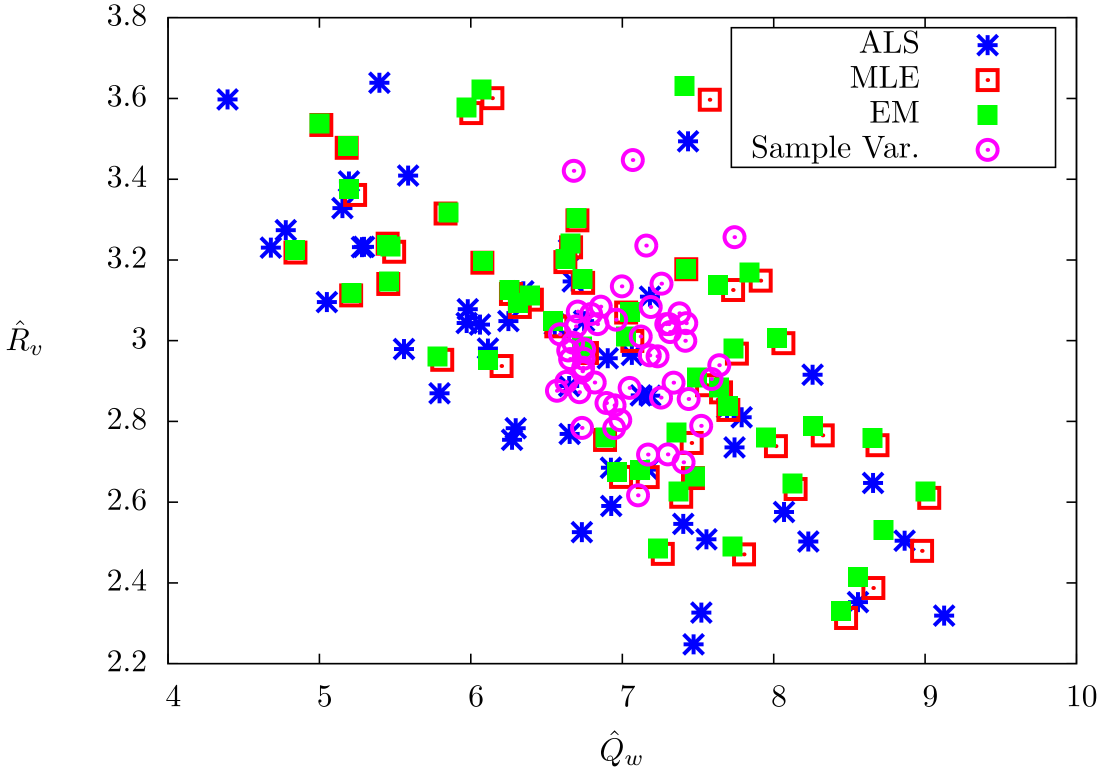

Figure 2: Noise variance estimates for ALS, MLE, and EM methods. The estimates from each method have a similar mean and variance. The EM and MLE methods produce approximately the same results.

the same results at each iteration, they have nearly the same objective function values. Therefore, the EM estimates approximate the maximum likelihood solution more accurately than do ALS estimates. For this problem, the ALS method is the fastest of the three options. The EM method is slower than either the ALS or MLE methods, due to its iterative nature. However, the EM method may scale better as the amount of data or system dimensions increase. [6] successfully applied the EM method to a larger problem on which the ALS method ran out of memory.

Table 2: Mean and variance of the estimates, average objective value, and average CPU time for the ALS, MLE, and EM methods.

	$\mathbb{E}(\hat{Q}_w)$	$\mathbb{E}(\hat{R}_v)$	$\text{var}(\hat{Q}_w)$	$\text{var}(\hat{R}_v)$	$\langle\phi\rangle$	$\langle$CPU Time (s)$\rangle$
ALS	6.699	2.916	1.306	0.113	2617.9279	0.904
MLE	6.944	2.988	1.208	0.105	2617.2499	2.771
EM	6.909	3.002	1.161	0.103	2617.2544	11.959
Sample Var.	7.044	2.969	0.099	0.027	2618.8098	

Table 3: Computation time for steps in the MLE method (in seconds of CPU time)

Quantity	Method	Full	Sparse
P		11.2	0.0436
$\log(\det(P))$	Eigenvalues	4.97	5.01
	Cholesky	0.418	0.0113
$Y'P^{-1}Y$	Inverse	74.9	77.4
	Left Division	0.503	0.0169

7.3 Example: $p = n = 2$

In this example, we illustrate how the methods mentioned in Section 6, including utilizing sparse matrices, significantly reduce the computation time. We consider the example

$$A = \begin{bmatrix} 0.600 & 0 \\ 0 & 0.338 \end{bmatrix} \qquad C = \begin{bmatrix} 0.887 & 0.309 \\ 0.238 & 0.732 \end{bmatrix}$$

$$Q_w = \begin{bmatrix} 17.9 & 10.5 \\ 10.5 & 6.99 \end{bmatrix} \qquad R_v = \begin{bmatrix} 6.62 & 0 \\ 0 & 5.22 \end{bmatrix}$$

Table 3 summarizes the time spent in each step. By using efficient numerical methods, the computation time for each iteration is reduced from approximately 91s to 0.072s.

We also show the computation time for the optimal innovation MLE method proposed by [2] in Table 4. In this table, the time to compute $Y'P^{-1}Y$ includes the time to calculate the innovations; left division was used to avoid directly inverting P. Comparing Table 3 to Table 4, we see that the optimal innovation method requires additional time to compute the innovations but reduces the computation time for the other steps in each iteration.

In Table 5 we compare the solutions and solution time of the "slow MLE" (full matrices, eigenvalues, and inverse), "fast MLE" (sparse matrices, Cholesky factorization, and left division), optimal innovation MLE, and ALS techniques for $N = 1000$. All MLE methods give identical results, however, the improvements to the code decrease the computation time from several hours to less than a minute. The MLE method based on the optimal

Table 4: Computation time for steps in the MLE method based on optimal innovations (in seconds of CPU time)

Quantity	Full	Sparse
P	0.0137	0.0118
$\log(\det(P))$	7.30×10^{-5}	7.30×10^{-5}
$Y'P^{-1}Y$	0.550	0.0518

innovations only slightly decreases the computation time for this problem size. The ALS method gives similar results with the smallest computation time (around 1s), but has a slightly higher objective value.

7.4 Example: $\mathbf{p = n = 5}$

In this example we consider a larger system, with 5 states and outputs. The data were generated using

$$Q_w = \begin{bmatrix} 8.92 & 9.12 & 14.44 & 5.82 & 12.54 \\ 9.12 & 13.07 & 14.90 & 10.41 & 17.13 \\ 14.44 & 14.90 & 25.11 & 11.32 & 21.50 \\ 5.82 & 10.41 & 11.32 & 11.98 & 14.73 \\ 12.54 & 17.13 & 21.50 & 14.73 & 24.20 \end{bmatrix}$$
$$R_v = \mathrm{diag}\left(\begin{bmatrix} 1.51 & 2.10 & 1.39 & 3.78 & 1.11 \end{bmatrix}\right)$$

We used two initial conditions to solve the MLE problem: (1) $Q_w = R_v = I$ and (2) the ALS estimates. The MLE solution yields a lower objective value than the ALS solution or the sample covariances of the noises. Changing the initial condition has a negligible effect on the MLE results but reduces the computation time. We also solved the MLE problem using the optimal innovations method, starting from both initial conditions. For this example, the optimal innovations MLE method significantly reduces the computation time and reaches the same solution as the output MLE method.

Table 5: Results for Second Example

	Q_w	$\mathrm{diag}(R_v)$	ϕ	Time (s)
"Slow" MLE	$\begin{bmatrix} 16.9 & 10.8 \\ 10.8 & 6.88 \end{bmatrix}$	$\begin{bmatrix} 7.12 \\ 5.08 \end{bmatrix}$	7326.5	17953
"Fast" MLE	$\begin{bmatrix} 16.9 & 10.8 \\ 10.8 & 6.88 \end{bmatrix}$	$\begin{bmatrix} 7.12 \\ 5.08 \end{bmatrix}$	7326.5	49.3
Optimal Innovation MLE	$\begin{bmatrix} 16.9 & 10.8 \\ 10.8 & 6.90 \end{bmatrix}$	$\begin{bmatrix} 7.13 \\ 5.07 \end{bmatrix}$	7326.5	43.9
ALS	$\begin{bmatrix} 17.2 & 10.55 \\ 10.55 & 6.47 \end{bmatrix}$	$\begin{bmatrix} 6.82 \\ 4.95 \end{bmatrix}$	7327.5	1.86
Sample Cov.	$\begin{bmatrix} 17.9 & 10.4 \\ 10.4 & 6.91 \end{bmatrix}$	$\begin{bmatrix} 6.63 \\ 4.92 \end{bmatrix}$	7328.4	

Table 6: Results for 5 State Example

Method	Results
MLE $Q_0 = I$	$Q_w = \begin{bmatrix} 8.50 & 8.19 & 9.97 & 2.84 & 10.31 \\ 8.19 & 14.29 & 17.02 & 10.40 & 18.45 \\ 9.97 & 17.02 & 28.58 & 10.92 & 21.76 \\ 2.84 & 10.40 & 10.92 & 11.19 & 14.29 \\ 10.31 & 18.45 & 21.76 & 14.29 & 24.48 \end{bmatrix}$ $\mathrm{diag}(R_v) = \begin{bmatrix} 1.53 & 1.90 & 1.58 & 2.99 & 0.801 \end{bmatrix}$ $\phi = 15375$ Time (min) = 34.8
MLE $Q_0 = \hat{Q}_{ALS}$	$Q_w = \begin{bmatrix} 8.34 & 8.32 & 10.02 & 2.85 & 10.32 \\ 8.32 & 14.25 & 17.06 & 10.35 & 18.45 \\ 10.02 & 17.06 & 28.60 & 10.86 & 21.75 \\ 2.85 & 10.35 & 10.86 & 11.25 & 14.26 \\ 10.32 & 18.45 & 21.75 & 14.26 & 24.25 \end{bmatrix}$ $\mathrm{diag}(R_v) = \begin{bmatrix} 1.52 & 1.91 & 1.57 & 3.00 & 0.822 \end{bmatrix}$ $\phi = 15375$ Time (min) = 20.5
MLE Optimal Innovations $Q_0 = I$	$Q_w = \begin{bmatrix} 8.36 & 8.31 & 9.99 & 2.81 & 10.29 \\ 8.31 & 14.17 & 17.03 & 10.43 & 18.47 \\ 9.99 & 17.03 & 28.56 & 10.84 & 21.70 \\ 2.81 & 10.43 & 10.84 & 11.14 & 14.19 \\ 10.29 & 18.47 & 21.70 & 14.19 & 24.17 \end{bmatrix}$ $\mathrm{diag}(R_v) = \begin{bmatrix} 1.53 & 1.91 & 1.57 & 3.00 & 0.818 \end{bmatrix}$ $\phi = 15375$ Time (min) = 6.75
MLE Optimal Innovations $Q_0 = \hat{Q}_{ALS}$	$Q_w = \begin{bmatrix} 8.36 & 8.31 & 9.99 & 2.81 & 10.29 \\ 8.31 & 14.18 & 17.04 & 10.43 & 18.47 \\ 9.99 & 17.04 & 28.56 & 10.84 & 21.70 \\ 2.81 & 10.43 & 10.84 & 11.14 & 14.19 \\ 10.29 & 18.47 & 21.70 & 14.19 & 24.18 \end{bmatrix}$ $\mathrm{diag}(R_v) = \begin{bmatrix} 1.53 & 1.90 & 1.57 & 3.00 & 0.817 \end{bmatrix}$ $\phi = 15375$ Time (min) = 4.22

ALS	$Q_w = \begin{bmatrix} 6.42 & 7.28 & 6.25 & 1.58 & 7.73 \\ 7.28 & 14.55 & 16.80 & 7.11 & 15.47 \\ 6.25 & 16.80 & 25.01 & 6.48 & 16.42 \\ 1.58 & 7.11 & 6.48 & 8.78 & 9.54 \\ 7.73 & 15.47 & 16.42 & 9.54 & 17.48 \end{bmatrix}$ $\text{diag}(R_v) = \begin{bmatrix} 1.38 & 1.70 & 1.14 & 2.33 & 0.94 \end{bmatrix}$ $\phi = 15465$ Time (min) = 0.173
Sample Covariances	$Q_w = \begin{bmatrix} 8.32 & 8.48 & 13.59 & 5.50 & 11.70 \\ 8.48 & 12.51 & 13.97 & 10.21 & 16.32 \\ 13.59 & 13.97 & 23.82 & 10.70 & 20.16 \\ 5.50 & 10.21 & 10.69 & 11.85 & 14.21 \\ 11.70 & 16.32 & 20.16 & 14.21 & 22.92 \end{bmatrix}$ $\text{diag}(R_v) = \begin{bmatrix} 1.40 & 2.05 & 1.34 & 3.88 & 1.07 \end{bmatrix}$ $\phi = 15387$

8 Conclusions

We formulate a direct maximum likelihood optimization problem to estimate the process and measurement noise covariance matrices for a linear time invariant system from the output measurements. We provide sufficient conditions under which a solution to the optimization problem exists. Uniqueness remains an open issue but is compared to the uniqueness conditions for the ALS technique. For small scale systems, the MLE problem can be solved in Octave; several improvements make the optimization more efficient. Solving the MLE problem in terms of the optimal innovations also reduces the computational time. Many open areas exist for further research, especially establishing sufficient conditions for uniqueness, solving the problem for $n > p$, and developing more efficient ways to perform the optimization on large-scale problems.

9 Appendices

A Null space of P_Q

Proposition 5. *Given P_Q and P_R as defined in* (6)*, then either (1) null$(P_R) \subseteq$ null(P_Q) and null$(P) =$ null(P_R), or (2) null$(P_R) \subseteq$ range(P_Q) and null$(P) = \{0\}$.*

Proof. To prove the proposition, first we show that any non-zero vector in the null space of P_R is also in the null space of P_Q if and only if $w_j := (A')^{K-j}C'v_1$ is in the null space of Q_w for all $1 \leq j \leq K$.

We write any (non-zero) vector in the null space of P_R as

$$V = \begin{bmatrix} \alpha_1 v_1 \\ \alpha_2 v_1 \\ \vdots \\ \alpha_N v_1 \end{bmatrix}$$

in which $\alpha_1 \dots \alpha_N$ are scalars. α_i may be zero, but at least one α_i must be non-zero.

If $V \in \text{null}(P_Q)$, we must have $X_i = \mathbb{O}'V$ in the null space of $(I \otimes Q_w)$. Note that $\mathbb{O}'$ takes the form

$$\mathbb{O}' = \begin{bmatrix} (A')^{K-1}C' & & & \\ \vdots & (A')^{K-1}C' & & \\ C' & \vdots & & \\ & C' & \ddots & (A')^{K-1}C' \\ & & & \vdots \\ & & & C' \end{bmatrix}$$

Let $X = \begin{bmatrix} x_1' & \dots & x_{N+K-1}' \end{bmatrix}'$. If X is in the null space of $(I \otimes Q_w)$, then each x_i must be in the null space of Q_w.

To prove that $w_j \in \text{null}(Q_w)$ implies $V \in \text{null}(P_Q)$, note that each x_i is a linear combination of the w_j. Therefore, if all w_j are in the null space of Q_w, each x_i is in the null space of Q_w, and V is in the null space of Q_w.

To prove that $V \in \text{null}(P_Q)$ implies $w_j \in \text{null}(Q_w)$, assume V is in the null space of P_Q. Define the index m such that $\alpha_j = 0$ for $j = 1, \dots, m-1$ and $\alpha_m \neq 0$. Then $x_i = 0$ for $i < m$, and $x_m = \alpha_m w_1$. Therefore, w_1 is in the null space of Q_w. We write each x_{m+j} as $x_{m+j} = \alpha_m w_{j+1} + \alpha_{m+1} w_j + \cdots + \alpha_{m+j} w_1$, for all $j = 0...K$. If w_i is in the null space of Q_w for all $1 \leq i \leq j$, then w_{j+1} must also be in the null space of Q_w. Since w_1 is in the null space of Q_w, by induction every w_j must lie in the null space of Q_w.

Therefore, V_i is in the null space of P_Q if and only if all the w_i are in the null space of Q_w. Since this condition is true for any vector in the null space of P_R, it must be true for all vectors in the null space. Thus, the null space of P_Q either contains the null space of P_R, or else the null spaces have no non-zero vectors in common.

Since the null space of P is the intersection of the null spaces of P_Q and P_R, it is equal to the null space of P_R when v_1 is in the null space of R_v and $Q_w(A')^iC'$ for $0 \leq i < K$, or else the null space of P contains only the zero element.

■

B Boundary Relationship Between (Q_w, R_v) and P

Proposition 6. *Given C full rank, (A, C) observable, and $N \geq n$,*

1. *$P > 0$ if $R_v > 0$*
2. *$P > 0$ if $Q_w > 0$*

3. $P \to \infty$ if and only if $Q_w \to \infty$ or $R_v \to \infty$

Proof. From (6), $P = P_Q + P_R$. Each term in P is positive semidefinite, so P is strictly positive definite provided that P_Q or P_R is positive definite. If $R > 0$, then $P_R = (I_N \otimes R_v) > 0$, so $P > 0$. Due to its structure, $\mathbb{O}$ is full row rank provided C is full row rank. To see that $\mathbb{O}$ is full row rank, we show that $\mathbb{O}'Y = 0$ only if $Y = 0$. Since the last block row of $\mathbb{O}'$ is $\begin{bmatrix} 0 & \dots & 0 & C' \end{bmatrix}$, for C full row rank, the last p elements of Y must be zero. Likewise, the last $2p$ elements of Y must be zero to enforce that the last two block rows of $\mathbb{O}'$ are zero, and the pattern continues. Therefore, since $(I_{N+K-1} \otimes Q_w) > 0$ when $Q_w > 0$, $P_Q = \mathbb{O}\,(I_{N+K-1} \otimes Q_w)\,\mathbb{O}' > 0$ for $Q_w > 0$. Note that (A, C) observable is not required for these conditions.

We say that $P \to \infty$ if and only if $\|P\|_2 \to \infty$, which implies that the largest eigenvalue of P goes to infinity. To prove that $P \to \infty$, it is sufficient to show that there exists some finite x such that $x'Px \to \infty$.

$P > 0$ implies $x'Px > 0$ for all $x \neq 0$. From (4),

$$x'Px = \sum_i x'\mathbb{O}_i Q_w \mathbb{O}_i' x + \sum_j x'\mathbb{I}_j R_v \mathbb{I}_j' x \tag{18}$$

Let α_k be (one of) the eigenvalues of Q_w that goes to infinity and v_k be the corresponding normalized eigenvector. Then $v_k' Q_w v_k = \alpha_k \to \infty$. For (A, C) observable and $N \geq n$, the block matrix $\mathbb{O}_K$ is full column rank. Therefore, we can always find some x such that $v_k = \mathbb{O}_K' x$. Then $x'\mathbb{O}_K Q_w \mathbb{O}_K' x = v_k' Q_w v_k = \alpha_k \to \infty$. Since (at least) one term in (18) tends to infinity and the other terms are non-negative, $x'Px \to \infty$ and therefore $P \to \infty$. By the same logic, $P \to \infty$ if $R_v \to \infty$. To prove that $P \to \infty$ only if Q_w or $R_v \to \infty$, we choose a finite x such that $x'Px \to \infty$. Then at least one term in (18) tends to infinity. By eigenvalue decomposition, we see that no term can go to infinity unless one of the eigenvalues of Q_w or R_v also goes to infinity. ■

C Rank of Data Matrix

Proposition 7 (Full rank of data matrix)**.** *Let the random variable $y \in \mathbb{R}^p$ be distributed as $N(0, R)$ with $R \in \mathbb{R}^{p \times p}$ positive definite, and let y_i, $i = 1, 2, \dots, N$ be N independent samples of y with $N \geq p$. Arrange the samples as the columns in the data matrix $\mathbb{Y} := \begin{bmatrix} y_1 & y_2 & \cdots & y_N \end{bmatrix}$. Then rank$(\mathbb{Y}) = p$ with probability one.*

Proof. Consider first a data matrix with one or more rows of zeros so that it has rank less than p. Assume without loss of generality that the elements of y are ordered so that the last row of $\mathbb{Y}$ is zero. We note that there is probability zero of achieving this matrix by sampling y. In order to zero the p^{th} component in all the samples, one must have a singular normal in which the unit vector e_p is an eigenvector of R with corresponding eigenvalue $\lambda_p = 0$. For such a semi-definite R, there is probability one of having a zero last row in $\mathbb{Y}$. For positive definite R, however, the probability of a zero row is zero.

To prove the proposition, we consider the (reduced) SVD of $\mathbb{Y}$

$$\mathbb{Y} = U \Sigma V'$$

with $U \in \mathbb{R}^{p \times p}, \Sigma \in \mathbb{R}^{p \times p}, V \in \mathbb{R}^{N \times p}$, in which U is orthonormal, Σ is diagonal, and V has orthonormal columns. Assume for contradiction that $\mathbb{Y}$ has rank less than p. Then consider the transformed random variable $z := U'y$, which has distribution $z \sim N(0, \tilde{R})$ with $\tilde{R} = U'RU$. Since U is nonsingular and R is positive definite, $\tilde{R}$ is also positive definite. If we form the data matrix from $z_i = U'y_i$, we have

$$\mathbb{Z} = U'\mathbb{Y} = \Sigma V'$$

Since $\mathbb{Y}$ has rank less than p, we know that $\sigma_p = 0$. Therefore the last row of $\mathbb{Z}$ is zero, and, combined with $\tilde{R}$ being positive definite, there is a contradiction and the proposition is established. ■

From Proposition 7, $\mathbb{Y}$ is full row rank in Propositions 1 and 2. However, in Propositions 3 and 4, the samples of y_i are not independent, so Proposition 7 does not directly apply. Instead, we prove the following proposition:

Proposition 8. *Let* $Y = \begin{bmatrix} y_1' & \dots & y_N' \end{bmatrix}'$ *be generated from a normal distribution with mean zero and covariance* $P^* > 0$*. Then* $\mathbb{Y} = \begin{bmatrix} y_1 & \dots y_N \end{bmatrix}$ *has full row rank.*

Proof. As in the previous proof it is sufficient to prove that with probability one, $\mathbb{Y}$ does not contain a row of zeros. For proof by contradiction, assume that the last row of $\mathbb{Y}$ is zero. For this row to occur with nonzero probability, the covariance matrix P^* must be singular with zero rows and columns at indices p, $2p$, $3p$, $\dots$, Np. But this covariance contradicts the assumption that $P^* > 0$, and the result is established. ■

D Matrix Differentials

The following matrix differentials come from [8]:

$$\begin{aligned} d(\det(X)) &= \det(X)\mathrm{tr}\left(X^{-1}\left(dX\right)\right) && X \in \mathbb{R}^{n\times n}\text{,invertible} \\ d(\mathrm{tr}(AX)) &= \mathrm{tr}\left(A\left(dX\right)\right) && X \text{ real} \\ dX^{-1} &= -X^{-1}(dX)X^{-1} && X \in \mathbb{R}^{n\times n}\text{,invertible} \end{aligned}$$

Acknowledgment

The authors would like to thank Drs. T.A. Badgwell and W. Li of ExxonMobil for helpful discussion of estimating noise variances from data.

References

[1] T. W. Anderson and I. Olkin. Maximum-likelihood estimation of the parameters of a multivariate normal distribution. *Linear Algebra Appl.*, 70:147–171, 1985.

[2] V. A. Bavdekar, A. P. Deshpande, and S. C. Patwardhan. Identification of process and measurement noise covariance for state and parameter estimation using extended Kalman filter. *J. Proc. Cont.*, 21:585–601, 2011.

[3] T. Bohlin and S. F. Graebe. Issues in nonlinear stochastic grey box identification. *Int. J. Adaptive Cont. Signal Proc.*, 9:465–490, 1995.

[4] R. L. Kashyap. Maximum likelihood identification of stochastic linear systems. *IEEE Trans. Auto. Cont.*, 15(1):25–34, 1970.

[5] N. R. Kristensen, H. Madsen, and S. B. Jørgensen. Parameter estimation in stochastic grey-box models. *Automatica*, 40:225–237, 2004.

[6] W. Li and T. A. Badgwell. Structured covariance estimation for state prediction. Accepted for publication in 53rd IEEE Conference on Decision and Control, 2014.

[7] J. R. Magnus. Maximum likelihood estimation of the GLS model with unknown parameters in the disturbance covariance matrix. *J. Econometrics*, 7(3):281–312, 1978.

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

[9] P. Matisko and V. Havlena. Noise covariance estimation for Kalman filter tuning using Bayesian approach and Monte Carlo. *Int. J. Adaptive Cont. Signal Proc.*, 27(11):957–973, 2013.

[10] R. K. Mehra. Identification of stochastic linear dynamic systems. In *1969 IEEE Symposium on Adaptive Processes (8th) Decision and Control*, 1969.

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

[12] 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.

[13] S. J. Qin. An overview of subspace identification. *Comput. Chem. Eng.*, 30:1502–1513, 2006.

[14] M. R. Rajamani. *Data-based Techniques to Improve State Estimation in Model Predictive Control.* PhD thesis, University of Wisconsin–Madison, October 2007.

[15] M. R. Rajamani and J. B. Rawlings. Estimation of the disturbance structure from data using semidefinite programming and optimal weighting. *Automatica*, 45:142–148, 2009.

[16] 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.