Practical Improvements to Autocovariance Least-Squares

Megan A. Zagrobelny and James B. Rawlings
Dept. of Chemical and Biological Engineering, University of Wisconsin-Madison, Madison, WI 53706

DOI 10.1002/aic.14771
Published online March 11, 2015 in Wiley Online Library (wileyonlinelibrary.com)

Identifying disturbance covariances from data is a critical step in estimator design and controller performance monitoring. Here, the autocovariance least-squares (ALS) method for this identification is examined. For large industrial models with poorly observable states, the process noise covariance is high dimensional and the optimization problem is poorly conditioned. Also, weighting the least-squares problem with the identity matrix does not provide minimum variance estimates. Here, ALS method to resolve these two challenges is modified. Poorly observable states using the singular value decomposition (SVD) of the observability matrix is identified and removed, thus decreasing the computational time. Using a new feasible-generalized least-squares estimator that approximates the optimal weighting from data, the variance of the estimates is significantly reduced. The new approach on industrial data sets provided by Praxair is successfully demonstrated. The disturbance model identified by the ALS method produces an estimator that performs optimally over a year-long period. © 2015 American Institute of Chemical Engineers *AIChE J*, 61: 1840–1855, 2015
Keywords: model predictive control, disturbance model identification, least-squares

Introduction

Identifying an appropriate noise model is essential for designing and monitoring a model predictive controller. Knowledge of the disturbances affecting the process is necessary for optimal estimator design. The disturbance model must also be known to calculate a meaningful theoretical benchmark for controller performance monitoring purposes.[1]

The autocovariance least-squares (ALS) method estimates the disturbance covariance matrices from routine operating data[2,3] and is an improvement on earlier correlation techniques.[4,5] The ALS method uses the L-innovation autocovariances and the closed-loop model to form a least-squares optimization problem for the noise covariances. This method can also be used to estimate the optimal noises for integrating disturbances used to provide offset-free control.[6]

An alternative to the ALS technique, subspace identification (ID) methods identify both the process model and noise statistics.[7,8] These methods use least-squares regression to identify a characteristic subspace of the input–output data from which the system matrices and noise statistics are extracted.[9] These methods focus on determining the innovation covariance and the optimal estimator gain rather than estimating the driving noises.[8] Finding the driving process and measurement noises allows more flexibility in estimator design and provides more information toward performance monitoring.[3] In addition, because subspace ID methods identify the system matrices as well as the noise statistics, the input must be persistently exciting to accurately identify the input matrix.[7] Subspace ID methods also have not been used in the literature to identify the disturbance model for a system containing integrators. Such a method would require using a gray-box model to identify the system matrices.

In this article, two improvements to the ALS technique are proposed. First, a method of removing poorly observable states is presented. This improvement reduces the computational time while leading to equally accurate results. Second, a data-based weighting approach is presented as an alternative to the optimal weighting discussed in the literature.[3] This weighting significantly reduces the variance of ALS results from multiple data sets. These improvements are demonstrated on both simulated and industrial data sets.

Problem Formulation

We write our system and estimator as

$$x^+ = Ax + Gw \qquad \hat{x}^+ = A\hat{x} + AL\mathscr{y}$$
$$y = Cx + v \qquad \mathscr{y} = y - C\hat{x}$$

in which $x \in \mathbb{R}^n$ is the unmeasured state, $w \in \mathbb{R}^g$ is the process noise, $y \in \mathbb{R}^p$ is the measured output, $v \in \mathbb{R}^p$ is the measurement noise, $\hat{x} \in \mathbb{R}^n$ is the state estimate, and $\mathscr{y} \in \mathbb{R}^p$ is the L-innovation (or one-step-ahead prediction error). The system is described by the matrices $A \in \mathbb{R}^{n\times n}$ and $C \in \mathbb{R}^{p\times n}$. $L \in \mathbb{R}^{n\times p}$ is the estimator gain. $G \in \mathbb{R}^{n\times g}$ is the noise-shaping matrix; unless otherwise specified, we assume that $G = I$ (and therefore $g = n$) to better represent industrial systems where the structure of the process noise is entirely unknown. We assume that w and v are normally distributed white noises with zero mean and covariances Q_w and R_v, respectively. We also assume w and v are independent. As we do not assume that L is optimal, $\mathscr{y}$ is normally distributed but not necessarily white. Note that the state x may be augmented to include integrators.[6] We then define the ALS problem as

Correspondence concerning this article should be addressed to J. B. Rawlings at james.rawlings@wisc.edu.

$$\min_{Q_w,R_v} \phi = ||A\begin{bmatrix}(Q_w)_{ss}\\(R_v)_{ss}\end{bmatrix} - \hat{b}||^2 \quad \text{subject to } Q_w \geq 0, R_v \geq 0 \quad (1)$$

in which

$$\mathscr{A} = [\mathscr{A}_1 \quad \mathscr{A}_2]$$

$$\mathscr{A}_1 = (C \otimes \mathscr{O})(I_{n^2} - \bar{A} \otimes \bar{A})^{-1}(G \otimes G)\mathscr{D}_g$$

$$\mathscr{A}_2 = \Big((C \otimes \mathscr{O})(I_{n^2} - \bar{A} \otimes \bar{A})^{-1}(AL \otimes AL) + (I_p \otimes \Gamma)\Big)\mathscr{D}_p$$

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

and $\bar{A} := A - ALC$. The notation $(X)_{ss}$ refers to the column stacking of the lower triangular elements of X; the duplication matrix $\mathscr{D}_n$ is such that $\text{vec}(X) = \mathscr{D}_n(X)_{ss}$ for $n \times n$ matrix X. The vector $\hat{b}$ is the vectorized form of the estimated L-innovation autocovariances

$$\hat{b} := \text{vec}\left(\begin{bmatrix} \frac{1}{N_d}\sum_{i=1}^{N_d} \mathscr{Y}(i)\mathscr{Y}(i)' \\ \vdots \\ \frac{1}{N_d - N + 1}\sum_{i=1}^{N_d - N + 1} \mathscr{Y}(i+N-1)\mathscr{Y}(i)' \end{bmatrix}\right)$$

in which N_d is the number of data points. N is a user-defined parameter chosen such that the sample autocovariance is approximately zero for lags of N or larger ($N = 15$ was used for the examples studied here). Here, we consider the single column ALS formulation.[3] However, as shown in Appendix A, the uniqueness conditions for this form also apply to the full matrix ALS formulation.[2]

ALS Method for Unobservable and Weakly Observable Systems

Unobservable systems

In industrial settings, the use of large models with many unobservable or poorly observable states limits the applicability of the ALS method. Intuitively, we expect that an unobservable system does not have a unique ALS solution, as the noises affecting the unobservable states have no effect on the outputs; here, we prove that this intuition is correct.

Theorem 1. *Assume that $\bar{A}$ is stable and A is nonsingular. When G is unknown ($G = I$), the ALS problem has a unique solution ($\mathscr{A}$ is full column rank) if and only if (A, C) is observable and* rank$(C) = n$.

Proof. The proof that these conditions lead to a unique ALS estimate is given in Corollary 3.1 of Rajamani (2007).[10] ■

Next, we prove that (A, C) observable and rank$(C) = n$ are necessary conditions. First, consider (A, C) observable and rank$(C) = \bar{p} < n$. From Lemma 13 of Rajamani and Rawlings (2009).[3]*

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

Thus, for $n > \bar{p}$, null$(\mathscr{A})$ contains at least one nonzero vector, and therefore, the ALS problem does not have a unique solution.

Next, consider the case where (A, C) is unobservable. Let n_o and n_u be the number of observable and unobservable modes, respectively. We transform the system into observability canonical form so that

$$\bar{A} = \begin{bmatrix} \bar{A}_{11} & 0_{n_o \times n_u} \\ \bar{A}_{21} & \bar{A}_{22} \end{bmatrix} \quad C = [C_1 \quad 0_{p \times n_u}]$$

and therefore

$$(C \otimes \mathscr{O}) = [C_1 \otimes \mathscr{O} \quad 0_{Np^2 \times nn_u}]$$

$$\bar{A} \otimes \bar{A} = \begin{bmatrix} \bar{A}_{11} \otimes \bar{A} & 0_{nn_o \times nn_u} \\ \bar{A}_{21} \otimes \bar{A} & \bar{A}_{22} \otimes \bar{A} \end{bmatrix}$$

As $(I_{n^2} - \bar{A} \otimes \bar{A})^{-1}$ has a zero block in the same location as $(\bar{A} \otimes \bar{A})$, we have

$$\mathscr{A}_1 = \Big[(C \otimes \mathscr{O})(I_{n^2} - \bar{A} \otimes \bar{A})^{-1}\Big]\mathscr{D}_n = [\mathscr{A}_{11} \quad 0_{Np^2 \times nn_u}]\mathscr{D}_n$$

which loses column rank. Although multiplication by $\mathscr{D}_n$ reduces the number of columns, the matrix remains rank deficient. As the last column of $\mathscr{D}_n$ is $[0 \quad \ldots 0 \quad 1]'$, the last column of $\mathscr{A}_1$ must be zero. Thus, a unique solution does not exist. □

This theorem also applies to any square matrix G of rank n, as by performing a similarity transform, we can write an equivalent system with $G = I_n$.

Notes on the Assumptions. The condition that A is nonsingular means that full column rank $\mathscr{A}_1$ implies that $\mathscr{A}$ is full column rank. When A is singular, $\mathscr{A}_2$ is still full column rank, but $\mathscr{A}$ loses rank (when $G = I_n$). The stability of $\bar{A}$ is always necessary to ensure that $(I_n^2 - \bar{A} \otimes \bar{A})$ is invertible.

We can always reduce an unobservable system to an equivalent observable subsystem by first removing the unobservable states. Applying the ALS method to the observable subsystem gives the noise model with the fewest number of independent noises, as no noises affect the unobservable modes. Rajamani and Rawlings (2009) proposed the following optimization problem to find the solution with the smallest number of independent process noises

$$\min_{Q_w,R_v} \phi + \rho \text{tr}(Q_w) \quad \text{subject to } Q_w \geq 0, R_v \geq 0 \qquad (2)$$

in which ϕ is the least-squares objective function defined in (1) and ρ is a nonnegative scalar chosen by the user.[3] Here, we show that for any $\rho > 0$, the optimization problem (2) is equivalent for the full model and the observable subsystem.

Theorem 2. *For an unobservable system (A, B, C), let T be an orthogonal transformation matrix such that*

*The cited lemma assumes that $\bar{p} = p$. Here, we also cover the case in which the measurements are not linearly independent ($p > \bar{p}$). From Corollary 2 of Hua (1990),[11] which was used to derive the cited lemma, we can substitute $\bar{p}$ for p in the rank condition.

$$\tilde{A}=TAT'=\begin{bmatrix}A_{11} & 0\\ A_{21} & A_{22}\end{bmatrix}\ \tilde{B}=TB=\begin{bmatrix}B_1\\ B_2\end{bmatrix}\ \tilde{C}=CT'=[C_1 \quad 0]$$

Let ρ be any strictly positive scalar. Then, the optimization problem in (2) using the reduced model (A_{11}, B_1, C_1) and that using the original model (A, B, C) have the same objective function values and solutions R_v for the measurement noise. The optimal process noise covariances are related as

$$TQ_{w,\mathrm{opt}}T'=\begin{bmatrix}Q_{11,\mathrm{opt}} & 0\\ 0 & 0\end{bmatrix}$$

Proof. First, we note that there exists an orthogonal T to transform the system into observability canonical form. We construct an *invertible* transformation matrix $T=[T_1 \quad T_2]'$ by choosing the n_o columns of T_1 to be a basis for the range of $\mathcal{O}'$ and the n_u columns of T_2 to be a basis for null(T_1') [or equivalently, a basis for null$\mathcal{O}$].[12] As we can choose T_1 and T_2 as orthogonal bases, we can produce an orthogonal transformation for any unobservable system. ■

Next, we note the equivalence of the two systems

$$x^+=Ax+Bu+w \tag{3a}$$

$$y=Cx+v \tag{3b}$$

$$\tilde{x}^+=\tilde{A}\tilde{x}+\tilde{B}u+\tilde{w} \tag{4a}$$

$$\tilde{y}=\tilde{C}\tilde{x}+\tilde{v} \tag{4b}$$

Let $\tilde{w}(k)=Tw(k)$ and $\tilde{v}(k)=v(k)$ for $k\geq 0$. Provided that $\tilde{x}(0)=Tx(0)$, then $\tilde{x}(k)=Tx(k)$ and $y(k)=\tilde{y}(k)$ for all $k\geq 0$. The covariance for the process noise of the transformed system is

$$\mathrm{cov}(\tilde{w})=\mathrm{cov}(Tw)=TQ_wT'=\begin{bmatrix}Q_{11} & Q_{12}\\ Q_{21} & Q_{22}\end{bmatrix}\ \mathrm{cov}(\tilde{v})=\mathrm{cov}(v)=R_v$$

If $\tilde{L}=TL$, then the state estimates for (3) and (4) are also related through the similarity transform.[6]

As the unobservable states in (4) do not affect y, we equivalently write the system as

$$x_1^+=A_{11}x_1+B_1u+\tilde{w}_1 \tag{5a}$$

$$y=C_1x_1+v \tag{5b}$$

in which $\tilde{w}_1\sim N(0, Q_{11})$.
Define the matrices

$$Q_w^*=T'\tilde{Q}_w^*T\ \tilde{Q}_w^*=\begin{bmatrix}Q_{11} & Q_{12}\\ Q_{21} & Q_{22}\end{bmatrix} Q_w^o=T'\tilde{Q}_w^oT\ \tilde{Q}_w^o=\begin{bmatrix}Q_{11} & \hat{Q}_{12}\\ \hat{Q}_{21} & \hat{Q}_{22}\end{bmatrix}$$

in which (Q_w^*, R_v^*) is a specific minimizer of (1). We define $\tilde{Q}_w^o$ such that the submatrix Q_{11} is the same as that in $\tilde{Q}_w^*$ but the other blocks of $\tilde{Q}_w^o$ are arbitrarily chosen (provided that $\tilde{Q}_w^o\geq 0$). Q_w^o is a transformation of $\tilde{Q}_w^o$. As the system is unobservable, Q_{12} and Q_{22} have no effect on y and, therefore, $\mathscr{A}_1(Q_w^*)_{\mathrm{ss}}=\mathscr{A}_1(Q_w^o)_{\mathrm{ss}}$. Thus, there exist an infinite number of Q_w^o such that (Q_w^o, R_v^*) also minimizes (1).

Consider instead the solution to (2) for $\rho>0$. As Q_w and $\tilde{Q}_w$ are similar matrices, they have the same trace.[13] Because we require $Q_{22}\geq 0$, any solution Q_w which minimizes (2) is a transformation of $\tilde{Q}_w=\begin{bmatrix}Q_{11} & 0\\ 0 & 0\end{bmatrix}$, as choosing some $Q_{22}>0$ would increase $\mathrm{tr}(Q_w)$ without decreasing ϕ. The optimization problem (2), therefore, reduces to

$$\min_{Q_w,R_v}\ \phi+\rho\mathrm{tr}(Q_{11}) \quad \text{subject to } Q_{11}\geq 0, R_v\geq 0, Q_{12}=0, Q_{22}=0.$$

Alternatively, we apply the ALS method to the reduced system (5). As the L-innovations (and therefore their autocovariances) are identical for the full and the reduced systems, we have

$$\tilde{\mathscr{A}}_1(Q_{11})_{\mathrm{ss}}+\tilde{\mathscr{A}}_2(R_v)_{\mathrm{ss}}=\mathscr{A}_1(Q_w)_{\mathrm{ss}}+\mathscr{A}_2(R_v)_{\mathrm{ss}}$$

in which $\tilde{\mathscr{A}}_1$ and $\tilde{\mathscr{A}}_2$ are formed using the reduced model. Thus, both the least-squares part of the objective ϕ and the $\mathrm{tr}(Q_w)$ penalty are equal for the two systems. Both forms of the ALS problem have identical objective values and yield the same solution Q_{11} and R_v. □

Note on the Choice of Transformation. Even with the constraint of orthogonality, the choice of T is not unique (unless $n_o=n_u=1$). Therefore, there are multiple systems (A_{11}, B_1, C_1) that represent (A, B, C). Each system has a different optimal Q_{11}, but the process noise covariances are all similarity transformations of each other and the systems have identical objective function values.

Weakly observable systems

As discussed above, the unobservable states have no effect on the output. However, many industrial models include some states which have little effect on the output relative to the other states, and thus are difficult to observe from the outputs. We refer to these systems (states) as weakly observable systems (states). We identify these systems and states through the observability matrix. For a weakly observable system, the observability matrix is poorly conditioned and has at least one singular value that is close to zero. The weakly observable modes correspond to those singular values that are near zero.

We transform the system into observability canonical form as follows. Let $\mathcal{O}=USV'$ be the singular decomposition of the observability matrix, and choose $T=V'$. Then, the observability matrix of the transformed system is $\tilde{\mathcal{O}}=\mathcal{O}T'=US$. As the singular values are ordered largest to smallest, the norm of the columns of $\tilde{\mathcal{O}}$ decrease from left to right, and the modes of the transformed system go from most observable to least observable. The transformed system takes the form

$$\tilde{A}=TAT'=\begin{bmatrix}A_{11} & \delta A_{12}\\ A_{21} & A_{22}\end{bmatrix}\ \tilde{B}=TB=\begin{bmatrix}B_1\\ B_2\end{bmatrix}\ \tilde{C}=CT'=[C_1 \quad \delta C_2] \tag{6}$$

in which the magnitude of the scalar $\delta\geq 0$ depends on the magnitude of the singular values corresponding to weakly observable modes. If $\delta=0$, then the system is unobservable. By choosing an orthogonal transformation, the singular values and condition number are unaffected by transforming the system.

Lima, Rawlings, Rajamani, and Soderstrom (2013) also discuss applying the ALS method on systems with unobservable or weakly observable states by removing these states before solving the ALS problem.[14] However, they do not discuss in detail how to transform the system. They also do not compare the ALS problem for the full and reduced

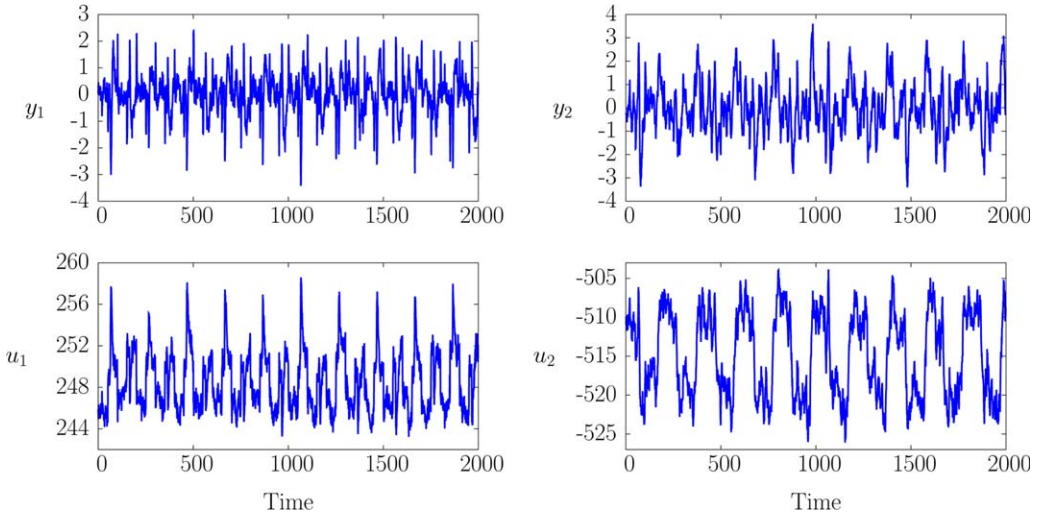

Figure 1. Inputs and outputs for example simulation.

Note that the outputs show the deviation from set point. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

models, and they only consider an example where G is known and there are few independent process noises ($g \leq 3$).

Note on Systems with Integrated Disturbances.. When the system is augmented with integrating disturbances to ensure offset-free control, it is essential that the integrator modes are maintained when the system is reduced. Although these modes are unaffected by similarity transforms (which maintain the same eigenvalues), the integrators may be lost when the weakly observable states are removed from the transformed model. To avoid this problem, we reduce the nonaugmented system and then add the integrators to the reduced model, that is, find (A_{11}, B_1, C_1) from the unaugmented (A, B, C), and then form

$$A_{\text{aug}} = \begin{bmatrix} A_{11} & B_d \\ 0 & I \end{bmatrix} \quad C_{\text{aug}} = [C_1 \quad C_d]$$

In addition, we penalize the trace of the unaugmented process noise covariance rather than that of the entire augmented covariance matrix to ensure that the integrators contain adequate noise.

Applying the ALS Method to Weakly Observable Systems. We summarize the method for applying the ALS method to poorly observable systems in the following steps:

1. Use singular value decomposition (SVD) on the observability matrix to obtain the transformation matrix $T = V'$, and then transform the system into observability canonical form.

2. Generate reduced models with the number of observable states ranging from $n_o = p$ to $n_o = n$, and augment the reduced models with integrators.

3. Solve the ALS problem on each augmented reduced model without penalizing the trace and without including the semidefinite constraints.†

4. Choose a reduced model that has a well-conditioned observability matrix but does not significantly increase the objective function value compared with the full model.

5. Using this reduced model, solve the ALS problem with semidefinite constraints and penalizing $\text{tr}(Q_w)$ as necessary.

6. Write the noise model for the full transformed model by assuming that no noise affects the weakly observable states.‡

7. Calculate the estimator gain for the full transformed system and convert to the original coordinates (or transform the process noise covariance matrix to the original coordinates and then calculate the estimator gain).

These steps are illustrated in a later example.

Feasible-Generalized ALS Technique

The standard ALS method, which we refer to as the "ordinary ALS" technique, uses the identity matrix to weight the least-squares problem. However, this weighting is chosen for practical reasons, and it does not produce minimum variance estimates for Q_w and R_v.[3] The minimum variance estimates are obtained from the generalized least-squares problem, where the variance of $\hat{b}$ is used as the weighting.[15,16] However, computing this variance has two major barriers.[3] First, calculating the variance is intractable for large sets of data, even if the dimensions of the state and output are small. This challenge arises because the fourth moment of the entire vector $\left[\mathcal{Y}(1)' \ldots \mathcal{Y}(N_d)'\right]'$ must be computed. Second, calculating this variance requires knowledge of Q_w and R_v, the unknowns to be found. Rajamani and Rawlings (2009) proposed iteratively solving the ALS problem for Q_w and R_v and then updating the weighting based on these values and resolving the ALS problem.[3] However, this iterative scheme is not guaranteed to converge. Therefore, rather than addressing the tractability of computing the optimal weighting, we propose using a feasible-generalized least-squares method to approximate it. The feasible-generalized least-squares method that uses an approximation of the variance is used to weight the least-squares problem.[16] We apply feasible-generalized least-squares to the ALS problem as follows:

Let S denote the covariance of $\hat{b}$ and $W = S^{-1}$ be the optimal weighting for the least-squares problem.

†We recommend solving the simpler problem here rather than the complete ALS problem as described in (2) to reduce the computation time.

‡Because the reduced model may not be sufficiently accurate for predictions over a longer horizon, it is recommended to continue to use the original model in the regulator. The question of whether or not the original model contains unnecessary states for the regulator problem is outside of the scope of this work.

We estimate S by the steps:

1. Let $t = 2N$ and $N_s = \frac{N_d - N + 1}{t}$. Then let

$$\mathbb{Y}_1 = \begin{bmatrix} \mathscr{y}_1 & \mathscr{y}_{2N+1} & \cdots & \mathscr{y}_{N_d-3N} \\ \mathscr{y}_2 & \mathscr{y}_{2N+2} & \cdots & \mathscr{y}_{N_d-3N+1} \\ \vdots & \vdots & \vdots & \vdots \\ \mathscr{y}_N & \mathscr{y}_{3N} & \cdots & \mathscr{y}_{N_d-2N+1} \end{bmatrix}$$

$$\mathbb{Y}_2 = \begin{bmatrix} \mathscr{y}_2 & \mathscr{y}_{2N+2} & \cdots & \mathscr{y}_{N_d-3N+1} \\ \mathscr{y}_3 & \mathscr{y}_{2N+3} & \cdots & \mathscr{y}_{N_d-3N+2} \\ \vdots & \vdots & \vdots & \vdots \\ \mathscr{y}_{N+1} & \mathscr{y}_{3N+1} & \cdots & \mathscr{y}_{N_d-2N+2} \end{bmatrix}$$

$$\vdots$$

$$\mathbb{Y}_t = \begin{bmatrix} \mathscr{y}_{2N} & \mathscr{y}_{4N} & \cdots & \mathscr{y}_{N_d-N+1} \\ \mathscr{y}_{2N+1} & \mathscr{y}_{4N+1} & \cdots & \mathscr{y}_{N_d-N+2} \\ \vdots & \vdots & \vdots & \vdots \\ \mathscr{y}_{3N-1} & \mathscr{y}_{5N-1} & \cdots & \mathscr{y}_{N_d} \end{bmatrix}$$

(For brevity, we show the time-index of $\mathscr{y}$ using a subscript rather than parentheses). As we assume that $\mathscr{y}_k$ and $\mathscr{y}_{k+N+i}$ are uncorrelated for $i \geq 0$, each $\mathbb{Y}_i$ is composed of columns that are approximately independent.

2. Let $\hat{P}_{y,i}$ be the sample variance of the columns of $\mathbb{Y}_i$. Then, each $\hat{P}_{y,i}$ gives an approximately unbiased approximation for $P_y := \mathrm{var}\left(\begin{bmatrix} \mathscr{y}_k' & \cdots & \mathscr{y}_{k+N-1}' \end{bmatrix}'\right)$ with the only bias due to the slight correlations between the columns. We approximate P_y as $\hat{P}_y = \frac{1}{N_s}\sum_{i=1}^{N_s}\left(\hat{P}_{y,i}\right)$.

3. Let $P_0 := \mathrm{var}(\mathscr{y}_k)$ and $P_{y,0} := \mathrm{cov}\left(\begin{bmatrix} \mathscr{y}_k \\ \vdots \\ \mathscr{y}_{k+N-1} \end{bmatrix}, \mathscr{y}_k\right)$. Then, we approximate $\hat{P}_0$ as the first $p \times p$ submatrix of $\hat{P}_y$ and $\hat{P}_{y,0}$ as the first p columns of $\hat{P}_y$.

4. Let S be the covariance of $\hat{b}$. Based on the Wishart distribution, we approximate S as

$$\hat{S} = \frac{1}{N_s}\left(\left(\hat{P}_0 \otimes \hat{P}_y\right) + K_{p,\vec{p}}\left(\hat{P}_{y,0} \otimes \hat{P}_{y,0}'\right)\right) \tag{7}$$

and find $\hat{W}$ as the inverse of $\hat{S}$. Equation (7) is derived in Appendix B.

We require at least N columns in each $\mathbb{Y}_i$ to compute the sample variance. Therefore, we need the number of data points to satisfy $N_d \geq 2N^2p + N - 1$. An alternative method to approximate S would be to divide the data into several smaller samples of length $N_s < N_d$, calculate $\hat{b}_i$ for each sample, and let $\hat{S}$ be the sample variance of $\hat{b}_i$. However, this approximation requires the number of data points to be on the order of N^2p^2 and does not produce independent samples of $\hat{b}$. Simulations indicate that this approximation is less effective at decreasing the variance of the ALS estimates compared with the approximation method outlined earlier.

Examples[§]

Example 1—Weakly observable systems and model reduction

We demonstrate the benefit of reducing the model to include only observable states by studying the system

$$A = \begin{bmatrix} 4.2\times10^{-17} & 0.15 & 0 & 0 & 0 & 0 & 0 \\ -0.1 & 0.84 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -4.2\times10^{-17} & 0.15 & 0 & 0 & 0 \\ 0 & 0 & -0.1 & 0.84 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.8 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1.1\times10^{-16} & 0.64 \\ 0 & 0 & 0 & 0 & 0 & -1 & 1.6 \end{bmatrix}$$

$$B = \begin{bmatrix} -0.78 & 0 \\ 0.28 & 0 \\ 0 & 0.39 \\ 0 & -0.14 \\ 0.2 & 0 \\ 0 & 0.017 \\ 0 & -0.019 \end{bmatrix} \quad C = \begin{bmatrix} 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 \end{bmatrix}$$

The observability matrix has condition number 4.78×10^{17} and singular values

$$\sigma = \begin{bmatrix} 6.6 & 2.4 & 1.3 & 0.14 & 0.00058 & 3.5\times10^{-7} & 1.4\times10^{-17} \end{bmatrix}$$

The mode corresponding to the smallest singular value is clearly unobservable, but the singular values alone do not indicate whether any additional modes can be removed from the system. Using SVD, we transformed the system into observability canonical form

$$\tilde{A} = \begin{bmatrix} 1 & 8.6e\text{-}18 & -0.062 & 4.1e\text{-}17 & -3.1e\text{-}18 & -2.1e\text{-}08 & -3.4e\text{-}19 \\ 2.6e\text{-}17 & 0.83 & -6.6e\text{-}18 & 0.034 & -2.5e\text{-}05 & 4e\text{-}23 & 2.1e\text{-}17 \\ 0.95 & -8e\text{-}18 & 0.56 & 2.1e\text{-}17 & -2e\text{-}17 & 2.2e\text{-}07 & -7.9e\text{-}19 \\ -2.6e\text{-}17 & -0.22 & -1.3e\text{-}18 & 0.0093 & -0.0033 & 2.5e\text{-}17 & 6.1e\text{-}17 \\ 4.6e\text{-}17 & -0.001 & -2.1e\text{-}17 & -0.0032 & 0.82 & -9.4e\text{-}18 & 1.7e\text{-}15 \\ -1.3 & -2.4e\text{-}15 & 0.32 & -1.5e\text{-}14 & 3.8e\text{-}12 & 0.8 & 1.2e\text{-}11 \\ -1.9e\text{-}11 & 0.00016 & 4.9e\text{-}12 & 0.001 & -0.25 & 1.2e\text{-}11 & 0.018 \end{bmatrix}$$

$$\tilde{C} = \begin{bmatrix} -1.3e\text{-}17 & -1.4 & 3.5e\text{-}17 & 0.11 & 6.2e\text{-}06 & -1.9e\text{-}23 & -2.8e\text{-}17 \\ 0.92 & 2.5e\text{-}17 & 1.1 & 6.9e\text{-}18 & 4.2e\text{-}17 & -1.6e\text{-}07 & -1e\text{-}17 \end{bmatrix}$$

and augmented the model with integrators on the inputs. We generated the data shown in Figure 1 by simulating the

[§]The ALS toolbox for Octave or Matlab was used in these examples and is available online at http://jbrwww.che.wisc.edu/software/als/. This toolbox has been updated to include the feasible-generalized ALS method.

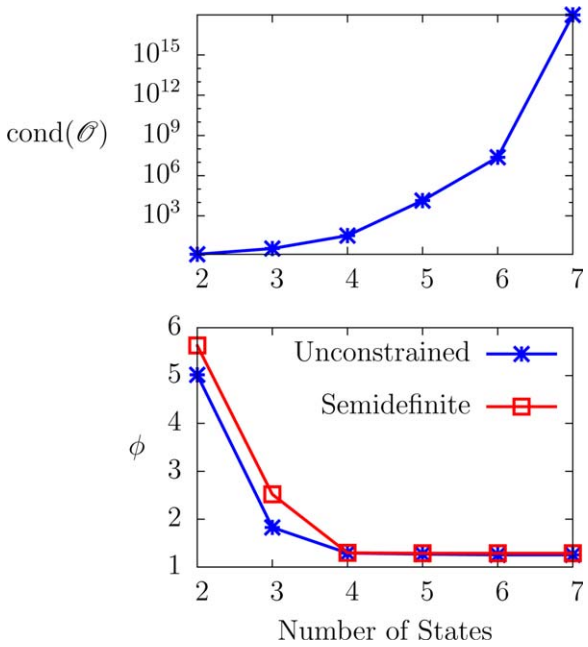

Figure 2. Condition number and ALS objective vs. number of states for the poorly observable model.

As more states are included in the model, the condition number increases and the ALS objective value decreases. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

system in closed-loop control against white noise disturbances added to the states and outputs as well as repeated step disturbances to the inputs.

Next, we formed reduced models from the canonical form, letting n_o range from $p = 2$ to $n = 7$. Using the simulated data, we compared the condition number of the observability matrix and the ALS objective for each of the models, as shown in Figure 2.

The condition number of the observability matrix increases as we include more states, with the most dramatic change when we increase the number of states from six to seven, as is expected from the singular values of the full observability matrix. The ALS objective function value decreases significantly when we include three states. It also decreases slightly for four states, but adding the last three states has no noticeable effect on the objective value. This behavior is also consistent with the singular values of the observability matrix.

Figure 2 illustrates that adding the semidefinite constraints does not affect the choice of reduced model size. However, we did use the feasible-generalized ALS method to avoid placing too much emphasis on matching zero covariances. The full ALS problem, including penalizing $\mathrm{tr}(Q)$ and the semidefinite constraints, was solved after the model size was chosen.

We compared the accuracy of the noise covariances estimated from the models with $n_o = 2,4$, and 7. Because of the step disturbances, we do not have a true theoretical value for Q_w against which to compare the ALS estimate. Instead, we examined the quality of the estimator produced by the ALS estimates. We first designed a new estimator from $\hat{Q}_w$ and $\hat{R}_v$ and computed the innovations using this estimator. We then examined whether the resulting innovations are white, as an optimal estimator produces white innovations, meaning that the autocovariances and cross-covariances of the innovations are significantly greater than zero only at lag zero.

As shown in Figure 3, the autocovariances of the two-state model remain significantly above zero for lags greater than zero. This behavior indicates that the system is

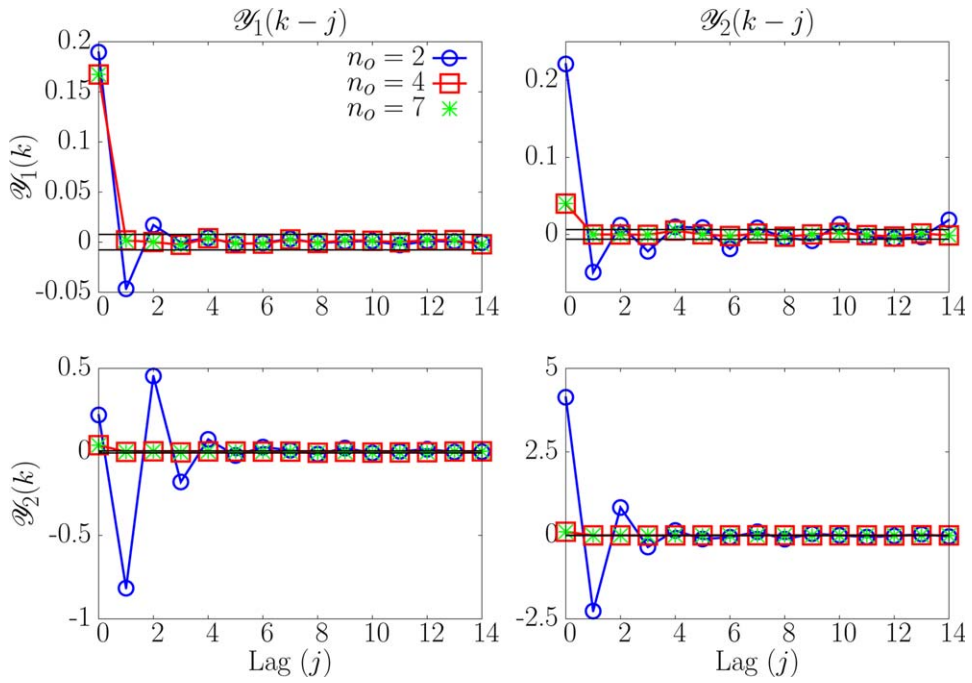

Figure 3. Innovation autocovariances using estimators designed from the ALS results for 2, 4, and 7 states.

Both the four- and seven-state models produce white innovations. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Table 1. $\hat{Q}_w$ and $\hat{R}_v$ Estimated by the ALS Method for Models Containing Two, Four, and Seven States

$n_o = 2$

$$\hat{Q}_w = \begin{bmatrix} 6.59\text{e-}02 & -1.09\text{e-}02 & 1.96\text{e+}00 & 3.58\text{e+}00 \\ -1.09\text{e-}02 & 1.43\text{e-}02 & -1.50\text{e+}00 & -3.06\text{e+}00 \\ 1.96\text{e+}00 & -1.50\text{e+}00 & 1.69\text{e+}02 & 3.39\text{e+}02 \\ 3.58\text{e+}00 & -3.06\text{e+}00 & 3.39\text{e+}02 & 6.81\text{e+}02 \end{bmatrix}$$

$$\hat{R}_v = \begin{bmatrix} 6.81\text{e-}10 & -2.08\text{e-}18 \\ -2.08\text{e-}18 & 8.08\text{e-}03 \end{bmatrix}$$

$n_o = 4$

$$\hat{Q}_w = \begin{bmatrix} 3.19\text{e-}03 & 3.50\text{e-}03 & 2.16\text{e-}03 & -3.37\text{e-}03 & 1.76\text{e-}02 & 1.99\text{e-}02 \\ 3.50\text{e-}03 & 1.01\text{e-}02 & 3.18\text{e-}03 & -4.75\text{e-}03 & -8.03\text{e-}03 & -2.32\text{e-}02 \\ 2.16\text{e-}03 & 3.18\text{e-}03 & 1.67\text{e-}03 & -2.59\text{e-}03 & 4.10\text{e-}03 & 1.50\text{e-}02 \\ -3.37\text{e-}03 & -4.75\text{e-}03 & -2.59\text{e-}03 & 4.05\text{e-}03 & -6.46\text{e-}03 & -2.65\text{e-}02 \\ 1.76\text{e-}02 & -8.03\text{e-}03 & 4.10\text{e-}03 & -6.46\text{e-}03 & 4.08\text{e-}01 & -2.13\text{e-}02 \\ 1.99\text{e-}02 & -2.32\text{e-}02 & 1.50\text{e-}02 & -2.65\text{e-}02 & -2.13\text{e-}02 & 1.01\text{e+}00 \end{bmatrix}$$

$$\hat{R}_v = \begin{bmatrix} 1.90\text{e-}02 & 4.24\text{e-}17 \\ 4.24\text{e-}17 & 2.74\text{e-}02 \end{bmatrix}$$

$n_o = 7$

$$\hat{Q}_w = \begin{bmatrix} \mathbf{3.15e\text{-}03} & \mathbf{3.48e\text{-}03} & \mathbf{2.16e\text{-}03} & \mathbf{-3.39e\text{-}03} & -1.10\text{e-}05 & 7.41\text{e-}09 & 1.60\text{e-}18 & \mathbf{1.77e\text{-}02} & \mathbf{2.02e\text{-}02} \\ \mathbf{3.48e\text{-}03} & \mathbf{1.01e\text{-}02} & \mathbf{3.21e\text{-}03} & \mathbf{-4.81e\text{-}03} & -8.37\text{e-}06 & 3.59\text{e-}08 & 8.59\text{e-}18 & \mathbf{-7.82e\text{-}03} & \mathbf{-2.22e\text{-}02} \\ \mathbf{2.16e\text{-}03} & \mathbf{3.21e\text{-}03} & \mathbf{1.69e\text{-}03} & \mathbf{-2.64e\text{-}03} & -3.91\text{e-}06 & 2.57\text{e-}08 & 2.01\text{e-}18 & \mathbf{4.25e\text{-}03} & \mathbf{1.54e\text{-}02} \\ \mathbf{-3.39e\text{-}03} & \mathbf{-4.81e\text{-}03} & \mathbf{-2.64e\text{-}03} & \mathbf{4.15e\text{-}03} & 5.58\text{e-}06 & -4.31\text{e-}08 & -2.87\text{e-}18 & \mathbf{-6.72e\text{-}03} & \mathbf{-2.75e\text{-}02} \\ -1.10\text{e-}05 & -8.37\text{e-}06 & -3.91\text{e-}06 & 5.58\text{e-}06 & 1.15\text{e-}06 & 5.39\text{e-}10 & -4.05\text{e-}20 & -2.17\text{e-}04 & 1.42\text{e-}04 \\ 7.41\text{e}-09 & 3.59\text{e-}08 & 2.57\text{e-}08 & -4.31\text{e-}08 & 5.39\text{e-}10 & 1.01\text{e-}06 & -2.86\text{e-}21 & -8.32\text{e-}07 & 1.14\text{e-}06 \\ 1.60\text{e}-18 & 8.59\text{e-}18 & 2.01\text{e-}18 & -2.87\text{e-}18 & -4.05\text{e-}20 & -2.86\text{e-}21 & 1.01\text{e-}06 & -2.67\text{e-}17 & -2.85\text{e-}17 \\ \mathbf{1.77e\text{-}02} & \mathbf{-7.82e\text{-}03} & \mathbf{4.25e\text{-}03} & \mathbf{-6.72e\text{-}03} & -2.17\text{e-}04 & -8.32\text{e-}07 & -2.67\text{e-}17 & \mathbf{4.09e\text{-}01} & \mathbf{-2.09e\text{-}02} \\ \mathbf{2.02e\text{-}02} & \mathbf{-2.22e\text{-}02} & \mathbf{1.54e\text{-}02} & \mathbf{-2.75e\text{-}02} & 1.42\text{e-}04 & 1.14\text{e-}06 & -2.85\text{e-}17 & \mathbf{-2.09e\text{-}02} & \mathbf{1.01e+00} \end{bmatrix}$$

$$\hat{R}_v = \begin{bmatrix} 1.90\text{e-}02 & 8.45\text{e-}17 \\ 8.45\text{e-}17 & 2.74\text{e-}02 \end{bmatrix}$$

Note that the bolded elements in the seven states $\hat{Q}_w$ are the same as $\hat{Q}_w$ for four states, while the remaining seven-state elements are near zero.

undermodeled. In contrast, the estimators for $n_o = 4$ and $n_o = 7$ both behave optimally. The estimator performance is unchanged when all seven states are included rather than only four states.

Finally, we compare $\hat{Q}_w$ and $\hat{R}_v$ for each of the three models, shown in Table 1. The four- and seven-state models produce approximately the same results. In the seven-state model, the elements of $\hat{Q}_w$ corresponding to the three poorly observable modes are approximately zero. In contrast to the four- and seven-state solutions, the two-state model produces completely different results for both $\hat{Q}_w$ and $\hat{R}_v$.

Although the four- and seven-state models produce identical results, the advantage of using the smaller observable model lies in the computational time required to solve the ALS problem. By reducing the model from seven to four

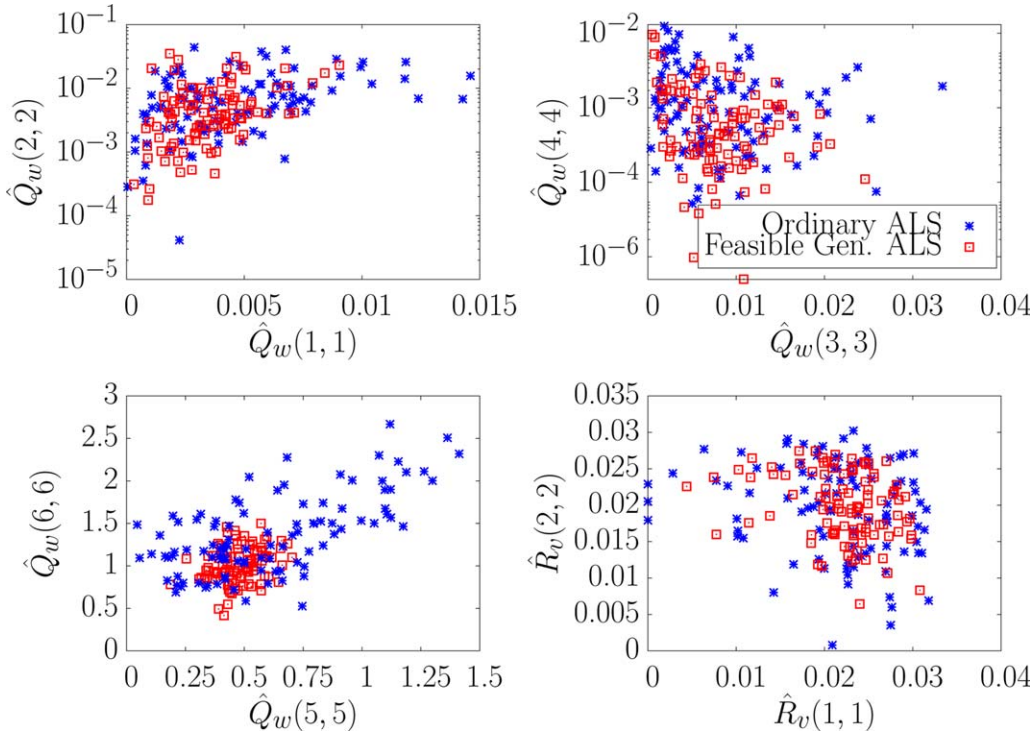

Figure 4. Noise variance estimates with data-based weighting and identity weighting.

The variance is noticeably reduced using the feasible-generalized ALS technique. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

states, the computational time was reduced from 5.98 to 3.63 s. Simulations indicate that the time to solve the ALS problem depends on the conditioning of the observability matrix as well as the number of states. An ill-conditioned problem also makes it difficult to choose appropriate stopping criteria for the optimizer; as a result, the optimizer may terminate before the true minimum is reached. A better conditioned system is less susceptible to this problem. For larger systems, model reduction can eliminate hours of computational time.

Example 2—Comparison of ordinary ALS and feasible-generalized ALS

In the previous example, we used the feasible-generalized ALS method. Here, we compare the feasible-generalized ALS method to the ordinary ALS method using the four-state model of the system. We generate multiple sets of data and apply both the feasible-generalized ALS and ordinary ALS methods to each dataset. For simplicity in presenting the results, we show only the diagonal elements of $\hat{Q}_w$ and $\hat{R}_v$ (although $\hat{Q}_w$ is nondiagonal) for the reduced four-state model. Note that $\hat{Q}_w$ includes the noises on the integrating disturbances. Figure 4 shows the diagonal elements of the noise covariance matrices estimated from each data set; their

Table 2. Variance of Each Diagonal Element of $\hat{Q}_w$ and $\hat{R}_v$

	Data-Based Weighting	Identity Weighting
$\hat{Q}_w(1,1)$	3.04×10^{-6}	9.90×10^{-6}
$\hat{Q}_w(2,2)$	4.92×10^{-5}	8.04×10^{-5}
$\hat{Q}_w(3,3)$	2.29×10^{-5}	4.34×10^{-5}
$\hat{Q}_w(4,4)$	5.08×10^{-5}	1.47×10^{-4}
$\hat{Q}_w(5,5)$	9.30×10^{-3}	1.10×10^{-1}
$\hat{Q}_w(6,6)$	4.53×10^{-2}	2.12×10^{-1}
$\hat{R}_v(1,1)$	2.55×10^{-5}	5.31×10^{-5}
$\hat{R}_v(2,2)$	2.23×10^{-5}	4.10×10^{-5}

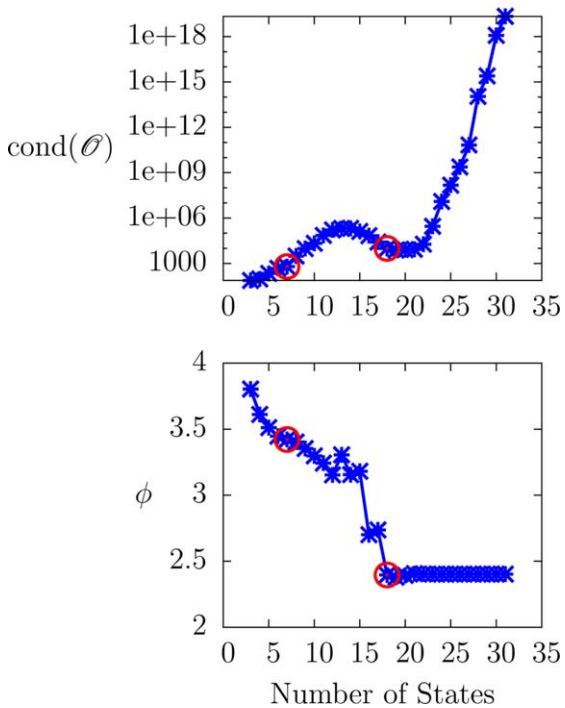

Figure 5. Condition number and ALS objective vs. number of states for the industrial example.

Seven states were chosen to have a balance between the condition number, number of states, and ALS objective function value. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

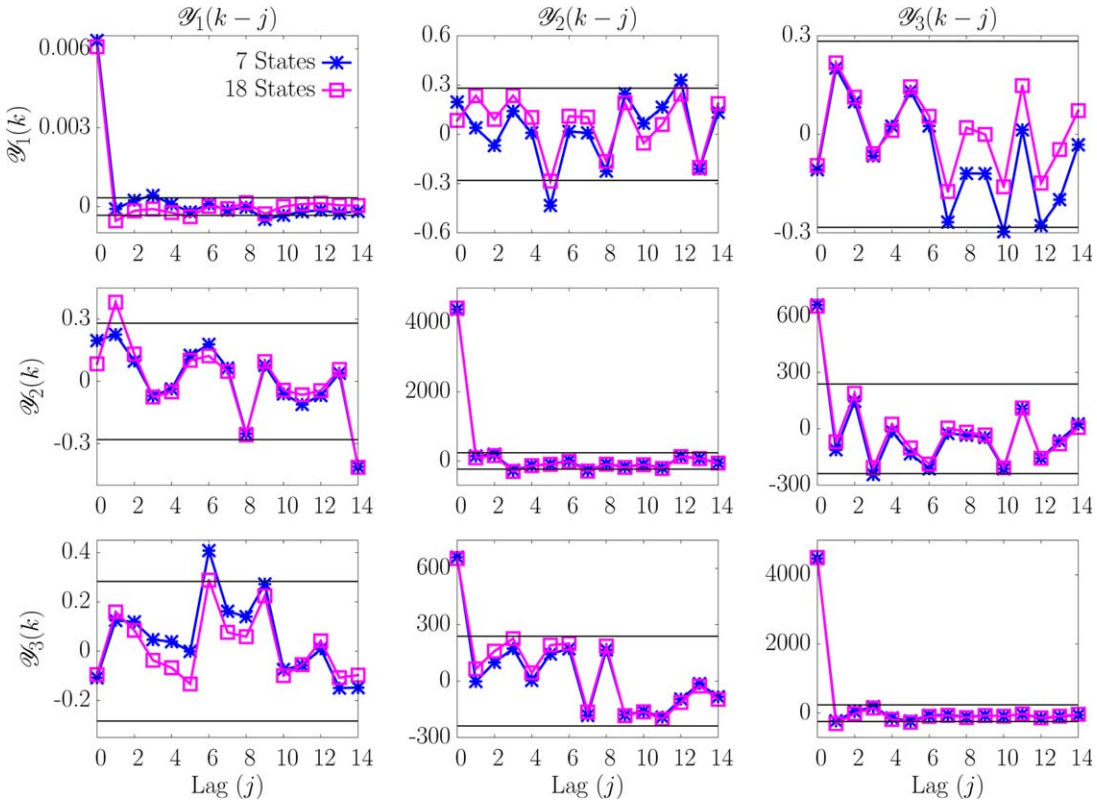

Figure 6. Innovation autocovariances for the seven- and 18-state models.

Both models produce excellent estimators when tuned by the ALS results. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

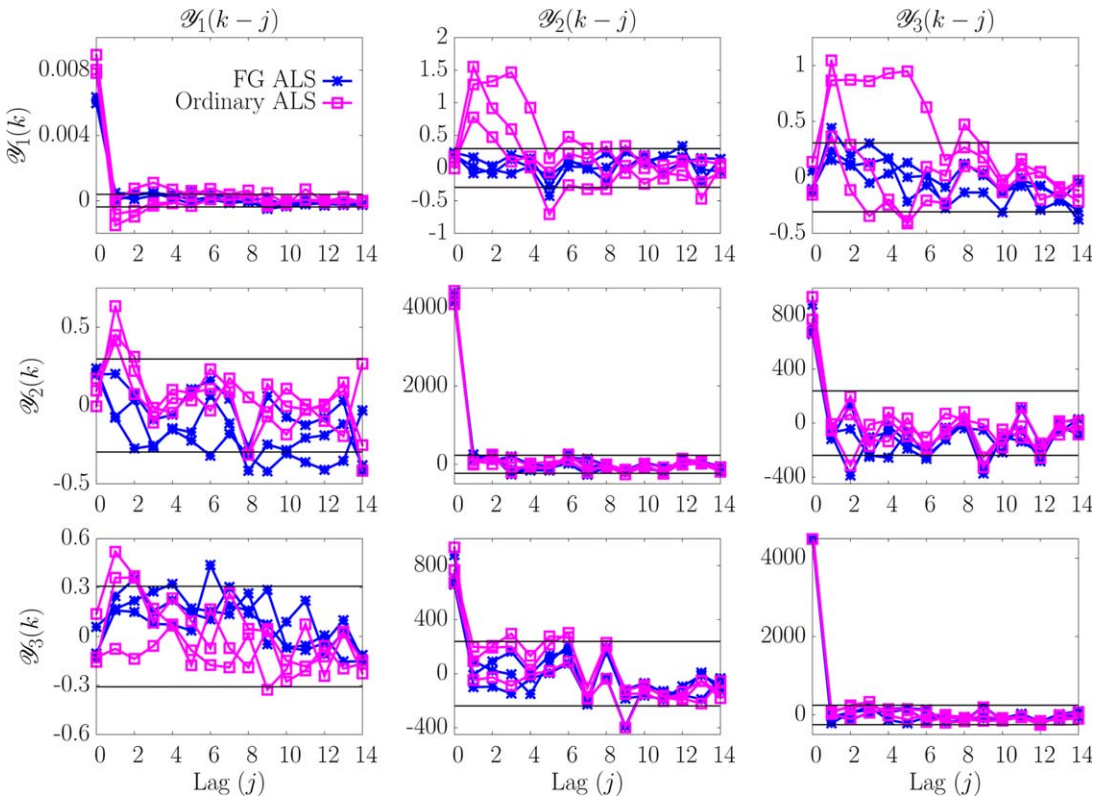

Figure 7. Innovation autocovariances for the feasible-generalized and ordinary ALS methods.

The feasible-generalized ALS method results in smaller autocovariances and cross-covariances for y_1. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

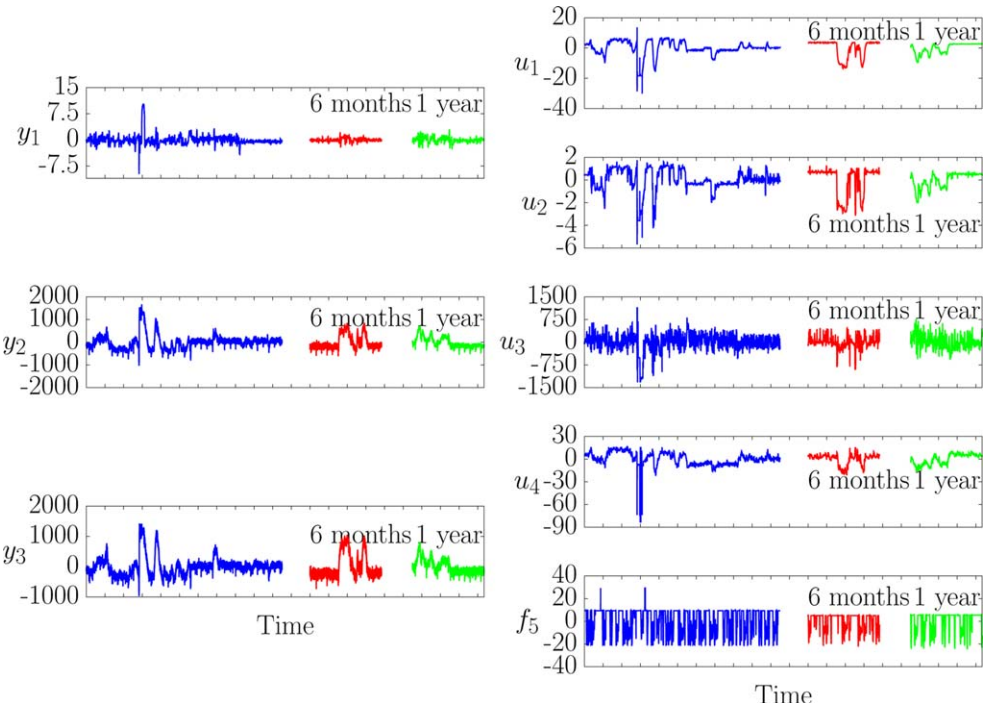

Figure 8. Industrial data analyzed in this work.

[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

variances are presented in Table 2. These results show that the feasible-generalized ALS technique significantly reduces the variance of the estimates.

Example 3—Industrial data set

In this example, we apply the ALS method to an air separation unit operated by Praxair. We analyze a subset of the

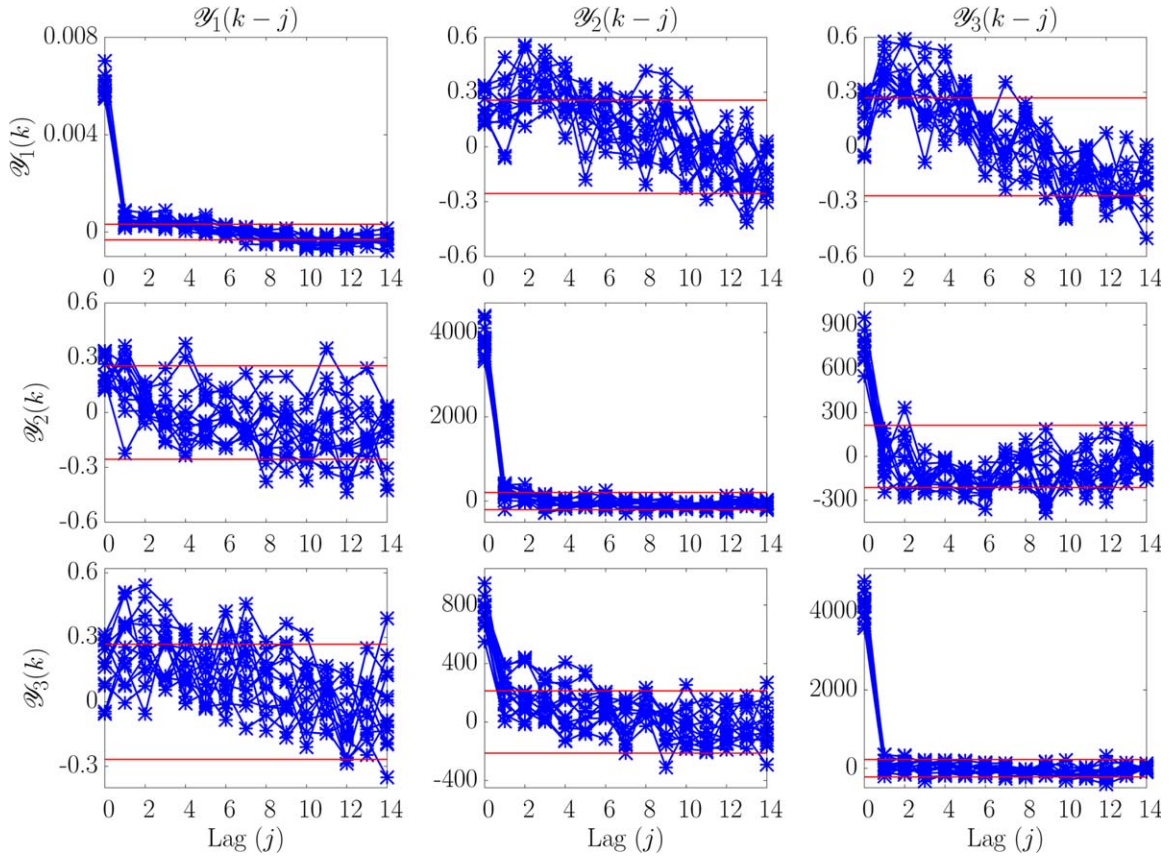

Figure 9. Autocovariances for L_1 innovations of data from the first time period.

[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

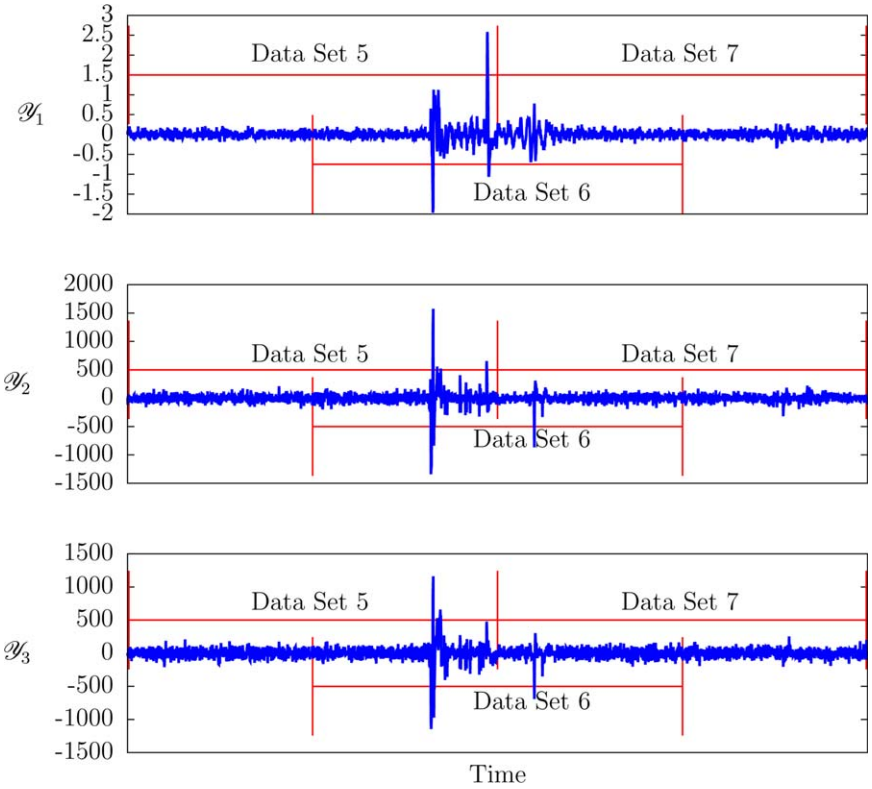

Figure 10. Innovations for data sets 5–7 in the first time period. A large disturbance is evident in the innovations.

[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

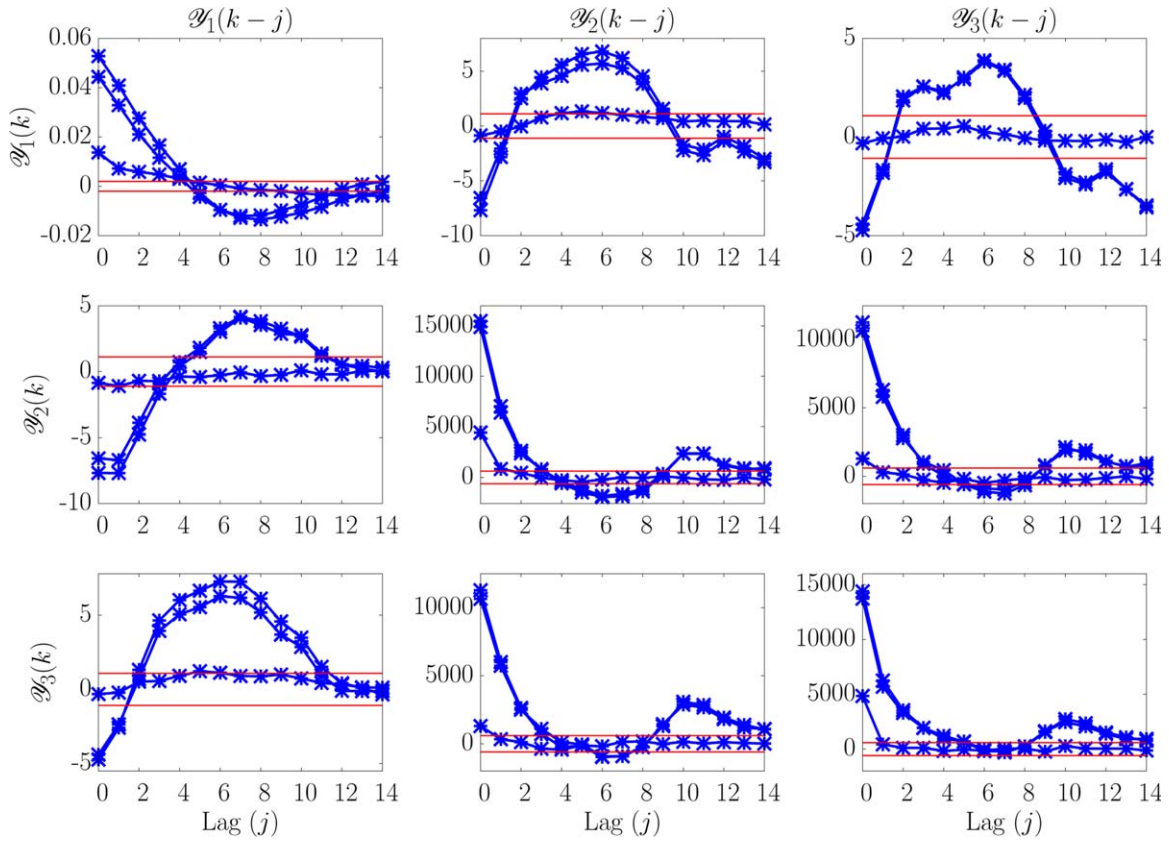

Figure 11. Autocovariance of innovations for data sets 5–7 in the first time period.

Due to the unexplained disturbance, the estimator is clearly suboptimal. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

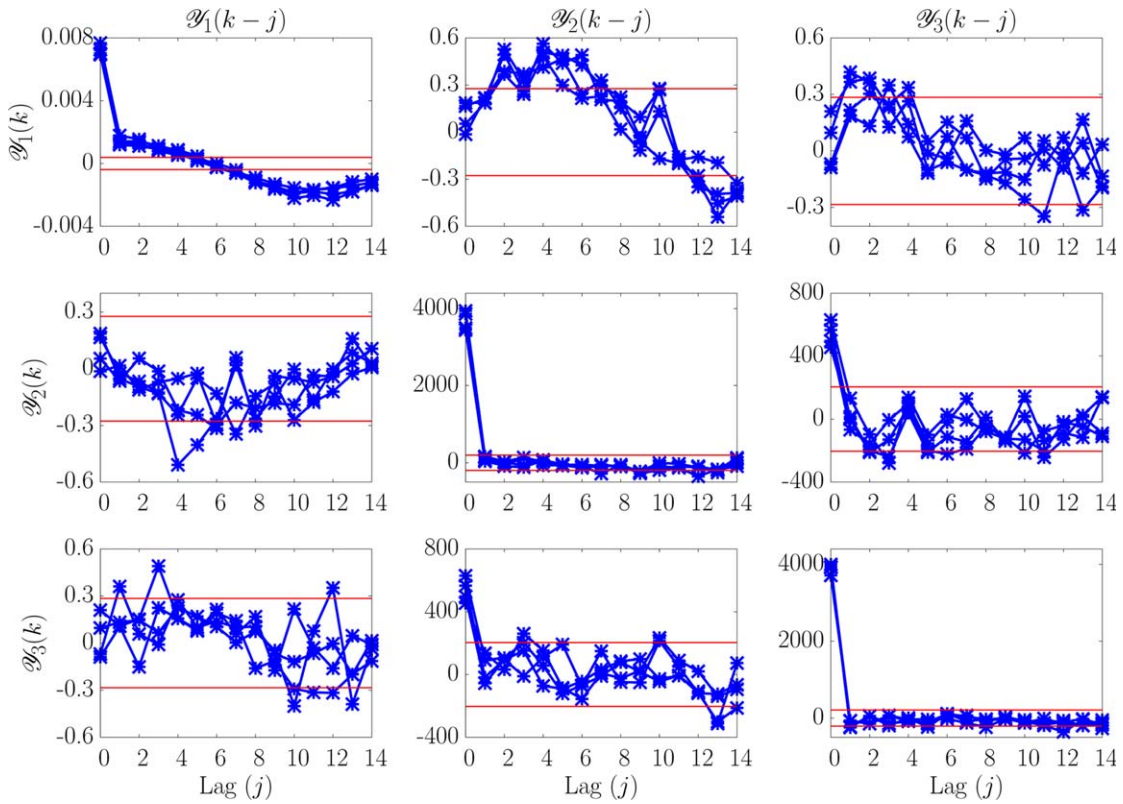

Figure 12. Autocovariance of innovations for data sets 17–20 in the first time period.

The data were processed with L_1 which is slightly suboptimal for these data. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

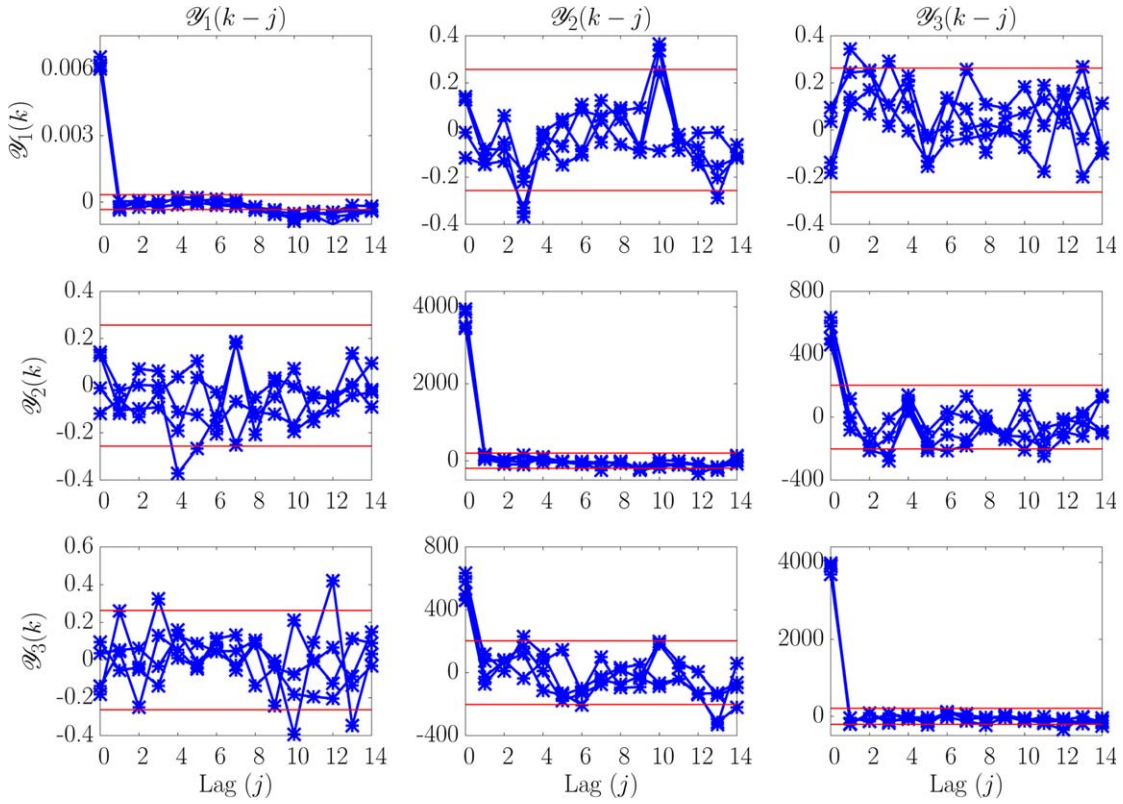

Figure 13. Autocovariance of innovations for data sets 17–20 in the first time period using a reidentified noise model.

By solving the ALS problem using data set 18, a more accurate noise model was identified for these data sets. The estimator designed from this noise model behaves optimally. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

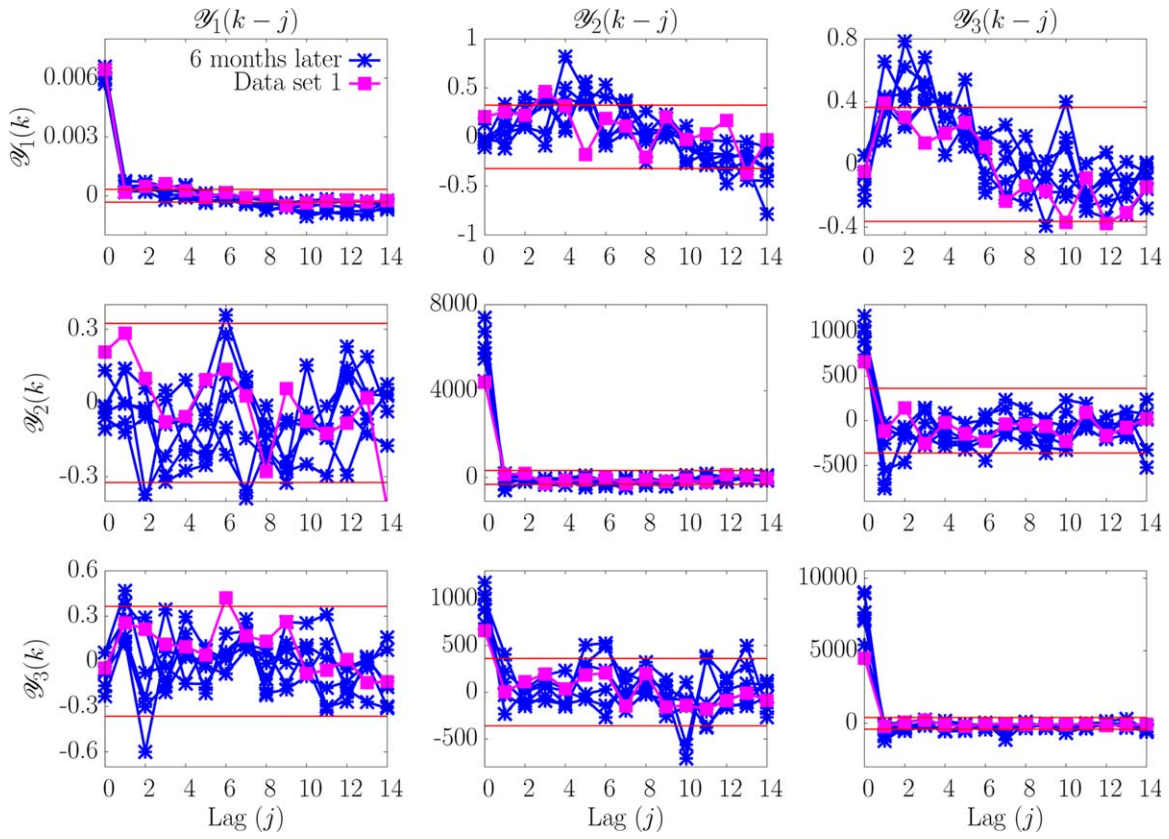

Figure 14. Autocovariance of innovations for data sets collected 6 months later.

The estimator L_1 performs nearly optimally on these data. For comparison, the autocovariances for data set 1 (from the first time period) are presented as well. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

variables used by the model predictive control (MPC), consisting of three outputs, four inputs, and one feed-forward variable. When solving the ALS problem, the feed-forward variable is handled in the same manner as the manipulated variables, as both types of inputs are measured.

The MPC uses an finite impulse response (FIR) model for the system. To obtain a state space model, transfer functions were fit to each input–output step response. These transfer functions were discretized, converted into state space, and combined to produce a single state space model for the system.

The variances of the outputs are different orders of magnitude; the variance of y_2 and y_3 are in the order of 10^4, whereas y_1 has a variance around 1. Therefore, we normalized the data by dividing each output and each row of C by the standard deviation of y_i. The state, input, and process noise remain the same in the normalized model. The ALS estimate of $\hat{Q}_w$ applies to both the normalized and original data, and the estimate of $\hat{R}_v$ corresponds to the normalized outputs. As we assume R_v is diagonal, we convert the estimated covariance back to the original scaling by multiplying each diagonal entry by the variance of the original output.

The scaled state space model was augmented with integrated disturbances to the outputs. However, analysis of the data showed that this disturbance model is insufficient for the ALS results to produce an optimal estimator for y_1. Instead, a double integrator model of the form

$$d^+ = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} d + w_d \qquad y_1 = C_1 x + \begin{bmatrix} 1 & 0 \end{bmatrix} d$$

was added to y_1. Simulations demonstrate that this type of disturbance model is more effective at estimating ramp-like or sinusoidal disturbances.

The full state space model contains 31 states and is unobservable, with a condition number of 3.24×10^{15}. Figure 5 shows the condition number and ALS objective function value vs. the number of states. Although the general trend is as expected, the shape of the curve does not give an obvious choice as to the number of states that should be retained. We compared the model with $n_o = 7$, which corresponds to a flattening in both curves before the condition number rises again, to the model with $n_o = 18$, which corresponds to the number of states after which ϕ no longer drops. The condition numbers are 611 and 9.67×10^3, respectively.

We compared the two models by solving the ALS problem and then designing an estimator based on the results for each model. We processed the data with each of these estimators and computed the autocovariances. Figure 6 shows that both estimators have near-optimal performance. However, the computation time increased from 19.5 to 487 s as the number of states increased. Therefore, the seven-state model was selected, as it gives nearly the same results in a much shorter time.

Using the seven-state model, we compared the feasible-generalized ALS technique with the ordinary ALS technique using each method on three data sets. The three sets overlap in the manner as shown in Figure 10. As seen in Figure 7, the feasible-generalized ALS method reduces the variance of the innovations for y_1 and also reduces the cross-covariance between y_1 and the other outputs at higher lags.

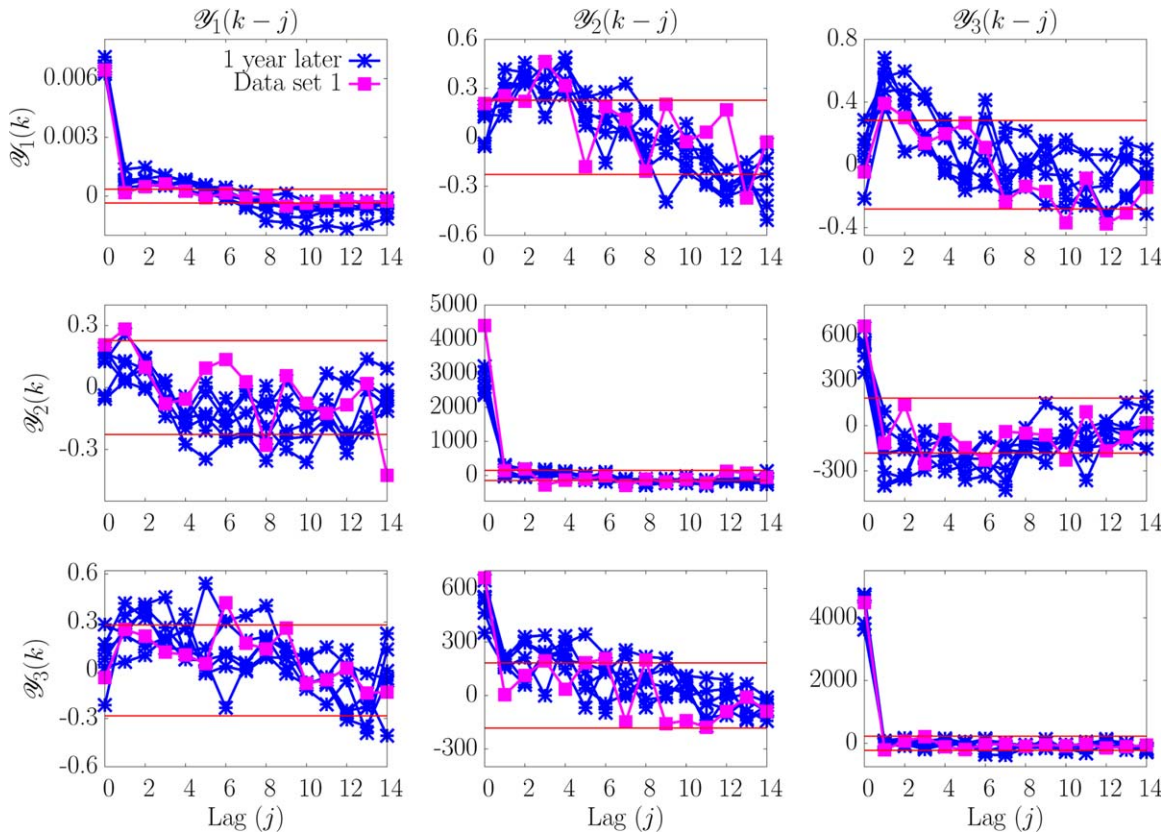

Figure 15. Autocovariance of innovations for data sets collected 1 year later.

Again, the estimator L_1 performs nearly optimally on these data. Slightly better performance is achieved when a new noise model and estimator are identified. For comparison, the autocovariances for data set 1 (from the first time period) are presented as well. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Finally, we examined the consistency of the ALS results. The data studied is shown in Figure 8. First, we considered the data at the start of year 1 and divided this data into 20 sets of 1500 data points. We applied the feasible-generalized ALS method to the first of these subsets and used the results to design an estimator L_1. We then processed the rest of the data using L_1 and calculated the L_1-innovation autocovariances. As shown in Figure 9, the estimator performed optimally on most (13 out of 20) of the data sets.

Three data sets in the middle of the time period are affected by a large disturbance which is not characteristic of the process (such a disturbance never reappears in several months of data). As a result, the innovations have large spikes, shown in Figure 10, and significant correlation remains in the L_1-innovations, shown in Figure 11. Solving the ALS problem for these sets of data does not produce an optimal estimator, as disturbances of this magnitude are not repeated in the data.

In addition, to the data sets with the large disturbance, the data at the end of the first time period exhibit different disturbance characteristics than the rest of the data. This change is most clearly visible in the y_1 data, where the variance decreases. As a result, the estimator L_1 is no longer optimal, as shown in Figure 12, although its behavior may still be considered acceptable. After applying the ALS method to one of these data sets, we produced a new estimator L_2. When we process the data with L_2, the innovations are white (Figure 13). Thus, we conclude that the ALS method works well on these data sets but the disturbances affecting the system have changed slightly.

Finally, we obtained data for the same process from 6 months later and 1 year later. The data from each time period were divided into six data sets and processed with L_1. The resulting innovation autocovariances are shown in Figures 14 and 15. Again, the estimator produces near-optimal results on all data sets. The results indicate that the disturbances to the system remain relatively constant over an extended time period, as the disturbance model identified at the beginning of the year produces an estimator which is nearly optimal throughout the year. We expect that the same disturbance model would also be reliable for use in calculating performance monitoring benchmarks although such an illustration is beyond the scope of this work.

Conclusions

By reducing large models and estimating the optimal weighting from data, the ALS method produces more reliable results and can be applied to industrial data. By eliminating poorly observable modes, the computational time decreases significantly and the results remain accurate. Estimating the optimal weighting from data provides a feasible alternative to the intractable computation of the optimal weighting. Both simulation and industrial data demonstrate the effectiveness of these methods. Several areas lie open for further research, including a detailed comparison with other methods of noise model identification, application to performance monitoring, determining how to choose an appropriate integrating disturbance model, and extending the ALS technique to nonlinear systems.

Acknowledgments

The authors would like to thank Drs. G. Hu and J. Flores-Cerrillo of Praxair for providing the industrial data and for their ongoing collaboration. 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.

Literature Cited

1. Zagrobelny MA, Ji L, Rawlings JB. Quis custodiet ipsos custodes? *Annu Rev Control*. 2013;37:260–270.
2. Odelson BJ, Rajamani MR, Rawlings JB. A new autocovariance least-squares method for estimating noise covariances. *Automatica*. 2006;42(2):303–308.
3. Rajamani MR, Rawlings JB. Estimation of the disturbance structure from data using semidefinite programming and optimal weighting. *Automatica*. 2009;45:142–148.
4. Mehra RK. On the identification of variances and adaptive Kalman filtering. *IEEE Trans Automat Control*. 1970;15(12):175–184.
5. Mehra RK. Approaches to adaptive filtering. *IEEE Trans Automat Control*. 1972;17:903–908.
6. Rajamani MR, Rawlings JB, Qin SJ. Achieving state estimation equivalence for misassigned disturbances in offset-free model predictive control. *AIChE J*. 2009;55(2):396–407.
7. Ljung L. *System Identification—Theory for the User*, 2nd ed. New Jersey: Prentice Hall, 1999.
8. Qin SJ. An overview of subspace identification. *Comput Chem Eng*. 2006;30:1502–1513.
9. Van Overschee P, De Moor B. A unifying theorem for three subspace system identification algorithms. *Automatica*. 1995;31(12):1853–1864.
10. Rajamani MR. Data-based techniques to improve state estimation in model predictive control. Ph.D. thesis, University of Wisconsin–Madison. 2007. Available at: URL http://jbrwww.che.wisc.edu/theses/rajamani.pdf.
11. Hua D. On the symmetric solutions of linear matrix equations. *Linear Algebra Appl*. 1990;131:1–7.
12. Aplevich JD. *The Essentials of Linear State-Space Systems*, chapter 9. New York: Wiley, 2000.
13. Lancaster P, Tismenetsky M. Computer science and applied mathematics. *The Theory of Matrices: With Applications*. San Diego: Academic Press, 1985.
14. Lima FV, Rawlings JB, Rajamani MR, Soderstrom TA. Covariance and state estimation of weakly observable systems: application to polymerization processes. *IEEE Trans Control Syst Technol*. 2013;21(4):1249–1257.
15. Magnus JR, Neudecker H. *Matrix Differential Calculus with Applications in Statistics and Econometrics*, chapter 13.5. New York: Wiley, 1999.
16. Schmidt P. *Econometrics*, chapter 2.5. New York: Marcel Dekker, Inc., 1976.
17. Anderson TW. *An introduction to multivariate statistical analysis*, 3rd ed. New York: Wiley, 2003.
18. Ghosh M, Sinha BK. A simple derivation of the Wishart distribution. *Am Stat*. 2002;56(2):100–101.
19. Magnus JR, Neudecker H. The commutation matrix: some properties and applications. *Ann Stat*. 1979;7(2):381–394.

Appendix A: Proof of Equivalence Between Single Column and Full Matrix ALS Techniques

We define the autocovariance matrix as

$$\mathscr{R}(N)=\begin{bmatrix} \mathscr{y}_k\mathscr{y}_k' & \dots & \mathscr{y}_k\mathscr{y}_{k+N-1}' \\ \vdots & \ddots & \vdots \\ \mathscr{y}_{k+N-1}\mathscr{y}_k' & \dots & \mathscr{y}_k\mathscr{y}_k' \end{bmatrix}$$

$$=[\mathscr{R}_1(N)\ \ \mathscr{R}_2(N)\dots\ \mathscr{R}_N(N)]$$

in which $\mathscr{R}_i(N)$ are the block columns of $\mathscr{R}(N)$. $\mathscr{R}(N)$ vectorizes to

$$\operatorname{vec}(\mathscr{R}(N))=\begin{bmatrix}\operatorname{vec}(\mathscr{R}_1(N))' & \operatorname{vec}(\mathscr{R}_2(N))' & \dots & \operatorname{vec}(\mathscr{R}_N(N))'\end{bmatrix}'$$
$$=\mathscr{A}_{\text{full}}\begin{bmatrix}\operatorname{vec}(Q_w)\\ \operatorname{vec}(R_v)\end{bmatrix}$$

in which $\mathscr{A}_{\text{full}}$ is the matrix $\mathscr{R}$ for the full matrix ALS method.[2] As the autocovariances in $\mathscr{R}_2\dots\mathscr{R}_N$ are duplicates of the autocovariances in $\mathscr{R}_1$, we can write each $\operatorname{vec}(\mathscr{R}_i(N))$ as a linear combination of the elements of $\operatorname{vec}(\mathscr{R}_1(N))$

$$\operatorname{vec}(\mathscr{R}(N))=\begin{bmatrix}I_{Np^2}\\ J_2\\ \vdots\\ J_N\end{bmatrix}\operatorname{vec}(\mathscr{R}_1(N))=\begin{bmatrix}I_{Np^2}\\ J_2\\ \vdots\\ J_N\end{bmatrix}\mathscr{R}\begin{bmatrix}\operatorname{vec}(Q_w)\\ \operatorname{vec}(R_v)\end{bmatrix}$$

in which $\mathscr{A}$ is the matrix for the single column ALS method and the J_i are chosen such that $\operatorname{vec}(\mathscr{R}_i(N))=J_i\operatorname{vec}(\mathscr{R}_1(N))$. Therefore, the matrices for the full matrix and single column ALS methods are related as

$$\mathscr{A}_{\text{full}}=\begin{bmatrix}I_{Np^2}\\ J_2\\ \vdots\\ J_N\end{bmatrix}\mathscr{A}$$

As the matrix $\begin{bmatrix}I_{Np^2} & J_2' & \dots & J_N'\end{bmatrix}'$ is always full column rank, $\mathscr{A}_{\text{full}}$ is full column rank if and only if $\mathscr{A}$ is full column rank.

Appendix B: Derivation of Formula for $S = \operatorname{cov}(\hat{b})$

Let

$$\hat{b}=\left(\frac{1}{N_s}\sum_{i=1}^{N_s}\left(\mathscr{y}_i y_i(1)'\right)\right)_s$$

in which $\mathscr{y}_i=\begin{bmatrix}y_i(1)' & \dots & y_i(N)'\end{bmatrix}'$ are i.i.d. normal variables with zero mean and variance $P_y=\begin{bmatrix}P_0 & P_{y,0}'\\ P_{y,0} & P_{y2}\end{bmatrix}$. Then, the variance of $\hat{b}$ is

$$\operatorname{cov}(\hat{b})=\frac{1}{N_s}\left(\left(P_0\otimes P_y\right)+K_{p,\tilde{p}}\left(P_{y,0}\otimes P_{y,0}'\right)\right) \tag{A1}$$

where p is the dimension of y and $\tilde{p}=Np$ is the dimension of Y. Defining Y_i and y_i appropriately in (A1), we arrive at (7). To derive (A1), we begin by noting that the sample variance of Y_i, $\hat{P}_y=\frac{1}{N_s}\sum_{i=1}^{N_s}\left(Y_iY_i'\right)$, is distributed according to the Wishart distribution with pdf[17,18]

$$p(\hat{P}_y|P,N_s+1)=\frac{|\hat{P}_y|^{\frac{1}{2}(N_s-\tilde{p})}\exp\left(-\frac{1}{2}\operatorname{tr}\left(P_y^{-1}\hat{P}_y\right)\right)}{2^{\frac{1}{2}(N_s+1)\tilde{p}}|P_y|^{\frac{N_s+1}{2}}\pi^{\frac{\tilde{p}(\tilde{p}-1)}{4}}\prod_{i=1}^{p}\Gamma\left(\frac{N_s-i}{2}\right)}$$

As a result[19]

$$\mathrm{var}((\hat{P}_y)_s)=\frac{1}{N_s}\left(I_{\tilde{p}^2}+K_{\tilde{p},\tilde{p}}\right)(P_y\otimes P_y) \tag{A2}$$

The estimated variance of $\hat{b}$ is the $p\tilde{p}\times p\tilde{p}$ matrix in the upper-left corner of $\mathrm{var}((\hat{P}_y)_s)$. From (A2), we can write this submatrix as

$$\mathrm{cov}(\hat{b})=\frac{1}{N_\mathrm{s}}\left(\left(P_0\otimes P_y\right)+K_{p,\tilde{p}}\left(P_{y,0}\otimes P'_{y,0}\right)\right)$$

which is the formula in (A1). Note that the denominator in (A1) and (A2) contains N_s rather than $N_\mathrm{s}-1$ because the mean of Y_i is known to be zero.

Manuscript received Dec. 12, 2014, and revision received Feb. 12, 2015.