MPC PERFORMANCE MONITORING AND DISTURBANCE MODEL IDENTIFICATION

by

MEGAN ANN ZAGROBELNY

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

(Chemical Engineering)

at the

UNIVERSITY OF WISCONSIN-MADISON

2014

Date of final oral examination: December 9, 2014

The dissertation is approved by the following members of the Final Oral Committee:

James B. Rawlings, Professor, Chemical and Biological Engineering
Christos T. Maravelias, Professor, Chemical and Biological Engineering
Ross E. Swaney, Associate Professor, Chemical and Biological Engineering
Michael D. Graham, Professor, Chemical and Biological Engineering
Robert D. Nowak, Professor, Electrical and Computer Engineering

To Mom, Dad, Tim, Beth, and Kathryn. And to my mother in Heaven, the Blessed Virgin Mary.

ACKNOWLEDGMENTS

Above all else, I thank God for creating me and blessing me with so many opportunities in my life. The Saints and Angels in Heaven, who have been my intercessors and companions, also deserve credit, especially my Guardian Angel, who has been working alongside me.

I also thank my advisor, Dr. James Rawlings, who has shaped my graduate school experience and education in so many ways. I have learned so much from him, and he has helped me to grow and to think for myself.

I also thank all the members of our research group. The work that Luo Ji did on performance monitoring gave me a good starting point for my project, and I enjoyed working alongside him during my time in Madison. The other experienced members of the group, Rishi Amrit, Rishi Srivastava, Kaushik Subramanian, and Ankur Gupta all helped me in countless ways. Ankur was also a valuable resource for his understanding of our computers. It was a pleasure to join the group alongside Cuyler Bates; I thank him for his friendship and for helping me get Octave back on my computer after a bad upgrade. It was also a joy to welcome new members into the group and to get to work with them. I wish Min Yao, Michael Risbeck, and Nishith Patel all the best as they continue their work. I also thank Mary Heimbecker for her help, her treats, and her cheerfulness.

My two summer internships were great experiences for me and also shaped the course of my research. I am grateful to the advanced controls groups at Eastman Chemical Company and Praxair, Inc. for giving me the opportunity to learn more and to step outside of the academic setting. Drs. James Downs and Ernie Vogel at Eastman were outstanding teachers and examples for me. Drs. Gangshi Hu and Jesus Flores-Cerrillo at Praxair not only helped me during my internship, but also provided data to me and continued our collaboration for the rest of my time in Madison. I am truly grateful to all of those with whom I interacted during those summers.

There are so many friends who deserve thanks, both in Madison and beyond. To all of my Madison roommates - Jamie, Kelly, Mary Clare, Val, Emily, Nicola, and Rosi - thank you for supporting me, putting up with me, feeding me, and eating my goodies. Val and Emily were also great mentors for me. Thanks also to the "Dark Side," Sara, Hannah, Marco, and Cuyler, for our many lunches, ice cream breaks, and experiences together. I am grateful to my Bible studies for

always being so loving, welcoming, and willing to put up with me. My spiritual directors helped my soul to grow along with my mind. The nuns of the Carmel of St. Therese in Loretto, PA also deserve much thanks for their prayers, which have helped me so much in these last few months. Thanks also goes to Amy, who has been going through grad school with me, even though we're not in the same school or even the same state. Thank you for visiting me, listening to me, and keeping me calm.

I also thank all of my teachers over the years. Without you all, I wouldn't have been able to begin this process. Mrs. Huckestein from 6th grade particularly requested acknowledgement. I am also grateful to Prof. Ogunnaike from the University of Delaware for teaching the undergraduate class that made me interested in process control.

Finally, I thank my family, especially my parents. I wouldn't be here without you. My sister Beth deserves much credit for her grammar help and for taking the time to read my thesis. She, Kathryn, and Tim all have supported me during this process. Thanks to my Aunt Helen for visiting me and staying in contact, and my Uncle John Klincewicz for encouraging me in my research.

I'm sure I've only mentioned a few of the many people who deserve credit in one way or another. If I missed someone, I'm very sorry, and I thank you for your help.

<div align="right">

Megan Ann Zagrobelny
Madison, WI
December 2014

</div>

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Although model predictive control (MPC) has been widely implemented in industry and shown to have significant economic benefit, currently no systematic method exists to assess if these controllers are performing optimally or to monitor their performance over time. Here we address this problem by proposing a benchmark called the key performance index (KPI), which is the expectation of the stage cost. Since the MPC regulator minimizes the stage cost, this benchmark inherently matches the controller's objective. For a linear, unconstrained system, the stage cost is a quadratic form of a normal variable and therefore has a generalized chi-squared distribution. The plant KPI is calculated as the time average of the stage cost and is shown to have a normal distribution. The mean and the variance of the stage cost and plant KPI are calculated from knowledge of the process model, controller, and disturbance model.

Calculation of the KPI requires accurate knowledge of the disturbances affecting the system. An accurate disturbance model also is necessary for designing an optimal estimator. The autocovariance least squares (ALS) method estimates these disturbance covariances from data using a modified least-squares problem. However, the standard ALS methods are not easily applicable to industrial data. The large-dimensional models used in industrial systems, which often contain poorly observable states, result in large optimization problems that are poorly conditioned and have a large number of unknowns. Directly solving these large optimization problems is computationally inefficient. In addition, because the original ALS formulation weights the least squares problem with the identity matrix, the resulting estimates may have a large variance.

We resolve the first of these challenges by reducing the model based on a singular value decomposition of the observability matrix to contain only the necessary observable states. As the optimal weighting for the least squares problem cannot be calculated in practice, a feasible generalized least squares technique is developed to estimate the optimal weighting from data. Application of the improved ALS method to an industrial data set demonstrates that the resulting covariances produce an optimal estimator. These improvements reduce the computational time and produce more reliable estimates as compared to the original ALS method.

As an alternative to the ALS method, a maximum likelihood estimation (MLE)

method is proposed. Whereas the ALS method requires estimating the optimal weighting from data, the MLE method does not have this requirement. Instead, the process and measurement noise covariances are estimated by maximizing the probability of observing the measured outputs. Thus this optimization problem has a more sound theoretical basis. Sufficient conditions for the existence of a solution to the MLE problem are given. The conditions for uniqueness are compared to those of the ALS method. Although the computational burden is large compared to the ALS method, the MLE method was applied to several small-scale examples and shown to maximize the likelihood compared to the ALS method.

Further research in applying the improved ALS method to performance monitoring, developing statistical tests to detect changes in the KPI, and adapting the MLE method to larger-scale systems is recommended.

1

INTRODUCTION

1.1 MOTIVATION

Model predictive control (MPC) is a common form of advanced process control and play an essential role in process operation in the chemical industry and other areas (García, Prett, and Morari, 1989; Qin and Badgwell, 2003; Bauer and Craig, 2008). MPC is especially advantageous on systems that are difficult to control due to constraints and interaction among variables. By directly using a process model within the controller, MPC chooses the optimal strategy to control these difficult systems and therefore improves overall operation and profitability of the plant compared to more basic control methods.

Although MPC provides many advantages, numerous factors, including a poor process model or a poor disturbance model, can inhibit the MPC controller from achieving its optimal performance and thus reduce its economic benefits. These factors may be present from the startup of the control system, or they may arise as the process conditions or equipment change over time. A controller performance monitoring system compares the achieved performance of each controller to an optimal benchmark and identifies those controllers that are not performing

well. Monitoring the performance of the controllers over time would alert the operators and control engineers when controller performance has degraded. A diagnosis step then identifies the cause of performance degradation so that the operators and engineers can resolve the problem. Even when the system is performing optimally given the constraints and disturbances, MPC monitoring can indicate what factors are limiting the performance: for example, the benefit of relaxing a constraint or adjusting the process to reduce the effects of a certain disturbance.

While controller performance monitoring has been a popular subject in the literature since the work of Åström (1970), MPC performance monitoring remains an open area of research. The size and complexity of the systems, as well as the more complicated controller objective function, limit the use of standard controller performance monitoring on MPC systems.

In order to assess the performance of any controller, there must be a benchmark for acceptable performance. While a type of benchmark could be determined purely from past data recorded during a period of satisfactory performance, a theoretical benchmark provides the most information about the optimal controller performance. A prerequisite for calculating such a benchmark is a disturbance model of the system. Without understanding the disturbances facing the system, it is impossible to judge how well the controller responds to these disturbances.

A disturbance model must be general enough to account for all the numerous disturbances that may affect the process. The standard approach of modeling all these unknown disturbances as white noise is justified by the central limit theorem. As this basic noise model does not guarantee that the controlled variables will achieve their setpoints, an additional stochastic term must be included, usually in the form of integrated white noise. Under this general structure, the disturbance identification problem becomes a question of identifying the covariances

for each of the noise terms. Accurately identifying each noise covariance is also necessary for estimator design — without knowing the extent of the random effects, unmeasured states of the system cannot be accurately estimated. Therefore, identifying the disturbance model not only allows the performance monitoring benchmark to be calculated, but also may improve controller performance by improving the estimator.

1.2 GOALS

To address these challenges to MPC performance monitoring, we focus on the three major goals of this work:

1. Propose a realistic benchmark for MPC performance monitoring and establish the distribution of this benchmark.

2. Present a modified autocovariance least squares method for identifying the disturbance model and illustrate this method on industrial data.

3. Formulate a maximum likelihood method for disturbance model identification and establish necessary conditions for the existence and uniqueness of the optimal solution.

1.3 NOTATION

We summarize the parts of the model predictive controller in Figure 1.1. In this work, we assume that the plant is described by the following discrete time linear model:

$$x^+ = Ax + Bu + Gw$$

$$y = Cx + v$$
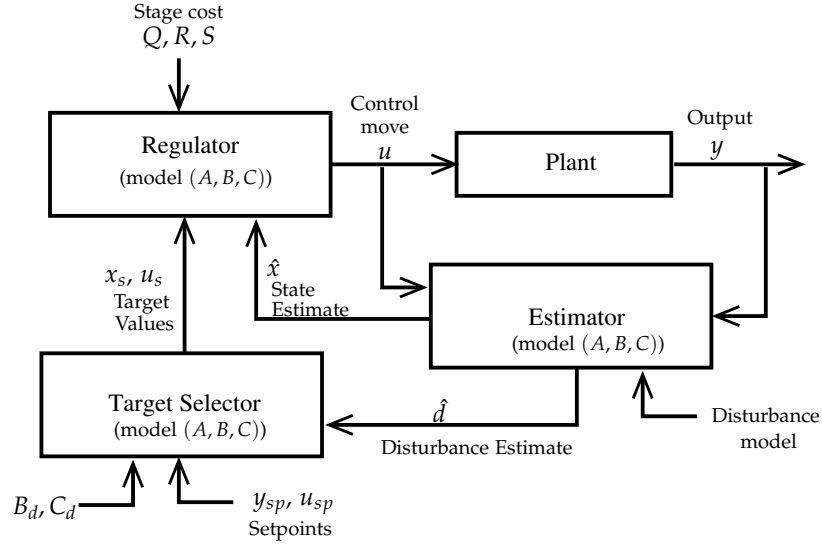
Figure 1.1: Diagram summarizing the key elements of an MPC controller.

in which the state, $x \in \mathbb{R}^n$, is an unmeasured vector of quantities describing the system, $u \in \mathbb{R}^m$ is the manipulated input, $w \in \mathbb{R}^g$ is the process noise (which affects the states), $y \in \mathbb{R}^p$ is the measured output, and $v \in \mathbb{R}^p$ is the measurement noise. The process model or deterministic system consists of the state transition matrix, $A \in \mathbb{R}^{n \times n}$, the input matrix, $B \in \mathbb{R}^{n \times m}$, and the output matrix, $C \in \mathbb{R}^{n \times p}$. The noise model or stochastic part of the system refers to the noise-shaping matrix, $G \in \mathbb{R}^{n \times g}$, and the covariances of w and v, which are denoted as Q_w and R_v, respectively. The noises w and v are assumed to be zero mean, normally distributed white noises and are assumed to be independent of each other.

The role of the estimator is to estimate the unmeasured state from the measured output (and input) data. For a linear system, the estimator is called the Kalman filter and takes the form:

$$\hat{x}(k+1|k) = A\hat{x}(k|k) + Bu(k)$$

$$\hat{x}(k|k) = \hat{x}(k|k-1) + L\varepsilon(k)$$

$$\varepsilon(k) = y(k) - C\hat{x}(k|k-1)$$

in which $\hat{x}(k|k-1)$ denotes the prediction of $x(k)$ given the data up to time $k-1$ and $\hat{x}(k|k)$ indicates the estimate of $x(k)$ given the data up to time k. To obtain only the estimate $\hat{x}(k|k-1)$, we alternately write the Kalman predictor as

$$\hat{x}(k+1|k) = A\hat{x}(k|k-1) + Bu(k) + AL\varepsilon(k)$$

We refer to L as the filter gain, and AL as the predictor gain. For simplicity, we use $\hat{x}(k)$ to denote $\hat{x}(k|k-1)$.

We refer to ε as the innovation. When the filter is optimal, ε is a white noise, *i.e.* $\varepsilon(k)$ and $\varepsilon(k+j)$ are uncorrelated for $j \neq 0$. We use the term L-innovation to refer to the innovation when the filter gain is not necessarily optimal; the L-innovations may not be white. The optimal filter gain is calculated from the noise covariances Q_w and R_v.

We often augment the model with an integrating disturbance model so that it takes the form

$$\begin{bmatrix} x \\ d \end{bmatrix}^+ = \begin{bmatrix} A & B_d \\ 0 & I \end{bmatrix} \begin{bmatrix} x \\ d \end{bmatrix} + \begin{bmatrix} B \\ 0 \end{bmatrix} u + \begin{bmatrix} G & 0 \\ 0 & G_d \end{bmatrix} \begin{bmatrix} w \\ w_d \end{bmatrix}$$

$$y = \begin{bmatrix} C & C_d \end{bmatrix} \begin{bmatrix} x \\ d \end{bmatrix} + v$$

We refer to this system as the augmented system. We can also write the Kalman filter for the augmented system. The estimator then provides both the state estimate, $\hat{x}$, and the disturbance estimate, $\hat{d}$.

For an augmented system, the disturbance model also includes the matrices B_d and C_d, a second noise-shaping matrix, G_d, and the combined variance of $\begin{bmatrix} w' & w'_d \end{bmatrix}'$.

The target selector uses the disturbance estimate $\hat{d}$ and the setpoints for y and u to choose steady-state targets x_s and u_s for the state and input. When

more outputs than inputs are present, not all outputs can be controlled to their setpoints, so we choose a subset of the outputs to be controlled to their setpoints. When more inputs than outputs are present, the target selector also seeks to keep the inputs near their setpoint values. The target selector is described in more detail in Appendix 3.A of Chapter 3.

The regulator chooses the optimal sequence of inputs $\mathbf{u} = \{u(0), u(1), \ldots u(N)\}$ to bring the state and input to their target values. Only the first input $u(0)$ is applied to the system, as the regulator finds a new optimal sequence of inputs at the next time step. For the linear, unconstrained system, we often use the linear quadratic regulator. In the infinite horizon case ($\mathbf{u} = \{u(0), \ldots u(\infty)\}$) the linear quadratic regulator solves the problem

$$
\min_{\mathbf{u}} \quad V(x(0), \mathbf{u}) = \sum_{k=0}^{\infty} \ell(k)
$$

$$
\text{subject to} \quad \ell(k) = \frac{1}{2} \Big((x(k) - x_s)' Q (x(k) - x_s) + (u(k) - u_s)' R (u(k) - u_s)
$$

$$
+ (u(k+1) - u(k))' S (u(k+1) - u(k)) \Big)
$$

$$
x(k+1) = Ax(k) + Bu(k) \qquad k = 0, 1, \ldots \infty
$$

We refer to the function $\ell(k)$ as the stage cost. The matrices Q, R, and S are weights chosen by the user; the choice of these tuning parameters is a tradeoff between keeping the states near their target values, keeping the inputs near their target values, and limiting the rapid change of inputs. The Kalman filter or predictor is used in place of $x(0)$, since the true state is unknown. This regulator results in a control law of the form

$$
u(k) = K \begin{bmatrix} \hat{x}(k) - x_s \\ u(k) - u_s \end{bmatrix} + u_s
$$

The linear quadratic Gaussian (LQG controller) refers to the MPC controller

with the estimator and regulator defined by the Kalman filter and linear quadratic regulator. A more detailed explanation of the LQG controller is given in Rawlings and Mayne (2009, Ch. 1).

1.4 DISSERTATION OVERVIEW

This dissertation is organized as follows:

In Chapter 2 a review of the controller performance monitoring literature is presented. The standard controller performance monitoring method, the minimum variance benchmark, is presented in the framework of state space MPC.

In Chapter 3, a performance monitoring benchmark that accurately reflects the objectives of the controller is presented. An analytical formula is derived for the theoretical benchmark for a linear unconstrained controller. The distribution of the benchmark is derived based on this analytical formula. The analytical formula is extended to include a general deterministic disturbance and is modified for models in innovation form.

In Chapter 4, a literature review is presented for disturbance model identification, and a brief derivation and explanation of the ALS method is given. The conditions under which the ALS problem has a unique solution are discussed. Methods for assessing the accuracy of the ALS solution are presented.

In Chapter 5, the use of integrating disturbance models is discussed and an alternative double integrator disturbance model is presented.

In Chapter 6, two methods are presented to make the ALS approach more applicable to industrial data: reducing the state space model by removing weakly observable states and using a feasible generalized least squares approach by estimating the optimal weighting from data. These method are demonstrated on simple examples.

In Chapter 7, the modified ALS method is applied to industrial data.

In Chapter 8, disturbance model identification is posed as a maximum likelihood estimation problem. The existence and uniqueness of solutions to the MLE problem are discussed, and the MLE problem is compared to the ALS problem with optimal weighting. Several recommendations to reduce the computation time are presented, and the MLE method is demonstrated on several low-dimensional examples. The advantages and disadvantages of the MLE and ALS methods are compared

In Chapter 9, the major contributions of this work are summarized and areas of future research are presented.

2

PERFORMANCE MONITORING BACKGROUND

2.1 REVIEW OF CONTROLLER PERFORMANCE MONITORING

THE MINIMUM VARIANCE BENCHMARK. The minimum variance method developed by Harris (1989) marks a starting point for controller performance monitoring. This method uses closed-loop data to estimate the minimum output variance possible for a specific system under any controller and compares this variance to the actual output variance. The groundwork for the minimum variance method was developed earlier by Åström (1970), who gave an expression for the minimum variance control law (the sequence of inputs that minimizes the output variance of a single-input single-output (SISO) system). Åström predicted the application of this concept to performance monitoring: since the order of the minimum variance output is determined by the time delay, the order of the controller error can be used to determine whether or not the system is under optimal control. Harris's primary contribution comes from estimating the minimum variance from closed-loop data without *a priori* knowledge of a process or disturbance model. Whereas Åström's derivation relies on a time series model of the process and disturbance, Harris's method uses standard techniques to fit a time series model to

routine operating data and uses this closed-loop model to estimate the minimum variance. With Harris's method, no *a priori* process knowledge is needed except for the time delay. Since only routine operating data is required, assessing the controller does not interrupt process operation. The minimum variance benchmark was normalized by Desborough and Harris (1992) to range from 0 to 1, while Qin (1998) related it to standard PID controllers.

EXTENSIONS OF THE MINIMUM VARIANCE BENCHMARK. Despite the appealing simplicity of Harris's method, the minimum variance benchmark required much development before it could be widely applied in practice, including extending the benchmark to multiple-input multiple-output (MIMO) systems, automatically identifying of a closed-loop process model, handling constrained processes, diagnosing causes of poor performance, and integrating the minimum variance technique with other signal processing tools (Harris, Seppala, and Desborough, 1999).

Investigation into extending the minimum variance method to MIMO systems occupies a prominent place in monitoring research. While knowing the SISO time delay is simple, the multivariate time delay matrix, called the interactor, cannot be determined solely from the time delays of each input-output pair (Harris, Boudreau, and MacGregor, 1996a). The complexity of the MIMO benchmark further increases when the covariances between the interacting outputs are considered (Qin and Yu, 2007). Following the direct extension to MIMO systems by Harris et al. (1996a), the algorithm of Huang, Shah, and Kwok (1997) reduces the amount of process knowledge needed, as well as the computational requirements. In the case of an unknown interactor, Kamrunnahar, Fisher, and Huang (2004) used a combined parametric and non-parametric model. Huang, Ding, and Thornhill (2005) estimated a suboptimal benchmark from the largest delay in the interactor. Huang, Ding, and Thornhill (2006) and Yu and Qin (2009) examined

systems with a diagonal interactor, which require less process knowledge.

Since the minimum variance benchmark does not represent a realistic benchmark for many processes, alternative benchmarks allow greater flexibility in how the optimal performance is defined. Huang and Shah (1998), Xu, Lee, and Huang (2006), and Horch and Isaksson (1999) added a user-defined component to the minimum variance output. Yuan, Lennox, and McEwan (2009) developed bounds for a MIMO user-defined benchmark. Rather than changing the desired output, Bezergianni and Georgakis (2000, 2003) and Yuan and Lennox (2009) defined a new benchmark that also incorporated the open-loop variance. Grimble (2002) estimated a generalized minimum variance benchmark with weighting on the input variance. For processes with step-disturbances, Eriksson and Isaksson (1994) modified the minimum variance benchmark to encourage integral control. Hugo (2006) accounted for the simplified disturbance models used by controllers, while Liu, Huang, and Wang (2011) considered several user-specified performance requirements.

A natural extension to the minimum variance benchmark, the linear-quadratic-Gaussian (LQG) benchmark identifies the minimum output variance possible provided that the input variance does not exceed a certain threshold (Huang and Shah, 1999, chap. 13). This benchmark uses a trade-off curve between the output and input variances. A suboptimal controller lies above the trade-off curve, while an optimal controller lies on the trade-off curve. Performance is measured by the distance between the current variances and the ideal curve. Since this benchmark requires a complete process model, Kadali and Huang (2002) and Dai and Yang (2004) developed methods to calculate the model using subspace identification.

As the minimum variance benchmark was designed for unconstrained, linear time invariant processes under feedback control, other research relaxed these conditions. Desborough and Harris (1993) expanded the minimum variance control law to feed-forward/feedback systems. Huang (2002), Olaleye, Huang, and

Tamayo (2004), and Xu, Huang, and Tamayo (2008) applied minimum variance control to different linear time variant cases. Bergh and MacGregor (1987) and Harrison and Qin (2009) developed expressions for constrained minimum variance. Ko and Edgar (2001) developed a monitoring system for the constrained case by using simulations to estimate the constrained minimum variance bounds.

ALTERNATIVES TO MINIMUM VARIANCE. Due to the limitations of the minimum variance benchmark as previously discussed, data-driven methods were developed as alternate ways to assess controller performance. These methods are based on analysis of the data with subspace-based or statistical techniques. Subspace projections provide a means to identify the directions with poorest performance and isolate the effects of measured and unmeasured disturbances (McNabb and Qin, 2003, 2005). Yu and Qin (2008a,b) extracted the directions with changes in control performance through generalized eigenvalue analysis and identified the responsible loops by calculating the contribution of each loop to these directions. Xie, Kruger, Lieftucht, Littler, Chen, and Wang (2006) developed a method to remove auto- and cross-correlation between the outputs before applying principal component analysis. Rather than a subspace-based method, Salsbury (2005) used the autocorrelation of the error signal to assess performance during step-wise load changes. Similarly, Badwe, Gudi, Patwardhan, Shah, and Patwardhan (2009) used the partial correlations between model residuals and inputs to identify the parts of the model that need re-identification. The normalized multivariate impulse response curve of Shah, Patwardhan, and Huang (2002) and the closed-loop potential plots used by Huang et al. (2006) and Zhao, Chu, Su, and Huang (2010) measure the decay rate of the output predictability, but rely on the user analysis of individual plots.

Fault detection methods have also been proposed as monitoring tools. Kesavan and Lee (1997) modeled possible causes of poor performance by creating a fault

parameter vector and then formulated an estimation problem to detect the true cause. However, this estimation problem requires a defined list of faults with prior probabilities. Ge, Yang, Song, and Wang (2008) used least-squares support vector regression to remove the effects of external variables before using statistical methods to determine if a fault occurred. Performance monitoring was formulated as a hypothesis testing problem by Tyler and Morari (1996) to determine if the closed-loop impulse response coefficients violated a desired bound and by Huang and Tamayo (2000) to determine if a new model better fit the process than did the old model.

STATISTICS OF PERFORMANCE METRICS. Understanding the statistics of the performance metrics is necessary to compare accurately the achieved and theoretical performance measures. Harris (1989) analyzed the sampling properties of the minimum variance estimate in order to determine the number of samples needed for a good estimate. Desborough and Harris (1992) computed the mean, variance, and kernel density estimates of the normalized minimum variance benchmark and showed that the variance depends both on the number of samples and the autocorrelation of the process. In addition, they discussed the need to choose a sample length that balances susceptibility to outliers, distribution of the metric, and ability to detect changes in the metric. Harris (2004) provided a more general discussion of statistics of performance metrics by recognizing that many performance indices are expressed as quadratic forms or as a ratio of quadratic forms. He focused specifically on a minimum variance type metric that is expressed as a ratio of quadratic forms. Since closed-form probability distributions do not exist for this metric, an iterative method as well as several approximate methods are developed to compute its confidence intervals.

MODEL PREDICTIVE CONTROL MONITORING. MPC performance monitoring is more complicated than the monitoring of simpler controllers, since MPC loops

tend to have many variables and constraints. In addition, since the process model is used directly by the controller, improving performance might require re-identifying the process model rather than simply adjusting the controller tuning parameters. Patwardhan (1999) proposed using the MPC design objective as the performance measure. Using this concept, Julien, Foley, and Cluett (2004) developed plots of the best achievable MPC performance. They used closed-loop data to re-identify the model and plotted performance curves first based on the new model and old controller, then based on the new model with the proposed new controller. This comparison shows the potential benefits of re-identifying the model used by the controller. Schafer and Cinar (2004) used a ratio of historic and achieved performance to detect *changes* in controller performance and diagnosed the cause of any changes by using a ratio of achieved and designed performance. Tsakalis and Dash (2007) proposed a design-based benchmark using robust stability conditions determined from the controller design. However, this method relies on using external excitation and suffers from false alarms.

An alternate method in Dumont, Kammer, Allison, Ettaleb, and Roche (2002) compared the controller prediction error to the open-loop output, since both signals are realizations of the same stochastic process under a perfect model. Using a similar idea, Sun, Qin, Singhal, and Megan (2013) compared the prediction error to the open-loop disturbance innovations, which are estimated through a projection method. Xu, Huang, and Akande (2007) and Lee, Tamayo, and Huang (2010) assessed the potential economic benefit of relaxing constraints or reducing the variability of certain variables. However, this method is valid only with perfect models.

Other research focused on diagnosing the underlying causes of poor MPC performance. Patwardhan and Shah (2002) gave bounds on the effects of constraints, modeling uncertainty, disturbance uncertainty, and nonlinearity on MPC performance. Badwe, Patwardhan, Shah, Patwardhan, and Gudi (2010) determined the

effect of model-plant mismatch by estimating relationships between the model and the plant from closed-loop data. To isolate the performance degradation caused by changes in the process/controller (as opposed to changes in the disturbances), Rato and Reis (2010) developed a modified index based on historic data. Tian, Chen, and Chen (2011) diagnosed performance by identifying the subspaces affected by each possible cause. Despite this research on MPC monitoring, a complete automated monitoring and diagnosis scheme has yet to be developed.

INDUSTRIAL APPLICATIONS.   Before implementing a monitoring scheme on industrial processes, factors such as ease of use for engineers/operators and clear connections between the controller benchmark and economic objectives must be considered (Desborough and Miller, 2001). To increase the ease of use, Thornhill, Oettinger, and Fedenczuk (1999) calculated parameters for use in monitoring according to the general loop type. In contrast, Ingimundarson and Hagglund (2005) recommended choosing such parameters based on the tuning of the controller, not the type of loop. They also discussed the challenge of using filtered and smoothed archived industrial data. Wang, Huang, and Chen (2007) developed a MIMO minimum variance method for systems with different rates of input and output sampling. To integrate all steps of monitoring, Ahsan, Grosvenor, and Prickett (2004) focused on combining controller monitoring with process monitoring by using a distributed architecture to share data. Shardt, Zhao, Qi, Lee, Yu, Huang, and Shah (2011) highlighted the integration of monitoring schemes for all regulatory loops, along with the supervisory loops that provide the setpoints for the regulatory loops.

Several studies have examined the implementation of monitoring systems in industry. Harris, Seppala, Jofriet, and Surgenor (1996b) and Paulonis and Cox (2003) implemented plant-wide monitoring of SISO loops in a newsprint mill and chemical plants, respectively. Other studies on commercial software for perfor-

mance monitoring also focused on SISO PID controllers (Badmus, Banks, Vishnubhotla, Huang, and Shah, 1998; Bonavita, Bovero, and Martini, 2004). Haarsma and Nikolaou (2000) used the MIMO minimum variance benchmark on a snack food frying process, but the applicability of the benchmark was severely limited due to constraints on the process. Huang, Kadali, Zhao, Tamayo, and Hanafi (2000) successfully assessed and diagnosed a model predictive controller for an industrial combined gas oil coker. However, since their method requires individually examining a number of plots, it is not effective for automatically monitoring a large number of MPC loops. Gao, Patwardhan, Akamatsu, Hashimoto, Emoto, Shah, and Huang (2003) evaluated the performance of two MPC loops. They used a variety of methods on the first controller, including the MIMO minimum variance benchmarks described in Huang and Shah (1999) and the normalized impulse response curve from Shah et al. (2002). However, they simply compared the performance with and without MPC, rather than evaluating the controller performance based on an ideal benchmark. On the second controller, setpoint settling time, a historical benchmark, and prediction error analysis indicated that the performance deteriorated after a load change, but these methods failed to identify why the controller could not compensate for this change.

2.2 MPC AND MINIMUM VARIANCE CONTROL

2.2.1 *Minimum variance as a form of MPC*

The minimum variance controller was derived by Åström based on a time series model and was presented as a theoretical controller, which may not be realizable in practice. In order to better understand how the minimum variance benchmark can be used in MPC performance monitoring and how it relates to other MPC monitoring techniques, we look at past studies to discuss how a model predictive controller can be designed to achieve minimum output variance. Kwong (1991)

determined the minimum variance control law for the SISO state space system and showed that this control law is equivalent to that of Åström (1970). McNabb and Qin (2003) further discuss minimum variance control of a MIMO MPC controller. Here, we use the results of Kwong (1991) to develop a simplified form of the control law and to write an expression for the output under minimum variance control.

Rather than using a state space model, Åström developed the minimum variance controller using a linear, time-invariant SISO time-series model, written as

$$
\begin{aligned}
y(k) = & - a_1 y(k-1) - a_2 y(k-2) - \cdots - a_n y(k-n) \\
& + b_1 u(k-1) + b_2 u(k-2) + \cdots + b_n u(k-n) \\
& + e(k) + c_1 e(k-1) + c_2 e(k-2) + \cdots c_n e(k-n).
\end{aligned}
\tag{2.1}
$$

The parameters a_i and b_i describe the system dynamics, the parameters c_i describe the noise dynamics, and n is the order of the system.

We convert a SISO state space system to the time series model as follows. We begin with the general state space model as defined in Section 1.3:

$$
\begin{aligned}
x^+ &= \tilde{A}x + \tilde{B}u + \tilde{G}w \\
y &= \tilde{C}x + v
\end{aligned}
\tag{2.2}
$$

in which $m = p = 1$. Provided that (A, C) is observable, we can transform this system into the observable canonical form (Callier and Desoer, 1991, p. 312, Chen, 1999, p. 187-189):

$$
\begin{aligned}
x^+ &= Ax + Bu + Gw \\
y &= Cx + v
\end{aligned}
$$

in which

$$A = \begin{bmatrix} -a_1 & 1 & 0 & \cdots & 0 \\ -a_2 & 0 & 1 & \cdots & 0 \\ \vdots & & & & \\ -a_{n-1} & 0 & 0 & \cdots & 1 \\ -a_n & 0 & 0 & \cdots & 0 \end{bmatrix} \qquad B = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix} \qquad C = \begin{bmatrix} 1 & 0 & \ldots & 0 \end{bmatrix}$$

Letting L be the optimal gain of the Kalman predictor, we write the system (2.2) in *innovation form* as

$$\hat{x}(k+1) = A\hat{x}(k) + Bu(k) + Le(k)$$
$$y(k) = C\hat{x}(k) + e(k) \qquad (2.3)$$

Defining c_i coefficients so that

$$L = \begin{bmatrix} c_1 - a_1 \\ c_2 - a_2 \\ \ldots \\ c_3 - a_3 \end{bmatrix}$$

(2.3) is equivalent to the time series model in (2.1). When L is the optimal filter gain, then the innovation, e, is white noise and distributed as $N(0, R_e)$.

Using an infinite horizon linear quadratic regulator, the control action is chosen by solving the optimization problem

$$\min_{\mathbf{u}} \quad V(x(0), \mathbf{u}) = \frac{1}{2} \sum_{k=0}^{\infty} \Big(x(k)'Qx(k) + u(k)'Ru(k)$$
$$+ (u(k+1) - u(k))' S (u(k+1) - u(k)) \Big)$$
$$\text{subject to} \quad x(k+1) = Ax(k) + Bu(k) \qquad k = 0, 1, \ldots \infty \quad (2.4)$$

in which **u** is the sequence of inputs, $u(0), u(1), \ldots u(\infty)$. The matrices Q, R, and S are weights chosen by the user. To achieve minimum variance control, we choose $Q = C'C$ and $R = S = 0$, so that $V(x(0), \mathbf{u}) = \frac{1}{2} \sum_{k=0}^{\infty} y(k)'y(k)$ is minimized by the controller. The control law takes the form

$$u(k) = Kx(k) \qquad\qquad K = -\left(B'\Pi B\right)^{-1} B'\Pi A$$

in which Π solves the Riccati equation

$$\Pi = C'C + A'\Pi A - A'\Pi B \left(B'\Pi B\right)^{-1} B'\Pi A \tag{2.5}$$

This Riccati equation has the solution

$$\Pi = \sum_{i=0}^{d-1} \left(\left(A^i\right)' C'CA^i\right) \tag{2.6}$$

in which d is the overall delay of the system ($b_i = 0$ for $i =< d$ in (2.1)). A delay of d means that the input $u(k)$ affects $y(k+d)$ but not $y(k) \ldots y(k+d-1)$. In the SISO state space model, the first $d-1$ rows of B are zero. Since we assume a causal system, $y(k)$ can only be affected by previous inputs, and we must have $d \geq 1$.

Due to the location of zeros in C and B, the following relations hold, as presented in Kwong (1991):

$$CA^{j-1}B = \begin{cases} 0 & j < d \\ b_d & j = d \end{cases} \tag{2.7}$$

With Π defined by (2.6), we derive the following relationships based on (2.7):

$$B'\Pi B = b_d^2 \qquad\qquad B'\Pi A = b_d CA^d$$

Thus the optimal control law has the form

$$u(k) = -\frac{CA^d}{b_d}x(k) \tag{2.8}$$

Since the true state, $x(k)$, is unknown, we use the optimal estimator for $x(k)$ in (2.8). For the linear system, this estimate comes from the Kalman filter. Note that $\hat{x}(k)$ as defined in (2.3) is the Kalman *predictor*, not the Kalman *filter*, so $\hat{x}(k) = \hat{x}(k|k-1)$ does not include any information about the measurement $y(k)$. We obtain a better estimate of $x(k)$ by using the Kalman filter, $\hat{x}(k|k) = \hat{x}(k) + L_f(y(k) - C\hat{x}(k))$. L_f is the Kalman filter gain; for A invertible we have $L_f = A^{-1}L$, where L is the Kalman predictor gain. We are guaranteed to have A invertible in this formulation, provided that $a_n \neq 0$ (the system order is minimal). Then we write the Kalman filter for $x(k)$ as

$$\hat{x}(k|k) = \hat{x}(k) + A^{-1}Le(k)$$

and the control law as

$$u(k) = -\frac{CA^d}{b_d}\hat{x}(k) - \frac{CA^{d-1}L}{b_d}e(k) \tag{2.9}$$

We then write $\hat{x}(k+d)$ and $y(k+d)$ as

$$\hat{x}(k+d) = A^d\hat{x}(k) + \sum_{i=0}^{d-1}A^{d-i-1}Bu(k+i) + \sum_{i=0}^{d-1}A^{d-i-1}Le(k+i)$$

$$y(k+d) = CA^d\hat{x}(k) + C\sum_{i=0}^{d-1}A^{d-i-1}Bu(k+i) + C\sum_{i=0}^{d-1}A^{d-i-1}Le(k+i)$$

$$+ e(k+d) \tag{2.10}$$

Due to the time delay, $CA^iB = 0$ for all $i < d-1$, and therefore (2.10) reduces to

$$y(k+d) = CA^d\hat{x}(k) + CA^{d-1}Bu(k) + C\sum_{i=0}^{d-1} A^{d-i-1}Le(k+i) + e(k+d) \qquad (2.11)$$

Letting $u(k)$ be the minimum variance control law in (2.9) and noting that $\frac{CA^{d-1}B}{b_d} = 1$, then we rewrite the second term in (2.11) as

$$CA^{d-1}Bu(k) = -CA^dx(k) - CA^{d-1}Le(k)$$

From (2.11), shifting the time index to k rather than $k+d$, we write the output under minimum variance control as

$$y(k) := y_{\mathrm{mv}}(k) = C\sum_{i=1}^{d-1} A^{d-i-1}Le(i) + e(k) \qquad (2.12)$$

The variance of y_{mv}, σ_{mv}^2, is the minimum output variance possible under any feedback control law and has the form

$$\sigma_{\mathrm{mv}}^2 = R_e \sum_{i=1}^{d-1}\left(CA^{d-1-i}LL'A'^{d-1-i}C'\right) + R_e \qquad (2.13)$$

in which $R_e := \mathrm{var}(e)$.

2.2.2   *Feedback invariant perspective*

Harris's approach views the minimum variance output as a feedback invariant term: a portion of the output that cannot be affected by any feedback controller due to the time delays of the system. Many extensions of the minimum variance benchmark have also been based on the concept of identifying the feedback invariant portion of the output. Here, we derive the minimum variance control law from the feedback invariant perspective.

We begin by examining the output as written in (2.11). Since at time k, we have

no information about $e(k+i)$ for $i > 0$, we cannot use the input $u(k)$ to remove these disturbances from $y(k+d)$. Therefore, the best possible output we could obtain is

$$y_{mv}(k) = C \sum_{i=1}^{d-1} A^{d-i-1} Le(i) + e(k)$$

Note that this output is the same as in (2.12). In order to achieve this output, we must have the remaining terms in (2.11) equal to zero:

$$CA^d x(k) + CA^{d-1} Bu(k) + CA^{d-1} Le(k) = 0.$$

Noting that $CA^{d-1}B = b_d$ and solving for $u(k)$, we arrive at the control law

$$u(k) = -\frac{CA^d}{b_d} x(k) - \frac{CA^{d-1}G}{b_d} v(k)$$

which is identical to the control law in (2.9) derived from the Riccati equation.

2.2.3 *Lower bound on minimum variance*

From (2.13), we observe that the minimum variance is always greater than or equal to the variance of the innovation, R_e. We write the variance of the innovation in terms of the process and measurement noise:

$$R_e = P_{11} + R_v \qquad P = APA' - APC'(P_{11} + R_v)^{-1} CPA' + Q_w$$

in which P_{11} is the $(1,1)$ element of P defined by the Riccati equation above. We make the following observations:

1. If $d = 1$, then $y_{mv} = e(k)$ and $\sigma_{mv}^2 = R_e$

2. If $d = 1$ and $Q_w = 0$, then the minimum variance equals to the measurement noise.

3. If $d > 0$, then in general we expect that the additional terms in (2.13) be strictly positive and $\sigma_{mv}^2 > R_e$ (although given free choice of Q_w and R_v, we could choose a specific noise model such that $\sigma_{mv}^2 = R_e$).

2.2.4 *Non-invertible zeros*

Thus far, the stability of the closed-loop system has not been discussed. In state space form, the system is closed-loop stable if every eigenvalue of $A + BK$ has magnitude less than one. In the time series equation, stability is determined by the zeros of polynomials written from the coefficients. In the time series form, the minimum variance control law given in Åström (1970) leads to a stable closed-loop system if and only if the polynomial defined by the b_i coefficients

$$b_1 z^{n-1} + b_2 z^{n-2} + \cdots + b_n - 1z + b_n$$

has all its roots within the unit circle. If any roots lie outside the unit circle, the minimum variance control law is not be stable (Åström and Wittenmark, 1984). If the state space model is written for such a time series equation, the solution to the Riccati equation given by (2.6) no longer produces a stabilizing control law (Kwong, 1991). A unique stabilizing solution to the Riccati equation may still exist; however, the control law based on this solution always results in a variance greater than R_e, even for a unit time delay (Hewer, 1971, 1973).

3

PERFORMANCE MONITORING BENCHMARK AND
DISTRIBUTION[1]

3.1 KEY CONCEPTS FOR LQG MONITORING

Controller performance monitoring for a linear quadratic Gaussian (LQG) system
can be summarized by the following three fundamental relationships:

$$\text{central limit theorem} \implies \text{normal distribution}$$

$$\text{normal} + \text{linear system} \implies \text{normal distribution}$$

$$\text{normal} + \text{quadratic stage cost} \implies \text{chi-squared distribution}$$

Control engineers implicitly use the first relationship to develop the noise model
for the system by assuming that the disturbances affecting the process and mea-
surement have a normal distribution. It is necessary to assume a distribution for
these noises in order to calculate the distribution of any signals of the system.
If the process and measurement disturbances are the result of many unmodeled
random effects, then the central limit theorem justifies the choice of a normal dis-

---

[1]Portions of this chapter have been published in Zagrobelny, Ji, and Rawlings (2013)

tribution. The common choice of a *linear* input to output process model is based mainly on user convenience and the expectation that the controller can successfully maintain the process close to some desired operating point. Near this point, the linear model may provide a good approximation, even while the system may be non-linear over a wider range of operating conditions. The identification of this linear model from data usually justifies assuming the normally distributed disturbances are zero mean. If non-stationary disturbances are present, they cannot be approximated well by the central limit theorem. Therefore, these disturbances must be known or approximated in order to determine the distribution of any signals in the system.

The second fundamental relationship states that for a linear process model with normally distributed noises, any signal generated by this system also has a normal distribution. This step follows directly from the property that a linear transformation of a normally distributed random variable is also normal. This relationship is powerful because it gives us knowledge of the distribution of *all* signals in the system (output, input, innovation, state, state estimate, *etc*.). When the system is stable, the signals also have time-independent distributions and finite variances.

Finally, the third fundamental relationship allows us to characterize the regulator's stage cost, *i.e.*, the function of the system's variables that the controller optimizes. When the stage cost is a quadratic form of normal signals from the process, the stage cost has a *generalized chi-squared distribution*. We use the expectation of the stage cost as the performance metric to monitor the controller. By understanding its distribution, we can develop confidence intervals and make more accurate comparisons between the ideal and achieved performances.

We refer to the controller performance benchmark as the key performance index, or KPI. To derive the expressions for KPI expectation and variance, we follow the steps below:

1. Write general expressions for the state/disturbance and their estimates.

2. Express the input, u, in terms of the state and estimate (by assuming a linear control law).

3. Write a closed loop expression for the state/disturbance and their estimates in which the only external input is the random noise affecting the system.

4. Derive a probability distribution for the state/disturbance and their estimates.

5. Derive a probability distribution for the signal used in KPI calculation.

6. Derive the probability distribution for the KPI.

3.2.1 *Closed-loop expression for the linear system*

We begin by expressing the the system and its estimator in state space form as

$$x^+ = Ax + Bu + B_d d + Gw \qquad \hat{x}^+ = A\hat{x} + AL_x \varepsilon + Bu + B_d \hat{d}$$
$$+ B_d L_d \varepsilon$$
$$d^+ = d + G_d w_d \qquad \qquad \hat{d}^+ = \hat{d} + L_d \varepsilon$$
$$y = Cx + C_d d + v \qquad \qquad \varepsilon = y - C\hat{x} - C_d \hat{d} \qquad (3.1)$$

in which the state x is augmented to include the past input. This augmented state is convenient for including the rate-of-change penalty in the regulator. We

assume that the process noise w and the measurement noise v are independent, zero mean, and normally distributed with variances Q_w and R_v. The integrating disturbance d, affected by white noise w_d, is added to achieve the desired zero-offset properties. We denote the variance of $\begin{bmatrix} w' & w_d' \end{bmatrix}'$ as Q_W. We discuss how to estimate the variances Q_W and R_v from data in Chapters 4-8. The estimator in (3.1) should match what is currently used by the controller and does not need to be optimal.

The following steps utilize the fact that all signals for the linear system are distributed normally. We also take $G_d = 0$ to ensure that the variance of the disturbance d is bounded.

We augment the system so that both the state and disturbance are treated as a single vector. Defining the augmented state and its estimate, $X := [x' \quad d']'$, $\hat{X} := [\hat{x}' \quad \hat{d}']'$, we have:

$$X^+ = A_{\text{aug}}X + B_{\text{aug}}u + G_{\text{aug}}W \quad \hat{X}^+ = A_{\text{aug}}\hat{X} + A_{\text{aug}}L_{\text{aug}}\varepsilon + B_{\text{aug}}u$$

$$y = C_{\text{aug}}X + v \qquad\qquad \varepsilon = y - C_{\text{aug}}\hat{X}$$

in which $W := \begin{bmatrix} w \\ w_d \end{bmatrix}$, $A_{\text{aug}} := \begin{bmatrix} A & B_d \\ 0 & I \end{bmatrix}$, $B_{\text{aug}} := \begin{bmatrix} B \\ 0 \end{bmatrix}$, $G_{\text{aug}} := \begin{bmatrix} G & 0 \\ 0 & G_d \end{bmatrix}$, $C_{\text{aug}} := \begin{bmatrix} C & C_d \end{bmatrix}$, and $L_{\text{aug}} := \begin{bmatrix} L_x \\ L_d \end{bmatrix}$. When a linear feedback control law is applied, the closed-loop system remains linear, and we can express this control law as

$$u(k) = K_{\text{aug}}\hat{X} + K_{\text{aug}}L_{\text{aug}}C_{\text{aug}}\tilde{X} + K_{\text{aug}}L_{\text{aug}}v + u_d$$

in which $\tilde{X} := X - \hat{X}$. K_{aug} and u_d are defined in (3.13) in Appendix 3.A. u_d is the constant term in every input and is only nonzero when nonzero setpoints are

present. We eliminate u from the system and write the closed-loop equation as

$$z^+ = \tilde{A}z + \tilde{G}\tilde{w} + z_d$$

in which

$$z := \begin{bmatrix} \hat{x} \\ \hat{d} \\ \hat{x} - x \\ \hat{d} - d \end{bmatrix} \quad \tilde{w} := \begin{bmatrix} w \\ w_d \\ v \end{bmatrix} \quad z_d := \begin{bmatrix} B_{\text{aug}} \\ 0 \end{bmatrix} u_d$$

$$\tilde{A} := \begin{bmatrix} A_{\text{aug}} + B_{\text{aug}}K_{\text{aug}} & (A_{\text{aug}} + B_{\text{aug}}K_{\text{aug}})LC_{\text{aug}} \\ 0 & A_{\text{aug}} - A_{\text{aug}}L_{\text{aug}}C_{\text{aug}} \end{bmatrix}$$

$$\tilde{G} := \begin{bmatrix} 0 & (A_{\text{aug}} + B_{\text{aug}}K_{\text{aug}})L \\ G_{\text{aug}} & -A_{\text{aug}}L \end{bmatrix} \tag{3.2}$$

The matrices $\tilde{A}$ and $\tilde{G}$ depend on the system matrices, as well as the estimator and regulator gains. The constant z_d depends on the solution to the target tracking problem and is nonzero for a nonzero setpoint in y or u. More details of this derivation are provided in Zagrobelny et al. (2013).

Since the only external input to z is the normally distributed vector $\tilde{w}$ and the constant term z_d, z is also a normally distributed vector (provided $z(0)$ is also normal). Letting $Q_{\tilde{w}} = \text{diag}(Q_W, R_v)$ be the covariance of $\tilde{w}$, then $z(k) \sim N(m(k), S(k))$ in which m and S satisfy the dynamic equations

$$m^+ = \tilde{A}m + z_d$$

$$S^+ = \tilde{A}S\tilde{A}' + \tilde{G}Q_{\tilde{w}}\tilde{G}'$$

Assuming that the system is stable[1] once the sample time is large enough to re-

---

[1]$\tilde{A}$ is stable as long as (A, B) is controllable and (A, C) is detectable; therefore $(I - \tilde{A})$ is also

move effects of the distribution of $z(0)$, z converges to the asymptotic distribution $z \sim N\left(m_\infty, S_\infty\right)$, with expectation and variance

$$m_\infty = (I - \tilde{A})^{-1} z_d$$

$$S_\infty = \tilde{A} S_\infty \tilde{A}' + \tilde{G} Q_{\tilde{w}} \tilde{G}'$$

Any signal from the system can be derived as a linear transformation of z and therefore also has a normal distribution.

3.2.2 *Distribution of stage cost and plant KPI*

We choose the stage cost as

$$\ell(k) := \ell\left(x(k), u(k)\right) = |y(k) - y_{sp}|_{Q_y}^2 + |u(k) - u_{sp}|_R^2 + |u(k) - u(k-1)|_S^2$$

Since the regulator is designed to minimize this stage cost, $\ell(k)$ serves as a natural choice for the controller performance metric. The ideal stage cost can be compared to the time-averaged stage cost from the controller data:

$$\langle \ell(k) \rangle = \frac{1}{k} \sum_{j=0}^{k-1} \ell(j) \tag{3.3}$$

We refer to both the expectation of the stage cost (from the model) and sample average of the stage cost (from the data) as the key performance index (KPI). Based on the linear model of the process and the normal distribution of the signals, we derive an analytical expression of the expectation of the stage cost, $\mathbb{E}(\ell(k))$. When the controller is performing optimally, $\langle \ell(k) \rangle$ converges to this value.

---

invertible and m_∞ and S_∞ exist.

The stage cost is equivalently expressed as the quadratic form

$$\ell(k) = f(k)'\tilde{Q}f(k) \tag{3.4}$$

$$f(k) := \begin{bmatrix} y(k) - y_{sp} \\ u(k) - u_{sp} \\ (u(k) - u(k-1)) \end{bmatrix} \qquad \tilde{Q} := \mathrm{diag}(Q_y, R, S)$$

in which $\tilde{Q} := \mathrm{diag}(Q_y, R, S)$. The vector $f(k)$ is simply a linear transformation of the modified state, z, and measurement noise, v:

$$f(k) = F_1 z(k) + F_2 v(k) + F_3 u_d - m_{sp}$$

in which

$$F_1 = \begin{bmatrix} C_{\mathrm{aug}} & C_{\mathrm{aug}} \\ K_{\mathrm{aug}} & K_{\mathrm{aug}} L_{\mathrm{aug}} C_{\mathrm{aug}} \\ K_{\mathrm{aug}} - h_1 & K_{\mathrm{aug}} L_{\mathrm{aug}} C_{\mathrm{aug}} - h_1 \end{bmatrix} \qquad F_2 = \begin{bmatrix} I \\ K_{\mathrm{aug}} L_{\mathrm{aug}} \\ K_{\mathrm{aug}} L_{\mathrm{aug}} \end{bmatrix} \qquad F_3 = \begin{bmatrix} 0 \\ I \\ I \end{bmatrix} \tag{3.5}$$

and

$$h_1 = \begin{bmatrix} 0_{m \times n} & I_m & 0_{m \times p} \end{bmatrix} \qquad m_{sp} = \begin{bmatrix} y'_{sp} & u'_{sp} & 0 \end{bmatrix}'$$

Therefore, $f(k)$ has the distribution

$$f(k) \sim N(\tilde{m}(k), P(k))$$

$$\tilde{m}(k) = F_1 m(k) + F_3 u_d - m_{sp}$$

$$P(k) = F_1 S(k) F_1' + F_2 R_v F_2'$$

When m_∞ and S_∞ exist, then we calculate the time-invariant mean and variance

of f as

$$\tilde{m}_\infty = F_1 m_\infty + F_3 u_d - m_{sp}$$

$$P_\infty = F_1 S_\infty F_1' + F_2 R_v F_2'$$

As the quadratic form of a normal variable, the stage cost has a generalized chi-squared distribution, whose parameters are the matrix of the quadratic form, $\tilde{Q}$; the mean of the signal, $\tilde{m}$; and variance of the signal, P (Cacoullos and Koutras, 1984). Although the generalized chi-squared distribution does not have a simple explicit expression, from (3.4) we calculate the mean and variance of the stage cost at time k as

$$\mathbb{E}(\ell(k)) = \text{tr}(\tilde{Q}P(k)) + \tilde{m}(k)'\tilde{Q}\tilde{m}(k) \tag{3.6}$$

$$\text{var}(\ell(k)) = 2\text{tr}(\tilde{Q}P(k)\tilde{Q}P(k)) + 4\tilde{m}(k)'\tilde{Q}P(K)\tilde{Q}\tilde{m}(k). \tag{3.7}$$

If the conditions are satisfied for z to have a time-independent distribution, then the distribution of ℓ is also time independent. The time-independent expressions for the expectation of the stage cost and variance take the same form, but $\tilde{m}(k)$ and $P(k)$ are replaced by $\tilde{m}_\infty$ and P_∞:

$$\mathbb{E}(\ell_\infty) = \text{tr}(\tilde{Q}P_\infty) + \tilde{m}_\infty'\tilde{Q}\tilde{m}_\infty \tag{3.8}$$

$$\text{var}(\ell_\infty) = 2\text{tr}(\tilde{Q}P_\infty\tilde{Q}P_\infty) + 4\tilde{m}_\infty'\tilde{Q}P_\infty\tilde{Q}\tilde{m}_\infty \tag{3.9}$$

in which ℓ_∞ denotes the steady-state stage cost (Searle, 1971 p.55–57, Mathai and Provost, 1992 chap. 3-4). While (3.6) and (3.8) hold true regardless of the distribution of z, the expression for the variance given in (3.7) and (3.9) requires that the signal be normally distributed.

The plant KPI is the sample average of the stage cost, so by the central limit theorem, as the number of samples increases, the variance of the plant KPI ap-

proaches a normal distribution. Letting $\mathcal{K}$ denote the plant KPI,

$$\mathcal{K} = \frac{1}{T} \sum_{k=1}^{T} \ell(k)$$

in which T samples of the stage cost are used. The mean of the KPI is equal to the mean of the stage cost:

$$\mathbb{E}(\mathcal{K}) = \frac{1}{T} \sum_{k=1}^{T} \mathbb{E}(\ell(k)) = \text{tr}(\tilde{Q}P_{\infty}) + \tilde{m}'\tilde{Q}\tilde{m}$$

If the samples of the stage cost were independent, then the variance would be equally simple:

$$\text{var}(\mathcal{K}) = \text{var}\left(\frac{1}{T} \sum_{k=1}^{T} \ell(k)\right) = \frac{1}{T^2} \sum_{k=1}^{T} \text{var}\left(\ell(k)\right))$$
$$= \frac{1}{T} \left(2\text{tr}(\tilde{Q}P_{\infty}\tilde{Q}P_{\infty}) + 4\tilde{m}'\tilde{Q}P_{\infty}\tilde{Q}\tilde{m}\right)$$

However, in reality, each $\ell(k)$ and $\ell(k+j)$ are correlated, since each $f(k)$ and $f(k+j)$ are correlated. Noting that

$$f(k) = F_1 z(k) + F_2 v(k) + \text{ constant}$$
$$f(k+j) = F_1 \tilde{A}^j z(k) + F_1 \sum_{i=0}^{j-1} A t^{j-i-1} \tilde{G}\tilde{w}(k+i) + F_2 v(k+j) + \text{ constant}$$

we see that the correlation between $f(k)$ and $f(k+j)$ has two contributions:

(1) $z(k)$ affects both $f(k)$ and $f(k+j)$

(2) $v(k)$ affects $f(k)$ directly and affects $f(k+j)$ through the noise term $\tilde{w}(k)$

Accounting for these correlations, the variance of the KPI is written as

$$
\begin{aligned}
\text{var}(\mathcal{K}) = {} & \frac{2}{T}\text{tr}(\tilde{Q}P_\infty\tilde{Q}P_\infty) + \frac{4}{T}\left(\tilde{m}'\tilde{Q}P_\infty\tilde{Q}\tilde{m}\right) \\
& + \frac{4}{T^2}\left(\sum_{j=1}^{T-1}(T-j)\text{tr}(\tilde{Q}(S_{fj}+S_{wj})\tilde{Q}(S_{fj}+S_{wj})')\right) \\
& + \frac{8}{T^2}\sum_{j=1}^{T-1}(T-j)\left(\tilde{m}'\tilde{Q}(S_{fj}+S_{wj})\tilde{Q}\tilde{m}\right)
\end{aligned}
\tag{3.10}
$$

in which $S_{fj} := F_1 S_\infty(\tilde{A}')^j F_1'$ and $S_{wj} := F_2 I_w \tilde{Q}_w \tilde{G}'(\tilde{A}')^{j-1}F_1'$. Note that S_{fj} and S_{wj} depend on the index j and therefore cannot be factored out of the sum in (3.10). S_{fj} and S_{wj} correspond to the correlation between $f(k)$ and $f(k+j)$ due to $z(k)$ and $v(k)$, respectively. (3.10) is derived in Appendix 3.B.

3.3 SIMULATION OF BENCHMARK DISTRIBUTION

While no closed-form probability distribution can be written for the stage cost, we use a simulation to demonstrate its distribution. A linear unconstrained system, with a perfect model and no deterministic disturbances, was simulated to generate a set of outputs and inputs. From these data points, the stage cost, $\ell(k)$, was calculated at each time step according to (3.4). The stage costs are presented as a histogram in Figure 3.1, with the sample mean and variance indicated. The histogram shows the asymmetric density of the stage cost. Because the density is highest close to zero, only an upper limit should be necessary when estimating a confidence interval. As shown in Table 3.1, the sample mean and variance agree well with their theoretical values calculated from the analytical formulas.

We used the same example to illustrate the distribution of the plant KPI. To compute this histogram, we performed 1000 independent simulations for the same system. In each simulation, we calculated $\ell(k)$ at each time step k, and then found the time average, $\langle\ell(k)\rangle$. A histogram of these sample averages is plotted in Figure
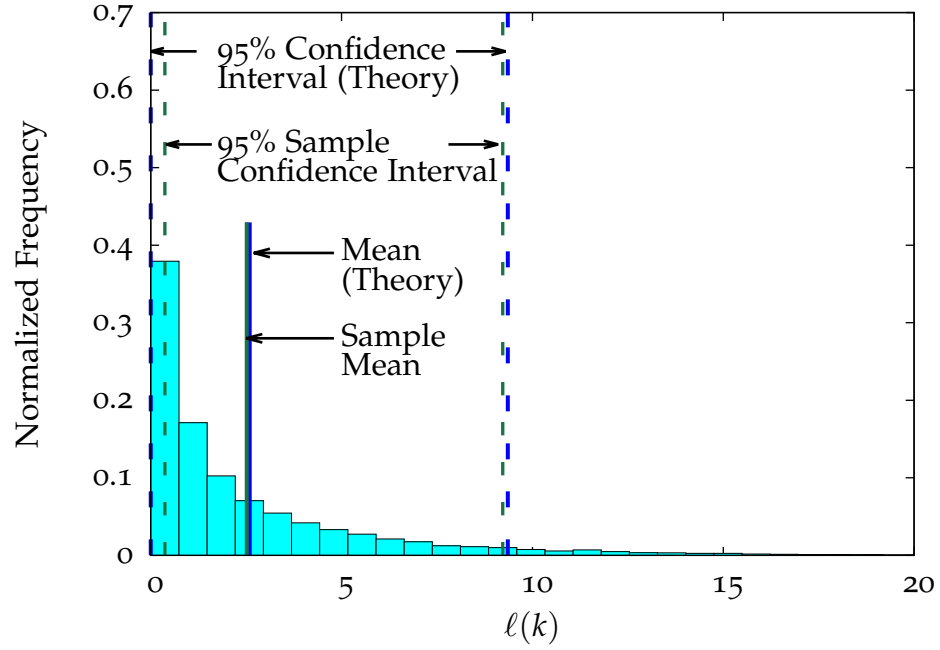
Figure 3.1: Histogram of the stage cost, $\ell(k)$ for 10000 samples. The solid green line indicates the sample mean (achieved KPI) and is close to the theoretical expectation (solid blue line). The thick dashed blue lines show two standard deviations from the mean (using the theoretical values), and the thin dashed green lines lines indicate the bins that contain 95% of the points.

3.2. The histogram shows that despite the fact that $\ell(k)$ is distributed according to the generalized chi-squared distribution, the sample average approaches a normal distribution. We overlay two normal probability density functions on the histogram. Both densities use the theoretical mean for the KPI. The first density uses the theoretical variance calculated under the (false) assumption that $\ell(k)$ and $\ell(j)$ are independent. This variance is clearly too small compared to the histogram data. The second normal distribution uses the theoretical variance given in (3.10),

Table 3.1: Sample and theoretical mean and variance for the stage cost example.

	Mean	Variance
Sample	2.52	11.39
Theoretical	2.56	11.55

which correctly accounts for the correlations between $\ell(k)$ and $\ell(j)$. This second density provides an excellent fit for the data in the histogram.
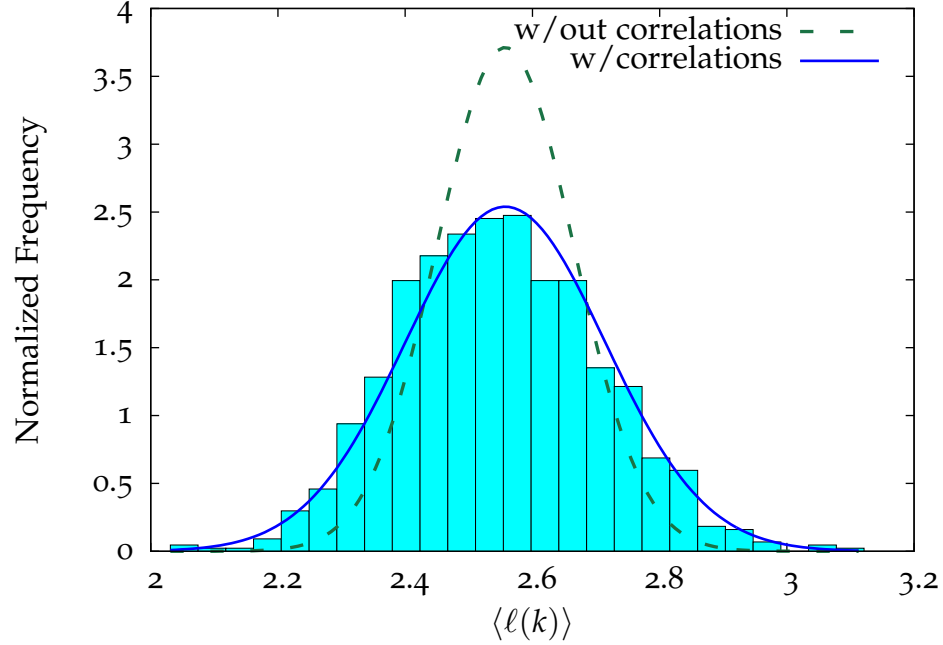
Figure 3.2: Histogram of the time-averaged stage cost (plant KPI) for 1000 simulations. The data looks approximately normally distributed and is compared to theoretical probability density functions for two normal distributions. The dashed green line uses the theoretical variance that does not account for correlations between the stage cost samples, and therefore is too narrow. The blue line uses the theoretical variance that correctly accounts for the correlations; this density has an excellent fit to the histogram data.

3.4 BENCHMARK CALCULATION WITH UNMODELED DETERMINISTIC DISTURBANCES

Zagrobelny et al. (2013) derive the theoretical KPI for the case when there is an unmodeled deterministic disturbance corrupting the output, i.e.:

$$y = Cx + p + v$$

Here we generalize this derivation for the case of a deterministic disturbance that

may affect both the input and the output:

$$x^+ = Ax + Bu + B_p p + Gw \qquad\qquad y = Cx + C_p p + v$$

Although p affects y in a manner analogous to d, we treat p in a deterministic manner, *i.e.* we assume that we know or can approximate p, rather than trying describing this disturbance through a probability distribution. If we know or can approximate the deterministic disturbance, p, for a limited number of samples $(0, T)$, the theoretical KPI can be compared with the actual plant KPI averaged over a much longer time period, by the assuming the disturbance is periodic. The KPI expectation in the presence of p represents the best performance achievable given that the plant is affected by this disturbance.

The dynamic model including the disturbance p is written as

$$z^+ = \tilde{A}z + \tilde{B}_p p + \tilde{G}\tilde{w} + z_d$$

in which

$$\tilde{B}_p = \begin{bmatrix} \left(A_{\text{aug}} + B_{\text{aug}}K_{\text{aug}}\right) LC_p \\ B_{p,\text{aug}} - A_{\text{aug}}LC_p \end{bmatrix} \qquad\qquad B_{p,\text{aug}} = \begin{bmatrix} B_p \\ 0 \end{bmatrix} \qquad (3.11)$$

$B_{p,\text{aug}}$ is the augmented version of B_p, and $\tilde{A}$, $\tilde{G}$, and z_d have the same form as previously given in (3.2). In Zagrobelny et al. (2013), $\tilde{B}_p$ takes the form

$$\tilde{B}_p = \begin{bmatrix} \left(A_{\text{aug}} + B_{\text{aug}}K_{\text{aug}}\right) L_{\text{aug}} \\ -A_{\text{aug}}L_{\text{aug}} \end{bmatrix}$$

which is equivalent to (3.11) when we assume an output disturbance model ($B_p = 0$ and $C_p = I$).

Because p is deterministic, it affects the expectation of z but not its variance:

$$m^+ = \tilde{A}m + \tilde{B}_p p + z_d$$

$$S_\infty^+ = \tilde{A}S_\infty \tilde{A}' + \tilde{G}Q_{\tilde{w}}\tilde{G}'$$

We rewrite signal of interest, $f(k)$ as

$$f(k) = F_1 z(k) + F_2 C_p p(k) + F_2 v(k) + F_3 u_d - m_{sp}$$

in which F_1, F_2, F_3, and m_{sp} are as previously defined in (3.5). This signal still has a normal distribution, with the time varying mean and variance expressed as

$$\tilde{m}(k) = F_1 m(k) + F_2 C_p p(k) + F_3 u_d - m_{sp}$$

$$P(k) = F_1 S(k) F_1' + F_2 R_v F_2'$$

Assuming that we know (or can estimate) the disturbance for T time points, we write the expectation of the stage cost at time k as

$$\mathbb{E}(\ell_\infty) = \frac{1}{T}\sum_{k=1}^{T}\mathbb{E}(\ell(k))$$

Since the variance of z converges to a constant value, we write this expectation as

$$\mathbb{E}(\ell_\infty) = \text{tr}(\tilde{Q}P) + \frac{1}{T}\sum_{k=1}^{T}\mathbb{E}(\tilde{m}(k)'\tilde{Q}\tilde{m}(k))$$

The first term accounts for the effects of the white process and measurement noises, whereas the second term accounts for the effects of the deterministic disturbances.

The KPI for deterministic disturbances may also be useful for performance monitoring during plant transitions, as changes in setpoints can be considered as deterministic disturbances. In this case, the "disturbance" is known exactly and

need not be approximated.

3.5 BENCHMARK CALCULATION FROM THE INNOVATION FORM OF THE MODEL

As discussed in Section 4.1, several methods for disturbance model identification focus on estimating the optimal filter gain and corresponding innovation variance, rather than estimating the process and measurement noises Q_w and R_v. Thus, we seek to derive the KPI for the system in innovation form, which is written as

$$x^+ = Ax + Bu + B_d d + M_x e$$

$$d^+ = d + M_d e$$

$$y = Cx + C_d d + e$$

in which e is white noise with zero mean and variance R_e. M_x and M_d denote the optimal gains for the Kalman predictor, which we distinguish from the estimator gains currently used by the controller, L_x and L_d. L_x and L_d may be suboptimal, and therefore ε defined below may not be white. We write the current estimator for the system as before:

$$\hat{x}^+ = A\hat{x} + Bu + B_d \hat{d} + L_x \varepsilon$$

$$\hat{d}^+ = \hat{d} + L_d \varepsilon$$

$$\varepsilon = y - C\hat{x} - C_d \hat{d}$$

Again we define the augmented state $X := \begin{bmatrix} x' & d' \end{bmatrix}'$ and let $\tilde{X} = X - \hat{X}$. Then we write the augmented system as

$$X = A_{\text{aug}} X + B_{\text{aug}} u + M_{\text{aug}} e \qquad \hat{X} = A_{\text{aug}} \hat{X} + A_{\text{aug}} L_{\text{aug}} C_{\text{aug}} \varepsilon$$

$$Y = C_{\text{aug}} X + e \qquad \varepsilon = y - C_{\text{aug}} \hat{X}$$

in which A_{aug}, B_{aug}, and L_{aug} are defined in (3.2) and $M_{\text{aug}} = \begin{bmatrix} M'_x & M'_d \end{bmatrix}'$. Note that we must have $M_d = 0$ for the system to be stable. The input has essentially the same form as before:

$$u = K_{\text{aug}}\hat{X} + K_{\text{aug}}L_{\text{aug}}C_{\text{aug}}\tilde{X} + K_{\text{aug}}L_{\text{aug}}e + u_d$$

in which K_{aug} and u_d are as defined in (3.13) in Appendix 3.A.

Again, letting $z = \begin{bmatrix} \hat{X}' & \tilde{X}' \end{bmatrix}'$ we write the closed-loop system as

$$z^+ = \tilde{A}z + \tilde{M}e + z_d \qquad \tilde{M} = \begin{bmatrix} \left(A_{\text{aug}} + B_{\text{aug}}K_{\text{aug}}\right)L_{\text{aug}} \\ M_{\text{aug}} - A_{\text{aug}}C_{\text{aug}} \end{bmatrix}$$

in which $\tilde{A}$ and z_d are as previously defined. Once again we define

$$m^+ = \tilde{A}m + z_d \qquad S^+ = \tilde{A}S\tilde{A}' + \tilde{M}R_e\tilde{M}'$$

Note that the Lyapunov equation for S has changed slightly to account for the different noise model. Using this definition of S, we derive the distribution of the signal of interest and of the KPI in the same manner as before, replacing $v(k)$ with $e(k)$ and R_v with R_e as necessary.

3.6 BENCHMARK CALCULATION WITH PLANT MODEL MISMATCH

Zagrobelny et al. (2013) also derive the KPI expectation for a system with plant model mismatch. In this derivation, (A_p, B_p, C_p) denotes the true plant behavior and (A_m, B_m, C_m) denotes the model behavior. We do not repeat the derivation here but suggest another application for it (in addition to giving a better theoretical understanding of the behavior under plant model mismatch). Suppose the model (A_m, B_m, C_m) is currently used in the controller, but a more accurate model for the

system has been identified, which we call (A_{m2}, B_{m2}, C_{m2}). We would not expect the plant KPI to equal the theoretical KPI using the m-model, as this model does not accurately describe the system. Neither would we expect the theoretical KPI with the $m2$-model to match the plant KPI, as the $m2$-model is not used in the controller. Instead, we propose using the mismatch formulas to calculate the KPI given that the system is described by the m-model and the controller uses the $m2$-model. This theoretical KPI corresponds to the expected performance of the current controller. Since the KPI calculated from the $m2$-model corresponds to the theoretical performance when the model used by the controller has been updated, comparing the theoretical $m/m2$ KPI with the KPI using $m2$ alone would give an idea as to the benefit implementing the new model in the controller.

3.7 CONSTRAINED AND NONLINEAR MPC MONITORING

In addition to the LQG monitoring problem, the monitoring of constrained and nonlinear MPC controllers is necessary in industrial applications. The previously established fundamental relationships no longer apply because the second relationship is no longer true:

$$\text{central limit theorem} \implies \text{normal distribution}$$
$$\text{normal} + \text{nonlinear system} \;\not\!\!\!\implies\; \text{normal distribution}$$

Therefore we no longer have a distribution for the signals of the system and cannot calculate the expectation or variance of the monitoring benchmark.

When we have systems with constraints or nonlinear systems, we cannot take advantage of convenient analytical formulas in calculating metrics or describing their statistics. In place of these formulas, however, we can use simulations and Monte Carlo methods to estimate the metrics. By simulating the model, we can use the sample average of the simulation stage cost as our theoretical KPI. Zagrobelny

et al. (2013)) shows that the average stage cost for a linear, constrained system still converges to a constant value, despite the more complicated nature of the system. The simulated KPI serves as a theoretical benchmark based on the process model. We can also use Monte Carlo methods to estimate the variance and confidence intervals for the KPI. These confidence intervals allow comparison between the simulated ideal KPI and the plant KPI.

3.8 APPENDICES

3.A *Steady-state target problem*

We derive the linear control law that was used in Section 3.2.1 by defining the target selector problem as

$$\min_{(x_s, u_s)} \frac{1}{2} \left(|u_s - u_{sp}|^2_{R_s} + |Cx_s + C_d \hat{d}_s - y_{sp}|^2_{Q_s} \right) \tag{3.12}$$

$$\text{s.t.} \quad \begin{bmatrix} I - A & -B \\ HC & 0 \end{bmatrix} \begin{bmatrix} x_s \\ u_s \end{bmatrix} = \begin{bmatrix} B_d \hat{d}_s \\ r_{sp} - HC_d \hat{d}_s \end{bmatrix}$$

y_{sp} and u_{sp} are the external setpoints provided to the controller, and x_s and u_s are the target values for the state and input given to the regulator. $\hat{d}_s$ is the estimated steady-state value for the disturbance, which is approximated as $\hat{d} + L_d \varepsilon$. In the case that we have more outputs than inputs, we select a subset of the outputs (equal to the number of inputs), $r = Hy$, to control to the setpoint. If the number of outputs is less than or equal to the number of inputs, we choose $r = y$. Using the method of Lagrange multipliers, we solve (3.12) to express x_s and u_s as

$$\begin{bmatrix} x_s \\ u_s \end{bmatrix} = T_1 \begin{bmatrix} r_{sp} \\ y_{sp} \\ u_{sp} \end{bmatrix} + T_2 \hat{d}_s$$

in which T_1 and T_2 are given by

$$T_1 = \begin{bmatrix} I & 0 & 0 & 0 \\ 0 & I & 0 & 0 \end{bmatrix} \begin{bmatrix} C'Q_sC & 0 & -(I-A') & -C'H' \\ 0 & R_s & B' & 0 \\ I-A & -B & 0 & 0 \\ HC & 0 & 0 & 0 \end{bmatrix}^{-1} \begin{bmatrix} 0 & C'Q_s & 0 \\ 0 & 0 & R_s \\ 0 & 0 & 0 \\ I & 0 & 0 \end{bmatrix}$$

$$T_2 = \begin{bmatrix} I & 0 & 0 & 0 \\ 0 & I & 0 & 0 \end{bmatrix} \begin{bmatrix} C'Q_sC & 0 & -(I-A') & -C'H' \\ 0 & R_s & B' & 0 \\ I-A & -B & 0 & 0 \\ HC & 0 & 0 & 0 \end{bmatrix}^{-1} \begin{bmatrix} -C'Q_sC_d \\ 0 \\ B_d \\ -HC_d \end{bmatrix}$$

Then we write u as

$$u = K_{\text{aug}}\hat{X} + K_{\text{aug}}LC_{\text{aug}}\tilde{X} + K_{\text{aug}}Lv + u_d$$

$$u_d = \left(\begin{bmatrix} 0 & I \end{bmatrix} - K \right) T_1 \begin{bmatrix} r_{sp} \\ y_{sp} \\ u_{sp} \end{bmatrix}$$

$$K_{\text{aug}} = \begin{bmatrix} K & K_d \end{bmatrix}, \quad K_d = \left(\begin{bmatrix} 0 & I \end{bmatrix} - K \right) T_2 \tag{3.13}$$

u_d is the constant term in every input and is only nonzero when nonzero set-points are present. The inclusion of K_d is necessary to reject disturbances. See Zagrobelny et al. (2013) for a more detailed derivation.

3.B   *Derivation of KPI variance*

Before deriving the variance of the KPI, we note that for a vector $x \sim N(0, P)$, a vector y that is independent of x, and constant matrices A and B,

$$\mathbb{E}(x' A y y' B y) = 0 \tag{3.14}$$

$$\mathbb{E}(x' A x x' B y) = 0 \tag{3.15}$$

$$\mathbb{E}(x' A x x' B x) = \text{tr}(AP)\text{tr}(BP) + 2\text{tr}(APBP) \tag{3.16}$$

The first equality follows from the expectation of x. The second and third equations are derived from Theorems 3.2d.2 and 3.2d.3 of Mathai and Provost (1992), respectively, by considering the special case where the normal variable is zero mean.

We begin our derivation by defining the signals

$$f(k) = F_1 z(k) + F_2 v(k) + f_d \qquad z(k+1) = \tilde{A} z(k) + \tilde{G} \tilde{w}(k) + z_d$$

We assume that $f(k)$ and $z(k)$ have time-invariant distributions:

$$f \sim N(\tilde{m}, P) \qquad z \sim N(m, S)$$

in which we neglect the ∞ subscript on the variances. Note that $v(k)$ and $\tilde{w}(k)$ are correlated. Letting $I_w := \begin{bmatrix} 0 & 0 & I \end{bmatrix}$, then $v(k) = I_w \tilde{w}(k)$.

The stage cost, ℓ, is a quadratic form of the signal f and the KPI, $\mathcal{K}$, is the time average of ℓ

$$\ell(k) = f(k)' \tilde{Q} f(k) \qquad \mathcal{K} = \frac{1}{T} \sum_{k=1}^{T} \ell(k)$$

We write the variance of $\mathcal{K}$ as

$$\text{var}(\mathcal{K}) = \mathbb{E}(\mathcal{K}^2) - \mathbb{E}(\mathcal{K})^2 = \mathbb{E}(\mathcal{K}^2) - \left(\text{tr}(\tilde{Q}P) + \tilde{m}'\tilde{Q}\tilde{m}\right)^2 \tag{3.17}$$

From the definition of $\mathcal{K}$,

$$\mathbb{E}(\mathcal{K}^2) = \frac{1}{T^2} \sum_{k=1}^{T} \sum_{j=1}^{T} \mathbb{E}(\ell(k)\ell(j)) \tag{3.18}$$

Since $\ell(K)$ is a scalar, $\ell(k)\ell(j) = \ell(j)\ell(k)$, and since the distribution of f (and ℓ) is time independent, $\ell(k+i)\ell(k) = \ell(j+i)\ell(j)$. Therefore, we rewrite (3.18) as

$$\mathbb{E}(\mathcal{K}^2) = \frac{1}{T}\mathbb{E}\left(\ell(k)^2\right) + \frac{2}{T^2}\left[\sum_{j=1}^{T-1}(T-j)\,\mathbb{E}(\ell(k)\ell(k+j))\right] \tag{3.19}$$

Let $\bar{f}(k) = f(k) - \tilde{m}$ and $\bar{z}(k) = z - m$, so that these signals are zero mean. Then $\bar{z}^+ = \tilde{A}\bar{z} + \tilde{G}\tilde{w}$ and

$$\ell = \left(\bar{f} + \tilde{m}\right)' \tilde{Q} \left(\bar{f} + \tilde{m}\right) = \bar{\ell} + 2\bar{f}'\tilde{Q}\tilde{m} + \tilde{m}'\tilde{Q}\tilde{m}$$

in which $\bar{\ell} = \bar{f}'\tilde{Q}\bar{f}$. To simplify our notation, we drop the time index (k) and use the subscript j to denote the index $(k+j)$.

We next evaluate the term $\mathbb{E}(\ell\ell_j)$ appearing in (3.19):

$$\mathbb{E}(\ell\ell_j) = \mathbb{E}\left(\bar{\ell}\bar{\ell}_j + 2\bar{\ell}\bar{f}_j'\tilde{Q}\tilde{m} + \bar{\ell}\tilde{m}'\tilde{Q}\tilde{m} + 2\bar{f}'\tilde{Q}\tilde{m}\bar{\ell}_j + 4\bar{f}'\tilde{Q}\tilde{m}\bar{f}_j'\tilde{Q}\tilde{m} + 2\bar{f}'\tilde{Q}\tilde{m}\tilde{m}'\tilde{Q}\tilde{m}\right.$$
$$\left. + \tilde{m}'\tilde{Q}\tilde{m}\bar{\ell}_j + 2\tilde{m}'\tilde{Q}\tilde{m}\bar{f}_j'\tilde{Q}\tilde{m} + (\tilde{m}'\tilde{Q}\tilde{m})^2\right) \tag{3.20}$$

From (3.14), $\mathbb{E}(\bar{f}'\tilde{Q}\tilde{m}\tilde{m}'\tilde{Q}\tilde{m}) = \mathbb{E}(mt'\tilde{Q}\tilde{m}\bar{f}_j'\tilde{Q}\tilde{m}) = 0$, since $\mathbb{E}(\bar{f}) = \mathbb{E}(\bar{f}_j) = 0$. By writing out the terms $\mathbb{E}(\bar{\ell}\bar{f}_j'\tilde{Q}\tilde{m})$ and $\mathbb{E}(\bar{f}'\tilde{Q}\tilde{m}\bar{\ell}_j)$ in terms of z and w_j (which we do not show for the sake of brevity), we see that these terms are also zero, by

(3.14) and (3.15). Since $\mathbb{E}(\bar{\ell}) = \mathbb{E}(\bar{\ell}_j)$, (3.20) reduces to

$$\mathbb{E}(\ell\ell_j) = \mathbb{E}(\bar{\ell}\bar{\ell}_j) + 2\mathbb{E}(\bar{\ell})\tilde{m}'\tilde{Q}\tilde{m} + 4\tilde{m}'\tilde{Q}\mathbb{E}(\bar{f}'\bar{f}'_j)\tilde{Q}\tilde{m} + \left(\tilde{m}'\tilde{Q}\tilde{m}\right)^2 \qquad (3.21)$$

Further, since

$$\bar{f} = F_1\bar{z} + F_2 v \qquad\qquad \bar{f}_j = F_1\tilde{A}^j\bar{z} + F_1\sum_{i=0}^{j-1}\tilde{A}^{j-i-1}\tilde{G}\tilde{w}_i + F_2 v_j$$

then

$$\mathbb{E}(\bar{f}\bar{f}'_j) = F_1\mathbb{E}(\bar{z}\bar{z}')(\tilde{A}')^j F'_1 + F_2 I_w\mathbb{E}(\tilde{w}\tilde{w}')\tilde{G}'(\tilde{A}')^{j-1}F'_1$$

$$= F_1 S(\tilde{A}')^j F'_1 + F_2 I_w \tilde{Q}_w \tilde{G}'(\tilde{A}')^{j-1}F'_1$$

Let $S_{fj} := F_1 S(\tilde{A}')^j F'_1$ and $S_{wj} = F_2 I_w \tilde{Q}_w \tilde{G}'(\tilde{A}')^{j-1}F'_1$. Then $\mathbb{E}(\bar{f}\bar{f}'_j) = S_{fj} + S_{wj}$. Since $\mathbb{E}(\ell) = \mathrm{tr}(\tilde{Q}P)$, we write (3.21) as

$$\mathbb{E}(\ell\ell_j) = \mathbb{E}(\bar{\ell}\bar{\ell}_j) + 2\mathrm{tr}(\tilde{Q}P)\tilde{m}'\tilde{Q}\tilde{m} + 4\tilde{m}'\tilde{Q}S_{fj}\tilde{Q}\tilde{m} + 4\tilde{m}'\tilde{Q}S_{wj}\tilde{Q}\tilde{m} + (\tilde{m}\tilde{Q}\tilde{m})^2$$

$$\mathbb{E}(\ell\ell) = \mathbb{E}(\bar{\ell}\bar{\ell}) + 2\mathrm{tr}(\tilde{Q}P)\tilde{m}'\tilde{Q}\tilde{m} + 4\tilde{m}'\tilde{Q}P\tilde{Q}\tilde{m} + (\tilde{m}'\tilde{Q}\tilde{m})^2$$

From these expressions and the fact that $\frac{1}{T} + \frac{2}{T^2}\left(\sum_{j=1}^{T-1}(T-j)\right) = 1$, (3.19) reduces to

$$\mathbb{E}(\mathcal{K}^2) = \frac{1}{T}\mathbb{E}\left(\bar{\ell}^2\right) + \frac{4}{T}\left(\tilde{m}'\tilde{Q}P\tilde{Q}\tilde{m}\right) + \frac{2}{T^2}\left(\sum_{j=1}^{T-1}(T-j)\bar{\ell}\bar{\ell}_j\right)$$

$$+ \frac{8}{T^2}\sum_{j=1}^{T-1}(T-j)\mathrm{tr}\left(\tilde{Q}(S_{fj}+S_{wj})\tilde{Q}\tilde{m}\tilde{m}'\right)$$

$$+ 2\mathrm{tr}(\tilde{Q}P)m'\tilde{Q}m + (\tilde{m}'\tilde{Q}\tilde{m})^2 \qquad (3.22)$$

To further reduce (3.22), we find $\mathbb{E}(\bar{\ell}\bar{\ell}_j)$. Letting $\tilde{\omega} = \sum_{i=0}^{j-1}\tilde{A}^{j-i-1}\tilde{G}\tilde{w}$, we write

$\bar{\ell}$ and $\bar{\ell}_j$ as

$$\bar{\ell} = \bar{z}' F_1' \tilde{Q} F_1 \bar{z} + 2\bar{z}' F_1' \tilde{Q} F_2 I_w \tilde{w} + \tilde{w}' I_w' F_2' \tilde{Q} F_2 I_w \tilde{w}$$

$$\bar{\ell}_j = \bar{z}' (\tilde{A}')^j F_1' \tilde{Q} F_1 \tilde{A}^j \bar{z} + 2\bar{z}' (\tilde{A}')^j F_1' \tilde{Q} F_1 \tilde{\omega} + 2\bar{z}' (\tilde{A}')^j F_1' \tilde{Q} F_2 v_j + \tilde{\omega}' F_1' \tilde{Q} F_1 \tilde{\omega}$$

$$+ 2\tilde{\omega} F_1' \tilde{Q} F_2 v_j + v_j' F_2' \tilde{Q} F_2 v_j$$

Then we have

$$\mathbb{E}(\bar{\ell}\bar{\ell}_J) = \underbrace{\mathbb{E}\left(\bar{z}' F_1' \tilde{Q} F_1 \bar{z} \bar{z}' (\tilde{A}')^j F_1' \tilde{Q} F_1 \tilde{A}^j \bar{z}\right)}_{q_1} + \underbrace{\mathbb{E}\left(\bar{z}' F_1' \tilde{Q} F_1 \bar{z} \tilde{\omega}' F_1' \tilde{Q} F_1 \tilde{\omega}\right)}_{q_2}$$

$$+ \underbrace{\mathbb{E}\left(\bar{z}' F_1' \tilde{Q} F_1 \bar{z} v_j' F_2' \tilde{Q} F_2 v_j\right)}_{q_3} + 4\underbrace{\mathbb{E}\left(\bar{z}' F_1' \tilde{Q} F_2 I_w \tilde{w} \bar{z}' (\tilde{A}^j)' F_1' \tilde{Q} F_1 \tilde{\omega}\right)}_{q_4}$$

$$+ \underbrace{\mathbb{E}\left(\tilde{w}' I_w' F_2' \tilde{Q} F_2' I_w \tilde{w} \bar{z} (\tilde{A}')^j F_1' \tilde{Q} F_1 \bar{z} \tilde{A}^j \bar{z}\right)}_{q_5} + \underbrace{\mathbb{E}\left(\tilde{w}' I_w' F_2' \tilde{Q} F_2 I_w \tilde{w} \tilde{\omega}' F_1' \tilde{Q} F_1 \tilde{\omega}\right)}_{q_6}$$

$$+ \underbrace{\mathbb{E}\left(\tilde{w}' I_w' F_2' \tilde{Q} F_2' I_w \tilde{w} v_j' F_2' \tilde{Q} F_2 v_j\right)}_{q_7} \tag{3.23}$$

We reduce the odd terms as

$$q_1 = \text{tr}((\tilde{A}')^j F_1' \tilde{Q} F_1 \tilde{A}^j S) \text{tr}(F_1 \tilde{Q} F_1 S) + 2\text{tr}(F_1' \tilde{Q} F_1 S (\tilde{A}')^j F_1' \tilde{Q} F_1 \tilde{A}^j S)$$

$$q_3 = \text{tr}(F_1' \tilde{Q} F_1 S) \text{tr}(F_2' \tilde{Q} F_2 R_v)$$

$$q_5 = \text{tr}(F_2' \tilde{Q} F_2 R_v) \text{tr}((\tilde{A}')^j F_1' \tilde{Q} F_1 \tilde{A}^j S)$$

$$q_7 = \text{tr}^2(F_2' \tilde{Q} F_2 R_v)$$

To simplify q_2, we begin with

$$\mathbb{E}(\tilde{\omega}\tilde{\omega}') = \sum_{i=0}^{j-1} \tilde{A}^{j-i-1} \tilde{G} \tilde{Q}_w \tilde{G}' (\tilde{A}')^{j-i-1} = \sum_{i=0}^{j-1} \tilde{A}^i \tilde{G} \tilde{Q}_w \tilde{G}' (\tilde{A}')^i = S - \tilde{A}^j S (\tilde{A}')^j$$

in which the last step follows from the Lyapunov equation for S. Then q_2 reduces

as follows:

$$q_2 = \text{tr}(F_1'\tilde{Q}F_1S)\text{tr}(F_1'\tilde{Q}F_1\mathbb{E}(\tilde{\omega}\tilde{\omega}')) = \text{tr}^2(F_1'\tilde{Q}F_1S) - \text{tr}(F_1'\tilde{Q}F_1S)\text{tr}(F_1'\tilde{Q}F_1S\tilde{A}^jS(\tilde{A}')^j)$$

q_4 reduces according to

$$\begin{aligned}
q_4 &= \mathbb{E}\left(\tilde{z}'F_1'\tilde{Q}F_2 I_w\tilde{w}\tilde{z}'(\tilde{A}^j)'F_1'\tilde{Q}F_1\tilde{\omega}\right) = \mathbb{E}\left(\tilde{z}'F_1'\tilde{Q}F_2 I_w\tilde{w}\tilde{\omega}'F_1'\tilde{Q}F_1\tilde{A}^j\tilde{z}\right)\\
&= \text{tr}(F_1'\tilde{Q}F_2 I_w\tilde{Q}_w\tilde{G}'(\tilde{A}')^{j-1}F_1'\tilde{Q}F_1\tilde{A}^jS)
\end{aligned}$$

To simplify q_6, first we separate the term of the form $\tilde{w}'A'\tilde{w}\tilde{w}'B\tilde{w}$ from the terms of the form $\tilde{w}'A\tilde{w}wt_j'B\tilde{w}_j$. These last terms simplify easily since $\tilde{w}$ and $\tilde{w}_j$ are independent:

$$\begin{aligned}
q_6 &= \mathbb{E}\left(\tilde{w}'I_w'F_2'\tilde{Q}F_2 I_w\tilde{w}\tilde{\omega}'F_1'\tilde{Q}F_1\tilde{\omega}\right)\\
&= \mathbb{E}\left(\tilde{w}'I_w'F_2'\tilde{Q}F_2 I_w\tilde{w}\tilde{w}'\tilde{G}'(\tilde{A}')^{j-1}F_1'\tilde{Q}F_1\tilde{A}^{j-1}\tilde{G}\tilde{w}\right)\\
&\quad + \text{tr}(F_2'\tilde{Q}F_2 R_v)\sum_{i=1}^{j-1}\text{tr}(F_1'\tilde{Q}F_1\tilde{A}^{j-i-1}\tilde{G}\tilde{Q}_w\tilde{G}'(\tilde{A}')^{j-i-1})
\end{aligned}$$

We write out the first term using (3.16):

$$\begin{aligned}
q_6 &= 2\text{tr}\left(I_w'F_2'\tilde{Q}F_2 I_w\tilde{Q}_w\tilde{G}'(\tilde{A})'^{j-1}F_1'\tilde{Q}F_1\tilde{A}^{j-1}\tilde{G}\tilde{Q}_w\right)\\
&\quad + \text{tr}(F_2'\tilde{Q}F_2 R_v)\text{tr}((\tilde{A})'^{j-1}F_1'\tilde{Q}F_1\tilde{A}^{j-1}\tilde{G}\tilde{Q}_w\tilde{G}')\\
&\quad + \text{tr}(F_2'\tilde{Q}F_2 R_v)\sum_{i=0}^{j-2}\text{tr}(F_1'\tilde{Q}F_1\tilde{A}^i\tilde{G}\tilde{Q}_w\tilde{G}'(\tilde{A}')^i)
\end{aligned}$$

We further simplify q_6 based on the Lyapunov equation for S:

$$\begin{aligned}
q_6 &= 2\text{tr}\left(I_w'F_2'\tilde{Q}F_2 I_w\tilde{Q}_w\tilde{G}'(\tilde{A})'^{j-1}F_1'\tilde{Q}F_1\tilde{A}^{j-1}\tilde{G}\tilde{Q}_w\right)\\
&\quad + \text{tr}(F_2'\tilde{Q}F_2 R_v)\text{tr}(F_1'\tilde{Q}F_1S) - \text{tr}(F_2'\tilde{Q}F_2 R_v)\text{tr}(F_1'\tilde{Q}F_1\tilde{A}^jS(\tilde{A}')^j)
\end{aligned}$$

Next we replace each q_i term in (3.23). As we do so, note that the first term in q_1 cancels with the second term in q_2 and q_5 cancels with the last term in q_6. Then we have

$$
\begin{aligned}
\mathbb{E}(\bar{\ell}\bar{\ell}_J) &= 2\text{tr}(F_1'\tilde{Q}F_1S(\tilde{A}')^j F_1'\tilde{Q}F_1\tilde{A}^jS) + 4\text{tr}(F_1'\tilde{Q}F_2 I_w\tilde{Q}_w\tilde{G}'(\tilde{A}')^{j-1}F_1'\tilde{Q}F_1\tilde{A}^jS) \\
&\quad + 2\text{tr}\left(I_w'F_2'\tilde{Q}F_2 I_w\tilde{Q}_w\tilde{G}'(\tilde{A})'^{j-1}F_1'\tilde{Q}F_1\tilde{A}^{j-1}\tilde{G}\tilde{Q}_w\right) + \text{tr}^2(F_1'\tilde{Q}F_1S) \\
&\quad + \text{tr}(F_2'\tilde{Q}F_2 R_v)\text{tr}(F_1'\tilde{Q}F_1S) + \text{tr}(F_1'\tilde{Q}F_1S)\text{tr}(F_2'\tilde{Q}F_2 R_v) + \text{tr}^2(F_2'\tilde{Q}F_2 R_v)
\end{aligned}
$$

The last four terms simplify to $\text{tr}^2(F_1'\tilde{Q}F_1S + F_2'\tilde{Q}F_2 R_v) = \text{tr}^2(\tilde{Q}P)$ from the definition of P. Therefore,

$$
\begin{aligned}
\mathbb{E}(\bar{\ell}\bar{\ell}_j) &= 2\text{tr}(F_1'\tilde{Q}S_{fj}\tilde{Q}F_1\tilde{A}^jS) + 4\text{tr}(F_1'\tilde{Q}S_{wj}\tilde{Q}F_1\tilde{A}^jS) + 2\text{tr}(I_w'F_2'\tilde{Q}S_{wj}\tilde{Q}F_1\tilde{A}^{j-1}\tilde{G}\tilde{Q}_w) \\
&\quad + \text{tr}^2(\tilde{Q}P) \\
&= 2\text{tr}(\tilde{Q}(S_{fj} + S_{wj})\tilde{Q}(S_{fj} + S_{wj})') + \text{tr}^2(\tilde{Q}P)
\end{aligned}
$$

Since $\mathbb{E}(\bar{\ell}^2) = \mathbb{E}\left(\bar{f}'\tilde{Q}\bar{f}\bar{f}'\tilde{Q}\bar{f}\right) = 2\text{tr}(\tilde{Q}P\tilde{Q}P) + \text{tr}^2(\tilde{Q}P)$, (3.22) simplifies to

$$
\begin{aligned}
\mathbb{E}(\mathcal{K}^2) &= \frac{2}{T}\text{tr}(\tilde{Q}P\tilde{Q}P) + \frac{4}{T}\left(\tilde{m}'\tilde{Q}P\tilde{Q}\tilde{m}\right) \\
&\quad + \frac{4}{T^2}\left(\sum_{j=1}^{T-1}(T-j)\text{tr}(\tilde{Q}(S_{fj} + S_{wj})\tilde{Q}(S_{fj} + S_{wj})')\right) \\
&\quad + \frac{8}{T^2}\sum_{j=1}^{T-1}(T-j)\left(\tilde{m}'\tilde{Q}(S_{fj} + S_{wj})\tilde{Q}\tilde{m}\right) \\
&\quad + \text{tr}^2(\tilde{Q}P) + 2\text{tr}(\tilde{Q}P)m'\tilde{Q}m + (\tilde{m}'\tilde{Q}\tilde{m})^2 \quad\quad (3.24)
\end{aligned}
$$

and from (3.17), we have the variance of the plant KPI:

$$
\begin{aligned}
\mathrm{var}(\mathcal{K}) = {} & \frac{2}{T}\mathrm{tr}(\tilde{Q}P\tilde{Q}P) + \frac{4}{T}\left(\tilde{m}'\tilde{Q}P\tilde{Q}\tilde{m}\right) \\
& + \frac{4}{T^2}\left(\sum_{j=1}^{T-1}(T-j)\mathrm{tr}(\tilde{Q}(S_{fj}+S_{wj})\tilde{Q}(S_{fj}+S_{wj})')\right) \\
& + \frac{8}{T^2}\sum_{j=1}^{T-1}(T-j)\left(\tilde{m}'\tilde{Q}(S_{fj}+S_{wj})\tilde{Q}\tilde{m}\right)
\end{aligned}
\tag{3.25}
$$

4

DISTURBANCE MODEL IDENTIFICATION BACKGROUND

4.1 REVIEW OF DISTURBANCE MODEL IDENTIFICATION

Methods for identifying noise covariances fall into the categories of Bayesian estimation, maximum likelihood estimation, covariance matching, correlation techniques (including the ALS method), and subspace ID methods. Before reviewing the other methods, we discuss subspace ID techniques, which have recently gained popularity. These methods are primarily designed for process model identification but also identify the noise statistics (Ljung, 1999; Qin, 2006). Subspace ID methods use least-squares regression to identify a characteristic subspace of the input-output data; the system matrices and noise statistics are then extracted from this subspace (Van Overschee and De Moor, 1995). Originally designed for open-loop data, these ID methods have been modified for use on closed-loop data (Qin, 2006). Because subspace ID methods identify the system matrices as well as the noise statistics, the input must be persistently exciting in order to accurately identify the input matrix B (Ljung, 1999). Rather than finding the driving process and measurement noises, these methods identify the innovation covariance and the optimal estimator gain (Qin, 2006). While this noise model can be used in perfor-

mance monitoring (as discussed in Section 3.5) and provides the optimal Kalman filter, knowledge of the process and measurement noise covariances provides a more inherent understanding of the disturbances affecting the system. Therefore, finding these noise covariances allows more flexibility in estimator design and provides more information towards performance monitoring (Rajamani and Rawlings, 2009). In addition, subspace ID methods have not been used to identify the disturbance model for a system containing integrators. Such a method would require using a grey-box model to identify the system matrices.

Like subspace ID methods, the general Bayesian estimation problem presented by Mehra (1972) may also include both unknown deterministic and unknown stochastic parameters. However, this method is easily reduced to the case in which the only unknowns are the elements of Q_w and R_v. Implementing this method in practice is challenging both because of the *a priori* knowledge required, as the user must choose prior probabilities for Q_w and R_v, and because of the computational requirements, as the method requires integration over a large dimensional space. Matisko and Havlena (2013) proposed a Bayesian method in which Q_w and R_v are the only unknown parameters. In this method, a grid of possible (Q_w, R_v) pairs is created, and the Kalman filter is designed for each (Q_w, R_v) pair. Then state estimation is performed using each Kalman filter, and the likelihood and posterior probability are computed. The covariance estimates are chosen as either the maximum *a posteriori* estimate or the mean-square estimate. The authors proposed using a Monte Carlo method to generate a (Q_w, R_v) grid with more density near the highest probability.

Unlike Bayesian estimation, maximum likelihood methods do not rely on knowledge of a prior distribution. Like subspace ID methods, several early maximum likelihood methods also focus on finding filter parameters (Mehra, 1969, 1972; Kashyap, 1970). The process and measurement noise covariances are then extracted from these results (under certain conditions). Mehra (1969) wrote the

likelihood for a SISO system in terms of the optimal innovation, which is maximized with respect to unknown deterministic parameters, the optimal filter gain, and the innovation variance. He proposed using correlation techniques to find the initial guess for the MLE problem. Kashyap (1970) proposed a three part maximum likelihood estimation scheme for a multivariable time series system. First the deterministic parameters are estimated, then the optimal filter parameters are estimated using these results, and finally the noise covariances are estimated from the optimal filter parameters (under certain identifiability assumptions). Likewise, in the maximum likelihood method of Mehra (1972) first finds the Kalman filter and innovation variance are estimated for a state space model, and then Q_w and R_v are found when uniqueness conditions are met.

More recently, Bohlin and Graebe (1995) and Kristensen, Madsen, and Jørgensen (2004) used 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, which can include the noise covariances, are unknown. The system discussed is a set of stochastic ordinary differential equations linearized using the extended Kalman filter.

Since direct maximum likelihood methods require solving a nonlinear optimization problem, Shumway and Stoffer (1982) proposed an iterative method using the expectation maximization (EM) technique. In the EM method, an initial guess of the unknown parameters is chosen, and the states are estimated from the current parameter estimate via the Kalman smoother. Then the unknown parameters are updated 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.

Bavdekar, Deshpande, and Patwardhan (2011) developed both a direct max-

imum 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 can be separated into 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 Shumway and Stoffer (1982) estimated the state transition matrix, A, as well as the noise covariances. Both the maximum likelihood and expectation maximization methods of Bavdekar et al. (2011) accurately identified Q_w and R_v in simulation and led to improved estimation for laboratory data. They applied both methods to systems with measurements sampled at multiple rates. Li and Badgwell (2014) applied the EM method of Bavdekar et al. (2011) 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 of the estimates compared to the ALS method.

The maximum likelihood estimation problem is discussed in detail in Chapter 8. By assuming that the deterministic system matrices are known, we reduce the maximum likelihood problem to estimation of parameters affecting the covariance of a normally distributed signal. The estimation of the covariance of a normal distribution when the entire covariance matrix is unknown is discussed in detail in Anderson and Olkin (1985). Whereas Anderson and Olkin (1985) assumed complete freedom in the covariance matrix, Magnus (1978) studied the case in which the covariance matrix is a function of some number of unknown parameters. He derived first and second order conditions for the maximum likelihood estimator of the mean and covariance. An iterative method was proposed for the case when these equations cannot be solved analytically.

Covariance matching techniques are less computationally expensive than Bayesian

or maximum likelihood estimation. In these methods, the disturbance model is identified by matching the covariances of the optimal innovations with their theoretical values (Myers and Tapley, 1976). The iterative method proposed in Myers and Tapley (1976) is not guaranteed to converge or to produce semidefinite results. Odelson (2003) presents a more detailed analysis of these techniques, which shows that they produce biased results, as the covariances do not provide enough information to estimate the noise matrices. In more recent modifications to this method, the measurement noise covariance is assumed to be known, and only the process noise covariance matrix is estimated (Valappil and Georgakis, 2000; Tzou and Lin, 2001).

Rather than considering only the covariances of the innovations, in the correlation techniques of Mehra (1970, 1972), the noise model is chosen to fit the autocorrelations (or autocovariances) of the innovations at different lags. By considering different lags, these techniques extracts more information from the measurement than does covariance matching. These methods need not be iterative, although using the optimal estimator gain may reduce the variance of the results (Mehra, 1972). While the original method involved multiple steps, the ALS technique reduces the correlation-based method to a single least-squares problem (Odelson, Rajamani, and Rawlings, 2006; Rajamani and Rawlings, 2009). The covariances are forced to be positive semidefinite by using a log-barrier penalty method (Odelson et al., 2006). The resulting optimization problem is convex, although the objective is no longer quadratic (Odelson et al., 2006). When there is not enough information to estimate the full process noise covariance matrix, the problem is modified to find the solution with the fewest number of independent noises (Rajamani and Rawlings, 2006). The ALS approach is also used to estimate the optimal noises for integrating disturbances used to provide offset-free control (Rajamani, Rawlings, and Qin, 2009). The ALS method is described in more detail in the following section.

Letting $\hat{x}(k)$ denote the Kalman predictor of $x(k)$, we define the state estimate error $\hat{x}$ and L-innovation $\mathscr{Y}$ as

$$\varepsilon(k) := x(k) - \hat{x}(k) \qquad\qquad \mathscr{Y}(k) := y(k) - C\hat{x}(k)$$

Then we write the evolution of these errors as

$$\varepsilon(k+1) = \bar{A}\varepsilon(k) + Gw(k) - ALv(k) \qquad\qquad \mathscr{Y} = C\varepsilon(k) + v(k)$$

in which $\bar{A} := A - ALC$. Letting P be the covariance of ε, the autocovariances of $\mathscr{Y}$ are

$$\mathbb{E}\left(\mathscr{Y}(k)\mathscr{Y}(k)'\right) = CPC' + R$$

$$\mathbb{E}\left(\mathscr{Y}(k+j)\mathscr{Y}(k)\right) = C\bar{A}^{j}PC' - C\bar{A}^{j-1}ALR_v$$

We combine the autocovariances from lag 0 to lag $N-1$ into a single equation:

$$\mathbb{E}\begin{bmatrix} \mathscr{Y}(k)\mathscr{Y}(k)' \\ \vdots \\ \mathscr{Y}(k+N-1)\mathscr{Y}(k)' \end{bmatrix} = \mathscr{O}PC' + \Gamma R_v \qquad (4.1)$$

in which

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

P is related to Q_w and R_v through the Lyapunov equation

$$P = \bar{A}P\bar{A}' + GQ_wG' + ALR_vL'A' \tag{4.2}$$

Next we rewrite (4.1) in terms of $\text{vec}(Q_w)$ and $\text{vec}(R_v)$, in which the vec operator stacks the column of a matrix. In doing so, we use the identity

$$\text{vec}\left(AXB\right) = \left(B' \otimes A\right)\text{vec}(X)$$

in which $\otimes$ denotes the Kronecker product. To eliminate P from (4.1), we vectorize (4.2), solve for $\text{vec}(P)$, and substitute this value of $\text{vec}(P)$ into the vectorized form of (4.1). Thus we obtain the equation

$$b := \text{vec}\left(\mathbb{E}\begin{bmatrix} \mathscr{Y}(k)\mathscr{Y}(k)' \\ \vdots \\ \mathscr{Y}(k+N-1)\mathscr{Y}(k)' \end{bmatrix}\right) = \mathscr{A}\begin{bmatrix} (Q_w)_{ss} \\ (R_v)_{ss} \end{bmatrix}$$

in which

$$\mathscr{A} = \begin{bmatrix} \mathscr{A}_1 & \mathscr{A}_2 \end{bmatrix}$$

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

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

and $(Q_w)_{ss}$ denotes the vector containing the lower-triangular elements of Q_w. For a symmetric matrix, $(Q_w)_{ss}$ completely describes Q_w. The duplication $\mathscr{D}_n$ is defined such that it satisfies $\text{vec}(Q_w) = \mathscr{D}_n(Q_w)_{ss}$ (Magnus and Neudecker, 1999, p.49).

The matrix $\mathscr{A}$ is completely known from the state space model and estimator. We use the model and estimator to calculate the state estimate $\hat{x}$ and then find

$\mathscr{Y} = y - C\hat{x}$. We calculate estimate b using the sample autocovariances of $\mathscr{Y}$; we refer to this estimate as $\hat{b}$. Then we solve the least-squares problem

$$\min_{Q_w, R_v} \left\| \mathscr{A} \begin{bmatrix} (Q_w)_{ss} \\ (R_v)_{ss} \end{bmatrix} - \hat{b} \right\|_W^2$$

The choice of the weight W has significant impact on the variance of the estimates; choosing this weight is discussed in Section 6.2.

The estimator that we use to calculate $\mathscr{Y}$ does not need to be same estimator that the controller is currently using, provided that the same estimator is used to calculate both $\mathscr{Y}$ and $\mathscr{A}$. When the estimator used by the controller is unknown, or when the autocovariances of the innovations decay too slowly, we recommend choosing a different initial estimator for the ALS problem.

We add semidefinite constraints to this problem in order to guarantee that Q_w and R_v are feasible. As discussed in more detail in Section 4.3, the problem has a unique solution if and only if $\mathscr{A}$ is full rank. In the case that $\mathscr{A}$ loses rank, Rajamani and Rawlings (2009) suggest seeking the minimum number of independent process noises by penalizing $\text{tr}(Q_w)$. This method uses trace as a substitute for rank in order to maintain a convex problem. The complete ALS problem is then written as

$$\min_{Q_w, R_v} \left\| \mathscr{A} \begin{bmatrix} (Q_w)_{ss} \\ (R_v)_{ss} \end{bmatrix} - \hat{b} \right\|_W^2 + \text{tr}(Q_w) \quad \text{subject to } Q_w \geq 0, \ R_v \geq 0$$

4.3 UNIQUENESS CONDITIONS FOR THE ALS METHOD

The necessary conditions for the ALS estimate to be unique are discussed in both Odelson et al. (2006) and Rajamani and Rawlings (2009). Here we offer a more complete discussion especially as concerns unobservable systems. Before

we begin this discussion, we note that for any given system, the full matrix ALS problem has a unique solution if and only if there is a unique solution to the single matrix ALS column, as shown in Appendix 4.A. Because of this property, the uniqueness conditions discussed below hold true for both formulations of the ALS problem. We also note that for $\mathscr{A}$ to be full rank, then both $\mathscr{A}_1$ and $\mathscr{A}_2$ must be full column rank (although the converse does not necessarily hold). Therefore, if a condition is necessary for $\mathscr{A}_1$ to be full column rank, then that condition is necessary for $\mathscr{A}$ to be full column rank.

Before discussing the conditions for uniqueness given in the literature, we note that a unique solution to the ALS problem exists only if the following two necessary conditions are satisfied:

1. G is full column rank. (Note that this condition implies $g \leq n$.)

2. $\frac{g(g+1)}{2} \leq np$

To prove Condition 1, we first observe that if G is not full column rank, then $G \otimes G$ has a non-zero null space. Further, from Hua (1990, Theorem 2), if G is not full column rank, there exists a symmetric matrix $X \neq 0$ such that $(G \otimes G)\text{vec}(X) = 0$. Therefore, $(X)_{ss}$ lies in the null space of $\mathscr{A}_1$, and the ALS problem does not have a unique solution.

Condition 2 holds because $\mathscr{A}_1$ is the product of three matrices and therefore its rank is less than or equal to the rank of each of those matrices:

$$\text{rank}(\mathscr{A}_1) \leq \min\left(\text{rank}(C \otimes \mathscr{O}), \text{rank}(I_{n^2} - \bar{A} \otimes \bar{A}), \text{rank}((G \otimes G)\mathscr{D}_g)\right)$$

For $n \geq p$, this condition simplifies to

$$\text{rank}(\mathscr{A}_1) \leq \min\left(np, n^2, g(g+1)/2\right) = \min\left(np, g(g+1)/2\right)$$

Since the number of columns in $\mathscr{A}_1$ equals the number of unknowns in Q_w, $g(g+$

$1)/2$, if this quantity is larger than np, the matrix $\mathscr{A}_1$ is not full column rank and the ALS estimates are not unique. Although these conditions are necessary, they are not sufficient, as shown in the following discussion.

In the correlation-based method in Mehra (1970), the following conditions are assumed:

A. (A, C) is observable.

B. (A, G) is controllable.

C. A is full rank.

D. The number of unknowns in Q_w is less than or equal to np.

Odelson et al. (2006) further examined these conditions and demonstrated that they are not sufficient.

The condition that (A, C) be observable is necessary in the case when the noise shaping matrix G is unknown, as is disccused in more detail in Section 6.1.1. However, if G is known, then (A, C) need not always be observable because the process noise covariance may be estimated from the observable states alone (Odelson et al., 2006).

The condition that (A, G) be controllable is not necessary for the ALS problem to have a unique solution. As a counter example, consider

$$A = \begin{bmatrix} 0.1 & 0 \\ 0 & 0.2 \end{bmatrix} \qquad G = \begin{bmatrix} 2 \\ 1 \end{bmatrix} \qquad C = \begin{bmatrix} 1 & 2 \end{bmatrix}$$

Although (A, G) is uncontrollable, the ALS matrix is still full rank. In general, having (A, G) uncontrollable may mean that there are fewer independent unknowns to consider, and so we can more easily solve the problem.

The condition that A be full rank allows the conditions for uniqueness to be greatly simplified as in Rajamani and Rawlings (2009). As mentioned in Rajamani

and Rawlings (2009), if A is singular, then the singular modes can be removed, and the system can be redefined with a non-singular A. If G is unknown and A is singular, the ALS estimate is not unique, as discussed in Theorem 1 in Section 6.1.1.

Note that Mehra's Condition D is the same as Condition 1 above and is therefore necessary. When G is unknown, we apply more stringent condition on the number of unknowns given in Lemma 13 of Rajamani and Rawlings (2009), which implies that the solution is not unique whenever $n > p$. This more strict condition is also given in (Mehra, 1972). From this condition, it follows that whenever $p \leq n$ in the non-augmented system and integrators are added to the states, there is not a unique solution to the ALS problem (when G is unknown). Odelson (2003) claims the condition $n > p$ is unnecessary; however, his counter-example was based on the assumption that Q_w is known to be diagonal.

As mentioned in Odelson et al. (2006), Conditions A-D are not sufficient. Whether or not the ALS problem has a unique solution depends also on the structure of the matrices A, C, and G. For example, consider the system

$$A = \begin{bmatrix} 0.09 & & & \\ & 0.27 & & \\ & & 0.08 & \\ & & & 0.05 \end{bmatrix} \qquad C = \begin{bmatrix} 3 & 4 & 0 & 0 \\ 0 & 0 & 1 & 7 \end{bmatrix}$$

with two different noise-shaping matrices,

$$G_1 = \begin{bmatrix} 0.3 & 0 & 0 \\ 0 & 0.9 & 0 \\ 0 & 0 & 0.5 \\ 0 & 0 & 1 \end{bmatrix} \qquad G_2 = \begin{bmatrix} 0.3 & 0 & 0 \\ 0 & 0.9 & 0 \\ -1 & 0 & 0.5 \\ 0 & 0 & 1 \end{bmatrix}$$

In both cases, (A, C) is observable, (A, G) is controllable, and A is full rank. The condition on the number of unknowns is satisfied for both systems since $np = 8$ and $g(g + 1)/2 = 6$. However, we have a unique ALS solution when we use G_2, but not when we use G_1. Therefore, the uniquess of the solution depends both on the structure of the system and on the number of unknowns relative to the number of outputs.

4.4 METHODS TO ASSESS THE ALS RESULTS

When using the ALS method in practice, it is necessary to evaluate the accuracy of the solution before implementing any changes to the controller or using the co-variances in MPC monitoring. In some simple simulations, the estimated values of Q_w and R_v can be compared to their true values. However, whenever determin-istic disturbances are present (even in simulation), we deliberately introduce plant model mismatch by approximating these disturbances as integrated white noise. Therefore, we no longer have a "true" value of Q_w against which to compare the ALS results. When using industrial data sets, we also do not know the true noise covariances. Therefore, we propose two methods to assess the accuracy of the ALS results:

1. Assess the goodness-of-fit of the least-squares problem.

2. Assess the performance of the redesigned estimator.

4.4.1 *Assess the goodness-of-fit of the least-squares problem*

In the ALS algorithm, we first calculate the sample autocovariances of the L-innovations, which form the vector $\hat{b}$. Let

$$b_{\text{ALS}} := \mathscr{A} \begin{bmatrix} (Q_w)_{ss} \\ (R_v)_{ss} \end{bmatrix}$$

In the ALS method, we choose Q_w and R_v such that b_{ALS} is as close to $\hat{b}$ as possible in a least-squares sense. One way to qualitatively assess this result is to plot the elements of $\hat{b}$ and the elements of b_{ALS} on the same axis and compare the two curves. To make the plots easier to view, we separate $\hat{b}$ and b_{ALS} into the autoco-variances of each output and cross-covariances between each pair of outputs. We create a $p \times p$ grid of plots. The diagonal plots show the autocovariances of each output. The plot in the $(i,j)^{\text{th}}$ position shows the cross-covariance between $\mathscr{Y}(i)$ and $Yin(j)$, $\text{cov}(\mathscr{Y}_i(k), \mathscr{Y}_j(k-l))$. The x-axis of each plot is the lag l.

An example of the autocovariance plots is shown in Figure 4.1. The top four subplots show the sample autocovariances and fits produced by the ALS method using a perfect model. In each of the four plots, the curves are in good agreement with each other, indicating that the ALS solution has fit the data well. The bottom set of plots shows the results of the ALS problem with a large amount of model mismatch; in this case, the fit is poor for the cross-covariances. The theoretical autocovariances at lag zero also do not fit the sample autocovariances (seen in the diagonal plots). Note that when studying these plots, we only consider is whether the theoretical and sample autocovariances match each other well; we do not consider whether or not the estimator is behaving optimally. Since we use the suboptimal initial estimator to form these plots, we would not expect to see optimal estimator behavior. We discuss the optimal estimator behavior in the following section.

4.4.2   *Assess the performance of the redesigned estimator*

We also assess the quality of the ALS results by using these results to design a new estimator and then assessing the quality of the estimator performance. In the absence of plant-model mismatch, when the noises are white and the noise covariances are known accurately, the Kalman filter produces white innovations. The innovations are white because the filter extracts all the useful information
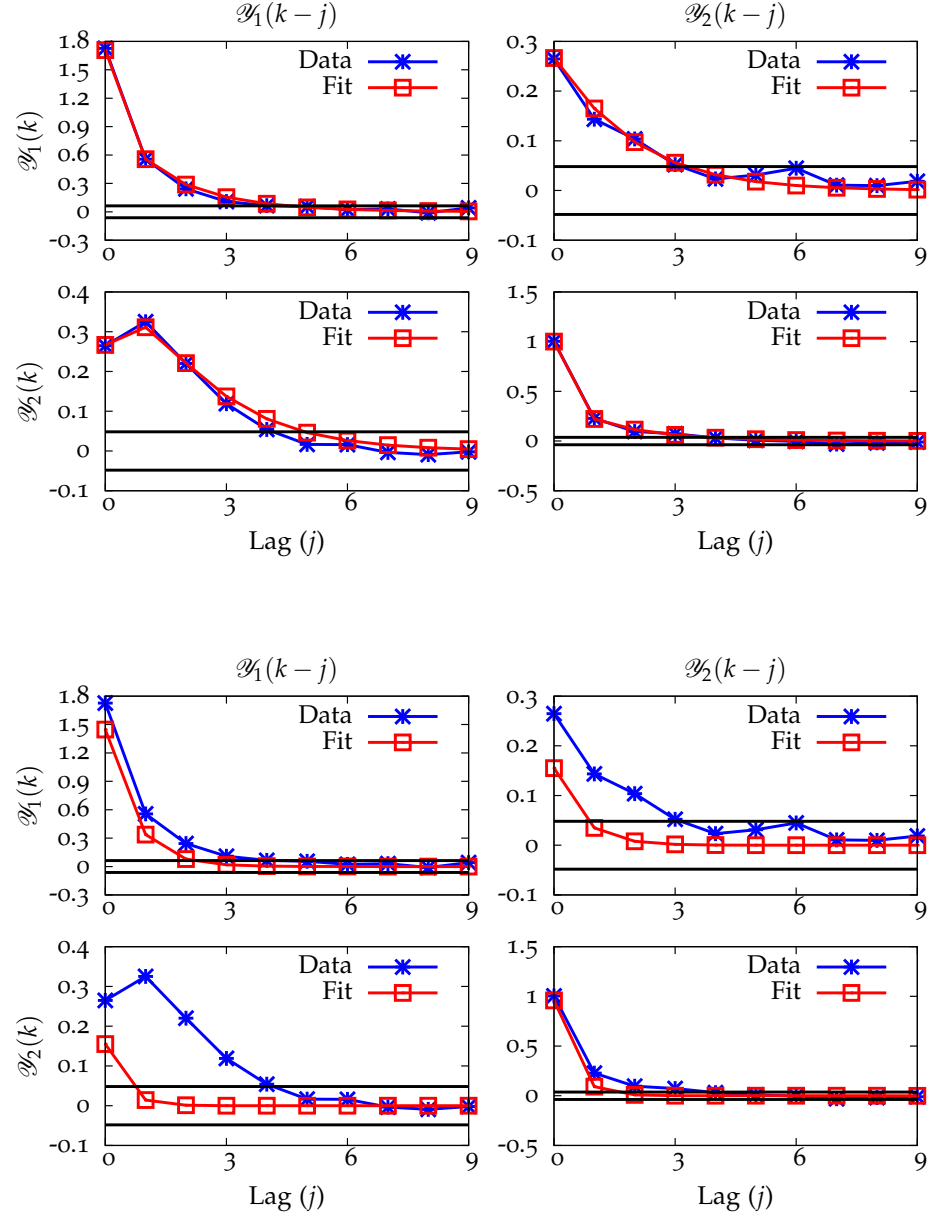
Figure 4.1: Example of plots to assess the goodness-of-fit for the ALS method. The top four plots show that the theoretical autocovariances using the ALS estimates match the sample autocovariances from data. In the bottom four plots, plant model mismatch has caused significant error in the ALS estimates. The theoretical autocovariances no longer match the sample autocovariances.

from the measurements but does not over-predict based on the random noises corrupting the measurements. Therefore, we assess the accuracy of the ALS estimates by determining if they produce a Kalman filter with white innovations. To do so, we tune the Kalman filter based on the estimates Q_w and R_v, process the data using this estimator, and study the autocovariances of the resulting innovations. In procesing the data, we can consider either the same data set that was used to estimate Q_w and R_v or a new data set from the same process.

When processing the data, we compute the new innovations, $\mathscr{Y}_{\text{ALS}}(k) = y(k) - C\hat{x}_{\text{ALS}}(k)$, where the state estimates come from the new Kalman filter. We then calculate the autocovariances and cross-covariances at different lags. For the ideal estimator, any auto- or cross-covariance may be nonzero at lag zero, but should be zero for all higher lags.

When estimator gain is optimal, this optimality does not guarantee that Q_w and R_v are correct. For example, the estimator gain remains constant when both noise covariances are scaled by the same factor. However, assessing the goodness-of-fit as described in the previous section ensures that the measurement noise is on the correct order of magnitude. We further discuss optimal estiamtor behavior in the following section.

4.4.3 *Guide to autocovariance plots*

To understand how to interperet the autocovariance plots, we study Figure , which shows ideal estimator behavior for the system

$$A = \begin{bmatrix} 0.5 & 0 & 0 & 0 \\ 0 & 0.5 & 0.1 & 0 \\ 0 & 0.1 & 0.5 & 0 \\ 0 & 0 & 0 & 0.15 \end{bmatrix} \quad C = I_4 \quad G = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & -0.15 \end{bmatrix} \quad Q_w = I_2 \quad R_v = 0.5I_4$$

In the determinstic part of the system, (A, C), y_1 and y_4 are completely independent from the other outputs, whereas y_2 and y_3 are correlated with each other. The measurement noises corrupting each output are uncorrelated with each other. Two process noises affect the system, w_1 and w_2, corresponding to the two columns of G. The noise w_1 affects the first state but not the other states (due to the structure of A). Thus, since $y_1 = x_1$, w_1 affects only the output y_1. Likewise, the noise w_2 affects states y_2, y_3, and y_4 but not y_1.

In Figure 4.2, the plots along the main diagonal in the figure show the autocovariances of each innovation with itself. For each innovation, there is significant autocovariance at lag zero, but the autocovariances are zero at higher lags. The cross-covariances between y_1 and the other outputs (the plots in the first column and the first row) remain within the confidence intervals at *all* lags, including lag zero. This complete lack of correlation indicates that y_1 is independent from all of the other outputs.

In this system, y_2 and y_3 are identical to each other except for the measurement noise. The strong correlation is evident in the cross-covariance plots between y_2 and y_3 (the plots in the (2,3) and (3,2) positions). These cross-covariances are non-zero at lag zero, but again the covariances are zero at higher lags. Although in the deterministic system, y_4 is independent from the other outputs, it is correlated to y_2 and y_3 through the process noise. Since the correlation is negative (as can be seen from the matrix G), the cross-covariance between y_4 and y_2 or y_3 is negative. As in all the plots, the correlation is only present at lag zero, indicating that the estimator is optimal.

Note that the autocovariances at higher lag are not symmetric. To illustrate this point, consider the cross-covariances between y_1 and y_4. The plot in the (1,4) position shows the covariances between $\mathscr{Y}_1(k)$ and $\mathscr{Y}_4(k-j)$ and the plot in the (4,1) position shows the covariances between $\mathscr{Y}_4(k)$ and $\mathscr{Y}_1(k-j)$. These two covariances are identical at lag zero but differ at higher lags.
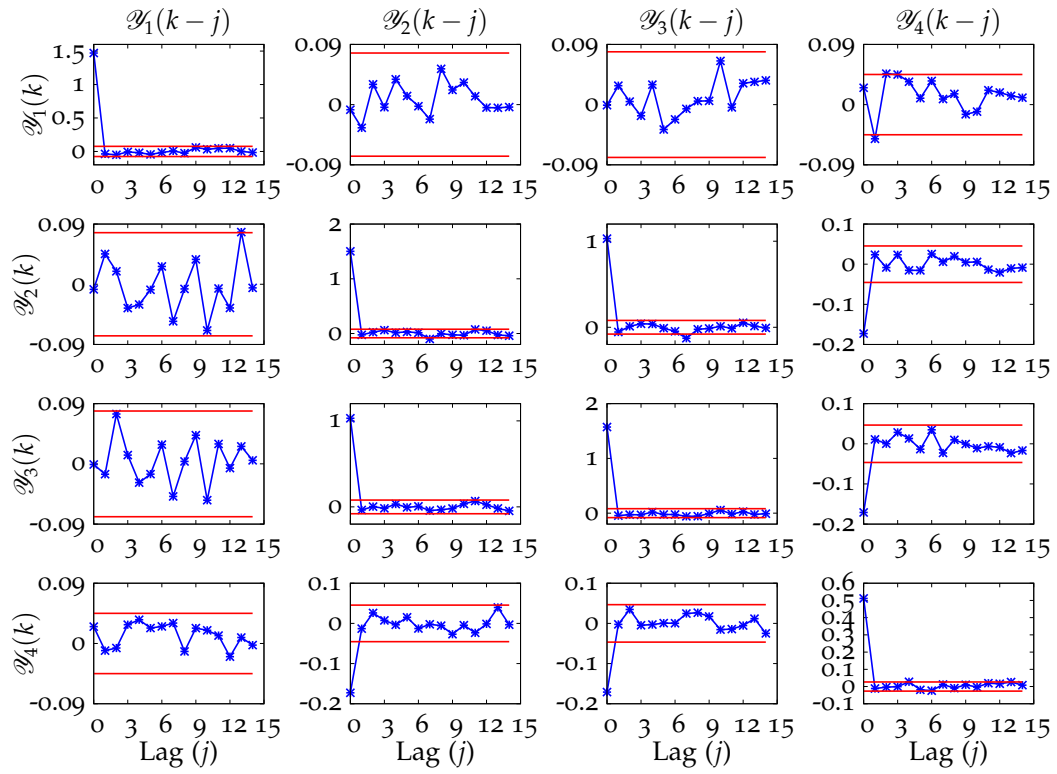
Figure 4.2: Example of autocovariances for an optimal estimator. The auto- and cross-covariances are non-zero only at lag zero.
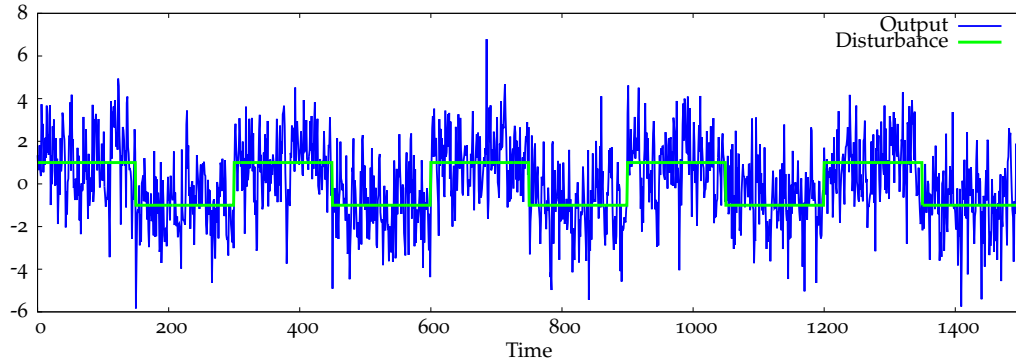
Figure 4.3: Output and disturbance for the simple example.

Next we illustrate suboptimal estimator behavior by studying a small model consisting of one state and one output with an integrating disturbance added to the output. The true system is affected by a repeated step disturbance, which can be estimated well by the integrated disturbance model. The output and disturbance are shown in Figure 4.3. We compare the performance of the optimal estimator with suboptimal estimators. This comparison gives a qualitative idea of the autocovariance patterns that result from different incorrect noise models.

The top plot shows the optimal behavior for the estimator (as determined from the ALS results), in which the only significant autocovariance is at lag zero. The second row shows the effect of having an incorrect measurement noise. When the measurement noise is too large, the effect of the measurement is undermodeled, *i.e.* the estimator makes little use of the current measurement. As a result, correlation remains present at higher lags. When R_v is too small, we are essentially overmodeling the effect of the measurement. The measurement noise is not adequately filtered and has too strong of an effect on the state estimate, resulting in the negative autocovariance at lag one. When Q_w is too large, we also see this characterstic negative autocovariance at lag one, since the filter treats the measurements as being more accurate than they are in reality. In addition, the autocovariances do not decay to zero, because the integrator is hardly used. When

Q_w is too small, the autocovariances are slightly oscillatory, as the information from the measurements is not being fully used. The autocovariances when Q_d is too large look similar to those when R_v is too small, but they also rise to zero more slowly after lag one. When Q_d is too small, the autocovariances are characterized by a very slow decay to zero at higher lags.

From inspection of the autocovariance plots we obtain some idea as to what errors are present. However, it would be difficult to develop an accurate noise model by inspection alone. Even with the simple example above, many errors in the noise model have similiar effects (for example when R_v is too small and when Q_d is too large). For larger systems with many states and outputs, the numerous variance and covariances would be impossible to identify by inspecting the autocovariances alone.

4.5 APPENDICES

4.A *Proof of equivalence between single column and full matrix ALS techniques*

We define the autocovariance matrix as

$$
\mathscr{R}(N) = \begin{bmatrix} \mathscr{Y}_k \mathscr{Y}_k' & \cdots & \mathscr{Y}_k \mathscr{Y}_{k+N-1}' \\ \vdots & \ddots & \vdots \\ \mathscr{Y}_{k+N-1} \mathscr{Y}_k' & \cdots & \mathscr{Y}_k \mathscr{Y}_k' \end{bmatrix} = \begin{bmatrix} \mathscr{R}_1(N) & \mathscr{R}_2(N) & \cdots & \mathscr{R}_N(N) \end{bmatrix}
$$

which vectorizes to

$$
\mathrm{vec}\,(\mathscr{R}(N)) = \begin{bmatrix} \mathrm{vec}(\mathscr{R}_1(N))' & \mathrm{vec}(\mathscr{R}_2(N))' & \cdots & \mathrm{vec}(\mathscr{R}_N(N))' \end{bmatrix}'
$$

$$
= \mathscr{A}_{\mathrm{full}} \begin{bmatrix} \mathrm{vec}(Q_w) \\ \mathrm{vec}(R_v) \end{bmatrix}
$$
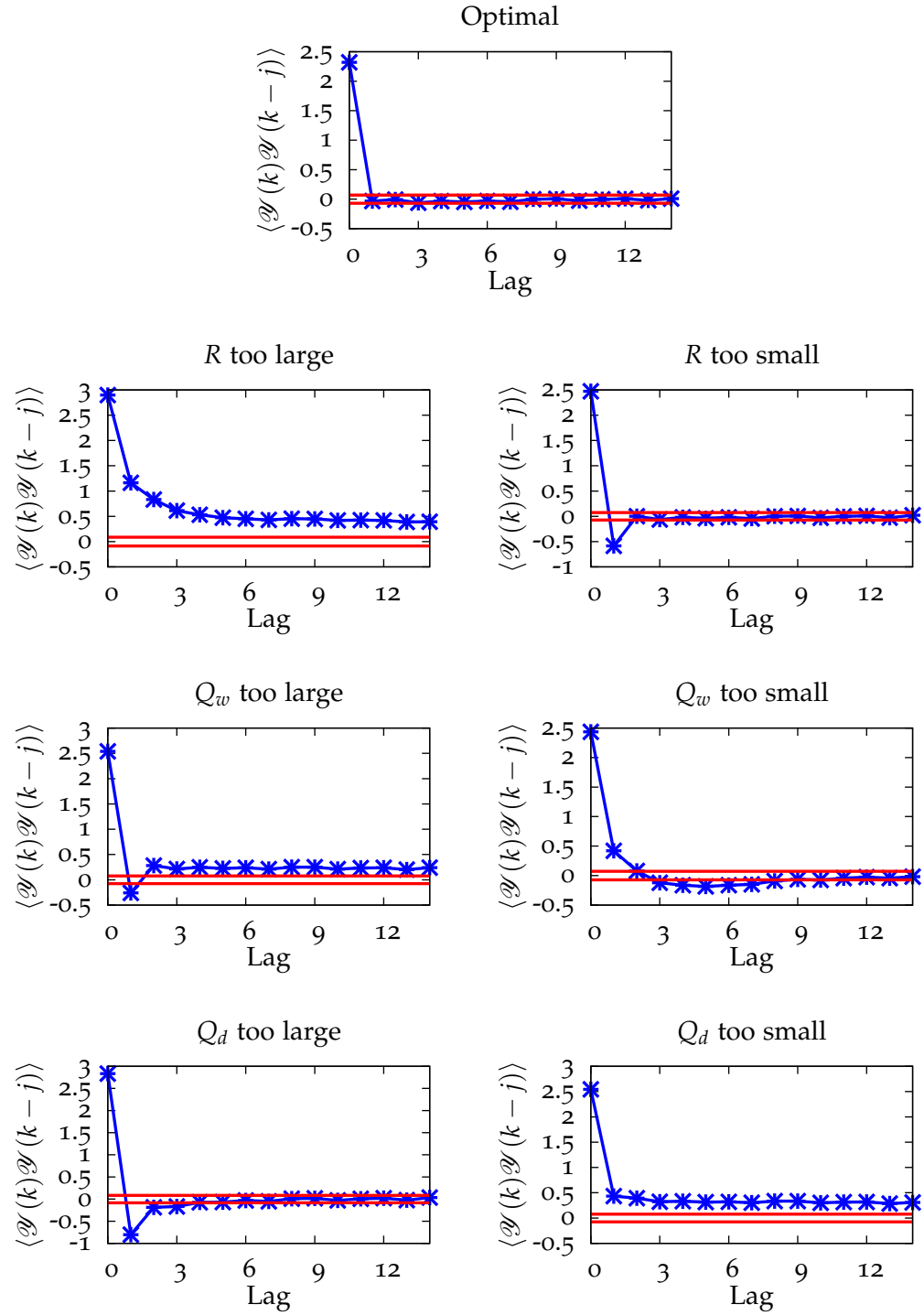
Figure 4.4: Autocovariances under different noise model mismatch scenarios.

in which $\mathscr{A}_{full}$ is the matrix $\mathscr{A}$ for the full matrix ALS method as defined in Odelson et al. (2006). Since the autocovariances in $\mathscr{R}_2 \ldots \mathscr{R}_N$ are duplicates of the autocovariances in $\mathscr{R}_1$, we can write each $\text{vec}(\mathscr{R}_i(N))$ as a linear combination of the elements of $\text{vec}(\mathscr{R}_1(N))$:

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

in which $\mathscr{A}$ is the matrix for the single column ALS method. Therefore, the matrices for the full matrix and single column ALS methods are related as

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

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

5

INTEGRATING DISTURBANCE MODELS

5.1 PURPOSE OF INTEGRATING DISTURBANCE MODELS

Although the Kalman filter (with correct noise covariances) calculates optimal state estimates in the presence of zero-mean white noises, it does not accurately estimate the states in the presence of non-zero mean disturbances. As a result, the regulator is not able to remove the effect of these disturbances. To obtain offset-free control in the presence of unmeasured disturbances, we need to estimate the disturbances as well as the state. We account for these disturbances in the model by augmenting the system with integrators as described in Rawlings and Mayne (2009) and Pannocchia and Rawlings (2003). Without these integrators, the controller would not compensate for any non-zero disturbances or model mismatch. We write the augmented system as

$$X^+ = \begin{bmatrix} A & B_d \\ 0 & I \end{bmatrix} X + W \qquad y = \begin{bmatrix} C & C_d \end{bmatrix} X + v$$

in which $X := \begin{bmatrix} x' & d' \end{bmatrix}'$. The augmented noise matrix $W := \begin{bmatrix} w' & w'_d \end{bmatrix}'$ is distributed as

$$W \sim N\left(0, Q_W\right) \quad Q_W = \begin{bmatrix} Q_w & Q_{wd} \\ Q'_{wd} & Q_d \end{bmatrix}$$

We then design the Kalman filter for the augmented system; this tuning of the estimator depends on the variance of the augmented covariance matrix Q_W.

This disturbance model is *not* meant to model accurately the disturbances entering the system. In fact, since the variance of d is always growing, if the plant faced a true integrated white noise disturbance, the variance of the input or output would become unbounded in response to this disturbance. Thus, by incorporating this disturbance model, we are deliberately adding introducing plant model mismatch because we are adding a disturbance that does not occur in the plant.

Despite the mismatch caused by the integrating disturbance model, we choose to use this model because it is simpler than estimating a more detailed disturbance model. By using this simple form, we can apply the same disturbance model to a variety of systems; we only need to choose an appropriate augment process noise covariance. We also are able to keep the same disturbance model for a given system even though the disturbance does not remain completely constant.

In the absence of the integrating disturbance model, any non-zero mean disturbance or plant model mismatch causes offset in the controlled variables. With the integrating disturbance model, we obtain offset-free control by using the Kalman filter for the augmented system. Since d does not decay in the integrator model, the disturbance estimate $\hat{d}$ remains non-zero to compensate for the disturbance. The noise term w_d allows $\hat{d}$ in the augmented Kalman filter to respond to changes in the disturbances.

As discussed in Rajamani et al. (2009), if we create two augmented systems by augmenting the same original state space model with two disturbance models, (B_{d1}, C_{d1}) and (B_{d2}, C_{d2}), the augmented systems are similarity transforms of

each other (provided that we add p integrators and that each augmented system is detectable). Therefore, the choice of (B_d, C_d) is not critical to obtain good estimator behavior, provided that the augmented process noise covariance is chosen appropriately, such as by using the ALS method.

We find the augmented process noise covariance by solving the ALS problem using the augmented model (Rajamani et al., 2009). Although A_{aug} is unstable, an estimator can always be chosen such that the closed loop $\bar{A}_{\text{aug}}$ is unstable (provided the augmented system is observable).

We also use the augmented model in calculating the theoretical KPI. Since the KPI depends both on the disturbance as well as the disturbance estimate error, the integrated white noise must be assumed to have zero variance. If the variance were non-zero, the plant would be responding to a growing disturbance, and the KPI would be infinite. In order to account for these non-zero mean disturbances, we must have a deterministic estimate of them to use in the KPI calculation, as discussed in Section 3.4.

5.2 STEP DISTURBANCES

The integrator model estimates step disturbances well because step disturbances and integrated white noise have similar autocovariances. We demonstrate this property by examining the system whose true dynamics are $y(k) = d(k)$, in which $d(k)$ is a repeated step disturbance. When we model the system as an integrator, the estimator has the form

$$\hat{d}^+ = \hat{d} + L_d \mathcal{Y}_d \qquad\qquad \mathcal{Y}_d = y - \hat{d}$$

When no measurement noise is present, the optimal estimator gain is $L_d = 1$, which gives the disturbance estimate $\hat{d}(k) = y(k-1)$.

The output, innovation, and innovation autocovariance for this system are

shown in Figure 5.1. This figure compares the behavior of the step disturbance to an actual integrated white noise. Although the disturbances and innovations are clearly different, the innovation autocovariances are nearly identical. The only difference is that whereas the innovation of the integrated white noise is completely white, the step disturbance has a non-zero autocovariance at lags corresponding to the frequency of the step. If these larger lags are not considered, the autocovariances of both systems are identical. Hence, if the maximum lag used in the ALS problem is smaller than the time between the step changes, then the ALS method fits the autocovariances of the step disturbance perfectly by using an integrated white noise model. Therefore, integrated white noise is a good stochastic model to use in approximating the step disturbance.
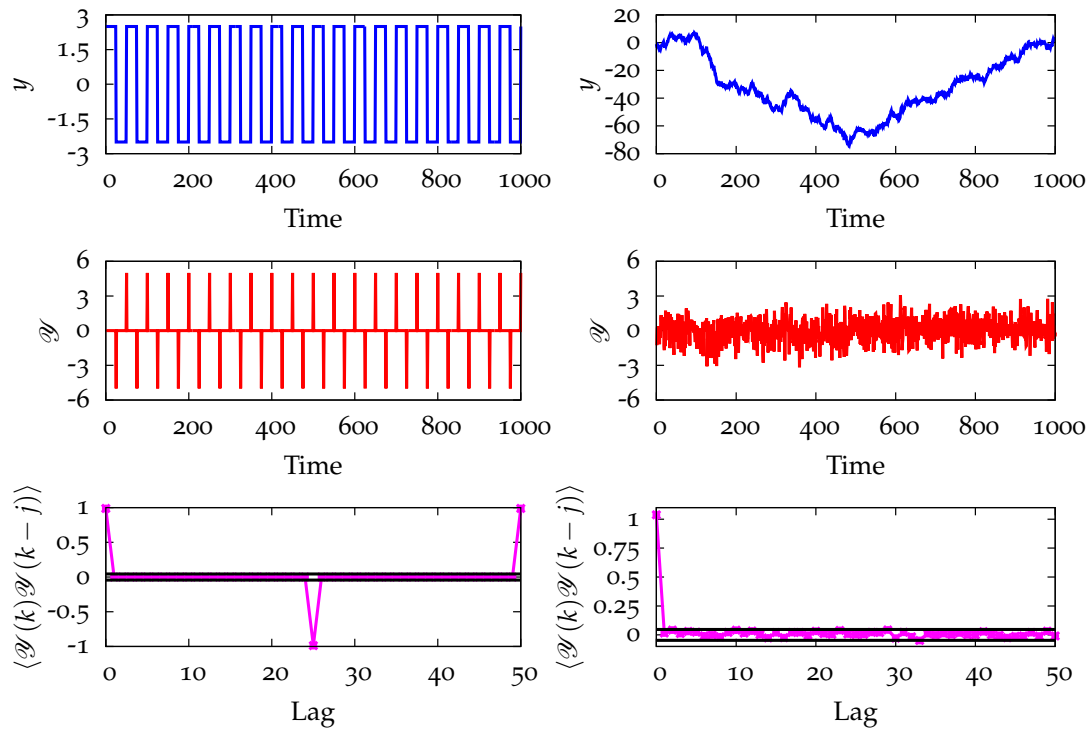
Figure 5.1: Disturbance (top), innovations (middle), and autocovariances (bottom) for a repeated step disturbance (left) and integrated white noise (right).

Further, the covariance of the driving noise, Q_d, (and therefore the integrator gain) is related to the frequency and magnitude of the step disturbance. From the

equation for the integrator, we write the variance as $Q_d = \text{var}\,(d(k) - d(k-1))$. Since the innovation is equal to $d(k) - d(k-1)$ in the absence of measurement noise, Q_d is also the variance of the innovation. For the step disturbance (without noise), the sample variance of the innovations is proportional to the frequency of the step changes and to the square of the magnitude of the step change. Thus, the ALS estimate of Q_d for the step disturbance is also proportional to the frequency of the step and the magnitude squared. We use the relationship between Q_d and the step disturbance to calculate the optimal estimator gain. When no noise is present in the system, $L = 1$ is the best estimator gain for this model type, regardless of the value of Q_d. When measurement noise is present, the optimal gain is $0 < L_d < 1$, and its precise value depends on the measurement noise variance as well as the size and frequency of the step.

To illustrate the optimal Q_d for a step disturbance, Figure 5.2 compares the step disturbances of different magnitudes and frequencies. Identical measurement noises with $R_v = 0.5$ were been added to each step disturbance. The plots in the top row show the noisy step disturbances as well as the predictions from the optimal estimator. The middle plots show the innovations, and the bottom plots show the autocovariances. The sample autocovariances for the step disturbance are plotted along with the theoretical autocovariance for an integrated white noise. The ALS results, as well as the optimal estimator gain for each disturbance sequence, are summarized in Table 5.1. As the magnitude or frequency of the step disturbance increases, the estimator gain also increases. This change allows the estimator to respond more quickly to the step changes. The cost of this quick response is that the innovations are larger during the flat periods where no steps occur.
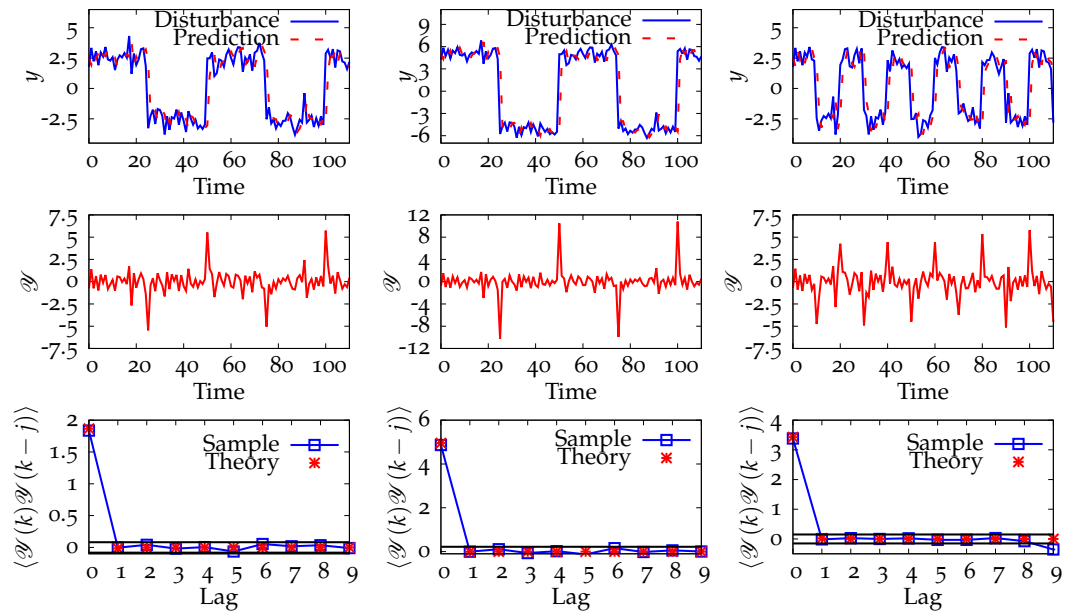
Figure 5.2: Disturbance and estimate (top), innovations (middle), and autocovariances (bottom) for steps of different magnitudes and frequencies. Compared to the plots on the left, the step disturbance in the middle plots has a larger magnitude and the step disturbance in the right plots has a higher frequency. As the magnitude and frequency increase, the sample autocovariance at lag zero also increases.

Table 5.1: ALS estimates and estimator gain for each step disturbance shown in Figure 5.2.

Magnitude	Frequency	$\hat{Q}$	$\hat{R}$	$\hat{L}$
5	1/25	1.02	0.48	0.74
10	1/25	4.01	0.49	0.90
5	1/10	2.45	0.52	0.85

5.3 RAMP DISTURBANCES AND DOUBLE INTEGRATOR MODELS

Although the integrated white noise model is a good approximation for a step disturbance, this model does not adequately account for all disturbances that may affect the plant. To illustrate a case when this model is inadequate, we consider a ramp-type disturbance as shown in Figure 5.3. When we estimate this disturbance with the integrator model, our estimates always differ from the true value by the slope of the ramp, as shown in Figure 5.3. We can tune the estimator by decreasing the gain but this tuning only makes the approximation worse, as it uses less information from the data. Our system is undermodeled — the integrator model does not allow us to capture the slope of the ramp. As a result, although the innovations are zero-mean, they remain above or below zero for significant periods of time. The resulting autocovariances have a clear trend, indicating that the system is undermodeled.

To eliminate this undermodeling, we propose a double integrator disturbance model to approximate the ramp:

$$y_1 = \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} d_1 \\ d_2 \end{bmatrix} + v \qquad \begin{bmatrix} d_1 \\ d_2 \end{bmatrix}^+ = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} d_1 \\ d_2 \end{bmatrix} + \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$$

In this model the slope is approximated as a single integrator, d_2. In the case of a pure ramp, the only "noise" in the system is due to the change in slope, so
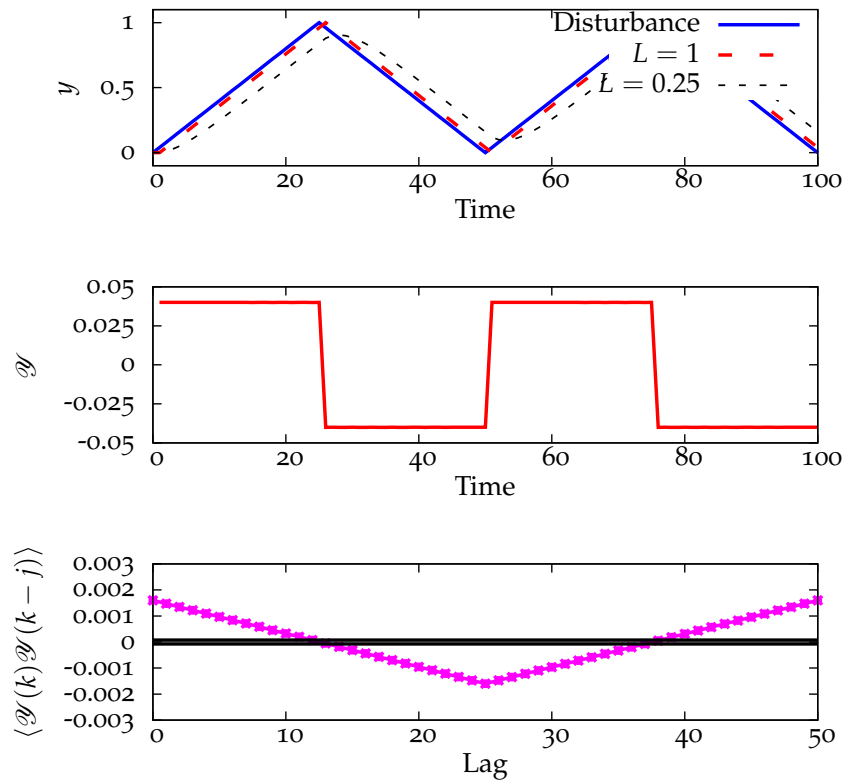
Figure 5.3: Disturbance and prediction (top), innovations (middle), and autocovariances (bottom) for the ramp disturbance modeled as a single integrator. The disturbance is clearly undermodeled.

$$Q_d = \begin{bmatrix} 0 & 0 \\ 0 & Q_{d2} \end{bmatrix}.$$ As there is no measurement noise, the estimator gain for the double integrator model is $L = \begin{bmatrix} 1 & 1 \end{bmatrix}'$. Figure 5.4 shows the optimal prediction for the ramp disturbance using a double integrator model. As shown in this figure, the innovations follow the same pattern as the innovations for step disturbances (with a single integrator model), and the autocovariance looks white at low lags. Since the slope of the ramp is a step disturbance and is being modeled as a single integrator, the double integrator is a good model for the ramp disturbance. For comparison, a true double integrator system is shown on the right of Figure 5.4.

Most disturbances affecting a system cannot be clearly classified as steps or
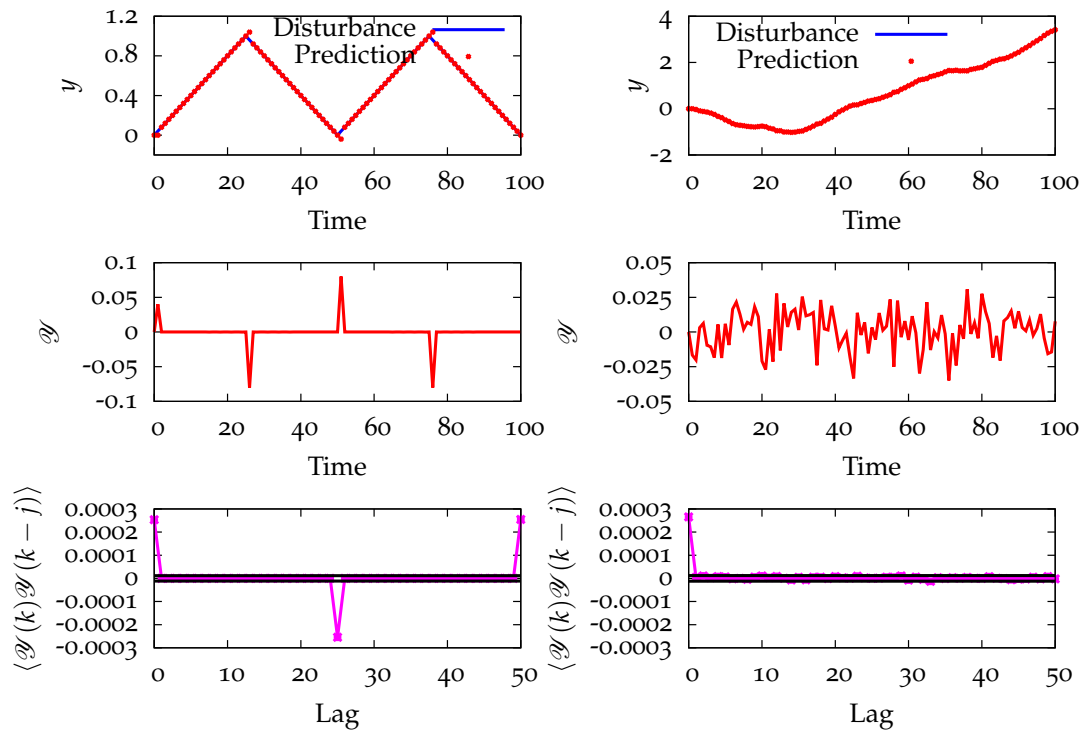
Figure 5.4: Disturbance and prediction (top), innovations (middle), and autocovariances (bottom) for the ramp disturbance modeled as a double integrator (left) and for a double integrator (right). The ramp disturbance is well modeled, and therefore the innovations appear white at low lags.

ramps. Figure 5.5 shows an example of a disturbance estimated from plant data. We apply the ALS method to this disturbance using both single and double integrator disturbance models and process the data using the "optimal" estimators from the ALS results. The double integrator model, shown in the left plots, works well; there is no clear trend in the innovations, and the autocovariances show that the estimator is approximately optimal. In contrast, the innovations from the single integrator model appear to be non-white. The autocovariances confirm that the single integrator model does not adequately model this disturbance. These autocovariances decay very slowly, indicating undermodeled dynamics. The predictions from each model are compared for part of the disturbances in Figure 5.6. As with the ramp disturbance, the single integrator model is not able to account
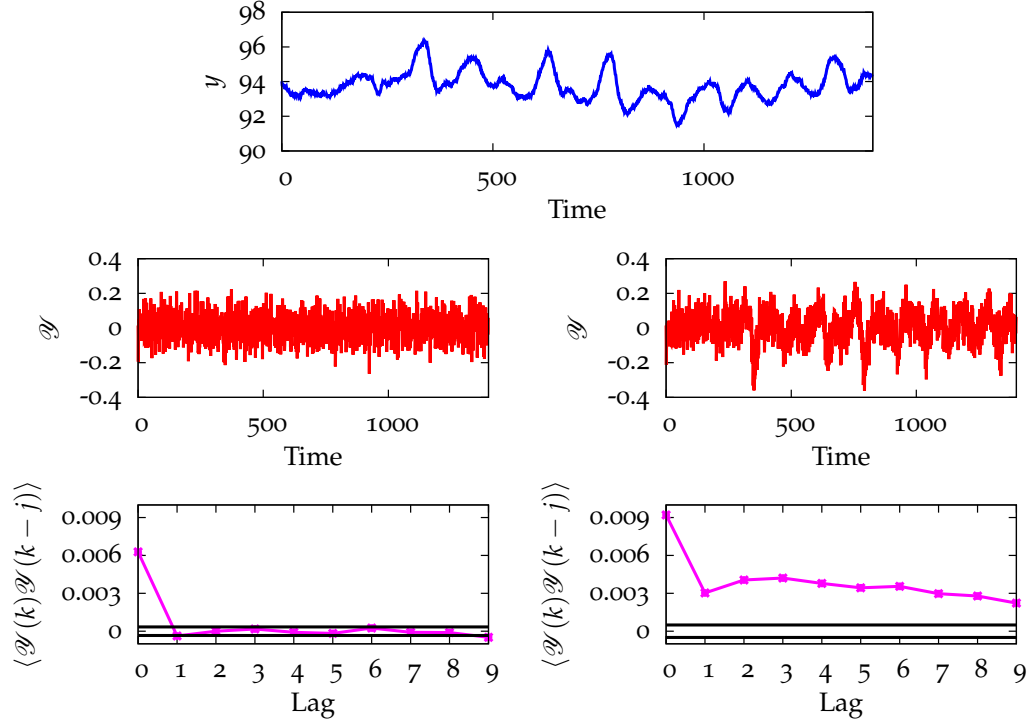
Figure 5.5: Disturbance estimated from plant data (top), innovations (middle), and autocovariances (bottom). The double integrator model (left plots) produces white innovations and is a significant improvement over the single integrator model (right plots).

for the slope of the constantly changing disturbance.

When the noise model includes double integrators, the augmented system takes the form

$$
A_{\text{aug}} = \begin{bmatrix} A & B_d \\ 0 & A_d \end{bmatrix} \qquad B_{\text{aug}} = \begin{bmatrix} B \\ 0 \end{bmatrix} \qquad C_{\text{aug}} = \begin{bmatrix} C & C_d \end{bmatrix}
$$

In this system, A_d is a block diagonal matrix with dimension $n_d \times n_d$, $p \leq n_d \leq 2p$. The blocks of A_d are either double or single integrators . We choose C_d and B_d such that the system is observable. When double integrators are added, we choose the C_d and B_d to have a zero column corresponding to d_2 of the double integrator. For example, suppose $p = 2$ and we choose to add a single integrator to the first
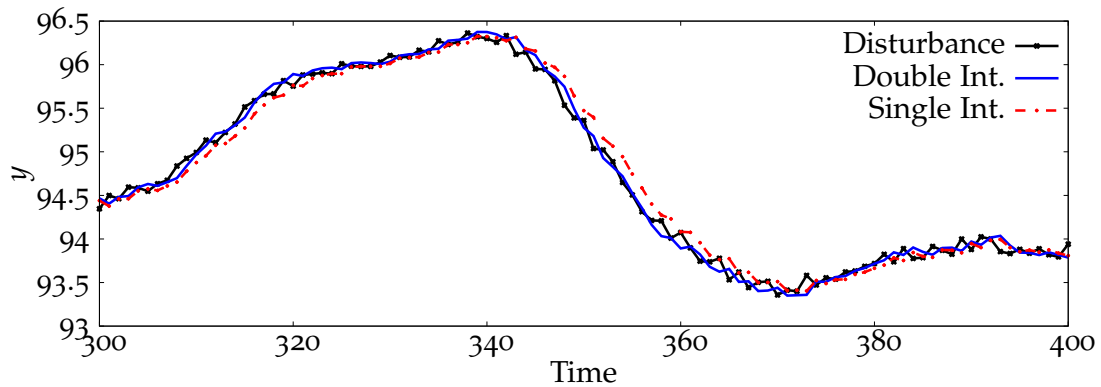
Figure 5.6: Comparison of the single and double integrator model estimates for the disturbance from plant data. The double integrator model shows noticeable improvement over the single integrator model, which cannot account for the slope of the disturbance.

output and a double integrator to the second output. Then $B_d = 0$ and

$$A_d = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix} \qquad\qquad C_d = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

5.4 CAUTIONS ON DOUBLE INTEGRATOR MODELS

Although double integrators are better able to estimate ramps and other distur-bances, we should avoid using double integrators when the are unnecessary, for several reasons:

1. Lack of steady state target: With the single integrator model, the target selec-tor portion of the controller chooses a steady state target for u and x based on the setpoints for the system and the disturbance estimate. The regulator then chooses the control action based on these steady state values. However, the double integrator only has a steady state solution for x and u when the slope is zero. Therefore, the controller must use a different method to re-ject the disturbances, such as choosing a steady slope for the disturbance, or

using the setpoints and the disturbance estimates directly in the regulator.

2. Lack of similarity transform: Using a single integrator disturbance model, the tuning of the estimator can compensate for misassigned disturbances (Rajamani et al., 2009). In other words, with appropriate tuning, an estimator that treats the disturbance as entering through the output behaves optimally even when the disturbance enters through the input, and vice-versa. This property is due to the fact that the input and output disturbance models are related through a similarity transform. For double integrator disturbance models, such a similarity transform no longer exists. Therefore, an inappropriate choice of B_d and C_d may lead to suboptimal estimator behavior, although this property has not been studied extensively.

3. Additional noise covariance elements: The double integrator has three noise elements to be obtained via the ALS problem, whereas the single integrator has only a single driving noise. When the correlations between all the states and disturbances are included, the double integrator model adds significantly more noise elements. As we already do not have enough information in general to obtain a unique ALS solution for the single integrator model, we do not want to add more unknowns to the ALS problem without good reason.

Users also should be cautious to avoid complicating the noise model when in reality the process model needs to be adjusted. In model assessment and identification, it is difficult but necessary to distinguish disturbance dynamics from the plant dynamics. Without a proper disturbance model, disturbances affecting the system may be accounted for as model error in the u-y model, when the error actually lies in the disturbance model. However, the opposite problem may occur as well — a complicated disturbance model may be created to account for a poor u-y model. The ALS technique assumes that the process model is known accurately,

but distinguishing between plant and disturbance dynamics is a difficult task and remains an active area of research.

6

IMPROVEMENTS TO THE ALS METHOD[1]

6.1 APPLYING THE ALS METHOD TO UNOBSERVABLE AND WEAKLY OBSERVABLE SYSTEMS

6.1.1 *Unobservable systems*

In industrial settings, the use of large models with many unobservable or poorly observable states limits the applicability of the ALS method. Since the noises affecting unobservable states have no effect on the outputs, intuitively we expect that when G is unknown, an unobservable system does not have a unique ALS solution. Here we prove that this intuition is correct. We define the ALS problem as

$$\min_{Q_w, R_v} \phi = \|\mathscr{A} \begin{bmatrix} (Q_w)_{ss} \\ (R_v)_{ss} \end{bmatrix} - \hat{b}\|^2 \quad \text{subject to} \quad Q_w \geq 0, R_v \geq 0 \qquad (6.1)$$

---

[1]Portions of this chapter will be published in Zagrobelny and Rawlings (2014a)

in which

$$\mathscr{A} = \begin{bmatrix} \mathscr{A}_1 & \mathscr{A}_2 \end{bmatrix}$$

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

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

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

The ALS problem was derived in Section 4.2; here the identity matrix is used to weight the least-squares term.

The matrix $\mathscr{O}$ in the ALS problem is the extend observability matrix. Provided that $N \geq n$, $\mathscr{O}$ has rank n (full column rank) when the system is observable, and loses rank when the system is unobservable.

Theorem 1. *Assume that $\bar{A}$ is stable. When G is a square matrix of rank n, the ALS problem has a unique solution ($\mathscr{A}$ is full column rank) if and only if (A, C) is observable, A is non-singular, and $\mathrm{rank}(C) = n$.*

Proof. Before proving this theorem, we show that we can restrict our attention to the case in which $G = I$. When $G \neq I$, we take a similarity transformation of the original system, so that it is written in terms of the transformed state $\tilde{x} = G^{-1}x$. The transformed system is described by the matrices $\tilde{A} = G^{-1}AG$, $\tilde{B} = G^{-1}B$, $\tilde{C} = CG$, $\tilde{L} = G^{-1}L$, and $\tilde{G} = I$. The same noises w and v affect both the original and transformed systems. Since the matrix $\mathscr{A}$ is identical in both cases, the original system has a unique ALS solution if and only if the transformed system has a unique solution. Therefore, we limit our discussion to the case when $G = I$. The

proof that these conditions lead to a unique ALS estimate is given in Corollary 3.1 of Rajamani (2007) and is repeated here. From Lemma 7 in Rajamani and Rawlings (2009), when (A, C) is observable, A is non-singular, and $G = I_n$, the null space of $\mathscr{A}$ is equal to the null space of $M = (C \otimes I_n) A^\dagger \mathscr{D}_n$, in which $A^\dagger = (I_{n^2} - \bar{A} \otimes \bar{A})^{-1}$. Since $A^\dagger$ is always full rank (for $\bar{A}$ stable) and $(C \otimes I_n)$ is also full column rank when C is full column rank (rank$(C) = n$), the rank of M is the rank of $\mathscr{D}_n$, or $n(n+1)/2$. Since $\mathscr{D}_n$ is full column rank, then M and therefore $\mathscr{A}$ are also full column rank.

Next we prove that (A, C) observable, A non-singular, and rank$(C) = n$ are necessary conditions. We utilize the fact that $\mathscr{A}$ loses rank if $\mathscr{A}_1$ loses rank. First, we prove the necessity of A being non-singular. (Hua, 1990, Corollary 3) implies that $(C \otimes A) (I_{n^2} - \bar{A} \otimes \bar{A})^{-1}$ is full rank only if $r = \text{rank}\left(\begin{bmatrix} A' & C \end{bmatrix}\right) = n$ and rank$(A) = r$. These conditions can only be satisfied when A is full rank. A more detailed proof is presented in Appendix 6.A.

Next we examine the rank condition on C. Let (A, C) be observable and let rank$(C) = \bar{p}$. From Lemma 13 of Rajamani and Rawlings (2009)[1],

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

Thus for $n > \bar{p}$, the dimension of the null space of $\mathscr{A}$ is greater than zero, and therefore the ALS problem does not have a unique solution. This condition is proved in more detail in Appendix 6.A.

Finally 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

---

[1] 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 Hua (1990, Corollary 2), which was used to derive the cited lemma, we can substitute $\bar{p}$ for p in the rank condition.

system into observability canonical form:

$$(C \otimes \mathscr{O}) = \begin{bmatrix} C_1 \otimes \mathscr{O} & 0_{Np^2 \times nn_u} \end{bmatrix} \qquad \bar{A} \otimes \bar{A} = \begin{bmatrix} \bar{A}_{11} \otimes \bar{A} & 0_{n_o \otimes n_u} \\ \bar{A}_{21} \otimes \bar{A} & \bar{A}_{22} \otimes \bar{A} \end{bmatrix}$$

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

$$\mathscr{A}_1 = \left[(C \otimes \mathscr{O}) \left(I_{n^2} - \bar{A} \otimes \bar{A} \right)^{-1} \right] \mathscr{D}_n = \begin{bmatrix} \mathscr{A}_{11} & 0_{Np^2 \times nn_u} \end{bmatrix} \mathscr{D}_n$$

which loses column rank. Note that although multiplication by $\mathscr{D}_n$ reduces the number of columns, the matrix remains rank deficient. This fact can be clearly seen since the last column of $\mathscr{D}_n$ is $\begin{bmatrix} 0 & \dots 0 & 1 \end{bmatrix}'$ which forces the last column of $\mathscr{A}_1$ to be 0. Thus, a unique solution does not exist. $\qquad \square$

Note: The assumption that $\bar{A}$ is stable in Theorem 1 is always necessary to ensure that $(I_n^2 - \bar{A} \otimes \bar{A})$ is invertible. However, for any detectable (A, C), we can always choose L such that $\bar{A}$ is stable.

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:

$$\min_{Q_w, R_v} \phi + \rho \text{tr}(Q_w) \qquad \qquad \text{subject to } Q_w \geq 0, R_v \geq 0 \qquad (6.2)$$

in which ϕ is the least-squares objective funtion defined in (6.1). Here we show that for any $\rho > 0$, the optimization problem (6.2) 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

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

Let ρ be any strictly positive scalar. Then the optimization problem in (6.2) 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

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

Proof. First we note that there exists an orthogonal T to transform the system into observability canonical form. We can construct an *invertible* transformation matrix $T = \begin{bmatrix} T_1 & T_2 \end{bmatrix}'$ by choosing the n_o columns of T_1 so that they form a basis for the range of $\mathcal{O}'$ and by choosing the n_u columns of T_2 so that they form a basis for null(T_1') (or equivalently, a basis for null$(\mathcal{O})$) (Aplevich, 2000, Ch. 9). Since 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

$$x^+ = Ax + Bu + w \qquad (6.3a) \qquad\qquad \tilde{x}^+ = \tilde{A}\tilde{x} + \tilde{B}u + \tilde{w} \qquad (6.4a)$$

$$y = Cx + v \qquad (6.3b) \qquad\qquad \tilde{y} = \tilde{C}\tilde{x} + \tilde{v} \qquad (6.4b)$$

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 process the noise of the transformed system is

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

As discussed in Rajamani et al. (2009), if $\tilde{L} = TL$, then the state estimates for (6.3) and (6.4) are also related through the similarity transform.

Since the unobservable states in (6.4) do not affect y, we equivalently write the system as

$$x_1^+ = A_{11}x_1 + B_1u + \tilde{w}_1 \tag{6.5a}$$

$$y = C_1x_1 + v \tag{6.5b}$$

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

Define the matrices

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

in which (Q_w^*, R_v^*) minimizes (6.1). Since the system is unobservable, Q_{12} and Q_{22} have no effect on y and therefore $\mathscr{A}_1(Q_w^*)_{ss} = \mathscr{A}_1(Q_w^o)_{ss}$. Thus there exist an infinite number of Q_w^o such that (Q_w^o, R_v^*) minimizes (6.1).

Consider instead the solution to (6.2) for $\rho > 0$. Since Q_w and $\tilde{Q}_w$ are similar matrices, they have the same trace (Lancaster and Tismenetsky, 1985). Since we require $Q_{22} \geq 0$, any solution Q_w that minimizes (6.2) is a transformation of $\tilde{Q}_w = \begin{bmatrix} Q_{11} & 0 \\ 0 & 0 \end{bmatrix}$, as choosing some $Q_{22} > 0$ would increase $\text{tr}(Q_w)$ without decreasing ϕ. The optimization problem (6.2) therefore reduces to

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

Alternatively, we apply the ALS method to the reduced system (6.5). As the L-innovations (and therefore their autocovariances) are identical for the full and

the reduced systems, we have

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

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 $\text{tr}(Q_w)$ 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 we can use to 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.

6.1.2 *Weakly observable systems*

As discussed above, the unobservable states have *no* effect on the output. However, many industrial models include some states that 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. Whereas $\mathscr{O}$ loses rank for the unobservable system (has at least one zero singular value), in the weakly observable case, $\mathscr{O}$ 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 $\mathscr{O} = USV'$ be the singular decomposition of the observability matrix, and choose $T = V'$. Then the observability matrix of the transformed system is $\tilde{\mathscr{O}} = \mathscr{O}T' =$

US. Since 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

$$\tilde{A} = TAT' = \begin{bmatrix} A_{11} & \delta A_{12} \\ A_{21} & A_{22} \end{bmatrix} \qquad \tilde{B} = TB = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} \qquad \tilde{C} = CT' = \begin{bmatrix} C_1 & \delta C_2 \end{bmatrix} \qquad (6.6)$$

in which the magnitude of the scalar $\delta \geq 0$ depends upon 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. 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 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), there is a possibility of the integrators being lost when the weakly observable states are removed from the transformed model. To avoid this problem, we recommend reducing the non-augmented system and then adding the integrators to the reduced model *i.e.* find (A_{11}, B_1, C_1) from the unaug-

mented (A, B, C), and then form

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

In addition, we recommend penalizing 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 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 semi-definite constraints.[2]

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

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

---

[2]We recommend solving the simpler problem here rather than the complete ALS problem as described in (6.2) to reduce the computation time.

6. Write the noise model for the full transformed model by assuming that no noise affects the unobservable states.[3]

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 the example in Section 6.3.1.

6.2 FEASIBLE GENERALIZED ALS TECHNIQUE

The standard ALS method, which we refer to as the "ordinary" ALS method uses the identity matrix to weight the least-squares problem. However, as noted in Rajamani and Rawlings (2009), this weighting is chosen for practical reasons, and it does not produce minimum variance estimates for Q_w and R_v. The minimum variance estimates are obtained from the generalized least-squares problem, where the variance of $\hat{b}$ is used as the weighting (Magnus and Neudecker, 1999, Section 13.5; Schmidt, 1976, Section 2.5). However, computing this variance has two major barriers (Rajamani and Rawlings, 2009). 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 $\left[\mathscr{Y}(1)' \quad \dots \quad \mathscr{Y}(N_d)' \right]'$ must be computed. Secondly, calculating this variance requires knowledge of Q_w and R_v, the unknowns to be found. Rajamani and Rawlings (2009) propose iteratively solving the ALS problem for Q_w and R_v, updating the weighting based on these values, and resolving the ALS problem. However, this iterative scheme is not guaranteed to converge. Therefore, rather than ad-

---

[3]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.

dressing the tractability of computing the optimal weighting, we propose using a feasible generalized least-squares method to approximate it. Feasible generalized least-squares refers to the method in which an approximation of the variance is used to weight the least-squares problem (Schmidt, 1976, Section 2.5). We apply feasible generalized least-squares to the ALS approach as follows.

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

We estimate S by the steps:

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

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

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

$$\vdots$$

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

Since we assume that $\mathcal{Y}_k$ and $\mathcal{Y}_{k+N+i}$ 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 $P_y := \text{var}\left(\begin{bmatrix} \mathscr{Y}'_k & \cdots & \mathscr{Y}'_{k+N-1} \end{bmatrix}'\right)$ 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 $P_0 = \text{var}\left(\mathscr{Y}_k\right)$ and $P_{y,0} = \text{cov}\left(\begin{bmatrix} \mathscr{Y}_k \\ \vdots \\ \mathscr{Y}_{k+N-1} \end{bmatrix}, \mathscr{Y}_k\right)$. 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. Based on the Wishart distribution, we calculate $\hat{S}$ as

$$\hat{S} = \text{cov}(\hat{b}) = \frac{1}{N_s}\left(\left(\hat{P}_0 \otimes \hat{P}_y\right) + K_{p,\tilde{p}}\left(\hat{P}_{y,0} \otimes \hat{P}'_{y,0}\right)\right) \tag{6.7}$$

and find $\hat{W}$ as the inverse of $\hat{S}$. (6.7) is derived in Appendix 6.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 to the approximation method outlined above.

6.3 EXAMPLES[1]

6.3.1 *Example: Weakly observable systems and model reduction*

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

$$
A = \begin{bmatrix}
4.2 \times 10^{-17} & 0.15 & 0 & 0 & 0 & 0 & 0 \\
-0.1 & 0.84 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & -4.2 \times 10^{-17} & 0.15 & 0 & 0 & 0 \\
0 & 0 & -0.1 & 0.84 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0.8 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & -1.1 \times 10^{-16} & 0.64 \\
0 & 0 & 0 & 0 & 0 & -1 & 1.6
\end{bmatrix}
$$

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

The observability matrix has condition number

$$
\sigma = \begin{bmatrix} 6.6 & 2.4 & 1.3 & 0.14 & 0.000\,58 & 3.5 \times 10^{-7} & 1.4 \times 10^{-17} \end{bmatrix}
$$

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 observ-

---

[1] 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.
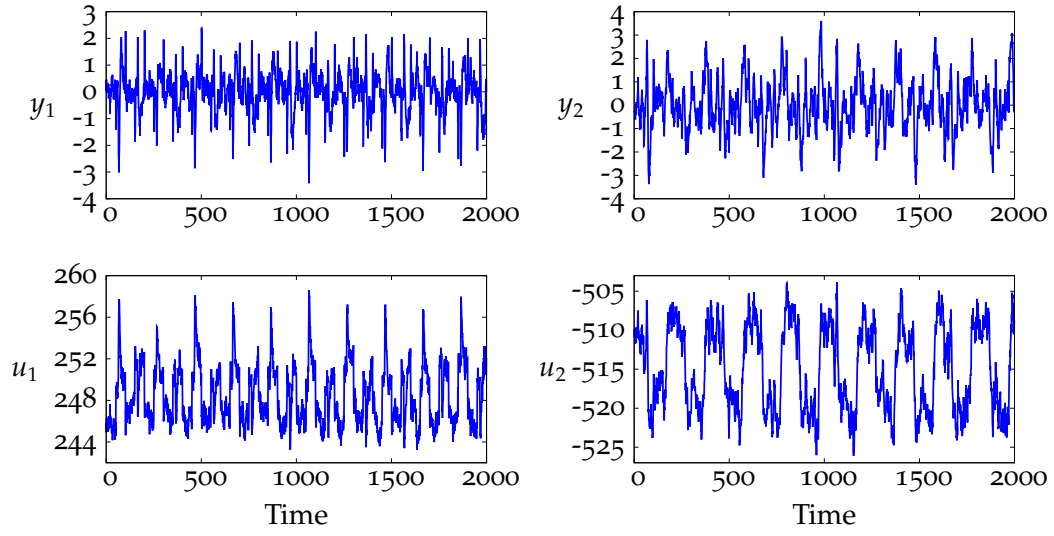
Figure 6.1: Simulated inputs and outputs used in the examples. The plots of the outputs show the deviation from set point.

ability canonical form:

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

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

and augmented the model with integrators on the inputs. We generated the data shown in Figure 6.1 by simulating the 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,
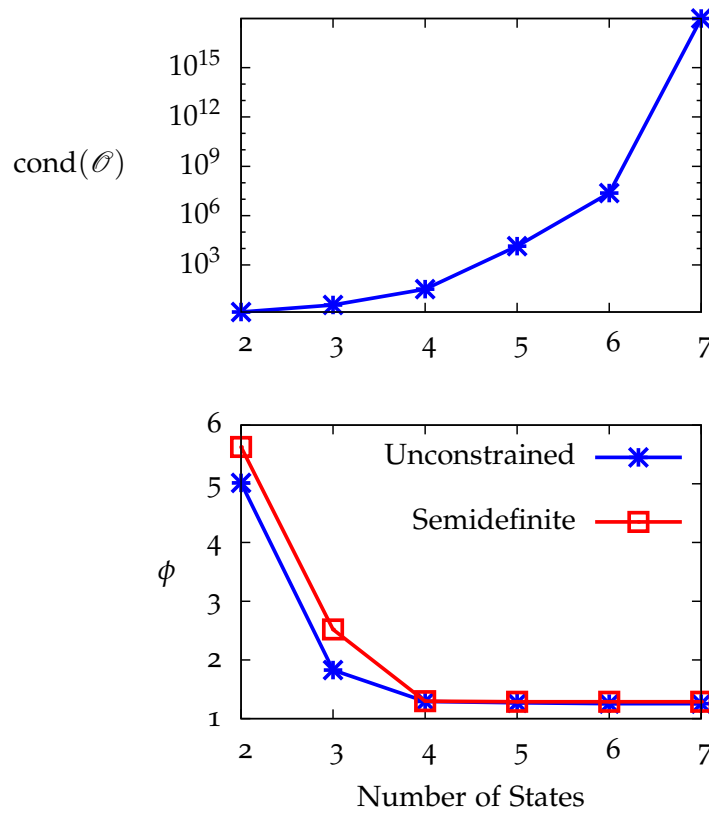
Figure 6.2: Condition number of the observability matrix and ALS objective function value vs. the number of states. As more states are included in the model, the condition number of the observability matrix increases and the ALS objective function value decreases.

as shown in Figure 6.2.

The condition number 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 6.2 illustrates that adding the semidefinite constraints does not affect the choice of reduced model size, as including these constraints does not affect

the trend in the ALS objective function curve. However, we did use the feasible-generalized ALS method in order to avoid placing too much emphasis on matching zero covariances. The full ALS problem, including penalizing $\text{tr}(Q)$ and the semidefinite constraints, was solved after the model size was chosen.

We further illustrate the effect of including more states on the quality the ALS estimates by plotting the estimated autocovariances (calculated from the data) alongside the theoretical autocovariances (calculated from the model and the ALS estimates of Q_w and R_v). This comparison is discussed in Section 4.4.1. Figures 6.3-6.5 show these fits for differently sized models. Since these plots use the original estimator, we did not consider whether or not the estimator performs optimally. Instead, we looked for a model that is sufficiently accurate, so that the estimated and theoretical covariances match each other well. As shown in Figures 6.3 and 6.4, increasing the number of states from two to four leads to a better fit of the data. However, there is no clear difference between the plots with four and seven states (Figures 6.4 and 6.5), indicating that the last three states are poorly observable and need not be included in the model. This conclusion is consistent with the singular value decomposition and also with the plot of the ALS objective function versus the number of states in Figure 6.2.

We next 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 studied whether these innovations are white (meaning that the auto- and cross-covariances of the innovations are significantly greater than zero only at lag zero), as would be the case for an optimal estimator.

As shown in Figure 6.6, the autocovariances of the two-state model remain significantly above zero for lags greater than zero. This behavior indicates that the
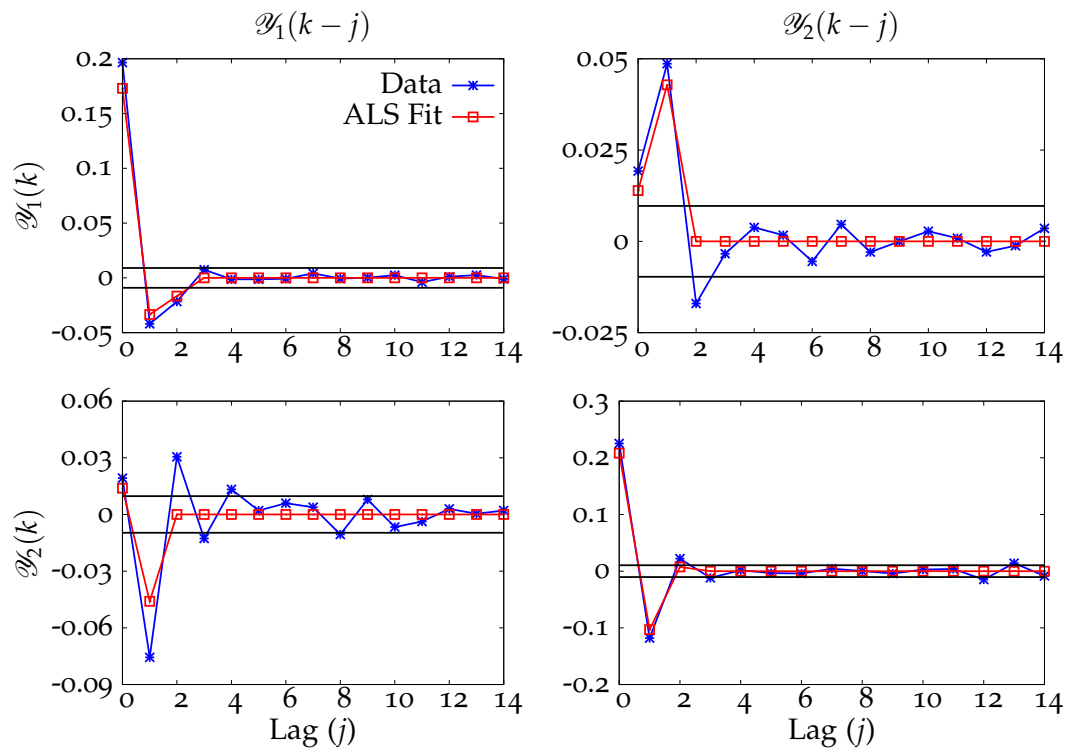
Figure 6.3: Sample and theoretical autocovariances for the two-state model, without semidefinite constraints. The theoretical autocovariances do not fit the data well, indicating that the model is inadequate.
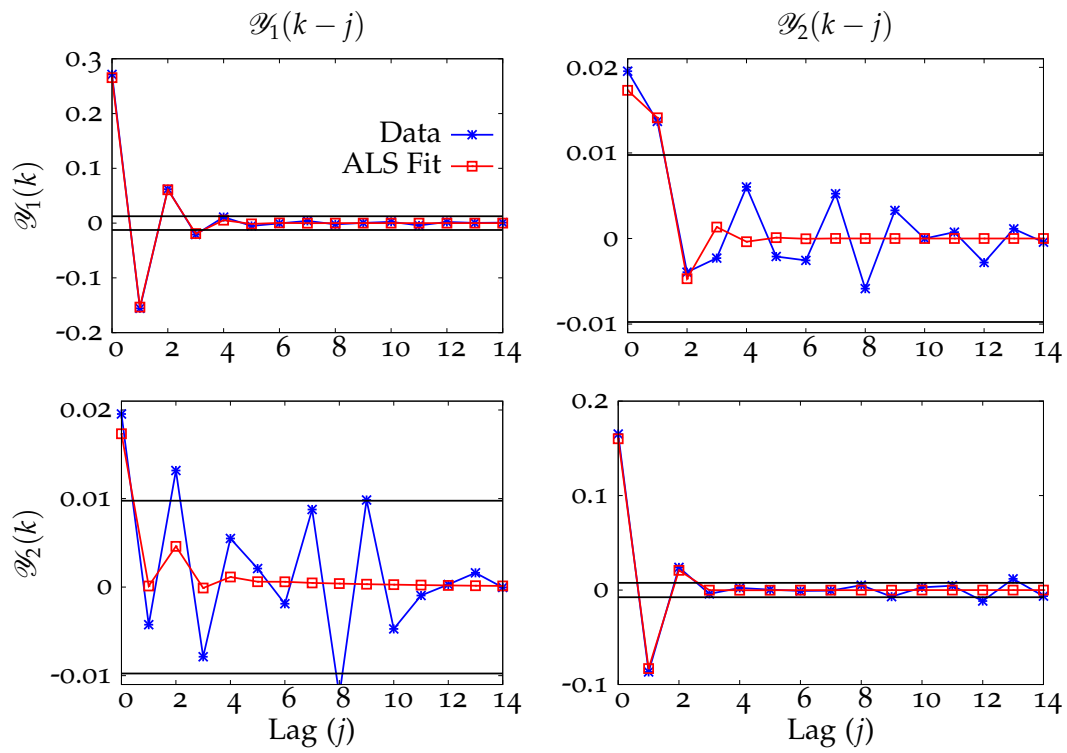
Figure 6.4: Sample and theoretical autocovariances for the four-state model, without semidefinite constraints. The theoretical autocovariances agree well with the data.
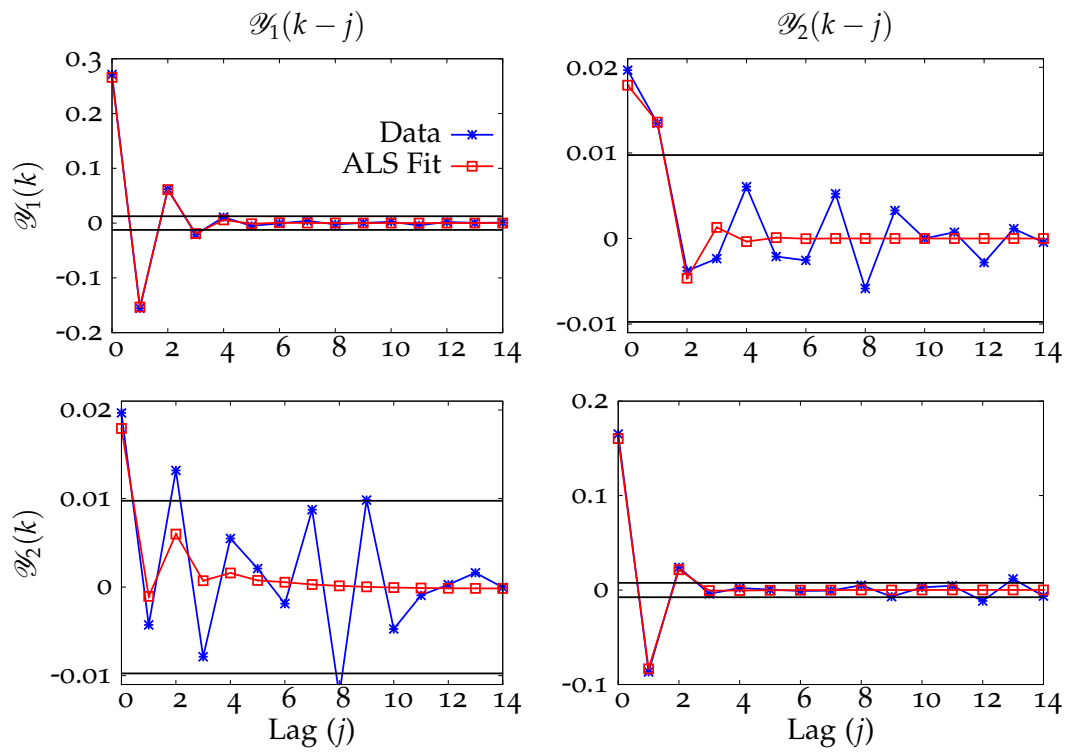
Figure 6.5: Sample and theoretical autocovariances for the seven-state model, without semidefinite constraints. There is no noticeable change in the autocovariances compared to those of the four-state model.
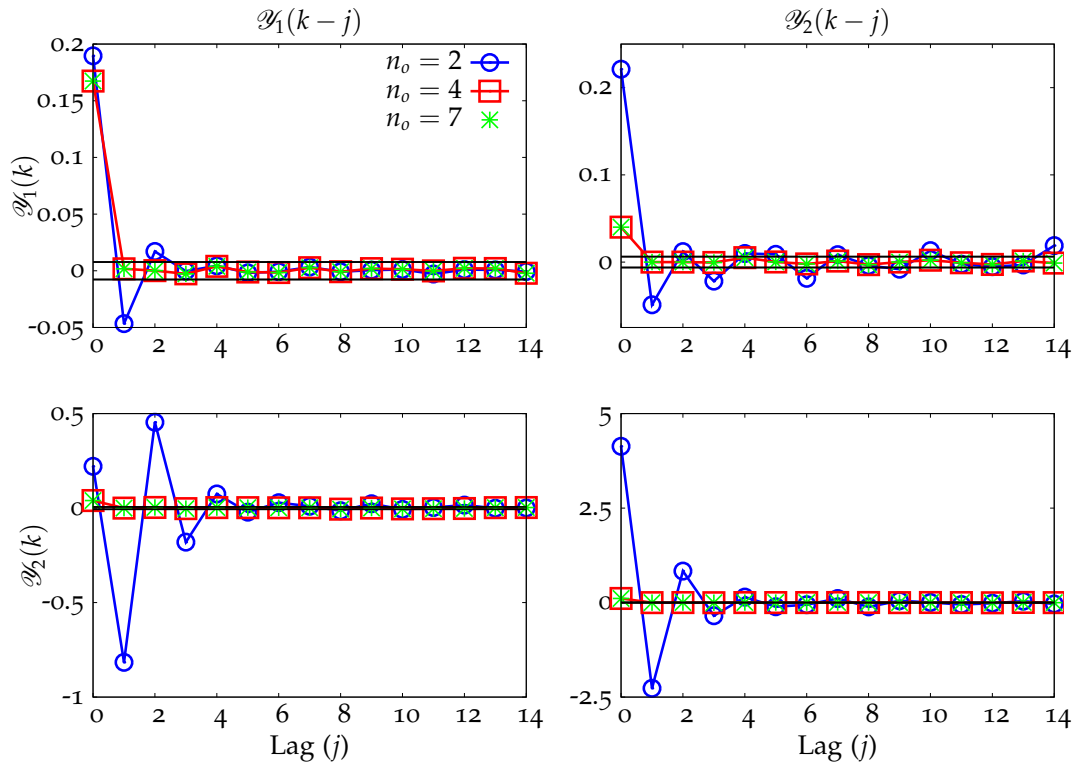
Figure 6.6: Autocovariances calculated using estimators designed from the ALS results for each model. The two-state model leads to suboptimal estimator behavior. The four- and seven-state models both lead to optimal estimators.

system is 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 6.1. 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 $\hat{R}_v$.

Altough the four- and seven-state models produce identical results, the advantage of using the smaller obsevable model lies in the computational time required

Table 6.1: $\hat{Q}_w$ and $\hat{R}_v$ estimated by the ALS method for models containing two, four, and seven states. (Note that the bolded elements in the seven-state $\hat{Q}_w$ are the same as $\hat{Q}_w$ for four states, while the remaining seven-state elements are near zero.)

$$n_o = 2$$

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

$$\hat{R}_v = \begin{bmatrix} 6.81\text{e-}10 & 3.95\text{e-}19 \\ 3.95\text{e-}19 & 8.08\text{e-}03 \end{bmatrix}$$

$$n_o = 4$$

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

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

$$n_o = 7$$

$$\hat{Q}_w =$$

$$\begin{bmatrix} \mathbf{3.15\text{e-}03} & \mathbf{3.48\text{e-}03} & \mathbf{2.16\text{e-}03} & \mathbf{-3.39\text{e-}03} & -1.10\text{e-}05 & 7.41\text{e-}09 & 1.49\text{e-}18 & \mathbf{1.77\text{e-}02} & \mathbf{2.02\text{e-}02} \\ \mathbf{3.48\text{e-}03} & \mathbf{1.01\text{e-}02} & \mathbf{3.21\text{e-}03} & \mathbf{-4.81\text{e-}03} & -8.37\text{e-}06 & 3.59\text{e-}08 & 8.51\text{e-}18 & \mathbf{-7.82\text{e-}03} & \mathbf{-2.22\text{e-}02} \\ \mathbf{2.16\text{e-}03} & \mathbf{3.21\text{e-}03} & \mathbf{1.69\text{e-}03} & \mathbf{-2.64\text{e-}03} & -3.91\text{e-}06 & 2.57\text{e-}08 & 2.06\text{e-}18 & \mathbf{4.25\text{e-}03} & \mathbf{1.54\text{e-}02} \\ \mathbf{-3.39\text{e-}03} & \mathbf{-4.81\text{e-}03} & \mathbf{-2.64\text{e-}03} & \mathbf{4.15\text{e-}03} & 5.58\text{e-}06 & -4.31\text{e-}08 & -3.01\text{e-}18 & \mathbf{-6.72\text{e-}03} & \mathbf{-2.75\text{e-}02} \\ -1.10\text{e-}05 & -8.37\text{e-}06 & -3.91\text{e-}06 & 5.58\text{e-}06 & 1.15\text{e-}06 & 5.39\text{e-}10 & -1.74\text{e-}21 & -2.17\text{e-}04 & 1.42\text{e-}04 \\ 7.41\text{e-}09 & 3.59\text{e-}08 & 2.57\text{e-}08 & -4.31\text{e-}08 & 5.39\text{e-}10 & 1.01\text{e-}06 & 4.07\text{e-}20 & -8.32\text{e-}07 & 1.14\text{e-}06 \\ 1.49\text{e-}18 & 8.51\text{e-}18 & 2.06\text{e-}18 & -3.01\text{e-}18 & -1.74\text{e-}21 & 4.07\text{e-}20 & 1.01\text{e-}06 & -2.49\text{e-}17 & -3.30\text{e-}17 \\ \mathbf{1.77\text{e-}02} & \mathbf{-7.82\text{e-}03} & \mathbf{4.25\text{e-}03} & \mathbf{-6.72\text{e-}03} & -2.17\text{e-}04 & -8.32\text{e-}07 & -2.49\text{e-}17 & \mathbf{4.09\text{e-}01} & \mathbf{-2.09\text{e-}02} \\ \mathbf{2.02\text{e-}02} & \mathbf{-2.22\text{e-}02} & \mathbf{1.54\text{e-}02} & \mathbf{-2.75\text{e-}02} & 1.42\text{e-}04 & 1.14\text{e-}06 & -3.30\text{e-}17 & \mathbf{-2.09\text{e-}02} & \mathbf{1.01\text{e+}00} \end{bmatrix}$$

$$\hat{R}_v = \begin{bmatrix} 1.90\text{e-}02 & -1.28\text{e-}16 \\ -1.28\text{e-}16 & 2.74\text{e-}02 \end{bmatrix}$$

to solve the ALS problem. By reducing the model from seven to four states, the computational time was reduced from 96.73 seconds to 42.89 seconds. 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.

6.3.2   *Example: Comparison of the ordinary ALS and feasible generalized ALS methods*

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 non-diagonal) for the reduced four-state model. Note that $\hat{Q}_w$ includes the noises on the integrating disturbances. Figure 6.7 shows the diagonal elements of the noise covariance matrices estimated from each data set; their variances are presented in Table 6.2. These results show that the feasible generalized ALS technique significantly reduces the variance of the estimates.
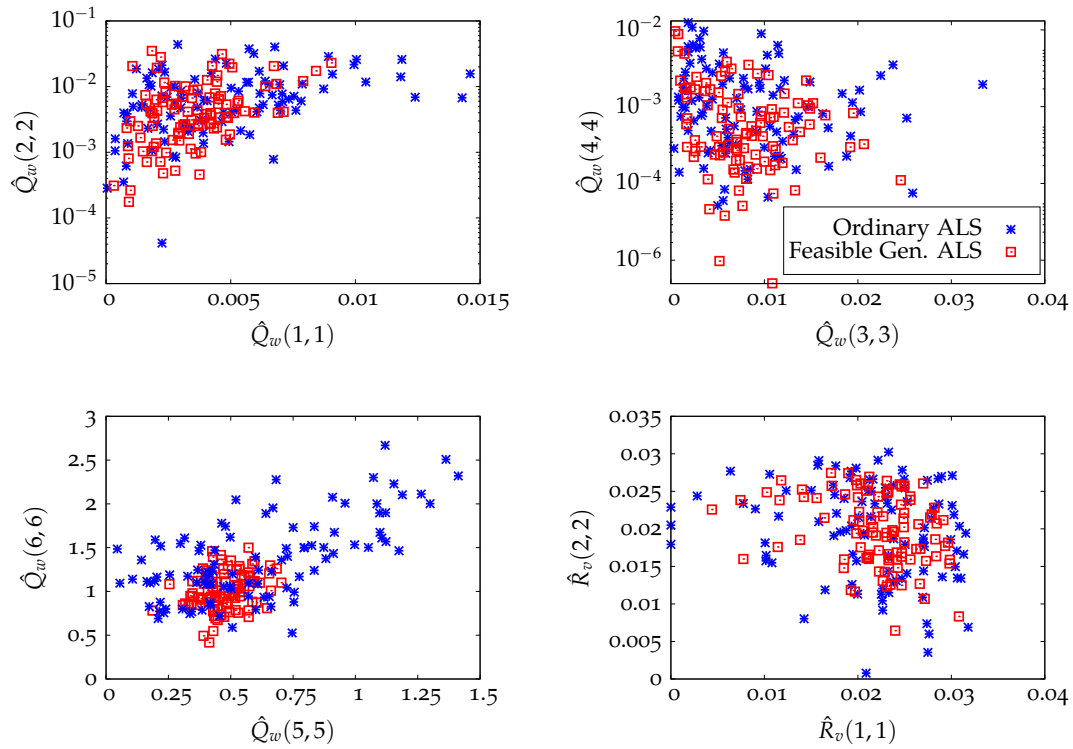
Figure 6.7: Noise variances estimated from the feasible generalized ALS and ordinary ALS methods. The variance is noticeably reduced by using the feasible generalized ALS method.

Table 6.2: Variance of each diagonal element of $\hat{Q}_w$ and $\hat{R}_v$ using the feasible generalized and ordinary ALS methods.

	Feasible Generalized ALS	Ordinary ALS
$\hat{Q}_w(1,1)$	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}

6.4 APPENDICES

6.A *Necessity of full rank A and C in Theorem 1*

Assume that $G = I$. From the simplification of equation (A.3) in Rajamani and Rawlings (2009, Appendix A), if q_N lies in the null space of $\mathscr{A}_1$, then

$$\left(I_p \otimes \mathscr{O}_1 A \right) \left(C \otimes I_n \right) \left(I_{n^2} - \bar{A} \otimes \bar{A} \right)^{-1} \mathscr{D}_n q_N = 0$$

in which $\mathscr{O}_1 = \begin{bmatrix} C' & \cdots & (C\bar{A}^{N-2})' \end{bmatrix}'$. We rewrite this condition as

$$\left(C \otimes \mathscr{O}_1 A \right) \left(I_{n^2} - \bar{A} \otimes \bar{A} \right)^{-1} \mathscr{D}_n q_N = 0$$

Let n_a be the dimension of the null space of A and $V_2 \in \mathbb{R}^{n \times n_a}$ be the null space of A. Following the logic in Hua (1990), for $n_a > 0$, we choose a full rank, symmetric matrix $Z \in \mathbb{R}^{n_a \times n_a}$ and construct a symmetric matrix $X_N = V_2 Z V_2'$. Since $A X_N = 0$, $\text{vec}(X_N)$ lies in the null space of $(C \otimes \mathscr{O}_1 A)$. Letting $Q_N = X_N - \bar{A} X_N \bar{A}$, then $\text{vec}(Q_N) = (I - \bar{A} \otimes \bar{A})\text{vec}(X_N)$. Since Q_N is symmetric, $q_N = (Q_N)_{ss}$ lies in the null space of $\mathscr{A}_1$. Since Z is full rank and V_2 is full column rank, X_N and q_N are non-zero. Therefore, for $G = I$, $\mathscr{A}_1$ is full rank only if A is full rank.

Likewise, let $\bar{p}$ be the rank of C and $V_3 \in \mathbb{R}^{n \times n - \bar{p}}$ be the null space of C. Then for $\bar{p} < n$, choose a full rank, symmetric matrix $Y \in \mathbb{R}^{n - \bar{p} \times n - \bar{p}}$ and construct a symmetric matrix $W_N = V_3 Y V_3'$. Since $W_N C' = 0$, $\text{vec}(W_N)$ lies in the null space of $(C \otimes \mathscr{O}_1 A)$. Letting $Q_N = W_N + \bar{A} W_N \bar{A}$, then $\text{vec}(Q_N) = (I - \bar{A} \otimes \bar{A})\text{vec}(W_N)$. Since Q_N is symmetric, $q_N = (Q_N)_{ss}$ lies in the null space of $\mathscr{A}_1$. Since Y is full rank and V_3 is full column rank, W_N and q_N are non-zero. Therefore, for $G = I$, $\mathscr{A}_1$ is full rank only if $\text{rank}(C) = n$.

6.B   *Derivation of formula for $S = cov(\hat{b})$*

Let

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

in which $Y_i = \left[y_i(1)' \quad \dots \quad y_i(N)' \right]'$ are independent and identicallly distributed

normal variables with zero mean and variance $P_y = \begin{bmatrix} P_0 & P'_{y,0} \\ P_{y,0} & P_{y2} \end{bmatrix}$. In this appendix,

we show that the variance of $\hat{b}$ is

$$cov(\hat{b}) = \frac{1}{N_s} \left((P_0 \otimes P_y) + K_{p,\tilde{p}} \left(P_{y,0} \otimes P'_{y,0} \right) \right) \tag{6.8}$$

where p is the dimension of y and $\tilde{p} = Np$ is the dimension of Y. Defining Y_i

and y_i appropriately in (6.8), we arrive at (6.7). To derive (6.8), we begin by noting

that the sample variance of Y_i, $\hat{P}_y = \frac{1}{N_s} \sum_{i=1}^{N_s} (Y_i Y_i')$, is distributed according to the

Wishart distribution with pdf

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

(Anderson, 2003; Ghosh and Sinha, 2002). As a result,

$$var((\hat{P}_y)_s) = \frac{1}{N_s} \left(I_{\tilde{p}^2} + K_{\tilde{p},\tilde{p}} \right) (P_y \otimes P_y) \tag{6.9}$$

(Magnus and Neudecker, 1979). The estimated variance of $\hat{b}$ is the $p\tilde{p} \times p\tilde{p}$ matrix

in the upper-left corner of $var((\hat{P}_y)_s)$. From (6.9), we can write this submatrix as

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

which is the formula in (6.8). Note that the denominator in (6.9) and (6.8) is N_s rather than $N_s - 1$ because the mean of Y_i is known to be zero.

7

APPLICATION OF THE ALS METHOD TO AN INDUSTRIAL
DATA SET[1]

7.1 NOISE COVARIANCE ESTIMATION

In this chapter, we apply the ALS method to an air separation unit operated by Praxair, Inc. We analyze a subset of the variables included in the 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 (since we have measurements available for both variable types).

The MPC controller uses an 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 on the order of 10^4, where as y_1 has a variance around 1. Therefore, we normalized the data by dividing each output and each row of C by the stan-

---

[1] Portions of this chapter will be published in Zagrobelny and Rawlings (2014a)

dard deviation of y_i. The state, input, and process noise remain the same in the normalized model. The ALS estimate of $\hat{Q}_w$ is consistent for both the normalized and original data, and the estimate of $\hat{R}_v$ corresponds to the normalized outputs. Since we assume R_v is diagonal, we can easily 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 then 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

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

was added to y_1. The use of double integrator disturbance models is discussed in 5.3.

The full state space model contains 31 states and is unobservable, with a condition number of 3.24×10^{15}. Figure 7.1 shows the condition number and ALS objective function value versus 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 first compared these two models by plotting the estimated autocovariances (calculated from the data) alongside the theoretical autocovariances (calculated from the model and the ALS estimates of Q_w and R_v). Since these plots use the original estimator, we considered whether or not the estimated and theoretical
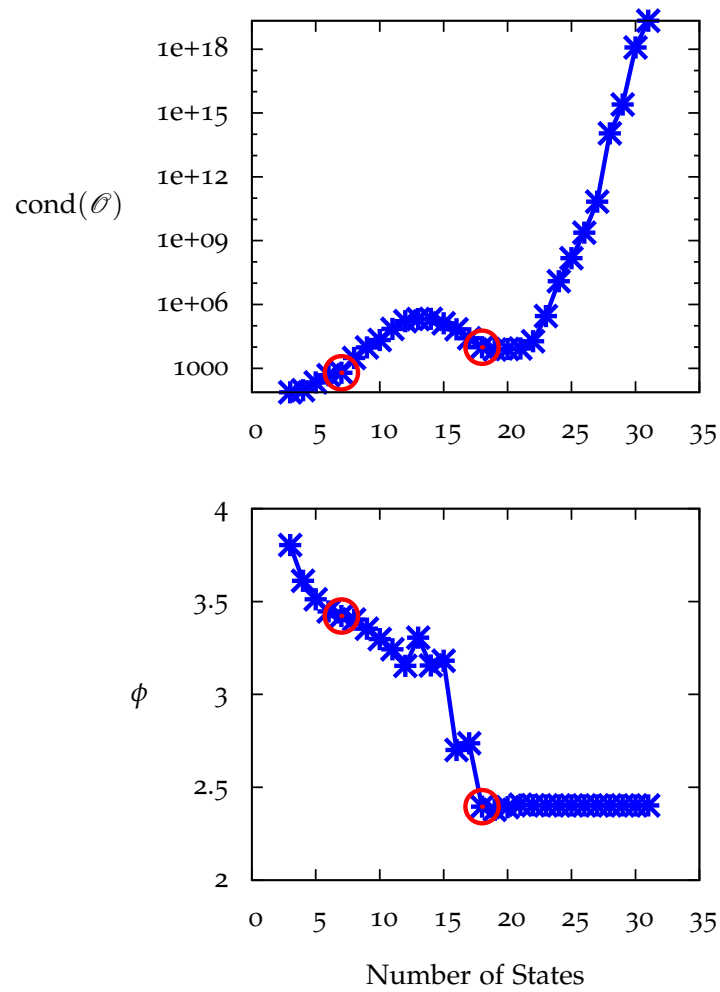
Figure 7.1: Condition number of the observability matrix and ALS objective value vs. the number of states. The seven-state model was chosen as a trade-off between the condition number, the number of states, and the ALS objective function value. The 18-state model was also examined.
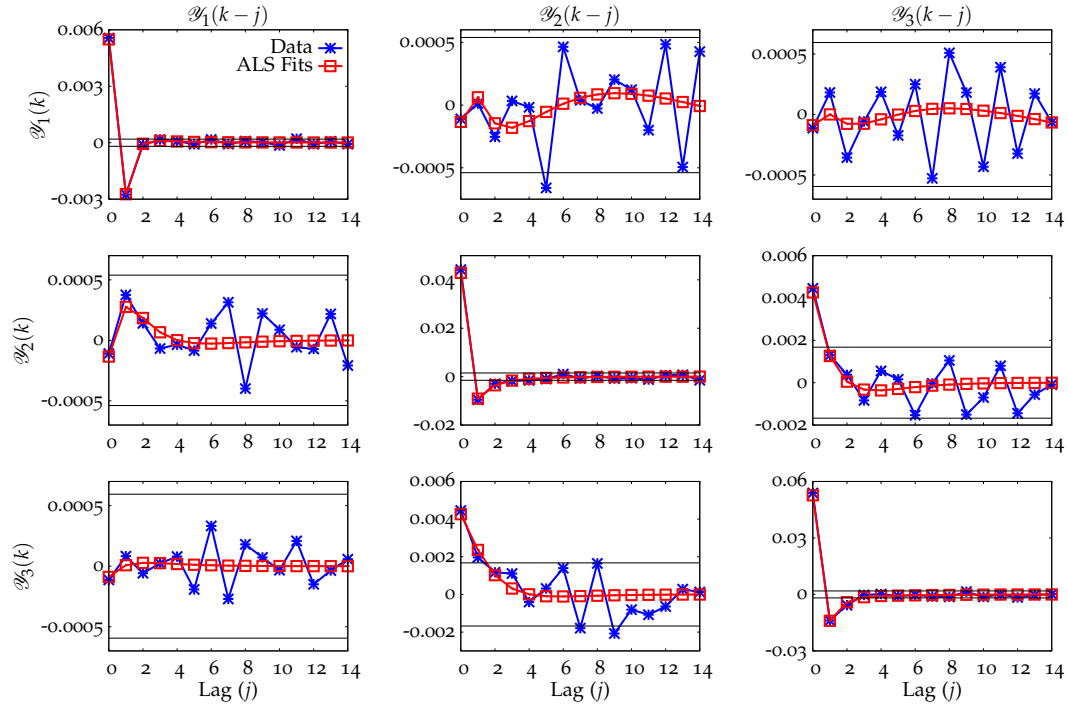
Figure 7.2: Sample and theoretical autocovariances for the seven-state model, without semidefinite constraints. The seven-state model adequately fits the data.

autocovariances match well and did not examine whether or not the estimator is performing optimally. Figures 7.2 and 7.3 show that the ALS estimates fit the data well in both models. The 18-state model does fit the cross-covariance terms with more accuracy, but as these terms are near zero, such an accurate fit may not be necessary. Therefore, we also compare the optimal estimator performance of the models.

To compare the optimal estimator performance, we solved the ALS problem for both models and designed an estimator based on the results for each model. We processed the data with each of these estimators and computed the autocovariances. Figure 7.4 shows that both estimators have near-optimal performance. However, the computation time increased from 19.2 seconds to 465 seconds 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.
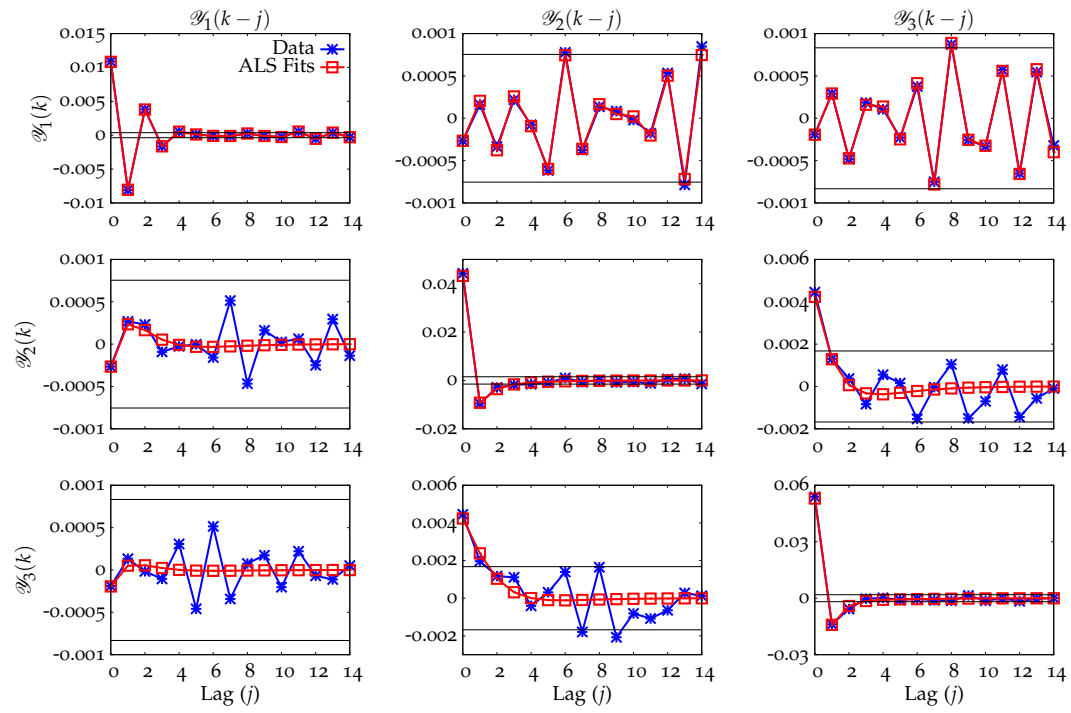
Figure 7.3: Sample and theoretical autocovariances for the 18-state model, without semidefinite constraints. The model fits the data well; increasing the number of states improves the fits of the cross-correlations.
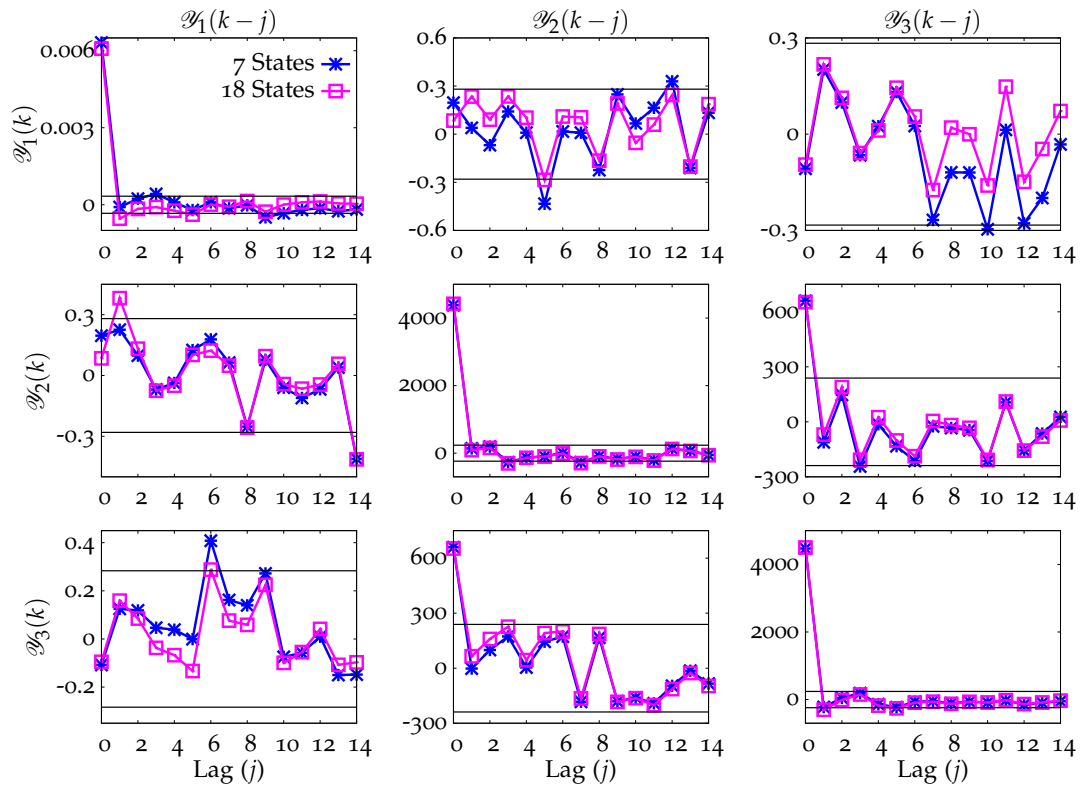
Figure 7.4: Autocovariances for the seven- and 18-state models, using estimators designed from the ALS results. Both models produce excellent estimators.
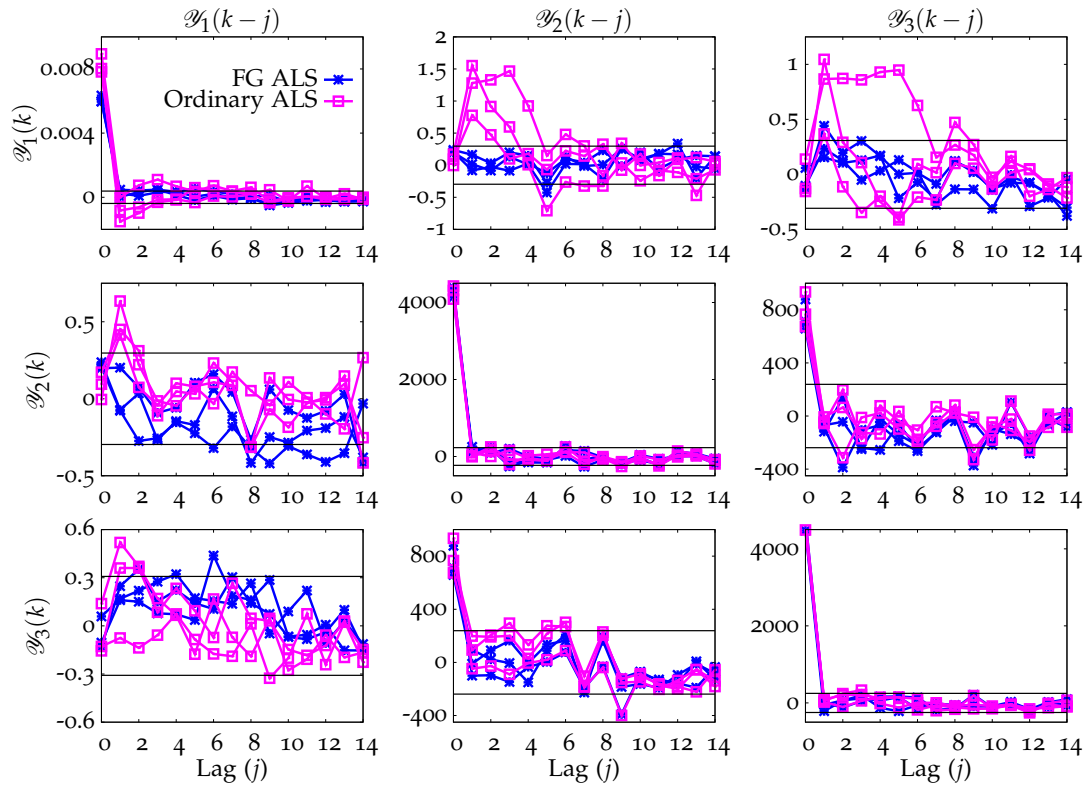
Figure 7.5: Autocovariances for the feasible generalized and ordinary ALS methods. The feasible generalized ALS method results reduces the covariance of $\mathscr{Y}_1$ and whitens the cross-correlations of $\mathscr{Y}_1$ with $\mathscr{Y}_2$ and $\mathscr{Y}_3$.

Using the seven-state model, we compared the feasible generalized ALS technique with the ordinary ALS technique by using each method on three data sets. The three sets overlap in the manner shown in Figure 7.8. As seen in Figure 7.5, using the feasible generalized ALS method reduces the variance of the innovations for y_1 and also reduces the cross-correlation between y_1 and the other outputs at higher lags.

Finally, we examined the consistency of the ALS results. The data studied is shown in Figure 7.6. 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, which we call L_1. We then processed the rest of the data using L_1 and calculated the L_1-innovation autocovariances. As shown in Figure 7.7, the estimator
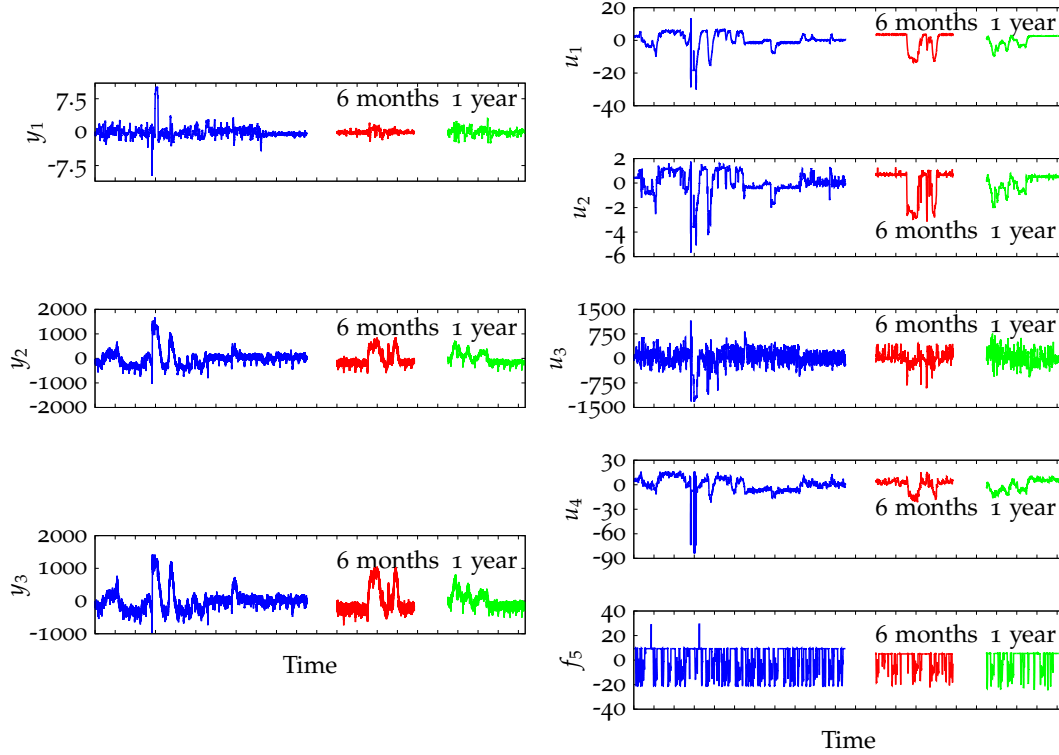
Figure 7.6: Industrial data analyzed in this work.

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 that 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 7.8) and significant correlation remains in the L_1-innovations (shown in Figure 7.9). 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 these data sets, 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 7.10, although its behavior may still be considered acceptable. After applying the ALS method to one of
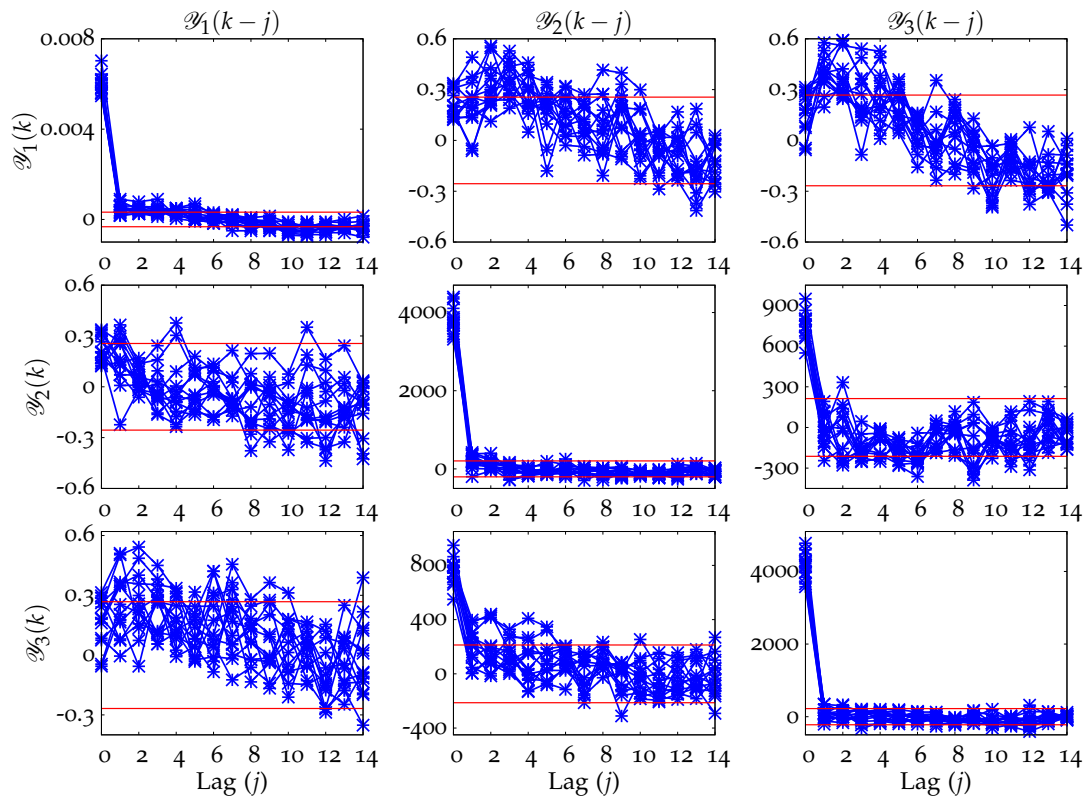
Figure 7.7: Autocovariances for the data from the first time period, using an ALS-based estimator. The estimator was calculated from data set 1 and applied to all the data sets. It performs optimally on the 13 out of 20 data sets from this time period.

Figure 7.8: Innovations for data sets 5-7 in the first time period. A large disturbance is evident in the innovations.
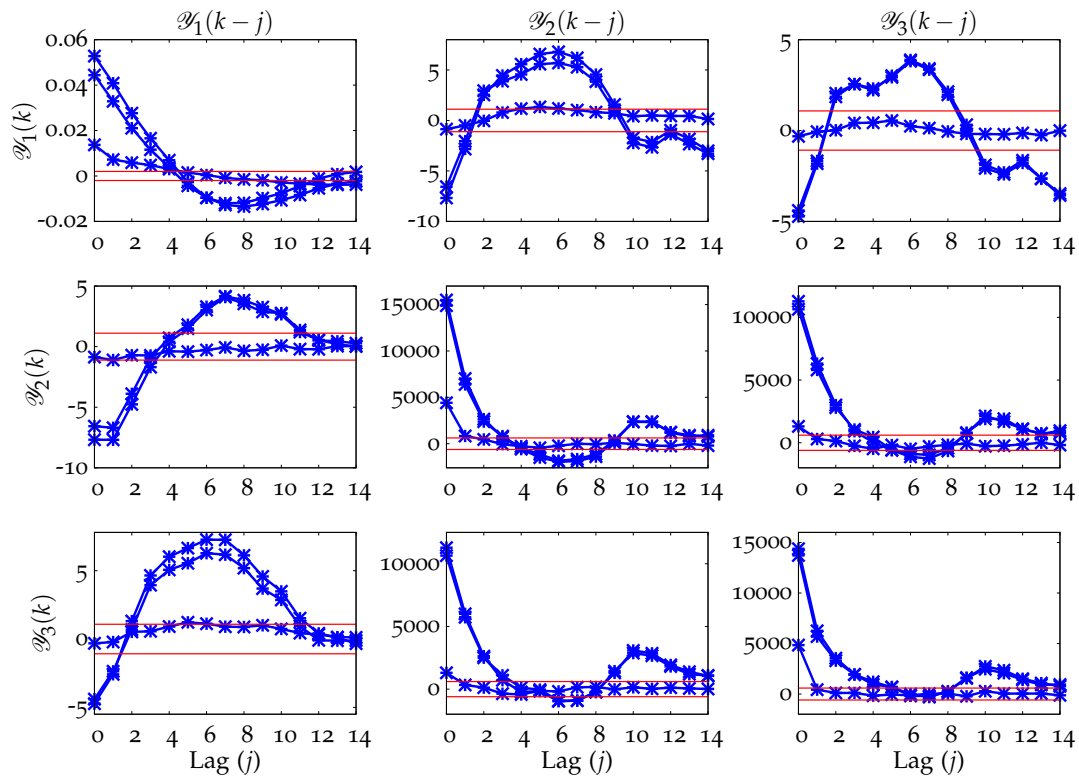
Figure 7.9: Autocovariances for data sets 5-7 in the first time period. Due to the unexplained disturbance, the estimator from data set 1 performs poorly.
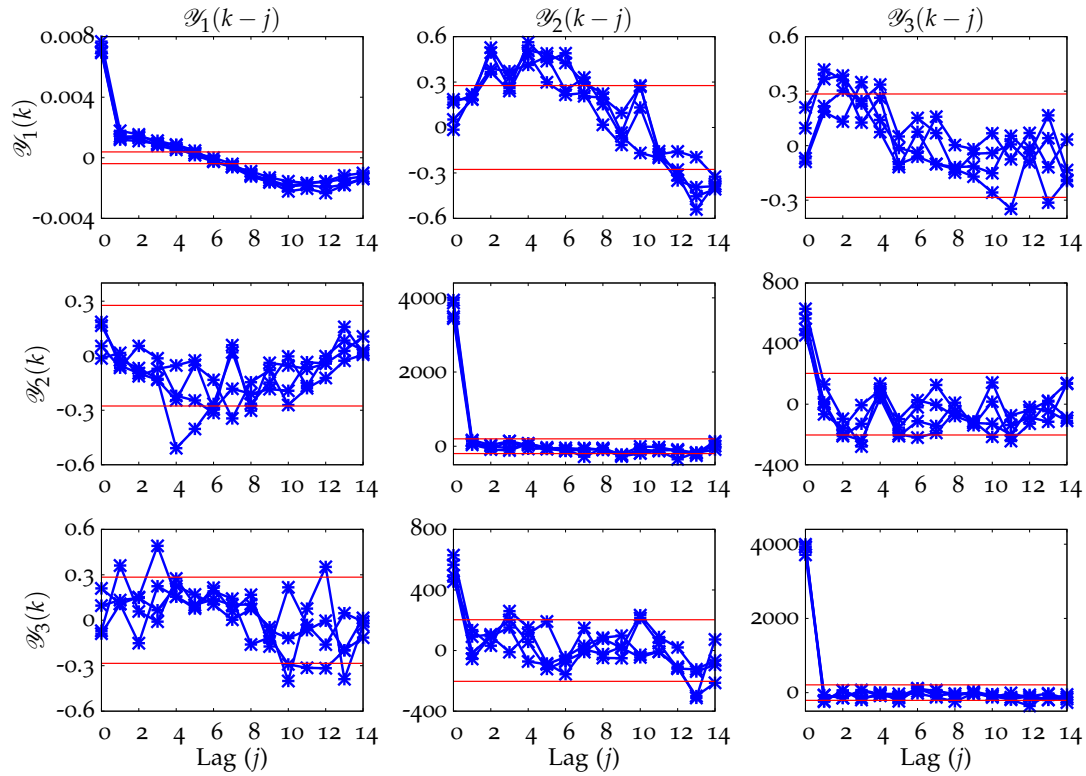
Figure 7.10: Autocovariances for data sets 17-20 in the first time period. The data were processed using the estimator from data set 1; this estimator is slightly suboptimal for these data.

these data sets, we produced a new estimator L_2. When we process the data with L_2, the innovations are white (Figure 7.11). 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 six months later and one 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 7.12 and 7.13. 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, since the disturbance model identified at the beginning of the year produces an estimator that is nearly optimal throughout the year. We expect that the same disturbance model would also be reliable for use

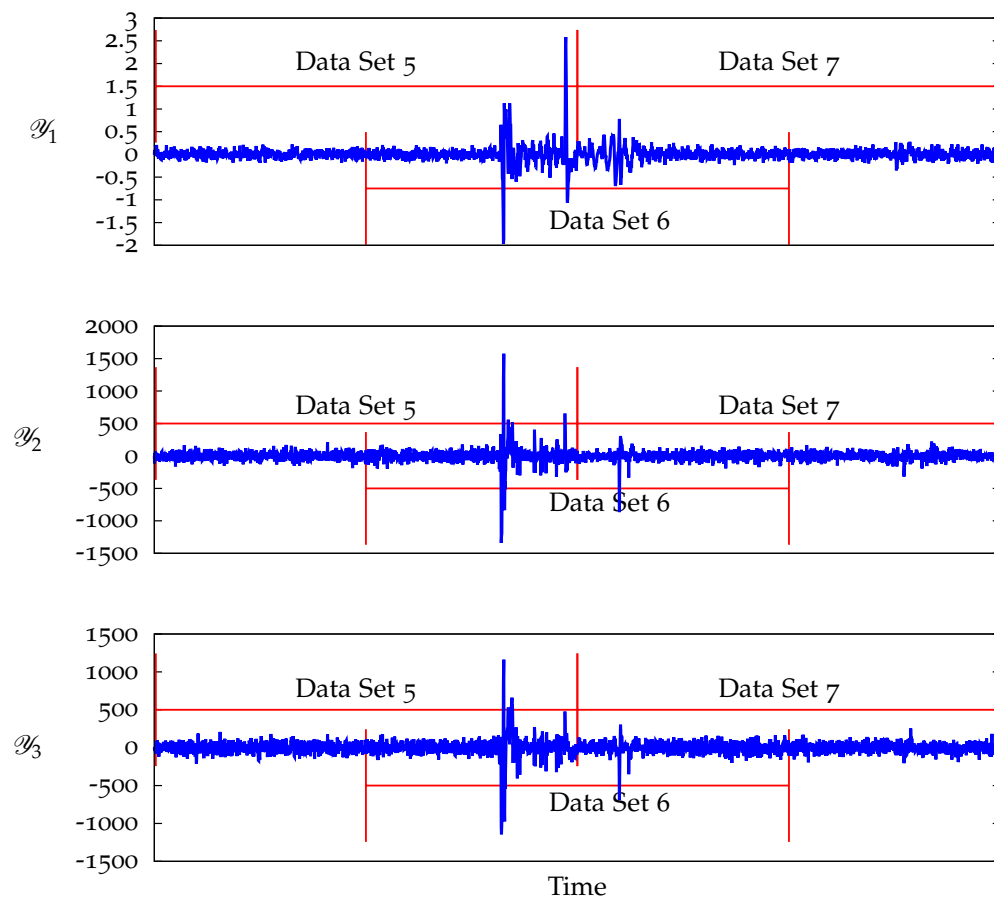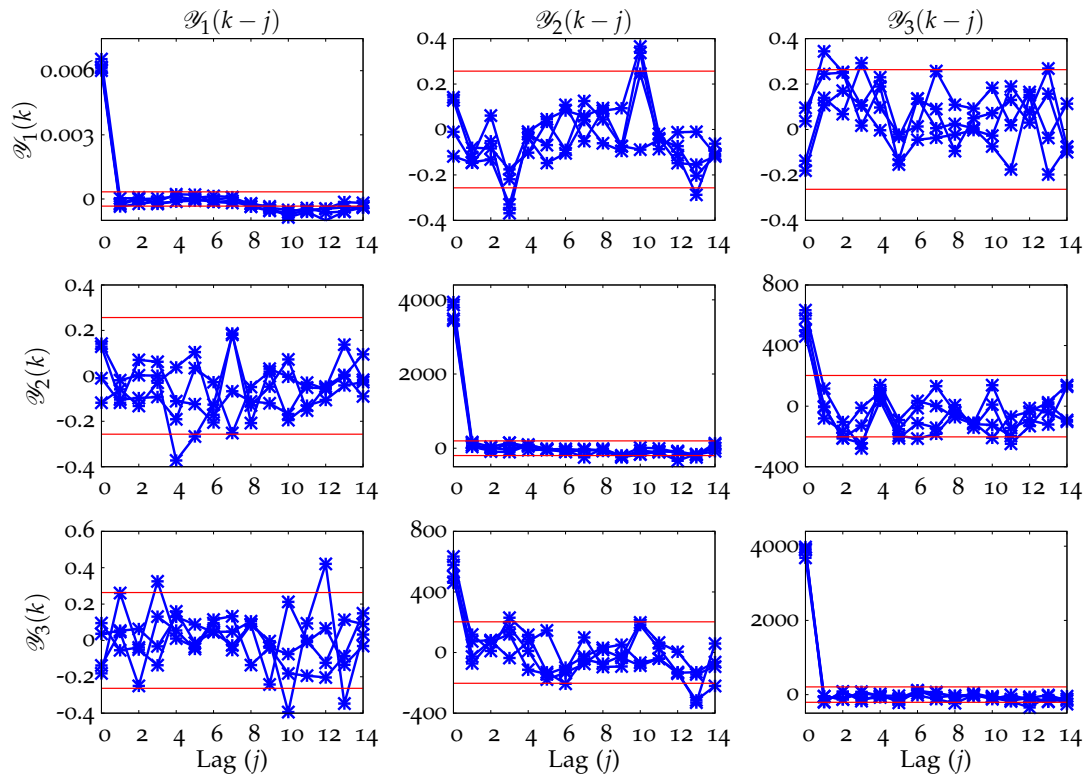Figure 7.11: Autocovariances for data sets 17-20 in the first time period using a re-identified noise model. The estimator was designed from the ALS results on data set 18. The noise model is more accurate, and the estimator behaves optimally.
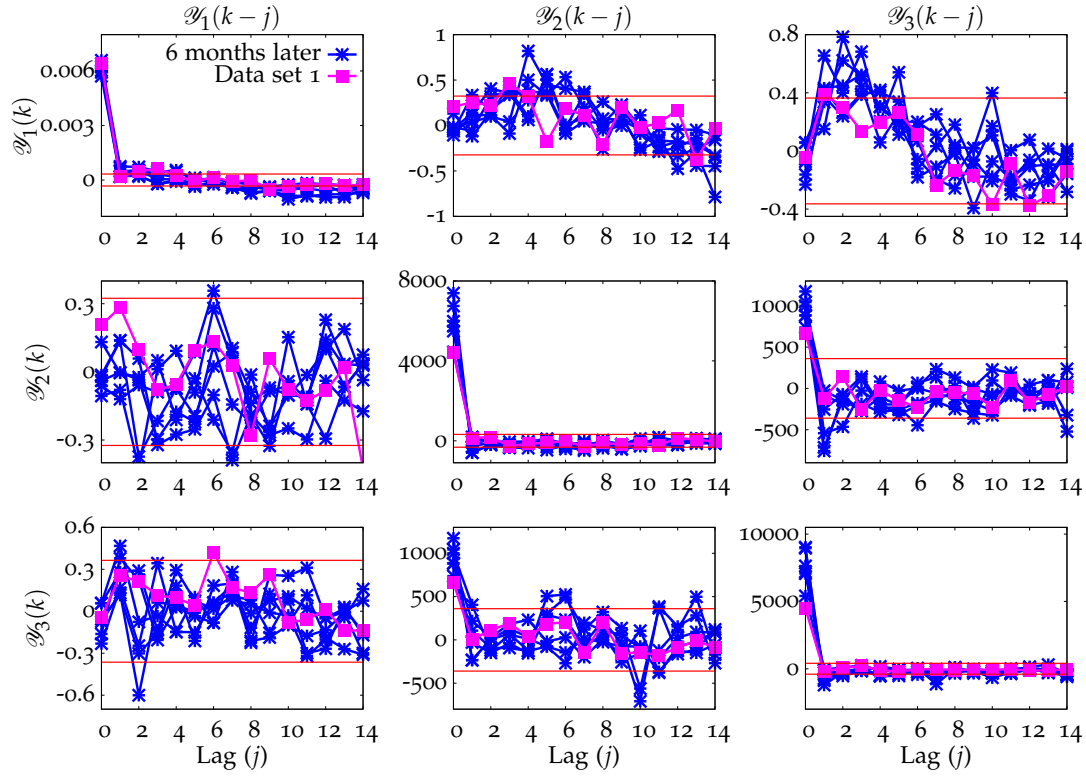
Figure 7.12: Autocovariances for the data sets collected six months later. The estimator from data set 1 (of the first time period) is nearly optimal for the later data. For comparison, the autocovariances for data set 1 are plotted with pink squares.

in calculating performance monitoring benchmarks although such an illustration is beyond the scope of this work.

7.2 CLOSED-LOOP SIMULATION

We performed a simulation study to compare the performance of the optimal ALS disturbance model to a DMC-type disturbance model in a closed-loop controller. The ALS-based disturbance model contains a double integrator on y_1 and single integrators on the other two outputs; the ALS estimates are used as the process and measurement noise covariances. In the DMC-type disturbance model, a single integrator is placed on each of the three outputs. Measurement noise is included
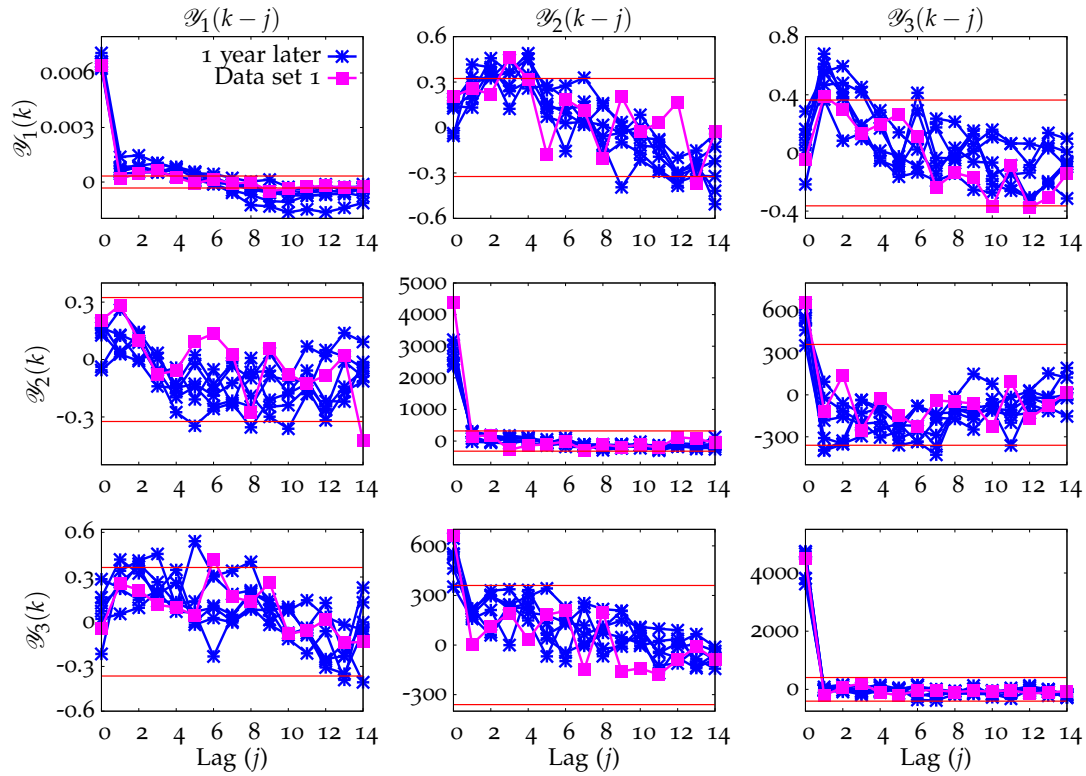
Figure 7.13: Autocovariances for the data sets collected one year later. Again, the estimator from data set 1 (of the first time period) performs nearly optimally on these data. Slightly better performance is achieved when a new noise model and estimator are identified. For comparison, the autocovariances for data set 1 are plotted with pink squares.

as DMC-type controllers often filter the output, but no process noise (except that on the integrators) is added. The ratio of measurement noise to integrator noise was chosen based on the ALS results to give the best possible performance for that disturbance model type. We designed estimators using both the ALS-based and the DMC-type disturbance models.

To compare the closed-loop performance of the estimators, we simulated the plant assuming no model mismatch. We added white noise disturbances with covariances equal to the ALS results. We also added deterministic disturbances, which were estimated from the plant data, to each of the outputs. We used three different schemes for the regulator: minimum variance, approximately equal

penalties on y and Δu, and very high penalty on Δu. We assessed the performance of the controllers by using the KPI as defined in Chapter 3. In addition to showing the total plant KPI in Table 7.1. Note that the KPIs are weighted based on the regulator tuning.

To further compare the disturbance models, we plot the outputs and inputs for each of the controllers. Note that the fifth input in each plot is the feed-forward variable and is not manipulated by the controller. We also plot the deterministic disturbances and the integrator estimates. Since the disturbances affect the output directly and the integrators are added to the outputs, we expect the integrators to track the disturbances, so we plot the estimates along with the disturbances. (When there is a mismatch in the disturbance location, the integrators track a transformation of the disturbances, so these plots would not be informative). We also plot the estimate of the slope of the disturbance in the double integrator case.

Under minimum variance tuning, the ALS-based controller yields much better closed-loop performance than does the DMC-type controller, as shown in Figure 7.14. Because the ALS-based estimator provides better state and disturbance estimates, the regulator rejects the disturbances more completely, and the controller achieves a much lower output variance, as compared to the controller with the DMC-type disturbance model. The lower output variance comes at the cost of increased use of the inputs (Figure 7.15). However, under this control strategy, reducing the output variance is far more important than reducing the input variance.

Figures 7.17-7.19 depict the behavior for a more realistic regulator, in which both the output variance and the rate of change of inputs are minimized. For this regulator tuning, the DMC-type and ALS-based controllers behave much more similarly, although the ALS estimator still results in improved performance.

The last control strategy, shown in Figures 7.20-7.22, has high move suppression. Therefore, the inputs move little and the outputs are highly correlated with

Table 7.1: Performance metrics for the ALS-based and DMC-type estimators in closed-loop simulations.

	KPI_y		$\text{KPI}_{\Delta u}$		KPI	
	ALS	DMC	ALS	DMC	ALS	DMC
Minimum variance	0.190	0.240	9.62e-8	9.70e-9	0.190	0.239
Balanced	0.317	0.423	0.0182	0.106	0.335	0.434
High move suppression	1.15	1.19	0.665	0.547	1.82	1.74

the deterministic disturbances, which are not rejected well. In this case, there is essentially no difference in the performance of the two controllers because the estimate has little effect on the control action.

In these simulations, the double integrator model was used only in the estimator, not in the regulator. The regulator chose the control action based on steady-state targets for the state and input, but these targets were calculated under the assumption that the slope of the double integrator is zero. The performance would have a greater improvement if the regulator properly accounted for the estimated slope of the disturbance.
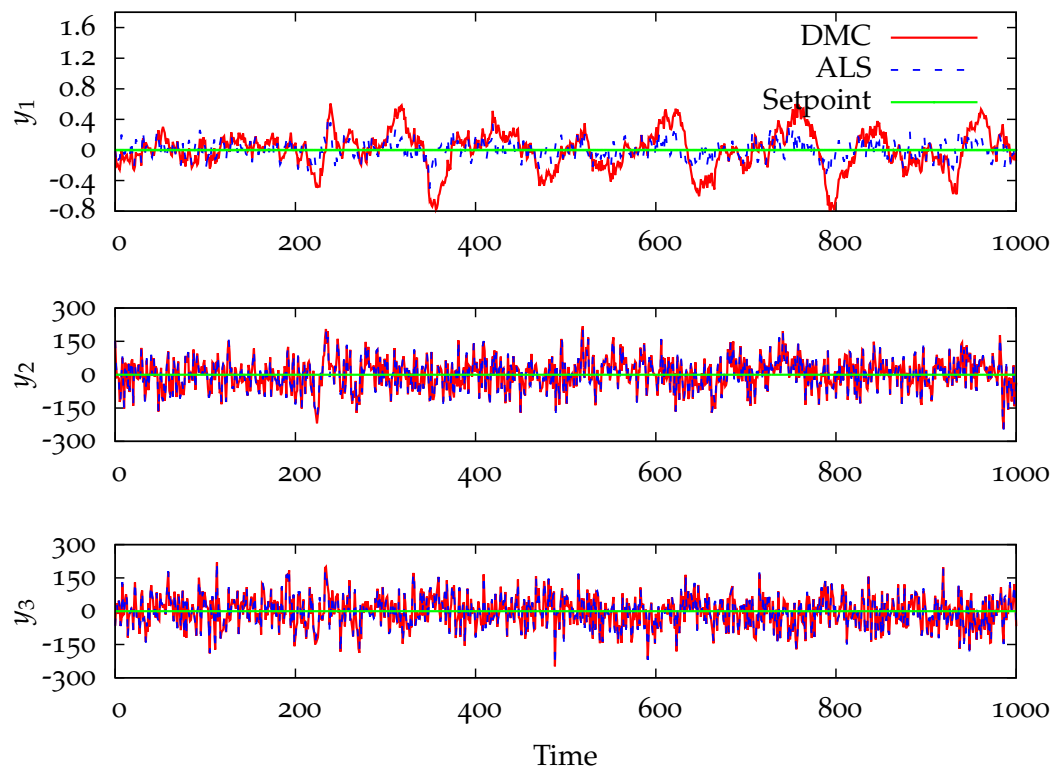
Figure 7.14: Outputs under minimum variance control, using the ALS-based and DMC-type disturbance models. The ALS-based disturbance model significantly reduces the variance of y_1.
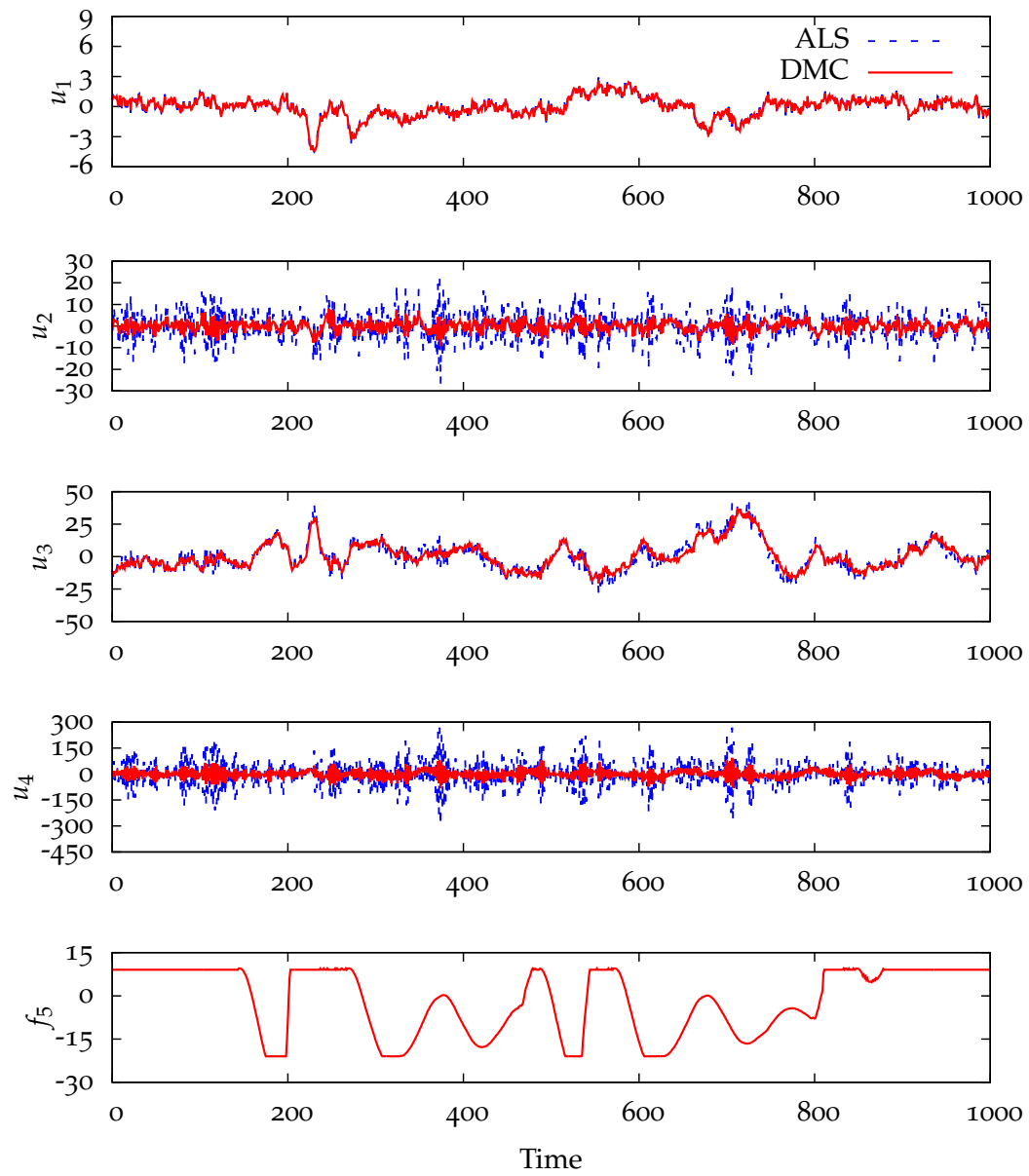
Figure 7.15: Inputs under minimum variance control, using the ALS-based and DMC-type disturbance models. Since it better reduces the output variance, the ALS-based disturbance model has significantly higher input variance than does the DMC-type disturbance model.
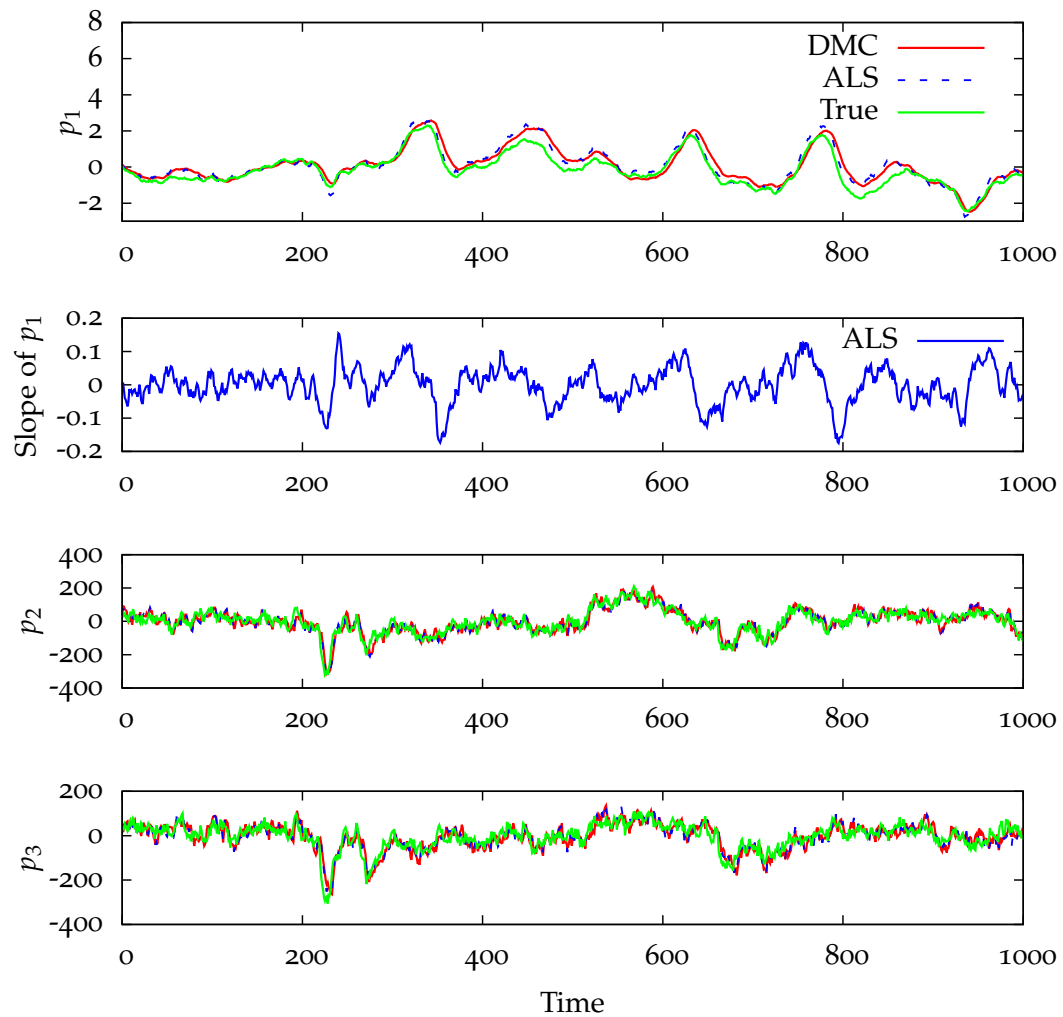
Figure 7.16: Disturbances and estimates under minimum variance control, using the ALS-based and DMC-type disturbance models. The ALS-based disturbance model better estimates p_1.
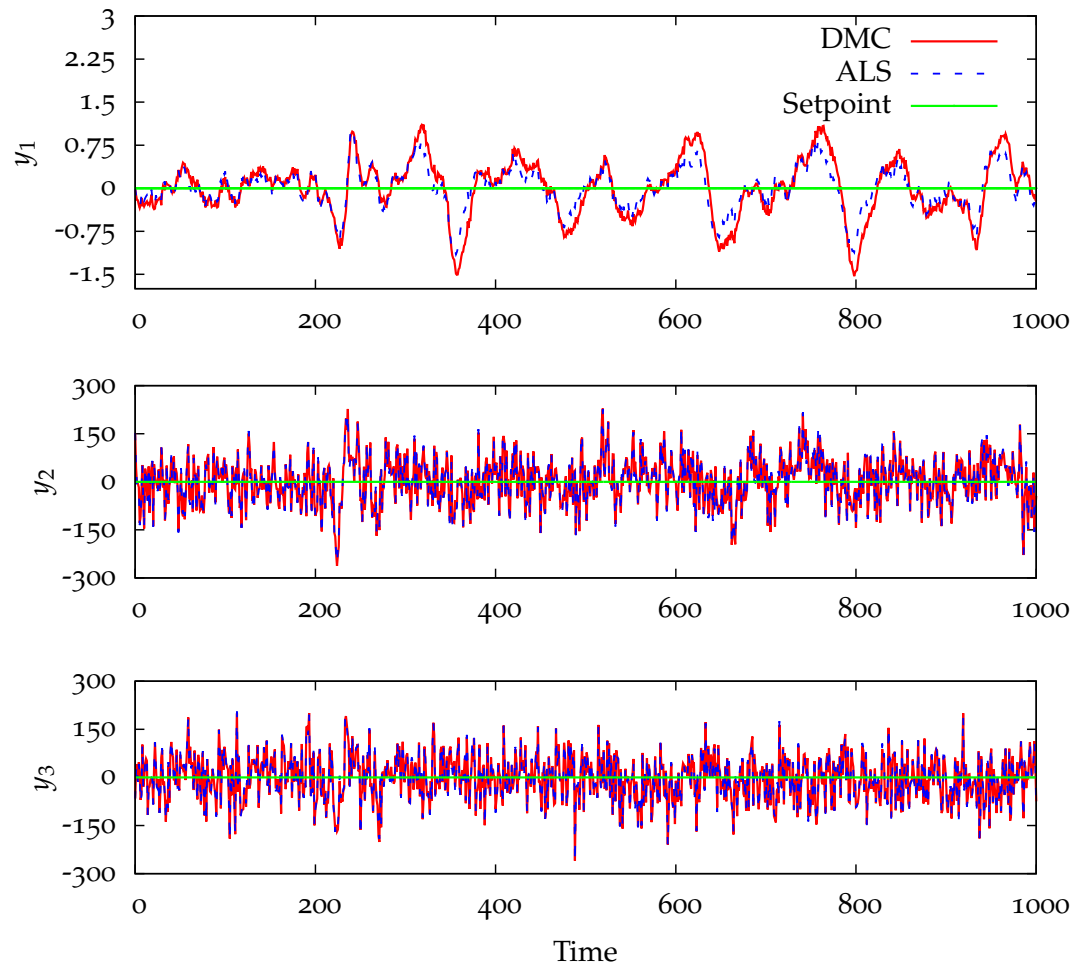
Figure 7.17: Outputs under the second regulator tuning, using the ALS-based and DMC-type disturbance models. The ALS-based disturbance model consistently reduces the variance of y_1.
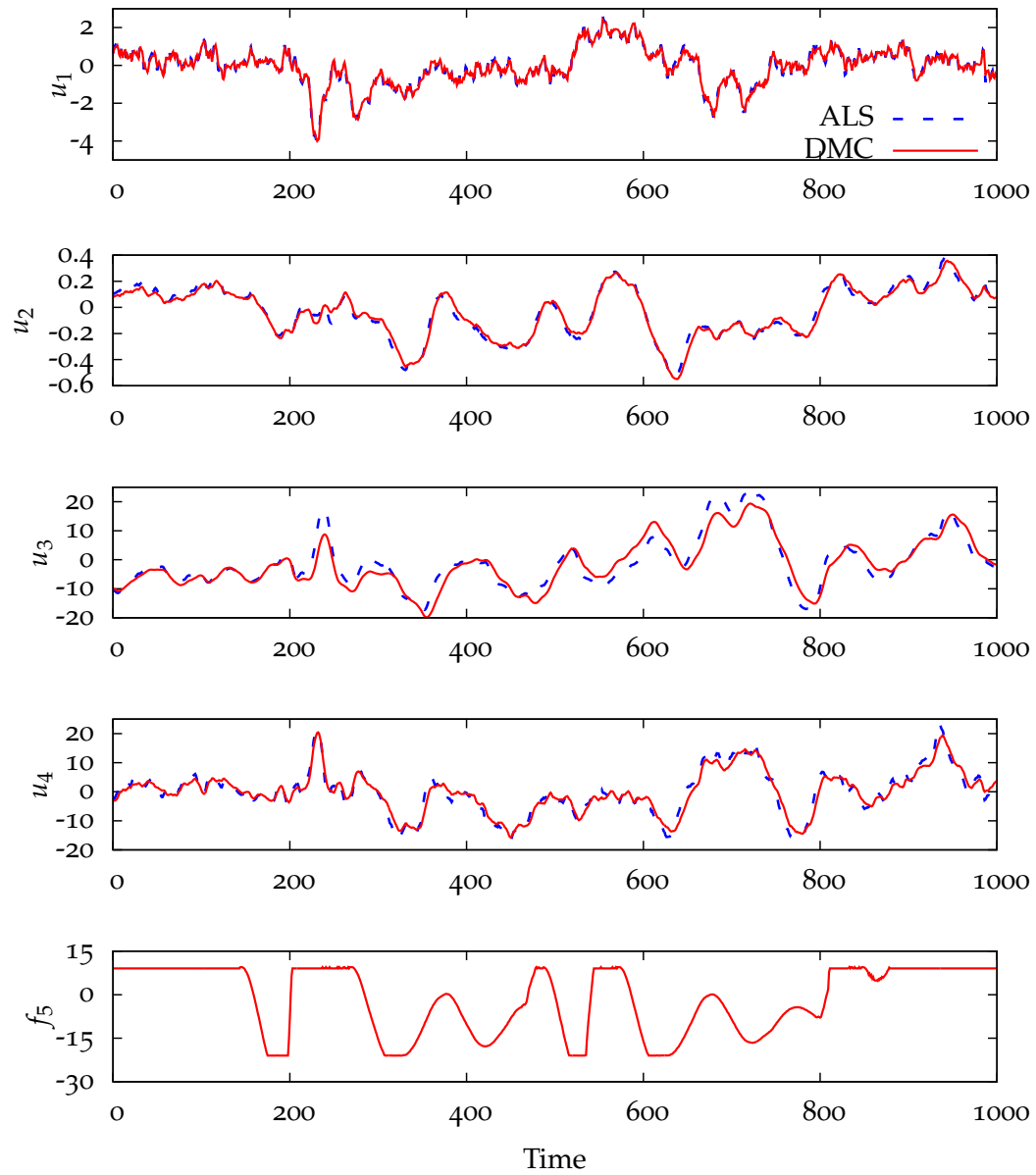
Figure 7.18: Inputs under the second regulator tuning, using the ALS-based and DMC-type disturbance models. The inputs are similar, although the ALS-based disturbance model has slightly higher input variance.
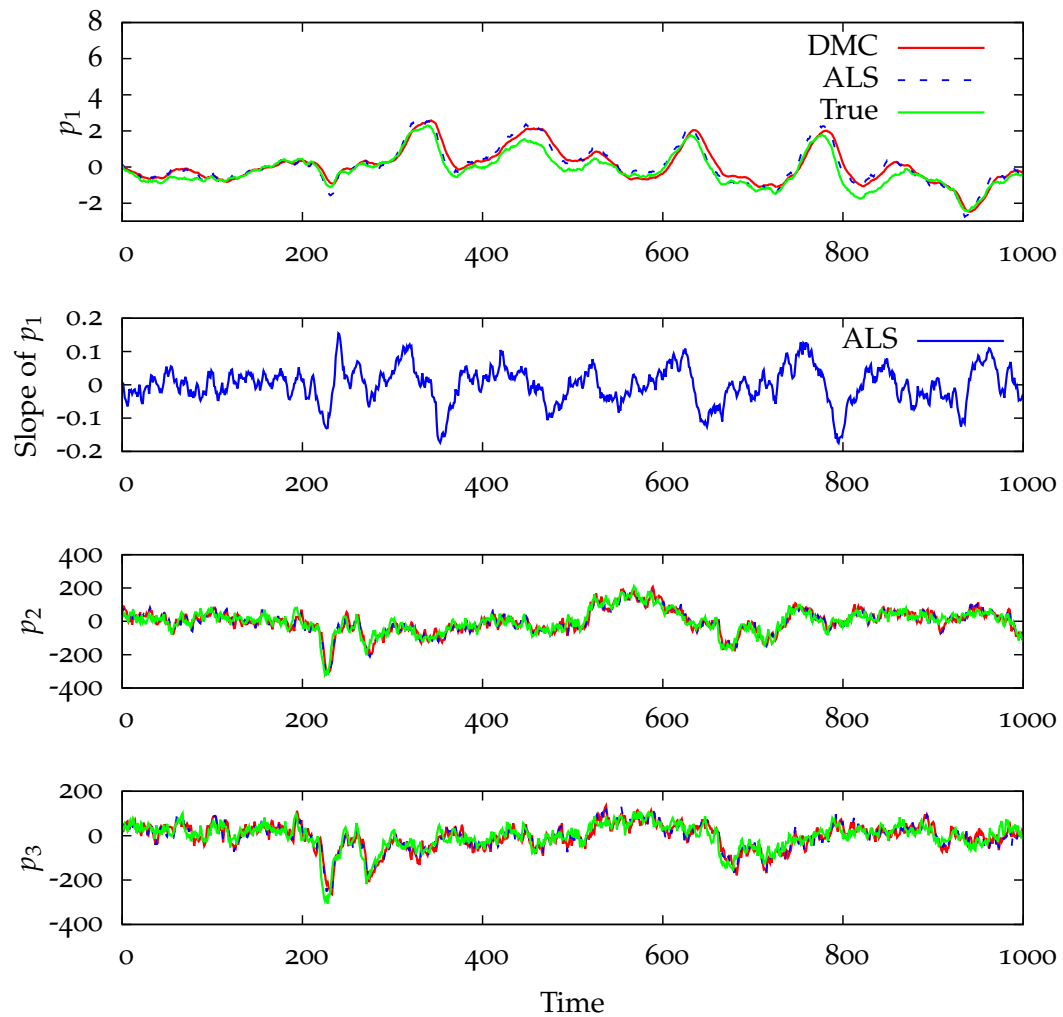
Figure 7.19: Disturbances and estimates under the second regulator tuning, using the ALS-based and DMC-type disturbance models. The ALS-based disturbance model gives improved performance.
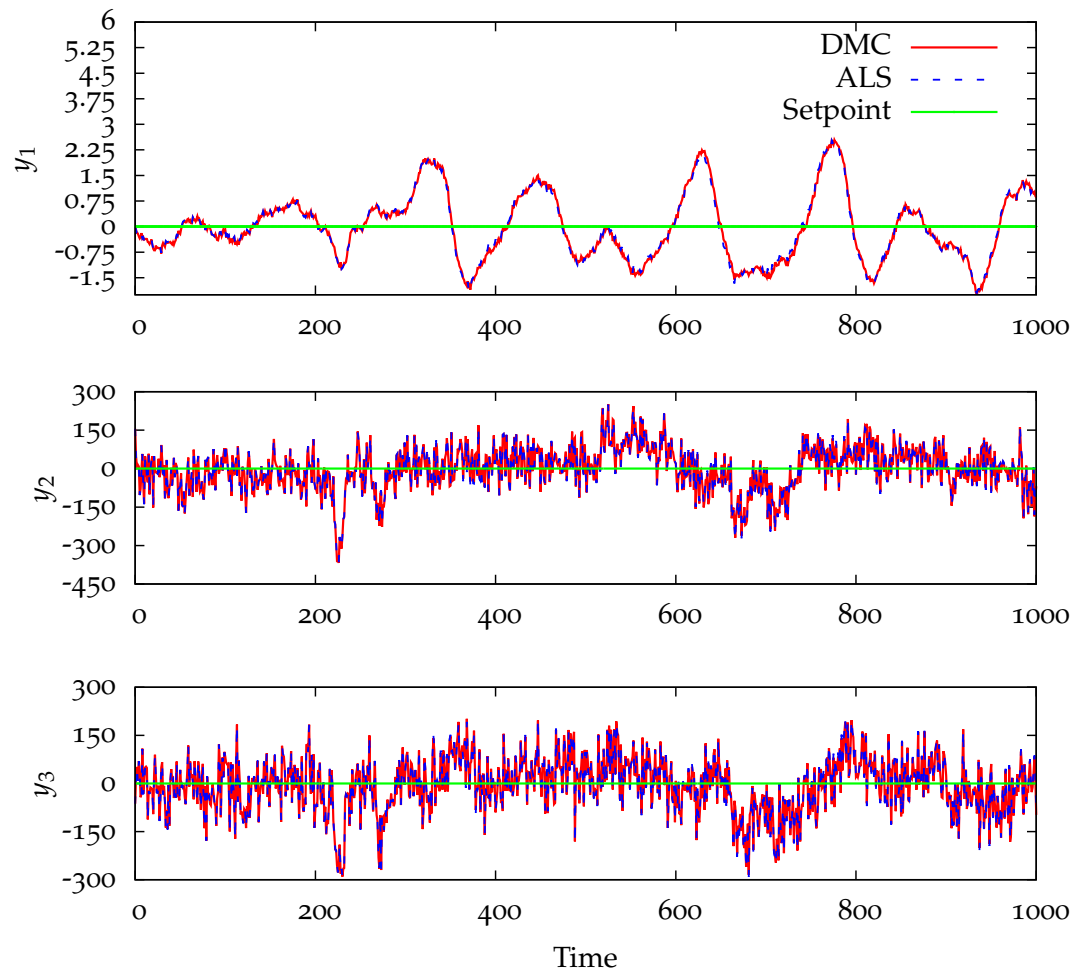
Figure 7.20: Outputs under high move suppression, using the ALS-based and DMC-type disturbance models. The controllers give nearly identical performance.
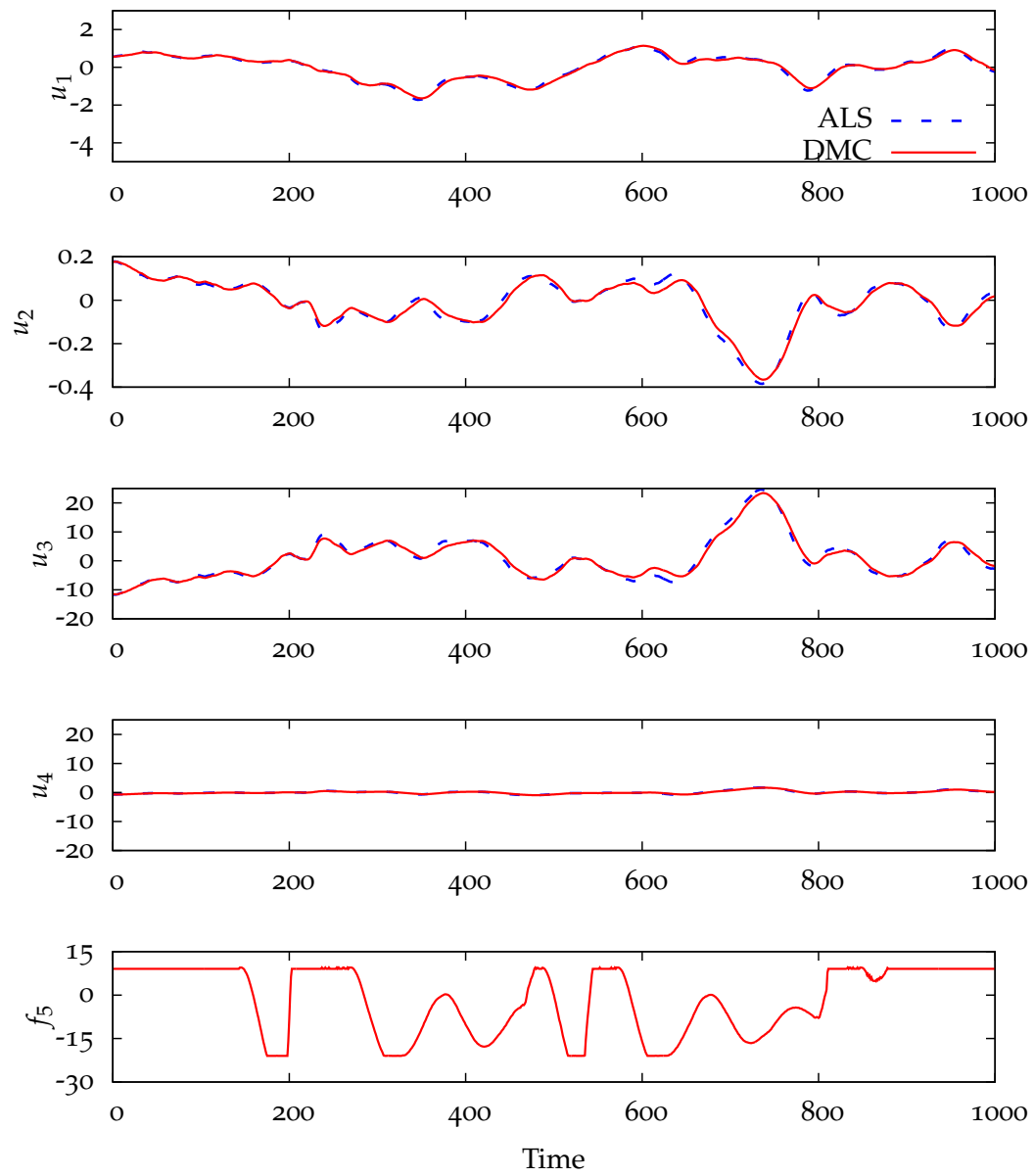
Figure 7.21: Inputs under high move suppression, using the ALS-based and DMC-type disturbance models. The controllers give nearly identical performance.
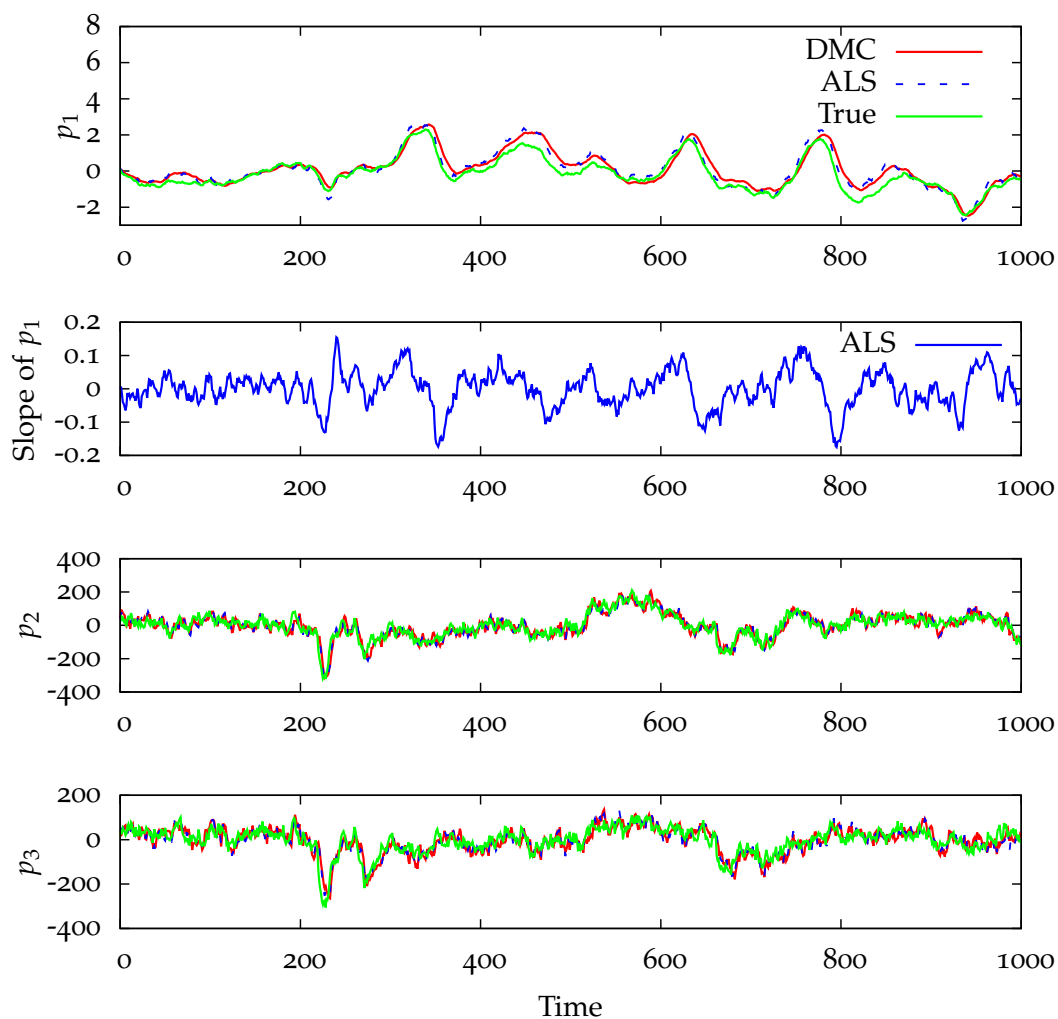
Figure 7.22: Disturbances and estimates under high move suppression, using the ALS-based and DMC-type disturbance models. The disturbance estimates have little effect on the performance under this controller tuning.

8

MAXIMUM LIKELIHOOD ESTIMATION[1]

8.1 FORMING THE MLE PROBLEM

In this chapter, we propose a maximum likelihood estimation method for estimating the process and measurement noise covariances from data. As discussed in greater detail in Section 8.7, this method has a stronger theoretical basis and a simpler derivation compared to the ALS method.

We begin with the state space model

$$x^+ = Ax + w$$

$$y = Cx + v$$

$$\begin{bmatrix} w \\ v \end{bmatrix} \sim N \left(0, \begin{bmatrix} Q_w & 0 \\ 0 & R_v \end{bmatrix} \right)$$

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

---

[1]Portions of this chapter will be published in Zagrobelny and Rawlings (2014b)

the system matrices A and C and a sequence of measurements $y(0), \ldots, y(N-1)$:

$$\max_{Q_w, R_v} \ln p_y(y(0) \ldots y(N-1)|Q_w, R_v)$$

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

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 $x(0)$:

$$
\begin{bmatrix} y(0) \\ y(1) \\ \vdots \\ y(\tilde{N}-1) \end{bmatrix} = \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{\tilde{N}-1} \end{bmatrix} x(0) + \begin{bmatrix} v(0) \\ v(1) \\ \vdots \\ v(\tilde{N}-1) \end{bmatrix}
$$
$$
+ \begin{bmatrix} 0 & 0 & \ldots & 0 \\ C & 0 & \ldots & 0 \\ \ldots & \ddots & & \vdots \\ CA^{\tilde{N}-2} & CA^{\tilde{N}-3} & \ldots & C \end{bmatrix} \begin{bmatrix} w(0) \\ w(1) \\ \vdots \\ w(\tilde{N}-2) \end{bmatrix} \tag{8.1}
$$

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

---

[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.

or later, (8.1) simplifies to

$$
\begin{bmatrix} y(K) \\ y(K+1) \\ \vdots \\ y(\tilde{N}-1) \end{bmatrix} \approx \begin{bmatrix} v(K) \\ v(K+1) \\ \vdots \\ v(\tilde{N}-1) \end{bmatrix}
$$

$$
+ \underbrace{\begin{bmatrix} CA^{K-1} & \cdots & C & 0 & \cdots & 0 \\ 0 & CA^{K-1} & \cdots & C & \cdots & 0 \\ \vdots & \ddots & & & & \vdots \\ 0 & \cdots & 0 & CA^{K-1} & \cdots & C \end{bmatrix}}_{\mathbb{O}} \begin{bmatrix} w(0) \\ w(1) \\ \vdots \\ w(\tilde{N}-2) \end{bmatrix} \tag{8.2}
$$

Since all of the noises are normally distributed, $\begin{bmatrix} y(K)' & \cdots & y(\tilde{N}-1)' \end{bmatrix}'$ also has a normal distribution. As the indices in (8.2) are arbitrary, we have the distribution

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

$$
P = \mathbb{O} \begin{bmatrix} Q_w & & \\ & \ddots & \\ & & Q_w \end{bmatrix} \mathbb{O}' + \begin{bmatrix} R_v & & \\ & \ddots & \\ & & R_v \end{bmatrix} \tag{8.3}
$$

Note that we can also write P as

$$
P = \sum_{i=1}^{N+K-1} \mathbb{O}_i Q_w \mathbb{O}_i' + \sum_{j=1}^{N} \mathbb{I}_j R_v \mathbb{I}_j' \tag{8.4}
$$

in which $\mathbb{O}_i$ is the ith $pN \times n$ block column of $\mathbb{O}$ and $\mathbb{I}_i$ is the ith $pN \times p$ block

column of I_{Np}. Finally, we write the maximum likelihood problem as

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

$$\text{subject to } Q_w, R_v \geq 0 \tag{8.5}$$

in which P is defined in (8.3) and (8.4). Note that $\phi(Q_w, R_v)$ is equal to $-2 \ln p_Y(Y | Q_w, R_v)$ without the constant term.

8.2 EXISTENCE OF THE SOLUTION

We next consider under what conditions a solution to the maximum likelihood optimization problem in (8.5) 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 3. *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

$$\sum_{i=1}^{N} y_i' W \Lambda^{-1} W' y_i = \sum_{i=1}^{N} y_i' \left(\sum_{j=1}^{p} \frac{1}{\lambda_j} w_j w_j' \right) y_i = \sum_j \frac{1}{\lambda_j} \sum_i y_i' w_j w_j' y_i$$

$$= \sum_j \frac{1}{\lambda_j} \sum_i w_j' y_i y_i' w_j = \sum_j \frac{1}{\lambda_j} w_j' \mathbb{Y} \mathbb{Y}' w_j$$

$$= \sum_{j=1}^{p} \frac{1}{\lambda_j} r_j' r_j$$

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

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

We next consider the behavior of f 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 4. *Given* $\mathbb{Y}$ *and* $f(R)$ *as defined in Proposition* 3, *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 that 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 (8.5). 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 3 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 5. *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* $\begin{bmatrix} y_1 & \cdots & y_N \end{bmatrix}$ *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 ith 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 that 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

$$P = P_Q + P_R \qquad P_Q = \mathbb{O} \begin{bmatrix} Q_w & & \\ & \ddots & \\ & & Q_w \end{bmatrix} \mathbb{O}' \qquad P_R = \begin{bmatrix} R_v & & \\ & \ddots & \\ & & R_v \end{bmatrix} \tag{8.6}$$

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 8.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:

$$R_{mr} = R_m + rW_R \text{diag} \left(\begin{bmatrix} 0 & 1 & \frac{1}{2} & \cdots & \frac{1}{m-1} & 0 & \ldots 0 \end{bmatrix} \right) W_R' \qquad (8.7)$$

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 (8.7), R_m is arbitrarily close to R_{mr}. Since ϕ is continuous in Q and R and R_{mr} is continuous in r, then $\phi(Q_r, R_{mr})$ 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, $\phi(Q_w, R_v) \to \infty$. $\square$

Proposition 6. *Given that the assumptions in Proposition 5 are satisfied, a solution exists to the maximum likelihood problem defined in (8.5).*

Proof. As Q_w or $R_v \to \infty$, $P \to \infty$ (see Appendix 8.B) and $\phi \to \infty$ (by Proposition 5). 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 8.B and Proposition 5). 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 (8.5), 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 (8.5) exists. $\square$

Note that Propositions 3-6 rely on the assumption that $\begin{bmatrix} y_1 \ldots y_N \end{bmatrix}$ is full row rank. As shown in Appendix 8.C, this condition is satisfied with probability one when Y is generated from a normal distribution with a positive definite covariance matrix.

8.3 UNIQUENESS OF THE SOLUTION

We find first and second differentials of $\phi(P) = \phi(Q_w, R_v)$ defined in (8.5). Several matrix differentials, which are given in Appendix 8.D, are used in these derivations. From (8.4), we write dP as

$$dP = \sum_{i=1}^{N+K-1} \mathbb{O}_i \, (dQ_w) \, \mathbb{O}_i' + \sum_{i=1}^{N} \mathbb{I}_i \, (dR_v) \, \mathbb{I}_i' \tag{8.8}$$

We then write $d\phi$ as

$$d\phi = \text{tr}\left((dP) \, P^{-1} \left(P - YY' \right) P^{-1} \right) \tag{8.9}$$

Using dP as defined in (8.8), we write $d\phi$ as:

$$d\phi = \text{tr}\left((dQ_w)\sum_i O'_i P^{-1}(P - YY')P^{-1}O_i\right)$$
$$+ \text{tr}\left((dR_v)\sum_i \mathbb{I}'_i P^{-1}(P - YY')P^{-1}\mathbb{I}_i\right)$$

Any solution to (8.5) 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

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

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

We have for the second differential

$$d^2\phi = \text{tr}((dP)\,P^{-1}\,(dP)\,P^{-1})$$
$$- 2\text{tr}\left((dP)\,P^{-1}\,(dP)\,P^{-1}(P - YY')P^{-1}\right) \qquad (8.10)$$

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 (8.10) 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.

8.4 CONNECTION TO THE ALS TECHNIQUE

Here we follow the derivation as in Rajamani (2007). For simplicity, we assume that $L = 0$ in both the MLE and ALS problems.

We rewrite the MLE first order condition in (8.9) as

$$\text{tr}\left(P^{-1}\left(dP\right)\left(I_{pN} - P^{-1}YY'\right)\right) = 0 \tag{8.11}$$

Let $\mathcal{I}_n := \text{vec}(I_n)$. Noting that for any $n \times n$ matrix A, $\text{tr}(A) = \mathcal{I}_n'\text{vec}(A)$, we rewrite (8.11) as

$$\mathcal{I}_{Np}'\left(\left(I_{pN} - YY'P^{-1}\right) \otimes P^{-1}\right)\text{vec}(dP) = 0 \tag{8.12}$$

Note that there was an error in (Rajamani, 2007, p. 128), which is fixed here. Next we write (8.12) in terms of dQ_w and dR_v. Starting with Y in terms of $x(0)$ as in (8.1) 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

$$P = \mathscr{O}P_x\mathscr{O}' + \Gamma_f\bigoplus_{i=1}^{N}Q_w\Gamma_f' + \bigoplus_{i=1}^{N}R_v \tag{8.13}$$

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 (8.13) to obtain

$$\text{vec}(P) = \begin{bmatrix} \mathcal{A}_1 & \mathcal{A}_2 \end{bmatrix} \begin{bmatrix} \text{vec}(Q_w) \\ \text{vec}(R_v) \end{bmatrix} \tag{8.14}$$

$$\mathcal{A}_1 = (\mathcal{O} \otimes \mathcal{O})(I_{n^2} - A \otimes A)^{-1} + (\Gamma_f \otimes \Gamma_f)\mathcal{J}_{n,N}$$

$$\mathcal{A}_2 = \mathcal{J}_{p,N}$$

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 a $m \times m$ 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 $\mathbb{A}_0 = \mathcal{I}'_{Np}\left((I_{pN} - YY'P^{-1}) \otimes P^{-1}\right)$ and using (8.14) in (8.12), we write the first order condition as

$$\mathbb{A}_0 \begin{bmatrix} \mathcal{A}_1 & \mathcal{A}_2 \end{bmatrix} = \begin{bmatrix} 0 & \ldots 0 \end{bmatrix} \tag{8.15}$$

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 (8.15), we write the first order condition for the maximum likelihood problem as

$$0 = \begin{bmatrix} \mathcal{A}'_1 \\ \mathcal{A}'_2 \end{bmatrix} \text{vec}(P^{-1}) - \frac{1}{2}\begin{bmatrix} \mathcal{A}'_1 \\ \mathcal{A}'_2 \end{bmatrix}\left(P^{-1} \otimes P^{-1}\right)\left(I_{(pN)^2} + K_{(pN)(pN)}\right)\text{vec}(YY') \tag{8.16}$$

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

$$\begin{bmatrix} \mathcal{A}'_1 \\ \mathcal{A}'_2 \end{bmatrix} W^\dagger \begin{bmatrix} \mathcal{A}_1 & \mathcal{A}_2 \end{bmatrix} \begin{bmatrix} \text{vec}(Q_w) \\ \text{vec}(R_v) \end{bmatrix} - \begin{bmatrix} \mathcal{A}'_1 \\ \mathcal{A}'_2 \end{bmatrix} W^\dagger \text{vec}(YY') \qquad (8.17)$$

We define W and its psuedoinverse as

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

Using this value of $W^\dagger$ in (8.17) 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 (8.16).

From equation (18) in Rajamani and Rawlings (2009), 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 Rajamani and Rawlings (2009).

2. From (8.14), 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 G is unknown, both the following conditions are necessary for either the ALS or the MLE problem to have a unique solution:

   (a) (A, C) observable

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

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

8.5 SOLVING THE MLE 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 (8.2) and (8.3). 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 System 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 $\backslash$ 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.

8.5.1 *Optimal innovations MLE method*

The MLE method proposed by Bavdekar et al. (2011) 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

$$\min_{Q_w, R_v} N \ln(\det(\Sigma_e)) + \sum_{i=0}^{N-1} (y(k) - \hat{y}(k|k-1))' \Sigma_e^{-1} (y(k) - \hat{y}(k|k-1))$$

subject to: Kalman filter equations

$$Q_w, R_v \geq 0$$

in which Σ_e is the covariance 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 covariance (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].

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 $Y'_{\text{inn}} P^{-1} Y_{\text{inn}}$ is computed more quickly than $Y' P^{-1} Y$ in the output based-method. 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 Q_w and R_v are updated. However, for larger systems, the optimal innovation

---

[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 $(y(k) - \hat{y}(k|k-1))' \Sigma_e^{-1} (y(k) - \hat{y}(k|k-1))$

MLE method significantly reduces the computational time. Both formulations of the MLE problem lead to the same estimates of Q_w and R_v.

8.6 EXAMPLES

8.6.1 *Scalar example*

Consider the example

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

We use $N = 1000$ data points and $K = 23$ (placing a threshold of 10^{-5} on the norm of A). We solve the MLE problem in Octave using the built-in function sqp. We also solve the ALS problem for comparison, in which the optimal weighting is approximated from the data and the window is fixed at $N_{\text{ALS}} = 15$. The results are summarized in Table 8.1 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 8.1 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.

Table 8.1: MLE and ALS results for the scalar example.

$N = 1000$	Q_w	R_v	ϕ	Time (s)
MLE	8.69	2.74	2656.85	2.86
ALS	8.66	2.65	2657.09	1.36
Sample Var.	6.79	3.07	2660.04	

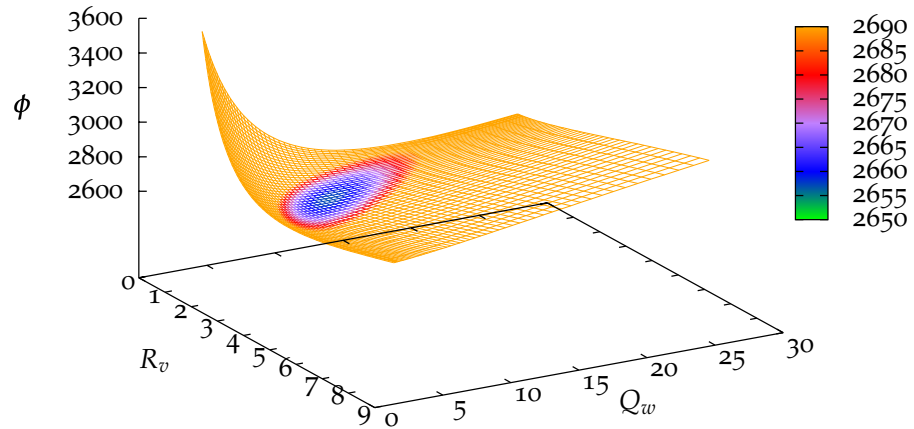| $N = 10000$	Q_w	R_v	ϕ	Time (s)
MLE	6.76	3.07	26277.25	178
ALS	6.64	3.10	26277.39	1.88
Sample Var.	6.93	3.04	26277.52	

Figure 8.1: MLE objective function value vs. Q_w and R_v for the scalar example with $N = 1000$ data points. The objective has a unique minimum and goes to infinity on the boundaries of P.

Table 8.2: Mean and variance of the estimates, average objective value, and average CPU time for the ALS, MLE, and EM methods.

	$\mathbb{E}(\hat{Q}_w)$	$\mathbb{E}(\hat{R}_v)$	$\text{var}(\hat{Q}_w)$	$\text{var}(\hat{R}_v)$	$\langle\phi\rangle$	$\langle\text{CPU Time (s)}\rangle$
ALS	6.699	2.916	1.306	0.113	2617.9279	0.903
MLE	6.944	2.988	1.208	0.105	2617.2499	2.756
EM	6.909	3.002	1.161	0.103	2617.2544	11.965
Sample Var.	7.044	2.969	0.099	0.027	2618.8098	

8.6.2 *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 Li and Badgwell (2014) and Bavdekar et al. (2011). We simulate 50 instances of the problem and calculate Q_w and R_v using all three approaches. Figure 8.2 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 8.2. 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 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 the 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, as it performs Kalman filtering and smoothing at each iteration. However, the EM method may scale better as the amount of data or system dimensions increase. Li and Badgwell (2014) successfully applied the EM method to a larger problem on which the ALS method ran out of memory.
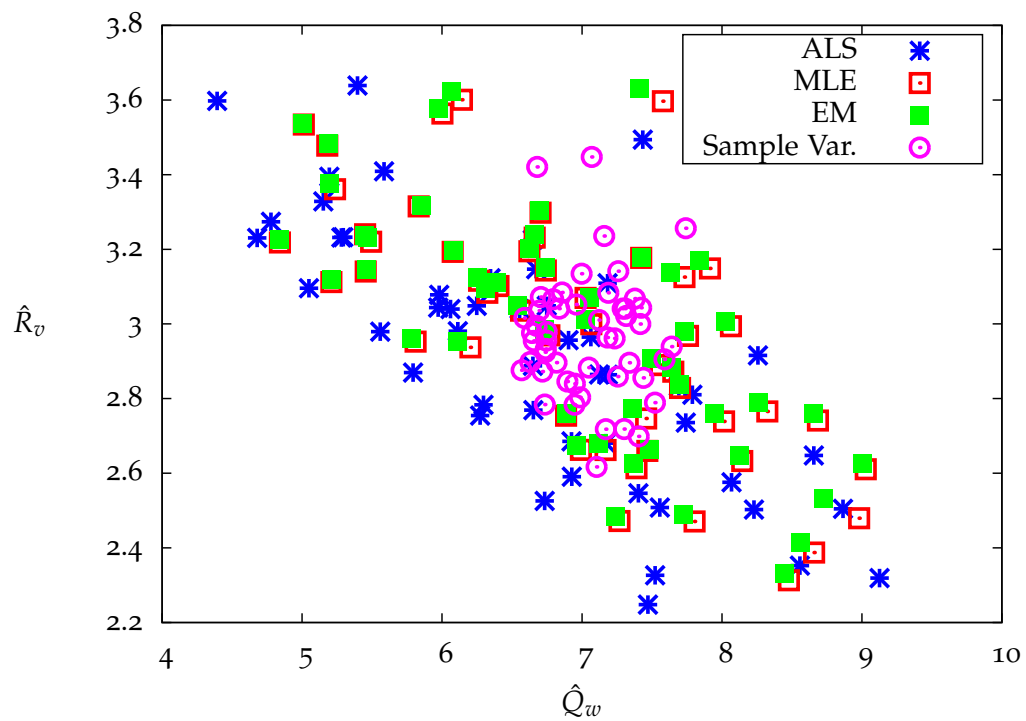
Figure 8.2: Noise variance estimates for the ALS, MLE, and EM methods. The estimates from each method have a similar mean and variance. The EM and MLE methods produce approximately the same results.

Table 8.3: 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

8.6.3 *Example: $p = n = 2$*

In this example, we illustrate how the methods mentioned in Section 8.5, 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 8.3 summarizes the time spent in each step. By using efficient numerical methods, the computational time for each iteration is reduced from approximately 91s to 0.072s.

We also show the computational time for the optimal innovation MLE method proposed by Bavdekar et al. (2011) in Table 8.4. 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 8.3 to Table 8.4, we see that the optimal innovation method requires additional time to compute the innovations but reduces the computational time for the other steps in each iteration.

In Table 8.5 we compare the solutions and solution time of the "slow MLE"

Table 8.4: Computation time for steps in the optimal innovations MLE method (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

Table 8.5: MLE and ALS results for the two-state 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	$\begin{bmatrix} 17.2 & 10.6 \\ 10.6 & 6.47 \end{bmatrix}$	$\begin{bmatrix} 6.82 \\ 4.95 \end{bmatrix}$	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	

(full matrices, eigenvalues, and inverse), "fast MLE" (sparse matrices, Cholesky factorization, and left division), optimal innovations MLE, and ALS techniques for $N = 1000$. All MLE methods give identical results, however, the "fast" MLE and optimal innovations MLE techniques decrease the computation time from several hours to less than a minute. The ALS method gives similar results with the smallest computation time (around 1s) but has a slightly higher objective value.

8.6.4   *Example: $p = n = 5$*

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

$$Q_w = \begin{bmatrix} 8.92 & 9.12 & 14.44 & 5.82 & 12.54 \\ 9.12 & 13.07 & 14.90 & 10.41 & 17.13 \\ 14.44 & 14.90 & 25.11 & 11.32 & 21.50 \\ 5.82 & 10.41 & 11.32 & 11.98 & 14.73 \\ 12.54 & 17.13 & 21.50 & 14.73 & 24.20 \end{bmatrix}$$

$$R_v = \text{diag}\left(\begin{bmatrix} 1.51 & 2.10 & 1.39 & 3.78 & 1.11 \end{bmatrix}\right)$$

We use 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 and 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 solve the MLE problem using the optimal innovations method, starting from each initial condition. For this example, the optimal innovations MLE method significantly reduces the computation time by more than a factor of five and reaches the same solution as the output MLE method.

Table 8.6: MLE and ALS results for the five-state example.

Method	Results

MLE
$Q_0 = I$

$$Q_w = \begin{bmatrix} 8.50 & 8.19 & 9.97 & 2.84 & 10.31 \\ 8.19 & 14.29 & 17.02 & 10.40 & 18.45 \\ 9.97 & 17.02 & 28.58 & 10.92 & 21.76 \\ 2.84 & 10.40 & 10.92 & 11.19 & 14.29 \\ 10.31 & 18.45 & 21.76 & 14.29 & 24.48 \end{bmatrix}$$

$$\text{diag}(R_v) = \begin{bmatrix} 1.53 & 1.90 & 1.58 & 2.99 & 0.801 \end{bmatrix}$$

$$\phi = 15375$$

Time (min) = 34.8

MLE
$Q_0 = \hat{Q}_{ALS}$

$$Q_w = \begin{bmatrix} 8.34 & 8.32 & 10.02 & 2.85 & 10.32 \\ 8.32 & 14.25 & 17.06 & 10.35 & 18.45 \\ 10.02 & 17.06 & 28.60 & 10.86 & 21.75 \\ 2.85 & 10.35 & 10.86 & 11.25 & 14.26 \\ 10.32 & 18.45 & 21.75 & 14.26 & 24.25 \end{bmatrix}$$

$$\text{diag}(R_v) = \begin{bmatrix} 1.52 & 1.91 & 1.57 & 3.00 & 0.822 \end{bmatrix}$$

$$\phi = 15375$$

Time (min) = 20.5

MLE
Optimal
Innovations
$Q_0 = I$

$$Q_w = \begin{bmatrix} 8.36 & 8.31 & 9.99 & 2.81 & 10.29 \\ 8.31 & 14.17 & 17.03 & 10.43 & 18.47 \\ 9.99 & 17.03 & 28.56 & 10.84 & 21.70 \\ 2.81 & 10.43 & 10.84 & 11.14 & 14.19 \\ 10.29 & 18.47 & 21.70 & 14.19 & 24.17 \end{bmatrix}$$

$$\text{diag}(R_v) = \begin{bmatrix} 1.53 & 1.91 & 1.57 & 3.00 & 0.818 \end{bmatrix}$$

$$\phi = 15375$$
Time (min) = 6.75

MLE
Optimal
Innovations
$Q_0 = \hat{Q}_{ALS}$

$$Q_w = \begin{bmatrix} 8.36 & 8.31 & 9.99 & 2.81 & 10.29 \\ 8.31 & 14.18 & 17.04 & 10.43 & 18.47 \\ 9.99 & 17.04 & 28.56 & 10.84 & 21.70 \\ 2.81 & 10.43 & 10.84 & 11.14 & 14.19 \\ 10.29 & 18.47 & 21.70 & 14.19 & 24.18 \end{bmatrix}$$

$$\text{diag}(R_v) = \begin{bmatrix} 1.53 & 1.90 & 1.57 & 3.00 & 0.817 \end{bmatrix}$$

$$\phi = 15375$$
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

$$Q_w = \begin{bmatrix} 8.32 & 8.48 & 13.59 & 5.50 & 11.70 \\ 8.48 & 12.51 & 13.97 & 10.21 & 16.32 \\ 13.59 & 13.97 & 23.82 & 10.70 & 20.16 \\ 5.50 & 10.21 & 10.69 & 11.85 & 14.21 \\ 11.70 & 16.32 & 20.16 & 14.21 & 22.92 \end{bmatrix}$$

Sample Covariances

$$\text{diag}(R_v) = \begin{bmatrix} 1.40 & 2.05 & 1.34 & 3.88 & 1.07 \end{bmatrix}$$

$$\phi = 15387$$

8.7 COMPARISON OF THE MLE AND ALS APPROACHES

The maximum likelihood approach to disturbance identification is preferable to the ALS method in several ways. From a theoretical point of view, the MLE estimates have a clear statistical meaning, whereas the ALS estimates do not. The ALS problem becomes increasingly arbitrary as we approximate the optimal weighting and choose a penalty to place on $\text{tr}(Q_w)$. In addition, the derivation of the MLE problem is more simple and straightforward.

However, from a practical point of view, the ALS technique is much better equipped to handle industrial data than is the MLE technique. Numerical methods have not been developed to solve the MLE problem on large systems, and the computational time increases significantly when more data is added. The MLE method also has not been developed for the case when the solution is not unique. In this case, some effort must be made to ensure a realistic solution is reached, for example, using the ALS solution as the starting point. Overall, solving the MLE problem requires more computational time and memory than does solving the ALS problem, even when the semidefinite constraints, trace penalty, and approximate weighting are added to the ALS objective. Whereas we have demonstrated the MLE approach only on small simulated systems, we have successfully applied the ALS method to industrial data sets, using the improvements discussed in this

dissertation. Therefore, we recommend using ALS on industrial data sets or on any large systems until the MLE method has been improved. However, given the power of modern computers and the efficiency of many algorithms, the MLE approach has the potential to become useful in practice. In these results we have already shown how choosing efficient computational methods can decrease the time significantly; with more effort this method can be made increasingly more efficient.

8.8 APPENDICES

8.A *Null space of P_Q*

Proposition 7. *Given P_Q and P_R as defined in* (8.6), *then either (1) null$(P_R) \in$ null(P_Q) and null$(P) =$ null(P_R), or (2) null$(P_R) \in$ range(P_Q) and null$(P) = \{0\}$.*

Proof. To prove the proposition, first we show that $V \in$ null(P_R) is 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 $\alpha_1 \ldots \alpha_N$ are scalars. α_i may be zero, but at least one α_i must be non-zero.

If $V \in$ null(P_Q), we must have $X = O'V$ 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 $X = \begin{bmatrix} x'_1 & \cdots & x'_{N+K-1} \end{bmatrix}'$. 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 $j = 1, \ldots, m-1$ 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 $Q_w(A')^i C'$

for $0 \leq i < K$, or else the null space of P contains only the zero element.

$\square$

8.b   *Relationship between (Q_w, R_v) and P on the boundary*

Proposition 8. *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 (8.6), $P = P_Q + P_R$. 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 its structure, $\mathbb{O}$ is full row rank provided C is full row rank. To see that $\mathbb{O}$ is full row rank, we shot 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 (8.4),

$$x'Px = \sum_i x'\mathbb{O}_i Q_w \mathbb{O}_i' x + \sum_j x'\mathbb{I}_j R_v \mathbb{I}_j' x \tag{8.18}$$

Let α_k be (one of) the eigenvalues of Q_w that goes to infinity and v_k be the corresponding normalized eigenvector. Then $v_k' Q_w v_k = \alpha_k \to \infty$. Since $\mathbb{O}_K$ contains the observability matrix when $N \geq n$, it is full rank when (A, C) is ob-

servable. 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 (8.18) 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 (8.18) 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. $\qquad\square$

8.c Rank of the data matrix

Proposition 9 (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, $i = 1, 2, \ldots, N$ 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 $\mathrm{rank}(\mathbb{Y}) = p$ 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 Y is zero. We note that there is probability zero of achieving this matrix by sampling y. In order to zero the pth 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 Y. For positive definite R, however, the probability of a zero row is zero.

To prove the proposition, we consider the (reduced) SVD of 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 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 Y has rank less than p, we know that its smallest singular value, σ_p, equals zero. Therefore the last row of $\mathbb{Z}$ is zero, and, combined with $\tilde{R}$ being positive definite, that is a contradiction and the proposition is established. $\qquad\square$

In Propositions 5 and 6, the samples of y_i are not independent, so Proposition 9 does not directly apply. However, by the same logic, $\mathbb{Y}$ has full rank when $Y = \begin{bmatrix} y'_1 & \cdots & y'_N \end{bmatrix}'$ is generated from a normal distribution with mean zero and covariance $P^* > 0$.

8.d *Matrix differentials*

The following matrix differentials come from Magnus and Neudecker (1999):

$$d(\det(X)) = \det(X)\mathrm{tr}\left(X^{-1}(dX)\right) \qquad X \in \mathbb{R}^{n\times n}, \text{invertible}$$

$$d(\mathrm{tr}(AX)) = \mathrm{tr}\left(A(dX)\right) \qquad\qquad X \text{ real}$$

$$dX^{-1} = -X^{-1}(dX)X^{-1} \qquad\qquad X \in \mathbb{R}^{n\times n}, \text{invertible}$$

9

CONCLUSIONS

9.1 CONTRIBUTIONS

MPC PERFORMANCE MONITORING We approach MPC performance monitoring by using to be the expectation of the regulator's stage cost as our monitoring benchmark. We show that the stage cost is the quadratic form of a normal variable and therefore has a generalized chi-squared distribution. We derive the expectation of the stage cost, as well as its variance, from the analytical formulas for the closed-loop system. This expectation serves as the ideal benchmark. As a time-average of the stage cost, the plant KPI approaches a normal distribution by the central limit theorem. We derive an expression for the variance of the plant KPI that accounts for the correlations between the stage cost samples. We illustrate the distributions of the stage cost and plant KPI with a simulation example. We further extend the analytical KPI formulas to account for a general deterministic disturbance and for systems in innovation form.

AUTOCOVARIANCE LEAST-SQUARES We present two major improvements to autocovariance least-squares: reducing the model to include only the fully observable states and approximating the optimal weighting from data. We use the

singular value decomposition of the observability matrix to identify the least observable states and then use the unconstrained ALS objective function value to choose the model order for the reduced system. By using this reduced model, we significantly improve the computational time for ALS. Using the optimal weighting for the least-squares portion of the ALS objective would reduce the variance of the ALS estimates, but the optimal weighting cannot be computed in practice. Therefore, we present a feasible-generalized ALS method in which this weighting is estimated from data. The feasible-generalized ALS method leads to significant reduction in the variance of the ALS estimates for a simulation example.

We apply the improved ALS method to industrial data. We significantly reduce the model to have a well-conditioned observability matrix. The reduced model has a faster computational time than a larger model and still results in an estimator that is approximately optimal. We use a double integrator model to estimate better the disturbances affecting one of the outputs. The feasible generalized ALS method shows significant improvement over ordinary ALS. The disturbances affecting the process remain largely consistent over a year-long period, and, through the results of ALS, we identify when changes in the disturbance dynamics occur.

MAXIMUM LIKELIHOOD ESTIMATION   We propose an alternative method for identifying the disturbance model using maximum likelihood estimation. By assuming the deterministic system is known, we present a direct maximum likelihood estimation problem in which the noise covariances are the only parameters being estimated. We prove that a solution to the MLE problem exists, since the data matrix is full rank with probability one. We also relate the maximum likelihood problem to the optimally-weighted ALS problem and thus relate the uniqueness conditions of the two methods. We reduce the computational time by using sparse matrices, computing the log determinant via Cholesky decomposition, and avoiding the direct computation of the matrix inverse. We apply the MLE method

to several small-scale examples.

9.2 FUTURE WORK

Due to the advances presented in this dissertation, several areas lie open for future research in MPC performance monitoring and disturbance identification:

MPC PERFORMANCE MONITORING   Since the theoretical monitoring benchmark is highly dependent upon the noise covariances, the improved ALS method should be applied to MPC monitoring. For both simulated and industrial systems, the noise covariances should be estimated using the improved ALS techniques, and the theoretical benchmark should be calculated using the ALS estimates.

In order to implement a monitoring scheme in practice, we require a statistical test for whether or not the plant KPI is significantly above its ideal benchmark. Therefore, the theoretical distribution of the plant KPI (which we derived) should be used to develop confidence intervals. An appropriate window size for calculating the KPI should be determined. The window should be large enough that the plant KPI is well-approximated by a normal distribution and is not susceptible to outliers, but small enough that the plant KPI responds quickly to changes in the controller performance. Once an appropriate window has been developed, a test should be designed to continuously monitor the controller performance and alert operators when a significant change in performance has occurred.

For linear systems, the analytical KPI formula should be extended to include feed-forward variables. For nonlinear and constrained systems, simulations should be performed to demonstrate the accuracy of Monte Carlo methods in estimating the optimal benchmark; this technique should be verified on industrial data.

When calculating the KPI for industrial data, the precise control law may be unknown. This problem is especially likely when industries rely on vendor-supplied software. This lack of knowledge inhibits the calculation of a theoretical

KPI. Therefore, another area of research is to identify the control law from data, *i.e.* fit a $y \rightarrow u$ model. The KPI calculations can then be modified to incorporate the estimated control law.

AUTOCOVARIANCE LEAST-SQUARES   While the ALS technique (with our improvements) has been successfully applied to industrial data, this method needs to be further developed in order to become more automated. Currently, the ALS method requires the user to specify many decisions: initial estimator gain, model size, trace penalty, *etc.* Thus the user must study each system in detail, rather than easily applying the ALS method to a large number of industrial systems. In addition, ALS techniques on non-linear and time varying systems should be further developed.

MAXIMUM LIKELIHOOD ESTIMATION   The computational algorithms for the MLE method require significant development before this method will be widely applicable to industrial data. These improvements may be as simple as using different software with more efficient methods to solve the optimization problem, or we may need to find more efficient algorithms to compute the likelihood. As these computational methods are developed, we should always maintain the goal of producing software that is simple to implement in practice.

The maximum likelihood method presented here was based on the open-loop problem, *i.e.* it was written in terms of the outputs rather than innovations. This MLE method should be compared in more detail to the MLE method based on the optimal innovations, which is faster to solve for larger systems. In particular, further research should demonstrate that a solution to the optimal innovations MLE problem exists and that the same necessary conditions for uniqueness apply. Studying a third MLE formulation written in terms of L-innovations that are fixed during the optimization may help connect the first two MLE methods.

We restricted our study here to problems with the same number of states and

outputs, as this condition is necessary for uniqueness. The MLE method should expanded to handle systems that do not have a unique solution. It is of note that since the problem is solved in terms of the factors of Q and R, this formulation naturally leads itself to restricting the rank of these noise matrices. Some study should be made as to how to choose the best initial guess and how sensitive the solution is to the initial guess. Once these problems have been addressed, other issues such as applicability to non-linear systems should be studied.

BIBLIOGRAPHY

Q. Ahsan, R. I. Grosvenor, and P. W. Prickett. Distributed control loop performance monitoring architecture. In *Proceedings of Control 2004*, 2004.

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

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

J. D. Aplevich. *The Essentials of Linear State-Space Systems*. John Wiley & Sons, Inc., New York, 2000.

K. J. Åström. *Introduction to Stochastic Control Theory*. Academic Press, San Diego, California, 1970.

K. J. Åström and B. Wittenmark. *Computer Controlled Systems: Theory and Design*. Prentice-Hall, Englewood Cliffs, New Jersey, 1984.

O. Badmus, D. Banks, A. Vishnubhotla, B. Huang, and S. L. Shah. Performance assessment: a requisite for maintaining your APC assets. In *Dynamic Modeling Control Applications for Industry Workshop, 1998. IEEE Industry Applications 1998*, pages 54 –58, 1998.

A. S. Badwe, R. D. Gudi, R. S. Patwardhan, S. L. Shah, and S. C. Patwardhan. Detection of model-plant mismatch in MPC applications. *J. Proc. Cont.*, 19(8): 1305–1313, 2009.

A. S. Badwe, R. S. Patwardhan, S. L. Shah, S. C. Patwardhan, and R. D. Gudi. Quantifying the impact of model-plant mismatch on controller performance. *J. Proc. Cont.*, 20(4):408–425, 2010.

M. Bauer and I. K. Craig. Economic assessment of advanced process control–A survey and framework. *J. Proc. Cont.*, 18(1):2–18, 2008.

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.

L. G. Bergh and J. F. MacGregor. Constrained minimum variance controllers: internal model structure and robustness properties. *Ind. Eng. Chem. Res.*, 26: 1558–1564, 1987.

S. Bezergianni and C. Georgakis. Controller performance assessment based on minimum and open-loop output variance. *Control Eng. Pract.*, 8(7):791–797, 2000.

S. Bezergianni and C. Georgakis. Evaluation of controller performance - use of models derived by subspace identification. *Int. J. Adaptive Cont. Signal Proc.*, 17 (7-9):527–552, 2003.

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

N. Bonavita, J. C. Bovero, and R. Martini. Control loops: Performance and diagnostics. In *Proceedings of the ANIPLA Conference*, 2004.

T. Cacoullos and M. Koutras. Quadratic forms in spherical random variables: Generalized noncentral χ^2 distribution. *Naval Res. Logist. Quart.*, 31(3):447–461, 1984.

F. M. Callier and C. A. Desoer. *Linear System Theory*. Springer-Verlag, New York, 1991.

C.-T. Chen. *Linear System Theory and Design*. Oxford University Press, third edition, 1999.

C. Dai and S. H. Yang. Controller performance assessment with a LQG benchmark obtained by using the subspace method. In *Proceedings of Control*, 2004.

L. Desborough and T. Harris. Performance assessment measures for univariate feedback control. *Can. J. Chem. Eng.*, 70:1186–1197, 1992.

L. Desborough and T. Harris. Performance assessment measures for univariate feedforward feedback control. *Can. J. Chem. Eng.*, 71:605–616, 1993.

L. Desborough and R. Miller. Increasing customer value of industrial control performance monitoring - Honeywell's experience. In *Chemical Process Control–VI: Sixth International Conference on Chemical Process Control*, pages 153–186. AIChE Symposium Series, Volume 98, Number 326, 2001.

G. A. Dumont, L. Kammer, B. J. Allison, L. Ettaleb, and A. A. Roche. Control performance monitoring: New developments and practical issues. In *Proccedings of the IFAC World Congress*, 2002.

P.-G. Eriksson and A. J. Isaksson. Some aspects of control loop performance monitoring. In *Control Applications, 1994., Proceedings of the Third IEEE Conference on*, volume 2, pages 1029 –1034, Aug. 1994.

J. Gao, R. Patwardhan, K. Akamatsu, Y. Hashimoto, G. Emoto, S. L. Shah, and B. Huang. Performance evaluation of two industrial MPC controllers. *Control Eng. Pract.*, 11(12):1371–1387, 2003.

C. E. García, D. M. Prett, and M. Morari. Model predictive control: Theory and practice—a survey. *Automatica*, 25(3):335–348, 1989.

Z. Ge, C. Yang, Z. Song, and H. Wang. Robust online monitoring for multimode

processes based on nonlinear external analysis. *Ind. Eng. Chem. Res.*, 47(14): 4775–4783, 2008.

M. Ghosh and B. K. Sinha. A simple derivation of the Wishart distribution. *Amer. Statist.*, 56(2):100–101, 2002.

M. J. Grimble. Controller performance benchmarking and tuning using generalised minimum variance control. *Automatica*, 38(12):2111–2119, 2002.

G. Haarsma and M. Nikolaou. Multivariate controller performance monitoring: Lessons from an application to snack food processes, 2000. URL `http://www.chee.uh.edu/faculty/nikolaou/FryerMonitoring.pdf`.

T. J. Harris. Assessment of control loop performance. *Can. J. Chem. Eng.*, 67:856–861, 1989.

T. J. Harris. Statistical properties of quadratic-type performance indices. *J. Proc. Cont.*, 14(8):899 – 914, 2004.

T. J. Harris, F. Boudreau, and J. F. MacGregor. Performance assessment of multivariate feedback controllers. *Automatica*, 32(11):1505–1518, 1996a.

T. J. Harris, C. T. Seppala, P. J. Jofriet, and B. W. Surgenor. Plant-wide feedback control performance assessment using an expert-system framework. *Control Eng. Pract.*, 4(9):1297–1303, 1996b.

T. J. Harris, C. T. Seppala, and L. D. Desborough. A review of performance monitoring and assessment techniques for univariate and multivariate control systems. *J. Proc. Cont.*, 9(1):1–17, 1999.

C. A. Harrison and S. J. Qin. Minimum variance performance map for constrained model predictive control. *J. Proc. Cont.*, 19(7):1199–1204, 2009.

G. Hewer. Analysis of a discrete matrix Riccati equation of linear control and Kalman filtering. *J. Math. Anal. Appl.*, 42(1):226–236, 1973.

G. A. Hewer. Iterative technique for computation of steady state gains for discrete optimal regulator. *IEEE Trans. Auto. Cont.*, AC16(4):382–384, 1971.

A. Horch and A. J. Isaksson. A modified index for control performance assessment. *J. Proc. Cont.*, 9(6):475–483, 1999.

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

B. Huang. Minimum variance control and performance assessment of time-variant processes. *J. Proc. Cont.*, 12(6):707–719, 2002.

B. Huang and S. L. Shah. Practical issues in multivariable feedback control performance assessment. *J. Proc. Cont.*, 8(5–6):421–430, 1998.

B. Huang and S. L. Shah. *Performance Assessment of Control Loops*. Springer-Verlag, London, 1999.

B. Huang and E. C. Tamayo. Model validation for industrial model predictive control systems. *Chem. Eng. Sci.*, 55(12):2315–2327, 2000.

B. Huang, S. L. Shah, and E. K. Kwok. Good, bad or optimal? Performance assessment of multivariable processes. *Automatica*, 33(6):1175–1183, 1997.

B. Huang, R. Kadali, X. Zhao, E. C. Tamayo, and A. Hanafi. An investigation into the poor performance of a model predictive control system on an industrial CGO coker. *Control Eng. Pract.*, 8(6):619–631, 2000.

B. Huang, S. X. Ding, and N. Thornhill. Practical solutions to multivariate feedback control performance assessment problem: reduced a priori knowledge of interactor matrices. *J. Proc. Cont.*, 15(5):573–583, 2005.

B. Huang, S. X. Ding, and N. Thornhill. Alternative solutions to multi-variate control performance assessment problems. *J. Proc. Cont.*, 16(5):457–471, 2006.

A. J. Hugo. Performance assessment of single-loop industrial controllers. *J. Proc. Cont.*, 16(8):785–794, 2006.

A. Ingimundarson and T. Hagglund. Closed-loop performance monitoring using loop tuning. *J. Proc. Cont.*, 15(2):127–133, 2005.

R. H. Julien, M. W. Foley, and W. R. Cluett. Performance assessment using a model predictive control benchmark. *J. Proc. Cont.*, 14(4):441–456, 2004.

R. Kadali and B. Huang. Estimation of the dynamic matrix and noise model for model predictive control using closed-loop data. *Ind. Eng. Chem. Res.*, 41:842–852, 2002.

M. Kamrunnahar, D. G. Fisher, and B. Huang. Performance assessment and robustness analysis using an ARMarkov approach. *J. Proc. Cont.*, 14(8):915–925, 2004.

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

P. Kesavan and J. H. Lee. Diagnostic tools for multivariable model-based control systems. *Ind. Eng. Chem. Res.*, 36(7):2725–2738, 1997.

B. Ko and T. F. Edgar. Performance assessment of constrained model predictive control systems. *Proc. Sys. Eng.*, 47(6):1363–1371, 2001.

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

R. H. Kwong. On the linear quadratic Gaussian problem with correlated noise and its relation to minimum variance control. *SIAM J. Cont. Opt.*, 29(1):139–152, 1991.

P. Lancaster and M. Tismenetsky. *The Theory of Matrices: With Applications*. Computer Science and Applied Mathematics. Academic Press, 1985.

K. H. Lee, E. C. Tamayo, and B. Huang. Industrial implementation of controller performance analysis technology. *Control Eng. Pract.*, 18(2, SI):147–158, 2010.

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

F. V. Lima, J. B. Rawlings, M. R. Rajamani, and T. A. Soderstrom. Covariance and state estimation of weakly observable systems: Application to polymerization processes. *IEEE Trans. Cont. Sys. Tech.*, 21(4):1249–1257, July 2013.

C. Liu, B. Huang, and Q. Wang. Control performance assessment subject to multi-objective user-specified performance characteristics. *IEEE Trans. Cont. Sys. Tech.*, 19(3):682–691, 2011.

L. Ljung. *System Identification - Theory for the User*. Prentice Hall, New Jersey, second edition, 1999.

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.

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

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

A. Mathai and S. B. Provost. *Quadratic Forms in Random Variables: Theory and Applications*. Marcel Dekker, Inc., 1992.

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.

C. A. McNabb and S. J. Qin. Projection based MIMO control performance monitoring: I-covariance monitoring in state space. *J. Proc. Cont.*, 13(8):739–757, 2003.

C. A. McNabb and S. J. Qin. Projection based MIMO control performance monitoring: II-measured disturbances and setpoint changes. *J. Proc. Cont.*, 15(1): 89–102, 2005.

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

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

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

K. A. Myers and B. D. Tapley. Adaptive sequential estimation with unknown noise statistics. *IEEE Trans. Auto. Cont.*, 21:520–523, 1976.

B. J. Odelson. *Estimating Disturbance Covariances From Data for Improved Control Performance*. PhD thesis, University of Wisconsin–Madison, 2003.

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.

F. B. Olaleye, B. Huang, and E. Tamayo. Feedforward and feedback controller performance assessment of linear time-variant processes. *Ind. Eng. Chem. Res.*, 43:589–596, 2004.

G. Pannocchia and J. B. Rawlings. Disturbance models for offset-free MPC control. *AIChE J.*, 49(2):426–437, 2003.

R. S. Patwardhan. *Studies in Synthesis and Analysis of Model Predictive Controllers*. PhD thesis, University of Alberta, Fall 1999. URL http://www.collectionscanada.ca/obj/s4/f2/dsk1/tape8/PQDD_0016/NQ46902.pdf.

R. S. Patwardhan and S. L. Shah. Issues in performance diagnostics of model-based controllers. *J. Proc. Cont.*, 12(3):413–427, 2002.

M. A. Paulonis and J. W. Cox. A practical approach for large-scale controller performance assessment, diagnosis, and improvement. *J. Proc. Cont.*, 13(2):155–168, 2003.

S. J. Qin. Controller performance monitoring – a review and assessment. *Comput. Chem. Eng.*, 23:173–186, 1998.

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

S. J. Qin and T. A. Badgwell. A survey of industrial model predictive control technology. *Control Eng. Pract.*, 11(7):733–764, 2003.

S. J. Qin and J. Yu. Recent developments in multivariable controller performance monitoring. *J. Proc. Cont.*, 17(3):221–227, 2007.

M. R. Rajamani. *Data-based Techniques to Improve State Estimation in Model Predictive Control*. PhD thesis, University of Wisconsin–Madison, October 2007. URL http://jbrwww.che.wisc.edu/theses/rajamani.pdf.

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

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

M. R. Rajamani, J. B. Rawlings, and S. J. Qin. Achieving state estimation equivalence for misassigned disturbances in offset-free model predictive control. *AIChE J.*, 55(2):396–407, February 2009.

T. J. Rato and M. S. Reis. Statistical monitoring of control loops performance: An improved historical-data benchmark index. *Qual. Reliab. Eng. Int.*, 26(8):831–844, 2010.

J. B. Rawlings and D. Q. Mayne. *Model Predictive Control: Theory and Design*. Nob Hill Publishing, Madison, WI, 2009. 576 pages, ISBN 978-0-9759377-0-9.

T. I. Salsbury. A practical method for assessing the performance of control loops subject to random load changes. *J. Proc. Cont.*, 15(4):393–405, 2005.

J. Schafer and A. Cinar. Multivariable MPC system performance assessment, monitoring, and diagnosis. *J. Proc. Cont.*, 14(2):113–129, 2004.

P. Schmidt. *Econometrics*. Marcel Dekker, Inc, 1976.

S. R. Searle. *Linear Models*. John Wiley & Sons, Inc., 1971.

S. L. Shah, R. Patwardhan, and B. Huang. Multivariable controller performance analysis: methods, applications and challenges. *Chem. Proc. Cont.*, 98:190–207, 2002.

Y. Shardt, Y. Zhao, F. Qi, K. Lee, X. Yu, B. Huang, and S. Shah. Determining the state of a process control system: Current trends and future challenges. *Can. J. Chem. Eng.*, 2011.

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.

Z. Sun, S. J. Qin, A. Singhal, and L. Megan. Performance monitoring of model-predictive controllers via model residual assessment. *J. Proc. Cont.*, 23(4):473 – 482, 2013.

N. F. Thornhill, M. Oettinger, and P. Fedenczuk. Refinery-wide control loop performance assessment. *J. Proc. Cont.*, 9(2):109–124, 1999.

X. Tian, G. Chen, and S. Chen. A data-based approach for multivariate model predictive control performance monitoring. *Neurocomputing*, 74(4):588 − 597, 2011.

K. Tsakalis and S. Dash. Multivariable controller performance monitoring using robust stability conditions. *J. Proc. Cont.*, 17(9):702–714, 2007.

M. L. Tyler and M. Morari. Performance monitoring of control systems using likelihood methods. *Automatica*, 32(8):1145–1162, 1996.

H. K. Tzou and Y. T. Lin. The tracking of a manoeuvring object by using an adaptive Kalman filter. *Proceedings of the Institute of Mechanical Engineers*, 215(2): 125–130, 2001.

J. Valappil and C. Georgakis. Systematic estimation of state noise statistics for extended Kalman filters. *AIChE J.*, 46(2):292–308, 2000.

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

X. Wang, B. Huang, and T. Chen. Multirate minimum variance control design and control performance assessment: a data-driven subspace approach. *IEEE Trans. Cont. Sys. Tech.*, 15(1):65–74, 2007.

L. Xie, U. Kruger, D. Lieftucht, T. Littler, Q. Chen, and S. Q. Wang. Statistical monitoring of dynamic multivariate processes-Part 1. Modeling autocorrelation and cross-correlation. *Ind. Eng. Chem. Res.*, 45(5):1659–1676, 2006.

F. Xu, K. H. Lee, and B. Huang. Monitoring control performance via sructured closed-loop response subject to output variance/covariance upper bound. *J. Proc. Cont.*, 16(9):971–984, 2006.

F. Xu, B. Huang, and S. Akande. Performance assessment of model pedictive control for variability and constraint tuning. *Ind. Eng. Chem. Res.*, 46(4):1208–1219, 2007.

F. Xu, B. Huang, and E. C. Tamayo. Performance assessment of MIMO control systems with time-variant disturbance dynamics. *Comput. Chem. Eng.*, 32(9): 2144–2154, 2008.

J. Yu and S. J. Qin. Statistical MIMO controller performance monitoring. Part I: Data-driven covariance benchmark. *J. Proc. Cont.*, 18(3–4):277–296, 2008a.

J. Yu and S. J. Qin. Statistical MIMO controller performance monitoring. Part II: performance diagnosis. *J. Proc. Cont.*, 18(3–4):297–319, 2008b.

J. Yu and S. J. Qin. MIMO control performance monitoring using left/right diagonal interactors. *J. Proc. Cont.*, 19(8):1267–1276, 2009.

Q. Yuan and B. Lennox. Control performance assessment for multivariable systems based on a modified relative variance technique. *J. Proc. Cont.*, 19(3):489–497, 2009.

Q. Yuan, B. Lennox, and M. McEwan. Analysis of multivariable control perfomance assessment techniques. *J. Proc. Cont.*, 19(5):751–760, 2009.

M. A. Zagrobelny and J. B. Rawlings. Practical improvements to autocovariance least-squares. In Preparation, 2014a.

M. A. Zagrobelny and J. B. Rawlings. Identification of disturbance covariances using maximum likelihood estimation. In Preparation, 2014b.

M. A. Zagrobelny, L. Ji, and J. B. Rawlings. Quis custodiet ipsos custodes? *Annual Rev. Control*, 37:260–270, 2013.

Y. Zhao, J. Chu, H. Su, and B. Huang. Multi-step prediction error approach for controller performance monitoring. *Control Eng. Pract.*, 18(1):1–12, 2010.