MANAGING THE EFFECTS OF INPUT EXCITATION ON MODEL-BASED FAULT DETECTION

by

Renee Marie Spinhirne

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 – 1346

A thesis submitted to the Faculty and the Board of Trustees of the Colorado School of Mines in partial fulfillment of the requirements for the degree of Doctor of Philosophy (Engineering Systems).

Golden, Colorado
Date 04/09/2007

Signed: Renee M Spinhirne
Renee Marie Spinhirne

Approved:
Dr. Tyrone L. Vincent
Associate Professor of Engineering
Thesis Advisor

Golden, Colorado
Date 04/09/2007

Dr. Terry Parker
Professor and Director
Division of Engineering

ABSTRACT

With the complexity and expense involved in industrial systems today, detecting failures or faults in the system and identifying these faults has become an important topic. Applying fault detection and identification (FDI) methods to complex systems can be both a time and cost saving measure. Because of the importance of an accurate system model for use in model-based FDI methods, system identification methods have also become an important aspect of the complete FDI system. This thesis investigates two topics related to the effect of input excitation on system identification for model-based fault detection. First, the topic of fault detection in the absence of sufficient input excitation is considered. For the case considered here the input can not be chosen arbitrarily during system identification. A method is introduced for residual generation based on subspace identification. Secondly, the case where the input can be chosen arbitrarily within system constraints is considered. This topic introduces a method of input design for system identification focusing on parameter identification and parameter variation fault detection.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ARX	AutoRegressive with eXtra input model
DPCA	Dynamic Principal Components Analysis
FD	Fault Detection
FDA	Fisher Discriminant Analysis
FDI	Fault Detection and Identification
FIR	Finite Impulse Response
MIMO	Multiple Input, Multiple Output
PCA	Principal Components Analysis
PLS	Partial Least Squares
PRBS	Pseudo Random Binary Sequence
SI	System Identification
SISO	Single Input, Single Output
SVD	Singular Value Decomposition
TE	Tennessee Eastman

ACKNOWLEDGMENTS

As this thesis has been a long process, there are many people to whom I must express my gratitude. First, I would like to thank Dr. Chris Debrunner, Dr. Mark Lusk, Dr. Dinesh Mehta, Dr. Kevin Moore, and Dr. Yaoguo Li for serving as my committee members. Their comments and suggestions have added to this thesis in valuable ways. And I would like to thank them for not making the defense process more nerve-wracking than necessary.

I would also like to thank my thesis advisor, Dr. Tyrone Vincent, for many years of instruction, support, advice, and endless tutorial sessions when I needed a little help with a particular idea. Without his support, and nudging, none of this work would have come to completion.

To the guys in lab 154 who made my first couple of years in the Engineering Department bearable. And to all my friends who helped me to remember to have a life during this process and have repeatedly helped me to blow off some steam when necessary. I owe you all for keeping me sane.

I owe a great deal of thanks to my parents, my sister, and the rest of my family for their support, love, and laughter over my many years in school. To my Dad for teaching me I could do anything the boys could do when I was in elementary school.

To my Mom for always reminding me that I would be loved no matter what. And to my sister for the constant friendship and the hours of chatting on the phone. They always stood by me and managed not to ask too many times over the last few years when I was going to manage to finally graduate. Without them, this thesis would have been impossible.

And lastly, I would like to thank my fiance for believing in me, even when I didn't believe in myself. And for reminding me to smile.

To all of those that said,

"You're getting a PhD in Engineering? Girls do that?"

Chapter 1

INTRODUCTION

As industrial systems become more complex it is important to have accurate methods for detecting and isolating failures or faults in these systems. Failures in any system can create costs due to down time and defective products. In complex systems, there are too many actions and sensors to be monitored manually. Therefore, it is necessary to have an automated monitoring system to alert the operator when a failure has occurred and to identify the location of this failure. Much of the Fault Detection and Identification (FDI) literature for dynamic systems has dealt with methods that require an analytical model, or "model-based" FDI. However in industrial systems it is usually too expensive or too complicated to develop a first principles model. Thus, System Identification (SI) is truly the first step in the FDI process.

When using model-based FDI methods, the accuracy of the SI method becomes very important to the success of any FDI algorithm. Modeling errors introduced by the SI method can appear as process faults to the FDI algorithm, leading to false alarms. Thus the effectiveness of the FDI algorithm depends, in part, on the accuracy of the SI method implemented. In many situations it may be more effective to treat

these separate actions as a combined effort of system monitoring to reap the most benefit.

To this end, this thesis will focus on two aspects of system identification for model-based FDI. Both concern the effect of input excitation on the fault detection process. However they differ in their approach. First, a method will be presented for handling fault detection in the absence of sufficient input excitation during the system identification stage. When the input to the system can not be chosen arbitrarily, system identification must be performed using production data. This data may not include a complete picture of the dynamics of the system if the input was not sufficiently exciting. It will be shown that when using a Dynamic Principal Component Analysis approach to system identification, if the correlation structure of the system identification input differs from that of the input during fault detection, the residual will scale with the size of the input instead of the level of noise in the system. This does not allow the expected noise level to be used as a bound on nominal residuals. In this case the resulting residual will indicate a fault has occurred based on modeling errors instead of actual failures in the system. However, this issue can be dealt with by adjusting the residual generation methods during fault detection. This thesis suggests a method of partial system identification during the residual generation process where the error resulting from the unmodeled dynamics is filtered from the residual resulting in a residual that scales with the level of noise in the system during nominal

operation. Thus the expect level of noise in the system can again be used as a bound on the nominal residual and a residual which grown above this bound will indicate a fault in the system. It is shown that this residual filtering allows for improved fault detection in the case where the input excitation varies between the data used for system identification and data monitored during the fault detection process.

In contrast, the second situation investigated occurs when the input can be chosen arbitrarily, within some system constraints. In this case, methods will be developed for an input design problem where the fault detection method utilizes parameter identification and parameter variation to detect faults. Specifically, we will propose a method for input design when nonlinear combinations of the parameters (such as products or ratios) are of interest. This work was originally motivated by the problem presented in Hilt *et al.* (2006) in which a thin film deposition process was described for which proper operating conditions should be verified through the ratio of certain parameters in a linear identification model. In Hilt *et al.* (2005), a simple input design method was proposed for estimation of this ratio. The work in this thesis extends this by considering other nonlinear combinations of parameters, allowing for nonlinear parameter constraints, and considering the effect of time correlations in the input sequence. In addition, the input is designed to allow maximize accuracy in detecting faults when monitoring the variation in the ratio of these parameters.

In summary, the contributions of this thesis are:

- A residual generation method for achieving fault detection when the input excitation during system identification is poor.
- A computable characterization of the variation of estimates for products and ratios of signals and parameters.
- A design process for arbitrary input design for parameter ratio estimation with ratio constraints that results in a convex optimization problem.

The remainder of this thesis is organized as follows. In Chapter 2, relevant prior research in the areas of SI, FD, input excitation, and input design are discussed. This chapter will give an overview of the important issues in these areas that are the motivation behind and the building blocks for the work introduced in the following chapters.

In Chapter 3, the situation of system modeling and fault detection in the absence of sufficient input excitation is discussed. A residual generation method for isolating system faults from the noise due to the lack of complete system identification is introduced.

Chapter 4 investigates input design for parameter estimation and parameter variation fault detection. A problem is presented in which the goal is to find the optimal input design for estimating parameter combinations such as ratios or products within system constraints. It is shown that this optimization can be formulated as a convex optimization problem.

Conclusions drawn from the previous chapters along with some future directions for research in the relevant areas are discussed in Chapter 5.

Chapter 2

SYSTEM IDENTIFICATION AND FAULT DETECTION

In this chapter a brief discussion of relevant information in the areas of SI, FDI, input excitation, and input design are presented. This information will not only give the foundation for the work done in the following chapters but will also allow a discussion of the issues that make the research in this thesis important. First some notation which will be used throughout this thesis is introduced.

2.1 Notation

x	vector in $\mathbb{R}^n$ where $x = [x_1 \;\; x_2 \;\; \ldots \;\; x_n]$
A	matrix in $\mathbb{R}^{m \times n}$.
I_n	identity matrix of size n x n.
x^T, A^T	transpose.
$A \in \mathbb{S}_{++}^n$	$A \in \mathbb{R}^{n \times n}$ is positive definite, i.e. $x^T A x > 0$ for all $x \neq 0$.
$A \in \mathbb{S}_{+}^n$	A is a positive semi-definite matrix.
$\|x\|$	$\sqrt{x^T x}$ for column vector, $\sqrt{x x^T}$ for row vector.

$\|x\|_A$	$\sqrt{x^T A x}$, weighted norm of x.
$\|A\|_F$	Frobenius matrix norm for $A \in \mathbb{R}^{m \times n}$, $\sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} \|a_{ij}\|^2}$.
$A^\perp$	orthogonal complement such that $A^\perp A = 0$ and $\|A^\perp\|_F = 1$.
$A = U \Sigma U^T$	eigenvalue decomposition where U contains the eigenvectors and Σ contains the eigenvalues on the diagonal.
$A = U \Sigma V^T$	singular value decomposition where U and V are unitary matrices and Σ is a non-negative diagonal matrix.
$A^{\frac{1}{2}}$	a matrix in $\mathbb{S}_+$ such that given $A \in \mathbb{S}_+$, $A^{\frac{1}{2}} A^{\frac{1}{2}} = A$. It can be found through the eigenvalue decomposition of A as $A^{\frac{1}{2}} = U \Sigma^{\frac{1}{2}} U^T$.
$A^{-\frac{1}{2}}$	square root of the inverse of A, if it exists.
$E[x]$	$\int_{-\infty}^{\infty} x f_x(x) dx$, where $f_x(x)$ is the probability density function of x.
$cov(x)$	covariance $E\left[(x - E[x])(x - E[x])^T\right]$, where E is expectation.
$\bar{E}[x_k]$	$\frac{1}{N} \sum_{k=1}^{N} x_k$, sample mean for a vector sequence x_k, $k = 1, \cdots, N$.
$\overline{cov}(x_k^1, x_k^2)$	$\frac{1}{N} \sum_{k=1}^{N} (x_k^1 - \bar{E}[x_k^1])(x_k^2 - \bar{E}[x_k^2])^T$ (biased) sample covariance.
$ratio(x)$	$\frac{x_1}{x_2}$, given $x \in \mathbb{R}^2$. If $x_2 = 0$, $ratio(x) = \infty$.
$prod(x)$	$x_1 x_2$, given $x \in \mathbb{R}^2$.

If the following is the singular value decomposition (Golub & Loan, 1989) of matrix M, with Σ invertible

$$M = \begin{bmatrix} U_M & \tilde{U}_M \end{bmatrix} \begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V_M^T \\ \tilde{V}_M^T \end{bmatrix} \tag{2.1}$$

then let

- $M^+ := V_M \Sigma^{-1} U_M^T$, the Moore-Penrose pseudo inverse.
- $\Pi_M^r := V_M V_M^T = (M)^+ M$, the projection on the row space of M ($range(M^T)$).
- $\bar{\Pi}_M^r := \tilde{V}_M \tilde{V}_M^T = I - \Pi_M^r$, the projection on the right nullspace of M ($null(M)$).
- $\Pi_M^\ell := U_M U_M^T = M(M)^+$, the projection on the column space of M ($range(M)$).
- $\bar{\Pi}_M^\ell := \tilde{U}_M \tilde{U}_M^T = I - \Pi_M^\ell$, the projection on the left nullspace of M ($null(M^T)$).
- $\underline{\sigma}(M) = \min(\Sigma)$, the smallest non-zero singular value.

2.2 Fault Detection and Identification

Fault Detection (FD) is the process of identifying when a system is operating outside of its normal operating parameters, thus indicating that a fault or failure has occurred in the system. Fault Identification (FI) involves isolating the location of the fault and possibly estimating its magnitude. With the complexity of today's industrial systems it is important to be able to detect faults quickly and have some knowledge

of where in the system the fault occurred. Fault Detection and Identification (FDI) can be both a time and cost saving measure. If a process fault is detected while a production run is in progress, the operator can halt that run to avoid wasting time and money by producing defective products. If a sensor fault is discovered, the system may be able to rely on other sensors to continue a successful production run allowing maintenance to be scheduled at a more convenient time.

In Gertler (1998), three important properties of an FDI method's effectiveness are described. These properties are sensitivity, reaction speed, and robustness. Sensitivity refers to the minimum magnitude of a fault necessary for it to be detected by the system. Reaction speed measures the length of time necessary to detect a fault after it first occurs. Finally, the robustness of the system is the ability to detect faults in the presence of noise and modeling errors. It is this last property, robustness, which will be the focus of this thesis.

2.3 Model vs Non-Model

Fault Detection and Identification methods fall into two main categories: model-based and non-model based. Non-model based methods are usually referred to as expert systems. These methods rely on people who have firsthand knowledge of the physical properties of a system to establish bounds for when the system is "in-control". Usually these bounds relate to a certain measurement. For example, if the operation

of a particular engine is being monitored, an expert on the physical properties of that engine may know the normal temperature readings. If that reading is too high then there could be a fault in the engine's cooling system. These limits are usually referred to as "threshold limits." Methods which employ this type of system to detect faults are generally called "limit checking" methods.

Model-based FDI methods rely on analytical models of the system to determine when a fault in the system has occurred. Several good references on model-based FDI methods are Gertler (1988), Gertler (1998), Frank (1990), Iserman (1984), and MacGregor (2003). When the model can not be developed from first principles, System Identification (SI) methods can be used. (See Ljung (1999) for a good overview of system identification.) SI methods, which will be discussed in more detail later, define a model of the system for "in-control" or fault free operation. In model-based FDI, methods for determining when a fault has occurred fall into two main classes. The first class compares the observed measurements of a system to predicted measurements based on a model of the system to determine if the observed measurements correspond to in-control operation. If not, a possible fault has occurred. These methods are called residual generation methods. Most of the model-based research has been focused on residual generation methods. The method considered in Chapter 3 falls into this class. A less studied class of model-based FDI, uses models which are parameterized. By checking the variations in the parameters based on newly observed

data, a change or fault in the system can be observed. This class of methods will be called parameter variation methods. The work in Chapter 4 falls into this class. Both residual generation and parameter variation methods will be discussed in more detail later in this chapter.

This thesis will be concerned with model-based fault detection in the case where the system is too complex for a first principles model to be developed cheaply. Thus system identification is an important part of the complete FDI system. A brief introduction to system identification methods will be presented first, followed by a look at how the model developed by the SI methods is used in FDI.

2.4 System Identification for Static Systems

Three popular methods of SI and model reduction are Principal Component Analysis (PCA), Partial Least Squares (PLS), and Fisher Discriminant Analysis (FDA). First, a typical static system which will be used to illustrate all three methods is defined.

2.4.1 Static Systems

For use throughout this thesis we now define what is meant by a static system. When the output of a system at any time, k, depends solely on the input at that time (with corruption by some noise), the system is considered a static system. In other words, the system has no memory. Measurements $k = 1, ..., N$ of a linear static

system can be described by

$$y(k) = Au(k) + e(k) \tag{2.2}$$

where

$$k = 1, ..., N.$$

In this system, $u(k) \in \mathbb{R}^m$ is the input and $y(k) \in \mathbb{R}^n$ is the output. The model of how the input is related to the output is represented as the matrix $A \in \mathbb{R}^{n \times m}$ plus some error, $e(k) \in \mathbb{R}^n$. For a graphical representation refer to Figure 2.1. The dimension of $y(k)$ and $u(k)$ will determine the number of outputs and inputs of the system, respectively. If both $u(k)$ and $y(k)$ are scalars then the system is called a single input, single output (SISO) system. If $u(k)$ and $y(k)$ are vectors in $\mathbb{R}^m$ and $\mathbb{R}^n$, respectively, with $m, n > 1$ then the system is termed a multiple input, multiple output (MIMO) system. The systems considered in this thesis will be MIMO systems.

Data for a "window" of time can be captured into a matrix using the following notation.

$$Y = \begin{bmatrix} y(1) & y(2) & \cdots & y(N) \end{bmatrix} \tag{2.3}$$

where the matrix contains N consecutive time steps. If the input, $u(k)$, and the error, $e(k)$, are written similarly as U and E respectively, an equation for the system can be written in the matrix form

$$Y = AU + E. \tag{2.4}$$

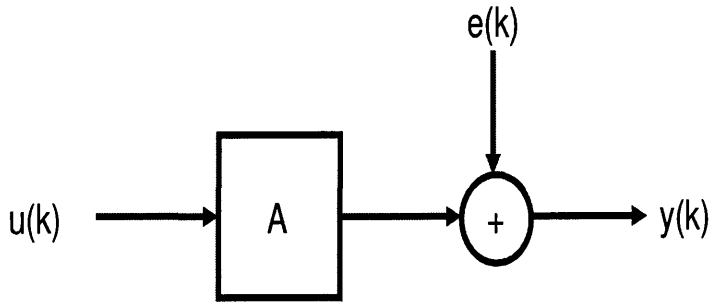

Figure 2.1. A static system with an input of $u(k)$, a parameter model A, an error term $e(k)$, and output data $y(k)$.

2.4.2 Least Squares

Now that the static system has been introduced, some methods for determining the parameter matrix, A, in equation (2.4) which models a static system can be discussed. While PCA, PLS, and FDA are three popular methods for determining A, least squares will be introduced first as it is a basis for these basic methods of SI. Least squares regression is the most common method for solving a system of equations. Consider equation (2.4) where we wish to estimate A. The least squares method finds the estimated parameter matrix, $\hat{A}$, by minimizing the deviations between the output data and the predicted output $\hat{A}U$. This assumes that all the error occurs in the Y direction because U is known. The least squares problem is to find the parameters,

A, that best predict Y in a least squares sense:

$$\hat{A} = \arg\min_{A} \|Y - AU\|^2 . \tag{2.5}$$

Then a prediction of Y is given by $\hat{A}U$.

An alternate formulation for the relationship between inputs and outputs is

$$\begin{bmatrix} Y \\ U \end{bmatrix} = Cs + E. \tag{2.6}$$

Here both U and Y are considered to be measured variables and as such may contain measurement errors. This is an extension of the least squares method called total least squares. The problem now consists of finding the matrix C which defines the relationship between the input/output matrix and a set of independent variables, s. This can be done using the method of Principal Component Analysis.

2.4.3 Principal Component Analysis

Principal Component Analysis (PCA) is a commonly used method for modeling data because of its simplicity and reliability (Lin *et al.*, 2000). Given a data set, X, PCA strives to find a set of latent variables and latent vectors to describe the data

matrix in terms of linear combinations. This is typically written as

$$X = PT \tag{2.7}$$

where X is the data matrix (which may contain both input and output variables), T represents the uncorrelated latent variables (or principal component scores), and P is a matrix comprised of the latent vectors (or the principal component loadings).

In PCA, each latent variable accounts for the variation in the data in a particular direction, with all the variation directions being orthogonal to one another. Consider the data set shown in Figure 2.2. Using methods discussed below, two latent vectors, represented in the figure as the two orthogonal lines, are found which represent the variation in the data set. The first latent vector lies in the direction containing the majority of the variation in the data set. The second latent vector describes the smaller variation in a direction orthogonal to the first. An eigenvector approach is commonly used to determine the latent variables. The eigenvectors and eigenvalues of the sample covariance matrix, $S = \frac{1}{n-1}X^T X$, are found as

$$S = V\Lambda V^T \tag{2.8}$$

where Λ is a diagonal matrix containing the eigenvalues and V is an orthogonal matrix containing the eigenvectors. These eigenvalues and associated eigenvectors allow the

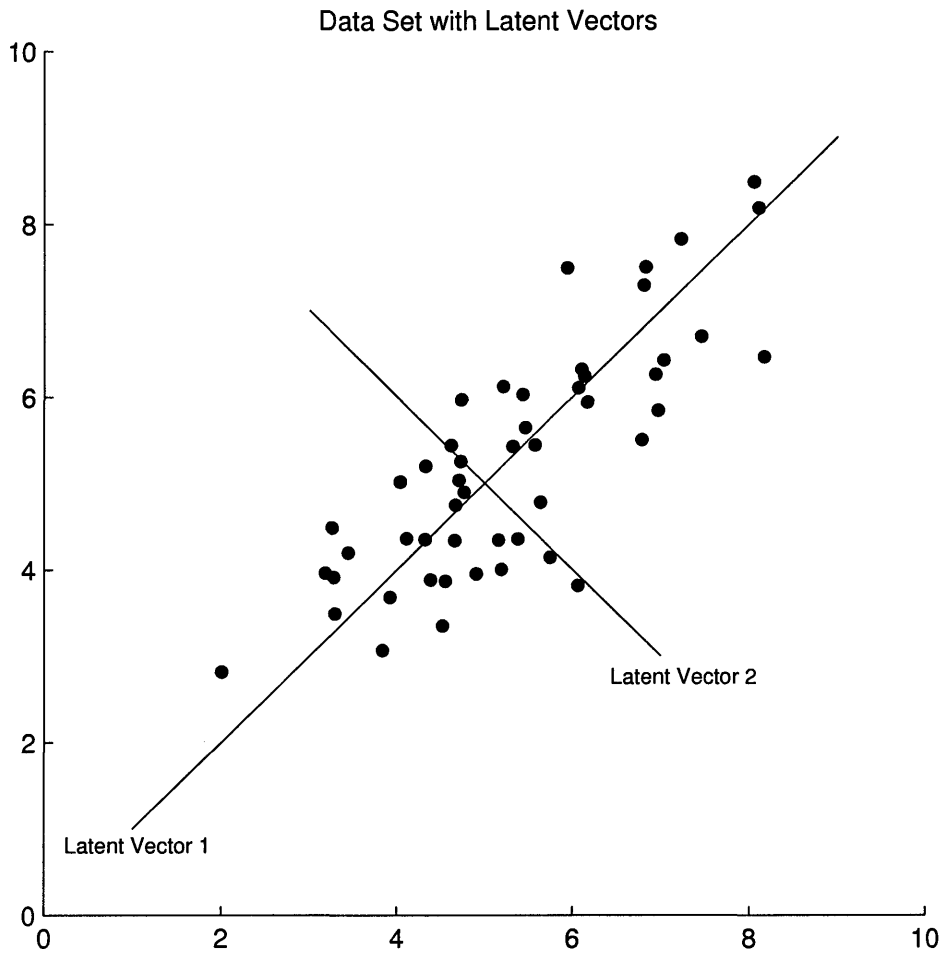

Figure 2.2. Data set with latent vectors found using Principal Component Analysis. It can be seen that the first latent vector describes the majority of the variation within the data set. The second latent vector describes the smaller variation in a direction orthogonal to the first.

directions which describe the variation in the data to be isolated. The eigenvector associated with the largest eigenvalue determines the direction of the largest variation.

Assuming there is some noise associated with the data set, those directions related to the smallest variation or smallest eigenvalues are likely due to this noise in the system and can be dropped from the model. Therefore, the model for the data is

$$X = \hat{P}\hat{T} + E$$

where $\hat{P}$ and $\hat{T}$ are the latent variables and latent vectors associated with the most significant eigenvalues and E is the residual due to noise. Methods for determining the optimal number of latent variables to include in the model will be discussed later. The PCA model can equivalently be found using a singular value decomposition of the data matrix.

Theorem 1. *Singular Value Decomposition: Every $m \times n$ matrix X can be decomposed into its Singular Value Decomposition (SVD):*

$$X = U_X \Sigma_X V_X$$

where

$$U_X \textit{ is } m \times m \textit{ and unitary,}$$

$$V_X \textit{ is } n \times n \textit{ and unitary, and}$$

$$\Sigma_X \textit{ is } m \times n \textit{ and diagonal.}$$

Algorithms for SVD are well established. Many software packages such as Matlab use the LAPACK routines for SVD. For more information on LAPACK see Anderson *et al.* (1999).

With the SVD of X available it can now be said that

$$X = PT$$

with

$$P = U_X, \text{ and}$$

$$T = \Sigma_X V_X.$$

Similar to the eigenvalue approach, the smallest singular values, those that are zero (or close to zero), are related to the noise in the system and can be discarded from

the system model. If we decompose the SVD of the data matrix as

$$X = \begin{bmatrix} U_X & \bar{U}_X \end{bmatrix} \begin{bmatrix} \Sigma_X & 0 \\ 0 & \bar{\Sigma}_X \end{bmatrix} \begin{bmatrix} V_X \\ \bar{V}_X \end{bmatrix},$$

assuming that $\bar{\Sigma}_X = 0$ (or close to zero), then

$$X = U_X \Sigma_X V_X + E$$

just as in the eigenvalue approach.

The number of latent vectors chosen for the model affects the accuracy of the model and there is a delicate balance between overfitting and inaccuracy. Determining the dimensions of a particular model will be discussed in a later section. If only one principal component is chosen, it should be the latent vector associated with the largest latent value, or singular value. It can be shown that the first principal component is also what is commonly referred to as the "line of closest fit" as seen in Figure 2.2 where the error is measured as the distance between each point and the latent vector on a line perpendicular to the latent vector.

2.4.4 Partial Least Squares

Like PCA, Partial Least Squares (also referred to as Projection to Latent Structures) finds loading vectors that capture the correlation within the data. The major

difference is that PLS splits the data into two sets. The predicted (dependent) set, Y, usually contains product quality data or some other form of output data. The predictor (independent) set, X, contains other process variables. The goal of PLS is to maximize the covariance between these two sets in contrast to PCA which strives to find the covariance structure of the data matrix. This method is not considered in this thesis due to the necessity of defining predictor and predicted data sets. More details on PLS can be found in Russell *et al.* (2000).

2.4.5 Fisher Discriminant Analysis

Fisher Discriminant Analysis (FDA) comes from the pattern classification community. In contrast to PCA or PLS it works by using data collected while a specific fault or classes of faults occurred. FDA strives to maximize the separation among the classes. This is done by determining projection vectors that maximize the scatter between the classes and minimize the scatter within the classes. Due to focus on the separation among fault classes, FDA is often used for fault isolation. However, we do not consider it in this thesis because of the necessity of having diagnosed data for a range of faults. For more information see Russell *et al.* (2000).

2.4.6 Determination of Model Order

All the above system identification techniques require a determination of the ideal order of the model. One popular method is Akaike's Information Criterion

(AIC). This criterion tries to balance the misclassification rate against a penalty for the size of the model. AIC finds the dimensionality, a, which minimizes

$$fm(a) + \frac{a}{\tilde{n}}$$

where $fm(a)$ is a measure of the misclassification rate and $\tilde{n}$ is the average number of observations per class. The misclassification rate, $fm(a)$, is a number between 0 and 1 indicating the percentage of misclassifications resulting from projecting the data set onto a model of order a (Russell *et al.*, 2000). Another common method to determine the dimensionality is Prediction Residual Sum of Squares or PRESS. Also known as the cross-validation methods, separate data sets (validation data) are used to determine the accuracy of the model developed. The model which predicts the validation data with the most accuracy is determined to have the ideal order.

One particular method of choosing the order of PCA models was developed in Dunia & Qin (1998a). The authors have done work in developing system subspace models for optimal fault detection. They proposed a method to determine the ideal PCA subspace model for use with fault detection. Their algorithm for determining the optimal set of sensors and size of the model uses a measure called the unreconstructed variance. This is a measure of the distance between a fault reconstruction and the principal component subspace. Minimizing this variance leads to complete reconstruction. This measure indicates the best PCA model for reconstruction. For

a survey of different methods for determining ideal model order see Jackson (1991).

2.5 System Identification for Dynamic Systems

The previous section discussed several methods of system identification for static systems, or systems that have no memory. However, most industrial systems fall into the category of dynamic systems. With some slight modifications the same methods of system identification previously introduced can be used for dynamic systems. First consider the structure of a typical dynamic system.

2.5.1 Dynamic Systems

In dynamic systems, as opposed to static systems, the input of the system at time k affects the future outputs of the system at times $k+1$, $k+2$, etc. A linear finite dimensional discrete time system can be represented in state space form as follows:

$$\begin{aligned} x(k+1) &= Ax(k) + Bu(k) \\ y(k) &= Cx(k) + Du(k) \end{aligned}$$

where $x(k)$ is the state vector, $y(k)$ is the output vector, and $u(k)$ is the input to the system. A, B, C, and D are matrices defining the effects of the state vector and the input on the output vector and the state at the next time. Therefore, the data at time k has some serial correlation with the data at time $k+n$. See Figure 2.3.

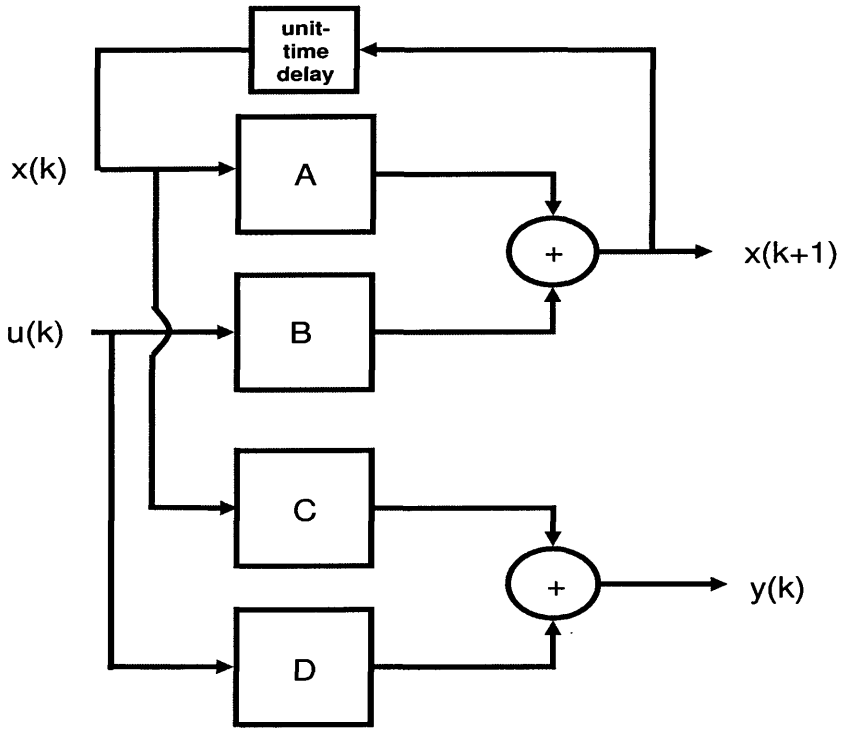

Figure 2.3. A dynamic system where the state, x, at any time, k, affects the future state of the system at time $k+1$, $k+2$, etc. $y(k)$ is the output data and $u(k)$ is the input data. The matrices A,B,C, and D describe the dynamics of the system.

In order to format the dynamic system as a matrix equation similar to the format of the static system shown in (2.4), data matrices are built where time lags are added to each sample of the data. This method is often called the "time lag shift" method. The data matrix for a measurement from a dynamic system is formed as follows:

$$Y^d = \begin{bmatrix} y(1) & y(2) & \cdots & y(N-d) \\ y(2) & \ddots & & \vdots \\ \vdots & & & \\ y(d) & & & y(N) \end{bmatrix}. \tag{2.9}$$

Here d is the number of time samples for which serial correlations should be modeled. The matrices that model the relationship of the input and the state to the output are then as follows:

$$\Gamma = \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{d-1} \end{bmatrix} \text{ and} \tag{2.10}$$

$$H = \begin{bmatrix} D & 0 & \cdots & 0 \\ CB & D & & \\ \vdots & & \ddots & \\ CA^{d-2}B & CA^{d-3}B & \cdots & D \end{bmatrix}. \tag{2.11}$$

The matrix equation describing the system becomes

$$Y^d = \Gamma X^1 + HU^d. \tag{2.12}$$

Since a matrix equation is available to describe the dynamic system, methods such as PCA can be used.

2.5.2 Dynamic Principal Component Analysis

Although the PCA method discussed previously was developed to model static systems, it is the basis for an extended method called Dynamic Principal Component Analysis (DPCA). Using the system equation (2.12), a combination of regression and latent variable methods can be used to identify Γ and H. Usually this involves multiplying Y by a correlation matrix to remove the U term. Then Γ can be determined using a system identification method such as those discussed above. This will be discussed in more detail later. For more details see Chou & Verhaegan (1997) and Ljung (1999).

2.5.3 Linear System Parameter Identification

While SI using DPCA assumes no a priori information, an alternative model of the system can be used if there is some knowledge as to the structure of the system to be identified. This model uses a known linear model structure with a set of parameters

that are found through system identification. For example, the ARX-model of a linear dynamic system is:

$$y(t) + a_1 y(t-1) + ... + a_n y(t-n) = b_1 u(t-1) + \ldots + b_m u(t-m). \tag{2.13}$$

The coefficients, a_i for $i = 1, .., n$ and b_j for $j = 1, .., m$, are considered unknown parameters for some known model structure and can be collected into the matrix

$$\theta = \begin{bmatrix} a_1 & \cdots & a_n & b_1 & \cdots & b_m \end{bmatrix}^T. \tag{2.14}$$

The system model (2.13) can now be written as:

$$A(q, \theta) y(t) = B(q, \theta) u(t) \tag{2.15}$$

where

$$A(q, \theta) = 1 + a_1 q^{-1} + ... + a_n q^{-n}, \text{and} \tag{2.16}$$

$$B(q, \theta) = b_1 q^{-1} + ... + b_m q^{-m} \tag{2.17}$$

with q^{-i} representing the delay operator. The equation for a measurement can be stated as:

$$y(t) = L_{yu}(\theta) u(t) + L_{ye}(\theta) e(t) \tag{2.18}$$

where $L_{yu}(\theta) = \frac{B(q,\theta)}{A(q,\theta)}$ and $L_{ye}(\theta) = \frac{1}{A(q,\theta)}$ are linear operators parameterized by θ, and $e(k)$ represents measurement noise. By writing the measurement equation in this way, we can take advantage of knowledge of the system structure even though the exact parameters, θ, are unknown.

Using experimental data, an estimate of the parameters can be found through a least squares regression similar to (2.5). The least squares problem for parameter estimation is:

$$\hat{\theta} = \arg\min_{\theta} \|Y - \theta^T \varphi\|^2 \tag{2.19}$$

where

$$\begin{aligned} \varphi &= [\ \varphi(1) \ \ldots \ \varphi(N)\ ], \text{ and} \\ \varphi(t) &= [\ -y(t-1) \ \cdots \ -y(t-n) \ \ u(t-1) \ \cdots \ \ u(t-m)\ ]^T. \end{aligned}$$

How this representation can be used for parameter variation fault detection will be discussed in a later section.

2.6 Model-Based FDI for Dynamic Systems

Using system models created by the methods discussed in the previous sections, model-based FDI can be performed. As discussed earlier there are two classes of model-based FDI, residual generation and parameter variation. We begin by dis-

cussing the main principles behind residual generation for dynamic systems.

2.6.1 Residual Generation for Dynamic Systems

Using models created by SI methods, residuals can be generated to monitor for faults in the system. In Frank (1990), the author determines several categories of residual generation methods. These categories are the parity space approach, the dedicated observer approach, and the fault detection filter approach. Residuals are normally generated from the difference between the observed measurements and the predicted measurement based on the controlled inputs and the system model determined through a system identification method.

Using the parity space approach as an example, assume Γ and H are determined by SI methods discussed Section 2.5. A new set of measurements, Y_n, are collected for a new known input:

$$Y_n = \Gamma X + HU_n. \tag{2.20}$$

The delay parameter, d, has been omitted for ease of notation. The degree to which the new data fits with the system model can be determined by calculating the residual for the new data. Since the residual is defined as the difference between the observed measurements and the predicted measurements, we can generate the residuals using

$$R = Y_n - \Gamma X - HU_n. \tag{2.21}$$

However, this formula for the residuals depends on the unknown state variable X and ideally the residuals should be decoupled from the state (Gertler, 1998). To solve this problem a matrix, W, is chosen so that $W\Gamma = 0$. This decouples R from X, leaving the residual equation as

$$R = W(Y_n - HU_n). \tag{2.22}$$

One possible choice for W is $\Gamma^{\perp}$ where $\Gamma^{\perp}\Gamma = 0$ and $\left\|\Gamma^{\perp}\right\|_F = 1$ (Spinhirne & Vincent, 2004). This will be discussed in more detail in Chapter 3.

If no fault has occurred and there is no measurement noise, model errors, or disturbances then the residual should be zero. Since it is rarely the case that no noise or disturbances exist and that the model has no errors, thresholds are set for the allowable magnitude of the residual before a fault is detected. Commonly a χ^2 test is used for fault detection. It is assumed that the noise in the system is Gaussian (Gertler, 1998). For this test a confidence region is determined in which a certain percentage of all the nominal residuals fall. If a residual is outside this confidence region, then a possible fault is detected.

The residuals generated above can also be used for the purpose of fault identification. By examining the contribution certain variables make to any residuals which fall outside the confidence region, the variables where a possible fault has occurred can be isolated. This type of examination is usually presented in a contribution plot. Another way of identifying the particular fault that has occurred is by using a bank

of fault models. A model is developed for the system when it is operating under a particular fault condition. The residuals are then generated for each fault model. The fault model which produces the smallest residuals indicates the most probable fault. An implementation of this method will be seen in Chapter 3.

2.6.2 Parameter Variation Method

Parameter variation is another class of methods for model-based FDI. In these methods, estimates of the system parameters are found as discussed in Section 2.5.3. Data collected in the presence of a fault would cause the SI method to derive different parameter estimates than those calculated using nominal data sets. Therefore, faults can be detected by comparing the parameter estimates found from nominal data and those found from experimental data. Similar to residual generation we must assume that some of the variation in the parameter estimates is due to noise in the system and establish a confidence region for when the variation is normal and when it indicates a possible fault.

2.7 System Identification for Optimal Fault Detection

Since model-based FDI is dependent on the quality of the system models, it is important to look at how the system identification affects the accuracy of fault detection. Traditionally SI and FDI have been treated as separate endeavors. However it is important to consider how system identification should be done to accomplish

optimal fault detection.

Consider the effect of input excitation on system identification and in turn on fault detection. SI requires a data set from the process to be modeled. However, industry is reluctant to allow their systems to be run solely for the sake of data collection because of the costs. Therefore, system identification must often be done using production data. It may be the case that this data does not cover all modes of the system. In this case the system models found during system identification may not fully describe the system. For example, the Γ and H found using regression and DPCA may not fully represent all possible outcomes of the system. If additional modes occur when doing FDI, this can cause accuracy problems for the fault detection method. The FDI methods will interpret these modes to be faults or failures in the system when in reality they may simply be normal operating conditions which were not encountered when collecting the SI data. This is referred to as a variation in the input excitation.

This thesis focuses on two methods of dealing with the problem of input excitation variation and its effects on model-based FDI. The first method looks at how to account for input excitation variation by filtering the residuals. The second method looks at how to design an input for optimal parameter estimation while staying within operating constraints.

2.8 Systems of Interest

As discussed in the previous sections, this thesis is focused on model-based fault detection and identification for dynamic MIMO systems. In this section two systems are discussed which will be used in the following chapters to demonstrate how the effect of input excitation on model-based FDI can de dealt with. The two systems are the Tennessee Eastman Plant Simulation and a thin film deposition process. We begin with a description of the Tennessee Eastman Plant Simulation.

2.8.1 Tennessee Eastman Plant Simulation

The Tennessee Eastman (TE) Plant Simulation was first proposed in Downs & Vogel (1993). The authors created the simulation while working at the Eastman Chemical Company as a realistic industrial simulation for use in the control and monitoring communities. Originally written in Fortran, the simulation produces data representing a two product process. The plant contains four reactants and one byproduct. There are a total of 41 measurements, 12 manipulated variables and 7 operating modes. All measurements include a Gaussian noise. While mainly used in the process control community (see Lyman & Georgakis (1995), McAvoy & Ye (1994), and Ricker (1996)) it has also been used for the testing of FDI methods. In Russell *et al.* (2000), the TE plant simulation was used as a basis for comparing different methods of fault detection and diagnosis. The simulation has 21 prepro-

grammed faults ranging from process variable step changes, to variability increases to slow drifts in reaction kinetics and sticking values.

2.8.2 Thin Film Deposition Process

The second system of interest is a thin film deposition process, as described in Hilt *et al.* (2006). This is a multi-zone co-evaporation process, with the following characteristics: a substrate is moved over multiple deposition zones, in which effusion boats create a metal vapor plume by heating a sample to high temperature. See Figure 2.4. The substrate is unrolled from one reel and taken up by another, and processing can continue until the reels need to be changed. The total thickness and relative composition of the deposited material is measured before final take-up. In multi-source co-evaporation systems the same metal may be deposited in more than one deposition zone, and different deposition rates may be desired. However, when the final thickness and composition is measured, only the total amount of a particular material deposited over all zones is reported. The exact amount deposited in any particular zone is unknown. So, for example, if the composition measurement indicates the total amount of indium deposited is lower than nominal, the set-point estimator will raise the temperature set-points on all effusion cells that deposit indium. If this decrease is due to only one particular effusion source, the relative amount of material deposited in each zone will remain out of balance, so that the desired amount of material deposited in a particular zone will not be achieved even if the sum over

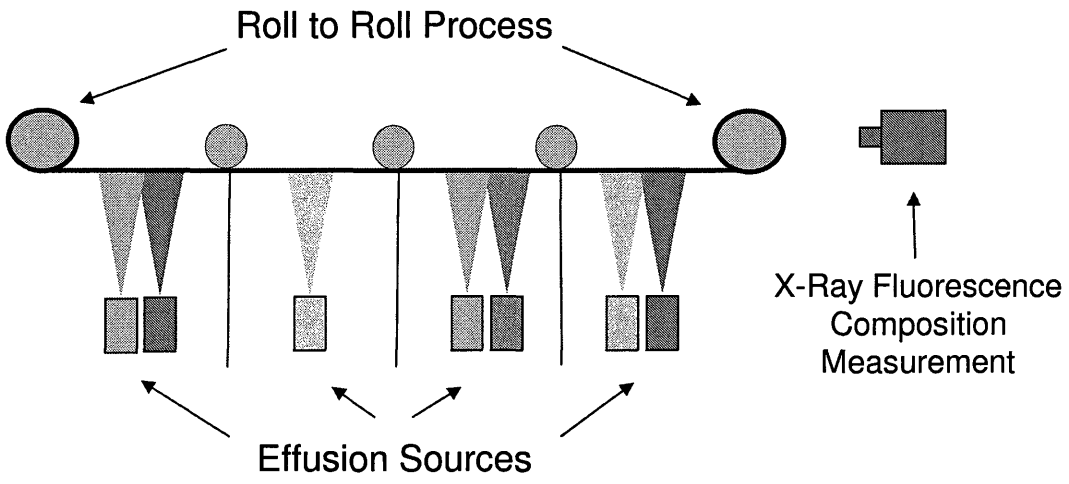

Figure 2.4. Thin Film Deposition Process in which a substrate is moved from roll to roll over several effusion sources. Each deposition zone may contain multiple sources and a particular material may be deposited in more than one zone. The total amount of each material is measured before the completed product is collected on the final roll.

all zones will be regulated correctly.

As discussed in Hilt *et al.* (2006), the relative deposition rates between two zones in the deposition system can be estimated from the ratio of the DC gain of a linearized model, identified using a system identification experiment. In order to be compatible with a manufacturing environment, the perturbation sequence utilized for the system identification process must satisfy some performance objectives. It turns out that for photovoltaic manufacturing, those performance objectives are in the form of ratios. For example, for photovoltaics based on $CuIn_{1-x}Ga_xSe_2$ films, the ratios Cu:(In+Ga) and Ga:(In+Ga) are considered important for achieving the

desired electronic properties (Birkmire & Eser (1997)).

This system is particularly relevant because of the interest in nonlinear combinations of the model parameters, such as the ratio of the DC grain. As discussed earlier, since this is an industry process, system identification must be done on production data. Therefore it is of interest to design inputs that allow for accurate estimating of the nonlinear combinations of the model parameters while still remaining within production constraints.

Chapter 3

FAULT DETECTION USING DPCA WITH VARYING INPUT EXCITATION

3.1 Introduction

As discussed in the previous chapter, one approach to fault detection is residual generation which involves testing system input/output data against assumed nominal and/or faulty behavior models to determine which best explains the data. When analytical models are available, residual generators can be built to calculate the distance from the input/output set to this model. As discussed in the previous chapter, it is often the case that the system does not have a well defined, compact model, so that information about the system must be derived from experimental data. In this case, an identification procedure can be used to obtain a nominal model. In this chapter, we examine the Dynamic Principal Component Analysis (DPCA) method for identification and analysis. For use with dynamic systems, DPCA extends traditional Principal Component Analysis (PCA) by creating a data vector which consists of values of the inputs and outputs over a window of time. This has been referred to as the "time lag shift" method of allowing a PCA model account for dynamic systems

(Ku *et al.*, 1995). The use of DPCA fault detection and isolation is discussed in Ku *et al.* (1995), Dunia & Qin (1998a), Li & Qin (2001), and Qin & Li (2001). The identification portion of DPCA is closely related to the family of subspace identification methods (see Overschee & Moor (1996) and the references therein). In fact, DPCA in essence simply stops at the identification of the extended observability matrix and input/output Toeplitz matrix, rather than converting to state space form. Although the PCA models give rise to windowed or dead-beat estimators which can be subject to problems like noise sensitivity, the resulting validation tests are quite simple with projection operators or least squares playing the central role.

The utilization of DPCA models for fault detection has some interesting features when the input excitation can vary. It is the objective of this chapter to point out one of those features and explore its effect on the fault detection process. In particular, we focus on the case when the input excitation is "sufficiently exciting" from the point of view of identifying a unique extended observability matrix but not to obtain a unique Toeplitz matrix. On the face of it, this restricts the use of the model for fault detection to conditions with the same (or less) degree of excitation. However, by keeping track of the identification input excitation conditions in the fault detection process, it is possible to still do fault detection and isolation without restricting the form of the input, although one necessarily loses some power to find faults (because some faults can become hidden.)

The degree of input excitation is an important issue in fault detection. There is usually opposition to performing dedicated identification experiments with arbitrarily defined inputs in industrial processes. Thus the fault detection process that is based on identified models must make the best use of available data.

A common theme throughout will be residual generation by on-line identification of parameters in model structures that characterize what is completely unknown about the model under consideration. For example, when the system model is not uniquely defined by the identification data (because the input sequence is not sufficiently exciting), the "unknown" part of the model is characterized by a particular null space. This can be used to remove unwanted parts of the residual.

3.2 PCA and Dynamic PCA

In this section, we discuss Principal Component Analysis (PCA) and Dynamic PCA (DPCA) in more detail. This material is standard, but is included for completeness. Good references are Ku *et al.* (1995), Shi & MacGregor (2001), and Dunia & Qin (1998a).

As discussed in the previous chapter although PCA does not necessarily relate to an underlying model of the process, it is useful to consider the identification of the system

$$y(k) = Cx(k) + e(k) \tag{3.1}$$

where $y(k) \in R^p$ are measurements, $x(k) \in R^n$ are free parameters defining the current system state, and $e(k) \in R^p$ is noise. To apply classical PCA, given a measurement sequence $y(k)$ $k = 1, \cdots, N$, we can form the following equation using the matrices related to each sequence:

$$Y^1 = \Gamma X^1 + E^1 \tag{3.2}$$

with $\Gamma = C$. The sequence $y(k)$ can contain measurements of both inputs and outputs. Note, for our purposes, the state representation coordinates are unimportant. With the coordinate transformation $\tilde{X}^1 = QX^1$, where Q is an invertible matrix, an equivalent model is

$$Y^1 = \tilde{\Gamma}\tilde{X}^1 + E^1$$

with $\tilde{\Gamma} = \Gamma Q^{-1}$. Thus, Γ will be considered known if it is found within a coordinate transformation.

Now, if the noise $E^1 = 0$, then from (3.2) the column space of Γ will be the same as Y^1. A singular value decomposition of Y^1 will have the form

$$Y^1 = U_{Y^1}\Sigma V_{Y^1}^T$$

where the terms that are zero have been left out so that Σ is full rank. Since U_{Y^1} is a matrix with full column rank that has the same column space as Y^1, appropri-

ate estimates are $\hat{\Gamma} = U_{Y^1}$ and $\hat{X}^1 = \Sigma V_{Y^1}^T$. Note, this step is typically stated as eigenvalue/eigenvector analysis of $Y^1 (Y^1)^T$, but the results are the same.

If the noise is small, then by setting the smallest singular values to zero, an approximate solution can be found as discussed in Section 2.4.6. The solution will have the following properties: if $\hat{\Gamma}$ is rank k, then so is $\hat{X}^1$ and $\left\| Y^1 - \hat{\Gamma}\hat{X}^1 \right\|_F$ is minimum over all other rank k choices for $\hat{\Gamma}$ and $\hat{X}^1$. In addition, the sample covariance of $\hat{X}^1$ is diagonal, i.e..$\hat{X}^1 \left(\hat{X}^1\right)^T = \Sigma^2$, so that the elements of $\hat{x}(k)$ are uncorrelated (Golub & Loan, 1989).

Dynamic systems have correlations between measurements at different sample times, so a natural extension of PCA to dynamic systems can occur. Consider the following linear dynamic system

$$x(k+1) = Ax(k) + Bu(k) \tag{3.3a}$$

$$y(k) = Cx(k) + Du(k) + e(k) \tag{3.3b}$$

where $e(k)$ represents both noise and disturbances and $u \in R^m$ is a known input.

Define

$$\Gamma^T = \begin{bmatrix} C^T & (CA)^T & \cdots & (CA^{d-1})^T \end{bmatrix}, \text{ and}$$

$$H = \begin{bmatrix} D & 0 & \cdots & 0 \\ CB & D & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ CA^{d-2}B & CA^{d-3}B & \cdots & D \end{bmatrix}$$

so that, for sequences $y(k)$, $u(k)$, $e(k)$, $k = 1 \cdots N$, and $x(k)$, $k = 1 \cdots N - s$, the associated matrices are related by

$$Y^d = \Gamma X^1 + HU^d + E^d. \tag{3.4}$$

There is still a column space relationship between Y^d and Γ, but it is complicated by the addition of H. Extraction of Γ and H takes place using the following steps:

- Correlation: Let $P = Y^1 Z$. It is most useful for Z to be a matrix that is correlated with ΓX^1, but for which $E^d Z$ is small. ($Z = I$ is the same as no correlation step.)

- Regression: Choose regression vector Φ, with assumption that $X^1 Z = M\Phi$ for some M. ($\Phi = I$ is the same as no regression step.) For technical reasons, we require Φ and $U^1 Z$ to be independent, in the sense that no row of $U^1 Z$ can be

written as a linear combination of rows of Φ. Find

$$\hat{\Theta}, \hat{H} = \arg\min_{\Theta,H} \left\| P - \begin{bmatrix} \Theta & H \end{bmatrix} \begin{bmatrix} \Phi \\ U^1 Z \end{bmatrix} \right\|_F^2 .$$

Note that a weighted least squares, or other regression method such as total least squares can be chosen, but for simplicity, we will use unweighted least squares.

- Extract Γ via a singular value decomposition of $\hat{\Theta}$, that is, $\Gamma = U_{\hat{\Theta}}$, the column space of $\hat{\Theta}$ associated with "significant" singular values. Notice that an estimate of H has already been obtained.

Subspace identification methods (see Overschee & Moor (1996), Ljung (1999) and the references therein) which were apparently developed independently of dynamic PCA, use a method equivalent to that described above to obtain Γ, but in this case it is only an intermediate result toward identification of the matrices defining the state space representation.

3.3 Fault detection and isolation in Dynamic PCA

As discussed previously, model-based fault detection occurs by determining if a new data set is compatible with the nominal, no-fault model in an appropriate sense. If a new data set is not compatible with this model, then a fault has been detected.

For DPCA models, a parity space approach can be easily applied. In this case, we find a matrix $\Gamma^{\perp}$ such that $\Gamma^{\perp}\Gamma = 0$, and $\left\|\Gamma^{\perp}\right\|_2 = 1$, for example $\Gamma^{\perp} = \frac{1}{\|\tilde{U}_{\Gamma}^T\|}\tilde{U}_{\Gamma}^T$. A correlation matrix Z can be selected, if desired. Then a residual matrix R_p is calculated as

$$R_p = \Gamma^{\perp}(Y_n^d Z - HU_n^d Z) = \begin{bmatrix} -\Gamma^{\perp}H & \Gamma^{\perp} \end{bmatrix} \begin{bmatrix} U_n^d Z \\ Y_n^d Z \end{bmatrix} \tag{3.5}$$

where Y_n^d and U_n^d are matrices formed from new recorded data. The size of R_p can be monitored, with a large residual (mean or variance) indicating that the data is incompatible with the model, and appropriate statistical and/or bound tests can be applied to recognize this.

One approach to fault isolation utilizing PCA was discussed in Dunia & Qin (1998b), where a fault model Ξ is assumed, with resulting model

$$Y^d = \Gamma X^1 + HU^d + E^d + \Xi F^d. \tag{3.6}$$

Once a large residual indicates a fault, further tests are performed to determine what fault models are compatible with the data. The residual for the fault model becomes

$$R_p = \begin{bmatrix} -\begin{bmatrix} \Gamma & \Xi \end{bmatrix}^{\perp} H & \begin{bmatrix} \Gamma & \Xi \end{bmatrix}^{\perp} \end{bmatrix} \begin{bmatrix} U^s Z \\ Y^s Z \end{bmatrix} \tag{3.7}$$

which is small if there exists an F^d that can explain the data within the noise bound.

When performing fault detection or isolation, there is the implicit assumption that Γ and H have been identified correctly, so that the parity checks (3.5) or (3.7) can be calculated. However, in the case of DPCA the correlation structure of the input can strongly influence the degree to which the complete model is identified. When the model is incomplete, in a sense to be defined more explicitly later, fault detection and isolation must be done with more care. This is the subject of the remainder of this chapter.

3.4 Properties of Identified DPCA Models

In this section, the main result of this chapter is stated, which is a characterization of the DPCA model with general conditions on the input excitation. This result will allow an explicit characterization of that part of the model which can be identified uniquely using the identification data, and that part which is completely unknown. It is assumed that the data is produced by a fixed linear system (3.3), or equivalently (3.4). One condition for usefulness is that the system parameters should be recovered as the noise goes to zero, or (with appropriate conditions on the noise and on an input sequence generated in open loop) as the data goes to infinity. The properties of the input, u, determine if the column space of Γ and H can be recovered. When the input satisfies conditions to allow the column space of Γ and H to be determined uniquely then the input will be said to be "sufficiently exciting." We

are also interested in the case when this is not so.

Let's examine in more detail the regression step when a unique solution is not guaranteed. The assumptions on the DPCA identification process will be as follows:

A1 Φ is full rank.

A2 $X^dZ = M\Phi$ for some matrix M of compatible size.

A3 No row of U^dZ can be written as a linear combination of the rows of Φ.

These assumptions are quite general. If A1 is not satisfied, we can simply remove the dependent rows from Φ. In practice, A2 will not hold exactly, but only at the cost of a small extra noise term that is not included for simplicity. Finally A3 can also be satisfied by a careful choice of Φ, and loosely speaking, will allow the non-uniqueness of the least squares problem to be divided between Γ and H. Note that one choice which automatically satisfies the first and third requirements is to choose the rows of Φ to be the transpose of vectors in the right null space of U^dZ.

Theorem 2. *Given data $y(k)$ and $u(k)$ generated by the system (3.4), and assumptions A1, A2 and A3,*

$$\min_{\eta} \left\| Y^dZ - \begin{bmatrix} \eta_1 & \eta_2 \end{bmatrix} \begin{bmatrix} \Phi \\ U^dZ \end{bmatrix} \right\|_F^2$$

has minimum norm solution

$$\begin{bmatrix} \eta_1^T \\ \eta_2^T \end{bmatrix} = \begin{bmatrix} (\Gamma X^1 Z \left(I - \left(U^d Z \bar{\Pi}_\Phi^r\right)^+ U^d Z\right) \Phi^+)^T \\ (H\Pi_{U^d Z}^\ell)^T \end{bmatrix} + O$$

where

$$O = E^d Z \begin{bmatrix} \Phi \\ U^d Z \end{bmatrix}^+$$

is a matrix that scales with the noise.

Proof. The minimum norm solution is given by

$$\begin{aligned} \begin{bmatrix} \eta_1 & \eta_2 \end{bmatrix} &= Y^d Z \begin{bmatrix} \Phi \\ U^d Z \end{bmatrix}^+ \\ &= \left(\Gamma X^d Z + H U^d Z + E^d Z\right) \begin{bmatrix} \Phi \\ U^d Z \end{bmatrix}^+ . \end{aligned}$$

Since $X^d Z = M\Phi$ we can write

$$\begin{bmatrix} \eta_1 & \eta_2 \end{bmatrix} = \begin{bmatrix} \Gamma M & H \end{bmatrix} \begin{bmatrix} \Phi \\ U^d Z \end{bmatrix} \begin{bmatrix} \Phi \\ U^d Z \end{bmatrix}^+ + O$$

where

$$O = E^d Z \begin{bmatrix} \Phi \\ U^d Z \end{bmatrix}^+ .$$

Now, since Φ has full row rank, and no row of $U^d Z$ can be written as a linear combination of the rows of Φ, we have

$$\begin{bmatrix} \Phi \\ U^d Z \end{bmatrix}^+ = \begin{bmatrix} \left(I - \left(U^d Z \bar{\Pi}^r_\Phi\right)^+ U^d Z\right) \Phi^+ & \left(U^d Z \bar{\Pi}^r_\Phi\right)^+ \end{bmatrix}$$

(this can be verified by showing that the right hand side satisfies the properties of the pseudo-inverse) and thus

$$\begin{bmatrix} \eta_1 & \eta_2 \end{bmatrix} = \begin{bmatrix} \Gamma M & H \end{bmatrix} \Lambda + O$$

where

$$\Lambda = \begin{bmatrix} \Phi \left(I - \left(U^d Z \bar{\Pi}^r_\Phi\right)^+ U^d Z\right) \Phi^+ & \Phi \left(U^d Z \bar{\Pi}^r_\Phi\right)^+ \\ U^d Z \left(I - \left(U^d Z \bar{\Pi}^r_\Phi\right)^+ U^d Z\right) \Phi^+ & U^d Z \left(U^d Z \bar{\Pi}^r_\Phi\right)^+ \end{bmatrix}.$$

But $\Phi \left(U^d Z \bar{\Pi}^r_\Phi\right)^+ = 0$ and if no row of $U^d Z$ can be written as a linear combination of the rows of Φ, $U^d Z \left(U^d Z \bar{\Pi}^r_\Phi\right)^+ = \Pi^\ell_{U^d Z}$. This further implies that

$$U^d Z \left(I - \left(U^d Z \bar{\Pi}^r_\Phi\right)^+ U^d Z\right) \Phi^+ = 0,$$

so we arrive at

$$\begin{bmatrix} \eta_1^T \\ \eta_2^T \end{bmatrix} = \begin{bmatrix} (\Gamma X^1 Z \left(I - \left(U^d Z \bar{\Pi}_\Phi^r\right)^+ U^d Z\right) \Phi^+)^T \\ (H \Pi_{U^d Z}^\ell)^T \end{bmatrix} + O$$

as claimed. □

Now, if $X^1 Z \left(I - \left(U^d Z \bar{\Pi}_\Phi^r\right)^+ U^d Z\right) \Phi^+$ has full row rank, then Γ can be recovered by singular value decomposition of η_1. Also if $\Pi_{U^d Z}^\ell = I$, then η_2 is an estimate of H. The condition for $X^1 Z \left(I - \left(U^d Z \bar{\Pi}_\Phi^r\right)^+ U^d Z\right) \Phi^+$ to be full rank is mainly a condition on the state sequence X^1 - that is, the input must be sufficiently exciting so that all modes of the system are excited. However, this can be achieved without having $\Pi_{U^d Z}^\ell = I$. Consider a two input system fully controllable from each input, with one input white noise and the other input zero. The input is sufficiently exciting from the standpoint of identifying Γ, but the impulse response from the second input to the output is completely unknown. There are other, more interesting cases as well, when some inputs are only weakly exciting, or when the window size d is chosen larger than the degree of input excitation.

It is this intermediate case that will be of interest in the remainder of the chapter. In particular, a method for performing fault detection and isolation will be presented which accounts for the parts of the model that are completely unknown.

If we have some knowledge of the noise behavior, we can obtain bounds on the

matrix O. One typical assumption is that the noise is bounded and uncorrelated with the input and weighting matrix Φ, and that the magnitude of the input and Φ is bounded from below. We use the following, related assumptions that do not depend on a stochastic model of the noise:

B1 $$\left\| E^d Z \begin{bmatrix} \Phi \\ U^d Z \end{bmatrix}^T \right\|_F^2 < kN.$$

B2 $$\underline{\sigma}\left(\begin{bmatrix} \Phi \\ U^d Z \end{bmatrix} \right) > c\sqrt{N}.$$

Assumption $B1$ bounds the correlation between the noise and the regressors used in the regression step. If $E^d Z$ and the regressors were perfectly correlated, then this term would grow as N^2, rather than N. Assumption $B2$ ensures that the non-trivial part of the input does not decay as time increases. With these assumptions, we can show that the size of O decreases with N.

Theorem 3. *Given* $O = E^d Z \begin{bmatrix} \Phi \\ U^d Z \end{bmatrix}^+$. *Let* r *be the number of independent rows of* $\begin{bmatrix} \Phi \\ U^d Z \end{bmatrix}$. *If assumptions B1 and B2 hold, then* $\|O\|_F^2 \leq \frac{l}{N}$ *where* $l = \frac{rk}{c^4}$.

Proof. Let $A = \begin{bmatrix} \Phi \\ U^d Z \end{bmatrix}$, $E = E^d Z$, and let $A = \begin{bmatrix} U & \tilde{U} \end{bmatrix} \begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V^T \\ \tilde{V}^T \end{bmatrix}$ be the singular value decomposition of A with Σ invertible. Then $\left\| EA^T \right\|_F^2$ can be written

as

$$\begin{aligned}\left\|EA^T\right\|_F^2 &= tr\left(E^TEV\Sigma^2V^T\right)\\ &= tr\left(\Sigma^2V^TE^TEV\right)\\ &\geq \underline{\sigma}^2\left\|EV\right\|_F^2\end{aligned}$$

so that

$$\left\|EV\right\|_F^2 \leq \frac{\left\|EA^T\right\|_F^2}{\underline{\sigma}^2}$$

and given assumptions B1 and B2, $\left\|EV\right\|_F^2 \leq \frac{k}{c^2}$. Now,

$$\left\|EA^+\right\|_F^2 = \left\|EV\Sigma^{-1}U^T\right\|_F^2 \leq \left\|EV\right\|_F^2\left\|\Sigma^{-1}U^T\right\|_F^2.$$

But since

$$\left\|\Sigma^{-1}U^T\right\|_F^2 = tr(\Sigma^{-2}U^TU) \leq \frac{r}{\underline{\sigma}^2}$$

we have

$$\left\|EA^+\right\|_F^2 \leq \left(\frac{k}{c^2}\right)\left(\frac{r}{c^2N}\right)$$

$$\left\|EA^+\right\|_F^2 \leq \frac{l}{N} \text{ where } l = \frac{rk}{c^4}.$$

□

3.5 Fault Detection when $\Pi^{\ell}_{U^dZ} \neq I$

What can be done when it is known Γ is correct, but only $H\Pi^{\ell}_{U^dZ}$ can be determined? There are two possibilities:

- Use knowledge of the underlying form of Γ and H: Extract C and A from Γ and then identify B and D from a least squares fit to data.
- Modify the residuals.

In the first case, we follow the subspace system identification methods to their conclusion. However, it is also possible to utilize the form of Γ and H within the DPCA framework, and we will be able to cover the case when even additional structural information does not yield a unique representation. With this motivation, we will examine the problem without appealing to the underlying state space representation.

Because we are primarily concerned with the effect of input excitation, rather than noise, we will assume that the correct Γ, and $H\Pi^{\ell}_{U^dZ}$ are found from the PCA step, and thus also $\Gamma^{\perp}$.

First, we note that if the new input does not satisfy the same correlation structure as the input used for identification, the residuals may be non-zero, even without noise. Suppose Y^d_n and U^d_n, are generated by system (3.4) with $E^d_n = 0$. Then the primary

residual becomes

$$R_p = \begin{bmatrix} -\Gamma^{\perp} H \Pi^{\ell}_{U^d Z} & \Gamma^{\perp} \end{bmatrix} \begin{bmatrix} U^d_n \\ Y^d_n \end{bmatrix}$$

$$= \Gamma^{\perp} H \bar{\Pi}^{\ell}_{U^d Z} U^d_n.$$

If $\bar{\Pi}^{\ell}_{U^d Z} U^d_n \neq 0$, then this residual may be non-zero! Clearly, in general, the residual evaluation can cause false alarms since the residual may be large even if no fault is present. This is due to possible changes in correlation structure of the new input relative to the input used for identification, which will result in $\bar{\Pi}^{\ell}_{U^d Z} U^d_n \neq 0$. However, with additional processing, it is possible to remove the effect of the new correlation structure, while still retaining some capability to detect faults. This is done by performing a partial system identification over the residual subspace which is known to be deficient.

Define:

$$R = R_p - \Gamma^{\perp} \hat{H} \bar{\Pi}^{l}_{U^d Z} U^d_n$$

where

$$\hat{H} = \arg\min_{H} \left\| R_p - \Gamma^{\perp} H \bar{\Pi}^{l}_{U^d Z} U^d_n \right\|^2. \tag{3.8}$$

Essentially, we are determining if there is a $H\bar{\Pi}^{l}_{U^d Z}$ (the part of H which is unknown) that could explain the residual within the noise bound. Note that this formulation

for the new residual is equivalent to:

$$R = R_p \bar{\Pi}^r_{\bar{\Pi}^l_{U^d Z} U^d_n}. \tag{3.9}$$

We can show that the size of this residual is bounded by the size of the noise sequence. With a statistical assumption on the noise, a bound on the expected residual size in the no fault case can also be easily calculated.

Theorem 4. *Given a new input/output sequence Y^d_n, U^d_n, and noise sequence E^d_n for a system satisfying (3.4), and residual generation equation (3.9) with $\left\|\Gamma^\perp\right\|_2 = 1$, then*

1. $\|R\|^2_F < \left\|E^d_n\right\|^2_F$

2. If $e(k)$ is an independent, identically distributed random sequence with zero mean and covariance $\sigma^2 I_p$, then

$$E\left[\|R\|^2_F\right] \leq dp\sigma^2((N-d+1) - \bar{\Pi}^l_{U^d Z} U^d_n).$$

Proof. Using the definition of the residuals

$$R = R_p \bar{\Pi}^r_{\bar{\Pi}^\ell_{U^d Z} U^d_n} = \Gamma^\perp E^d_n \bar{\Pi}^r_{\bar{\Pi}^\ell_{U^d Z} U^d_n}.$$

Clearly

$$\|R\|^2_F = \left\|\Gamma^\perp E^d_n \bar{\Pi}^r_{\bar{\Pi}^\ell_{U^d Z} U^d_n}\right\|^2_F \leq \left\|E^d_n\right\|^2_F.$$

If, in addition, the noise sequence is white with covariance σI_p, then with $M = \bar{\Pi}^l_{U^d Z} U^d_n$,

$$\begin{aligned} \|R\|_F^2 &\leq \left\| E_n^d \bar{\Pi}^r_M \right\|_F^2 \\ &\leq tr((E_n^d)^T E_n^d \bar{\Pi}^r_M). \end{aligned}$$

Since $E\left[(E_n^d)^T E_n^d\right] = dp\sigma^2 I_{(N-d+1)}$,

$$\begin{aligned} E\left[\left\| E_n^d \bar{\Pi}^r_M \right\|_F^2\right] &= tr(dp\sigma^2 I_{(N-d+1)} \bar{\Pi}^r_M) \\ &= dp\sigma^2 tr(\bar{\Pi}^r_M). \end{aligned}$$

Since $tr(\bar{\Pi}^r_M) = (N - d + 1) - rank(M)$,

$$E\left[\left\| E_n^d \bar{\Pi}^r_M \right\|_F^2\right] = dp\sigma^2((N - d + 1) - rank(M)).$$

Therefore

$$E\left[\|R\|_F^2\right] \leq dp\sigma^2((N - d + 1) - rank(M)).$$

□

3.6 Fault Isolation

When a large residual cannot be explained by a change in the input correlation, a fault is called. Fault isolation can take place for a certain class of faults.

Assuming the data is generated by a system that satisfies (3.6), the primary residual becomes

$$\begin{aligned} R_p &= \begin{bmatrix} -\Gamma^{\perp} H \Pi^{\ell}_{UZ} & \Gamma^{\perp} \end{bmatrix} \begin{bmatrix} U \\ Y \end{bmatrix} \\ &= \Gamma^{\perp} H \bar{\Pi}^{\ell}_{U^d Z} U^d_n + \Gamma^{\perp} \Xi F^d_n + \Gamma^{\perp} E^d_n. \end{aligned}$$

We will attempt to eliminate the effect of changes in the input correlation, while attempting to determine if a fault sequence F^d_n exists that can explain the data. In this case, there are unknown matrices which may be manipulated to determine if the data set is compatible with the fault model (3.6), namely $H\bar{\Pi}^{\ell}_{U^d Z}$ and F^d_n. Instead of using projection operators, we will utilize a particular least squares problem which takes advantage of the known shift structure of F^d_n. Let $f(k)$ be the sequence associated

with F_n^d, and let f_{ni}, be the ith column of F_n^d. Let

$$S = \begin{bmatrix} 0 & I & \cdots & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & \vdots & \ddots & I \\ \ddots & 0 & 0 & 0 \end{bmatrix}, \text{ and } T = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ I \end{bmatrix}$$

be matrices of appropriate structure such that f_{ni} can be formed by

$$f_{ni} = Sf_{n(i-1)} + Tf(i+d-1)$$

with appropriate initial conditions. Let $H = \begin{bmatrix} H_1 & H_2 \end{bmatrix} \begin{bmatrix} \tilde{U}_{U^dZ}^T \\ U_{U^dZ}^T \end{bmatrix}$. Therefore $H\bar{\Pi}_{U^dZ}^l = H_1\tilde{U}_{U^dZ}^T$. Let u_i be the ith column of $\bar{\Pi}_{U^dZ}^{\ell}U_n^d$, and let r_i be the ith column of R_p. Then, we solve the following constrained least squares problem:

$$\min_{H,f_1,\cdots,f_N} \sum \left\| r_i - \Gamma^{\perp}Hu_i + \Gamma^{\perp}\Xi f_i \right\|^2$$
$$\text{such that } f_i = Sf_{i-1} + Te(i)$$

where $e(i) = f(i+d-1)$. The solution is straightforward. See, e.g. (Golub & Loan,

1989, 12.1.4). The residual is then formed via

$$R = R_p - \Gamma^{\perp}\hat{H}\bar{\Pi}^{\ell}_{U^d Z}U^d_n - \Gamma^{\perp}\Xi\hat{F} \tag{3.10}$$

where $\hat{H}$ and $\hat{F}$ are the estimates obtained from the constrained least squares problem. When this residual is small, there exists a fault sequence (and an unmodeled portion of H) that can explain the data, and the modeled fault is included as a possible fault.

Using techniques from recursive least squares filtering a recursive solution to this optimization problem can be found. With the appropriate substitutions the problem can be written as:

$$\min_{H,f_1,\cdots,f_N} \sum \|r_i - Cf_i - D_i vec(H)\|^2 \tag{3.11}$$
$$\text{such that } f_i = Sf_{i-1} + Te(i)$$

where $C = \Gamma^{\perp}\Xi$ and $D_i = u_i \otimes \Gamma^{\perp}$. The symbol $\otimes$ represents the Kronecker product.

Theorem 5. *A recursive solution to the optimization problem 3.11 can be found using the following algorithm*

1 Initialize M to

$$M = \begin{bmatrix} 0 & 0 & -C & -D_1 & r_1 \\ -S & -T & I & 0 & 0 \end{bmatrix}.$$

2 Using a QR decomposition find matrices such that $M = Q \begin{bmatrix} R_{11} & R_{12} \\ 0 & R_{22} \end{bmatrix}$ *with* Q *orthogonal.*

3 Using the singular value decomposition find $R_{11} = U\Sigma V^T$. *Set*

$$G_i = -1 * (\Sigma V^T)^{-1}(U^T R_{12})_n, \text{ and}$$

$$e_i = \min \left\| \begin{bmatrix} (U^T R_{12})_z \\ R_{22} \end{bmatrix} \begin{bmatrix} f_i \\ h \\ 1 \end{bmatrix} \right\|^2 \tag{3.12}$$

where $(U^T R_{12})_n$ *are the rows of* $U^T R_{12}$ *associated with the nonzero singular values and* $(U^T R_{12})_z$ *are the rows of* $U^T R_{12}$ *that correspond with the zero singular values, rows* $k+1$ *to* N.

4 Let

$$\begin{bmatrix} (U^T R_{12})_z \\ R_{22} \end{bmatrix} = \begin{bmatrix} \Gamma_{1i} & \Gamma_{2i} & \Gamma_{3i} \end{bmatrix}, \text{ and}$$

$$M = \begin{bmatrix} \Gamma_{1i} & 0 & 0 & \Gamma_{2i} & \Gamma_{3i} \\ 0 & 0 & -C & -D_i & r_i \\ -S & -T & I & 0 & 0 \end{bmatrix}.$$

5 If $k < N$ *go to step 2.*

6 Let

$$\hat{h}, \hat{f}_N = \arg\min_{h, f_N} \left\| \begin{bmatrix} \Gamma_{1N} & \Gamma_{2N} & \Gamma_{3N} \end{bmatrix} \begin{bmatrix} f_N \\ h \\ 1 \end{bmatrix} \right\|^2 .$$

Find $\hat{f}_i, i = 0 \ldots N-1$ *by using*

$$\begin{bmatrix} \hat{f}_{i-1} \\ \hat{e}_i \end{bmatrix} = G_i \begin{bmatrix} \hat{f}_i \\ \hat{h} \\ 1 \end{bmatrix} .$$

Proof. See Appendix A □

3.7 Hidden Faults

Although the filtering described above can help eliminate the effect of a change in the input excitation, it may be possible for a fault sequence to exist without increasing the size of the residual beyond that expected by the noise. To examine this possibility, we assume that data is generated by a system that satisfies (3.6). Using the residual sequence defined by (3.9), we clearly have

$$R = \Gamma^{\perp} \Xi F_n^d \bar{\Pi}^r_{\bar{\Pi}_{U^d Z} U_n^d} + \Gamma^{\perp} E_n^d \bar{\Pi}^r_{\bar{\Pi}_{U^d Z} U_n^d}.$$

A fault will be called if the first term causes R to become larger than expected by the second term alone. However, if

$$\Gamma^{\perp}\Xi F_n^d \bar{\Pi}^r_{\bar{\Pi}_{U^d Z} U_n^d} = 0, \tag{3.13}$$

then a fault will not be called regardless of the size of F_n^d. These types of faults, for which the size of the residual is independent of the size of F_n^d will be called hidden faults. All possible hidden fault sequence F_n^d are given by solutions of (3.13). Equivalently, F_n^d is a hidden fault sequence if and only if it satisfies

$$\Gamma^{\perp}\Xi F_n^d = -\Gamma^{\perp} H \bar{\Pi}^l_{U^d Z} U_n^d. \tag{3.14}$$

Thus, a hidden fault is one whose contribution to the residual is equivalent to unmodeled dynamics.

Because of the possibility of hidden faults, some care must be taken when performing fault detection. However, due to the connection given by equation (3.14) there is some information that is known about the magnitude of a possible hidden fault. Using the estimate of H found from solving (3.8) the size of H is known. A bound on the magnitude of a possible hidden fault is given by

$$\left\| F_n^d \right\|_F \leq \frac{\left\| \Gamma^{\perp} H \bar{\Pi}^l_{U^d Z} U_n^d \right\|_F}{\underline{\sigma}(\Gamma^{\perp}\Xi)}$$

which follows directly from (3.14).

3.8 Simple Example

As an example, consider the following system:

$$
\begin{aligned}
x(k+1) &= \begin{bmatrix} 0 & 1 \\ -0.891 & 1.89 \end{bmatrix} x(k) + \begin{bmatrix} 1 & 0.5 \\ 0 & 0.5 \end{bmatrix} u(k) \\
y(k) &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} x(k) + e(k)
\end{aligned}
$$

where $e(k)$ is a white noise sequence with variance 0.01. An identification experiment is performed, for $k = 1, \cdots, 1000$ with input

$$
u(k) = \begin{bmatrix} \sin(0.4\pi k) \\ n(k) \end{bmatrix}
$$

where $n(k)$ is a simulation of a white Gaussian random sequence with unit variance. Dynamic PCA is applied with $d = 5$ to determine Γ and $H\Pi^{\ell}_{U^d Z}$ with $Z = I$ and $\Phi\Phi' = \bar{\Pi}^{r}_{U^d Z}$. Note that because of the sinusoid input, $\Pi^{\ell}_{U^d Z}$ has only rank 7, rather than 10, which is what is required for $\Pi^{\ell}_{U^d Z} = I$. However, since the system is only 2^{nd} order, Γ can be identified uniquely.

A new input is applied, $k = 1, \cdots, 1000$

$$u_n(k) = \begin{bmatrix} \sin(0.8\pi k) \\ n(k) \end{bmatrix}$$

where, because of the change in frequency, the associated matrix $\Pi^{\ell}_{U^d Z} U^d_n \neq U^d_n$. In addition, we add the sensor fault

$$y(k) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} x(k) + e(k) + \begin{bmatrix} f(k) \\ 0 \end{bmatrix}$$

where

$$f(k) = \begin{cases} 0 & k < 500 \\ \sin(0.4\pi k) & k \geq 500. \end{cases}$$

The primary and filtered residuals are shown in Figure 3.1. Given the noise variance, the expected bounds on the residuals are shown in both cases. Note that the primary residual is much greater than expected for noise alone, even before the fault is applied, thus a fault would be called incorrectly at time 0. Therefore the identification input, $u(k)$, was not sufficiently exciting to determine H completely. This has caused false alarms to occur. In order to eliminate the false alarms and correctly detect the fault, the residual is filtered using (3.9). It can clearly be seen in the graph of the filtered residual in Figure 3.1 that the additional filtering has the

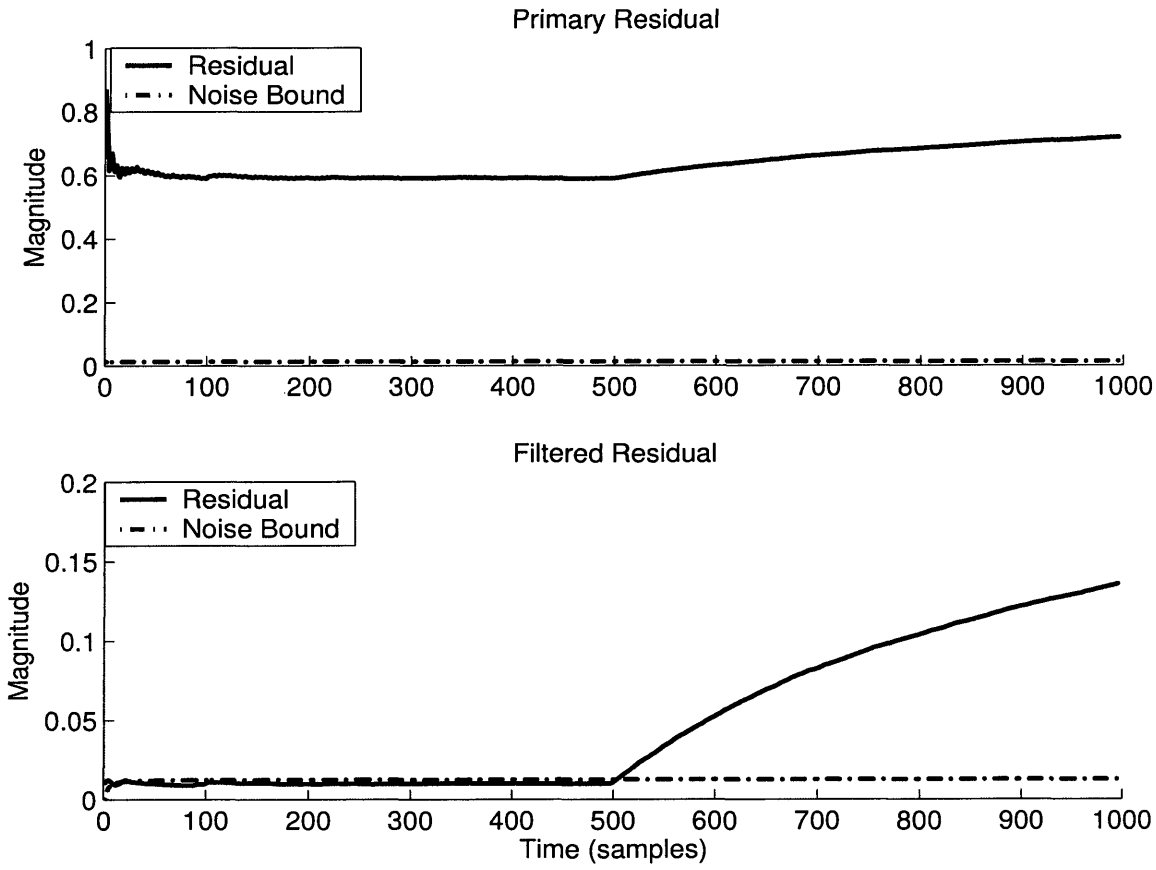

Figure 3.1. Primary residuals from input/output data with the input 'more' exciting than the input used in the identification data set (top) and the residuals after filtering (bottom). Note that the primary residual is always above the noise bound in contrast with the filtered residual which falls above the noise bound only when a fault has occurred.

desired behavior. The filtered residual corresponds with the expected noise bound until time 500 when the fault occurs. This indicates that the filtering has not only eliminated false alarms but also allowed the fault to be detected correctly.

For fault isolation, we determine if a fault model is compatible with the data, by using the fault isolation method described in section 3.6. Using the solution to the minimization problem, the error is tracked for each time step. The results are shown in Figure 3.2. In this case, the fact that the residual stays small for fault model 1 indicates that this fault model is compatible with the data, which is correct. However, the residual for fault model 2 still indicates a fault occurring, meaning that a system model indicating a fault in sensor 2 is not correct.

3.9 Fault Detection for the Tennessee Eastman Plant Simulation

Now consider the more complex example of the Tennessee Eastman Plant Simulation described in Section 2.8.1. A simulink model of the control strategy described in Ricker (1996) was used to generate nominal and fault data for the plant with different input excitation. The system identification step was done using the nominal case where the set points are constant through time. The process was run for a simulated time of 72 hours with no disturbances occurring during the simulation. Measurements where taken every 3 minutes. Using the data produced by the simulation and the Dynamic PCA methods discussed earlier with window parameter $d = 5$, Γ and

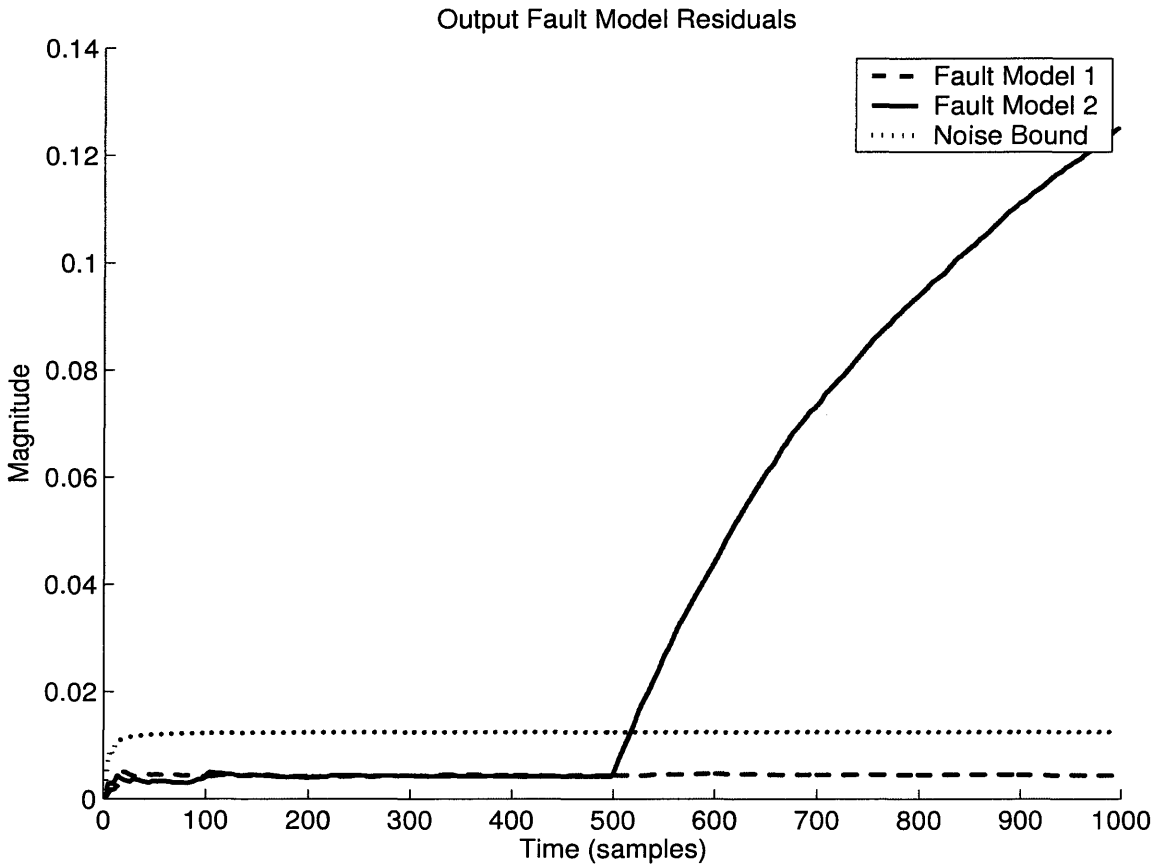

Figure 3.2. Residuals for two fault models capturing faults in outputs 1 and 2 respectively. Fault Model 1 stays below the noise bound indicating that this is the correct model for the fault that occurred in the system.

$H\Pi^{l}_{U^dZ}$ were found using $Z = I$ and $\Phi\Phi' = \bar{\Pi}^{r}_{U^dZ}$.

New data was then collected with a different level of input excitation and a fault occurring. The different level of input excitation was caused by the using a sinusoidal input for the production set point instead of a constant value. In addition, disturbance 14 which is a sticking value disturbance was introduced to the system around hour 30. A predicted noise bound was calculated using residuals as derived in (3.5) for known nominal data in which the input excitation matched that of the system identification data.

The primary and filtered residuals are shown in Figure 3.3. As in the previous example, the difference in input excitation causes the primary residual to be above the noise even in the absence of a fault. Therefore, it would be concluded that a fault is present for the entire process time. In contrast, the bottom graph shows the filtered residual which is clearly below the noise bound until the fault occurs midway through the process run time. This shows that the filtered residuals have removed the effect of differing input excitation between the system identification data and the fault detection data.

Fault Isolation can also be performed for this example. Due to the unknown structure of the fault occurring in the system, fault isolation can not be performed as in section 3.7. Instead, a bank of fault models are used to isolate the model which describes the new data correctly. Two models were produced using the method dis-

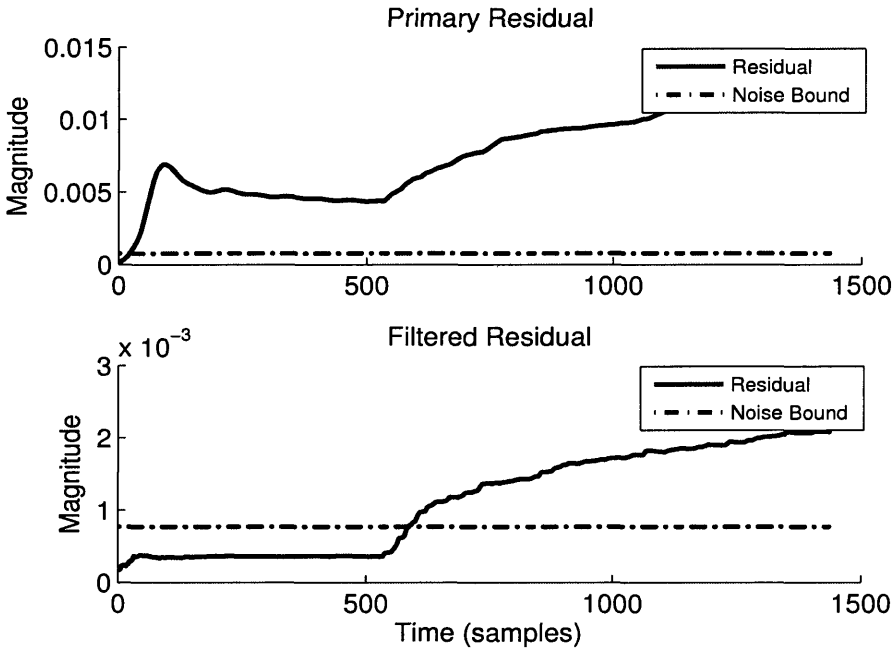

Figure 3.3. Primary and Filtered Residuals with the system identification data collected using constant set points and fault detection performed on data with a sinusoidal production set point, this causing a difference in the input excitation.

cussed in section 3.5 and system data collected during the occurrence of disturbances 14 and 15 separately. Residuals comparing the fault detection data with these two models were calculated using (3.9).

The results are shown in Figure 3.4. The residual produced using fault model 14 stays below the noise bound for the length of the simulation. This indicates that the new data matches fault model 14. In contrast, the residual for fault model 15 increases above the noise bound after sample 500, indicating that this model does not match the new data. Therefore we have correctly identified fault model 14 as the appropriate model for the new data.

3.10 Conclusion

In this chapter, a fault detection process was described for dynamic PCA with independent inputs where the input satisfied only a partial excitation condition, and no restrictions were placed on the new data to be tested for faults. It was shown that although the identified model only partially described the input/output Toeplitz matrix, fault detection and isolation could still occur, although at the cost of the possibility of some hidden faults.

Although an alternative formulation could have used a state space representation, the fault detection filter remains in a very simple form, retaining the advantage of the DPCA/Parity space system identification/fault detection procedure. The fault

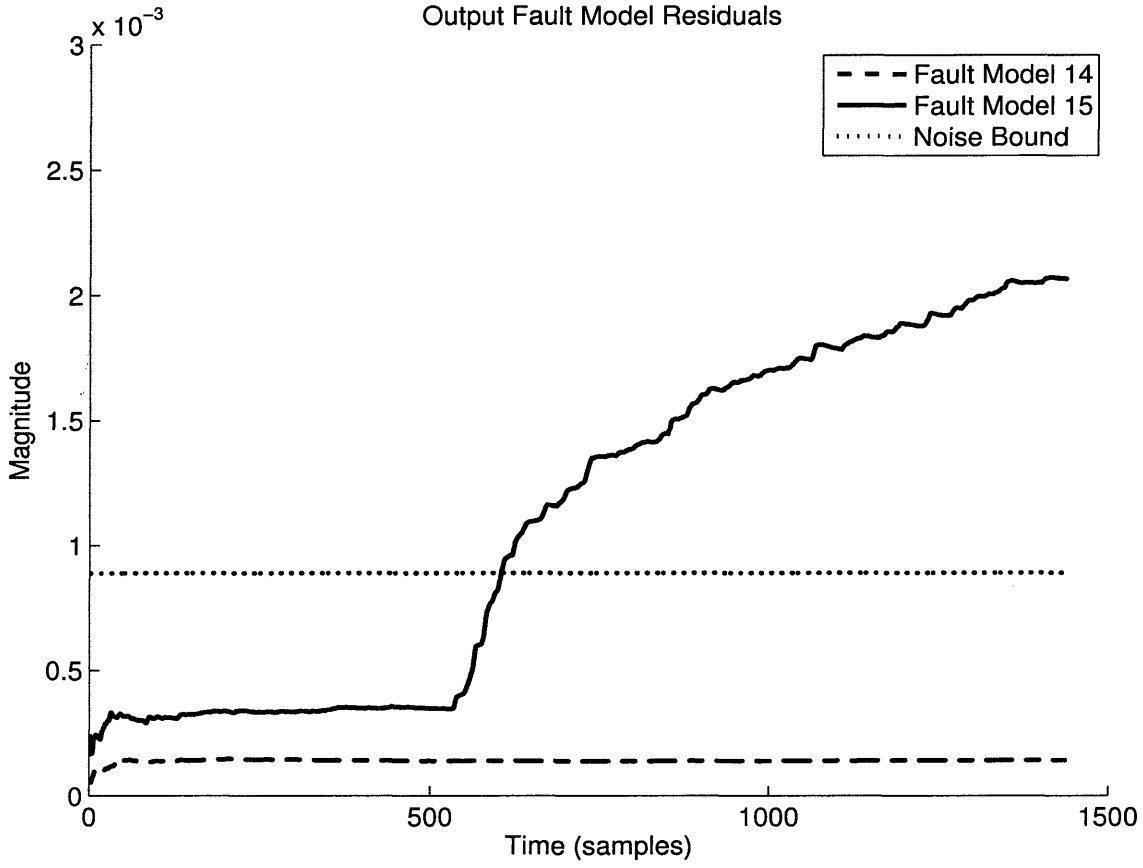

Figure 3.4. Residuals for two fault models capturing disturbance 14 and 15 respectively. Fault Model 14 stays below the noise bound indicating that this is the correct model for the fault that occurred in the system.

identification process requires solving a constrained least squares problem. Standard fast, recursive algorithms are available to solve these kinds of problems, but it remains to be seen if the specific structure of the shift equations could yield particulary efficient algorithms.

In contrast to this method, the next chapter looks at the case where the operator has some control over the input excitation. In this case we would like to choose the optimal input excitation for model parameter estimation and parameter variation fault detection.

Chapter 4

EXPERIMENT DESIGN FOR PARAMETER VARIATION FAULT DETECTION OF RATIO ESTIMATES

4.1 Introduction

In this chapter, we discuss methods to perform an open-loop experiment on a linear system for parameter identification and parameter variation fault detection. This system, as illustrated in Figure 4.1, has available an input for manipulation and measured output subject to noise. There are additional signals, that may or may not be actual measurements, that represent variables that must meet certain constraints. The results of this experiment will be applied to a linear estimator which produces parameter estimates. However, our true interest will be in combinations of the parameter estimates and performance variables, rather than their absolute values. Specifically, we will be interested in ratios or products of parameter estimates and performance variables. Our goal is to design an experiment which delivers the best ratio or product estimates for parameter variation fault detection, while minimizing the impact on constraint variables, which may also be expressed as ratios.

For this chapter we wish to focus on the thin film deposition system described

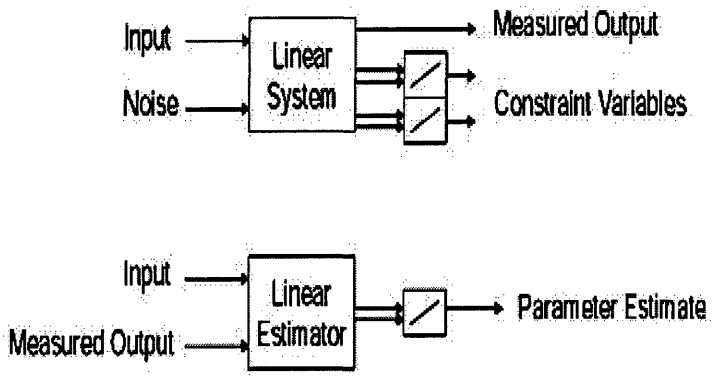

Figure 4.1. Input Optimization Problem. The linear system has an input for manipulation and a measured output. It is desired to have a linear estimator that, given the input and measured output, gives an estimate of a nonlinear combination of parameters, specifically either a ratio or product of the parameters.

in Section 2.8.2. In Hilt *et al.* (2006), the optimization of the system identification experiment for ratio estimates was discussed, but without considering ratios for the performance constraints. In this chapter, we will generalize the performance variables to ratios and products, as well as extend both the class of applicable systems and complexity of the input sequence generation.

Experiment design for parameter estimation of dynamic systems has been of interest for some time, in particular see Goodwin & Payne (1977) and Gevers (2005) and the references therein. In the standard case, the input is assumed to be small and operates over a long period of time, which allows asymptotic results on estimate distributions to be applied. The formulation is then in terms of minimizing the parameter estimate covariance. More recently, in Campbell & Nikoukhah (2004), the

problem was formulated as a multi-model trajectory separation problem. The smallest input is found that can guarantee separation of output trajectories over a finite set of possible models, subject to (indefinite) quadratic constraints on disturbances.

The problem considered here falls into the standard case, as a very small input (operating over a long period of time) fits the manufacturing environment well. The difference from previous work is the optimization with respect to nonlinear functions of the system parameters, although for a very specific case. However, as will be shown, the restriction to ratios allows solutions to be found using a convex optimization problem.

4.2 Problem Description

The problem considered here is to design or find an "optimal" input for parameter based fault detection when the fault condition is determined by a ratio or product of the parameters. The design problem will be solved by finding suitable metrics for both estimate quality and process upset, such that with an appropriate parametrization of the input, optimization of the estimate quality can occur while respecting constraints of the process variables. While it is possible to set up an optimization problem in which the input at each time t is a separate optimization variable, this approach does not lead to convex optimization problems, and it may be difficult to obtain an optimal solution. Therefore, an alternative method is used.

4.2.1 Spectral Factorization

The input will be described in terms of its autocorrelation. Given a random sequence u_k, the autocorrelation is defined to be

$$R_u(\tau) = E[u_k u_{k-\tau}^T]. \tag{4.1}$$

Since $R(\tau) = R(-\tau)$, only autocorrelations for $\tau \geq 0$ are needed to completely characterize the random sequence. In practice, $R_u(\tau) \to 0$ as $\tau \to \infty$. Therefore, the autocorrelation can be restricted to R_u for $0 \leq \tau \leq n-1$ where n is an appropriate number. The vector

$$\Phi_u = [\ R_u(0) \quad R_u(1) \quad \ldots \quad R_u(n-1)\ ] \tag{4.2}$$

is used as the parametrization of the input. This is similar to the approach taken in Wu *et al.* (1999) in the design of an input, however our design problem is quite different. Once the optimal Φ_u is obtained, the input sequence is produced to have this autocorrelation. One way to do this is to use an appropriate filter with a white noise input as show in Figure 4.2. Because $R_u(\tau) = 0$ for $\tau \geq n$, a Finite Impulse Response, or FIR, filter is used.

For this FIR filter, the desired input is u_k and w_k is a white noise sequence with

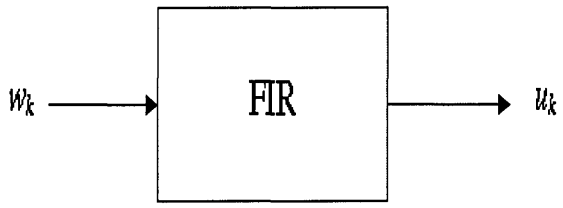

Figure 4.2. Finite Impulse Response Filter where w_k is a white noise sequence and u_k is the desired input produced by a FIR filter with the appropriate filter coefficients.

$E[w_k] = 0$ and $cov(w_k) = \lambda I$. The result of the FIR filter is written as

$$u_k = \sum_{i=1}^{n} F_i w_{k-i} \tag{4.3}$$

where F_i, for $i = 1, \ldots, n$, are called the filter coefficients. The process of determining these filter coefficients, F_i, to produce the desired autocorrelation is called spectral factorization.

Now, the autocorrelation of the input sequence can be parameterized by the filter coefficients

$$R_u(\tau) = E[u_k u_{k-\tau}^T] = E\left[\left(\sum_{i=1}^{n} F_i w_{k-i}\right)\left(\sum_{i=1}^{n} F_i w_{k-\tau-i}\right)\right] = \lambda \sum_{i=1}^{n-\tau} F_i F_{i+\tau}. \tag{4.4}$$

As long as there exists a way of decomposing the optimization variable Φ_u into the appropriate filter coefficients, F_i, an input sequence with the optimal autocorrelation can be determined.

One way to decompose the optimization variable into filter coefficients is to perform a Cholesky Decomposition on a block Toeplitz matrix containing the autocorrelation coefficients $R_u(\tau)$. This matrix is formed such that

$$T_N(\Phi_u) = \begin{bmatrix} R_u(0) & \cdots & R_u(1-n) & 0 & \cdots \\ \vdots & \ddots & & & \\ R_u(n-1) & & & & \\ 0 & & & & \\ \vdots & & & & \end{bmatrix}$$

where N determines the size of the matrix. The following is the well known Cholesky Decomposition Theorem.

Theorem 6. *(Cholesky Decomposition) If $A \in \mathbb{R}^{nxn}$ is symmetric positive definite, then there exists a unique lower triangular $G \in \mathbb{R}^{nxn}$ with positive diagonal entries such that $A = GG^T$ (Golub & Loan, 1989).*

The symmetric positive definite constraint placed on $T_N(\Phi_u)$ in order to use the Cholesky Decomposition ensures that there exists a set of filter coefficients that will produce an input with the autocorrelation represented by Φ_u. If this is the case, then it has been shown that the Cholesky Decomposition may be used to recover the filter coefficients (Wu *et al.*, 1999).

Theorem 7. *Given $T_N(\Phi_u)$, a block Toeplitz matrix containing the autocorrelation*

coefficients $R_u(\tau)$. If UU^T is the Cholesky decomposition of $T_N(\Phi_u)$ then the nonzero elements of the rows of U converge to the filter coefficients F_i as $N \to \infty$.

To Summarize: The input is parameterized by an autocorrelation vector, Φ_u, and is used as the optimization variable with a constraint on the block Toeplitz matrix, $T_N(\Phi_u)$. Then a Cholesky decomposition can be used on the optimal autocorrelation vector to determine the filter coefficients that will produce the optimal input sequence.

4.2.2 System Identification

The input to be designed is assumed to be applied to a linear system, resulting in a measured output y, subject to measurement noise e. In addition, we assume that a linear model is available of a signal z that characterizes the constraints that must be satisfied for the process to result in an acceptable product. Thus, we have

$$y_k = L_{yu}(\theta)u_k + e_k, \;\; z_k = L_{zu}u_k$$

where L_{yu} and L_{zu} are linear operators with L_{yu} parameterized by θ. This is the same as (2.18) with $A(q,\theta) = I$. After an experiment is performed, parameter estimates $\hat{\theta}$ of a linear model are obtained using least squares regression as in (2.19). Let P_θ =cov($\hat{\theta}$) and P_z = cov(z_k). Because we have assumed both a linear system and a linear estimate, P_z is an affine function of Φ_u and P_θ^{-1} is an affine function of Φ_u. We

represent these mappings as $P_\theta^{-1} = f(\Phi_u)$ and $P_z = g(\Phi_u)$ where

$$f(\Phi_u) = \frac{N}{\lambda}\begin{bmatrix} R_u(0) & R_u(1) & R_u(2) \\ R_u(1)^T & R_u(0) & R_u(1) \\ R_u(2)^T & R_u(1)^T & R_u(0) \end{bmatrix}, \text{ and}$$

$$g(\Phi_u) = \Delta_y \begin{bmatrix} R_u(0) & R_u(1) & R_u(2) \\ R_u(1)^T & R_u(0) & R_u(1) \\ R_u(2)^T & R_u(1)^T & R_u(0) \end{bmatrix} \Delta_y^T$$

with Δ_y containing the system impulse response. Therefore, we have an explicit relationship between the optimization variables, Φ_u, which contain the autocorrelation of the input and the covariance of the parameter estimates.

4.2.3 Quality Measures of Important Parameters and Signals

For the thin film process of interest, we have identified parameter and signal ratios as important for our application. The true performance of our experiment will thus be measured in terms of the distribution of $ratio(C_\theta\theta + d_\theta)$ and $ratio(C_z z + d_z)$, where C_z and C_θ are matrices mapping from the appropriate dimension to a length 2 vector, and $d_\theta, d_z \in \mathbb{R}^2$.

For some applications polynomial combinations of parameters and signals may be important. In particular we will look at products of parameters and signals. For

this case, the distribution of $prod(C_\theta\theta + d_\theta)$ and $prod(C_z z + d_z)$, where C_θ, C_z, d_θ, and d_z have the above definitions, will be a measure of the true performance of our experiment. In order to define either optimization problem, we need to establish metrics that measure the "spread" or width of distributions that are made up of these ratios or products. This is the subject of this section.

Ratios Throughout this chapter, we will explore measures of the width of the distribution for the ratio of two random variables, first introduced in Hilt *et al.* (2006). Let the joint probability distribution of the variables $x = \begin{bmatrix} x_1 & x_2 \end{bmatrix}^T$ be $f_x(x_1, x_2)$. If r =ratio$(x) := \frac{x_1}{x_2}$, then the probability density function of r will follow the ratio distribution given by

$$f_r(r) = \int_{-\infty}^{\infty} |x_2| f_x(rx_2, x_2)\ \mathrm{d}x_2. \tag{4.5}$$

We would like to choose input excitations that result in the distribution given by (4.5) being narrow. A method to evaluate the narrowness of these distributions must be established first. Normally the variance of a distribution would be used to gauge it's width. However, when x is jointly Gaussian, because the tail of this distribution decreases slowly, second moments of this distribution do not exist, and the variance is infinite. We propose an alternate characterization of distribution width based on an ellipsoidal bound.

For convenience, we define the set of points satisfying an ellipsoidal constraint

as follows

Definition 8. *Let*

$$\mathcal{E}(S, \bar{s}, \delta) = \left\{ s | (s - \bar{s})^T S (s - \bar{s}) \leq \delta^2 \right\}.$$

If $\delta > 0$ and $S \in \mathbb{S}_+^n$, then $\mathcal{E}(S, \bar{s}, \delta)$ is an ellipse (Ros *et al.*, 2002). Suppose that it is known that the parameters, s, lie inside an ellipse, that is, satisfy

$$(s - \bar{s})^T S (s - \bar{s}) \leq \delta^2 \tag{4.6}$$

for some $\delta > 0$ and $S \in \mathbb{S}_+^n$. Such a hard constraint can be derived using an a-priori bound on the measurement noise, or it can be given a stochastic interpretation. For example, in the latter case, if $n = 2$ and s is a Gaussian random vector, with covariance $P = S^{-1}$ and mean $\bar{s}$, then s satisfies this constraint with probability $1 - \exp\left(-\frac{\delta^2}{2}\right)$ (Van Trees, 1968, p. 77). Thus δ defines the level of confidence in the estimate that is desired. If δ is chosen to be 2, then the probability that s satisfies the constraint and therefore lies inside the ellipse is 0.86. We will assume that a particular choice of $\delta > 0$ has been made, and it will be fixed for the remainder of the chapter.

We will be interested in characterizing the distribution for ratios of the form

$ratio(x)$ where $x = Cs + d$, $C \in \mathbb{R}^{2\times n}$, and $d \in \mathbb{R}^2$. In order for the ratio to be well defined, we require that the denominator be positive. We define the set in which all ratios have a positive denominators as $\mathcal{D}_{C,d}$.

Definition 9. *Let $\mathcal{D}_{C,d}$ be the set of $(S, \bar{s}, \delta)$ such that $[Cs + d]_2 > 0$ for all $s \in \mathcal{E}(S, \bar{s}, \delta)$, where $[x]_2$ is the second element of the vector x.*

The following well known result is a useful characterization of points inside an ellipse.

Lemma 10. *If S is invertible, $s \in \mathcal{E}(S, \bar{s}, \delta)$ if and only if $s = \delta S^{-\frac{1}{2}} v + \bar{s}$ for some $\|v\| \leq 1$.*

Proof. If $s \in \mathcal{E}(S, \bar{s}, \delta)$, by definition

$$(s - \bar{s})^T S(s - \bar{s}) \leq \delta^2 \tag{4.7}$$

Let $s = \delta S^{-1/2} v + \bar{s}$ for some $\|v\| \leq 1$, then $s \in \mathcal{E}(S, \bar{s}, \delta)$ if and only if

$$\begin{aligned} (\delta S^{-1/2} v + \bar{s} - \bar{s})^T S(\delta S^{-1/2} v + \bar{s} - \bar{s}) &\leq \delta^2 && (4.8) \\ (\delta S^{-1/2} v)^T S(\delta S^{-1/2} v) &\leq \delta^2 && (4.9) \\ v^T (S^{-1/2})^T \delta^T S \delta S^{-1/2} v &\leq \delta^2 && (4.10) \end{aligned}$$

$S^{-1/2}$ is symmetric and δ is scalar so $S^{-1/2} = (S^{-1/2})^T$ and $\delta = \delta^T$. So we can simplify

(4.10) as:

$$\begin{aligned} v^T S^{-1/2} \delta S \delta S^{-1/2} v &\leq \delta^2 \\ \delta^2 v^T S^{-1/2} S S^{-1/2} v &\leq \delta^2 \end{aligned}$$

It can be seen that $S^{-1/2} S S^{-1/2} = I$. Therefore

$$\begin{aligned} \delta^2 v^T v &\leq \delta^2 \\ v^T v &\leq 1 \\ \|v\| &\leq 1 \end{aligned}$$

□

Using the ellipsoidal bounds on the parameters s, as in Definition 8, the following measures for the width of the distribution of $ratio(Cs + d)$ can be defined as:

Definition 11. *For* $(S, \bar{s}, \delta) \in \mathcal{D}_{C,d}$ *Let*

$$\phi^{C,d}_{\max}(S, \bar{s}, \delta) = \begin{cases} \max & ratio(Cs + d) \\ \text{subject to} & s \in \mathcal{E}(S, \bar{s}, \delta) \end{cases} \tag{4.11}$$

$$\phi^{C,d}_{\min}(S, \bar{s}, \delta) = \begin{cases} \min & ratio(Cs + d) \\ \text{subject to} & s \in \mathcal{E}(S, \bar{s}, \delta) \end{cases} \tag{4.12}$$

$$\phi^{C,d}(S, \bar{s}, \delta) = \phi^{C,d}_{\max}(S, \bar{s}, \delta) - \phi^{C,d}_{\min}(S, \bar{s}, \delta) \tag{4.13}$$

Let us consider an example of an ellipsoidal representation of the width of a probability density function of $ratio(Cs+d)$ with the following properties. Let $C = I_n$ and $d = 0$. Assume the covariance matrix of the parameters is

$$P = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}^{-1} \tag{4.14}$$

and the parameter used to describe the shape of the ellipse, S, is equal to P^{-1}. Let the mean be $\bar{s} = \begin{bmatrix} 1 & 1 \end{bmatrix}^T$. The maximum and minimum points are illustrated in Figure 4.3 for $\sigma_1 = 0.1, \sigma_2 = 0.15$ and $\rho = 0$ and in Figure 4.4 for $\sigma_1 = 0.1, \sigma_2 = 0.15$ and $\rho = 0.5$. The top graph in each figure shows the ellipse $\mathcal{E}$ for $\delta = 2$ (probability of 0.86) along with the locations of the minimizer and maximizer of $ratio(s)$, indicated by $s_{\max}$ and $s_{\min}$. The dashed lines denote values of s which also share the same ratio as the maximum or minimum ratio in $\mathcal{E}$. The bottom graph in each figure shows the corresponding ratio distribution if the random variables were jointly Gaussian, along with the value of ϕ, and the locations of $ratio(s_{\max})$ and $ratio(s_{\min})$. Note how the bounds from the elliptical representation in the top graph reasonably reflect the relative breadth of the ratio distributions as shown in the second graph. It is also interesting to note how the correlation of the two parameters effect the width of the distribution. While both Figure 4.3 and Figure 4.4 have the same variances σ_1 and σ_2, Figure 4.4 represents a correlation between the two parameter estimates causing

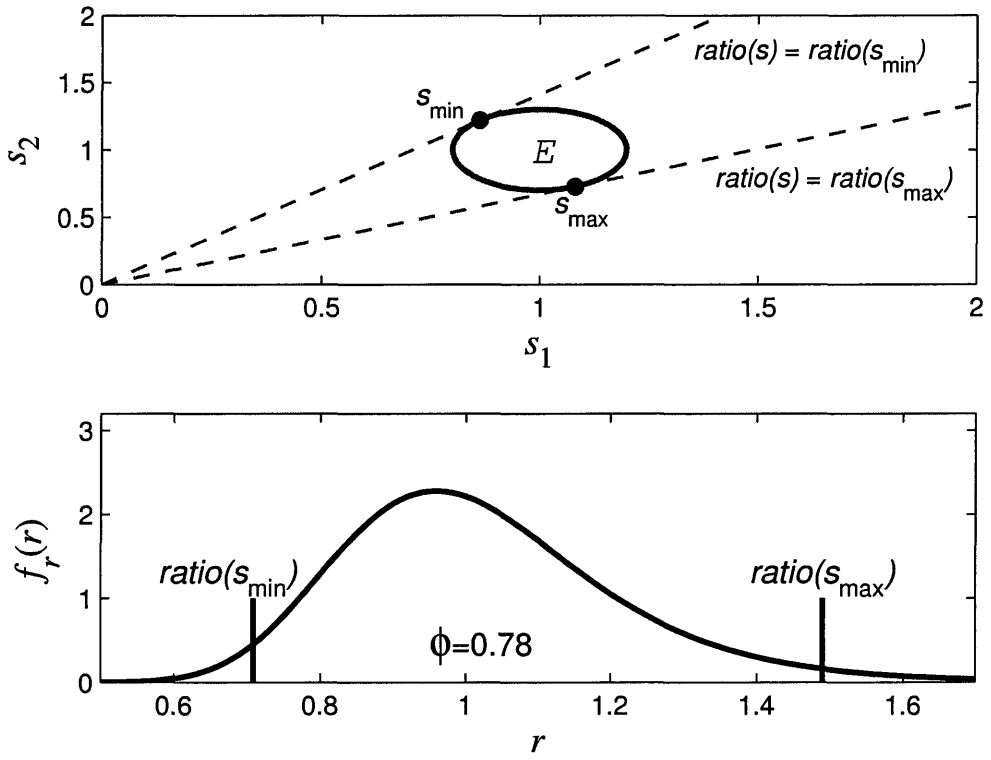

Figure 4.3. Ratio distribution with uncorrelated variables. The top figure shows the ellipse for $\delta = 2$, the mean $\bar{s} = [1 \quad 1]^T$, and P as in equation 4.14, with $\sigma_1 = 0.1, \sigma_2 = 0.15$ and $\rho = 0$. The bottom graph shows the ratio distribution if the random variables are jointly Gaussian with the bounds from the elliptical representation marked.

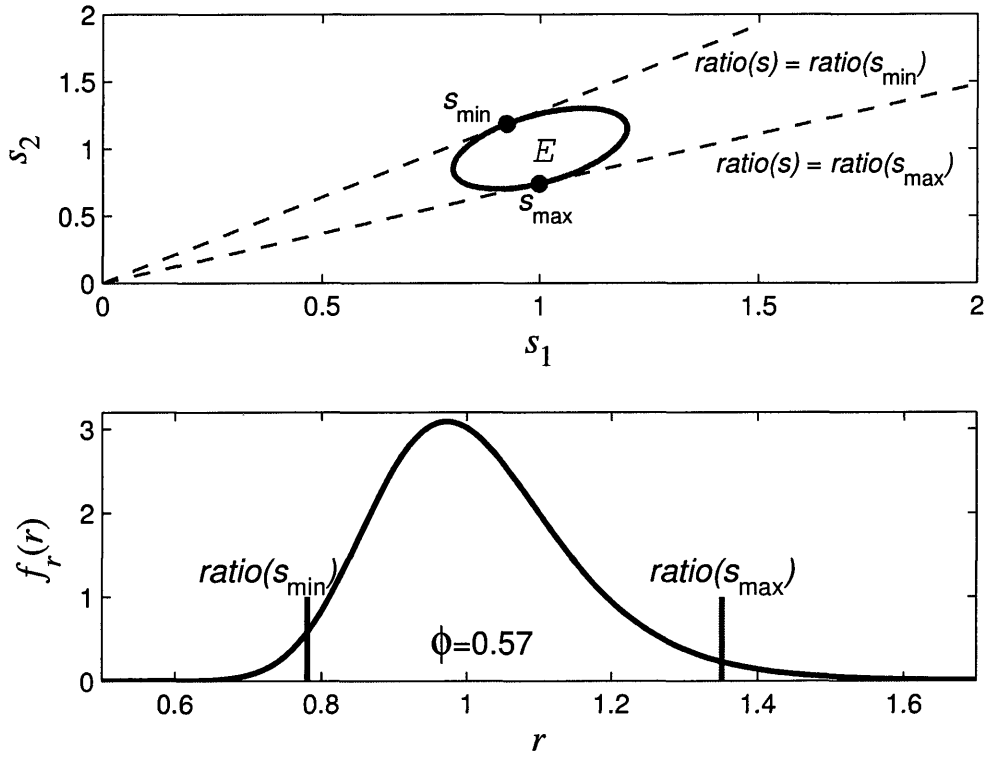

Figure 4.4. Ratio distribution with correlated variables. The top figure shows the ellipse for $\delta = 2$, the mean $\bar{s} = [1 \quad 1]^T$, and P as in equation 4.14, with $\sigma_1 = 0.1, \sigma_2 = 0.15$ and $\rho = 0.5$. The bottom graph shows the ratio distribution if the random variables are jointly Gaussian with the bounds from the elliptical representation marked.

the distribution to be narrower than the distribution in Figure 4.3. Now that we have defined an alternative characterization of the width of a ratio distribution based on an ellipsoidal bound, we would like to take a corresponding approach for product distributions.

Products When looking at products of parameters a similar approach to defining an alternate characterization of the distribution width can be used. As in the previous section, let the joint probability of the variables $x = [x_1 \;\; x_2]^T$ be $f_x(x_1, x_2)$. If $p = prod(x) := x_1 x_2$, then the probability density function of p will follow the product distribution given by

$$f_p(p) = \int_{-\infty}^{\infty} \frac{1}{|x_2|} f_x\left(\frac{p}{x_2}, x_2\right) dx_2. \tag{4.15}$$

Again we will use ellipsoidal bounds on the parameter s, to derive the measures for the width of the distribution of the $prod(Cs + d)$. These measures are defined as follows:

Definition 12. *For* $(S, \bar{s}, \delta) \in \mathcal{D}_{C,d}$ *Let*

$$\begin{aligned} \psi^{C,d}_{\max}(S, \bar{s}, \delta) \\ \psi^{C,d}_{\max}(S, \bar{s}, \delta) &= \begin{cases} \max & prod(Cs + d) \\ \text{subject to} & s \in \mathcal{E}(S, \bar{s}, \delta) \end{cases} \end{aligned} \tag{4.16}$$

$$\psi^{C,d}_{\min}(S, \bar{s}, \delta) = \begin{cases} \min & prod(Cs + d) \\ \text{subject to} & s \in \mathcal{E}(S, \bar{s}, \delta) \end{cases} \tag{4.17}$$

$$\psi^{C,d}(S, \bar{s}, \delta) = \psi^{C,d}_{\max}(S, \bar{s}, \delta) - \psi^{C,d}_{\min}(S, \bar{s}, \delta) \tag{4.18}$$

As in the previous section, consider an example of an ellipsoidal representation of the width of a probability density function of $prod(Cs + d)$ with the same properties as before including the covariance matrix of the parameters (4.14). The maximum and minimum points are illustrated in Figure 4.5 for $\sigma_1 = 0.1$, $\sigma_2 = 0.15$, and $\rho = 0$ and in Figure 4.6 for $\sigma_1 = 0.1$, $\sigma_2 = 0.15$, and $\rho = 0.5$. It is important to that note that the bounds from the ellipsoidal representation are a reasonable measure of the width of the product distributions shown in the bottom graph of each figure.

4.3 Analysis

The measures for the width of the distribution of $ratio(Cx+d)$ and $prod(Cs+d)$ established in the previous section have very nice properties that will be important in the optimization process. In particular, it will be shown in this section that these measures have properties which allow them to be easily computable.

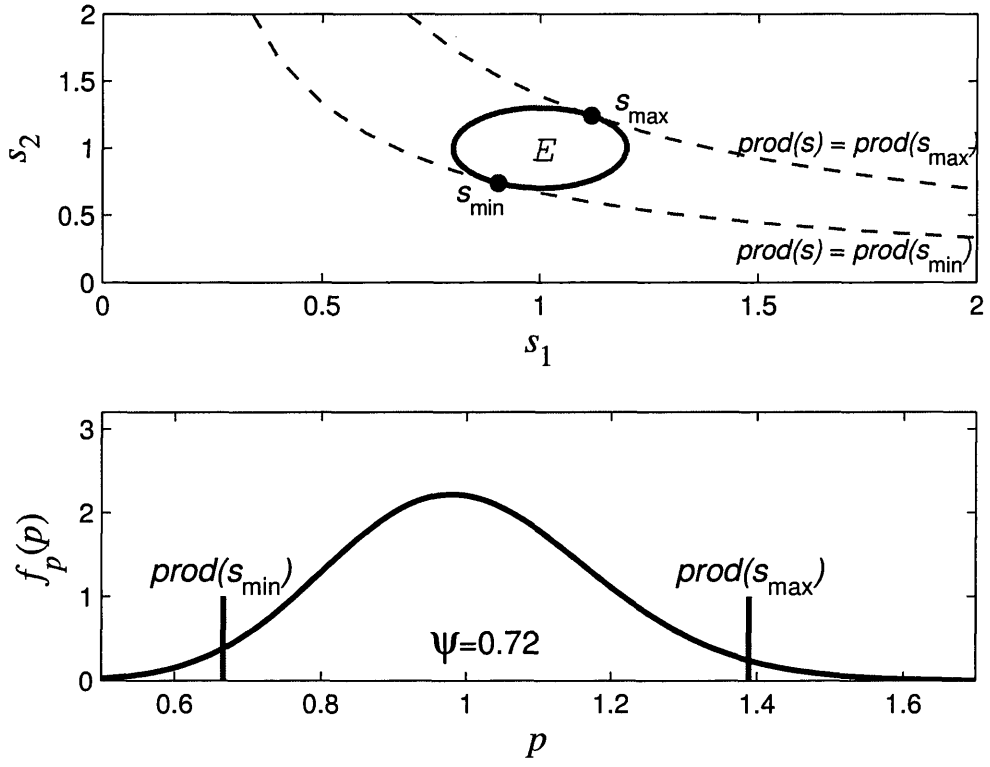

Figure 4.5. Product distribution with uncorrelated variables. The top figure shows the ellipse for $\delta = 2$, the mean $\bar{s} = [1 \quad 1]^T$, and P as in equation 4.14, with $\sigma_1 = 0.1, \sigma_2 = 0.15$ and $\rho = 0$. The bottom graph shows the product distribution if the random variables are jointly Gaussian with the bounds from the elliptical representation marked.

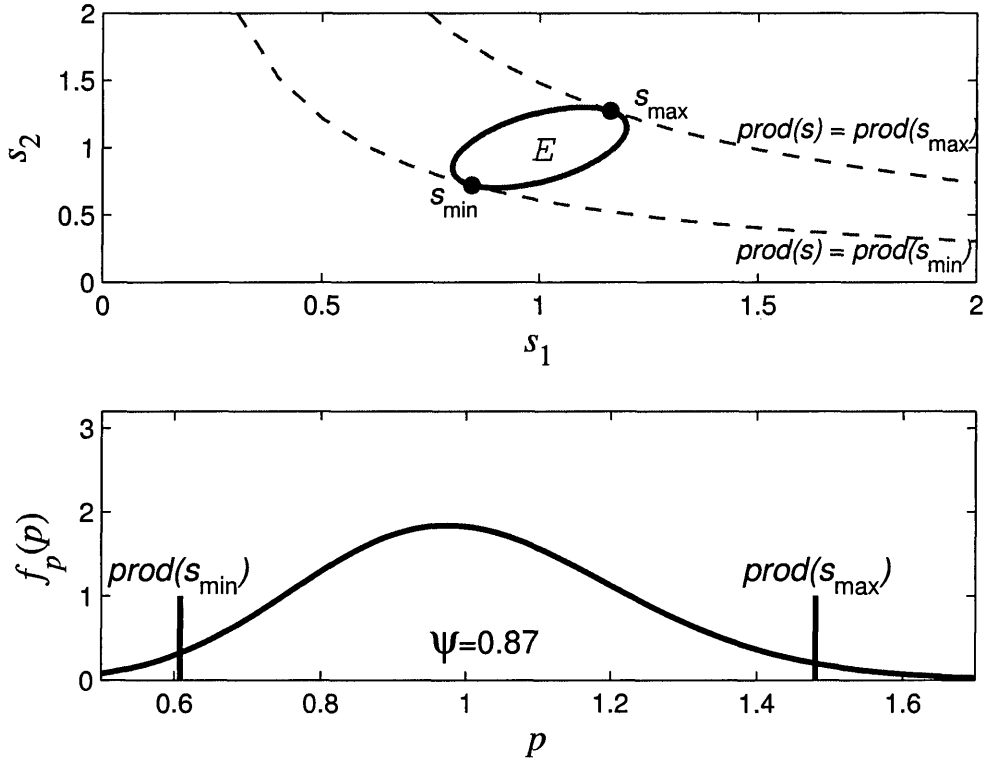

Figure 4.6. Product distribution with correlated variables. The top figure shows the ellipse for $\delta = 2$, the mean $\bar{s} = [1 \quad 1]^T$, and P as in equation 4.14, with $\sigma_1 = 0.1, \sigma_2 = 0.15$ and $\rho = 0.5$. The bottom graph shows the product distribution if the random variables are jointly Gaussian with the bounds from the elliptical representation marked.

4.3.1 Ratios

In this section we will establish the following results related to the width measures for ratios:

- Elements of an ellipse can be mapped to a half-space using a general affine function.
- Combining ellipse parameters can result in an ellipse in the same halfspace.
- The set of ellipses in a halfspace with the same center is convex.

We begin with a general lemma that gives a necessary and sufficient condition for the elements of $\mathcal{E}(S, \bar{s}, \delta)$ to be mapped by a general affine function into a particular half-space in $\mathbb{R}^m$. We will later specialize to $m = 2$, but these results are true in general.

Lemma 13. *Given* $S \in \mathbb{S}_+^n$, $\bar{s} \in \mathbb{R}^n$, $\delta > 0$, $C \in \mathbb{R}^{m \times n}, d \in \mathbb{R}^m$, *and halfspace* $\mathcal{H}$ *in* $\mathbb{R}^m$ *defined by* $\mathcal{H} = \left\{x \in \mathbb{R}^m : c^T x \geq a\right\}$. *If* $C\bar{s} + d \in \mathcal{H}$, *then the following are equivalent:*

1. $Cs + d \in \mathcal{H}\ \forall s \in \mathcal{E}(S, \bar{s}, \delta)$
2. $\left(c^T(C\bar{s} + d) - a\right)^2 \geq \delta^2 c^T C S^{-1} C^T c$

Proof. $1 \Rightarrow 2$: Given that $c^T(Cs+d) \geq a$ and $s = \delta S^{-1/2}v + \bar{s}$ for $\|v\| \leq 1$, then

$$\begin{aligned} Cs &= C\delta S^{-1/2}v + C\bar{s} \\ Cs+d &= C\delta S^{-1/2}v + C\bar{s} + d \\ c^T(Cs+d) &= c^TC\delta S^{-1/2}v + c^TC\bar{s} + c^Td. \end{aligned}$$

From $c^T(Cs+d) \geq a$,

$$\begin{aligned} c^TC\delta S^{-1/2}v + c^TC\bar{s} + c^Td &\geq a \\ \delta c^TCS^{-1/2}v + c^T(C\bar{s}+d) &\geq a \\ c^T(C\bar{s}+d) - a &\geq -\delta c^TCS^{-1/2}v \end{aligned}$$

for $\|v\| \leq 1$. Let $v = \frac{-\delta S^{-1/2}C^Tc}{\|\delta S^{-1/2}C^Tc\|}$, so that

$$\begin{aligned} c^T(C\bar{s}+d) - a &\geq -\delta c^TCS^{-1/2}\left(\frac{-\delta S^{-1/2}C^Tc}{\|\delta S^{-1/2}C^Tc\|}\right) \\ c^T(C\bar{s}+d) - a &\geq \frac{\delta^2 c^TCS^{-1}C^Tc}{\|\delta S^{-1/2}C^Tc\|} \end{aligned}$$

By definition $\|\delta S^{-1/2}C^Tc\| = \sqrt{\delta^2 c^TCS^{-1}C^Tc}$, therefore

$$\begin{aligned} c^T(C\bar{s}+d) - a &\geq \sqrt{\delta^2 c^TCS^{-1}C^Tc} \\ (c^T(C\bar{s}+d) - a)^2 &\geq \delta^2 c^TCS^{-1}C^Tc \end{aligned}$$

$2 \Rightarrow 1$: By contradiction. If we assume that

$$\begin{aligned} \left(c^T(C\bar{s}+d)-a\right)^2 &\geq \delta^2 c^T C S^{-1} C^T c \\ c^T(C\bar{s}+d)-a &\geq \delta\sqrt{c^T C S^{-1} C^T c} \end{aligned} \tag{4.19}$$

but $c^T(Cs+d) \notin \mathcal{H} \forall s \in E(S,\bar{s},\delta)$, implying that $c^T(Cs+d) < a$ for some $s = \delta S^{-1/2} v + \bar{s}$ with $\|v\| \leq 1$. Then

$$\begin{aligned} c^T(C\delta S^{-1/2}v + C\bar{s} + d) &< a \\ c^T(C\bar{s}+d)-a &< -c^T C\delta S^{-1/2} v \end{aligned} \tag{4.20}$$

Combining (4.19) and (4.20) we get

$$\delta\sqrt{c^T C S^{-1} C^T c} \leq c^T(C\bar{s}+d)-a < -c^T C\delta S^{-1/2} v$$

$$\sqrt{c^T C S^{-1} C^T c} < -c^T C S^{-1/2} v$$

for $\|v\| \leq 1$. Choose $v = \frac{-S^{-1/2}C^T c}{\|S^{-1/2}C^T c\|}$ giving:

$$\begin{aligned} \sqrt{c^T C S^{-1} C^T c} &< -c^T C S^{-1/2}\left(\frac{-S^{-1/2}C^T c}{\|S^{-1/2}C^T c\|}\right) \\ \sqrt{c^T C S^{-1} C^T c} &< \frac{c^T C S^{-1} C^T c}{\|S^{-1/2}C^T c\|} \\ \sqrt{c^T C S^{-1} C^T c} &< \sqrt{c^T C S^{-1} C^T c} \end{aligned}$$

which is a contradiction. □

The result of this lemma is that we now have a way to test whether an ellipse lies within a given halfspace. This result has some interesting corollaries. For example, the following result shows that when the parameters of multiple ellipses are combined via a convex sum to create a new ellipse, this new ellipse remains in a half space common to the primary ellipses, as long as the change in the ellipse centers is parallel to the hyperplane defining the half space.

Corollary 14. *Given ellipses* $\mathcal{E}_i(S_i, \bar{s}_i, \delta)$ $i = 1, \cdots, k$ *and half space* $\mathcal{H}$. *If*

- $\mathcal{E}_i(S_i, \bar{s}_i, \delta) \subset \mathcal{H}$ $\forall i$
- $c^T C(\bar{s}_i - \bar{s}_j) = 0$ $\forall i, j$

then if $S(\gamma) = \sum_{i=1}^{k} \gamma_i S_i$ *and* $\bar{s}(\gamma) = \sum_{i=1}^{k} \gamma_i \bar{s}_i$ *with* $\sum_{i=1}^{k} \gamma_i = 1$, *the ellipse* $\mathcal{E}_i(S(\gamma), \bar{s}(\gamma), \delta) \subset \mathcal{H}$.

Proof. First, note that $c^T C \bar{s}_i$ is constant for all ellipses and equal to $c^T C \bar{s}(\gamma)$. Then, since $S_i > 0$, by Schur complements the condition

$$\left(c^T(C\bar{s}_i + d) - a\right)^2 \geq \delta^2 c^T C S_i^{-1} C^T c$$

from Lemma 13 is equivalent to

$$\begin{bmatrix} S_i & \delta C^T c \\ \delta c^T C & \left(c^T(C\bar{s}_i + d) - a\right)^2 \end{bmatrix} \geq 0,$$

(see Boyd *et al.* (1994)) and clearly is also satisfied by $S(\gamma)$, $\bar{s}(\gamma)$ whenever $c^T C(\bar{S}(\gamma) - \bar{s}_i) = 0$. □

Remarkably, the same is true even if the ellipses are parameterized using $P = S^{-1}$ instead.

Corollary 15. *Given ellipses* $\mathcal{E}_i(P_i^{-1}, \bar{s}_i, \delta)$ $i = 1, \cdots, k$ *and half space* $\mathcal{H}$. *If*

- $\mathcal{E}_i(P_i^{-1}, \bar{s}_i, \delta) \subset \mathcal{H}$ $\forall i$
- $c^T C(\bar{s}_i - \bar{s}_j) = 0$ $\forall i, j$

then if $P(\gamma) = \sum_{i=1}^k \gamma_i P_i$, *and* $\bar{s}(\gamma) = \sum_{i=1}^k \gamma_i \bar{s}_i$, *with* $\sum_{i=1}^k \gamma_i = 1$, *the ellipse* $\mathcal{E}_i(P(\gamma)^{-1}, \bar{s}(\gamma), \delta) \subset \mathcal{H}$.

Proof. First, note that $c^T C\bar{s}_i$ is constant for all ellipses and equal to $c^T C\bar{s}(\gamma)$. Then, if $\mathcal{E}_i \subset \mathcal{H}$ $\forall i$,

$$\begin{aligned} \delta^2 c^T C P(\gamma) C^T c &= \delta^2 \sum_{i=1}^k \gamma_i c^T C P_i C^T c \\ &\leq \sum_{i=1}^k \gamma_i \left(c^T C\bar{s}_i + d - a\right)^2 \\ &= \left(c^T C\bar{s}(\gamma) + d - a\right)^2 \end{aligned}$$

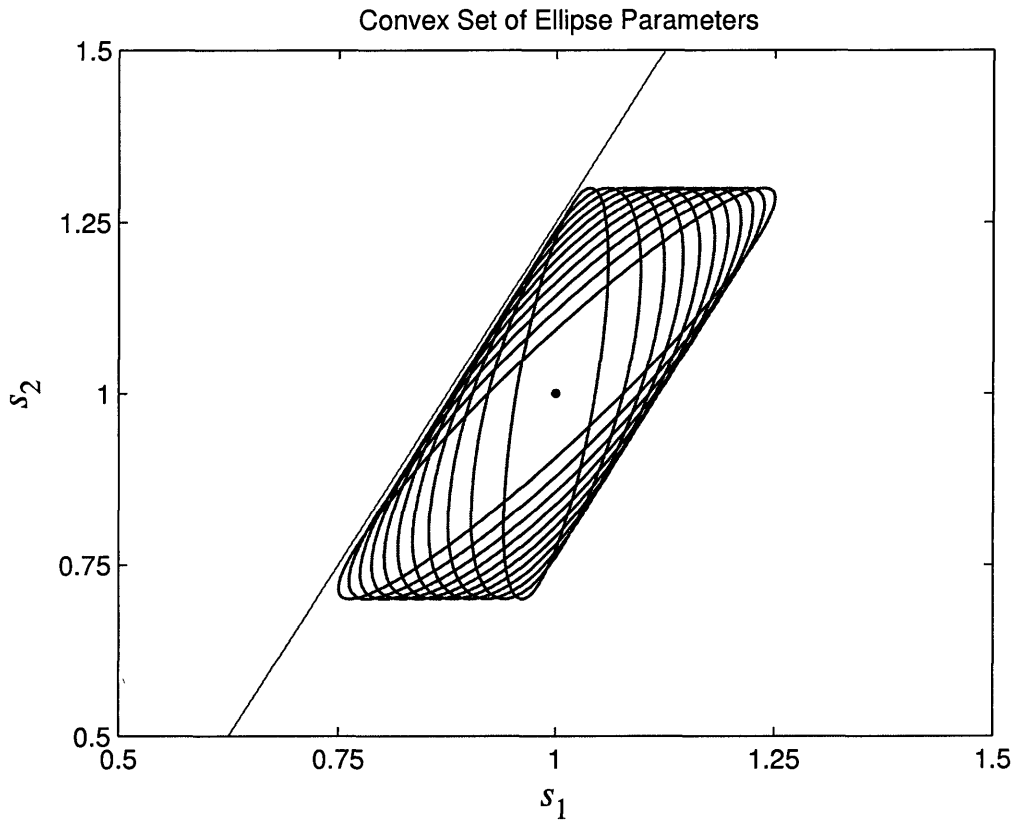

Figure 4.7. The set of convex combinations of ellipses remains on the same side of a common half-space, therefore the set of all ellipse parameters for which the ellipses remain inside a particular half-space forms a convex set.

Thus, $P(\gamma)^{-1}$, $\bar{s}(\gamma)$ satisfy the conditions of Lemma 13, and the ellipse $\mathcal{E}_i(P(\gamma)^{-1}, \bar{s}(\gamma), \delta)$ is contained in $\mathcal{H}$. □

Now, with $\bar{s}$ fixed, since "convex combinations" of ellipses remain on the same side of a common half-space, the set of all ellipse parameters for which the ellipses remain inside a particular half-space forms a convex set as shown in Figure 4.7.

Corollary 16. *Given halfspace $\mathcal{H}$ and fixed $\bar{s}$ with $C\bar{s}+d \in \mathcal{H}$, the set of S such that $Cs+d \in \mathcal{H}$ for all $s \in \mathcal{E}(S,\bar{s},\delta)$ is convex. Furthermore, the set of P such that $Cs+d \in \mathcal{H}$ for all $s \in \mathcal{E}(P^{-1},\bar{s},\delta)$ is convex.*

Proof. Follows from Corollaries 14 and 15. □

We are now in a position to establish the main results concerning the ϕ measures of the ratio distribution. Note that for these results, we are specializing to $C \in \mathbb{R}^{2\times n}$ and $d \in \mathbb{R}^2$.

Theorem 17. *For $(S,\bar{s},\delta) \in \mathcal{D}_{C,d}$, let $Q = \left(CS^{-1}C^T\right)^{-1}$ and $\gamma = \left\|Q^{\frac{1}{2}}(C\bar{s}+d)\right\|$. Then*

$$\phi_{\max}^{C,d}(S,\bar{s},\delta) = ratio(\delta Q^{-1}w_+ + C\bar{s} + d) \tag{4.21}$$

$$\phi_{\min}^{C,d}(S,\bar{s},\delta) = ratio(\delta Q^{-1}w_- + C\bar{s} + d) \tag{4.22}$$

where

$$w_\pm = -Q^{\frac{1}{2}}\begin{bmatrix} \delta & \mp\sqrt{\gamma^2-\delta^2} \\ \pm\sqrt{\gamma^2-\delta^2} & \delta \end{bmatrix}\frac{Q^{\frac{1}{2}}(C\bar{s}+d)}{\gamma^2}. \tag{4.23}$$

Furthermore,

- *$\phi_{\max}^{C,d}(S,\bar{s},\delta)$, $-\phi_{\min}^{C,d}(S,\bar{s},\delta)$, and $\phi^{C,d}(S,\bar{s},\delta)$ are convex functions of S.*
- *$\phi_{\max}^{C,d}(P^{-1},\bar{s},\delta)$ and $\phi_{\min}^{C,d}(P^{-1},\bar{s},\delta)$ are quasi-linear functions of P.*

Proof. See Appendix B. □

4.3.2 Products

As in the case of ratios, the measures for the width of the distribution of the $prod(Cs+d)$, ψ_{max}, ψ_{min}, and ψ have nice properties. Due to the polynomial properties of $prod(Cs+d)$, we can find ψ_{min} and ψ_{max} using a sequence of convex linear matrix inequality (LMI) problems as presented in Lasserre (2001). The following is an overview of this method.

We state the problem addressed in Lasserre (2001) as this: Given a real-valued polynomial, $p(x)$, we wish to find the global minimum and corresponding minimizer. In Lasserre (2001) it was shown that advantage can be taken of convex optimization methods by representing the minimization problem as a sequence of convex LMI optimization problems. We will start by introducing some notation.

Notation Define $p(x) : \mathbb{R}^n \rightarrow \mathbb{R}$ to be a real-valued polynomial mapping a vector of length n to a scalar. The order of the polynomial is m. The polynomial can also be represented as a sum of terms such that

$$p(x) = \sum_i p_i x^{\alpha_i} \tag{4.24}$$

with the following definitions. The exponent α_i is a vector $\begin{bmatrix} \alpha_{i,1} & \alpha_{i,2} & \cdots & \alpha_{i,n} \end{bmatrix}$ such that $x^{\alpha_i} = x_1^{\alpha_{i,1}} x_2^{\alpha_{i,2}} \ldots x_n^{\alpha_{i,n}}$ and $\sum_{j=1}^{n} \alpha_{i,j} \leq m$. The vector $p = p_i$ is the coefficient vector of $p(x)$. The vector $y = y_{\alpha_i}$, is called the moment vector where, for the two

dimensional case, $y_{i,j}$ is the $(i+j)$-order moment $\int x^i y^j \mu(d(x,y))$. The first element $y_{\alpha_0} = y_{0,\ldots,0} = 1$. $M_m(y)$ is the block moment matrix $M_{i,j}(y)_{0 \le i,j \le m}$ defined by

$$M_{i,j}(y) = \begin{bmatrix} y_{i+j,0} & y_{i+j-1,1} & \cdots & y_{i,j} \\ y_{i+j-1,1} & y_{i+j-2,2} & \cdots & \vdots \\ \vdots & \vdots & \cdots & \vdots \\ y_{j,i} & \cdots & \cdots & y_{0,i+j} \end{bmatrix}. \tag{4.25}$$

$M_m(y)$ can also be represented as $(\sum_{\alpha \neq 0} y_\alpha B_\alpha) + B_o$ where B_α and B_0 are the appropriate basis matrices.

Unconstrained Sum of Squares As shown in Lasserre (2001), the optimization problem

$$\mathbb{P} \mapsto p^* = \min_{x \in \mathbb{R}^n} p(x) \tag{4.26}$$

can be written as the LMI problem

$$\mathbb{Q} \mapsto \begin{cases} inf_y \sum_\alpha p_\alpha y_\alpha \\ M_m(y) \succeq 0 \end{cases} \tag{4.27}$$

or

$$\mathbb{Q} \mapsto \begin{cases} inf_y \sum_\alpha p_\alpha y_\alpha \\ \sum_{\alpha \neq 0}^{\alpha} y_\alpha B_\alpha \succeq -B_0 \end{cases} \tag{4.28}$$

Given a global minimum, $p^* = \min p(x)$, if $p(x) - p^*$ is a sum of squares of polynomials, then min $\mathbb{P}$ is equivalent to the convex LMI problem $\mathbb{Q}$. Therefore, by solving $\mathbb{Q}$ the global minimum of the polynomial $p(x)$ can be found.

Constrained General Polynomial Case An appropriate LMI problem can also be found to solve a constrained minimization problem such as

$$\mathbb{P}_K \mapsto p_K^* := \min_{X \in K} p(x) \tag{4.29}$$

where the following definitions apply. The function $p(x)$ is a real-valued polynomial mapping $\mathbb{R}^n \to \mathbb{R}$ with an order of m. K is a set of constraint inequalities of the form $g_i(x) \geq 0$ where $g_i(x) : \mathbb{R}^n \to \mathbb{R}$ is a real-valued polynomial. The degree of $g_i(x)$ is w_i. For each constraint define the matrix

$$M_m(g_i y) = g_i(x) M_m(y) = \begin{bmatrix} g(x) * M_m(y)(1,1) & \cdots & g(x) * M_m(y)(1,n) \\ \cdots & \ddots & \\ g(x) * M_m(y)(n,1) & & \end{bmatrix}. \tag{4.30}$$

For example, in the two dimensional case where $m = 2$ and $n = 2$,

$$M_2(y) = \begin{bmatrix} 1 & y_{1,0} & y_{0,1} & y_{2,0} & y_{1,1} & y_{0,2} \\ y_{1,0} & y_{2,0} & y_{1,1} & y_{3,0} & y_{2,1} & y_{1,2} \\ y_{0,1} & y_{1,1} & y_{0,2} & y_{2,1} & y_{1,2} & y_{0,3} \\ y_{2,0} & y_{3,0} & y_{2,1} & y_{4,0} & y_{3,1} & y_{2,2} \\ y_{1,1} & y_{2,1} & y_{1,2} & y_{3,1} & y_{2,2} & y_{1,3} \\ y_{0,2} & y_{1,2} & y_{0,3} & y_{2,2} & y_{1,3} & y_{0,4} \end{bmatrix}. \tag{4.31}$$

Given the constraint polynomial $g(x) = ax_1 + bx_2$ then

$$M_m(g(x)y) = (ax_1 + bx_2).M_m(y) \tag{4.32}$$

$$= \begin{bmatrix} g(x) * 1 & g(x) * y_{1,0} & \cdots \\ g(x) * y_{1,0} & g(x) * y_{2,0} & \cdots \\ \vdots & & \end{bmatrix} \tag{4.33}$$

$$= \begin{bmatrix} ay_{1,0} + by_{0,1} & ay_{2,0} + by_{1,1} & \cdots \\ ay_{2,0} + by_{1,1} & ay_{3,0} + by_{2,1} & \cdots \\ \vdots & & \end{bmatrix} \tag{4.34}$$

Now write the LMI problem

$$\mathbb{Q}_K^N \begin{cases} \inf_y \sum_\alpha p_\alpha y_\alpha, \\ M_n(y) \succeq 0, \\ M_{N-\widetilde{w}_i}(g_i y) \succeq 0, i = 1, \ldots, r \end{cases} \tag{4.35}$$

where $\widetilde{w}_i := \lceil w_i/2 \rceil$, $N \geq \max_i \widetilde{w}_i$, and $N \geq \lceil m/2 \rceil$.

Using a Convex LMI problem for products We can use the above LMI problem $\mathbb{Q}_K^N$ to find our measures of distribution width for products, $\psi_{\min}$ and $\psi_{\max}$ with the following definitions. We choose our polynomial function as $p(x) = prod(Cs+d)$. There is only one constraint, $s \in E(S, \bar{s}, \delta)$, which we represent, using equation (4.6), as $g_1(x) = (s-\bar{s})^T S(s-\bar{s}) - \delta^2$. Solving $\mathbb{Q}_K^N$ with these definitions will give us $\psi_{\min}^{C,d}(S, \bar{s}, \delta)$. Likewise, solving $\mathbb{Q}_K^N$ with $p(x) = -1 * prod(Cs+d)$ and the same constraint will give the measure $\psi_{\max}^{C,d}(S, \bar{s}, \delta)$.

While this formulation allows for calculation of $\psi_{\min}^{C,d}$ and $\psi_{\max}^{C,d}$, it lacks the convexity properties in Φ_u that are found when calculating the distribution width measures for parameter ratio estimates. Since the parameter ratios are relevant to our thin film deposition system and are the more interesting case due to the convexity properties, we will focus on parameter ratios for the remainder of the chapter.

4.4 Input Optimization for Ratios

We are now in a position to define the input optimization problem. The following are given:

- Parameter ratio to be estimated: $ratio(C_\theta\theta + d_\theta)$.
- Performance signal ratio to be constrained: $ratio(C_z z + d_z)$.
- Affine mappings for the parameter and performance signal covariances

$$P_\theta^{-1} \backsimeq f(\Phi_u),\ P_z = g(\Phi_u).$$

- Expected values for the parameters and performance signals: $\bar{\theta}$, $\bar{z}$.

Based on the discussion in the previous section, the width of the parameter ratio distribution will then be given by, $\phi^{C_\theta, d_\theta}\left(f(\Phi_u), \bar{\theta}, \delta\right)$ where δ is chosen depending on the certainty to which the constraint is expected to be met.

The constraint on the performance signal is one that should be met at all time points. In order to have a well defined problem, we must assume that the input

satisfies a uniformity property. In particular, we assume that for all k,

$$\begin{bmatrix} u(k+1) \\ u(k+2) \\ \vdots \\ u(k+\ell) \end{bmatrix}^T \Phi_u^{-1} \begin{bmatrix} u(k+1) \\ u(k+2) \\ \vdots \\ u(k+n-1) \end{bmatrix} \leq \beta^2. \tag{4.36}$$

Then the objective $\underline{m}_1 \leq ratio(C_z z_k + d_z) \leq \bar{m}_1$ will be met if

$$\phi_{\max}^{C_z,d_z}(g(\Phi_u), \bar{z}, \beta) \leq \bar{m}_1, \text{ and}$$
$$\phi_{\min}^{C_z,d_z}(g(\Phi_u), \bar{z}, \beta) \geq \underline{m}_1.$$

In addition, it is easy to add constraints on the variance of the performance signals. For example, if it is desired to achieve cov $(C_z^2 z) \leq b$, we can utilize the constraint $C_z^2 g(\Phi_u) \left(C_z^2\right)^T \leq b$.

We then arrive at the following optimization problem whose solution gives the correlation coefficients of the optimal input.

$$\begin{array}{ll} \min_{\Phi_u} & \phi^{C_\theta, d_\theta}\left(f(\Phi_u), \bar{\theta}, \delta\right) \\ \text{subject to} & \begin{cases} \phi_{\max}^{C_z, d_z}(g(\Phi_u), \bar{z}, \beta) \leq \bar{m}_1 \\ -\phi_{\min}^{C_z, d_z}(g(\Phi_u), \bar{z}, \beta) \leq -\underline{m}_1 \\ C_z^2 g(\Phi_u)\left(C_z^2\right)^T \leq b \\ T_N(\Phi_u) > 0 \end{cases} \end{array} \tag{4.37}$$

This optimization problem is convex, as the convexity or quasi-convexity of each function is preserved through the linear mappings f and g, so the objective function is convex, as are the constraints.

4.5 Experiment Design for a Thin Film Deposition System

In this section we describe how the input design problem developed in the previous sections can be used in a thin film deposition system such as the one described in Section 2.8.2. The following example was presented in Spinhirne & Vincent (2006).

As an example of a simple deposition system, consider a two-zone, two-material process, with nominal system model

$$\begin{bmatrix} y_k^1 \\ y_k^2 \end{bmatrix} = \begin{bmatrix} y_k^{11} + y_k^{12} \\ y_k^{21} + y_k^{22} \end{bmatrix} + \begin{bmatrix} d^{11} + d^{12} \\ d^{21} + d^{22} \end{bmatrix} + \begin{bmatrix} e_k^1 \\ e_k^2 \end{bmatrix}, \tag{4.38}$$

where $y_k^{ij} + d^{ij}$ is the amount of material i deposited in zone j at time k, y_k^i is the measured amount of deposited material, e_k^i is a zero mean measurement noise satisfying

$$cov\left(\begin{bmatrix} e_k^1 & e_k^2 \end{bmatrix}^T\right) = \lambda I.$$

In addition,

$$y_k^{ij} = G_{ij}(q^{-1})u_k^{ij},$$

where

$$\begin{aligned} G_{11}(q^{-1}) &= 0.4 + 3.2q^{-1} + 0.4q^{-2}, \\ G_{21}(q^{-1}) &= 0.8 + 6.4q^{-1} + 0.8q^{-2}, \\ G_{12}(q^{-1}) &= 0.8 + 6.4q^{-1} + 0.8q^{-2}, \\ G_{22}(q^{-1}) &= 1.5 + 12q^{-1} + 1.5q^{-2}, \text{ and} \\ \begin{bmatrix} d^{11} & d^{12} \\ d^{21} & d^{22} \end{bmatrix} &= \begin{bmatrix} 180 & 100 \\ 100 & 30 \end{bmatrix}. \end{aligned}$$

The objective is the estimation of the ratio $\frac{G_{11}(1)}{G_{12}(1)}$, that is, the ratio of DC gain between zone 1 and zone 2 for material 1. The process constraints will bound the variance of y^{ij} as well as the ratio of y^{1i} to the sum of y^{1i} and y^{2i} in each zone, as shown in Table 4.1. We took $N = 500$ and $\lambda = 20^2$. The problem data is as follows:

Variable	Desired Bound
y^{1i} variance in each zone	25^2
y^{2i} variance in each zone	25^2
$\frac{y^{1i}}{y^{1i}+y^{2i}}$ in Zone 1 and 2	± 0.1

Table 4.1. Performance Constraints.

$$
\begin{aligned}
\bar{\theta} &= \begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{21} & \theta_{22} \end{bmatrix}, \\
C_\theta &= \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & \cdots \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & \cdots \end{bmatrix}, \\
d_\theta &= \begin{bmatrix} 4 & 8 \end{bmatrix}^T, \\
\bar{z} &= \begin{bmatrix} 180 & 100 & 100 & 30 \end{bmatrix}, \\
C_z^1 &= \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\ 1 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & \cdots \end{bmatrix}, \\
d_z &= \begin{bmatrix} 180 & 280 \end{bmatrix}^T, \\
C_z^2 &= \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & \cdots & \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & \cdots & \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & \cdots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots \end{bmatrix}, \\
\delta &= 2, \text{ and} \beta = 1.
\end{aligned}
$$

After finding the optimal correlation parameters Φ_u, an input sequence was created by finding FIR filter parameters as discussed in Section 4.2.1, and then applying

independent pseudo-random-binary-sequences (PRBS) (Söderström & Stoica, 1989) as inputs to this filter. The resulting nominal performance signals are found by applying the input sequence to system (4.38). The sample covariance for y^{ij} was found to be

$$\overline{cov}(y^{11}) = 24.3^2, \overline{cov}(y^{12}) = 23.1^2,$$

$$\overline{cov}(y^{21}) = 16.8^2, \overline{cov}(y^{22}) = 5.8^2,$$

meeting the desired constraints. The ratios $\frac{y_k^{1i}}{y_k^{1i}+y_k^{2i}}$ are shown in Figure 4.8, along with the desired upper and lower limits. Note that the constraints are met at the majority of sample times. The constraints are violated whenever the PRBS input does not satisfy the input uniformity requirement (4.36), but clearly this is a good approximation.

To simulate the estimation process, 100 different noise sequences were added to the nominal response, and parameter estimates were found using least squares. The resulting distribution of estimate for $\frac{G_{11}(1)}{G_{12}(1)}$ is shown in Figure 4.9. Note that ϕ is a very reasonable estimate for the width of this distribution.

4.6 Experiment Design for Detecting Ratio Faults

In this section we describe how the input optimization problem we have introduced can be reformulated to design inputs for optimal detection of faults occurring

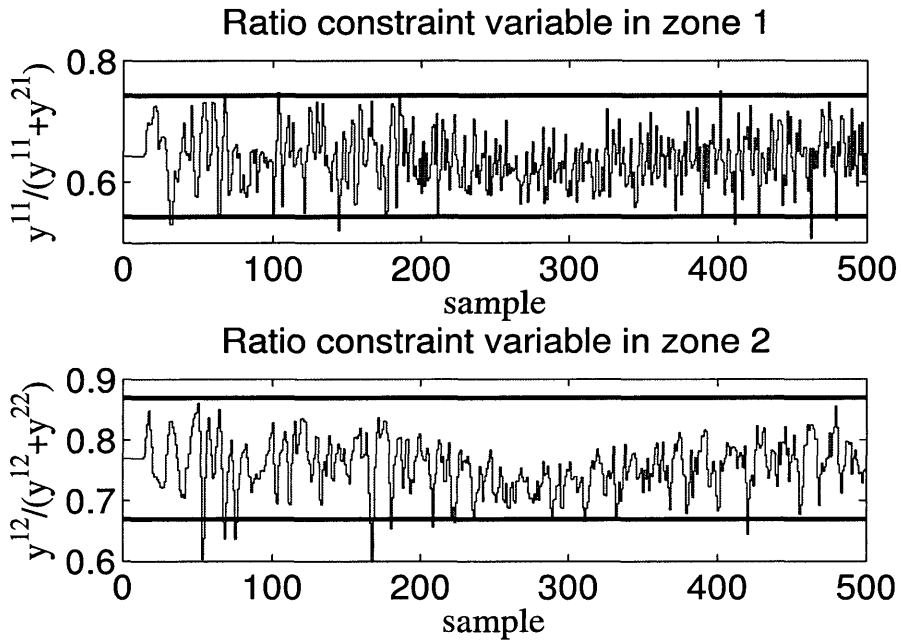

Figure 4.8. Time response of constraint signals. Desired limits are shown as solid lines. Note that in general the signals stay within the designated constraints.

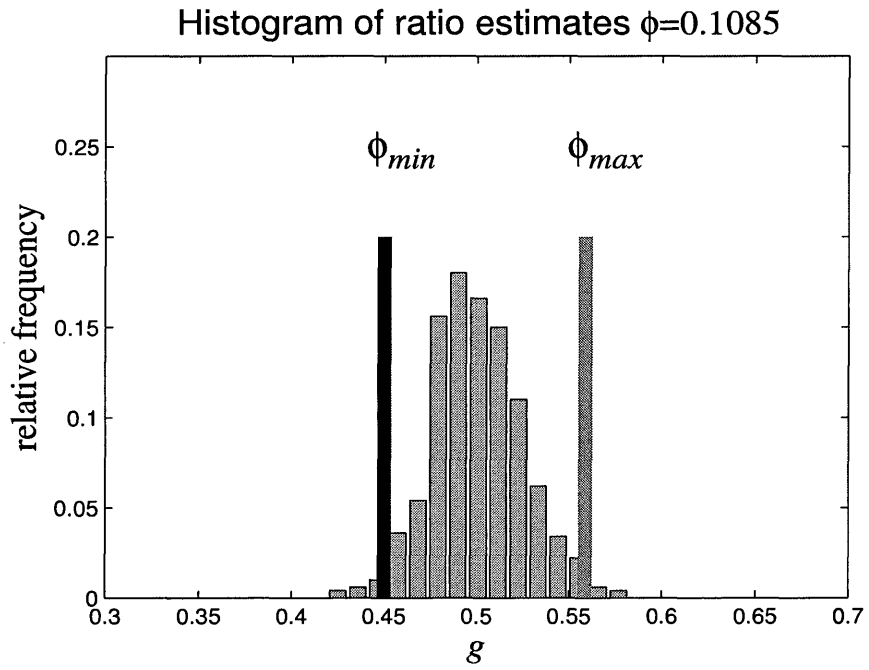

Figure 4.9. Histogram of estimates for $\frac{G_{11}(1)}{G_{12}(1)}$. with ϕ_{min} and ϕ_{max} based on the optimal input shown. Note that with optimal input, ϕ_{min} and ϕ_{max} designate reasonable bounds on the distribution of the estimate of the ratio of interest.

in the form of parameter ratios.

4.6.1 Optimization Problem

Given some fault line $c^T x = a$, defining a fault space from a non-fault space as in Figure 4.10, it can be said that a fault has occurred if the parameters, s, satisfy $c^T s \geq a$. Likewise no fault has occurred if $c^T s < a$.

However, because we can only estimate the parameters, s, within a confidence region (represented by an ellipse, $\mathcal{E}(S, \bar{s}, \delta)$) we have the possibility of misclassifying a parameter estimate as a fault or not a fault. Given a nominal parameter in the no-fault region, s_n, it is possible that part of the ellipse defining the estimate region may fall below the fault line, or in the fault space. If our estimate of s_n falls in this region, a fault would be detected where none exists. This is referred to as a false positive and is illustrated as the lower ellipse in Figure 4.11. The opposite case can also occur. Given a parameter in the fault region, s_f, the estimate may fall in the non-fault region and no fault is detected when one has occurred. This case is called a false negative and is represented in figure 4.11 by the upper ellipse.

If the system is such that production must be halted when a fault has occurred, a false positive causes unnecessary stops in production and lost time. Likewise, false negatives may result in defective products which also causes increased cost due to lost time and materials. Thus it is important to reduce the occurrence of both false positives and false negatives.

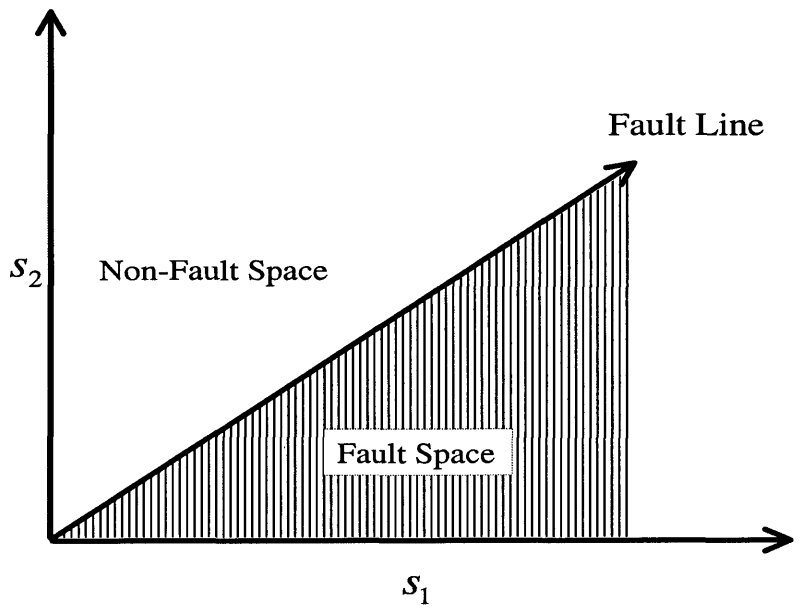

Figure 4.10. Fault Space for the *ratio*(s). Any ratio falling above the line is considered nominal while a ratio falling below indicates a fault in the system.

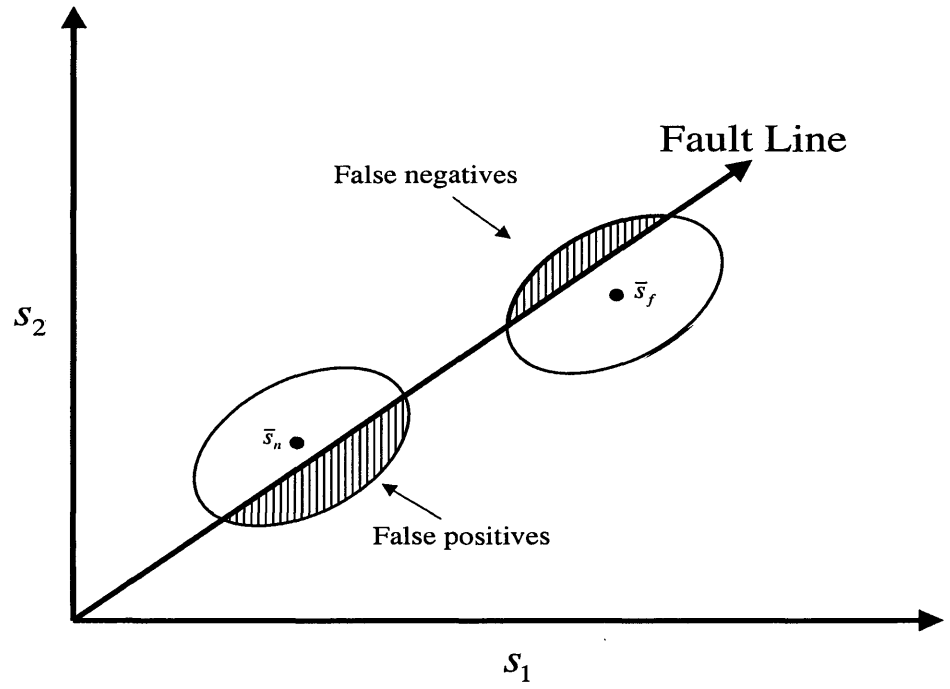

Figure 4.11. False positives and False negatives. Estimates of $\bar{s}_n$ falling below the fault line will incorrectly be classified as faults resulting in a false positive. Estimates of $\bar{s}_f$ falling above the fault line will be misclassified as nominal behavior resulting in a false negative.

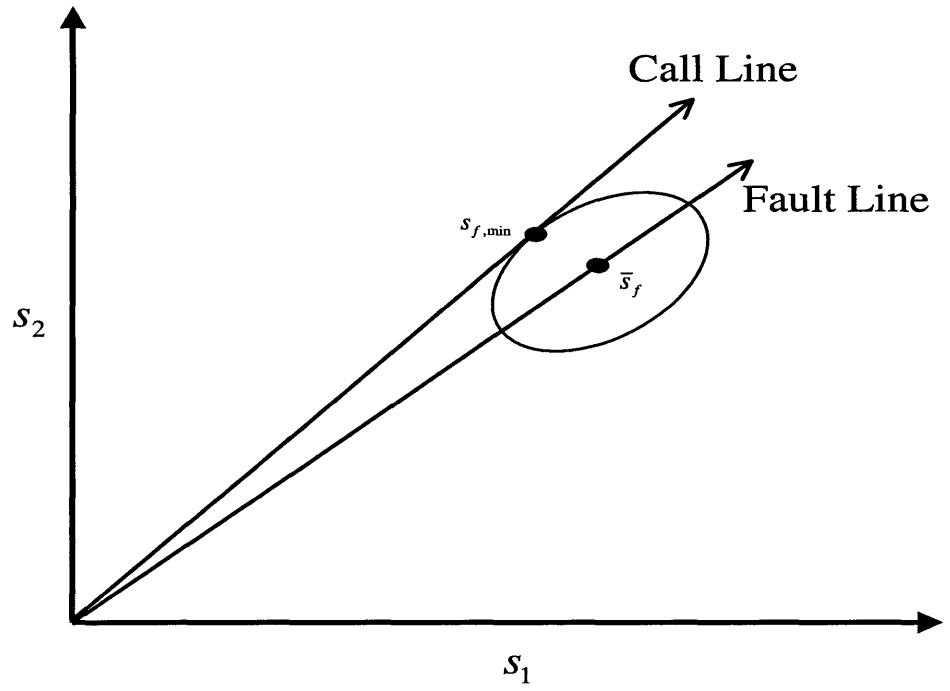

Figure 4.12. Fault call line based on $s_{f,min}$ will reduce the number of false negatives for a given confidence region.

To reduce the number of false negatives, one approach is to establish a call line at which a possible fault is detected that includes all estimates of $\bar{s}_f$ within a certain confidence interval. See Figure 4.12. This line is based on the $\phi_{\min}(S, \bar{s}_f, \delta)$. While the nominal parameter value, $\bar{s}_n$ is known from the system data, we must choose a fault parameter value, $\bar{s}_f$, to use in the input design problem. The worst case scenario for false negatives is when the faulty parameters $\bar{s}_f$ fall on the fault line as this puts the largest portion of the uncertainty ellipse in the no fault region, resulting in the largest number of false negatives. Therefore we want to design for when the faulty

parameters lie on the fault line. We choose $\bar{s}_f$ to be the point on the fault line such that the line between $\bar{s}_n$ and $\bar{s}_f$ is perpendicular to the fault line. This minimizes the distance between the nominal parameters and the fault parameters. This point is found via the following:

Given $\bar{s}_n = [\ \bar{s}_{n,1} \quad \bar{s}_{n,2}\ ]^T$ and a fault line $c^T x = a$:

- The distance from $\bar{s}_n$ to the line $c^T x = a$ can be found as

$$d = \left| \frac{c_1 \bar{s}_{n,1} + c_2 \bar{s}_{n,2} - a}{\sqrt{c_1^2 + c_2^2}} \right| = \left| \frac{c^T \bar{s}_n - a}{\|c\|} \right| \tag{4.39}$$

- Now $\bar{s}_f$ can be found. Let $l = \sqrt{\|\bar{s}_n\|^2 - d^2}$. We know that $\|\bar{s}_f\| = l$ and the slope of the fault line is $\frac{-c_1}{c_2} = \frac{1}{ratio(\bar{s}_f)}$. This gives:

$$\bar{s}_{f,1} = \frac{l}{\sqrt{1 + \left(-\frac{c_1}{c_2}\right)^2}} \tag{4.40}$$

$$\bar{s}_{f,2} = \frac{-c_1 \bar{s}_{f,1}}{c_2} \tag{4.41}$$

We now have the situation depicted in Figure 4.13. If a call line is based on the ellipse $\mathcal{E}(S, \bar{s}_f, \delta)$ where $\bar{s}_f$ is chosen as in Figure 4.13, any ellipse $\mathcal{E}(S, \bar{s}_f 2, \delta)$, where $\bar{s}_f$ falls on the fault line but further away from the origin than $\bar{s}_f$, will be below the call line. See Figure 4.14. Thus, by choosing $\bar{s}_f$ as a design parameter, we are ensured to also eliminate any false negatives with in our confidence level should the actual

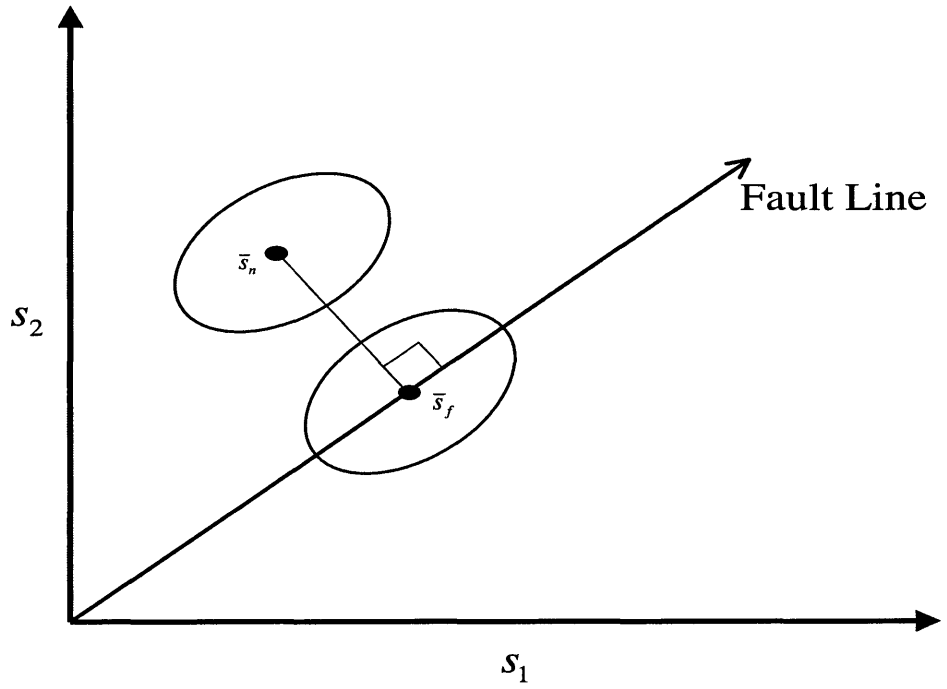

Figure 4.13. Choosing a fault parameter, $\bar{s}_f$, to lie on the fault line represents the worst case for estimates of $\bar{s}_f$ resulting in false negatives.

parameters lie further away from the origin than $\bar{s}_f$.

While this eliminates the number of false negatives, it can increase the number of false positives. Therefore, we wish to bring the call line as close to the fault line as possible. Thus, we will design an input such that $s_{f,min}$ is maximized and the call line is based on $\phi_{\min}(S, \bar{s}_f, \delta)$. This gives the fault test as $c_{cl}^T s = a_{cl}$ where $c_{cl}^T = \begin{bmatrix} \frac{1}{\phi_{min}(S,\bar{s}_f,\delta)} & -1 \end{bmatrix}$.

However, we do not want to design an input that allows the confidence ellipse for a non-fault set of parameters to grow such that it falls below the call line, increasing

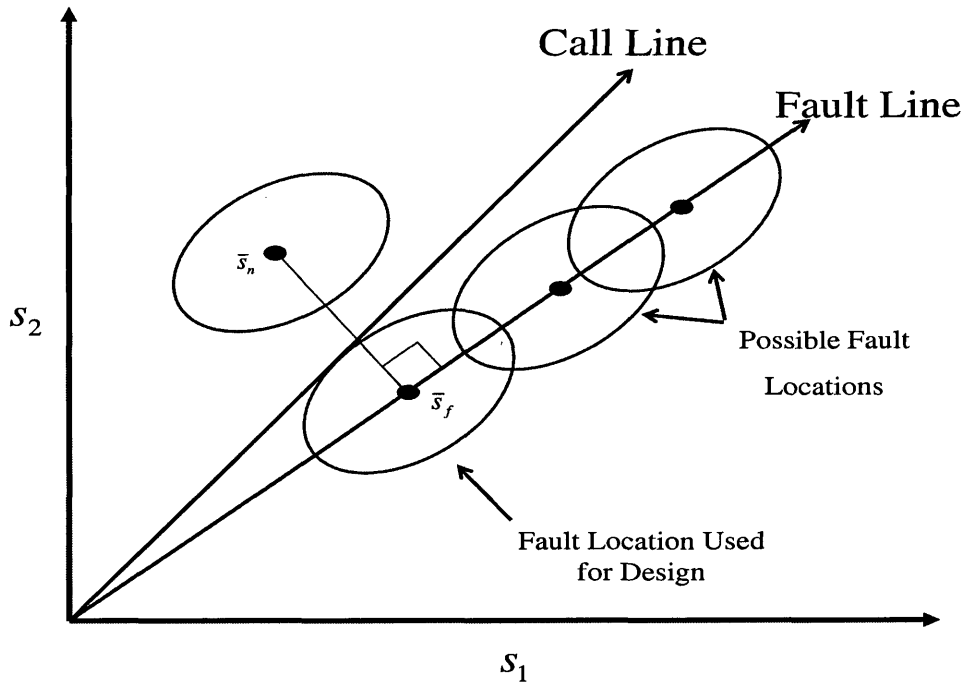

Figure 4.14. Alternate Fault Locations. Other possibilities for $\bar{s}_f$ further away from the origin along the fault line will still have estimate regions that fall under the call line, thus making the choice of $\bar{s}_f$ an appropriate one.

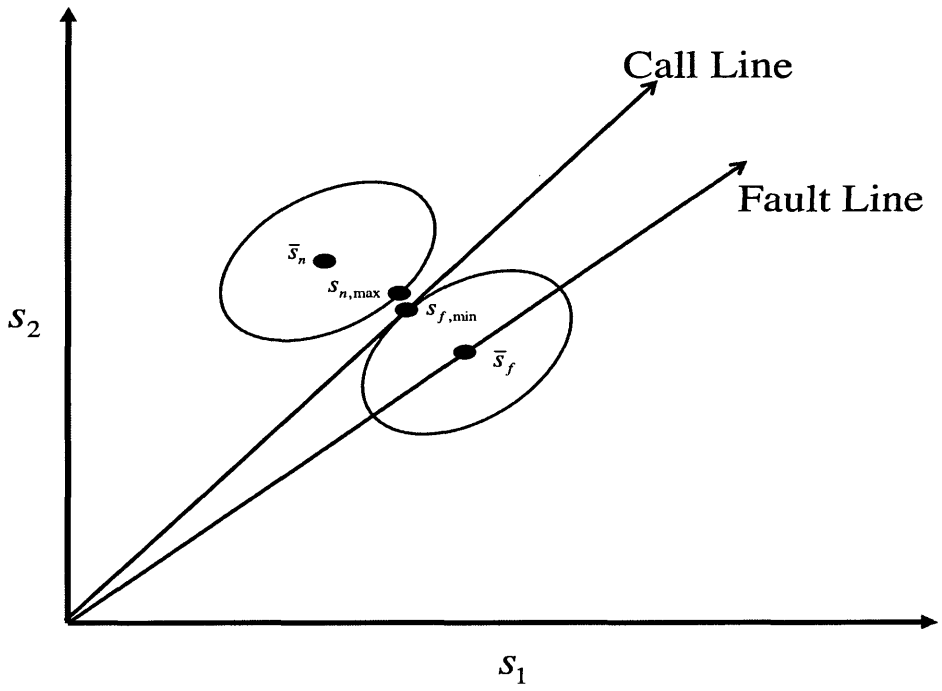

Figure 4.15. Depiction of the call line optimization problem where the goal is to maximize $\phi_{min}(S, \bar{s}_f, \delta)$ with the constraint that $\phi_{max}(S, \bar{s}_n, \delta)$ must lie above the call line thereby minimizing both false positives and false negatives.

the occurrence of false positives. Therefore we add a constraint to prevent this. Given δ and $\bar{s}_n$, this gives the optimization problem:

$$\begin{aligned} \max_S \quad & \phi_{min}(S, \bar{s}_f, \delta) \\ \text{subject to} \quad & c_{cl}^T s_{n,max} \leq a_{cl} \end{aligned} \tag{4.42}$$

where $\bar{s}_f$ and c_{cl}^T are found as discussed above.

Refer to Figure 4.15. While this will not eliminate false positives and false

negatives, it will help to greatly reduce them within the specified confidence regions.

4.6.2 Ratio Fault Example

Consider the two-zone, two-material process introduced in section 4.5.1. Using the same problem data, we wish to find the optimal correlation parameters Φ_u that give an optimal call line for detecting faults of the ratio $\frac{G_{11}(1)}{G_{12}(1)}$ within the same process constraints discussed above. In the previous example we saw that a nominal value for the ratio $\frac{G_{11}(1)}{G_{12}(1)}$, $\bar{s}_n$, is 0.5. A fault will be called if this ratio exceeds 0.65. We choose $\bar{s}_f$ to lie on the line which represents a ratio of 0.65 at a point where the line between $\bar{s}_n$ and $\bar{s}_f$ is perpendicular to the fault line. Now, the optimal input correlation parameters, and thereby the optimal call line, is found by solving the optimization problem (4.42) with the process constraints discussed previously. The input design problem becomes:

$$\begin{aligned} \min_{\Phi_u} \quad & -\phi_{\min}^{C_\theta, d_\theta}\left(f(\Phi_u), \bar{s}_f, \delta\right) \\ \text{subject to} \quad & \begin{cases} \phi_{\max}^{C_z, d_z}(g(\Phi_u), \bar{z}, \beta) \leq \bar{m}_1 \\ -\phi_{\min}^{C_z, d_z}(g(\Phi_u), \bar{z}, \beta) \leq -\underline{m}_1 \\ C_z^2 g(\Phi_u)\left(C_z^2\right)^T \leq b \\ T_N(\Phi_u) > 0 \\ c_{cl}^T s_{n,\max} \leq a_{cl}. \end{cases} \end{aligned} \tag{4.43}$$

Repeating the procedure discussed in 4.5.1, the optimal correlation parameters, Φ_u, were found and an input sequence was created by finding the FIR filter parameters and then applying independent pseudo-random-binary sequences (PRBS) as inputs to this filter. As in the previous example, the performance signals can be shown to meet the desired constraints. The sample covariances for the individual metals in each zone are as follows:

$$\overline{cov}(y^{11}) = 22.3^2, \overline{cov}(y^{12}) = 24.8^2,$$

$$\overline{cov}(y^{21}) = 18.4^2, \overline{cov}(y^{22}) = 8.3^2.$$

By looking at Figure 4.16 it can be seen that the process constraints on ratios also were met. This figure shows the constraint ratios produced from applying the input sequence to the system. The estimation process was simulated by adding 100 different noise sequences to the nominal response, and parameter estimates were found using least squares. Figure 4.17 shows the optimal call line for the ratio fault along with the estimates of the nominal parameters. It can be seen that all the nominal parameter estimates fall above the call line therefore giving no false positives. At the same time all faulty parameter estimates within the given confidence region will fall under the call line thus reducing the possibility of false negatives. Figure 4.18 shows the resulting distribution of the parameter ratio estimate. We note that ϕ is still a very reasonable estimate for the width of the distribution.

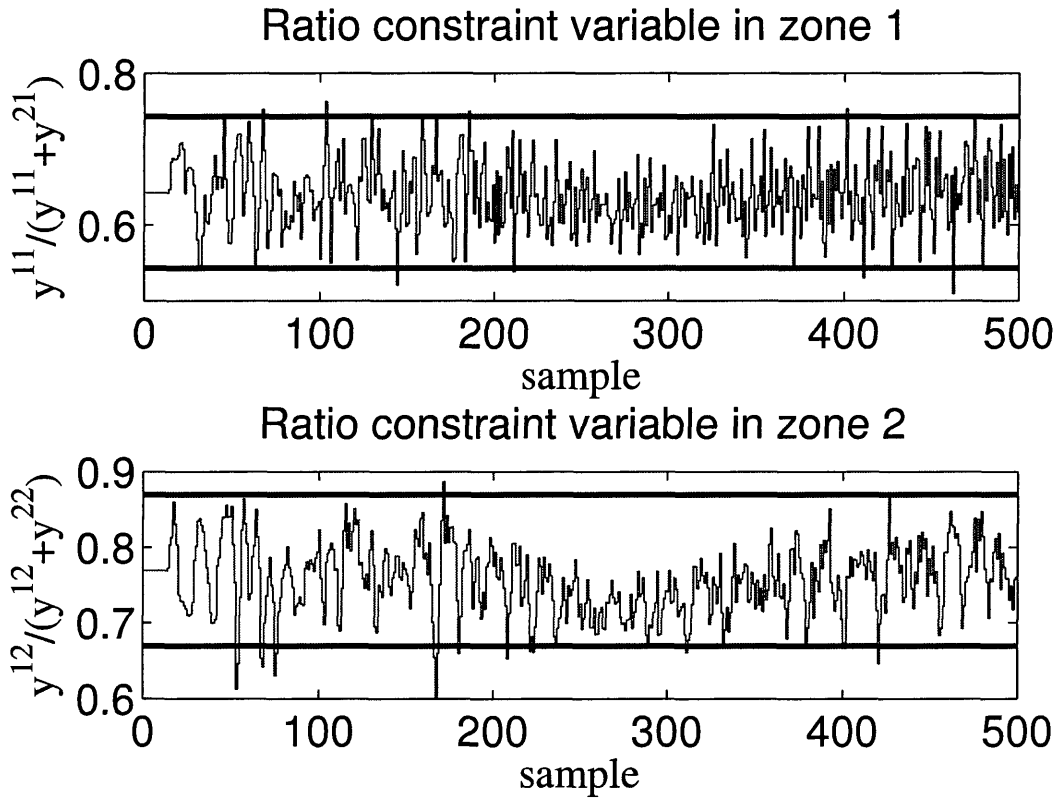

Figure 4.16. Time response of constraint signals for ratio fault optimization problem. Desired limits are shown as solid lines. Note that in general the signals stay within the designated constraints.

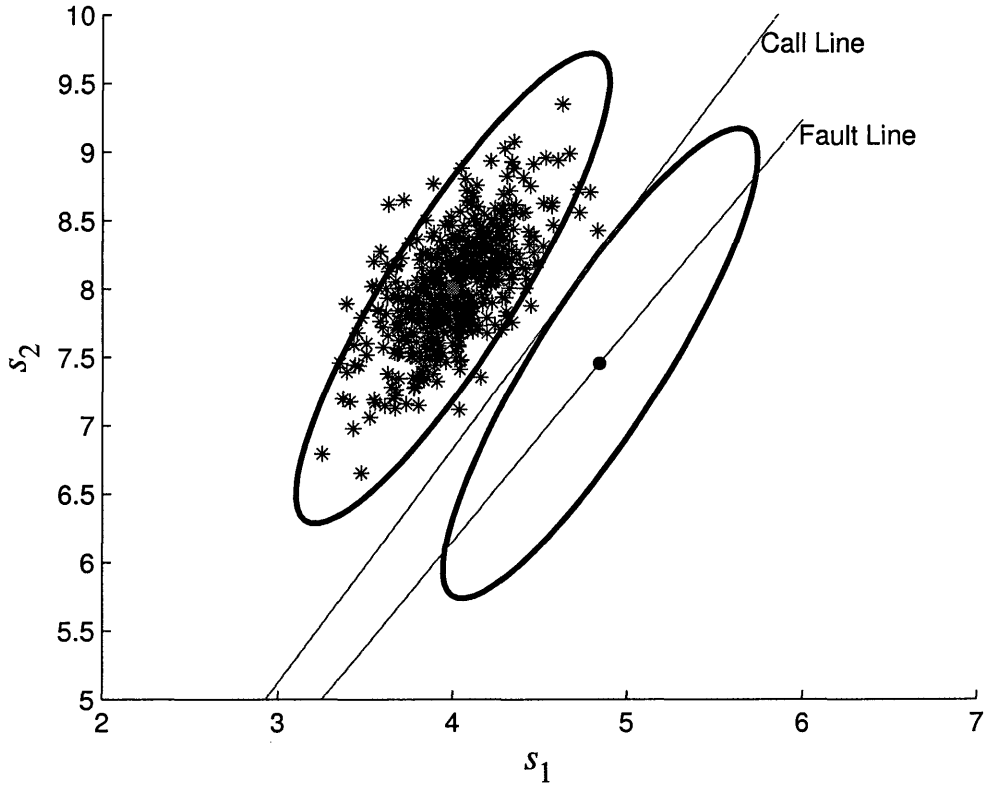

Figure 4.17. Results from the ratio fault optimization. Note that the estimates of the nominal ratio lie above the call line resulting in no false positives while the confidence region for the faulty ratio is oriented along the fault line reducing the number of false negatives.

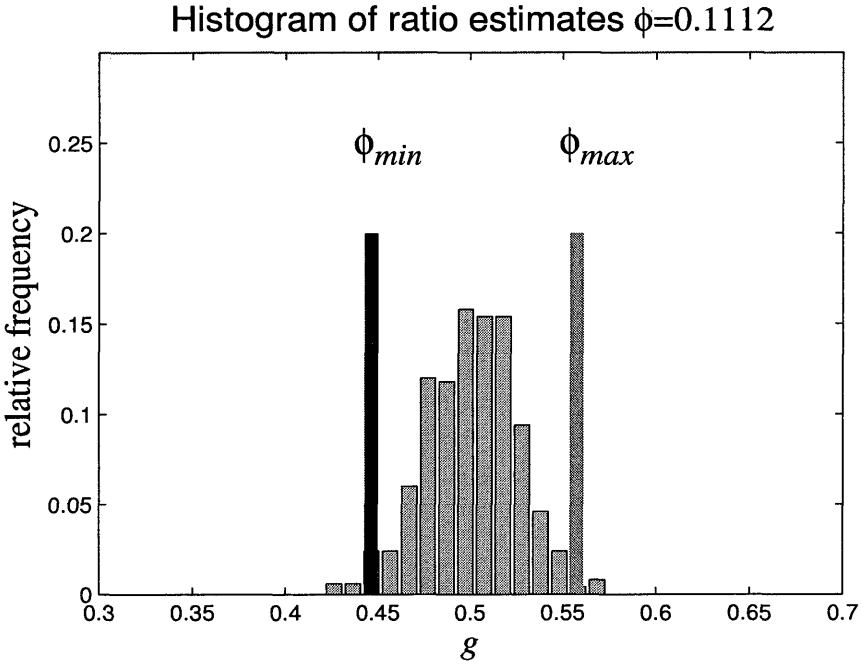

Figure 4.18. Histogram of estimates for $\frac{G_{11}(1)}{G_{12}(1)}$. with ϕ_{min} and ϕ_{max} based on the optimal input from the ratio fault optimization problem shown. Note that with optimal input, ϕ_{min} and ϕ_{max} designate reasonable bounds on the distribution of the estimate of the ratio of interest.

Chapter 5

CONCLUSIONS

Fault Detection and Identification is becoming more crucial in today's complex industrial systems. Due to the need for accurate system models for model-based FDI, system identification methods and their effects on the fault detection process become important components of system monitoring. One important consideration in system identification for fault detection is the input condition which existed during system identification. Two contrasting cases related to the input during system identification have been presented.

The first case considers a situation where the input during system identification can not be chosen arbitrarily. As this is a common case in industry, it is important for fault detection methods to take into account the situations that can result from using production data for system identification. If the input did not excite all modes of the system, the model produced by the system identification will not be complete. This is a case of the input not be sufficiently exciting. This thesis has developed a method for filtering residuals to remove the error due to incomplete system identification. This allows for increased accuracy in the fault detection process even though the system

identification data may have had a different level of input excitation than the data during system monitoring.

While the method of filtering residuals shown in this thesis greatly increases the accuracy of fault detection, there are still some limitations to this approach. It was shown that the filtering of the residual can lead to hidden faults, although a bound on the magnitude of these faults can be determined. Therefore, while the majority of the faults will be detected, there is a class of faults that will be obscured by choosing to filter the residuals. Another limitation is in the fault identification process. If there exists a model of how a sensor fault affects the output, then a minimization problem can be solved to determine the fault sequence that may have occurred. If the fault is not a sensor fault or can not be modeled so easily, then fault identification requires diagnosed data taken in the presence of known faults. In addition, this method is focused on a linear system model.

Future work in this area could include investigating the properties of hidden faults resulting from residual filtering to determine the effect these faults would have on the system. There is also a need for improved fault modeling to allow fault isolation with out the need for diagnosed production data which may not be available. Another area with interesting possibilities is to extend this method to nonlinear system models. Due to the nonlinearity of most industrial systems, this would allow for improved accuracy in fault detection.

The second case considered in this thesis assumes that the input can be chosen within some system constraints. Here the problem of input design for parameter estimation and parameter variation fault detection was investigated. It was shown that this problem could be formed as a convex optimization problem. An alternative characterization for the estimate quality based on ellipsoidal bounds was established which led to the convex properties of the optimization problem.

While the ability to form the input design problem as a convex optimization problem has definite advantages over a nonlinear problem, it still has some limitations due to the complexity of the problem. Even though the example problem considered in this thesis was fairly low in complexity, it was reaching a limit of what a gradient method can solve. In order to improve the computational cost and allow for increased complexity, alternative methods for solving convex problems such as interior-point methods or ellipsoidal methods should be investigated. The application of these methods to the design problem formulated in this thesis would allow application to a greater class of industrial problems.

An additional area of continued research could focus on the case where the optimization problem is infeasible due to the manufacturing constraints placed on the input. In these cases it may be necessary to modify some of the parameters that were held constant in this thesis such as the confidence level, δ, which defines the size of the estimate ellipse. Another possibility is to include the number of samples considered

when doing parameter estimation as a variable in the design problem. The larger the number of samples, the better the estimate of the parameter ratio that can be achieved. However, this number should be balanced with the need to do timely fault detection. If it is necessary to collect a large number of data samples before doing parameter variation fault detection, then faults will not be detected in a manner which saves time and money. It may be possible to determine some metric for balancing the increase in accuracy of the parameter estimation with timely fault detection that would allow for more flexibility in the input design problem. All of these issues would be interesting topics for future work.

REFERENCES

Anderson, E., Bai, Z., Bischof, C., Blackford, S., Demmel, J., Dongarra, F., Du Croz, J., Greenbaum, A., Hammarling, S., McKenney, A., & Sorensen, D. 1999. *LAPACK User's Guide.* 3 edn. SIAM.

Birkmire, Robert W., & Eser, Erten. 1997. Polycristalline Thin Film Solar Cells: Present Status and Future Potential. *Annu. Rev. Mater. Sci.*, **27**, 625–53.

Boyd, Stephen, El Ghaoui, Laurent, Feron, Eric, & Balakrishnan, Venkataramanan. 1994. *Linear Matrix Inequalities in System and Control Theory.* Philadelphia: SIAM.

Campbell, Stephen L., & Nikoukhah, Ramine. 2004. *Auxiliary Signal Design for Failure Detection.* Princeston NJ: Princeston University Press.

Chou, C.T., & Verhaegan, M. 1997. Subspace Algorithms for the Identification of Multivariable Dynamic Errors-in-Variables Model. *Automatica*, **33**(10), 1857–1869.

Downs, J.J., & Vogel, E.F. 1993. A plant-wide industrial process control problem. *Computers and Chemical Engineering*, **17**(3), 245–255.

Dunia, R., & Qin, S. J. 1998a. A subspace approach to multidimensional fault identification and reconstruction. *AIChE Journal*, **44**(8), 1813–1831.

Dunia, R., & Qin, S. J. 1998b. A unified geometric approach to process and sensor fault identification and reconstruction: The unidimensional fault case. *Computers and Chemical Engineering*, **22**(7-8), 927–943.

Frank, P. 1990. Fault Diagnosis in Dynamic Systems Using Analytical and Knowledge-based Redundancy - A Survey and Some New Results. *Automatica*, **26**(3), 459–474.

Gertler, J. 1988. Survey of Model-Based Failure Detection and Isolation in Complex Plants. *IEEE Control Systems*, **8**(6), 3–11.

Gertler, J. 1998. *Fault Detection and Diagnosis in Engineering Systems.* New York: Marcel Dekker, Inc.

Gevers, Michel. 2005. Identification for Control: From the Early Achievements to the Revival of Experiment Design. *European Journal of Control*, **11**, 335–352.

Golub, G.H., & Loan, C.F.V. 1989. *Matrix Computations.* 2 edn. Baltimore: Johns Hopkins University Press.

Goodwin, G. C., & Payne, R. L. 1977. *Dynamic System Identification: Experiment Design and Data Analysis.* New York: Academic Press.

Hilt, Matthew J., Vincent, Tyrone L., Joshi, Bharat S., & Simpson, Lin J. 2005. Estimating Relative Deposition in a Multi-Zone Process Using a Single Composition Sensor. *Pages 4525–4530 of: Proc. American Control Conference June 8-10. Portland, OR, USA.* Green Valley, AZ: AACC.

Hilt, Matthew J., Vincent, Tyrone L., Joshi, Bharat S., & Simpson, Lin J. 2006. Estimating Relative Deposition in a Multi-Zone Process Using a Single Composition Sensor. *IEEE Transactions on Control Systems Technology*, **14**(2), 247–259.

Iserman, R. 1984. Process Fault Detection Based on Modeling and Estimation Methods - A Survey. *Automatica*, **20**(4), 387–404.

Jackson, J.E. 1991. *A User's Guide to Principle Components.* New York: Wiley-Interscience.

Ku, W., Storer, R.H., & Georgakis, C. 1995. Disturbance detection and isolation by dynamic principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, **30**, 179–196.

Lasserre, Jean B. 2001. Global Optimization with Polynomials and the Problem of Moments. *SIAM Journal of Optimization*, **11**(3), 796–817.

Li, W., & Qin, S.J. 2001. Consistent Dynamic PCA Based on Errors-in-Variables Subspace Identification. *Journal of Process Control*, **11**(6), 661–678.

Lin, W., Qian, Y., & Li, X. 2000. Nonlinear dynamic principal component analysis for on-line process monitoring and diagnosis. *Computers and chemical Engineering*, **24**, 423–429.

Ljung, L. 1999. *System Identification: Theory for the User.* 2 edn. New Jersey: Prentice Hall.

Lyman, P.R., & Georgakis, C. 1995. Plant-wide control of the Tennessee Eastman problem. *Computers and Chemical Engineering*, **19**(3), 321–331.

MacGregor, J.F. 2003. Multivariate Statistical Approaches to Fault Detection and Isolation. *Pages 579–584 of: Safeprocess 2003.* IFAC Symposium on Fault Detection, Supervision, and Safety of Technical Processes.

McAvoy, T. J., & Ye, N. 1994. Base control for the Tennessee eastman problem. *Computers and Chemical Engineering*, **18**, 383–413.

Overschee, P.V., & Moor, B.D. 1996. *Subspace Identification for Linear Systems: Theory, Implementation, Applications.* Boston: Klumer Academic Publishers.

Qin, S.J., & Li, W. 2001. Detection and Identification of Faulty Sensors in Dynamic Processes with Maximized Sensitivity. *AIChE Journal*, **47**, 1581–1593.

Ricker, N.L. 1996. Decentralized Control of the Tennessee Eastman Challenge Process. *Journal of Process Control*, **6**, 205–221.

Ros, L., Sabater, A., & Thomas, F. 2002. An Ellipsoidal Calculus Based on Propagation and Fusion. *IEEE Transactions on Systems, Man, and Cybernetics-Part B: Cybernetics*, **32**(4), 430–442.

Russell, E.L., Chiang, L.H., & Braatz, R.D. 2000. *Data-Driven Techniques for Fault Detection and Diagnosis in Chemical Processes.* London: Springer-Verlag.

Shi, R., & MacGregor, J.F. 2001. A framework for subspace identification methods. *Pages 3678–3683 of: Proceedings of the American Control Conference.* AACC, Green Valley, AZ.

Söderström, Torsten, & Stoica, Petre. 1989. *System Identification.* New York: Prentice Hall.

Spinhirne, RM, & Vincent, TL. 2004 (November). On Fault Detection Using Dynamic PCA with Varying Input Excitation. *Pages 281–286 of: Proceedings of the IAR Advanced Control and Diagnostic Workshop.*

Spinhirne, RM, & Vincent, TL. 2006. Experiment Design for Ratio Estimates. *Pages 5165–5170 of: Proceedings of the 2006 American Controls Conference.*

Van Trees, Harry L. 1968. *Detection, Estimation, and Modulation Theory.* New York: Wiley.

Wu, Shao-Po, Boyd, Stephen, & Vandenberghe, Lieven. 1999. FIR Filter Design Via Spectral Factorization and Convex Optimization. *Chap. 1 of:* Datta, Biswa N. (ed), *Applied and Computational Control, Signals, and Circuits.* Birkhauser.

APPENDIX A

PROOF OF THEOREM 5

The optimization problem 3.11 can be rewritten to include the constraint as

$$\min_{h,f_0,\ldots,f_N} \sum_{i=1}^{N} \left\| \begin{bmatrix} 0 & 0 & -C & -D_i & r_i \\ -S & -T & I & 0 & 0 \end{bmatrix} \begin{bmatrix} f_{i-1} \\ e_i \\ f_i \\ h \\ 1 \end{bmatrix} \right\|_F^2 .$$

The first iteration can be separated so that

$$\min_{h,f_0,\ldots,f_N} \left[\left\| M \begin{bmatrix} f_0 \\ e_1 \\ f_1 \\ h \\ 1 \end{bmatrix} \right\|_F^2 + \sum_{i=2}^{N} \left\| \begin{bmatrix} 0 & 0 & -C & -D_i & r_i \\ -S & -T & I & 0 & 0 \end{bmatrix} \begin{bmatrix} f_{i-1} \\ e_i \\ f_i \\ h \\ 1 \end{bmatrix} \right\| \right]$$

where

$$M = \begin{bmatrix} 0 & 0 & -C & -D_1 & r_1 \\ -S & -T & I & 0 & 0 \end{bmatrix}.$$

Focusing on the first term, a QR decomposition can be used to find matrices such that $M = Q \begin{bmatrix} R_{11} & R_{12} \\ 0 & R_{22} \end{bmatrix}$ with Q orthogonal. The first term can now be written as

$$\left\| M \begin{bmatrix} f_0 \\ e_1 \\ f_1 \\ h \\ 1 \end{bmatrix} \right\|_F^2 = \left\| \begin{bmatrix} R_{11} & R_{12} \\ 0 & R_{22} \end{bmatrix} \begin{bmatrix} \begin{bmatrix} f_0 \\ e_1 \end{bmatrix} \\ \begin{bmatrix} f_1 \\ h \\ 1 \end{bmatrix} \end{bmatrix} \right\|_F^2.$$

Using the singular value decomposition find $R_{11} = U\Sigma V^T$.so that

$$\left\| M \begin{bmatrix} f_0 \\ e_1 \\ f_1 \\ h \\ 1 \end{bmatrix} \right\|_F^2 = \left\| \begin{bmatrix} \Sigma & 0 & (U^T R_{12})_n \\ 0 & 0 & (U^T R_{12})_z \\ 0 & & R_{22} \end{bmatrix} \begin{bmatrix} V^T \begin{bmatrix} f_0 \\ e_1 \end{bmatrix} \\ \begin{bmatrix} f_1 \\ h \\ 1 \end{bmatrix} \end{bmatrix} \right\|_F^2$$

where $(U^T R_{12})_n$ are the rows of $U^T R_{12}$ associated with the nonzero singular values and $(U^T R_{12})_z$ are the rows of $U^T R_{12}$ that correspond with the zero singular values, rows $k+1$ to N. Now

$$\left\| M \begin{bmatrix} f_0 \\ e_1 \\ f_1 \\ h \\ 1 \end{bmatrix} \right\|_F^2 = \left\| \begin{bmatrix} \Sigma & (U^T R_{12})_n \end{bmatrix} \begin{bmatrix} V^T \begin{bmatrix} f_0 \\ e_1 \end{bmatrix} \\ \begin{bmatrix} f_1 \\ h \\ 1 \end{bmatrix} \end{bmatrix} \right\|_F^2 + \left\| \begin{bmatrix} (U^T R_{12})_z \\ R_{22} \end{bmatrix} \begin{bmatrix} f_1 \\ h \\ 1 \end{bmatrix} \right\|_F^2 .$$

Since Σ is invertible, given f_1 and h, values for f_0 and e_1 can be found such that the first term is zero. Therefore the new minimization problem is reduced to

$$\min_{h, f_0, \ldots, f_N} \left[\left\| \begin{bmatrix} \Gamma_{11} & \Gamma_{21} & \Gamma_{31} \end{bmatrix} \begin{bmatrix} f_1 \\ h \\ 1 \end{bmatrix} \right\|_F^2 + \sum_{i=2}^{N} \left\| \begin{bmatrix} 0 & 0 & -C & -D_i & r_i \\ -S & -T & I & 0 & 0 \end{bmatrix} \begin{bmatrix} f_{i-1} \\ e_i \\ f_i \\ h \\ 1 \end{bmatrix} \right\| \right]$$

where

$$\begin{bmatrix} (U^T R_{12})_z \\ R_{22} \end{bmatrix} = \begin{bmatrix} \Gamma_{11} & \Gamma_{21} & \Gamma_{31} \end{bmatrix} .$$

Now separate another iteration so that the problem is

$$\min_{h,f_0,\ldots,f_N} \left\| M \begin{bmatrix} f_1 \\ e_2 \\ f_2 \\ h \\ 1 \end{bmatrix} \right\|_F^2 + \sum_{i=2}^{N} \left\| \begin{bmatrix} 0 & 0 & -C & -D_i & r_i \\ -S & -T & I & 0 & 0 \end{bmatrix} \begin{bmatrix} f_{i-1} \\ e_i \\ f_i \\ h \\ 1 \end{bmatrix} \right\|$$

where

$$M = \begin{bmatrix} \Gamma_{11} & 0 & 0 & \Gamma_{21} & \Gamma_{31} \\ 0 & 0 & -C & -D_2 & r_2 \\ -S & -T & I & 0 & 0 \end{bmatrix}.$$

This process can be repeated until $\begin{bmatrix} \Gamma_{1N} & \Gamma_{2N} & \Gamma_{3N} \end{bmatrix}$ is found. Then estimates of the fault can be found via

$$\hat{h}, \hat{f}_N = \arg\min_{h,f_N} \left\| \begin{bmatrix} \Gamma_{1N} & \Gamma_{2N} & \Gamma_{3N} \end{bmatrix} \begin{bmatrix} f_N \\ h \\ 1 \end{bmatrix} \right\|^2.$$

Find $\hat{f}_i, i = 0 \ldots N-1$ by using

$$\begin{bmatrix} \hat{f}_{i-1} \\ \\ \hat{e}_i \end{bmatrix} = G_i \begin{bmatrix} \hat{f}_i \\ \\ \hat{h} \\ \\ 1 \end{bmatrix}.$$

where

$$G_i = -1 * (\Sigma V^T)^{-1} (U^T R_{12})_n.$$

APPENDIX B

PROOF OF THEOREM 17

First, we note that $\mathcal{D}_{C,d}$ is convex in both S and P by Corollary 16.

The formulas for ϕ_{max} and ϕ_{min}, and the convexity of ϕ_{max} and ϕ_{min} with respect to S can be established using results similar to that in Hilt *et al.* (2006). The proof is repeated here for completeness.

Re-parameterize the ellipsoid $\mathcal{E}$. Let $w = S^{-1/2}v$ for some $\|v\| \leq 1$. Then $s \in \mathcal{E}(S, \bar{s}, \delta)$ if and only if $s = \delta S^{-1}w + \bar{s}$ for $\|w\|_{S^{-1}} \leq 1$. This gives an equivalent optimization problem

$$\phi^{C,d}_{\max}(S, \bar{s}, \delta) = \begin{cases} \max_w & ratio(\delta S^{-1}w + \bar{s}) \\ \text{subject to} & \|w\|_{S^{-1}} \leq 1 \end{cases} \tag{B.1}$$

$$\phi^{C,d}_{\min}(S, \bar{s}, \delta) = \begin{cases} \min_w & ratio(\delta S^{-1}w + \bar{s}) \\ \text{subject to} & \|w\|_{S^{-1}} \leq 1 \end{cases} \tag{B.2}$$

Some additional lemmas which will help in this proof need to be established.

Lemma 18. *If w is a solution to (B.1) or (B.2), then $\|w\|_{S^{-1}} = 1$ and $w^T\bar{s} = -\delta$ are satisfied.*

Proof. Consider the first equation. If w is a solution to (B.1) or (B.2) then the derivative must be equal to zero. However the derivative of $ratio(\delta S^{-1}w + \bar{s})$ does not equal 0 for $\|w\|_{S^{-1}} < 1$, therefore the solution must be on the boundary of the ellipse, giving $\|w\|_{S^{-1}} = 1$. Now for the second equation. If $\|w\|_{S^{-1}} = 1$ and $w = S^{1/2}v$, then $\|v\| = 1$ as well. Choosing $v = \begin{bmatrix} \sin(\theta) & \cos(\theta) \end{bmatrix}^T$, s can be written as:

$$s(\theta) = \delta S^{-1/2} \begin{bmatrix} \sin(\theta) & \cos(\theta) \end{bmatrix}^T + \bar{s}.$$

At values of θ corresponding the maximum or minimum solution, $\frac{\partial ratio(s(\theta))}{\partial \theta} = 0$. This can also be written as

$$s(\theta)^T \left[\frac{\partial s}{\partial \theta}\right]^{\perp}$$

where $\left[\frac{\partial s}{\partial \theta}\right]^{\perp}$ satisfies $\left(\frac{\partial s}{\partial \theta}\right)^T \left[\frac{\partial s}{\partial \theta}\right]^{\perp} = 0$

$$\begin{aligned}
\frac{\partial s}{\partial \theta} &= \delta S^{-1/2} \begin{bmatrix} \cos(\theta) & -\sin(\theta) \end{bmatrix}^T \\
\left[\frac{\partial s}{\partial \theta}\right]^{\perp} &= S^{1/2} \begin{bmatrix} \sin(\theta) & \cos(\theta) \end{bmatrix}^T \\
\left(\frac{\partial s}{\partial \theta}\right)^T \left[\frac{\partial s}{\partial \theta}\right]^{\perp} &= \left(\delta S^{-1/2} \begin{bmatrix} \cos(\theta) & -\sin(\theta) \end{bmatrix}^T\right)^T S^{1/2} \begin{bmatrix} \sin(\theta) & \cos(\theta) \end{bmatrix}^T \\
\left(\frac{\partial s}{\partial \theta}\right)^T \left[\frac{\partial s}{\partial \theta}\right]^{\perp} &= \begin{bmatrix} \cos(\theta) & -\sin(\theta) \end{bmatrix} (S^{-1/2})^T \delta S^{1/2} \begin{bmatrix} \sin(\theta) & \cos(\theta) \end{bmatrix}^T
\end{aligned}$$

The matrix $S^{-1/2}$ is symmetric so $(S^{-1/2})^T = S^{-1/2}$ and $S^{-1/2}S^{1/2} = I$.

$$\begin{aligned}
\left(\frac{\partial s}{\partial \theta}\right)^T \left[\frac{\partial s}{\partial \theta}\right]^{\perp} &= \delta \begin{bmatrix} \cos(\theta) \\ -\sin(\theta) \end{bmatrix} S^{-1/2}S^{1/2} \begin{bmatrix} \sin(\theta) \\ \cos(\theta) \end{bmatrix} \\
&= \delta \begin{bmatrix} \cos(\theta) \\ -\sin(\theta) \end{bmatrix} \begin{bmatrix} \sin(\theta) \\ \cos(\theta) \end{bmatrix} \\
&= \delta(\sin(\theta)\cos(\theta) - \sin(\theta)\cos(\theta)) \\
&= \delta(0) \\
&= 0
\end{aligned}$$

If s is a maximum or minimum solution, then there exist some θ such that

$$
\begin{aligned}
s(\theta)\left[\frac{\partial s}{\partial \theta}\right]^{\perp} &= 0 \\
\left(\delta S^{-1/2}\begin{bmatrix}\sin(\theta)\\ \cos(\theta)\end{bmatrix}+\bar{s}\right)^{T} S^{1/2}\begin{bmatrix}\sin(\theta)\\ \cos(\theta)\end{bmatrix} &= 0 \\
\left(\begin{bmatrix}\sin(\theta) & \cos(\theta)\end{bmatrix}(S^{-1/2})^{T}\delta+\bar{s}^{T}\right) S^{1/2}\begin{bmatrix}\sin(\theta)\\ \cos(\theta)\end{bmatrix} &= 0 \\
\delta\begin{bmatrix}\sin(\theta) & \cos(\theta)\end{bmatrix}\begin{bmatrix}\sin(\theta)\\ \cos(\theta)\end{bmatrix}+\bar{s}^{T}S^{1/2}\begin{bmatrix}\sin(\theta)\\ \cos(\theta)\end{bmatrix} &= 0 \\
\delta(\sin^{2}(\theta)+\cos^{2}(\theta))+\bar{s}^{T}S^{1/2}\begin{bmatrix}\sin(\theta)\\ \cos(\theta)\end{bmatrix} &= 0 \\
\delta+\bar{s}^{T}S^{1/2}\begin{bmatrix}\sin(\theta)\\ \cos(\theta)\end{bmatrix} &= 0 \\
\bar{s}^{T}S^{1/2}\begin{bmatrix}\sin(\theta)\\ \cos(\theta)\end{bmatrix} &= -\delta \\
\bar{s}^{T}S^{1/2}v &= -\delta \\
\bar{s}^{T}w &= -\delta \\
w^{T}\bar{s} &= -\delta
\end{aligned}
$$

□

Lemma 19. *The equations*

$$w^T \bar{s} = -\delta$$

$$\|w\|_{S^{-1}} = 1$$

$$w = \begin{bmatrix} w_1 & w_2 \end{bmatrix}^T$$

allow two solutions, w_+ and w_-, with $w_{+,1} \geq 0$, $w_{+,2} \leq 0$ and $w_{-,1} \leq 0$, $w_{-,2} \geq 0$. Furthermore, w_+ solves the max problem (B.1) with $\phi_{\max}(S) = \frac{\delta + w_{+,1}\bar{s}_1}{w_{+,1}\bar{s}_2}$, and w_- solves the min problem (B.2) with $\phi_{\min}(S) = \frac{\delta + w_{-,1}\bar{s}_1}{w_{-,1}\bar{s}_1}$.

Proof. The line $w^T\bar{s} = -\delta$ intersects the ellipse described by $\|w\|_{S^{-1}} = 1$ at zero, one or two points. If there is a point on the ellipse such that $w^T\bar{s} < -\delta$ then there are two points of intersection. Choose $w = \frac{-S\bar{s}}{\|S^{1/2}\bar{s}\|}$ which can be shown to be a point on the ellipse.

$$\begin{aligned}
\|w\|_{S^{-1}} &= \sqrt{\left(\frac{-S\bar{s}}{\|S^{1/2}\bar{s}\|}\right)^T S^{-1} \left(\frac{-S\bar{s}}{\|S^{1/2}\bar{s}\|}\right)} \\
&= \sqrt{\frac{(-\bar{s}^T S^T) S^{-1} (-S\bar{s})}{\left(\|S^{1/2}\bar{s}\|\right)^2}} \\
&= \left(\frac{1}{(\|S^{1/2}\bar{s}\|)}\right) \sqrt{(-\bar{s}^T S^T) S^{-1} (-S\bar{s})} \\
&= \frac{\sqrt{\bar{s}^T S \bar{s}}}{\sqrt{\bar{s}^T S \bar{s}}} \\
&= 1
\end{aligned}$$

Now show that $w^T\bar{s} < -\delta$,

$$\begin{aligned} w^T\bar{s} &= \left(\frac{-S\bar{s}}{\|S^{1/2}\bar{s}\|}\right)^T \bar{s} \\ &= \frac{-\bar{s}^T S\bar{s}}{\|S^{1/2}\bar{s}\|} \\ &= \frac{-\bar{s}^T S\bar{s}}{\sqrt{\bar{s}^T (S^{1/2})^T S^{1/2}\bar{s}}} \\ &= \frac{-\bar{s}^T S\bar{s}}{\sqrt{\bar{s}^T S\bar{s}}} \\ &= -\sqrt{\bar{s}^T S\bar{s}} \\ &= -\|S^{1/2}\bar{s}\| \\ &= \delta \end{aligned}$$

Therefore, there are two points of intersection, One solution, w_+, to (B.1) and one solution, w_-, to (B.2). These two solutions satisfy:

$$s_{\max} = \delta S^{-1} w_+ + \bar{s} \tag{B.3}$$

$$s_{\min} = \delta S^{-1} w_- + \bar{s} \tag{B.4}$$

From above, it is known that $w_+^T\bar{s} = -\delta$ and $\|w_+\|_{S^{-1}} = 1$. Therefore, w_+ is orthog-

onal to $s_{\max}$ as can be shown by:

$$\begin{aligned} w_+^T s_{\max} &= w_+^T(\delta S^{-1} w_+ + \bar{s}) \\ &= w_+^T \delta S^{-1} w_+ + w_+^T \bar{s} \\ &= \delta w_+^T S^{-1} w_+ - \delta \\ &= \delta \|w_+\|_{S^{-1}} - \delta \\ &= \delta - \delta \\ &= 0 \end{aligned}$$

If it is known that w_+ is orthogonal to $s_{\max}$ and w_- is orthogonal to $s_{\min}$, then

$$ratio(s_{\max}) = \frac{-1}{ratio(w_+)} = \frac{-w_{+,2}}{w_{+,1}} \tag{B.5}$$

and

$$ratio(s_{\min}) = \frac{-1}{ratio(w_-)} = \frac{-w_{+,2}}{w_{+,1}} \tag{B.6}$$

Since $s_{\max}$ and $s_{\min}$ are in the first quadrant, w_+ and w_- must be in the second or fourth quadrant. It is known from the definition of ratio that:

$$ratio(s_{\min}) \leq ratio(\bar{s}) \leq ratio(s_{\max}) \tag{B.7}$$

$$\frac{1}{-ratio(w_-)} \leq ratio(\bar{s}) \leq \frac{1}{-ratio(w_+)} \tag{B.8}$$

Therefore w_- must be in quadrant two and w_+ must be in quadrant four. It can also be said that

$$w^T \bar{s} = -\delta \tag{B.9}$$

$$w_{+,1}\bar{s}_1 + w_{+,2}\bar{s}_2 = -\delta \tag{B.10}$$

$$\frac{\bar{s}_1}{\bar{s}_2} + \frac{w_{+,2}}{w_{+,-}} = \frac{-\delta}{\bar{s}_2 w_{+,1}} < 0 \tag{B.11}$$

In order for (B.7) and (B.10) to be true, $w_{+,1} > 0$ and $w_{+,2} < 0$. Furthermore, from (B.10):

$$\begin{aligned} w_{+,1}\bar{s}_1 + w_{+,2}\bar{s}_2 &= -\delta \\ w_{+,2}\bar{s}_2 &= -\delta - w_{+,1}\bar{s}_1 \\ w_{+,2} &= \frac{-\delta - w_{+,1}\bar{s}_1}{\bar{s}_2} \end{aligned}$$

and $\phi_{\max}(S) = ratio(s_{\max})$. From (B.3):

$$s_{\max} = \delta S^{-1} w_+ + \bar{s}$$

From (B.5) it is known that

$$ratio(s_{\max}) = \frac{-w_{+,2}}{w_{+,1}} = \frac{\delta + w_{+,1}\bar{s}_1}{w_{+,1}\bar{s}_2} \tag{B.12}$$

The process is similar to show that $\phi_{\min}(S) = \frac{\delta + w_{-,1}\bar{s}_1}{w_{-,1}\bar{s}_2}$. □

Lemma 20. *$\phi_{\max}(S) = f(S)$ where*

$$f(S) = \begin{cases} \min_w & \frac{\delta + w_1\bar{s}_1}{w_1\bar{s}_2} \\ \textit{such that} & w^T\bar{s} = -\delta \\ & \|w\|_{S^{-1}} \leq 1 \\ & w_1 \geq 0 \\ & w_2 \leq 0 \end{cases} \tag{B.13}$$

Proof. The feasible points along hyperplane $w^T\bar{s} = -\delta$ in the fourth quadrant start with $w = \begin{bmatrix} 0 & -\frac{\delta}{\bar{s}_2} \end{bmatrix}^T$ where $ratio(w) = 0$. Then $ratio(w)$ will increase monotonically along $w^T\bar{s} = -\delta$ to a boundary of the ellipse described by $\|w\|_{S^{-1}} \leq 1$. Therefore the solution must satisfy $\|w\|_{S^{-1}} = 1$, which describes the ellipse boundary. By Lemma 19, the solution, (B.12), satisfies $w^T\bar{s} = -\delta$, $\|w\|_{S^{-1}} = 1$, $w_1 \geq 0$, and $w_2 \leq 0$, so the minimizing w must be w_+ and $f(S) = \frac{\delta + w_{+,1}\bar{s}}{w_{+,1}\bar{s}_2} = \phi_{\max}(S)$. □

Lemma 21. *$\phi_{\max}(S)$ is convex and $\phi_{\min}(S)$ is concave in S.*

Proof. By Lemma 20 $\phi_{\max}(S)$ is convex if $f(S)$ is convex. Let $g(w) = \frac{\delta + w_1\bar{s}_1}{w_1\bar{s}_2}$. It is known that $g(w)$ is convex in w for $w_1 > 0$. This can be proved by looking at the second derivative. If $g(w)$ is twice differentiable, $g(w)$ is convex if and only if

$g''(w) \geq 0$. This can be shown with the following sequence:

$$\begin{aligned} g(w) &= \frac{\delta}{w_1 \bar{s}_2} + \frac{w_1 \bar{s}_1}{w_1 \bar{s}_2} \\ &= \frac{\delta}{w_1 \bar{s}_2} + \frac{\bar{s}_1}{\bar{s}_2} \\ \frac{\partial g(w)}{\partial w_1} &= \left(\frac{\delta}{\bar{s}_2}\right)\left(\frac{-1}{w_1^{-2}}\right) \\ \frac{\partial^2 g(w)}{\partial w_1^2} &= \left(\frac{2\delta}{\bar{s}_2}\right)\left(\frac{1}{w_1^{-3}}\right) \\ \frac{\partial^2 g(w)}{\partial w_1^2} &= \frac{2\delta}{\bar{s}_2 w_1^{-3}} \end{aligned}$$

Since δ and $\bar{s}_2$ are greater than zero, if $w_1 > 0$ then $\frac{\partial^2 g(w)}{\partial w_1^2} > 0$ and therefore $g(w)$ is convex for $\bar{s}_1,\ w_1 > 0$. If $f(S') = g(w')$ and $f(S'') = g(w'')$ then a feasible solution for $f(\alpha S' + (1-\alpha)S'')$ is $\alpha w' + (1-\alpha)w''$ where $\alpha \in [0,1]$. Due to Lemma 20,

$$f(\alpha S' + (1-\alpha)S'') \leq g(\alpha w' + (1-\alpha)w'').$$

Since $g(w)$ is convex, $g(\alpha w' + (1-\alpha)w'') \leq \alpha g(w') + (1-\alpha)g(w'')$, and therefore

$$\begin{aligned} f(\alpha S' + (1-\alpha)S'') &\leq \alpha g(w') + (1-\alpha)g(w'') \\ f(\alpha S' + (1-\alpha)S'') &\leq \alpha f(S') + (1-\alpha)f(S'') \end{aligned}$$

Therefore $\phi_{\max}(S)$ is convex in S. □

Using Lemmas 18 - 21 established above, we can now proceed to the proof of

Theorem 17. By Lemma 19: two solutions satisfy $w^T\bar{s} = -\delta$ and $\|w\|_{S^{-1}} = 1$. Thus

$$w^T S^{-1/2} S^{1/2} \bar{s} = \|S^{-1/2}w\| \|S^{1/2}\bar{s}\| \cos(\theta)$$

where θ is the angle between $S^{-1/2}w$ and $S^{1/2}\bar{s}$ or $\cos(\theta) = \frac{-\delta}{\|S^{1/2}\bar{s}\|}$. Therefore,

$$\begin{aligned}
w^T S^{-1/2} S^{1/2} \bar{s} &= \|S^{-1/2}w\| \|S^{1/2}\bar{s}\| \cos(\theta) \\
&= \left(\sqrt{w^T (S^{-1/2})^T S^{-1/2} w}\right) \left(\sqrt{\bar{s}^T (S^{1/2})^T S^{1/2} \bar{s}}\right) \cos(\theta) \\
&= \left(\sqrt{w^T S^{-1} w}\right) \left(\sqrt{\bar{s}^T S \bar{s}}\right) \cos(\theta) \\
&= \|w\|_{S^{-1}} \left(\sqrt{\bar{s}^T S \bar{s}}\right) \cos(\theta) \\
&= \left(\sqrt{\bar{s}^T S \bar{s}}\right) \cos(\theta) \\
&= \left(\sqrt{\bar{s}^T S \bar{s}}\right) \frac{-\delta}{\|S^{1/2}\bar{s}\|} \\
&= -\delta
\end{aligned}$$

A solution to $S^{-1/2}w$ can be found by rotating a unit vector in the $S^{1/2}\bar{s}$ direction by θ. There are two solutions for θ in $[-\pi, \pi]$: w_+ in the counterclockwise (positive) direction and w_- in the clockwise direction. The matrix $\begin{bmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{bmatrix}$ rotates a two element vector by θ. Therefore $S^{1/2}w_+$ can be found using the following where

$\gamma = \|S^{1/2}\bar{s}\|$ so that $\cos(\theta) = \frac{-\delta}{\gamma}$ and $\sin(\theta) = \sqrt{1 - \frac{\delta^2}{\gamma^2}}$:

$$
\begin{aligned}
S^{-1/2}w_+ &= \begin{bmatrix} \frac{-\delta}{\gamma} & \sqrt{1-\frac{\delta^2}{\gamma^2}} \\ -\sqrt{1-\frac{\delta^2}{\gamma^2}} & \frac{-\delta}{\gamma} \end{bmatrix} \frac{S^{1/2}\bar{s}}{\gamma} \\
w_+ &= S^{1/2} \begin{bmatrix} -\delta & \sqrt{\gamma^2-\delta^2} \\ -\sqrt{\gamma^2-\delta^2} & -\delta \end{bmatrix} \frac{S^{1/2}\bar{s}}{\gamma^2} \\
w_+ &= -S^{1/2} \begin{bmatrix} \delta & -\sqrt{\gamma^2-\delta^2} \\ \sqrt{\gamma^2-\delta^2} & \delta \end{bmatrix} \frac{S^{1/2}\bar{s}}{\gamma^2}
\end{aligned}
$$

Now all that is left is to prove the quasi-linearity of $\phi_{\max}(P^{-1}, \bar{s}, \delta)$ in P. $(Cs + s) \leq t$ defines a half-space T in s. $\phi_{\max}(P^{-1}, \bar{s}, \delta) \leq t$ if and only if $\mathcal{E}(P^{-1}, \bar{s}, \delta) \in T$. Second, note that $\phi^{C,d}_{\max}(P^{-1}, \bar{s}, \delta) \leq t$ if and only if the ellipse defined by $(P^{-1}, \bar{s}, \delta)$ is contained in the half space defined by $ratio(Cs + d) \leq t$. By Corollary 16, $\phi^{C,d}_{\max}(P^{-1}, \bar{s}, \delta) \leq t$ defines a convex set in P, and so $\phi^{C,d}_{\max}$ is quasi-convex.

It can also be shown that $\phi^{C,d}_{\max}(P^{-1}, \bar{s}, \delta)$ is quasi-concave in P: First, note that $ratio_{C,d}(s)$ is a quasi-concave function of s when $[Cs + d]_2 > 0$. Consider the optimization problem (4.11) that defines $\phi^{C,d}_{\max}$. Let s_1 be the maximizer achieved for $(P_1, \bar{s})$ and s_2 the maximizer achieved for P_2. Since the constraints $(s-\bar{s})^T P^{-1}(s-\bar{s}) \leq \delta$ are convex in s and P, the vector $s_\theta = \theta s_1 + (1-\theta)s_2$ is feasible for $P_\theta = \theta P_1 + (1-\theta)P_2$

where $\theta \in [0,1]$. Thus

$$\begin{aligned}
\phi_{\max}^{C,d}(P_\theta^{-1}, \bar{s}, \delta) &\geq ratio_{C,d}(\theta s_1 + (1-\theta)s_2) \\
&\geq \min\left(ratio_{C,d}(s_1), ratio_{C,d}(s_2)\right) \\
&\geq \min(\phi_{\max}^{C,d}(P_1^{-1}, \bar{s}, \delta), \phi_{\max}^{C,d}(P_2^{-1}, \bar{s}, \delta))
\end{aligned}$$

Where the second line follows by the quasi-concavity of $ratio_{C,d}(s)$. Thus $\phi_{\max}^{C,d}$ is both quasi-convex and quasi-concave (i.e. quasi-linear). The quasi-linearity of $\phi_{\min}^{C,d}$ can be established similarly.