THE TREATMENT OF MISSING MEASUREMENTS

IN

PCA AND PLS MODELS

By

PHILIP R. C. NELSON, B.A.SC., M. ENG.

A Thesis

Submitted to the School of Graduate Studies

In Partial Fulfilment of the Requirements

For the Degree

Doctorate of Philosophy

McMaster University

TREATMENT OF MISSING MEASUREMENTS IN PCA AND PLS MODELS

Doctorate of Philosophy (2002)
(Chemical Engineering)

McMaster University
Hamilton, Ontario

TITLE: The Treatment Of Missing Measurements In PCA And PLS Models

AUTHOR: Philip R. C. Nelson, B.A.Sc. (University of Toronto), M. Eng.
(McMaster University)

SUPERVISORS: Professor P. A. Taylor and Professor J. F. MacGregor

NUMBER OF PAGES: xiv, 162

Abstract

This thesis investigates the building and application of principal components analysis (PCA) and projection to latent structures (PLS) models when some objects in the data set have missing measurements. Score calculation is treated first, followed by model application to prediction and monitoring. Model building is explored in the final part of the thesis.

The first problem treated in this work is that of estimating scores from an existing PCA or PLS model when new observation vectors are incomplete. Several methods for estimating scores from data with missing measurements are presented and analysed, including a novel method involving data replacement by the conditional mean. Expressions are developed for the error in the scores calculated by each method and the factors that lead to error are drawn from these expressions. The error analysis is illustrated using simulated data sets designed to highlight problem situations. A larger industrial data set and a simulated process data set are also used to compare the approaches. In general, all the methods perform reasonably well with moderate amounts of missing data (up to 20% of the measurements). However, in extreme cases where critical combinations of measurements are missing, the novel method is generally superior to the other approaches.

Uncertainty intervals arising from missing measurements are then developed for the squared prediction error (SPE), Hotelling T^2, PLS predictions and their contributions. These uncertainty intervals provide performance measures and diagnostics when measurements are missing in process monitoring and prediction. The uncertainty intervals derived agree well with the values calculated from the complete objects and give valuable information about the true level of knowledge about the process. The uncertainty in the contributions is used to correctly diagnose which missing measurement in a set gives the greatest reduction in uncertainty when it is recovered.

Insight gained in the analysis of score estimation and model application is applied to model building with the nonlinear iterative partial least squares (NIPALS), maximum likelihood principal components analysis (MLPCA), expectation maximisation (EM) and iterative replacement algorithms. Challenges unique to model building and factors that lead to error in individual steps of each model building algorithm are examined. Recommendations are made to improve the quality of the models obtained, and a procedure is proposed to screen data for objects and variables with missing measurements that would have an adverse effect on a model built using missing measurements.

Acknowledgements

I would like to thank Professors John MacGregor and Paul Taylor for their patience and hard work in supervising my thesis. I would also like to thank Professor Jim Reilly for providing a stimulating alternative point of view on my supervisory committee. I would like to thank Svante Wold for valuable information about the algorithms during the initial stages of this work. Finally, I would like to express my appreciation to my wife Katherine for all her support and for believing in this because I did.

Table Of Contents

Table of Figures

Table of Tables

Chapter 1: Introduction and Background Theory

1.1 Introduction

This thesis investigates the use of Principal Components Analysis (PCA) and Projection to Latent Structures (PLS) models when some of the measurements in the data set are missing. A novel missing measurement score calculation algorithm is proposed and the factors affecting the performance of both the novel and existing algorithms are identified and illustrated by designed and industrial process data sets. Performance measures for the application of PCA and PLS models are derived and illustrated on the industrial data set. These measures distinguish between situations where model performance with missing measurements will continue to be acceptable and situations where it will become unacceptable. In the latter case the measurements must be recovered or the application shut down. Diagnostics are developed to aid in determining which missing measurements are causing the greatest uncertainty in the application and will yield the greatest reduction in uncertainty by their recovery.

There are many reasons why measurements may be missing from a data set. Missing measurements can occur when sensors fail or are taken off-line for routine maintenance. In other situations, measurements are removed from a data set because gross measurement errors occur, manual samples are simply not

collected at the required time or sensors have different sampling periods. The measurements may be missed at random times, as for missed manual samples and sensor failures, or on a very regular basis, as for sensors with different sampling periods.

PCA and PLS have been widely used to develop models from data sets composed of observations on large numbers of highly correlated variables. Once a model has been built, it can be applied to future process data in inferential control schemes to predict process responses (Kresta et al., 1994 and Medjell and Skogestad, 1991), or in multivariate statistical process control schemes to monitor and diagnose future process operating performance (Kresta et al., 1991; Nomikos and MacGregor, 1994; Nomikos and MacGregor, 1995; Skagerberg et al., 1992; Wise et al., 1991; Wise and Ricker, 1991). In many of these situations, particularly those involving industrial processes, missing measurements are a common occurrence. To insist on using only complete data sets when building or applying PCA or PLS models would entail throwing away large amounts of the data. In an online application the sampling period would have to be the greatest common multiple of the measurement sampling periods and the application would have to shut down whenever any measurement was unavailable. This would reduce the benefits of the application or increase the capital and maintenance costs to meet a desired level of availability. Therefore, it is important that efficient methods for handling missing data be available, both for using such multivariate

models on future data and for analysing and building multivariate models from past data.

The balance of this chapter will introduce the nomenclature, literature and theory used in the thesis. Basic vector and matrix notation is introduced, followed by PCA and PLS model structures and properties. Model building, score calculation and application of models to prediction and monitoring with complete data are then presented. The following section presents these topics when objects have missing measurements. The final section introduces the theory necessary to calculate the cumulative distribution of quadratic forms of random variables.

Chapter 2 analyses the algorithms that are used for handling missing measurements when the underlying PCA or PLS model is assumed to be fixed and known. This is the situation when a PCA or PLS model has been built from a large amount of plant data and it is to be applied using new observations that have missing measurements. A novel algorithm for score calculation with missing measurements is developed and its properties analysed and compared with those of existing algorithms. The sources of error for each algorithm are laid out and illustrated using designed data sets. This gives specific quantities calculated from information available from model building to determine score calculation performance with each missing measurement algorithm. Recommendations are made on pruning variables and on the number of components to use in a model for good performance with missing measurements. An analysis of two process data sets from the literature, one simulated and the other industrial, is presented to

show application to realistic situations. While all methods perform well in most cases, it is shown that the novel method is best in certain critical situations identified by the analysis.

In chapter 3, we develop performance measures for PCA and PLS model applications which reveal the level of uncertainty caused by missing measurements and give diagnostics to assist in determining which of the missing measurements would give the maximum benefit if it were recovered. These measures distinguish between situations where model performance with missing measurements will continue to be acceptable and situations where it will become unacceptable. In the latter case, the measurements must be recovered or the application shut down. Current practice is to use the scores calculated using the missing measurement algorithms in the same manner as scores from complete objects. The performance measures are uncertainty intervals on the predictions, sum squared prediction error (SPE), Hotelling T^2 and scores. The prediction uncertainty interval combines the variance from parameter estimation together with the uncertainty arising from the missing measurements with the missing measurement uncertainty dominating in the example. All other uncertainty intervals developed are due to missing measurement uncertainty alone. Uncertainty intervals for the contributions to these quantities are developed to aid in determining which missing measurements play the largest role in the uncertainty interval.

Model building with missing measurements and model building for applications that need to be robust to missing measurements are considered in chapter 4. Factors that affect the NIPALS, EM, MLPCA and Iterative Replacement model building algorithms are developed and improvements to the MLPCA algorithm proposed. A guide for pruning variables during model building is given which considers the resulting model's robustness to missing measurements and a procedure proposed to apply the analysis of chapters 2 and 3 to improve model building with missing measurements. Also, the issues unique to model building are reviewed and how they apply to the analysis in the rest of this chapter discussed.

1.2 Nomenclature

This section will define the basic vector and matrix notation used in this thesis and the notation used to indicate missing and present measurements. Some notation concerning PCA and PLS will be briefly introduced as it relates to missing measurements but the notation and properties of PCA and PLS are defined in section 1.3.

Lower case bold variables, both Roman and Greek, are column vectors and upper bold case ones are matrices. Where an upper case Roman character is used, the lower case of that letter is used to represent the columns of that matrix with a subscript indicating the column index. So $\mathbf{X} = \begin{bmatrix} \mathbf{x}_1, & \mathbf{x}_2, & \cdots & \mathbf{x}_K \end{bmatrix}$ where $\mathbf{x}_k$ is the k^{th} column of $\mathbf{X}$. An individual element of a matrix is indicated by the

same character as the matrix but in lower case with two subscripts; the first subscript for the row number and the second for the column number. As an example, x_{ik} is the element of matrix $\mathbf{X}$ in the i^{th} row and k^{th} column. A subscript of two characters separated by a colon indicates which columns of a matrix or rows of a column vector are taken. The matrix $\mathbf{P}_{1:A}$ contains columns 1 to A of the matrix $\mathbf{P}$ and $\boldsymbol{\tau}_{A+1:K}$ contains the elements from rows A+1 to K of the column vector $\boldsymbol{\tau}$.

In some instances, a vector or matrix will change at each iteration of an algorithm. The index for the iteration of the algorithm will then be indicated in round brackets after the variable so that $\mathbf{x}(a)$ will be the vector $\mathbf{x}$ at the a^{th} iteration of the algorithm.

A multivariate vector of measurements is denoted $\mathbf{z}^T = [z_1, \ z_2, \ \cdots \ z_K]$ with K being the number of variables. A data matrix $\mathbf{X}$ is then a collection of N row vectors $\mathbf{z}_i^T$ (objects) or K column vectors $\mathbf{x}_k$ (variables).

$$\mathbf{X} = \begin{bmatrix} \mathbf{z}_1^T \\ \mathbf{z}_2^T \\ \vdots \\ \mathbf{z}_N^T \end{bmatrix} = [\mathbf{x}_1, \ \mathbf{x}_2, \ \cdots \ \mathbf{x}_K]$$

The letter i is used exclusively as an index for the objects and k exclusively for variables.

It will be shown in section 1.3 that score matrices have a row for each object and a column for each component. The matrix of scores is expressed as

$$\mathbf{T} = \begin{bmatrix} \boldsymbol{\tau}_1^T \\ \boldsymbol{\tau}_2^T \\ \vdots \\ \boldsymbol{\tau}_N^T \end{bmatrix} = [\mathbf{t}_1, \quad \mathbf{t}_2, \quad \cdots \quad \mathbf{t}_A]$$

where the $\boldsymbol{\tau}_i^T$ are row vectors corresponding to the values of the scores for the i^{th} observation and the $\mathbf{t}_j$ are column vectors corresponding to the scores for the j^{th} component. The letter j is used exclusively as an index for the components.

Loading matrices, such as P or W, have a row for each variable and a column for each component as will be shown in section 1.3. The matrix is a collection of column vectors $\mathbf{p}_j$.

$$\mathbf{P} = [\mathbf{p}_1, \quad \mathbf{p}_2, \quad \cdots \quad \mathbf{p}_A]$$

A shorthand notation for various matrices and vectors with elements corresponding to either missing or present measurements removed is needed. A single superscript asterisk indicates a vector or matrix with elements corresponding to missing measurements removed. It will be shown in section 1.3 that these elements are rows for vectors of measurements such as $\mathbf{z}$ or loading matrices such as P or W so if variables 2 and 5 are missing then $\mathbf{z}^*$ is obtained from $\mathbf{z}$ by removing the 2^{nd} and 5^{th} rows and $\mathbf{P}^*$ is obtained from P in the same manner. A single superscript pound (#) indicates a vector or matrix with elements corresponding to measured variables removed. The missing data can be taken to

be the first elements of the data vector without loss of generality, so the vector $\mathbf{z}$ can be partitioned as

$$\mathbf{z}^T = \left[\mathbf{z}^{\#T}, \quad \mathbf{z}^{*T}\right]$$

where $\mathbf{z}^{\#}$ denotes the missing measurements and $\mathbf{z}^{*}$ the present measurements. Correspondingly, the $\mathbf{P}$ matrix can be partitioned as $\mathbf{P}^T = \left[\mathbf{P}^{\#T}, \quad \mathbf{P}^{*T}\right]$.

The estimated covariance matrix $\mathbf{S}$ is a special case since both its rows and columns are related to specific measurements. It thus has two superscripts, the first for the rows and the second for the columns. Thus if the missing measurements are placed first in the matrix:

$$\mathbf{S} = \begin{bmatrix} \mathbf{S}^{\#\#} & \mathbf{S}^{\#*} \\ \mathbf{S}^{*\#} & \mathbf{S}^{**} \end{bmatrix}$$

The expression $\angle \mathbf{p}_i, \mathbf{p}_j$ is used to denote the angle between the two vectors in degrees.

1.3 PCA and PLS Model Structure and Properties

PCA and PLS are used to develop models from data sets composed of observations on large numbers of highly correlated variables. The structures of the two models and their properties will be shown together to highlight the similarities and differences. Theory and applications can be found in Wold et al. (1987) and Jackson (1991) for PCA and Wold et al. (1983) and Martens and Naes (1989) for PLS.

PCA is applied to a single data matrix **X** and PLS is applied to an input data matrix **X** and an output data matrix **Y**. The model structure for the matrix **X** for both methods is:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E}$$

and the model structure for the output matrix **Y** for PLS is:

$$\mathbf{Y} = \mathbf{T}\mathbf{Q}^T + \mathbf{F}$$

T is an N by A matrix of scores and **P** is a K by A matrix of loadings as stated in section 1.2. The number of dimensions of any PCA or PLS model in this work is A. PLS has a second K by A matrix of loadings **W** that is required to calculate the scores as shown in section 1.3.2. The scores are new variables that are defined by the loading matrices. They are defined to be independent and at each step the loading vector $\mathbf{p}_j$ for PCA is chosen to give maximum reduction in the variance of **X**. For PLS the component is chosen to give the maximal reduction in the covariance $\mathbf{X}^T\mathbf{Y}$. One motivation for using PCA and PLS is that usually A can be chosen to be much less than K and still represent the variation in the data well, giving fewer variables to treat. Proper choice of A will also remove rank deficiency or ill-conditioning in **X**. Another reason to use these models is to study the relationships between the variables by examination of the loading matrices.

The properties of the PCA and PLS matrices relevant to the work in this thesis will be listed here; additional properties can be found in Jackson (1991) for PCA and Hoskuldsson (1988) for PLS. For both PCA and PLS,

$$\mathbf{t}_m^T \mathbf{t}_j = 0 \quad m \neq j$$

For PCA,

$$\mathbf{p}_m^T \mathbf{p}_j = 0 \quad m \neq j$$
$$\mathbf{p}_m^T \mathbf{p}_n = 1 \quad m = n$$

and for PLS

$$\mathbf{w}_m^T \mathbf{w}_j = 0 \quad m \neq j$$
$$\mathbf{w}_m^T \mathbf{w}_j = 1 \quad m = j$$
$$\mathbf{w}_m^T \mathbf{p}_j = 0 \quad m < j$$

1.3.1 Model Building

There are a number of algorithms for calculating PCA and PLS models when no measurements are missing. A PCA model can be calculated using a power method algorithm or by doing an SVD of the **X** matrix. Standard notation for an SVD is

$$\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^T$$

and in the nomenclature of PCA $\mathbf{P} = \mathbf{V}_{1:A}$ and **T** is the first A columns of the product of **U** and **S**.

A PLS model can be calculated by the nonlinear iterative partial least squares (NIPALS) algorithm (a power method) or by one of the kernel methods published in Dayal and MacGregor (1997) or Rannar et al. (1994a) or de Jong and Ter Braak (1994). The NIPALS algorithm will be covered in section 1.5.1.2.

1.3.2 Score Calculation

Once a model has been calculated, scores $\boldsymbol{\tau}^T = [\tau_1, \quad \tau_2, \quad \cdots \quad \tau_A]$ for new objects **z** can be calculated using the following formula for PCA

$$\tau_j = \mathbf{z}^T(j)\mathbf{p}_j$$

For PLS the **W** matrix is used in calculating the score instead of **P**

$$\tau_j = \mathbf{z}^T(j)\mathbf{w}_j$$

And for both methods

$$\mathbf{z}(1) = \mathbf{z}$$

$$\mathbf{z}(j+1) = \mathbf{z}(j) - \tau_j \mathbf{p}_j$$

The calculation of all score values can also be done in one step, as

$$\boldsymbol{\tau} = \mathbf{R}^T\mathbf{z}$$

where $\mathbf{R} = \mathbf{P}$ for PCA and $\mathbf{R} = \mathbf{W}\left(\mathbf{P}^T\mathbf{W}\right)^{-1}$ for PLS. This form for calculating the scores allows the contribution to the scores to be defined directly. Each score has the form

$$\tau_j = \sum_{k=1}^{K} r_{kj} z_k$$

The contribution of variable k to score j is the term in the calculation of the score that involves that variable which can be seen from the equation above to be:

$$r_{kj} z_k \qquad \text{Equation 1-1}$$

1.4 Applications of PCA and PLS

1.4.1 Predictions

PLS is used to provide predictions for the variables in the **Y** matrix. The prediction equation with a new vector of measurements **z** is

$$\hat{\mathbf{y}} = \mathbf{Q}\boldsymbol{\tau} \qquad \text{Equation 1-2}$$

In this work we will consider the case where $\mathbf{Q}$ is a row vector. This is the case where there is only one variable being predicted or where the variable interactions in a multivariable model are ignored. In terms of the original variables the prediction for one variable is

$$\hat{\mathbf{y}} = \mathbf{Q}\boldsymbol{\tau} = \mathbf{Q}\mathbf{R}^T\mathbf{z} = \boldsymbol{\beta}^T\mathbf{z} \qquad \text{Equation 1-3}$$

with $\boldsymbol{\beta} = \mathbf{R}\mathbf{Q}^T$.

There are several papers dealing with PLS prediction error variance (Phatak et al., 1993; Hoy et al., 1998; Faber, 2000). The following definition is useful in several of the expressions:

$$h = \boldsymbol{\tau}^T\left[\mathbf{T}^T\mathbf{T}\right]^{-1}\boldsymbol{\tau}$$

Phatak et al. linearized the PLS estimator by truncating a Taylor series expansion for the estimator after a single term to develop an expression for the variance of the predictions. The expression assumes that the matrices are mean centred and is

$$\mathrm{var}\left(y - \hat{y}\right) = \mathrm{var}\left(\mathbf{z}^T\boldsymbol{\beta}\right) = \mathbf{z}^T\,\mathrm{var}\left(\boldsymbol{\beta}\right)\mathbf{z} = \mathbf{z}^T\mathbf{J}_o\mathbf{J}_o^T\mathbf{z}\sigma^2$$

$$\mathbf{J} = \left\{\left[\mathbf{s}^T\mathbf{V}_m\mathbf{R}_m \otimes \left(\mathbf{I}_p - \mathbf{H}_m\mathbf{S}\right)\right] + \left[\mathbf{V}_m\mathbf{R}_m \otimes \mathbf{s}^T\left(\mathbf{I}_p - \mathbf{H}_m\mathbf{S}\right)\right]\right\} \left(\left(\mathbf{M}^T\right)^{-1} \otimes \mathbf{I}_p\right)\mathbf{U}_m^T + \mathbf{H}_m\mathbf{X}^T$$

where m is the number of dimensions in the model, $\otimes$ denotes the Kronecker product and we can set $\mathbf{V}_m = \mathbf{W}$ without loss of generality. The other vectors and matrices are

$$\mathbf{s} = \mathbf{X}^T\mathbf{y}$$
$$\mathbf{S} = \mathbf{X}^T\mathbf{X}$$
$$\mathbf{R}_m = \left[\mathbf{V}_m^T\mathbf{S}\mathbf{V}_m\right]^{-1}$$
$$\mathbf{H}_m = \mathbf{V}_m\mathbf{R}\mathbf{V}_m^T$$
$$\mathbf{U}_m = \left[\mathbf{X}\middle|\mathbf{X}\mathbf{S}\middle|\cdots\middle|(\mathbf{X}\mathbf{S})^{(m-1)}\right]$$

and the matrix **M** satisfies

$$\left[\mathbf{X}^T\mathbf{y}\middle|(\mathbf{X}^T\mathbf{X})\mathbf{X}^T\mathbf{y}\middle|\ldots\middle|(\mathbf{X}^T\mathbf{X})^{(m-1)}\mathbf{X}^T\mathbf{y}\right] = \mathbf{V}_m\mathbf{M}$$

The expression in Hoy et al. (1998) with the correction due to De Vries and Ter Braak (1995) is

$$\mathrm{var}(y-\hat{y}) = \left[1-\frac{A+1}{I_{cal}}\right]\left[h+\frac{V_x}{V_{x,cal}}+\frac{2}{I_{cal}}\right]V_y \qquad \text{Equation 1-4}$$

where V_y is the mean squared prediction error of y, A is the number of components in the model, I_{cal} is the number of objects in the data set used to build the model, V_X is the mean squared measurement error and $V_{X,cal}$ is the mean squared measurement error of the modelling data set. Equation 1-4 is composed of four terms when V_y is multiplied into the second set of brackets:

$$1-\frac{A+1}{I_{cal}} \qquad \text{Equation 1-5}$$

$$hV_y = \boldsymbol{\tau}^{\mathrm{T}}\left[\mathbf{T}^{\mathrm{T}}\mathbf{T}\right]^{-1}\boldsymbol{\tau}V_y = \boldsymbol{\tau}^{\mathrm{T}}\,\mathrm{var}(\mathbf{Q})\boldsymbol{\tau} \qquad \text{Equation 1-6}$$

$$\frac{V_x}{V_{x,cal}}V_y \qquad \text{Equation 1-7}$$

$$\frac{2}{I_{cal}}V_y \qquad \text{Equation 1-8}$$

Equation 1-5 is a correction factor due to De Vries and Ter Braak (1995), Equation 1-6 is the variance assuming the scores are without error, Equation 1-7

is the variance assuming the loadings are without error and Equation 1-8 is the variance due to error in the mean estimate.

The expression in Faber (2000) results from an error-in-variables approach and is

$$\mathrm{var}(\hat{y}-y) = h\left[\mathrm{var}(\varepsilon)+\mathrm{var}(\Delta\mathbf{y})+\|\boldsymbol{\beta}\|^2\,\mathrm{var}(\Delta\mathbf{X})\right]+\mathrm{var}(\varepsilon) + \|\boldsymbol{\beta}\|^2\,\mathrm{var}(\Delta\mathbf{X}) \qquad \text{Equation 1-9}$$

where ε is the unmodelled part of y, $\boldsymbol{\beta}$ is the vector of PLS regression coefficients and Δy and $\Delta\mathbf{X}$ are measurement errors in the output and input respectively. The expression above has been simplified by assuming that the errors in the model building and prediction data are the same.

The differences in these expressions for the prediction variance are due to the different assumptions made about measurement and other errors in the underlying structure of the data and model. PLS was originally based on an algorithm, rather than an assumed data structure and objective function, so it is not clear which of these expressions best represents the variance of the PLS predictions.

1.4.2 Monitoring

PCA and PLS are used in statistical process monitoring due to their ability to handle large sets of correlated variables with singular or nearly singular covariance matrices. These methods extract information about how the variables change relative to one another and reduce the noise level due to averaging (Kourti

and MacGregor, 1995). This increases the effectiveness of monitoring as well as reducing the number of plots to monitor.

The PCA or PLS model defines new (latent) variables that are used for monitoring. The variation in these new variables and the residual variation in the original variables must be monitored. This is done by monitoring the SPE and Hotelling T^2 as described in the following sections.

1.4.2.1 Monitoring Model Residual Variation

The squared prediction error (SPE) for **X** gives a measure of the distance of the new observation from the space defined by the model. It is defined as

$$\begin{aligned} SPE &= \left[\mathbf{z}-\mathbf{P}\hat{\boldsymbol{\tau}}\right]^T\left[\mathbf{z}-\mathbf{P}\hat{\boldsymbol{\tau}}\right] \\ &= \left[\mathbf{z}-\mathbf{PRz}\right]^T\left[\mathbf{z}-\mathbf{PRz}\right] \\ &= \mathbf{z}^T\left[\mathbf{I}-\mathbf{PR}\right]^T\left[\mathbf{I}-\mathbf{PR}\right]\mathbf{z} \end{aligned} \qquad \text{Equation 1-10}$$

Approximate expressions for the control limits for the SPE have been laid out in Jackson and Mudholkar (1979) and Nomikos and MacGregor (1995). Since the SPE is the sum of the squared input matrix residuals, the contributions of the variables to the SPE are the residuals themselves.

1.4.2.2 Monitoring Variation in the Scores

The Hotelling T^2 is a statistic that can be plotted to determine when there has been a shift in the mean of the scores. It is calculated as

$$Hotelling\ T^2 = \boldsymbol{\tau}^T\left[\mathbf{T}^T\mathbf{T}\right]^1\ \boldsymbol{\tau} \qquad \text{Equation 1-11}$$

and the control limit is provided in Kourti and MacGregor (1995). The contributions to the Hotelling T^2 are the scores. Kourti and MacGregor (1996) define the Hotelling T^2 as

$$Hotelling\ T^2 = \sum_{k=1}^{K}\sum_{j=1}^{A}\frac{\tau_j}{E\left(\tau_j^2\right)}r_{kj}z_k \qquad \text{Equation 1-12}$$

so the contribution of variable k to the Hotelling T^2 is the inner summation over j for a fixed k.

If only two scores are to be monitored then the score-score plot can be used (Kresta et al., 1991).

1.5 PCA and PLS with Missing Measurements

1.5.1 Model Building

1.5.1.1 Complete Object Method

The complete object method consists of removing all objects that have missing measurements and using the remaining objects to build the model. The assumption is made that the complete objects provide a representative sample of the process data. The performance of this method will be poor when this is not true. When there is an abundance of data, the performance of that method is acceptable and its computational complexity is low. This method sets a minimum performance standard for the other methods.

1.5.1.2 NIPALS

The standard procedure for handling missing data in PCA during model building is based on H. Wold's NIPALS algorithm (Wold, 1966 and Geladi and Kowalski, 1986). In PCA model building, one iteration of the NIPALS algorithm consists of a linear regression of the columns of $\mathbf{X}$ on a score vector $\mathbf{t}$ to obtain a loading vector $\mathbf{p}$, followed by a linear regression of the rows of $\mathbf{X}$ on the loading vector to obtain a new estimate of $\mathbf{t}$. Convergence is reached when the mean square change in the scores falls below a threshold. When data in any column or row of $\mathbf{X}$ are missing, the iterative regressions are performed using the data that is present and the missing points are ignored (Wold et al., 1983 and Martens and Naes, 1989). This approach was first used by H. Wold in the analysis of horse racing results from the American Jockey Club around 1964 (Wold, 1995). This procedure can be interpreted in different ways. It is equivalent to setting the residuals for all missing elements in the least squares objective function to zero in each iteration. It can also be interpreted as replacing the missing values by their minimum distance projections onto the current estimate of the loading or score vector at each iteration.

As long as the number of variables present in any row or column is greater than or equal to the number of scores to be calculated then the NIPALS algorithm can obtain a solution. However, in practice one should have many more observations than the scores or loadings being estimated to obtain reliable results. A rule of thumb is to have 5 times as many observations in any row or column as

the number of dimensions (A) being calculated (Wold, 1995). The NIPALS algorithm is usually recommended only when the missing data pattern is random rather than structured (Wold, 1995). An example of a structured missing data pattern is one resulting from sensors with different sampling periods. In these cases, the measurements are missing in blocks.

With l representing the indices of missing measurements in the summations, one iteration of the NIPALS algorithm for the j^{th} component in PCA is then:

$$p_{kj} = \sum_{\substack{i=1 \\ i \neq l}}^{N} x_{ik} t_{ij} \Bigg/ \sum_{\substack{i=1 \\ i \neq l}}^{N} t_{ij}^2$$

$$t_{ij} = \sum_{\substack{k=1 \\ k \neq l}}^{N} x_{ik} p_{kj} \Bigg/ \sum_{\substack{k=1 \\ k \neq l}}^{N} p_{kj}^2 \qquad \text{Equation 1-13}$$

and for PLS it is:

$$w_{kj} = \sum_{\substack{i=1 \\ i \neq l}}^{N} x_{ik} u_{ij} \Bigg/ \sum_{\substack{i=1 \\ i \neq l}}^{N} u_{ij}^2$$

$$t_{ij} = \sum_{\substack{k=1 \\ k \neq l}}^{K} x_{ik} w_{kj} \Bigg/ \sum_{\substack{k=1 \\ k \neq l}}^{K} w_{kj}^2 \qquad \text{Equation 1-14}$$

$$q_{mj} = \sum_{\substack{i=1 \\ i \neq l}}^{N} y_{im} t_{ij} \Bigg/ \sum_{\substack{i=1 \\ i \neq l}}^{N} t_{ij}^2$$

$$u_{ij} = \sum_{\substack{k=1 \\ k \neq l}}^{M} y_{ik} q_{kj} \Bigg/ \sum_{\substack{k=1 \\ k \neq l}}^{M} q_{kj}^2$$

Once the vectors in the above iteration converge, the data is deflated for PCA by

$$\mathbf{X} = \mathbf{X} - \mathbf{t}_j \mathbf{p}_j^T \qquad \text{Equation 1-15}$$

and for PLS

$$p_{kj} = \sum_{\substack{i=1 \\ i \neq l}}^{N} x_{ik} t_{ij} \Bigg/ \sum_{\substack{i=1 \\ i \neq l}}^{N} t_{ij}^2$$

$$\mathbf{X} = \mathbf{X} - \mathbf{t}_j \mathbf{p}_j^T \qquad \text{Equation 1-16}$$

$$\mathbf{Y} = \mathbf{Y} - \mathbf{t}_j \mathbf{q}_j^T$$

The assumption in this algorithm at each component is that the missing measurement is represented by the appropriate loading vector and score product without taking into account any latent dimensions not yet calculated.

1.5.1.3 Iterative Replacement

This algorithm relies on the estimates from a PCA or PLS model to fill in the missing elements. It is outlined in Rannar et al. (1994b) where it is referred to as 'the EM algorithm'.

For PLS, the algorithm consists of :

1. replacing missing data by an initial value
2. optionally mean centring and scaling the matrices
3. calculating the PLS model from the resulting data matrices
4. obtaining new values for the missing variables
5. if convergence has not been reached, returning to step 2

The initial value suggested in Rannar et al. (1994b) is the mean of the row and column means. The new values of the missing measurements in step 4 are obtained from the entries in

$$\hat{\mathbf{X}} = \mathbf{T}\mathbf{P}^T \qquad \text{Equation 1-17}$$
$$\hat{\mathbf{Y}} = \mathbf{T}\mathbf{Q}^T$$

corresponding to the missing measurements and the convergence criterion used was

$$[abs(\mathbf{p}_{new}) - abs(\mathbf{p}_{old})]^2 / [\mathbf{p}_{old}]^2 < tol \qquad \text{Equation 1-18}$$

While not stated, it is reasonable to assume that the actual criterion was summed over all of the model dimensions and that the power 2 is interpreted as a sum of the squared elements of the vector. The cut-off value of the criterion that was used by Rannar et al. was 10^{-3}.

From the outline of this algorithm, an extension to PCA is clear. The PCA model is substituted for the PLS model and there is no output Y block to update.

The essential step in this algorithm is the calculation of the replacements for the missing measurements. Once the data matrix is fixed, the model obtained is not dependent on the method used to calculate it. Any of the kernel algorithms, NIPALS or other SVD or eigenvalue decomposition methods can be used.

1.5.1.4 The MLPCA Algorithm with Missing Measurements

The maximum likelihood principal components analysis (MLPCA) algorithm (Wentzell et al., 1997; Wentzell and Lohnes, 1999) produces a model that minimises the sum of squared fitting errors for the data, as does PCA, but incorporating information about the measurement errors. A PCA model is produced only when the errors are independent and have equal variance and this

was used in Andrews and Wentzell (1997) to produce a model that was similar to PCA from data with missing measurements. There are several versions of the algorithm given in Wentzell et al. (1997), the algorithm used in Andrews and Wentzell (1997) is one where the error varies by row and column, there are no error covariance terms and there are no intercept terms. The algorithm used in Andrews and Wentzell (1997) is the one that is covered here. The objective function is

$$S^2 = \sum_{i=1}^{N} \left[\mathbf{z}_i - \hat{\mathbf{z}}_i\right]^T \boldsymbol{\Sigma}_i^{-1} \left[\mathbf{z}_i - \hat{\mathbf{z}}_i\right] = \sum_{k=1}^{K} \left[\mathbf{x}_k - \hat{\mathbf{x}}_k\right]^T \boldsymbol{\Psi}_k^{-1} \left[\mathbf{x}_k - \hat{\mathbf{x}}_k\right] \qquad \text{Equation 1-19}$$

where $\boldsymbol{\Sigma}_i$ is the error covariance matrix for object $\mathbf{z}_i$ and $\boldsymbol{\Psi}_k$ is the error covariance matrix for variable $\mathbf{x}_k$. Both of these matrices are diagonal in this case because the errors are assumed uncorrelated. The estimates of the measurements $\hat{\mathbf{z}}_i$ and $\hat{\mathbf{x}}_k$ come from

$$\hat{\mathbf{z}}_i = \hat{\mathbf{V}}\left[\hat{\mathbf{V}}^T \boldsymbol{\Sigma}_i^{-1} \hat{\mathbf{V}}\right]^{-1} \hat{\mathbf{V}}^T \boldsymbol{\Sigma}_i^{-1} \mathbf{z}_i \qquad \text{Equation 1-20}$$

$$\hat{\mathbf{x}}_k = \hat{\mathbf{U}}\left[\hat{\mathbf{U}}^T \boldsymbol{\Psi}_k^{-1} \hat{\mathbf{U}}\right]^{-1} \hat{\mathbf{U}}^T \boldsymbol{\Psi}_k^{-1} \mathbf{x}_k \qquad \text{Equation 1-21}$$

and the matrices $\hat{\mathbf{U}}$ and $\hat{\mathbf{V}}$ come from a singular value decomposition of $\hat{\mathbf{X}}$

$$\hat{\mathbf{X}} = \left[\hat{\mathbf{x}}_1, \quad \hat{\mathbf{x}}_2, \quad \ldots \quad \hat{\mathbf{x}}_K\right] = \begin{bmatrix} \hat{\mathbf{z}}_1^T \\ \hat{\mathbf{z}}_2^T \\ \vdots \\ \hat{\mathbf{z}}_N^T \end{bmatrix} = \hat{\mathbf{U}}\hat{\mathbf{S}}\hat{\mathbf{V}}^T \qquad \text{Equation 1-22}$$

The algorithm starts with an initial value of $\hat{\mathbf{U}}$ and $\hat{\mathbf{V}}$ from an SVD of $\mathbf{X}$. One iteration of the algorithm consists of two parts, the first of which is projection on the rows $\mathbf{z}_i$ using Equation 1-20, calculation of the objective function Equation

1-19 using the rows and an update of $\hat{\mathbf{U}}$ and $\hat{\mathbf{V}}$ by an SVD of the projected row data $\hat{\mathbf{z}}_i$. The second part of the iteration is projection on the columns $\mathbf{x}_k$ using Equation 1-21, calculation of the objective function Equation 1-19 using the columns and an update of $\hat{\mathbf{U}}$ and $\hat{\mathbf{V}}$ by an SVD of the projected column data $\hat{\mathbf{x}}_k$. Convergence is obtained when the difference between the values of the objective function calculated in the two parts is less than a tolerance. A graphical representation of the algorithm is contained in Wentzell et al. (1997).

The MLPCA model with missing measurements in Andrews and Wentzell (1997) was calculated by setting the error variances to one for present measurements and 10^{10} for missing measurements. This weighting will yield a PCA model if there are no missing measurements. Andrews and Wentzell did not specify what values were used for the algorithm for the missing measurements but it will be shown in section 4.3.3 that these values do not have a strong effect on the results.

1.5.2 Score Calculation with Missing Measurements

The objective is to obtain estimates of the A elements of the vector of scores $\boldsymbol{\tau}$ using a new, incomplete multivariate observation $\mathbf{z}^*$.

1.5.2.1 Single Component Projection Algorithm

Once a PCA model has been built, and the loading vectors ($\mathbf{p}_i$'s) are fixed, a non-iterative approach analogous to NIPALS can be used to handle missing data

in new multivariate observations. The score calculation step of the NIPALS missing data model building algorithm is simply applied to each dimension sequentially. This is called the single component projection method (SCP) throughout the remainder of the thesis.

Consider the case where a PCA model has been built and a new multivariate observation vector $\mathbf{z}^T(1)=\left[\mathbf{z}^{\#T}(1),\ \ \mathbf{z}^{*T}(1)\right]$ becomes available. Here $\mathbf{z}^*(1)$ represents the variables for which observations are present. To calculate the j^{th} element of the vector score, τ_j, corresponding to this observation, the single component projection algorithm minimises

$$J=\frac{1}{2}\left[\mathbf{z}^*(j)-\tau_j\mathbf{p}_j^*\right]^T\left[\mathbf{z}^*(j)-\tau_j\mathbf{p}_j^*\right]$$

which yields

$$\hat{\tau}_j=\frac{\mathbf{p}_j^{*T}\mathbf{z}^*(j)}{\mathbf{p}_j^{*T}\mathbf{p}_j^*} \qquad \text{Equation 1-23}$$

as the least squares estimate of τ_j based on the observed variables. The portion of $\mathbf{z}^*(j)$ explained by this component is then subtracted to yield the deflated object

$$\mathbf{z}^*(j+1)=\mathbf{z}^*(j)-\hat{\tau}_j\mathbf{p}_j^* \qquad \text{Equation 1-24}$$

and the next component ($\hat{\tau}_{j+1}$) is then calculated.

Substituting $\mathbf{w}_j^*$ for $\mathbf{p}_j^*$ in Equation 1-23 gives the PLS score calculation formula. Deflation in the PLS score calculation is the same as Equation 1-24; the

loading vector $\mathbf{p}_j^*$ for deflation in PLS is not the same as the one used to calculate the scores but is calculated in Equation 1-16.

1.5.2.2 Handling Missing Data in PCA by Projection to the Model Plane

The objective function to be minimised for projection to a plane is

$$J = \frac{1}{2}\left[\mathbf{z}^* - \mathbf{P}_{1:A}^*\boldsymbol{\tau}_{1:A}\right]^T\left[\mathbf{z}^* - \mathbf{P}_{1:A}^*\boldsymbol{\tau}_{1:A}\right]$$

where $\mathbf{P}_{1:A}$ is the matrix of the first A loading vectors corresponding to $\boldsymbol{\tau}_{1:A}$ which are the scores being calculated. The optimal value for the score vector ($\hat{\boldsymbol{\tau}}_{1:A}$) is obtained by taking the derivative with respect to $\boldsymbol{\tau}_{1:A}$ and setting it equal to zero.

$$\frac{dJ}{d\boldsymbol{\tau}_{1:A}} = -\mathbf{P}_{1:A}^{*T}\left[\mathbf{z}^* - \mathbf{P}_{1:A}^*\hat{\boldsymbol{\tau}}_{1:A}\right] = \mathbf{0}$$

Equation 1-25

$$\hat{\boldsymbol{\tau}}_{1:A} = \left[\mathbf{P}_{1:A}^{*T}\mathbf{P}_{1:A}^*\right]^{-1}\mathbf{P}_{1:A}^{*T}\mathbf{z}^*$$

This projection method for obtaining all score estimates simultaneously by regressing $\mathbf{z}^*$ onto the plane defined by $\mathbf{P}^*$ has been advocated by Wold et al. (1983) and by Martens and Naes (1989). The regression based algorithm given by Wise and Ricker (1991) can also be shown to be equivalent to Equation 1-25. However, the latter algorithm is more difficult to implement and does not lend itself readily to the analysis performed in this thesis.

1.5.3 Prediction and Monitoring with Missing Measurements

1.5.3.1 Prediction

Section 1.5.2 gave methods for calculating scores with missing measurements. Once the scores are calculated the predictions are calculated using Equation 1-2.

1.5.3.2 Monitoring

Once values for the scores have been calculated using one of the methods in section 1.5.2, the Hotelling T^2 can be calculated by Equation 1-11. In Equation 1-10, values for all measurements are needed in addition to the scores before the SPE can be calculated. The approach that is used in this thesis is to take the score that is calculated by the appropriate score calculation method in section 1.5.2 and the loading vector provided by the PCA or PLS model and use that to calculate a value

$$\hat{\mathbf{z}}^{\#} = \mathbf{P}^{\#}\boldsymbol{\tau}$$

This value is then used for the missing measurements. Thus the residuals for the missing measurements will be zero if every latent variable in the model is used in monitoring. This is the approach taken in the MACSTAT software package (1995). The calculated SPE will be lower than the true SPE unless the missing measurements have zero residuals in the model.

1.6 EM Algorithm

The Expectation-Maximization (EM) algorithm (Little and Rubin, 1987) can be used to calculate a maximum likelihood estimate of a mean vector (μ) and covariance matrix (S) from incomplete data. Some of the properties of EM are that it increases the value of the likelihood function at each iteration and it has the maximum likelihood values of the statistics as a stationary point (Little and Rubin, 1987). At each iteration, the algorithm can be interpreted as replacing the missing variable values by the expected values from the conditional normal distribution given the present data and the current estimate of the means and covariance, that is

$$\hat{\mathbf{z}}^{\#} = E\left(\mathbf{z}^{\#} | \mathbf{z}^{*}, \mu, \mathrm{S}\right)$$

In model building, the estimate of $\mathbf{z}^{\#}$ is then combined with $\mathbf{z}^{*}$ and the current estimate of the mean vector and covariance matrix and used to calculate an updated mean vector and covariance matrix. This is repeated until the estimates converge.

1.6.1 EM and PCA and PLS Models

Since both PCA and PLS are based on the decomposition of covariance matrices (Hoskuldsson, 1988), EM can be combined with any covariance matrix based PCA or PLS algorithm to produce a maximum likelihood estimate of the model from incomplete data. With an input-output model, EM makes no distinction between input and output variables. In the equations below all

variables are considered to be placed in a single matrix **X** whose mean vector and variance matrix are to be calculated. After these quantities are calculated, the estimates can be partitioned to obtain the desired input and output grouping.

The expectation step in the case of estimating a mean vector $\boldsymbol{\mu}$ and a covariance matrix S (Little and Rubin, 1987) with iteration index n is

$$E\left(\sum_{i=1}^{N} x_{ik} \mid \mathbf{X}^*, \boldsymbol{\mu}(n), \mathrm{S}(n)\right) = \sum_{i=1}^{N} x_{ik}(n) \qquad \text{Equation 1-26}$$

$$E\left(\sum_{i=1}^{N} x_{ik_1} x_{ik_2} \mid \mathbf{X}^*, \boldsymbol{\mu}(n), \mathrm{S}(n)\right) = \sum_{i=1}^{N} \left[x_{ik_1}(n) x_{ik_2}(n) + c_{k_1 k_2 i}(n)\right] \qquad \text{Equation 1-27}$$

where

$$x_{ik}(n) = \begin{cases} x_{ik} & x_{ik} \text{ measured} \\ E\left(x_{ik} \mid \mathbf{z}_i^*, \boldsymbol{\mu}(n), \mathrm{S}(n)\right) & x_{ik} \text{ missing} \end{cases} \qquad \text{Equation 1-28}$$

$$c_{k_1 k_2 i}(n) = \begin{cases} 0 & x_{ik_1} \text{ or } x_{ik_2} \text{ measured} \\ \mathrm{cov}\left(x_{ik_2}, x_{ik_2} \mid \mathbf{z}_i^*, \boldsymbol{\mu}(n), \mathrm{S}(n)\right) & x_{ik_1} \text{ and } x_{ik_2} \text{ missing} \end{cases} \qquad \text{Equation 1-29}$$

The means and variances above are the estimates at the current iteration. Note that the expected value of the cross-products of the variables is not just the product of the estimates when both the elements are missing. A correction is calculated based on the current estimate of the covariance matrix to account for the sum of squares that can not be estimated from the measured variables.

The maximization (M) step updates the mean and covariance estimates for the next iteration

$$\mu_k(n+1) = \sum_{i=1}^{N} x_{ik}(n) \Big/ N \qquad \text{Equation 1-30}$$

$$s_{k_1k_2}(n+1) = \frac{E\left(\sum_{i=1}^{N} \left[x_{ik_1}(n) - \mu_{k_1}(n)\right]\left[x_{ik_2}(n) - \mu_{k_2}(n)\right] + c_{k_1k_2i}(n) \right)}{N}$$

Assuming a non-singular normal distribution for the data, the expectation in Equation 1-28 in matrix vector form is

$$E\left(\mathbf{z}_i^{\#} \mid \mathbf{z}_i^{*}, \boldsymbol{\mu}(n), \mathbf{S}(n)\right) = \boldsymbol{\mu}^{\#}(n) + \left[\mathbf{S}^{\#\#}(n)\right]^{-1}\mathbf{S}^{\#*}(n)\mathbf{z}_i^{*} \qquad \text{Equation 1-31}$$

The covariance term in Equation 1-29 in matrix vector form is

$$\operatorname{cov}\left(\mathbf{z}_i^{\#}, \mathbf{z}_i^{\#} \mid \mathbf{z}_i^{*}, \boldsymbol{\mu}(n), \mathbf{S}(n)\right) = \mathbf{S}^{\#\#}(n) - \mathbf{S}^{*\#}(n)\left[\mathbf{S}^{\#\#}(n)\right]^{-1}\mathbf{S}^{\#*}(n) \qquad \text{Equation 1-32}$$

These basic results can be found in Johnson and Wichern (1988).

1.7 Quadratic Forms

In this thesis it will be necessary to calculate the cumulative probability distribution of a quadratic form

$$Q = \mathbf{x}^T\mathbf{A}\mathbf{x} \qquad \text{Equation 1-33}$$

$$\mathbf{x} \sim \mathbf{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

Equation 1-10 and Equation 1-11 have this structure so the Hotelling T^2 and SPE are examples of quadratic forms. In both cases the mean vector of the random variables $\boldsymbol{\mu}$ is zero which is referred to as the central case. The standard results for the critical values of the forms referenced in sections 1.4.2.1 and 1.4.2.2 are for

the central case and can not be used where the mean of the variables (μ) is not zero.

Approximations to the cumulative distribution of the non-central quadratic form can be found in the literature (Jensen and Solomon, 1972; Kotz et al., 1967) when both **A** and Σ are positive definite. The parameters for the approximations are calculated after Equation 1-33 is put into the standard form

$$Q = \sum_{k=1}^{K} c_k \{x_k + a_k\}^2$$

$$\mathbf{x} \sim N(\mathbf{0}, \mathbf{I})$$

According to Imhof (1961), a similar form can be obtained when Σ is singular but **A** is not. A reduction to the form above when both **A** and Σ are singular is given in appendix 1.

The approximation that will be used in this work is that of Jensen and Solomon (1972). They apply a transformation to the quadratic value that yields an approximately standardised Gaussian variable

$$z = \frac{\theta_1 \left[\left[\frac{Q}{\theta_1} \right]^{h_o} - 1 - \frac{\theta_2 h_o [h_o - 1]}{\theta_1^2} \right]}{\left[2\theta_2 h_o^2 \right]^{\frac{1}{2}}}$$

$$\theta_n = \sum_{k=1}^{K} c_k^n \left[1 + n a_k^2 \right]$$

$$h_o = 1 - \frac{2\theta_1 \theta_3}{3\theta_2^2}$$

with approximate mean $1+\frac{\theta_2 h_0[h_0-1]}{\theta_1^2}$ and variance $\frac{2\theta_2 h_0^2}{\theta_1^2}$.

Chapter 2: Analysis of Score Errors from Missing Measurements

2.1 Introduction

Once a PCA or PLS model has been built it will often be applied to new objects which are not a part of the model building data set. Applications where this is done include prediction of process responses for inferential control and process monitoring. These new objects may have missing measurements due to faulty instruments, maintenance or missed manual samples. The model can only be applied to these incomplete objects if a value for the score vector τ can be calculated from the incomplete object which is close to the value that would be produced if all measurements were present. There are several algorithms in use for calculating scores when the model is fixed and known and some measurements in a new object $\mathbf{z}$ are missing but the factors affecting their performance have not been developed.

A novel missing measurement score calculation algorithm is proposed in this chapter and the factors affecting the performance of both the novel and existing algorithms are identified and illustrated by designed and industrial process data sets. The novel score calculation algorithm is developed from the EM algorithm. Recommendations for pruning variables and the number of dimensions in a model that lead to models that are more robust to missing

measurements are made. The designed data sets show the effect of individual factors on score calculation with missing measurements. The analysis of score calculation with missing measurements for two process data sets from the literature, one simulated and the other industrial, shows how the specific factors can be evaluated in practice to determine performance for a given algorithm. Much of this work was published in Nelson et al. (1996).

Expressions are developed for the score estimation error arising from the missing data with the SCP algorithm. This analysis reveals how errors enter and propagate in the SCP method, thereby providing justification for using simultaneous projection methods and supplying insight into the sources of error that arise during model building with sequential methods. Two approaches which are not limited to considering a single direction at a time are then treated in a similar manner: (i) projection to the model plane and (ii) data replacement using the conditional mean.

The mean squared score estimation errors are calculated for each of the methods when applied to simulation examples carefully constructed to accentuate the effects of the types of errors identified in the analysis. Finally, the methods are applied to a data set from a simulated distillation column and an industrial data set to illustrate how the methods work in practice. The examples reveal the application of the score error expressions and potential pitfalls in variable pruning during model building.

2.2 Handling Missing Data by Single Component Projection

An important point that distinguishes the single component projection approach from other methods of handling missing data is that it treats the missing data separately in the calculation of each latent dimension. It does not consider the impact on future dimensions, nor does it consider the effect that errors made in earlier dimensions will have on the calculation of the current dimension. It is the purpose of this section to carry out an error analysis of the single component projection approach which is outlined in section 1.5.2.1, and to illustrate the consequences of these errors using simple simulation examples specifically constructed to highlight problem situations.

2.2.1 Error Analysis for PCA

The estimate of the error in the score calculated using Equation 1-23 can be analysed by assuming a structure for the independent data vector **z**. Since a model has already been built which decomposes the covariance matrix, it can be used to give this structure. The measured portion of the new data vector **z** can be expressed as

$$\mathbf{z}^*(1) = \mathbf{P}^*\boldsymbol{\tau} + \mathbf{e}^* \qquad \text{Equation 2-1}$$

where $\boldsymbol{\tau}$ is the true value for the scores and the 1 in brackets denotes the original data. The number in brackets after **z** is incremented after each dimension as Equation 1-24 is applied. The residual term $\mathbf{e}^*$ contains measurement errors.

The structure in Equation 2-1 is that of a PCA model but the actual dimensionality of the data vector **z** may not be the same as the model that has

been calculated. Therefore the matrix **P** in Equation 2-1 has K columns and τ has K rows to give the most general representation possible. The first A columns of **P** are equal to the model loading vectors. This allows the true dimensionality of the data to differ from that of any model of it but requires that some of the elements τ_j be fixed at zero if the actual dimensionality of the data is less than K.

The error in the first score is $\tau_1 - \hat{\tau}_1$ where $\hat{\tau}_1$ is given by Equation 1-23. The expression for $\mathbf{z}^*(1)$ given in Equation 2-1 can be substituted into the error expression for the first score to obtain

$$\begin{aligned}
\tau_1 - \hat{\tau}_1 &= \tau_1 - \left[\mathbf{p}_1^{*T}\mathbf{p}_1^*\right]^{-1}\mathbf{p}_1^{*T}\mathbf{z}^*(1) \\
&= \tau_1 - \left[\mathbf{p}_1^{*T}\mathbf{p}_1^*\right]^{-1}\mathbf{p}_1^{*\mathrm{T}}\left[\mathbf{P}^*\tau + \mathbf{e}^*\right] \\
&= \tau_1 - \left[\mathbf{p}_1^{*T}\mathbf{p}_1^*\right]^{-1}\mathbf{p}_1^{*T}\left[\mathbf{p}_1^*,\ \mathbf{p}_2^*,\ \ldots,\ \mathbf{p}_K^*\right]\tau - \left[\mathbf{p}_1^{*T}\mathbf{p}_1^*\right]^{-1}\mathbf{p}_1^{*T}\mathbf{e}^* \\
&= -\sum_{j=2}^{K}\left[\mathbf{p}_1^{*T}\mathbf{p}_1^*\right]^{-1}\mathbf{p}_1^{*T}\mathbf{p}_j^*\tau_j - \left[\mathbf{p}_1^{*T}\mathbf{p}_1^*\right]^{-1}\mathbf{p}_1^{*T}\mathbf{e}^*
\end{aligned}$$

after deflation,

$$\mathbf{z}^*(j) = \mathbf{P}^*\tau + \mathbf{e}^* - \hat{\tau}_1\mathbf{p}_1^* - \hat{\tau}_2\mathbf{p}_2^* - \ldots - \hat{\tau}_{j-1}\mathbf{p}_{j-1}^*$$

and therefore

$$\begin{aligned}
\tau_j - \hat{\tau}_j &= \tau_j - \left[\mathbf{p}_j^{*T}\mathbf{p}_j^*\right]^{-1}\mathbf{p}_j^{*T}\mathbf{z}^*(j) \\
&= \tau_j - \left[\mathbf{p}_j^{*T}\mathbf{p}_j^*\right]^{-1}\mathbf{p}_j^{*\mathrm{T}}\left[\mathbf{P}^*\tau + \mathbf{e}^* - \sum_{m=1}^{j-1}\hat{\tau}_m\mathbf{p}_m^*\right] \\
&= \tau_j - \left[\mathbf{p}_j^{*T}\mathbf{p}_j^*\right]^{-1}\mathbf{p}_j^{*\mathrm{T}}\left[\mathbf{p}_1^*,\ \mathbf{p}_2^*,\ \ldots,\ \mathbf{p}_K^*\right]\tau \\
&\quad - \left[\mathbf{p}_j^{*T}\mathbf{p}_j^*\right]^{-1}\mathbf{p}_j^{*\mathrm{T}}\mathbf{e}^* + \left[\mathbf{p}_j^{*T}\mathbf{p}_j^*\right]^{-1}\mathbf{p}_j^{*\mathrm{T}}\sum_{m=1}^{j-1}\hat{\tau}_m\mathbf{p}_m^* \\
&= -\sum_{m=j+1}^{K}\left[\mathbf{p}_j^{*T}\mathbf{p}_j^*\right]^{-1}\mathbf{p}_j^{*\mathrm{T}}\mathbf{p}_m^*\tau_m - \left[\mathbf{p}_j^{*T}\mathbf{p}_j^*\right]^{-1}\mathbf{p}_j^{*\mathrm{T}}\mathbf{e}^* \\
&\quad - \left[\mathbf{p}_j^{*T}\mathbf{p}_j^*\right]^{-1}\sum_{m=1}^{j-1}\mathbf{p}_j^{*\mathrm{T}}\mathbf{p}_m^*\left[\tau_m - \hat{\tau}_m\right]
\end{aligned}$$

Equation 2-2

for j = 2, 3, ..., A. When there are no missing measurements, $\mathbf{p}_j^{*T}\mathbf{p}_m^* = \mathbf{p}_j^T\mathbf{p}_m = 0, j \neq m$ and $\mathbf{p}_j^{*T}\mathbf{p}_j^* = \mathbf{p}_j^T\mathbf{p}_j = 1$, and so Equation 2-2 reduces to

$$\tau_j - \hat{\tau}_j = -\left[\mathbf{p}_j^{\mathrm{T}}\mathbf{p}_j\right]^{-1}\mathbf{p}_j^{\mathrm{T}}\mathbf{e} \qquad \text{Equation 2-3}$$

This shows that the score estimation error with no missing measurements is only affected by the measurement errors. However, in the presence of missing data Equation 2-2 shows that two additional terms appear in the error expression for the j^{th} score estimate. The first term in Equation 2-2 represents the error in $\hat{\tau}_j$ arising from incorrectly attributing some of the variance in $\mathbf{z}^*$ that arises from later principal component scores τ_m (m > j) to the current score τ_j. This term will tend to be large whenever the subsequent scores explain a significant amount of the variation in $\mathbf{z}^*$ and when the missing elements make $\mathbf{p}_j^*$ and $\mathbf{p}_m^*$ (m > j) no longer orthogonal ($\mathbf{p}_j^{*T}\mathbf{p}_m^* \neq 0$). The last term in Equation 2-2 represents the propagation of the errors made in estimating the earlier scores (m < j) into the error of the present score estimate $\hat{\tau}_j$. Again this term will be important when the missing elements make $\mathbf{p}_j^*$ and $\mathbf{p}_m^*$ (m < j) no longer orthogonal. Note that whenever the missing measurements have large loadings in the current loading vector $\mathbf{p}_j$, the squared length of the $\mathbf{p}_j^*$ vector ($\mathbf{p}_j^{*T}\mathbf{p}_j^*$) will be small, and its inverse will be large, thereby increasing the size of all the error terms in Equation 2-2.

A general expression for the error in the SCP algorithm can be also be defined recursively in matrix-vector form to give the error in terms of the original data vector $\mathbf{z}^*$.

$$\begin{aligned}\mathbf{z}^*(j) &= \mathbf{z}^*(j-1) - \hat{\tau}_{j-1}\mathbf{p}^*_{j-1} \\ &= \mathbf{z}^*(j-1) - \left[\mathbf{p}^{*\mathrm{T}}_{j-1}\mathbf{p}^*_{j-1}\right]^{-1}\left[\mathbf{p}^{*\mathrm{T}}_{j-1}\mathbf{z}^*(i-1)\right]\mathbf{p}^*_{j-1} \\ &= \mathbf{z}^*(j-1) - \left[\mathbf{p}^{*\mathrm{T}}_{j-1}\mathbf{p}^*_{j-1}\right]^{-1}\mathbf{p}^*_{j-1}\left[\mathbf{p}^{*T}_{j-1}\mathbf{z}^*(j-1)\right] \\ &= \left(\mathbf{I} - \left[\mathbf{p}^{*\mathrm{T}}_{j-1}\mathbf{p}^*_{j-1}\right]^{-1}\mathbf{p}^*_{j-1}\mathbf{p}^{*T}_{j-1}\right)\mathbf{z}^*(j-1) \\ &= \Delta(j)\mathbf{z}^*(1)\end{aligned} \qquad \text{Equation 2-4}$$

where

$$\Delta(j) = \prod_{m=1}^{j-1}\left(\mathbf{I} - \left(\mathbf{p}^{*T}_{j-m}\mathbf{p}^*_{j-m}\right)^{-1}\mathbf{p}^*_{j-m}\mathbf{p}^{*T}_{j-m}\right) \qquad \text{Equation 2-5}$$

Substituting this definition into the score error equation for PCA, we obtain

$$\begin{aligned}\tau_j - \hat{\tau}_j &= \tau_j - \left[\mathbf{p}^{*T}_j\mathbf{p}^*_j\right]^{-1}\mathbf{p}^{*T}_j\mathbf{z}^*(j) \\ &= \tau_j - \left[\mathbf{p}^{*T}_j\mathbf{p}^*_j\right]^{-1}\mathbf{p}^{*\mathrm{T}}_j\Delta(j)\left[\mathbf{P}^*\boldsymbol{\tau} + \mathbf{e}^*\right]\end{aligned} \qquad \text{Equation 2-6}$$

2.2.2 Simulation Studies

Example data sets, each with 300 objects and 3 variables, were specifically created to illustrate the effect on the score estimation error of each of the terms in Equation 2-2. Only three variables were used to enable geometric interpretations of the results, but this means that 33% of the data will be missing when one measurement is deleted. In addition, the data sets are designed to deliberately accentuate the errors that arise in problem situations. Therefore, the results obtained from these simulated data sets will show extreme situations. Results that are more typical are shown in the industrial example treated in

section 2.5.2.

The simulated data were generated as follows. Three sets of sample scores were drawn from normal distributions with zero means and different variances. Several data sets were generated from each set of scores using specified loading vectors that are described below. The first variable was assumed to be missing in each of the 300 simulated multivariate observations in each data set and the first and second scores were estimated using the single component projection missing data algorithm. The differences between these estimated scores and the known true scores were calculated and the mean square of these errors displayed in Table 2-1 in the column labelled 'SCP Algorithm'. Mean squared errors for the other methods introduced and analysed later in this chapter are also listed in this table, and will be referenced later. The significance of the labels on the rows of Table 2-1 is discussed below.

Influential variables have large weights in the loading vector. When they are removed from the data vector the magnitude of $\mathbf{p}_j^{*T}\mathbf{p}_j^*$ decreases from a value of 1 with no missing measurements to a minimum of 0. Geometrically, if $\mathbf{p}_j^*$ is collinear with the plane of the missing measurements then there is no information about that component in the measured variables and if $\mathbf{p}_j^*$ is orthogonal to the plane of the missing measurements then there is full information in the observed variables. As $\mathbf{p}_j^*$ becomes more collinear with the plane of missing measurements, the term $\mathbf{p}_j^{*T}\mathbf{p}_j^*$ decreases and its inverse increases, thereby

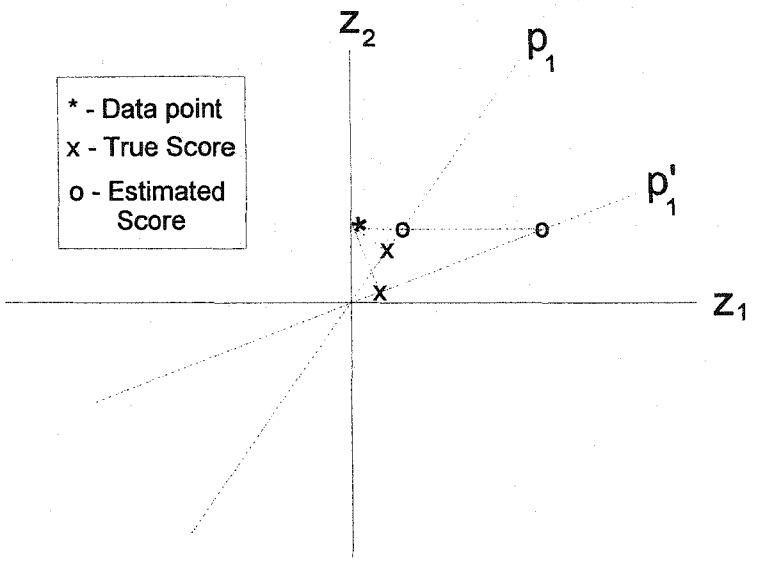


Figure 2-1: Score Estimation Error Increases as the Loading Vector Nears the Missing Variable Axis

increasing the size of each term in the score estimation error Equation 2-2. This effect is illustrated graphically in Figure 2-1 where the error in the score, both relative and absolute, increases when the loading vector becomes more collinear with the missing variable axis (z_1). The data sets in Table 2-1 with $\mathbf{p}_1^{*T}\mathbf{p}_1^* = 0.0196$ have a first loading vector that depends heavily on the missing variable (z_1). The first loading vector of data sets with $\mathbf{p}_1^{*T}\mathbf{p}_1^* = 0.6667$ is not as collinear with the missing variable axis but rather depends equally on all variables. Comparing the score estimate mean square errors in Table 2-1 shows that, with the rest of the characteristics the same, data sets with $\mathbf{p}_1^{*T}\mathbf{p}_1^* = 0.0196$ have much larger errors than data sets with $\mathbf{p}_1^{*T}\mathbf{p}_1^* = 0.6667$ for the first component scores.

Loading Vector Characteristics	Number of High Variance Dimensions	SCP Algorithm	Projection to Model Plane with OLS	Projection to Model Plane with PCR	Conditional Mean Replacement
$\mathbf{p}_1^{*T}\mathbf{p}_1^* = 0.6667$ $\angle \mathbf{p}_1^*, \mathbf{p}_2^* = 50.8°$	1	0.0040	0.0072	0.0072	0.0035
	2	0.2548	0.0142	0.0142	0.0137
	3	0.2868	0.6072	0.3079	0.2055
$\mathbf{p}_1^{*T}\mathbf{p}_1^* = 0.6667$ $\angle \mathbf{p}_1^*, \mathbf{p}_2^* = 7°$	1	0.0049	0.4815	0.0918	0.0048
	2	0.3785	0.9483	0.2924	0.2226
	3	0.3194	40.68	0.2337	0.2291
$\mathbf{p}_1^{*T}\mathbf{p}_1^* = 0.0196,$ $\angle \mathbf{p}_1^*, \mathbf{p}_2^* = 50.8°$	1	0.3343	0.4025	0.4025	0.2230
	2	15.64	0.7926	0.7926	0.4516
	3	25.66	34.00	0.8404	0.8188
$\mathbf{p}_1^{*T}\mathbf{p}_1^* = 0.0196,$ $\angle \mathbf{p}_1^*, \mathbf{p}_2^* = 7°$	1	0.4912	16.53	0.7857	0.2879
	2	37.49	32.55	0.9568	0.9395
	3	31.90	1397	0.8275	0.8246

Table 2-1: Mean Square Error of the First PCA Score Estimate for Example Data Sets

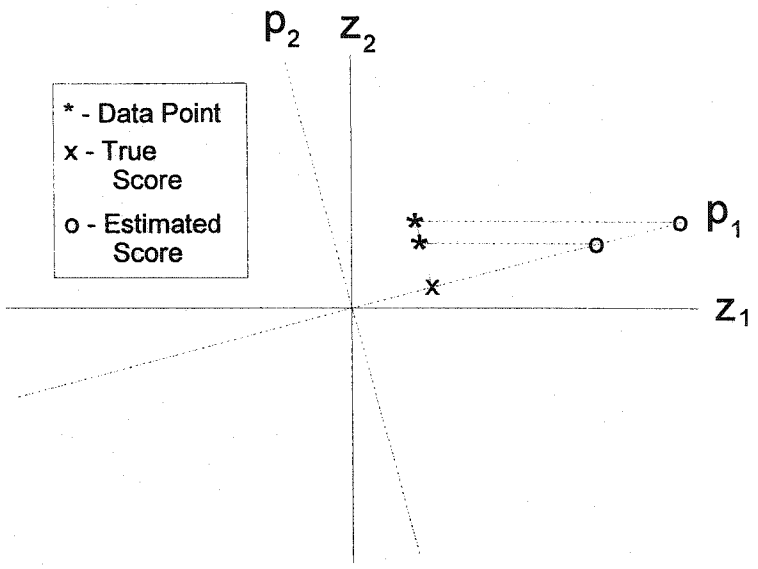


Figure 2-2: Score Estimation Error Increases as the Second Score Becomes Larger Relative to the First

If loading vectors become collinear because of missing measurements then the first of the loading vectors which is used to calculate a score will be used to explain some of the variance in $\mathbf{z}^*$ in the direction of the other. The expression $\mathbf{p}_j^{*T}\mathbf{p}_m^*\tau_m$ in the first term of Equation 2-2 is due to this effect and will be zero when there is no missing data. With missing data, it will increase in importance when τ_m is large compared to τ_j (data sets with two or more large scores) and when $\mathbf{p}_j^*$ is nearly collinear with $\mathbf{p}_m^*$ (data sets with $\angle\mathbf{p}_1^*,\mathbf{p}_2^* = 7^\circ$). The expression indicates how much of the variance in $\mathbf{z}^*$ associated with τ_m will be mistakenly attributed to τ_j when $\mathbf{p}_m^*$ is not orthogonal to $\mathbf{p}_j^*$. This is illustrated geometrically in Figure 2-2 where two objects with the same first score and different second

scores are regressed onto a loading vector. The larger the second score in comparison to the first, the larger the error will become.

One way that the data sets used to calculate the results in Table 2-1 can be classified is by the relative magnitude of the variance of their scores. There are three different classes of relative score variance. In the data sets with one large score dimension, the variance of the first score is almost two orders of magnitude greater than the others with values of 0.9, 0.01, and 0.005. The data sets with 2 large scores have the first two score variances close together and the third much smaller (0.9, 0.7, and 0.01). All 3 score variances are of similar magnitude (0.9, 0.7, and 0.5) in the data sets with all scores large. Comparing consecutive entries in Table 2-1 with constant $\mathbf{p}_1^{*T}\mathbf{p}_1^*$ and $\angle\mathbf{p}_1^*,\mathbf{p}_2^*$, the error tends to increase with the increase in size of the variances of the later scores compared to that of the first score as expected from the error expression in Equation 2-2.

The first and second missing variable loading vectors are almost collinear for half of the data sets, those where the angle between $\mathbf{p}_1^*$ and $\mathbf{p}_2^*$ is 7.0^O, and relatively distinct in the other half where the angle is 50.8^O. With no missing data, the angle is 90^O and the vectors are orthogonal. Compare data sets with a different angle between $\mathbf{p}_1^*$ and $\mathbf{p}_2^*$ but identical scores and $\mathbf{p}_1^{*T}\mathbf{p}_1^*$ to see that, with everything else constant, the score calculation error increases as $\mathbf{p}_2^*$ becomes more collinear with $\mathbf{p}_1^*$.

Loading Vector Characteristics	Number of High Variance Dimensions	SCP Algorithm	Projection to Model Plane with OLS	Projection to Model Plane with PCR	Conditional Mean Replacement
$\mathbf{p}_1^{*T}\mathbf{p}_1^* = 0.6667$ $\angle\mathbf{p}_1^*,\mathbf{p}_2^* = 50.8°$	1	0.0049	0.0096	0.0096	0.0047
	2	0.1280	0.0189	0.0189	0.0183
	3	0.4032	0.8096	0.4105	0.2740
$\mathbf{p}_1^{*T}\mathbf{p}_1^* = 0.6667$ $\angle\mathbf{p}_1^*,\mathbf{p}_2^* = 7°$	1	0.0098	0.9582	0.1827	0.0096
	2	0.7386	1.887	0.5820	0.4430
	3	0.6427	80.96	0.4651	0.4559
$\mathbf{p}_1^{*T}\mathbf{p}_1^* = 0.0196,$ $\angle\mathbf{p}_1^*,\mathbf{p}_2^* = 50.8°$	1	0.0027	0.0033	0.0033	0.0018
	2	0.1237	0.0064	0.0064	0.0037
	3	0.2668	0.2753	0.0068	0.0066
$\mathbf{p}_1^{*T}\mathbf{p}_1^* = 0.0196,$ $\angle\mathbf{p}_1^*,\mathbf{p}_2^* = 7°$	1	0.0097	0.3238	0.0155	0.0057
	2	0.7385	0.6416	0.0189	0.0185
	3	0.6281	27.52	0.0163	0.0163

Table 2-2: Mean Square Error of the Second PCA Score Estimate for Example Data Sets

The last term in Equation 2-2 shows how errors made in previous score calculations, $\tau_m - \hat{\tau}_m$ with m < j, affect the error in the score estimate ($\hat{\tau}_j$) currently being calculated. This term will be large in cases where there are large errors in the early scores, and when the missing data loading vector $\mathbf{p}_j^*$ is nearly collinear with the previous loading vector $\mathbf{p}_m^*$ (data sets with $\angle \mathbf{p}_1^*, \mathbf{p}_2^* = 7^\circ$). This effect is shown for the single component projection algorithm in Table 2-2 where the mean squared error in the second score vector estimate is seen to be largest for these cases.

These pathological examples are used only to illustrate the effects of various terms on the error in score estimates obtained from single component projection when there are missing measurements. Since there are only 3 variables, the effects will be exaggerated compared to real situations where a smaller percentage of the variables may be missing and such high collinearities may not be present.

2.2.3 Error Analysis for PLS

The structure of the deflated data vector at the j^{th} stage of the single component projection algorithm for PLS is

$$\mathbf{z}^*(j) = \mathbf{P}^*\boldsymbol{\tau} + \mathbf{e}^* + \mathbf{d}^* - \sum_{m=1}^{j-1} \hat{\tau}_m \mathbf{p}_m^* \qquad \text{Equation 2-7}$$

since PLS deflates using the **p** vectors. Once again **P** has K columns and $\boldsymbol{\tau}$ has K rows to allow the dimensionality of the data to be different than the model and

some of the elements of τ will be fixed at zero if the dimensionality of the data is less than K. If there is no score estimation error and the underlying dimensionality of the data is A then $\hat{\tau}_m = \tau_m$ and the first and last terms cancel after all A scores have been used in deflation. In this case, the residual data vector is composed of the sum of a vector of random error variables $\mathbf{e}^*$ and a deterministic remainder $\mathbf{d}^*$. The $\mathbf{d}^*$ term arises because PLS does not necessarily use all the non-random information in the independent data block that is of greater magnitude than the noise, as does PCA. Substituting this expression for $\mathbf{z}^*(j)$ into the expression for $\hat{\tau}_j$ for PLS gives

$$\begin{aligned}
\tau_j - \hat{\tau}_j &= \tau_j - \left[\mathbf{w}_j^{*T}\mathbf{w}_j^*\right]^{-1}\mathbf{w}_j^{*T}\mathbf{z}^*(j) \\
&= \tau_j - \left[\mathbf{w}_j^{*T}\mathbf{w}_j^*\right]^{-1}\mathbf{w}_j^{*T}\left[\mathbf{P}^*\tau + \mathbf{e}^* + \mathbf{d}^* - \sum_{m=1}^{j-1}\hat{\tau}_m\mathbf{p}_m^*\right] \\
&= \tau_j - \left[\mathbf{w}_j^{*T}\mathbf{w}_j^*\right]^{-1}\mathbf{w}_j^{*T}\left[\left[\mathbf{p}_1^*, \quad \mathbf{p}_2^*, \quad \ldots, \quad \mathbf{p}_K^*\right]\tau + \mathbf{e}^* + \mathbf{d}^*\right. \\
&\quad \left. - \sum_{m=1}^{j-1}\hat{\tau}_m\mathbf{p}_m^*\right] \\
&= \tau_j\left[1 - \left[\mathbf{w}_j^{*T}\mathbf{w}_j^*\right]^{-1}\mathbf{w}_j^{*T}\mathbf{p}_j^*\right] - \sum_{m=j+1}^{K}\left[\mathbf{w}_j^{*T}\mathbf{w}_j^*\right]^{-1}\mathbf{w}_j^{*T}\mathbf{p}_j^*\tau_j \\
&\quad - \left[\mathbf{w}_j^{*T}\mathbf{w}_j^*\right]^{-1}\mathbf{w}_j^{*T}\mathbf{e}^* - \left[\mathbf{w}_j^{*T}\mathbf{w}_j^*\right]^{-1}\mathbf{w}_j^{*T}\mathbf{d}^* \\
&\quad - \sum_{m=1}^{j-1}\left[\mathbf{w}_j^{*T}\mathbf{w}_j^*\right]^{-1}\mathbf{w}_j^{*T}\mathbf{p}_m^*\left[\tau_m - \hat{\tau}_m\right]
\end{aligned} \qquad \text{Equation 2-8}$$

When there are no missing measurements, the score estimation error reduces to

$$\tau_j - \hat{\tau}_j = -\left[\mathbf{w}_j^T\mathbf{w}_j\right]^{-1}\mathbf{w}_j^T\mathbf{e} - \sum_{m=1}^{j-1}\left[\mathbf{w}_j^T\mathbf{w}_j\right]^{-1}\mathbf{w}_j^T\mathbf{p}_m\left[\tau_m - \hat{\tau}_m\right] \qquad \text{Equation 2-9}$$

Unlike PCA in Equation 2-3, the PLS score estimation error with no missing data has an error propagation term. This means that score estimation errors originating in measurement noise are transmitted to later scores. Errors propagate because the loading vector $\mathbf{w}_j^*$ is not required to be orthogonal to $\mathbf{p}_m^*$ when m is less than j although deviations from orthogonality are penalised (Burnham et al., 1996).

It is instructive to look at the various contributions to the score estimation error in Equation 2-8. The term $\mathbf{w}_j^{*\mathrm{T}}\mathbf{w}_j^*$ is analogous to the $\mathbf{p}_j^{*\mathrm{T}}\mathbf{p}_j^*$ term for the PCA single component projection algorithm. All error terms are magnified by the inverse of this term, which is large when the loading vector is nearly collinear with the missing variable subspace. The last term in Equation 2-8 is analogous to the last term in Equation 2-2 for the PCA analysis. It represents the propagation of errors made in estimating earlier scores into the error in the present score estimate. In PLS, it is present to a slight degree even when there is no missing data (see Equation 2-9), but becomes much larger whenever the missing measurements are such that $\mathbf{w}_j^*$ becomes more collinear with some of the $\mathbf{p}_m^*$ (m < j) making $\mathbf{w}_j^{*T}\mathbf{p}_m^*$ large. The second term in Equation 2-8 is analogous to the first term in Equation 2-2 and shows how much of later scores τ_m (m > j) are erroneously captured by this score τ_j. This term is zero with no missing data, but becomes large when the missing measurements are such that $\mathbf{w}_j^*$ and some of the $\mathbf{p}_m^*$ (m > j) become more collinear, and when subsequent latent directions account

for a large amount of the variance in $\mathbf{z}^*$. The first term in Equation 2-8 will normally be small but will be non-zero whenever $\mathbf{w}_j^{*T}\mathbf{p}_j^*$ is not equal to $\mathbf{w}_j^{*T}\mathbf{w}_j^*$. Variables in $\mathbf{z}$, which have little correlation with the Y variables, will make it larger. This situation will also tend to increase the deterministic residual contribution term $\mathbf{w}_j^{*T}\mathbf{d}^*$.

The matrix-vector form of the error equation analogous to Equation 2-6 is

$$\begin{aligned}\boldsymbol{\tau}_j - \hat{\boldsymbol{\tau}}_j &= \boldsymbol{\tau}_j - \left[\mathbf{w}_j^{*T}\mathbf{w}_j^*\right]^{-1}\mathbf{w}_j^{*T}\mathbf{z}^*(j) \\ &= \boldsymbol{\tau}_j - \left[\mathbf{w}_j^{*T}\mathbf{w}_j^*\right]^{-1}\mathbf{w}_j^{*T}\prod_{m=1}^{j-1}\left[\mathbf{I} - \left[\mathbf{w}_{j-m}^{*T}\mathbf{w}_{j-m}^*\right]^{-1}\mathbf{p}_{j-m}^*\mathbf{w}_{j-m}^{*T}\right] \\ &\quad \left[\mathbf{P}^*\boldsymbol{\tau} + \mathbf{e}^* + \mathbf{d}^*\right]\end{aligned} \qquad \text{Equation 2-10}$$

2.3 Handling Missing Data in PCA by Projection to the Model Plane

One of the main sources of error in using the single component projection algorithm for treating new data vectors with missing measurements arises from some of the variance in $\mathbf{z}^*$ being assigned to the wrong score. This problem is most acute when consecutive scores are of similar magnitude and the associated $\mathbf{p}^*$ loading vectors are close to collinear. One way to alleviate this difficulty for PCA is to calculate all A of the scores at once by projecting onto the hyperplane formed by the $\mathbf{p}_j^*$ vectors j = 1, 2, …, A. This is equivalent to saying that all of the loading vectors will be fitted to the data at once. When there are no missing measurements, this will result in scores identical to those calculated using the

single component projection method since the loading vectors are orthogonal. It is this loss of orthogonality that causes many of the problems associated with missing measurements.

Using the structure for a new object $\mathbf{z}^*(1)$ given by Equation 2-1, we can calculate the score estimation error arising from Equation 1-25 as

$$\begin{aligned}
\boldsymbol{\tau}_{1:A} - \hat{\boldsymbol{\tau}}_{1:A} &= \boldsymbol{\tau}_{1:A} - \left[\mathbf{P}_{1:A}^{*T}\mathbf{P}_{1:A}^{*}\right]^{-1}\mathbf{P}_{1:A}^{*T}\left[\mathbf{P}^{*}\boldsymbol{\tau} + \mathbf{e}^{*}\right] \\
&= \boldsymbol{\tau}_{1:A} - \left[\mathbf{P}_{1:A}^{*T}\mathbf{P}_{1:A}^{*}\right]^{-1}\mathbf{P}_{1:A}^{*T}\left[\left[\mathbf{P}_{1:A}^{*}, \quad \mathbf{P}_{A+1:K}^{*}\right]\boldsymbol{\tau} + \mathbf{e}^{*}\right] \\
&= \boldsymbol{\tau}_{1:A} - \left[\mathbf{P}_{1:A}^{*T}\mathbf{P}_{1:A}^{*}\right]^{-1}\mathbf{P}_{1:A}^{*T}\mathbf{P}_{1:A}^{*}\boldsymbol{\tau}_{1:A} - \left[\mathbf{P}_{1:A}^{*T}\mathbf{P}_{1:A}^{*}\right]^{-1}\mathbf{P}_{1:A}^{*T}\mathbf{P}_{A+1:K}^{*}\boldsymbol{\tau}_{A+1:K} \\
&\quad - \left[\mathbf{P}_{1:A}^{*T}\mathbf{P}_{1:A}^{*}\right]^{-1}\mathbf{P}_{1:A}^{*T}\mathbf{e}^{*} \\
&= -\left[\mathbf{P}_{1:A}^{*T}\mathbf{P}_{1:A}^{*}\right]^{-1}\mathbf{P}_{1:A}^{*T}\mathbf{P}_{A+1:K}^{*}\boldsymbol{\tau}_{A+1:K} - \left[\mathbf{P}_{1:A}^{*T}\mathbf{P}_{1:A}^{*}\right]^{-1}\mathbf{P}_{1:A}^{*T}\mathbf{e}^{*}
\end{aligned}$$

Equation 2-11

Here $\mathbf{P}_{A+1:K}$ and $\boldsymbol{\tau}_{A+1:K}$ represent the latent variable space that has been neglected in the PCA model. If the model dimension A has been correctly chosen then all $\boldsymbol{\tau}_{A+1:K}$ are zero.

The terms in this error expression can be compared with those for the standard single component projection algorithm in Equation 2-2. The first two terms are directly analogous, but note that the third term in the single component projection error expression (Equation 2-2) which shows the propagation of errors from preceding scores into the current estimate is absent in Equation 2-11. The absence of this term is a direct result of the fact that all scores are being estimated simultaneously in the projection algorithm.

The first term in Equation 2-11 represents the errors arising from incorrectly attributing some of the variance in $\mathbf{z}^*$ to the scores that are being

estimated. This variance is in the direction of higher dimensional latent variables that have been ignored. Both terms in Equation 2-11 can become large when $\mathbf{P}_{1:A}^{*T}\mathbf{P}_{1:A}^{*}$ becomes ill-conditioned. This situation can arise whenever a combination of missing measurements makes some of the columns of $\mathbf{P}_{1:A}^{*}$ nearly collinear. When this occurs, a biased regression method such as principle components regression (PCR) or ridge regression can be used to calculate the scores. Equation 1-25 is replaced by the estimator for the chosen regression method and Equation 2-11 is based on this new estimator.

2.3.1 Simulation Examples

This projection algorithm, using only two latent vectors in the model (A=2), was applied to the illustrative examples described in section 2.2.2. The mean square of the first two score estimate errors from projection to the model plane with ordinary least squares (OLS) are presented in Table 2-1 and Table 2-2.

The most obvious source of error with the projection algorithm with OLS arises in all the cases with three large scores. In these cases, there are three large latent vectors and only two were used in the model. As a result, the first term in Equation 2-11 is large. Variance in $\mathbf{z}^{*}$ arising from the third latent variable is being wrongly attributed to the first two dimensions. This emphasises the point that all significant latent variable dimensions must be included when this projection algorithm is being used to handle missing data. This is true even in a process monitoring application which will only use the first few dimensions. The

single component projection algorithm is not affected by poor choice of number of dimensions in the model since it proceeds sequentially.

Ill-conditioning of the $\mathbf{P}_{1:A}^{*T}\mathbf{P}_{1:A}^{*}$ matrix is the other major factor influencing the score estimation error terms for projection with OLS. In the examples, this clearly shows up in cases with $\mathbf{p}_1^{*T}\mathbf{p}_1^{*} = 0.0196$ and $\angle\mathbf{p}_1^{*},\mathbf{p}_2^{*} = 7°$ where the missing first variable greatly affects the magnitude of the first column of $\mathbf{P}_{1:A}^{*}$, and where the first two columns of $\mathbf{P}_{1:A}^{*}$ are nearly linearly dependent.

The results were re-calculated using PCR rather than ordinary least squares and the results placed in Table 2-1 and Table 2-2. The singular directions that were used in PCR were chosen based on validation of the resulting model. The sample data was divided into two groups and scores were calculated for each group separately with the other group in each case functioning as the validation data. The vectors chosen were the same in each case. In practice, data from the model building stage would be used as the validation set. The mean squared score estimation errors are much improved from the OLS results.

Once again, it should be remembered that the illustrative examples are designed to show extremes. Application to actual data will likely produce much smaller errors, particularly when higher dimensional data sets with some redundancy are used.

2.4 Missing Data Replacement using Conditional Mean Replacement

In this chapter, we are assuming that we have already built a PCA or PLS model and are only interested in handling missing data in future multivariate observations. Therefore, when calculating scores, good estimates of the variable means and the covariance matrix are available from the modelling step. We assume that the addition of the information in one new object will not change these estimates appreciably and we therefore use only the expectation step of the EM algorithm to calculate replacement values for the missing measurements. To assist in developing an expression for the replacement values for the missing measurements, the missing measurements are grouped in the first elements of the data vector, that is

$$z = \begin{bmatrix} z^{\#} \\ z^{*} \end{bmatrix}$$

The estimate of the covariance matrix, S, is

$$\begin{aligned} \mathbf{S} &= \mathbf{X}^T\mathbf{X}/_{N-1} \\ &= \left[\mathbf{TP}^T\right]^T\mathbf{TP}^T/_{N-1} \\ &= \mathbf{PT}^T\mathbf{TP}^T/_{N-1} \\ &= \mathbf{P}\Theta\mathbf{P}^T \end{aligned}$$

where $\Theta = \mathbf{T}^T\mathbf{T}/_{N-1}$ is a K by K diagonal matrix with the diagonal containing the estimated variances of the latent variables.

2.4.1 PCA

Substituting

$$\boldsymbol{P} = \begin{bmatrix} \boldsymbol{P}^{\#} \\ \boldsymbol{P}^{*} \end{bmatrix}$$

into the expression for S gives

$$\mathbf{S} = \begin{bmatrix} \mathbf{S}^{\#\#} & \mathbf{S}^{\#*} \\ \mathbf{S}^{*\#} & \mathbf{S}^{**} \end{bmatrix} = \begin{bmatrix} \mathbf{P}^{\#}\boldsymbol{\Theta}\mathbf{P}^{\#T} & \mathbf{P}^{\#}\boldsymbol{\Theta}\mathbf{P}^{*T} \\ \mathbf{P}^{*}\boldsymbol{\Theta}\mathbf{P}^{\#T} & \mathbf{P}^{*}\boldsymbol{\Theta}\mathbf{P}^{*T} \end{bmatrix}$$

Using this expression for S, the conditional expectation of the missing measurements (Johnson and Wichern, 1988) is given by

$$\begin{aligned} \hat{\mathbf{z}}^{\#} &= \mathbf{S}^{\#*}\left[\mathbf{S}^{**}\right]^{-1}\mathbf{z}^{*} \\ &= \mathbf{P}^{\#}\boldsymbol{\Theta}\mathbf{P}^{*T}\left[\mathbf{P}^{*}\boldsymbol{\Theta}\mathbf{P}^{*T}\right]^{-1}\mathbf{z}^{*} \end{aligned} \qquad \text{Equation 2-12}$$

The estimated missing measurements can then be used in the score calculation along with the observed data as if no measurements were missing. For PCA this gives

$$\begin{aligned} \hat{\tau}_j &= \begin{bmatrix} \hat{\mathbf{z}}^{\#}(j) \\ \mathbf{z}^{*}(j) \end{bmatrix}^{T} \mathbf{p}_j \\ &= \hat{\mathbf{z}}^{\#T}(j)\mathbf{p}_j^{\#} + \mathbf{z}^{*T}(j)\mathbf{p}_j^{*} \\ &= \left[\mathbf{S}^{\#*}\left[\mathbf{S}^{**}\right]^{-1}\mathbf{z}^{*}(j)\right]^{T}\mathbf{p}_j^{\#} + \mathbf{z}^{*T}(j)\mathbf{p}_j^{*} \\ &= \mathbf{z}^{*T}(j)\left[\mathbf{S}^{**}\right]^{-1}\mathbf{S}^{*\#}\mathbf{p}_j^{\#} + \mathbf{z}^{*T}(j)\mathbf{p}_j^{*} \end{aligned} \qquad \text{Equation 2-13}$$

The score calculation for PCA can also be re-written in terms of the loading vectors and score variances:

$$\begin{aligned}\hat{\boldsymbol{\tau}}_{1:A} &= \mathbf{P}_{1:A}^{T}\begin{bmatrix}\hat{\mathbf{z}}^{\#}\\ \mathbf{z}^{*}\end{bmatrix}\\ &= \mathbf{P}_{1:A}^{T}\begin{bmatrix}\mathbf{P}^{\#}\boldsymbol{\Theta}\mathbf{P}^{*T}\left[\mathbf{P}^{*}\boldsymbol{\Theta}\mathbf{P}^{*T}\right]^{-1}\mathbf{z}^{*}\\ \left[\mathbf{P}^{*}\boldsymbol{\Theta}\mathbf{P}^{*T}\right]\left[\mathbf{P}^{*}\boldsymbol{\Theta}\mathbf{P}^{*T}\right]^{-1}\mathbf{z}^{*}\end{bmatrix}\\ &= \mathbf{P}_{1:A}^{T}\begin{bmatrix}\mathbf{P}^{\#}\\ \mathbf{P}^{*}\end{bmatrix}\boldsymbol{\Theta}\mathbf{P}^{*T}\left[\mathbf{P}^{*}\boldsymbol{\Theta}\mathbf{P}^{*T}\right]^{-1}\mathbf{z}^{*}\\ &= \mathbf{P}_{1:A}^{T}\mathbf{P}\boldsymbol{\Theta}\mathbf{P}^{*T}\left[\mathbf{P}^{*}\boldsymbol{\Theta}\mathbf{P}^{*T}\right]^{-1}\mathbf{z}^{*}\\ &= \left[\mathbf{I},\quad \mathbf{0}\right]\boldsymbol{\Theta}\mathbf{P}^{*T}\left[\mathbf{P}^{*}\boldsymbol{\Theta}\mathbf{P}^{*T}\right]^{-1}\mathbf{z}^{*}\end{aligned} \qquad \text{Equation 2-14}$$

where $\mathbf{I}$ is an A by A identity matrix and $\mathbf{0}$ is an A by (K-A) matrix of zeros. The error in these scores is

$$\begin{aligned}\boldsymbol{\tau}_{1:A} - \hat{\boldsymbol{\tau}}_{1:A} &= \boldsymbol{\tau}_{1:A} - \left[\mathbf{I},\quad \mathbf{0}\right]\boldsymbol{\Theta}\mathbf{P}^{*T}\left[\mathbf{P}^{*}\boldsymbol{\Theta}\mathbf{P}^{*T}\right]^{-1}\mathbf{z}^{*}\\ &= \boldsymbol{\tau}_{1:A} - \left[\mathbf{I},\quad \mathbf{0}\right]\boldsymbol{\Theta}\mathbf{P}^{*T}\left[\mathbf{P}^{*}\boldsymbol{\Theta}\mathbf{P}^{*T}\right]^{-1}\left[\mathbf{P}^{*}\boldsymbol{\tau}+\mathbf{e}^{*}\right]\end{aligned} \qquad \text{Equation 2-15}$$

Applying this score calculation formula to the illustrative examples described in section 2.2.2 gives the results shown in Table 2-1 for the first score and Table 2-2 for the second score. The mean square of the score estimation errors for conditional mean replacement is lower than that for any of the other methods considered in this paper in all of the test cases studied.

A potential problem with using this conditional expectation replacement approach is that one needs to obtain the inverse of the (K-K_M) by (K-K_M) matrix $\mathbf{X}^{*T}\mathbf{X}^{*}$ (or equivalently $\mathbf{P}^{*}\boldsymbol{\Theta}\mathbf{P}^{*T}$) where K_M is the number of missing measurements. This matrix may be very ill-conditioned with highly correlated data. In this situation, the projection to the model plane approach of Equation

1-25 has the advantage of only needing the inverse to a much smaller A by A matrix.

The score calculation by conditional mean replacement can be related to the least squares estimate of the scores based on the training data set, assuming the same variables to be missing in each row of X. Denote the new X matrix with $\mathbf{z}^{\#}$ missing from each row as $\mathbf{X}^{*}$. Let $\mathbf{a}_j$ be the PCA or PLS vector that is used to calculate the j^{th} score from the complete or full X matrix ($\mathbf{p}_j$ for PCA and the j^{th} column of $\mathbf{W}\left(\mathbf{P}^T\mathbf{W}\right)^{-1}$ for PLS). The sum of squares objective function for the deviation of the vector of $\mathbf{t}_j$ values based on the model using the full data set ($\mathbf{t}_j = \mathbf{X}\mathbf{a}_j$) from its estimate based on the incomplete matrix $\mathbf{X}^{*}$ is given by

$$J = \left[\mathbf{t}_j - \mathbf{X}^{*}\boldsymbol{\beta}_j\right]^T\left[\mathbf{t}_j - \mathbf{X}^{*}\boldsymbol{\beta}_j\right]$$

and the least squares estimate is given by

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_j &= \left[\mathbf{X}^{*T}\mathbf{X}^{*}\right]^{-1}\mathbf{X}^{*T}\mathbf{t}_j \\
&= \left[\mathbf{X}^{*T}\mathbf{X}^{*}\right]^{-1}\mathbf{X}^{*T}\mathbf{X}\mathbf{a}_j \\
&= \left[\mathbf{X}^{*T}\mathbf{X}^{*}\right]^{-1}\mathbf{X}^{*\mathrm{T}}\left[\mathbf{X}^{\#} \quad \mathbf{X}^{*}\right]\mathbf{a}_j \\
&= \left[\mathbf{X}^{*T}\mathbf{X}^{*}\right]^{-1}\left[\mathbf{X}^{*\mathrm{T}}\mathbf{X}^{\#} \quad \mathbf{X}^{*\mathrm{T}}\mathbf{X}^{*}\right]\mathbf{a}_j \\
&= \left[\mathbf{S}^{**}\right]^{-1}\left[\mathbf{S}^{*\#} \quad \mathbf{S}^{**}\right]\mathbf{a}_j \\
&= \left[\mathbf{S}^{**}\right]^{-1}\mathbf{S}^{*\#}\mathbf{a}_j^{\#} + \mathbf{a}_j^{*}
\end{aligned}$$

Thus, the least squares estimate of the parameter vector that should be used to calculate a score with missing data gives the same result as replacing the missing data with its conditional expectation, assuming a normal distribution and

calculating the score using the complete data method (Equation 2-13). This is an important analogy because, as mentioned above, it is possible that $\mathbf{S}^{**}$ will be ill-conditioned or its inverse will not exist. This problem will be particularly acute when large numbers of variables are missing. A number of traditional regression solutions, such as ridge regression, principal components regression and PLS, could then be used to provide improved estimates of $\left[\mathbf{S}^{**}\right]^{-1}$ and $\hat{\boldsymbol{\beta}}_j$ in these situations.

To check whether ill-conditioning had any affect on the results in Table 2-1 and Table 2-2, both PCR and PLS regression were applied to calculate a model between $\mathbf{X}^*$ and $\mathbf{T}$. The number of components in these models was validated as in projection to the model plane. The best results in both cases were obtained when the methods used all latent variable dimensions, thereby reducing to the OLS solution. We can conclude that ill-conditioning was not a problem in this case.

2.4.2 PLS

For the PLS score calculation analysis, the vectors that are chosen as the basis for $\mathbf{X}$ are the columns of $\mathbf{W}$. This basis is not unique but was chosen for the convenient form that it gives to the score vector estimation equation. The matrix $\boldsymbol{\Omega}$ contains the variance of scores calculated in the PCA fashion with $\mathbf{W}$ as the loading matrix. They differ from PLS scores in that $\mathbf{W}$ rather than $\mathbf{P}$ is used to deflate the data vector, but $\boldsymbol{\Omega}$ is still diagonal. Thus

$$\boldsymbol{\Omega} = \mathbf{W}^T\mathbf{X}^T\mathbf{X}\mathbf{W}/N-1$$

and

$$\begin{aligned}\hat{\mathbf{z}}^{\#} &= \mathbf{S}^{\#*}\left[\mathbf{S}^{**}\right]^{-1}\mathbf{z}^{*} \\ &= \mathbf{W}^{\#}\boldsymbol{\Omega}\mathbf{W}^{*T}\left[\mathbf{W}^{*}\boldsymbol{\Omega}\mathbf{W}^{*T}\right]^{-1}\mathbf{z}^{*}\end{aligned} \qquad \text{Equation 2-16}$$

and the PLS score estimates are

$$\begin{aligned}\hat{\boldsymbol{\tau}}_{1:A} &= \left[\mathbf{W}_{1:A}^T\mathbf{P}_{1:A}\right]^{-1}\mathbf{W}_{1:A}^T\begin{bmatrix}\hat{\mathbf{z}}^{\#}\\ \mathbf{z}^{*}\end{bmatrix} \\ &= \left[\mathbf{W}_{1:A}^T\mathbf{P}_{1:A}\right]^{-1}\mathbf{W}_{1:A}^T\begin{bmatrix}\mathbf{W}^{\#}\boldsymbol{\Omega}\mathbf{W}^{*T}\left[\mathbf{W}^{*}\boldsymbol{\Omega}\mathbf{W}^{*T}\right]^{-1}\mathbf{z}^{*}\\ \mathbf{z}^{*}\end{bmatrix} \\ &= \left[\mathbf{W}_{1:A}^T\mathbf{P}_{1:A}\right]^{-1}\mathbf{W}_{1:A}^T\begin{bmatrix}\mathbf{W}^{\#}\boldsymbol{\Omega}\mathbf{W}^{*T}\left[\mathbf{W}^{*}\boldsymbol{\Omega}\mathbf{W}^{*T}\right]^{-1}\mathbf{z}^{*}\\ \left[\mathbf{W}^{*}\boldsymbol{\Omega}\mathbf{W}^{*T}\right]\left[\mathbf{W}^{*}\boldsymbol{\Omega}\mathbf{W}^{*T}\right]^{-1}\mathbf{z}^{*}\end{bmatrix} \\ &= \left[\mathbf{W}_{1:A}^T\mathbf{P}_{1:A}\right]^{-1}\mathbf{W}_{1:A}^T\begin{bmatrix}\mathbf{W}^{\#}\\ \mathbf{W}^{*}\end{bmatrix}\boldsymbol{\Omega}\mathbf{W}^{*T}\left[\mathbf{W}^{*}\boldsymbol{\Omega}\mathbf{W}^{*T}\right]^{-1}\mathbf{z}^{*} \\ &= \left[\mathbf{W}_{1:A}^T\mathbf{P}_{1:A}\right]^{-1}\mathbf{W}_{1:A}^T\mathbf{W}\boldsymbol{\Omega}\mathbf{W}^{*T}\left[\mathbf{W}^{*}\boldsymbol{\Omega}\mathbf{W}^{*T}\right]^{-1}\mathbf{z}^{*} \\ &= \left[\mathbf{W}_{1:A}^T\mathbf{P}_{1:A}\right]^{-1}\left[\mathbf{I},\quad \mathbf{0}\right]\boldsymbol{\Omega}\mathbf{W}^{*T}\left[\mathbf{W}^{*}\boldsymbol{\Omega}\mathbf{W}^{*T}\right]^{-1}\mathbf{z}^{*}\end{aligned}$$

where $\mathbf{I}$ is an A by A identity matrix and $\mathbf{0}$ is an A by (K-A) matrix of zeros. The error in the calculated PLS scores is

$$\begin{aligned}\boldsymbol{\tau}_{1:A} - \hat{\boldsymbol{\tau}}_{1:A} &= \boldsymbol{\tau}_{1:A} - \left[\mathbf{W}_{1:A}^T\mathbf{P}_{1:A}\right]^{-1}\left[\mathbf{I},\quad \mathbf{0}\right]\boldsymbol{\Omega}\mathbf{W}^{*T}\left[\mathbf{W}^{*}\boldsymbol{\Omega}\mathbf{W}^{*T}\right]^{-1}\mathbf{z}^{*} \\ &= \boldsymbol{\tau}_{1:A} - \left[\mathbf{W}_{1:A}^T\mathbf{P}_{1:A}\right]^{-1}\left[\mathbf{I},\quad \mathbf{0}\right]\boldsymbol{\Omega}\mathbf{W}^{*T}\left[\mathbf{W}^{*}\boldsymbol{\Omega}\mathbf{W}^{*T}\right]^{-1} \\ &\quad \left[\mathbf{P}^{*}\boldsymbol{\tau} + \mathbf{e}^{*} + \mathbf{d}^{*}\right]\end{aligned} \qquad \text{Equation 2-17}$$

The error can arise from: any violations of the assumptions of least squares regression, the noise term, ill-conditioning in $\mathbf{S}^{**}$, or lack of information in the measured variables about the unmeasured variables. The matrix $\mathbf{S}^{\#*}$ shows the covariance between the present and missing measurements. Any zero row in

$S^{\#*}$ shows a complete lack of information about a missing variable; thus the variation in that variable is unique to it or to the group of missing measurements. This method has information about the magnitude of the expected squared scores through its use of the covariance matrix of the data and the conditional expectation. This gives it an advantage over the single component projection method and the projection to the model plane.

2.5 Examples

2.5.1 Distillation Column

In the first example, a PLS model which was published in Kresta et al. (1994) is analysed. The data set comes from a distillation column simulation with a methanol-acetone-water (MAW) feed, 13 trays and a reboiler and reflux condenser. The reflux drum temperature is numbered 1, the tray temperatures are numbered 2 through 14 and the reboiler temperature is number 15. A steady state empirical PLS model related these temperatures to the outlet compositions. The purpose of the model was to provide a prediction of the outputs for an inferential control scheme using only simple measurements.

Kresta et al. emphasised that the PLS model should not be converted to polynomial regression coefficients but rather should be left in standard PLS form. This allows the missing data handling feature of PLS to be used if a measurement is missing. To illustrate this, the reboiler temperature was assumed to be missing.

This variable had a large polynomial regression coefficient for the prediction of the bottoms composition which would lead one to believe that losing this measurement would seriously affect the performance of the model. Losing this variable did not adversely affect prediction with the PLS model form using the missing data handling feature but did affect the regression coefficient form when the missing variable was replaced by its mean value. This work will analyse the PLS model to show why it was robust to losing the reboiler temperature measurement.

An error in the prediction of the output must be caused by an error in the scores. If the errors in the scores are small compared to their magnitude then the prediction error will be minor. The terms in Equation 2-8 show the terms which must be small in order to have an insignificant error in the score estimate and in this case the first three score estimation errors are

$$\tau_1 - \hat{\tau}_1 = 0.0390\tau_1 + [-0.4826]\tau_2 + 0.3114\tau_3 + 0.0538\tau_4 + [-0.1193]\tau_5 - [\mathbf{w}_1^{*T}\mathbf{w}_1^*]^{-1}\mathbf{w}_1^{*T}\mathbf{e}^* - [\mathbf{w}_1^{*T}\mathbf{w}_1^*]^{-1}\mathbf{w}_1^{*T}\mathbf{d}^* \quad \text{Equation 2-18}$$

$$\tau_2 - \hat{\tau}_2 = 0.0884\tau_2 + [-0.5992]\tau_3 + [-0.1036]\tau_4 + 0.2296\tau_5 + [-0.3695][\tau_1 - \hat{\tau}_1] - [\mathbf{w}_2^{*T}\mathbf{w}_2^*]^{-1}\mathbf{w}_2^{*T}\mathbf{e}^* - [\mathbf{w}_2^{*T}\mathbf{w}_2^*]^{-1}\mathbf{w}_2^{*T}\mathbf{d}^* \quad \text{Equation 2-19}$$

$$\tau_3 - \hat{\tau}_3 = 0.0330\tau_3 + 0.0419\tau_4 + [-0.0930]\tau_5 + 0.4316[\tau_1 - \hat{\tau}_1] + [-0.1476][\tau_2 - \hat{\tau}_2] - [\mathbf{w}_3^{*T}\mathbf{w}_3^*]^{-1}\mathbf{w}_3^{*T}\mathbf{e}^* - [\mathbf{w}_3^{*T}\mathbf{w}_3^*]^{-1}\mathbf{w}_3^{*T}\mathbf{d}^* \quad \text{Equation 2-20}$$

The magnitudes of the successive scores being close together can lead to problems and we see in Table 2-3 that the expected squared value of the second

score is different from the first by a factor of only 2.9. This will be serious if $\mathbf{w}_1^{*T}\mathbf{p}_2^*$ divided by $\mathbf{w}_1^{*T}\mathbf{w}_1^*$, the coefficient on τ_2 in the error for τ_1, is large compared to the ratio of the first and second expected squared scores. It can be seen from Equation 2-21 that the error is not very large. To quantify the magnitude of the effect, the expectation of the square of this term can be taken. The contribution of this term to the overall sum squared error relative to the expected square of the score is

$$\frac{E\left(\left[\mathbf{w}_1^{*T}\mathbf{p}_2^*/\mathbf{w}_1^{*T}\mathbf{w}_1^*\tau_2\right]^2\right)}{E\left(\tau_1^2\right)} = \left[-0.4826\right]^2\frac{E\left(\tau_2^{\,2}\right)}{E\left(\tau_1^2\right)}$$
$$= \left[-0.4826\right]^2\left[0.3502\right]$$
$$= 0.0816$$

since the expectation of all terms involving different scores or scores and residuals are zero. The second and third expected squared scores are close in magnitude as well and $\mathbf{w}_2^{*T}\mathbf{p}_3^*$ divided by $\mathbf{w}_2^{*T}\mathbf{w}_2^*$ is large, indicating an appreciable error in $\hat{\tau}_2$. The contribution of this term to the sum squared score error is approximately

$$\left[\frac{\mathbf{w}_2^{*T}\mathbf{p}_3^*}{\mathbf{w}_2^{*T}\mathbf{w}_2^*}\right]^2\frac{E\left(\tau_3^2\right)}{E\left(\tau_2^2\right)} = \left[-0.5992\right]^2\frac{E\left(\tau_3^2\right)}{E\left(\tau_1^2\right)}\frac{E\left(\tau_1^2\right)}{E\left(\tau_2^2\right)}$$
$$= \left[-0.5992\right]^2\frac{0.1088}{0.3502}$$
$$= 0.1115$$

This is approximate since it ignores the cross-correlation with $\hat{\tau}_1$ which can either reduce or increase the effect of τ_3 on $\hat{\tau}_2$. The full effect is

$$\left[\frac{\mathbf{w}_2^{*T}\mathbf{p}_3^*}{\mathbf{w}_2^{*T}\mathbf{w}_2^*}+\frac{\mathbf{w}_2^{*T}\mathbf{p}_1^*}{\mathbf{w}_2^{*T}\mathbf{w}_2^*}\frac{\mathbf{w}_1^{*T}\mathbf{p}_3^*}{\mathbf{w}_1^{*T}\mathbf{w}_1^*}\right]^2\frac{E\left(\tau_3^2\right)}{E\left(\tau_2^2\right)}$$
$$=\left[-0.5992+\left[-0.3695\right]0.3114\right]^2\frac{0.1088}{0.3502}$$
$$=0.1585$$

The first order approximation is useful in order to identify potential problems; it will be used in place of the full effect here.

The contribution to the squared error from the deterministic residual $E\left(\mathbf{w}_j^{*T}\mathbf{d}^*\mathbf{d}^{*T}\mathbf{w}_j^*\right)$ is small. The expected square of $\mathbf{w}_j^{*T}\mathbf{d}^*$ is shown because it is independent of the other terms so that there are no cross-terms involving this quantity in the overall error. The error term $\mathbf{w}_j^{*T}\mathbf{e}^*$ is assumed to be zero since no information about it is known.

Propagation of error will not be a problem for the two dominant scores because the first score has, by definition, no error propagation and the small first score error will not have a large impact on the second score. The approximate effect of error propagation on the second score estimation error is $\left[\frac{\mathbf{w}_2^{*T}\mathbf{p}_1^*}{\mathbf{w}_2^{*T}\mathbf{w}_2^*}\right]^2\frac{E\left(\left[\tau_1-\hat{\tau}_1\right]^2\right)}{E\left(\tau_2^2\right)}=0.0366$. From Equation 2-20 it can be seen that the third and subsequent scores have little effect on the $\hat{\tau}_3$ error but the error from the first

score estimate propagates through strongly ($\left[\frac{\mathbf{w}_3^{*T}\mathbf{p}_1^*}{\mathbf{w}_3^{*T}\mathbf{w}_3^*}\right]^2 \frac{E\left(\left[\tau_1 - \hat{\tau}_1\right]^2\right)}{E\left(\tau_3^2\right)} = 0.1606$). The effect of the second score error is much smaller: $\left[\frac{\mathbf{w}_3^{*T}\mathbf{p}_2^*}{\mathbf{w}_3^{*T}\mathbf{w}_3^*}\right]^2 \frac{E\left(\left[\tau_2 - \hat{\tau}_2\right]^2\right)}{E\left(\tau_3^2\right)} = 0.0162$. The fourth and fifth scores also are affected almost entirely by error propagation. The approximate magnitudes of the error propagation terms in the sum squared errors for the fourth score estimate are $\left[\frac{\mathbf{w}_4^{*T}\mathbf{p}_1^*}{\mathbf{w}_4^{*T}\mathbf{w}_4^*}\right]^2 \frac{E\left(\left[\tau_1 - \hat{\tau}_1\right]^2\right)}{E\left(\tau_4^2\right)} = 0.032$ and $\left[\frac{\mathbf{w}_4^{*T}\mathbf{p}_2^*}{\mathbf{w}_4^{*T}\mathbf{w}_4^*}\right]^2 \frac{E\left(\left[\tau_2 - \hat{\tau}_2\right]^2\right)}{E\left(\tau_4^2\right)} = 0.0859$. These scores can have appreciable errors relative to their magnitude even with small error propagation terms because they have a small magnitude relative to the first two scores.

The fourth row of Table 2-3 lists the expected squared errors for each of the scores. This error is not the true error since the true scores are unknown. Since there is no information about the measurement errors, the difference between the score calculated with the complete data ($\tilde{\tau}_j$) and the score calculated with the incomplete data ($\hat{\tau}_j$) is used as the error. This squared error is therefore not an absolute error but a relative error and indicates an increase in the variability of the score. Examining the ratio of the expected squared error to the squared score in the fourth row of Table 2-3, it can be seen that, as expected from Kresta et al. (1994) and this analysis, the expected errors are small compared to the

magnitudes of the scores for the large, important scores. There are large relative errors in two of the three small scores as indicated by this analysis.

The squared score errors when score calculation with conditional mean replacement is used are listed in the last row of Table 2-3. They are lower for all scores than the standard method and are negligible, with the largest percentage error being approximately 3%. There must be sufficient information in the data without the reboiler temperature to calculate the scores and this method is more efficient at using it than the standard method. The condition number of the matrix to be inverted in this case is 8414. This shows a dangerous degree of ill-conditioning.

Information about the robustness of the PLS model to the loss of the reboiler temperature can also be obtained from plots of the loadings. PLS groups variables with similar information about the plotted components together or reflected in the origin. Important variables have large weights and thus are far from the origin. Thus variables near a common line through the origin and far from the origin are important and contain similar information for the components plotted. The plot of the first versus the second component loadings in Figure 2-3 shows that while the reboiler temperature (variable 15) is very important, variable 14 has much the same information and variables 9 and 10 are as important for component 1. Variable 14 is physically the closest measurement to the reboiler temperature and can be seen to contain much the same information. This is why it has a similar loading in the first two components of the PLS model. Variable 10 is

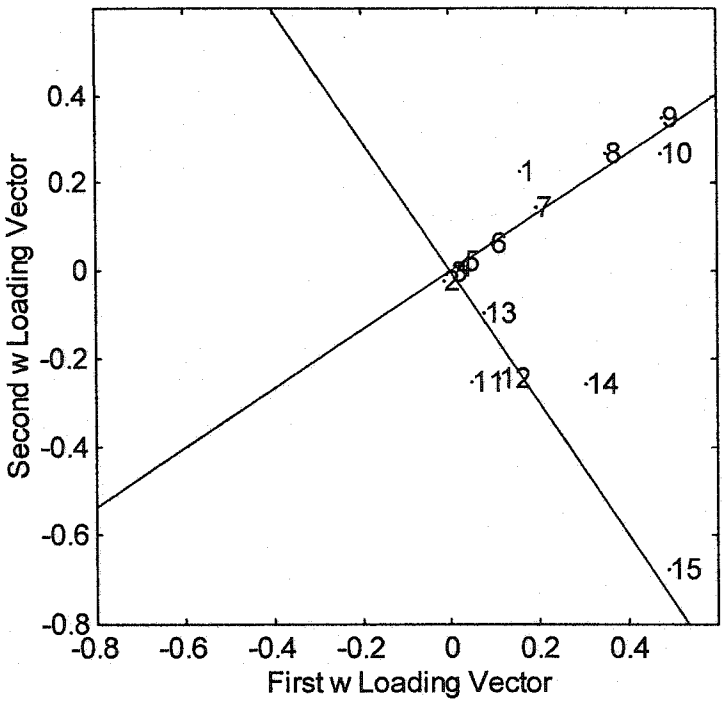


Figure 2-3: PLS Loading Plot for MAW Distillation Column

the feed tray temperature so it too affects the bottoms composition strongly and variable 9 is close to it and thus similar. Since important variables have large weights and are thus far from the origin, the term $\mathbf{w}_j^{*T}\mathbf{w}_j^*$ discussed above can be understood to decrease, and thus increase errors, when the amount of information about the output variance that vector represents decreases. If a variable with similar information to an important, but missing, variable is present then errors will be minimal.

Term	Component 1	Component 2	Component 3	Component 4	Component 5
$\frac{mean\left(\tilde{\tau}_j^2\right)}{mean\left(\tilde{\tau}_1^2\right)}$	1	0.3502	0.1088	0.0229	0.0027
$\mathbf{w}_j^{*T}\mathbf{w}_j^*$	0.7592	0.5434	0.8267	0.9857	0.9646
$mean\left(\mathbf{w}_j^{*T}\mathbf{d}^*\mathbf{d}^{*T}\mathbf{w}_j^*\right)$	2.8656e-6	5.4344e-6	2.0631e-6	1.7022e-7	4.2084e-7
$\frac{mean\left(\left[\tilde{\tau}_j-\hat{\tau}_j\right]^2\right)}{mean\left(\tilde{\tau}_j^2\right)}$ SCP	0.0938	0.2316	0.2774	0.0252	2.5714
$\frac{mean\left(\left[\tilde{\tau}_j-\hat{\tau}_j\right]^2\right)}{mean\left(\tilde{\tau}_j^2\right)}$ CMR	6.255e-4	2.596e-3	3.219e-3	1.176e-4	3.193e-2

Table 2-3: Score Error Terms for the PLS Model of the MAW Distillation Column with Bottoms Temperature Missing

2.5.2 The Kamyr Pulp Digester

This example uses historical data from a full scale, industrial process. The process is a continuous Kamyr pulp digester from which process measurements or process measurement averages were gathered hourly for a period of eight months. A PLS modelling study with this data was reported in Dayal et al. (1994) and Dayal (1992). The Kappa number, a variable which signifies the amount of lignin remaining in the pulp, was related to 21 process measurements and the Kappa number from the previous time period.

Missing measurements are common for this process, especially since two of the input measurements, upper cook zone active alkali (UCZAA) and past Kappa number, are laboratory measurements and are not measured every 12^{th} sample due to shift changes at the plant. Unfortunately, these variables are very important to the prediction of the current Kappa number. In fact, two of the three essential variables kept in the model when the number of variables was reduced were these laboratory measurements (Dayal, 1992).

A 5 component (A=5) PLS model was calculated using the complete objects. Cross-validation was used to determine the number of components. This model is slightly different from the model in Dayal et al. (1994) because the model reported here is calculated using only the 1919 complete objects.

When the number of latent variables is small, a simple qualitative method of assessing the impact of missing measurements in certain variables or sets of variables is to inspect the loading plots for various combinations of latent

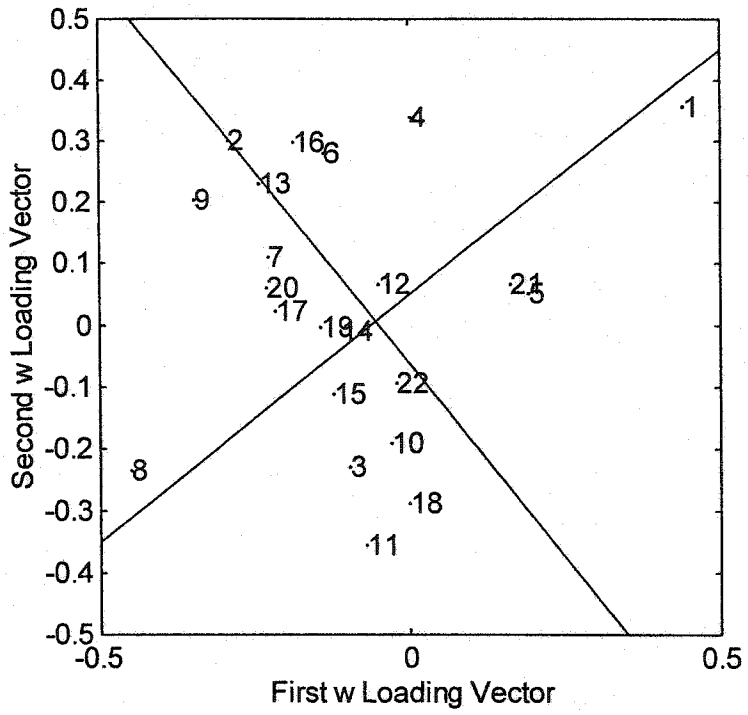


Figure 2-4: Plot of First Two PLS W Loading Vectors for Kamyr Digester

variables. This can also be used to determine which missing measurements will have the most impact on score calculation. Examining loading vector plots is less computationally demanding than brute force enumeration of possible missing measurement sets but may fail to indicate important combinations due to examining only 2 dimensional slices of the A dimensional plane. Other methods for determining high impact sets of missing measurements will be discussed in chapter 4.

Figure 2-4 shows the joint loading plot ($\mathbf{w}_1$ versus $\mathbf{w}_2$) for the first two latent variables. From this plot it is obvious that one of the dimensions in this two dimensional plane is strongly dependent upon variables 1 and 8 (past Kappa

number and UCZAA). A rotation of the axes indicated by the darker lines in Figure 2-4 has been added to illustrate how one of the dimensions is dominated by these two variables. Therefore, if a new multivariate observation has both variables 1 and 8 missing, most of the information defining one direction of the plane is lost. This can cause large errors in the estimates of τ_1 and τ_2.

2.5.2.1 A Study of the Effect of Combinations of Missing Measurements

Data sets were prepared with 2 or 4 variables missing (9% or 18% of the total number of variables) to illustrate the behaviour of the missing data algorithms on this industrial data. Two hundred unique data sets were generated for each number of missing variables. Each data set was formed by removing the columns in **X** corresponding to the missing variables. The scores and predicted Kappa number were then estimated for all objects using single component projection and conditional mean replacement and the mean squared difference between the estimates from incomplete data ($\hat{\tau}_j, \quad j = 1, 2, \ldots, 5$ and the predicted Kappa number $\hat{y}$) and those calculated with the complete data set ($\tilde{\tau}_j, \quad j = 1, 2, \ldots, 5$ and $\tilde{y}$) were tabulated. Histograms of the fractional mean squared score difference $1/1919 \sum_{i=1}^{1919} \left(\tilde{\tau}_{ij} - \hat{\tau}_{ij}\right)^2 / \tilde{\tau}_{ij}^2$ for the first three scores ($j = 1, 2, \ldots, 3$) and the fractional squared Kappa number differences $1/1919 \sum_{i=1}^{1919} \left(\tilde{y}_i - \hat{y}_i\right)^2 / \tilde{y}_i^2$ are shown in Figure 2-5 and Figure 2-6. The fractional mean

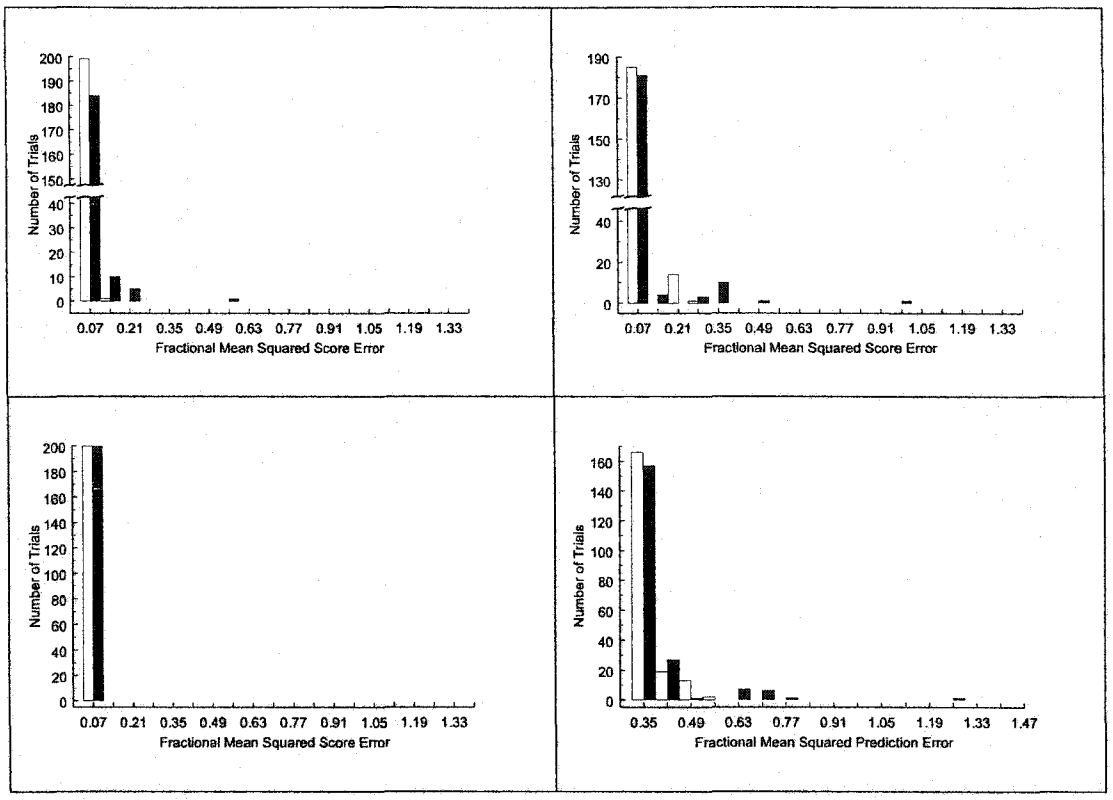


Figure 2-5: Histograms of Mean Squared Errors with 2 Variables Missing. Filled Bars are for Single Component Projection, Unfilled Bars for Conditional Mean Replacement

squared score difference was used to normalise the abscissa since the scores have different magnitudes (see Table 2-4).

All of the score histograms show that the fractional mean squared score differences for the majority of the combinations of missing measurements are small (less than 0.15 or 15%). Each algorithm has a small number of cases which have a higher mean squared difference. Comparing the two approaches, it can be seen that the single component projection method gives a higher percentage of large error cases, and a few cases of extreme error. In each plot in Figure 2-5, the most extreme estimation error arises when the combination of UCZAA and past

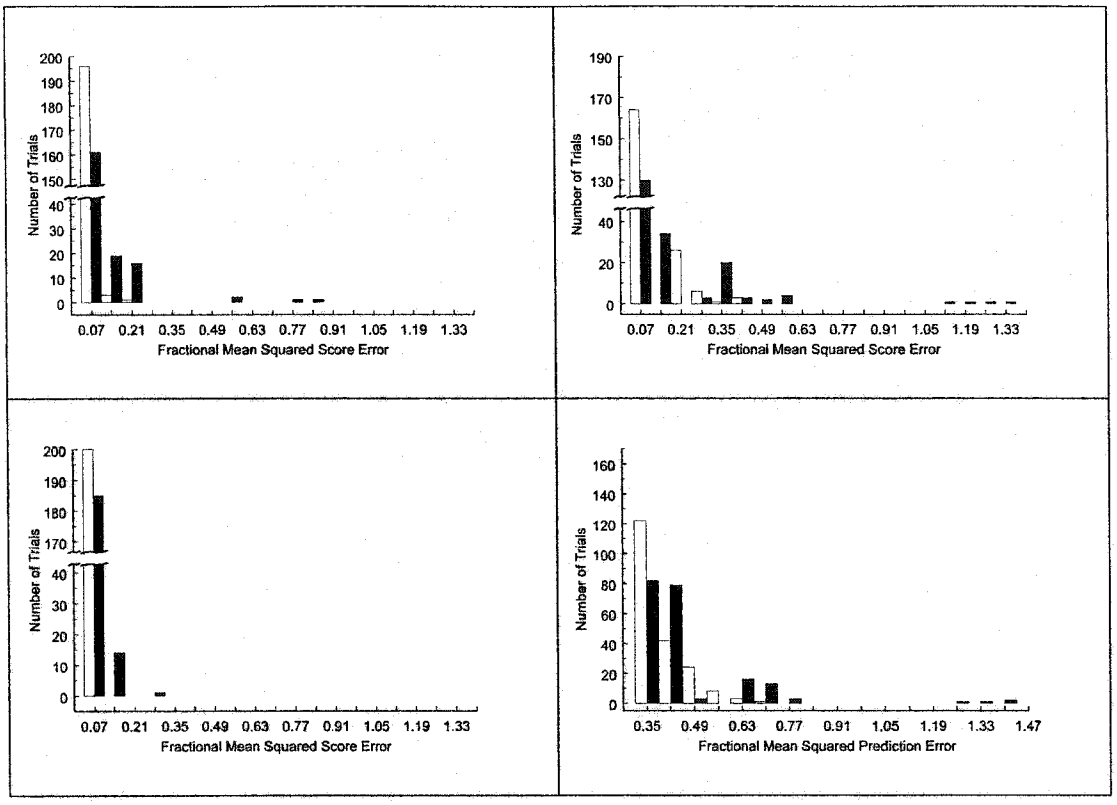


Figure 2-6: Histograms of Mean Squared Errors with 4 Variables Missing. Filled Bars are for Single Component Projection, Unfilled Bars for Conditional Mean Replacement

Kappa number is missing. This result was expected in light of the earlier discussion of the latent vector loading plot in Figure 2-4. This combination of missing variable measurements also produced the largest mean squared estimation differences in Figure 2-6 where 200 distinct sets of 4 measurements were missing. All of the extreme error cases for both the single component projection method and the conditional mean replacement method involved combinations which included the UCZAA and past Kappa number pair.

Term	Component 1	Component 2	Component 3	Component 4	Component 5
$\frac{mean\left(\tilde{\tau}_j^2\right)}{mean\left(\tilde{\tau}_1^2\right)}$	1	0.6100	0.9222	0.2909	0.2054
$\mathbf{w}_j^{*T}\mathbf{w}_j^*$	0.4626	0.7600	0.9356	0.9653	0.9594
$mean\left(\mathbf{w}_j^{*T}\mathbf{d}^*\mathbf{d}^{*T}\mathbf{w}_j^*\right)$	0.0561	0.0263	0.0311	0.0171	0.0041
$\frac{mean\left(\left[\tilde{\tau}_j-\hat{\tau}_j\right]^2\right)}{mean\left(\tilde{\tau}_j^2\right)}$ SCP	0.5768	0.9894	0.0742	0.0931	0.1022
$\frac{mean\left(\left[\tilde{\tau}_j-\hat{\tau}_j\right]^2\right)}{mean\left(\tilde{\tau}_j^2\right)}$ CMR	0.1323	0.2914	0.0301	0.0592	0.0419

Table 2-4: Score Estimation Error Terms for the 22 Variable PLS Model of the Kamyr Pulp Digester with UCZAA and Past Kappa Number Missing

Another interesting feature of Figure 2-5 and Figure 2-6 is the absence of large mean squared score differences for the third score. This implies that the error propagation term in Equation 2-8 for $\hat{\tau}_3$ must be small for all the missing data sets which cause a large first or second score estimation error.

2.5.2.2 Analysis of the Score Estimation Error for the Extreme Case

In this section, we examine the terms in the score estimation error equation for the single component projection method for the case where UCZAA and past Kappa number are missing. Substituting the values for this case into Equation 2-8, we get the following equations for the first three score estimation errors when the single component projection algorithm is used:

$$\tau_1 - \hat{\tau}_1 = -0.3744\tau_1 + 0.7449\tau_2 + 0.0255\tau_3 + 0.1204\tau_4 + 0.248 - \left[\mathbf{w}_1^{*T}\mathbf{w}_1^*\right]^{-1}\mathbf{w}_1^{*T}\mathbf{e}^* - \left[\mathbf{w}_1^{*T}\mathbf{w}_1^*\right]^{-1}\mathbf{w}_1^{*T}\mathbf{d}^* \quad \text{Equation 2-21}$$

$$\tau_2 - \hat{\tau}_2 = -0.0134\tau_2 + 0.0391\tau_3 + 0.0798\tau_4 - 0.0883\tau_5 + 0.9558\left[\tau_1 - \hat{\tau}_1\right] - \left[\mathbf{w}_2^{*T}\mathbf{w}_2^*\right]^{-1}\mathbf{w}_2^{*T}\mathbf{e}^* - \left[\mathbf{w}_2^{*T}\mathbf{w}_2^*\right]^{-1}\mathbf{w}_2^{*T}\mathbf{d}^* \quad \text{Equation 2-22}$$

$$\tau_3 - \hat{\tau}_3 = 0.0227\tau_3 + 0.0522\tau_4 + 0.0134\tau_5 - 0.0135\left[\tau_1 - \hat{\tau}_1\right] + 0.3168\left[\tau_2 - \hat{\tau}_2\right] - \left[\mathbf{w}_3^{*T}\mathbf{w}_3^*\right]^{-1}\mathbf{w}_3^{*T}\mathbf{e}^* - \left[\mathbf{w}_3^{*T}\mathbf{w}_3^*\right]^{-1}\mathbf{w}_3^{*T}\mathbf{d}^* \quad \text{Equation 2-23}$$

A large first score estimation difference (Equation 2-21) is expected since Table 2-4 shows that the first and second scores are both large and the value of $\mathbf{w}_1^{*T}\mathbf{w}_1^*$ is small. The inverse of the latter term is contained in each of the terms of Equation 2-21. The contribution of the deterministic residual $mean\left(\mathbf{w}_j^{*T}\mathbf{d}^*\mathbf{d}^{*T}\mathbf{w}_j^*\right)$

to the squared error is small and the measurement error contribution is unknown. The error in the first component affects all those that follow. It affects the error in $\hat{\tau}_2$ very heavily as we can see in Equation 2-22 where the coefficient on the first score estimation error is large at 0.9558. Since the rest of the coefficients in Equation 2-22 are small, the relative estimation difference shown for $\hat{\tau}_2$ in Table 2-4 is due almost entirely to propagation of the error in $\hat{\tau}_1$. The small to moderate coefficients in Equation 2-23, particularly those on the propagation terms, reveal why the third mean squared score difference is low.

The mean squared differences from the single component projection algorithm for the PLS model with both UCZAA and past Kappa number missing are listed in Table 2-4. These differences are obtained from the actual score estimates, not from Equation 2-8. The listed errors are large in the first two components, as expected from the earlier qualitative loading plot analysis and from the analysis of Equation 2-21 and Equation 2-22 in combination with the relative score magnitudes.

The mean squared score estimation errors obtained from missing data replacement by the conditional mean are also listed in Table 2-4. Although they are substantially less than those of the single component projection method, they still may be unacceptable in some applications. The condition number of the matrix to be inverted is 206.8, which indicates a sensitivity to noise and numerical error. Applying PCR and PLS to combat the ill-conditioning led to no improvement, which indicates that conditioning is not a problem in this case.

Term	Component 1	Component 2	Component 3
$\frac{mean(\tilde{\tau}_j^2)}{mean(\tilde{\tau}_1^2)}$	1	0.5754	0.3808
$\mathbf{w}_j^{*T}\mathbf{w}_j^*$	0.9219	0.3554	0.7227
$mean(\mathbf{w}_j^{*T}\mathbf{d}^*\mathbf{d}^{*T}\mathbf{w}_j^*)$	0.7632e-32	0.1635e-32	0.3704e-32
$\frac{mean([\tilde{\tau}_j - \hat{\tau}_j]^2)}{mean(\tilde{\tau}_j^2)}$ SCP	0.0476	1.1024	0.5545
$\frac{mean([\tilde{\tau}_j - \hat{\tau}_j]^2)}{mean(\tilde{\tau}_j^2)}$ CMR	0.0459	0.5961	0.4516

Table 2-5: Score Error Terms for the 3 Variable PLS Model of the Kamyr Pulp Digester with Chip Mass Flowrate Missing

2.5.2.3 Pruned PLS Models

Dayal et al. (1994) exploited the information in the loading plots to reduce the number of input variables used in the model without reducing its modelling power. This is desirable because reducing the number of variables reduces both the capital cost and the maintenance cost for sensors. Redundant measurements can, however, be desirable from the point of view of model prediction with missing variables. An analysis of the effects of the pruning on the expected score errors with missing variables follows.

The input variable set was reduced to 3 variables in Dayal (1992). The three variables were past Kappa number, UCZAA and chip mass feed rate (variables 1, 8 and 16). A 3 component model was calculated with these three

inputs and analysed when the chip mass feed rate measurement is missing with the results placed in Table 2-5 and Equation 2-24 to Equation 2-26.

$$\tau_1 - \hat{\tau}_1 = 0.034\tau_1 + [-0.2529]\tau_2 + 0.1597\tau_3 - [\mathbf{w}_1^{*T}\mathbf{w}_1^*]^{-1}\mathbf{w}_1^{*T}\mathbf{e}^* - [\mathbf{w}_1^{*T}\mathbf{w}_1^*]^{-1}\mathbf{w}_1^{*T}\mathbf{d}^* \quad \text{Equation 2-24}$$

$$\tau_2 - \hat{\tau}_2 = 0.0702\tau_2 + [-1.1896]\tau_3 + [-0.4916][\tau_1 - \hat{\tau}_1] - [\mathbf{w}_2^{*T}\mathbf{w}_2^*]^{-1}\mathbf{w}_2^{*T}\mathbf{e}^* - [\mathbf{w}_2^{*T}\mathbf{w}_2^*]^{-1}\mathbf{w}_2^{*T}\mathbf{d}^* \quad \text{Equation 2-25}$$

$$\tau_3 - \hat{\tau}_3 = 0\tau_3 + 0.2854[\tau_1 - \hat{\tau}_1] + [-0.5259][\tau_2 - \hat{\tau}_2] - [\mathbf{w}_3^{*T}\mathbf{w}_3^*]^{-1}\mathbf{w}_3^{*T}\mathbf{e}^* - [\mathbf{w}_3^{*T}\mathbf{w}_3^*]^{-1}\mathbf{w}_3^{*T}\mathbf{d}^* \quad \text{Equation 2-26}$$

The first score error should be small since the first coefficient in Equation 2-24 is small and the first score is twice as large as the second and the coefficients on the second and third scores are small. However, the second score error should be large since the second score is one and a half times the third and the coefficient multiplying the third score in Equation 2-25 is large. The approximate contribution to the sum squared error of the second score estimate by the third is very large at $[-1.1896]^2 \frac{E(\tau_3^2)}{E(\tau_2^2)} = 0.9365$. The third score error is expected to be large because the term multiplying the second score error in Equation 2-26 is large meaning that the error from the second score will propagate through. The approximate amount of sum squared error that is due to propagation from the second score error is $[-0.5259]^2 \frac{E([\tau_2 - \hat{\tau}_2]^2)}{E(\tau_3^2)} = 0.4607$. The contribution from deterministic residuals $mean(\mathbf{w}_j^{*T}\mathbf{d}^*\mathbf{d}^{*T}\mathbf{w}_j^*)$ is approximately the square of the

machine precision since the number of components is equal to the number of variables. Nothing is known about the measurement error term to estimate its size. We can see from the second to last row in the table that the second and then the third expected squared score errors are large compared to the full data calculated scores $\tilde{\tau}_j$.

The last row of the table shows that score calculation using conditional mean replacement also produces unacceptably large errors. The condition number of the matrix to be inverted for this example is 2.2 showing that ill-conditioning is not a factor. The pruning of variables has left this model very vulnerable to a failure in the chip mass feed rate sensor. The key indication of this is that $\mathbf{w}_2^{*T}\mathbf{w}_2^*$ is very small, indicating that much of the information about the second component is missing.

In Dayal et al. (1994), a five variable subset of the input set was also used in a PLS model. The variables used are drawn from the two directions in Figure 2-4; past Kappa number and UCZAA (variables 1 and 8) from the one direction and blow flow, white liquor flow rate and chip mass flow rate (variables 4, 9 and 16) from the other. A three component PLS model was calculated. The effect of losing the chip mass flow rate measurement is now quite different, as can be seen in Equation 2-27 to Equation 2-29 and Table 2-6.

$$\tau_1 - \hat{\tau}_1 = 0.0512\tau_1 + [-0.1466]\tau_2 + [-0.0409]\tau_3 - [\mathbf{w}_1^{*T}\mathbf{w}_1^*]^{-1}\mathbf{w}_1^{*T}\mathbf{e}^* - [\mathbf{w}_1^{*T}\mathbf{w}_1^*]^{-1}\mathbf{w}_1^{*T}\mathbf{d}^* \quad \text{Equation 2-27}$$

Term	Component 1	Component 2	Component 3
$\frac{E(\tilde{\tau}_j^2)}{E(\tilde{\tau}_1^2)}$	1	0.7020	0.2262
$\mathbf{w}_j^{*T}\mathbf{w}_j^*$	0.9355	0.6699	0.8782
$mean(\mathbf{w}_j^{*T}\mathbf{d}^*\mathbf{d}^{*T}\mathbf{w}_j^*)$	0.0128	0.0654	0.0241
$\frac{mean([\tilde{\tau}_j - \hat{\tau}_j]^2)}{mean(\tilde{\tau}_j^2)}$ SCP	0.0255	0.1051	0.2290
$\frac{mean([\tilde{\tau}_j - \hat{\tau}_j]^2)}{mean(\tilde{\tau}_j^2)}$ CMR	0.0128	0.0682	0.1388

Table 2-6: Score Error Terms for the 5 Variable PLS Model of the Kamyr Pulp Digester with Chip Mass Flowrate Missing

$$\tau_2 - \hat{\tau}_2 = [-0.0296]\tau_2 + 0.1292\tau_3 + 0.1107[\tau_1 - \hat{\tau}_1] - [\mathbf{w}_2^{*T}\mathbf{w}_2^*]^{-1}\mathbf{w}_2^{*T}\mathbf{e}^* - [\mathbf{w}_2^{*T}\mathbf{w}_2^*]^{-1}\mathbf{w}_2^{*T}\mathbf{d}^* \quad \text{Equation 2-28}$$

$$\tau_3 - \hat{\tau}_3 = [-0.0788]\tau_3 + [-0.1759][\tau_1 - \hat{\tau}_1] + 0.3273[\tau_2 - \hat{\tau}_2] - [\mathbf{w}_3^{*T}\mathbf{w}_3^*]^{-1}\mathbf{w}_3^{*T}\mathbf{e}^* - [\mathbf{w}_3^{*T}\mathbf{w}_3^*]^{-1}\mathbf{w}_3^{*T}\mathbf{d}^* \quad \text{Equation 2-29}$$

We can see that an improvement in the score prediction error can be expected over the 3 variable model. The first two score magnitudes are closer together but $\mathbf{w}_j^{*T}\mathbf{w}_j^*$ is larger in all cases and the coefficients multiplying the scores are all small. This is what is anticipated since problems in $\mathbf{w}_j^{*T}\mathbf{w}_j^*$ are associated with the loss of large weights and the weighting is spread out over more variables when redundant variables are present. The $\mathbf{w}_3^{*T}\mathbf{p}_j^*$ terms are smaller because more information is retained with the redundant variables and

with full information the terms reported will be zero. The term $\left[1-\frac{\mathbf{p}_j^{*T}\mathbf{w}_j^*}{\mathbf{w}_j^{*T}\mathbf{w}_j^*}\right]$ is larger in some of the cases but overall this term remains negligible. The contribution from the deterministic residuals $mean\left(\mathbf{w}_j^{*T}\mathbf{d}^*\mathbf{d}^{*T}\mathbf{w}_j^*\right)$ is small and the contribution from measurement error is unknown. We can see from the expected squared errors, where the full data calculated score $\tilde{\tau}_j$ is used in place of the unknown true score, that an improvement has indeed been made in the robustness of the model to chip mass feed rate sensor failure. The last row of the table shows that the results for score calculation with conditional mean replacement have also improved. The condition number of the matrix to be inverted in this score calculation has only increased marginally, to 5.6, with the addition of the two variables so ill-conditioning is not a problem.

2.6 Conclusions

A novel missing measurement score calculation algorithm called conditional mean replacement is proposed in this chapter, and the factors affecting the performance of both the novel and existing algorithms are identified and illustrated by designed and industrial process data sets. Conditional mean replacement of the missing measurements is found to be superior to the existing methods. Several approaches for estimating the scores in the presence of missing data are analysed: a single component projection method, a method of

simultaneous projection to the model plane, and conditional mean replacement of the missing measurements. Expressions are developed for the score errors arising from each of the methods, and an analysis of the major sources of error performed. Simulated data sets, designed specifically to accentuate various sources of error, are used to illustrate the nature of these errors. Two data sets, one a simulated process and the other industrial, are also analysed in order to show more typical results, as well as illustrating situations where errors become large.

An expression for the score estimation error from the single component projection missing data algorithm shows that the error increases with: (i) collinearity of the loading vectors, after loadings for the missing measurements are removed, and (ii) similarity in the magnitudes of the scores. An increase in the noise variance in the measurements will also increase the error. Large errors in a PLS score calculation using the single component projection method are shown to arise when the **w** and **p** loading vectors are significantly different. This is caused by the presence of variables in the independent data block which do not explain a significant amount of the dependent data. Errors are shown to propagate from estimated scores to subsequently calculated ones through deflation. The analysis is demonstrated on industrial process data and a simulated process taken from the literature.

Improvement over single component projection PCA score calculation is possible by fitting all loading vectors at once in a procedure called Projection to the Model Plane. The score estimation error equation developed for this method

showed the sources of the error to be: (i) under-estimating the dimensionality of the data, (ii) noise and (iii) ill-conditioning in the least squares projection. A biased regression algorithm is recommended for the projection and was shown to combat ill-conditioning.

Another, generally better method to calculate scores when there are missing measurements, which can be applied to both PCA and PLS, is to replace a missing variable by its conditional mean given the variables that are observed and then to use standard score estimation routines. This is termed Conditional Mean Replacement, and it had the lowest mean squared score estimation error in all examples evaluated here. This method has information about the expected squared magnitude of the scores, therefore the error is due only to lack of information, numerical ill-conditioning, noise, or violations of the assumptions made in the least squares regression between the measured and unmeasured variables. Identical score estimates are obtained if it is assumed either that all of the variables follow a multivariate normal distribution and the unmeasured variables are replaced by their conditional means (maximum likelihood estimates), or that the assumptions of least squares regression for a linear model between the measured variables and the scores hold. A biased regression method can be used in place of least squares to combat any ill-conditioning in the latter approach; this was not necessary in any of the examples in this chapter.

An example given in the literature as an illustration of PLS robustness to missing measurements was analysed. It was shown that the factors that cause

score calculation error with missing measurements identified in this chapter were small for the MAW column PLS model. This agrees with the conclusions in the literature.

The analysis of the industrial data set in this chapter agrees with the guideline that most of the missing data algorithms perform quite well with up to 20% of the measurements missing. However, the theoretical score error analysis shows that certain critical combinations of missing measurements can give rise to large errors. This is illustrated by the pathological simulated data sets and the industrial example. In these situations, the conditional mean replacement method is shown to be superior to the single component projection method and the simultaneous projection to the plane method. Pruning of variables during model building was shown to have potential to reduce the ability of the model to be used with missing measurements. Leaving in some redundant variables was shown to increase robustness while leaving potential for a reduction in the number of variables.

Chapter 3: Performance Measures for PCA and PLS Applications

3.1 Introduction

When all variables are measured, statistics for the prediction error variance and critical values for alarming have been developed to show how reliable the models are in use. These have been reviewed in chapter 1. The practice has been to continue using these statistics when measurements are missing even though the performance of the model will have degraded. System designers and operators need to know when performance will have been reduced by the presence of missing measurements to the point that the application has to be placed offline. Any increase in uncertainty in the results needs to be communicated to the operator so that they can have the appropriate level of confidence in the results that are presented. The purpose of this chapter is to develop expressions for the uncertainty introduced by the missing measurements and to develop diagnostics to aid in the determination of which missing measurements contribute the most to the uncertainty.

This chapter will characterise the uncertainty due to missing measurements in the prediction, SPE, Hotelling T^2 and score values in PCA and PLS applications so that conclusions can be made about the performance of the application with missing measurements. The performance measures developed

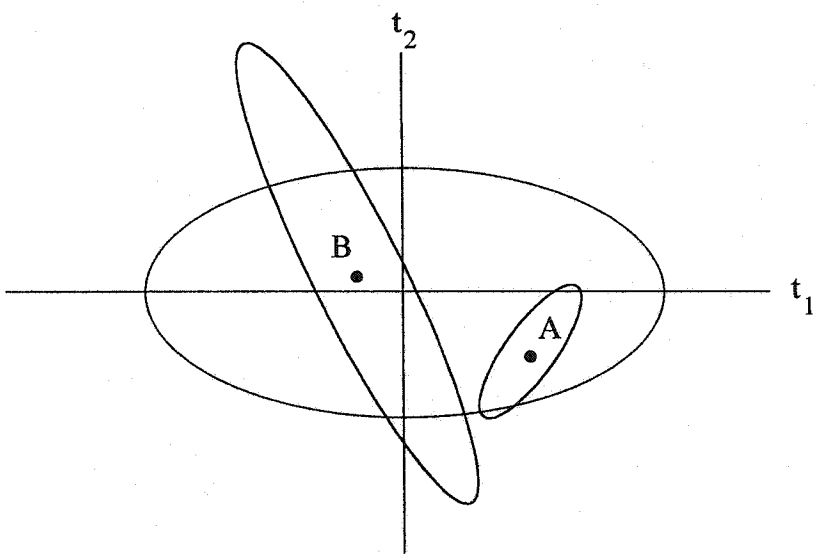


Figure 3-1: Score-score plot with SPC control chart limits and uncertainty regions for the scores due to missing measurements

distinguish between situations where model performance with missing measurements will continue to be acceptable and situations where it will become unacceptable. In the latter case, the measurements must be recovered or the application shut down. It will show how the uncertainty arising from missing measurements for these statistics gives additional information about the state of the process and our knowledge of it. A graphical example of this is given in Figure 3-1 where uncertainty regions are shown for two points on the score-score plot. The point marked A is closer to the control chart limit than point B but its uncertainty region caused by missing measurements is smaller and the probability that the actual value exceeds the control limit is much lower.

A PLS model from an industrial data set will be used to show how the missing measurement uncertainty intervals for the calculated statistics compare to

the values obtained when all of the measurements are present. Several missing measurement score calculation methods (CMR, SCP, mean replacement) will be used to calculate statistic values that will be plotted against the uncertainty intervals to show how they compare. This will allow the different methods' performance on industrial data to be compared.

The probability density for the missing measurements is mapped in this work into the score, Hotelling T^2, SPE and prediction spaces and uncertainty intervals and regions are generated. This is not dependent on the missing data algorithm, only the measured values and the data distribution. This is illustrated visually in Figure 3-2 where the top of the left hand column shows a plot of objects drawn at random from a 3 dimensional distribution with the scores and Hotelling T^2 values from these objects plotted underneath. The right hand column shows the effect of fixing one of the variables (z) before drawing the objects. The conditional distribution for x and y given z (shown in Figure 3-2 (d)) is propagated into the score and Hotelling T^2 spaces as shown in Figure 3-2 (e) and (f). This shows the joint distribution of t_1 and t_2 and the distribution of Hotelling T^2 for a fixed or observed value of z if x and y are missing at a given time. Note that the region occupied by the points in plot (e) forms an ellipse like the uncertainty regions in Figure 3-1. Rather than representing an object with missing measurements by selecting one point in each of the plots on the right hand side of Figure 3-2, this work will provide the uncertainty regions in the score, prediction, Hotelling T^2 and SPE spaces arising from the missing measurements.

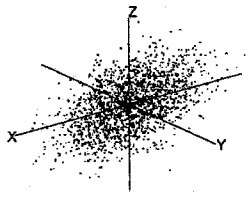
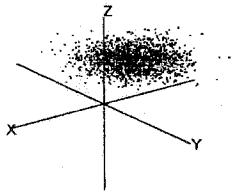
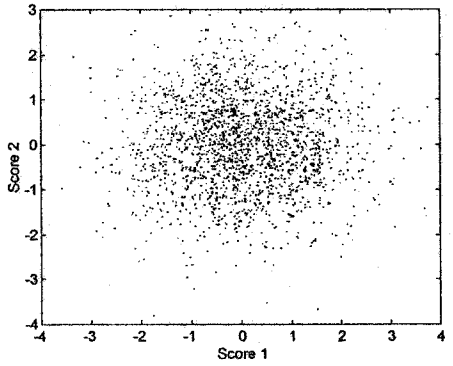
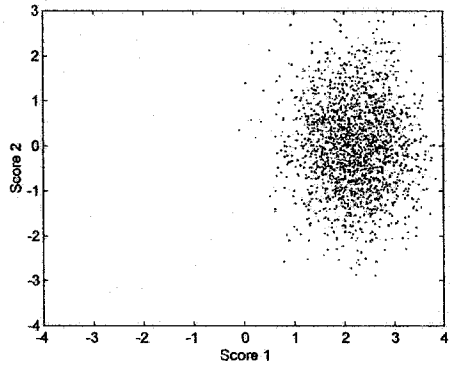
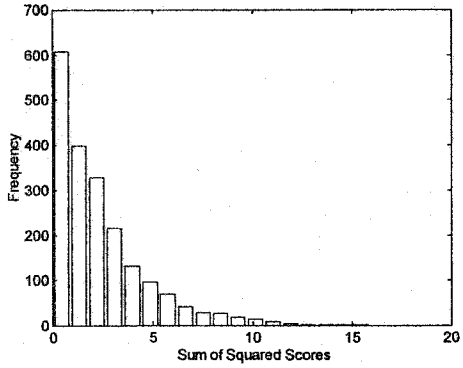
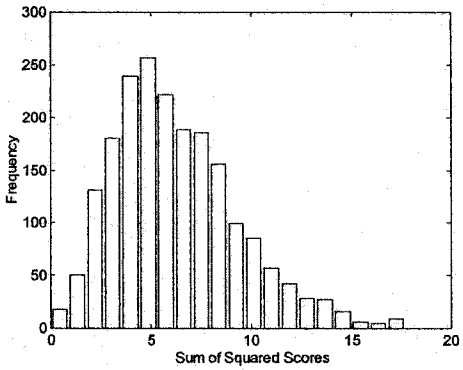
Randomly Drawn Objects	Objects Drawn from those with a Fixed z Value
	
(a) Plot of Objects Randomly Drawn from the Distribution	(d) Plot of Objects Randomly Drawn from those with a Fixed Z Value
	
(b) Plot of Scores from Objects Randomly Drawn from the Distribution	(e) Plot of Scores from Objects Randomly Drawn from those with a Fixed Z Value
	
(c) Histogram of T^2 Values from Objects Randomly Drawn from the Distribution	(f) Histogram of T^2 Values from Objects Randomly Drawn from those with a Fixed Z Value

Figure 3-2: Plots showing mapping of objects with and without restrictions on one of the variables.

As referenced in sections 1.3 and 1.4, prediction confidence limits and critical values in the literature have been generated from an assumption that the data is normally distributed and this assumption is used here as well. The conditional distribution of the missing measurements given the measured ones can be derived from this assumption and is used to associate a probability with each possible value of the missing measurements. A normal distribution conditional on fixed z was used to generate the objects for the plots on the right hand side of Figure 3-2. As shown in section 2.4.1, Equation 2-12 is valid when all data are normally distributed or the residuals of regressing the present measurements on the missing measurements are normally distributed. Alternatively, a historical data set could be used to provide a conditional distribution but this would require a very large amount of data since a representative number of objects would have to exist for each possible sample of the present measurements.

3.2 Distribution for Scores

The scores and the variable contributions to the scores are used in analysing both prediction and monitoring applications. This section of the thesis will develop the properties of the distributions that will be used in later sections. The distributions will be used to connect uncertainty in the prediction, Hotelling T^2 or SPE with individual missing measurements. This will aid in determining which missing measurements have a critical influence on the uncertainty in the results. If the uncertainty regions due to missing measurements are too large (for

example point B on Figure 3-1) and measurements for the most critical variables can not be recovered, monitoring or prediction may have to be discontinued until these critical measurements become available.

3.2.1 Uncertainty Region for Scores

The scores are computed as a linear combination of the measurements and are thus normally distributed under our assumed data distribution. The conditional distribution of the computed scores given the present measurements is then also normally distributed with mean and variance developed below:

$$\begin{aligned} E\left(\boldsymbol{\tau}|\mathbf{z}^*,\mathbf{S}\right) &= E\left(\mathbf{R}^{*T}\mathbf{z}^* + \mathbf{R}^{\#T}\mathbf{z}^{\#}|\mathbf{z}^*,\mathbf{S}\right) \\ &= \mathbf{R}^{*T}\mathbf{z}^* + \mathbf{R}^{\#T}E\left(\mathbf{z}^{\#}|\mathbf{z}^*,\mathbf{S}\right) \\ &= \mathbf{R}^{*T}\mathbf{z}^* + \mathbf{R}^{\#T}\mathbf{S}^{\#*}\left[\mathbf{S}^{**}\right]^{+}\mathbf{z}^* \\ &= \left[\mathbf{R}^{*T} + \mathbf{R}^{\#T}\mathbf{S}^{\#*}\left[\mathbf{S}^{**}\right]^{+}\right]\mathbf{z}^* \end{aligned} \qquad \text{Equation 3-1}$$

$$\begin{aligned} \mathrm{var}\left(\boldsymbol{\tau}|\mathbf{z}^*,\mathbf{S}\right) &= E\left(\left[\boldsymbol{\tau} - E\left(\boldsymbol{\tau}|\mathbf{z}^*,\mathbf{S}\right)\right]\left[\boldsymbol{\tau} - E\left(\boldsymbol{\tau}|\mathbf{z}^*,\mathbf{S}\right)\right]^T|\mathbf{z}^*,\mathbf{S}\right) \\ &= E\left(\left[\mathbf{R}^{*T}\mathbf{z}^* + \mathbf{R}^{\#T}\mathbf{z}^{\#} - \left[\mathbf{R}^{*T} + \mathbf{R}^{\#T}\mathbf{S}^{\#*}\left[\mathbf{S}^{**}\right]^{+}\right]\mathbf{z}^*\right]\right. \\ &\quad \left.\left[\mathbf{R}^{*T}\mathbf{z}^* + \mathbf{R}^{\#T}\mathbf{z}^{\#} - \left[\mathbf{R}^{*T} + \mathbf{R}^{\#T}\mathbf{S}^{\#*}\left[\mathbf{S}^{**}\right]^{+}\right]\mathbf{z}^*\right]^T|\mathbf{z}^*,\mathbf{S}\right) \\ &= E\left(\left[\mathbf{R}^{\#T}\mathbf{z}^{\#} - \mathbf{R}^{\#T}\mathbf{S}^{\#*}\left[\mathbf{S}^{**}\right]^{+}\mathbf{z}^*\right]\left[\mathbf{R}^{\#T}\mathbf{z}^{\#} - \mathbf{R}^{\#T}\mathbf{S}^{\#*}\left[\mathbf{S}^{**}\right]^{+}\mathbf{z}^*\right]^T|\mathbf{z}^*,\mathbf{S}\right) \\ &= E\left(\mathbf{R}^{\#T}\left[\mathbf{z}^{\#} - \mathbf{S}^{\#*}\left[\mathbf{S}^{**}\right]^{+}\mathbf{z}^*\right]\left[\mathbf{z}^{\#} - \mathbf{S}^{\#*}\left[\mathbf{S}^{**}\right]^{+}\mathbf{z}^*\right]^T\mathbf{R}^{\#}|\mathbf{z}^*,\mathbf{S}\right) \\ &= \mathbf{R}^{\#T}E\left(\left[\mathbf{z}^{\#} - \mathbf{S}^{\#*}\left[\mathbf{S}^{**}\right]^{+}\mathbf{z}^*\right]\left[\mathbf{z}^{\#} - \mathbf{S}^{\#*}\left[\mathbf{S}^{**}\right]^{+}\mathbf{z}^*\right]^T|\mathbf{z}^*,\mathbf{S}\right)\mathbf{R}^{\#} \\ &= \mathbf{R}^{\#T}\,\mathrm{var}\left(\mathbf{z}^{\#}|\mathbf{z}^*,\mathbf{S}\right)\mathbf{R}^{\#} \\ &= \mathbf{R}^{\#T}\left[\mathbf{S}^{\#\#} - \mathbf{S}^{\#*}\left[\mathbf{S}^{**}\right]^{+}\mathbf{S}^{*\#}\right]\mathbf{R}^{\#} \end{aligned} \qquad \text{Equation 3-2}$$

where $\left[\mathbf{S}^{**}\right]^{+}$ is the pseudo-inverse of $\mathbf{S}^{**}$.

The uncertainty in the scores arising from the missing measurements $\mathbf{z}^{\#}$ can be examined one score at a time by looking at the diagonal elements of Equation 3-2 or in pairs by plotting an uncertainty ellipse based on Equation 3-1 and Equation 3-2 on a score-score plot. Since the number of scores is usually small, examining them in pairs on the score-score plot is not an onerous task.

3.2.2 Uncertainty in Variable Contributions to the Scores with Missing Measurements

A contribution to a score is the term in the calculation that involves an individual measurement. We will examine the contribution of the k^{th} variable to the j^{th} score. When z_k is present, the contribution is a fixed value as shown in Equation 1-1. The contribution of a missing measurement (z_k missing) can not be calculated but we can calculate the expected value and the variance around that value conditional on the measured variables and the data distribution as shown below.

$$E\left(r_{kj}z_k\middle|\mathbf{z}^*,\mathbf{S}\right)=r_{kj}E\left(z_k\middle|\mathbf{z}^*,\mathbf{S}\right)=r_{kj}\mathbf{s}_k^{*T}\left[\mathbf{S}^{**}\right]^{+}\mathbf{z}^* \qquad \text{Equation 3-3}$$

$$\operatorname{var}\left(r_{kj}z_k\middle|\mathbf{z}^*,\mathbf{S}\right)=r_{kj}^2\operatorname{var}\left(z_k\middle|\mathbf{z}^*,\mathbf{S}\right)=r_{kj}^2\left[s_{kk}-\mathbf{s}_k^{*T}\left[\mathbf{S}^{**}\right]^{+}\mathbf{s}_k^{*}\right] \qquad \text{Equation 3-4}$$

The distribution of a contribution for an individual missing measurement gives an indication of how uncertainty from that unmeasured variable contributes

to the overall uncertainty in the score. This is not the total effect of the individual measurement on the uncertainty; the effect of regaining a missing measurement would also include reducing the conditional variance in missing measurements to which it is correlated. The distribution of the contribution for an individual missing measurement is an indication of the minimum effect of the measurement on the statistic and can be used to eliminate measurements that have little effect and to do a preliminary ranking of those that have a large effect.

3.3 Prediction

3.3.1 Diagnostic

The diagnostic that we will develop is the conditional variance of the prediction. There are two factors that contribute to the uncertainty in the prediction: the prediction error, which exists even when all measurements are present, and the uncertainty arising solely from the missing measurements.

In order to approximate the conditional variance of the prediction, we expand the prediction of y around the estimate of the PLS loading matrix $\mathbf{Q}$ ($\mathbf{Q}_o$) and the conditional mean score $\boldsymbol{\tau}_o = E\left(\boldsymbol{\tau}\middle|\mathbf{z}^*,\mathbf{S}\right)$. This approach separates the effects of the variance due to estimation of $\mathbf{Q}$ and the variance due to the scores.

$$\hat{\mathbf{y}} \approx \mathbf{Q}_o\boldsymbol{\tau}_o + \left[\mathbf{Q}-\mathbf{Q}_o\right]\boldsymbol{\tau}_o + \mathbf{Q}_o\left[\boldsymbol{\tau}-\boldsymbol{\tau}_o\right]$$

If $\mathbf{Q}$ and $\boldsymbol{\tau}$ are normally distributed then $\hat{\mathbf{y}}$ is normally distributed and $E(\hat{\mathbf{y}}) = \mathbf{Q}_o\boldsymbol{\tau}_o$ which is the prediction obtained from the CMR scores.

$$\hat{\mathbf{y}} - E(\hat{\mathbf{y}}) \approx \mathbf{Q}_o\boldsymbol{\tau}_o + [\mathbf{Q}-\mathbf{Q}_o]\boldsymbol{\tau}_o + \mathbf{Q}_o[\boldsymbol{\tau}-\boldsymbol{\tau}_o] - \mathbf{Q}_o\boldsymbol{\tau}_o$$
$$= [\mathbf{Q}-\mathbf{Q}_o]\boldsymbol{\tau}_o + \mathbf{Q}_o[\boldsymbol{\tau}-\boldsymbol{\tau}_o]$$

$$E\left([\hat{\mathbf{y}} - E(\hat{\mathbf{y}})][\hat{\mathbf{y}} - E(\hat{\mathbf{y}})]^T \middle| \mathbf{z}^*, \mathbf{S}\right)$$
$$= E\left([[\mathbf{Q}-\mathbf{Q}_o]\boldsymbol{\tau}_o + \mathbf{Q}_o[\boldsymbol{\tau}-\boldsymbol{\tau}_o]]\right.$$
$$\left.[[\mathbf{Q}-\mathbf{Q}_o]\boldsymbol{\tau}_o + \mathbf{Q}_o[\boldsymbol{\tau}-\boldsymbol{\tau}_o]]^T \middle| \mathbf{z}^*, \mathbf{S}\right)$$
$$= E\left([\mathbf{Q}-\mathbf{Q}_o]\boldsymbol{\tau}_o\boldsymbol{\tau}_o^T[\mathbf{Q}-\mathbf{Q}_o]^T \middle| \mathbf{z}^*, \mathbf{S}\right)$$
$$+ E\left(\mathbf{Q}_o[\boldsymbol{\tau}-\boldsymbol{\tau}_o]\boldsymbol{\tau}_o^T[\mathbf{Q}-\mathbf{Q}_o]^T \middle| \mathbf{z}^*, \mathbf{S}\right)$$
$$+ E\left([\mathbf{Q}-\mathbf{Q}_o]\boldsymbol{\tau}_o[\boldsymbol{\tau}-\boldsymbol{\tau}_o]^T\mathbf{Q}_o^T \middle| \mathbf{z}^*, \mathbf{S}\right)$$
$$+ E\left(\mathbf{Q}_o[\boldsymbol{\tau}-\boldsymbol{\tau}_o][\boldsymbol{\tau}-\boldsymbol{\tau}_o]^T\mathbf{Q}_o^T \middle| \mathbf{z}^*, \mathbf{S}\right)$$
$$= E\left([\mathbf{Q}-\mathbf{Q}_o]\boldsymbol{\tau}_o\boldsymbol{\tau}_o^T[\mathbf{Q}-\mathbf{Q}_o]^T \middle| \mathbf{z}^*, \mathbf{S}\right) \qquad \text{Equation 3-5}$$
$$+ E\left(\mathbf{Q}_o[\boldsymbol{\tau}-\boldsymbol{\tau}_o]\boldsymbol{\tau}_o^T[\mathbf{Q}-\mathbf{Q}_o]^T \middle| \mathbf{z}^*, \mathbf{S}\right)$$
$$+ E\left([\mathbf{Q}-\mathbf{Q}_o]\boldsymbol{\tau}_o[\boldsymbol{\tau}-\boldsymbol{\tau}_o]^T\mathbf{Q}_o^T \middle| \mathbf{z}^*, \mathbf{S}\right)$$
$$+ \mathbf{Q}_o E\left([\boldsymbol{\tau}-\boldsymbol{\tau}_o][\boldsymbol{\tau}-\boldsymbol{\tau}_o]^T \middle| \mathbf{z}^*, \mathbf{S}\right)\mathbf{Q}_o^T$$

Each of the terms in the last line of Equation 3-5 will now be treated. Since we are looking at a single output, $\mathbf{Q}$ is a row vector and $\mathbf{Q}\boldsymbol{\tau}$ is a scalar so the first term can be re-arranged to:

$$E\left([\mathbf{Q}-\mathbf{Q}_o]\boldsymbol{\tau}_o\boldsymbol{\tau}_o^T[\mathbf{Q}-\mathbf{Q}_o]^T \middle| \mathbf{z}^*, \mathbf{S}\right)$$
$$= E\left([[\mathbf{Q}-\mathbf{Q}_o]\boldsymbol{\tau}_o]^T\left[\boldsymbol{\tau}_o^T[\mathbf{Q}-\mathbf{Q}_o]^T\right]^T \middle| \mathbf{z}^*, \mathbf{S}\right)$$
$$= E\left(\boldsymbol{\tau}_o^T[\mathbf{Q}-\mathbf{Q}_o]^T[\mathbf{Q}-\mathbf{Q}_o]\boldsymbol{\tau}_o \middle| \mathbf{z}^*, \mathbf{S}\right)$$
$$= \boldsymbol{\tau}_o^T E\left([\mathbf{Q}-\mathbf{Q}_o]^T[\mathbf{Q}-\mathbf{Q}_o] \middle| \mathbf{z}^*, \mathbf{S}\right)\boldsymbol{\tau}_o \qquad \text{Equation 3-6}$$

since a scalar is equal to its transpose.

Likewise for the second and third term:

$$E\left(\mathbf{Q}_o[\boldsymbol{\tau}-\boldsymbol{\tau}_o]\boldsymbol{\tau}_o^T[\mathbf{Q}-\mathbf{Q}_o]^T \middle| \mathbf{z}^*, \mathbf{S}\right) = \mathbf{Q}_o E\left([\boldsymbol{\tau}-\boldsymbol{\tau}_o][\mathbf{Q}-\mathbf{Q}_o] \middle| \mathbf{z}^*, \mathbf{S}\right)\boldsymbol{\tau}_o \qquad \text{Equation 3-7}$$

$$E\left(\left[\mathbf{Q}-\mathbf{Q}_o\right]\boldsymbol{\tau}_o\left[\boldsymbol{\tau}-\boldsymbol{\tau}_o\right]^T\mathbf{Q}_o^T\middle|\mathbf{z}^*,\mathbf{S}\right) \qquad \text{Equation 3-8}$$
$$=\boldsymbol{\tau}_o^T E\left(\left[\mathbf{Q}-\mathbf{Q}_o\right]^T\left[\boldsymbol{\tau}-\boldsymbol{\tau}_o\right]^T\middle|\mathbf{z}^*,\mathbf{S}\right)\mathbf{Q}_o^T$$

These two terms are transformations of the covariance matrix between **Q** and the scores. This covariance is zero as long as the data used to calculate **Q** and the scores are independent. This will be the case except when operating on the data used to calculate the regression model. Substituting Equation 3-6, Equation 3-7 and Equation 3-8 into Equation 3-5 assuming **Q** and the scores are independent:

$$E\left(\left[\hat{\mathbf{y}}-E(\hat{\mathbf{y}})\right]\left[\hat{\mathbf{y}}-E(\hat{\mathbf{y}})\right]^T\middle|\mathbf{z}^*,\mathbf{S}\right)\approx\boldsymbol{\tau}_o^T E\left(\left[\mathbf{Q}-\mathbf{Q}_o\right]^T\left[\mathbf{Q}-\mathbf{Q}_o\right]\right)\boldsymbol{\tau}_o \qquad \text{Equation 3-9}$$
$$+\mathbf{Q}_o E\left(\left[\boldsymbol{\tau}-\boldsymbol{\tau}_o\right]\left[\boldsymbol{\tau}-\boldsymbol{\tau}_o\right]^T\middle|\mathbf{z}^*,\mathbf{S}\right)\mathbf{Q}_o^T$$

The first term in the variance expression arises from the estimation of **Q** in the modelling stage and the second term from the estimation of **P** and **W** in the modelling stage and uncertainty due to missing measurements in the new observations.

As shown in section 1.4.1, there are many expressions for the prediction error for PLS. In the balance of this chapter we will assume the uncertainty in the prediction is due to measurement error in y from the modelling stage and the uncertainty due to missing measurements in **z**. Measurement error in the independent variables in **X** is not treated but this is not a restriction imposed by the development of Equation 3-9. The variance of the **Q** matrix in the first term of Equation 3-9 under this assumption is shown in Equation 1-6. The variance of the scores with missing measurements was developed in Equation 3-2. Thus

$$E\left(\left[\hat{\mathbf{y}}-E(\hat{\mathbf{y}})\right]\left[\hat{\mathbf{y}}-E(\hat{\mathbf{y}})\right]^T\middle|\mathbf{z}^*,\mathbf{S}\right)\approx \boldsymbol{\tau}_o^T\left[\mathbf{T}^T\mathbf{T}\right]^1\boldsymbol{\tau}_o\sigma^2 + \mathbf{Q}_o\mathbf{R}^{\#T}\left[\mathbf{S}^{\#\#}-\mathbf{S}^{\#*}\mathbf{S}^{**+}\mathbf{S}^{*\#}\right]\mathbf{R}^{\#}\mathbf{Q}_o^T \qquad \text{Equation 3-10}$$

The first term of Equation 3-10 depends on the value of the present measurements $\mathbf{z}^*$ and the second on which measurements are missing. The prediction error variance and the contribution of missing measurement uncertainty relative to estimation error in that variance will therefore change from object to object.

In order to attribute the variance in a prediction $\hat{y}$ to individual measurements z_k, the contributions to the prediction must be examined. The mean and variance of these contributions for the missing measurements are developed in a similar manner to the contributions to the scores in Equation 3-3 and Equation 3-4.

$$E\left(contribution(z_k\ to\ \hat{y})\right)=E\left(\sum_{j=1}^{A}q_{1j}r_{kj}z_k\middle|\mathbf{z}^*,\mathbf{S}\right) \qquad \text{Equation 3-11}$$
$$=E\left(z_k\middle|\mathbf{z}^*,\mathbf{S}\right)\sum_{j=1}^{A}q_{1j}r_{kj}$$

$$\mathrm{var}\left(contribution(z_k\ to\ \hat{y})\right)=\mathrm{var}\left(\sum_{j=1}^{A}q_{1j}r_{kj}z_k\middle|\mathbf{z}^*,\mathbf{S}\right) \qquad \text{Equation 3-12}$$
$$=\mathrm{var}\left(z_k\middle|\mathbf{z}^*,\mathbf{S}\right)\left[\sum_{j=1}^{A}q_{1j}r_{kj}\right]^2$$

3.3.2 Example: Kamyr Digester

The process and overall data set for this example are the same as used in section 2.5.2. The first 200 objects of the data set were used to build a PLS model of the process and 100 objects from the rest of the data set were selected as a test set. The test data set was distinct from the modelling data set in order to meet the conditions for using Equation 3-9.

The measurements that are selected to be missing are in variables 1, 7, 8 and 17. These measurements were chosen to illustrate both typical behaviour and potential problems. The score error with measurements of variables 1 and 8 missing together was analysed in section 2.5.2.2 for the SCP and CMR score calculation algorithms and was shown to cause particular difficulty for the SCP algorithm.

The pattern of missing measurements in the test set objects has been graphically represented in Figure 3-3 with each missing measurement in an object represented by a point on the graph. An operator shift change in the plant caused measurements 1 and 8 to be missing every 12 hours, one hour apart in the original data set and this is reproduced in the complete objects used in this example. In addition in this example, measurement 8 is missing in objects 1 through 25, measurement 7 in objects 55 to 85 and measurement 17 in objects 20 to 40 representing sensor breakdowns or maintenance.

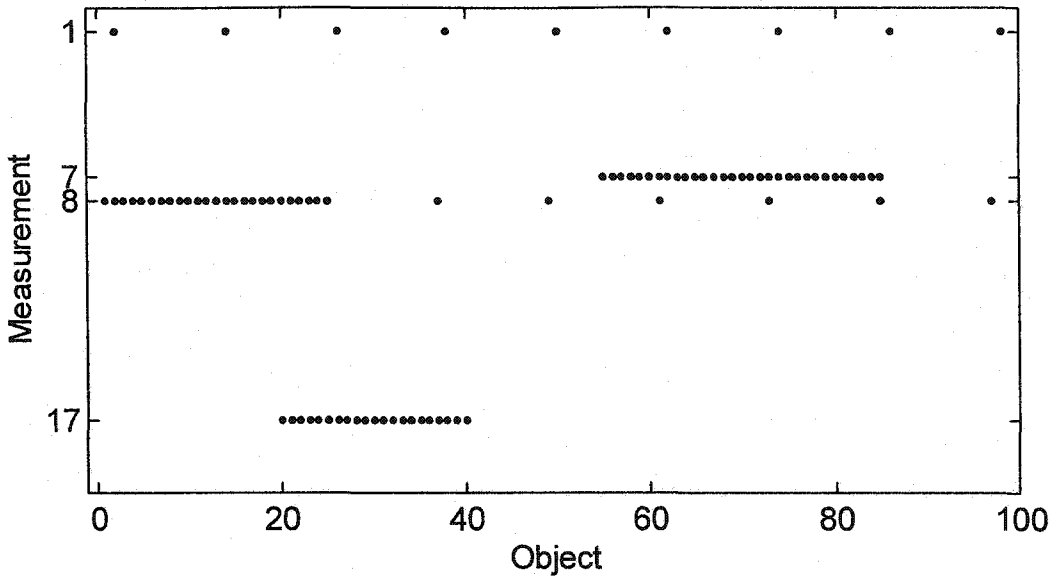


Figure 3-3: Plot with Dots Indicating Measurements Designated Missing in the Kamyr Digester Data Set.

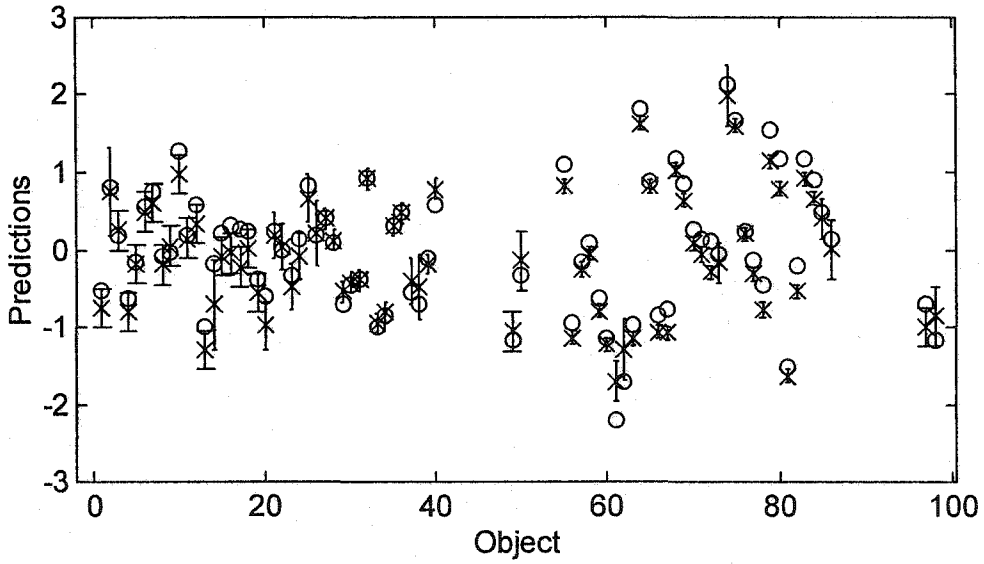


Figure 3-4: Prediction plot with Various Measurements Missing. Bars are 95% uncertainty interval, 'o' is the prediction with all the measurements, 'x' is the CMR prediction.

The full data predictions, predictions from CMR and an uncertainty interval based on Equation 3-9 are plotted in Figure 3-4. Only those objects with missing measurements are plotted as they are the only ones of interest. Comparing the full data predictions to the intervals will allow us to determine how well the assumptions and approximations made in developing the interval apply to this industrial data set.

Before the comparison between full data predictions and the uncertainty intervals is done we should eliminate the objects which have an SPE which is greater than the critical value since that indicates that the data is not following the same distribution as the training data. In practice, the SPE would be checked before the prediction is calculated. We will see in Figure 3-8 in section 3.4.3.1 that the SPE is outside its control limit at object 17 and around objects 60 and 80 and just below the limit at object 10 at a 99% confidence level. Therefore, any inferences such as our uncertainty intervals made using the modelling data distribution and these objects are suspect. Comparing only the objects with SPE's below the control limit, there is a good match between the uncertainty interval, centred on the CMR prediction, and the full data prediction. The full data predictions and uncertainty intervals do not match well for the objects with SPE's above the control limit.

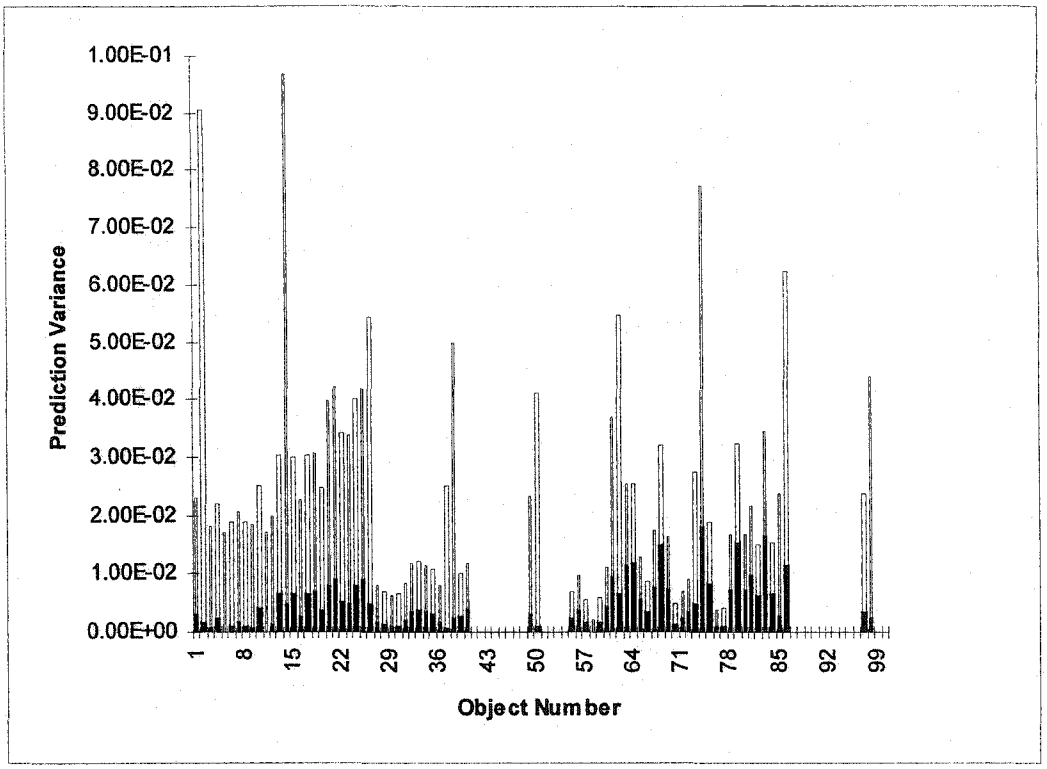


Figure 3-5: Stacked Bar Plot of Prediction Error Variance due to Parameter Variance and Approximate Combined Parameter Variance and Missing Measurement Uncertainty. Filled Bars are parameter variance and unfilled bars combined variance.

The difference between the uncertainty intervals from parameter variance only (filled bars) and the combined effect of missing data and parameter variance (unfilled bars) is shown in Figure 3-5. The parameter only variance is calculated using Equation 1-6 and once again the results for objects that have high SPE values in Figure 3-8 should be viewed with suspicion. Note that the parameter only variance is in general a small fraction of the combined variance so that the current practice of using this variance when there are missing measurements can

be very misleading. The relative contribution of missing measurement uncertainty and variance due to parameter estimation varies from object to object as expected.

The predictions from using alternative missing data replacement approaches, namely SCP and mean replacement, to calculate the scores are plotted in Figure 3-6 and Figure 3-7 respectively. The uncertainty intervals from Equation 3-9 are also plotted with the centre of the interval being the CMR estimate. The SCP and mean replacement predictions can be outside the calculated uncertainty intervals since the assumptions used in calculating the interval are not the same as those used to calculate the SCP and mean replacement scores. SCP is outside the interval more often than mean replacement. As shown in 2.5.2.1, certain missing measurement combinations can lead to calculated scores that are inconsistent with those calculated from complete data. On the whole, the predictions from both the SCP and mean replacement methods on those objects before number 60 are comparable to each other and to the CMR method.

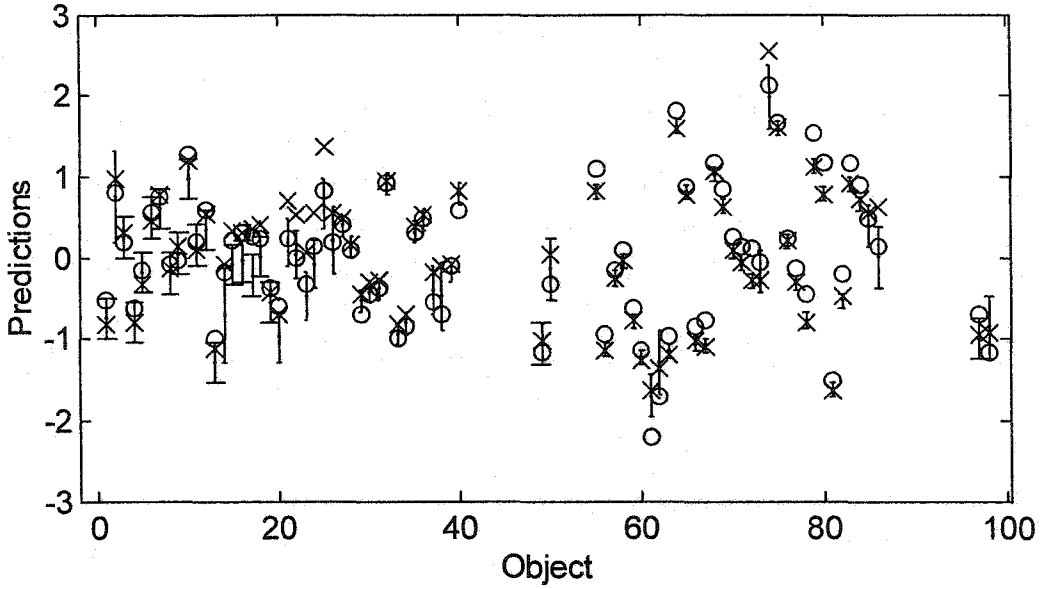


Figure 3-6: Prediction plot with Various Measurements Missing. Bars are 95% uncertainty interval, 'o' is the prediction with all the measurements, 'x' is the SCP prediction.

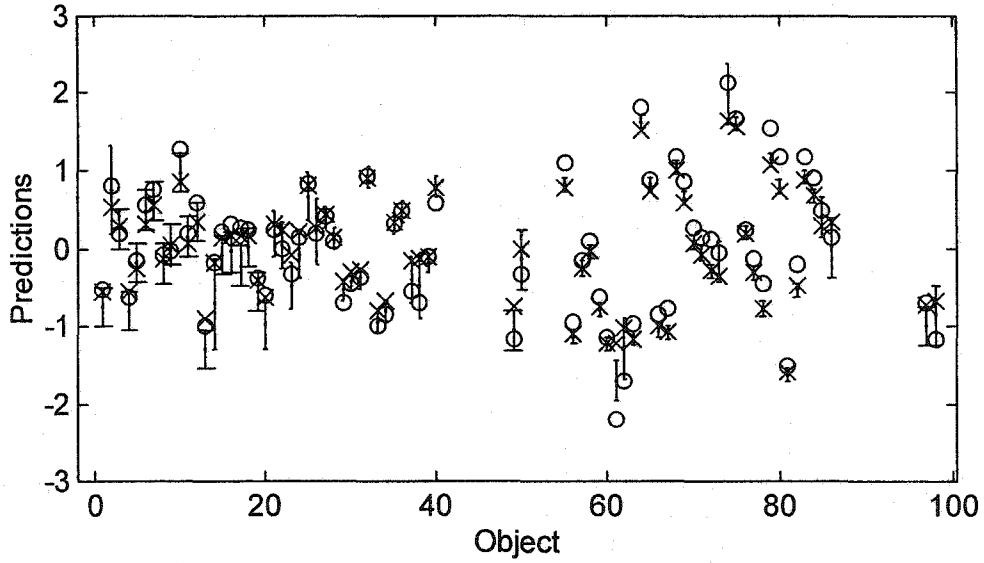


Figure 3-7: Prediction plot with various measurements missing. Bars are 95% uncertainty interval, 'o' is the prediction with all the measurements, 'x' is the mean replacement prediction.

3.4 Monitoring

3.4.1 Residuals

3.4.1.1 Uncertainty Region for the SPE

The uncertainty interval for the SPE due to missing measurements is the diagnostic that will be developed in this section. This will show the location of possible SPE values and the amount of uncertainty in the values arising from missing measurements. The uncertainty interval due to missing measurements for the SPE developed here is useful because it indicates the range of possible values of the SPE arising just from the missing measurements.

The process is assumed to be 'in control' in the development of the uncertainty interval. This is necessary because the distribution of the training data is used to calculate the uncertainty due to missing measurements. If the new objects are not drawn from this distribution then the uncertainty interval is invalid.

Care must be taken when using the uncertainty interval due to missing measurements to make conclusions about the validity of alarms. If the uncertainty interval due to missing measurements is clearly outside the alarm limit, an alarm is confirmed. An uncertainty interval due to missing measurements inside the alarm limit increases confidence but does not guarantee that an alarm condition

does not exist since the missing measurements may be the ones that would cause the SPE to be over the critical value.

To conduct this analysis, we must put the SPE with missing measurements in the standard quadratic form in Equation 1-33. Without considering missing measurements we have:

$$\begin{aligned} SPE &= [\mathbf{z}-\mathbf{P}\tau]^T[\mathbf{z}-\mathbf{P}\tau] \\ &= [\mathbf{z}-\mathbf{PRz}]^T[\mathbf{z}-\mathbf{PRz}] \\ &= [[\mathbf{I}-\mathbf{PR}]\mathbf{z}]^T[[\mathbf{I}-\mathbf{PR}]\mathbf{z}] \\ &= \mathbf{z}^T[\mathbf{I}-\mathbf{PR}]^T[\mathbf{I}-\mathbf{PR}]\mathbf{z} \end{aligned}$$

So the general quadratic form $\mathbf{x}\mathbf{A}^T\mathbf{x}$ has $\mathbf{x}=\mathbf{z}$, which is a sample from a distribution, and the weighting matrix $\mathbf{A}=[\mathbf{I}-\mathbf{PR}]^T[\mathbf{I}-\mathbf{PR}]$. When we do not have values for all of the measurements, the missing measurements do not have fixed values. We will characterise the missing measurements by their conditional distribution given the present measurements and the distribution from the training data. Since we are fixing the present measurements $\mathbf{z}^*$, the variables corresponding to $\mathbf{z}^*$ in $\mathbf{x}$ will be fixed (zero variance) and the variables corresponding to $\mathbf{z}^\#$ will be normally distributed. Thus

$$\mathbf{x} \sim N\left(\begin{bmatrix} \mathbf{S}^{\#*}\mathbf{S}^{**+}\mathbf{z}^* \\ \mathbf{z}^* \end{bmatrix}, \begin{bmatrix} \mathbf{S}^{\#\#}-\mathbf{S}^{\#*}\mathbf{S}^{**+}\mathbf{S}^{*\#} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}\right) \qquad \text{Equation 3-13}$$

$$\mathbf{A}=[\mathbf{I}-\mathbf{PR}]^T[\mathbf{I}-\mathbf{PR}] \qquad \text{Equation 3-14}$$

The variance matrix for $\mathbf{x}$ is not full rank and the weighting matrix $\mathbf{A}$ will be rank deficient for PCA and may be rank deficient for PLS. As mentioned in section 1.7, the approximations for the cumulative distribution of this statistic

must be transformed to the full rank equivalent as part of the transformation to the standard form Equation 1-33 which is shown in appendix 1. The approximation to the cumulative distribution of the quadratic form in Jensen and Solomon (1972) is used to calculate critical values for the SPE in this work.

3.4.1.2 Uncertainty Interval for Contributions to the SPE

The contributions to the SPE are the individual residuals. The uncertainty intervals for the residuals due to missing measurements can be examined to determine which missing measurements are causing the greatest uncertainty and hence will give the most benefit if the missing values are acquired. These residuals are normally distributed with mean and variance given by:

$$E\left(resid(\mathbf{z})\middle|\mathbf{z}^*,\mathbf{S}\right)=E\left(\mathbf{z}-\mathbf{P}\hat{\tau}\middle|\mathbf{z}^*,\mathbf{S}\right) \qquad \text{Equation 3-15}$$
$$=E\left(\left[\mathbf{I}-\mathbf{PR}\right]\mathbf{z}\middle|\mathbf{z}^*,\mathbf{S}\right)$$
$$=\left[\mathbf{I}-\mathbf{PR}\right]E\left(\mathbf{z}\middle|\mathbf{z}^*,\mathbf{S}\right)$$

$$\mathrm{var}\left(resid(\mathbf{z})\middle|\mathbf{z}^*,\mathbf{S}\right)=E\Big(\left[\mathbf{I}-\mathbf{PR}\right]\left[\mathbf{z}-E\left(\mathbf{z}\middle|\mathbf{z}^*,\mathbf{S}\right)\right]\left[\mathbf{z}-E\left(\mathbf{z}\middle|\mathbf{z}^*,\mathbf{S}\right)\right]^T \qquad \text{Equation 3-16}$$
$$\left[\mathbf{I}-\mathbf{PR}\right]^T\middle|\mathbf{z}^*,\mathbf{S}\Big)$$
$$=\left[\mathbf{I}-\mathbf{PR}\right]E\Big(\left[\mathbf{z}-E\left(\mathbf{z}\middle|\mathbf{z}^*,\mathbf{S}\right)\right]\left[\mathbf{z}-E\left(\mathbf{z}\middle|\mathbf{z}^*,\mathbf{S}\right)\right]^T$$
$$\middle|\mathbf{z}^*,\mathbf{S}\Big)\left[\mathbf{I}-\mathbf{PR}\right]^T$$
$$=\left[\mathbf{I}-\mathbf{PR}\right]\mathrm{var}\left(\mathbf{z}\middle|\mathbf{z}^*,\mathbf{S}\right)\left[\mathbf{I}-\mathbf{PR}\right]^T$$

There are two factors that complicate using the uncertainty in the contributions to the SPE to choose the best missing measurement to recover; they are correlation between missing measurements and uncertainty in the residuals of the present measurements.

If the missing measurements are correlated then recovering one measurement may reduce the uncertainty in the residual of another missing measurement. Take the case where two temperature measurements (A and B) in close physical proximity are missing as well as a flow measurement which does not enter strongly into the same loading vectors as the two temperatures. The two temperatures are correlated with each other but not with the flow measurement. The flow measurement has the largest uncertainty interval on its contribution to the SPE followed by temperature A and then temperature B. Temperature A may be the best measurement to recover to reduce the uncertainty in the SPE due to missing measurements since having a measurement for it will reduce the uncertainty in the contribution of temperature B as well as reducing the uncertainty in its own contribution.

Measurements in $\mathbf{z}^*$ which have values can have uncertainty associated with their residuals. This occurs because the residual is a function of the scores and the scores can have uncertainty arising from the missing measurements associated with them. If the uncertainty in the residuals for present measurements is small compared to those for missing measurements then we can directly infer which missing measurements are contributing most to the uncertainty in the SPE. If the uncertainty in the residuals for the present measurements is large and comparable to that for the missing measurements, the uncertainty in the scores cannot be neglected.

3.4.2 Scores

3.4.2.1 Uncertainty Interval for the Hotelling T^2

The uncertainty interval for the Hotelling T^2 with missing measurements is developed in a similar way to the SPE in section 3.4.1.1. We can see from Equation 1-11 that the Hotelling T^2 is a quadratic form $\mathbf{x}\mathbf{A}^T\mathbf{x}$ with $\mathbf{x}=\boldsymbol{\tau}$ and $\mathbf{A}=\left[\mathbf{T}^T\mathbf{T}\right]^{-1}$. With the present variables fixed and the missing measurements unknown, the distribution of the random variable is the conditional distribution of the score given the present measurements developed in section 3.2.1.

Again in approximating the cumulative distribution, one must beware a rank deficient distribution for the random variable ($\boldsymbol{\tau}$) but in this case the weighting matrix A is guaranteed to be full rank. The approximation to the cumulative distribution of the quadratic form in Jensen and Solomon (1972) is again used to calculate critical values.

As with the SPE, an interval entirely outside the alarm limit provides confidence that an alarm is present but an interval inside the alarm limit does not assure that an alarm condition is not present. The process must also be assumed to be 'in control' for the uncertainty interval due to missing measurements to be valid.

If there are only two latent variables in the model, the Hotelling T^2 may be replaced by the score-score plot and the uncertainty region for the scores can be

shown directly on this plot based on the conditional distribution of the scores developed in section 3.2.1.

3.4.2.2 Uncertainty Intervals for the Contributions to the Hotelling T^2

The total effect of a given score on the Hotelling T^2 is summed up by the score's individual missing measurement uncertainty interval. This occurs because **A** is diagonal due to the modelling data set scores being independently distributed. The uncertainty in the scores due to missing measurements is determined as detailed in section 3.2.1 and individual scores are identified for further investigation by examination of their contributions to the uncertainty in the Hotelling T^2. The uncertainty in the contributions to the identified scores due to missing measurements developed in section 3.2.2 will lead to the identification of the missing measurements that cause the most uncertainty. Ranking the missing measurements that cause uncertainty further depends on the correlation between the missing measurements and the relative magnitudes of the scores in which the missing measurements cause uncertainty which determines the weighting of the score in the Hotelling T^2.

3.4.3 Example: Kamyr Digester

In this section the uncertainty intervals for the SPE and Hotelling T^2 developed in the previous sections will be applied to an industrial data set and plotted along with the values calculated using all of the measurements (full data values). This will show that the assumptions that were made to develop the

intervals are realistic in application. The data set, model and pattern of missing data used in this section are the same as in section 3.3.2. The missing data pattern is depicted graphically in Figure 3-3.

Contributions to the uncertainty intervals will be calculated and used to illustrate how to determine which missing measurements will yield the largest reduction in uncertainty when they are recovered. The actual uncertainty intervals after the variables are recovered are calculated to verify that the correct results were obtained.

Finally, plots are made of the SPE and Hotelling T^2 calculated using scores from SCP and mean replacement along with the uncertainty intervals due to missing measurements. This allows the work that has been done here to be compared with current practice. The additional information the uncertainty intervals due to missing measurements contribute and the accuracy of the intervals compared to the point values calculated using the currently applied methods are shown.

3.4.3.1 SPE Monitoring

The full data SPE and the uncertainty intervals arising from the missing measurements are plotted in Figure 3-8. All objects are plotted since the trend in the SPE is also of interest in determining whether an alarm condition exists. Objects with no missing measurements have only a full data SPE value plotted.

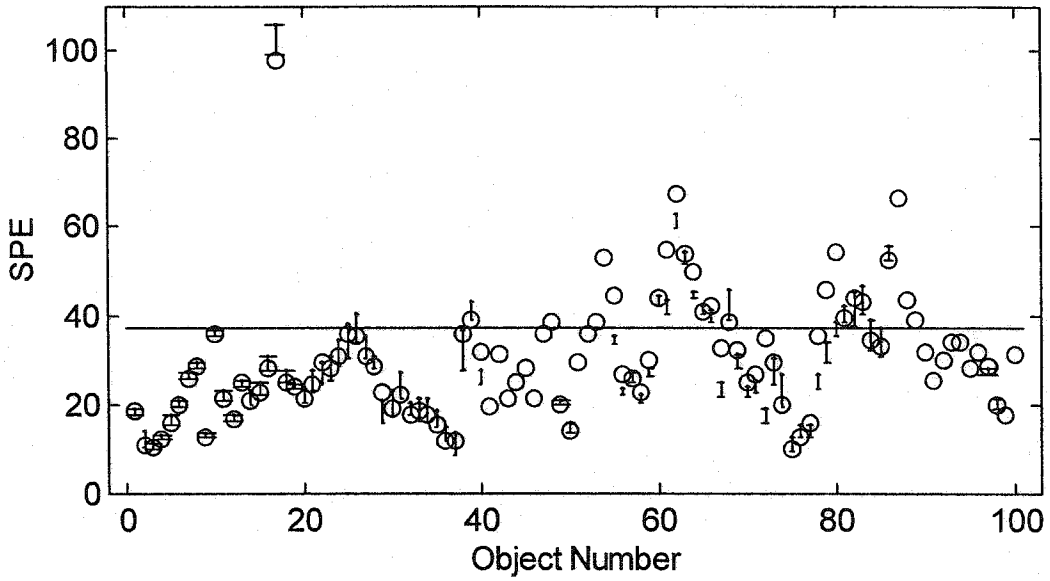


Figure 3-8: SPE Plot with Various Missing Measurements. The bars on some objects are 95% uncertainty intervals for the SPE arising from missing measurements and the 'o' is the full data value of the SPE. The solid horizontal line is the 99% control limit.

The size of the uncertainty interval does not change in a regular way with the magnitude of the SPE so the information provided by the interval is important. For objects 1 to 40 the agreement between the interval and the full data SPE is very good, even for object 17 which is well above the critical value. Between objects 55 and 80, there are many instances where the full data SPE is outside the interval but the SPE is above the critical value for much of this period. This is most likely due to a sustained disturbance that changed the data distribution and invalidated the assumptions on which the interval is based. The interval does not indicate an alarm condition at objects 55, 79 or 80 but does show the alarms at objects 60-66, 68, 81-83 and 86.

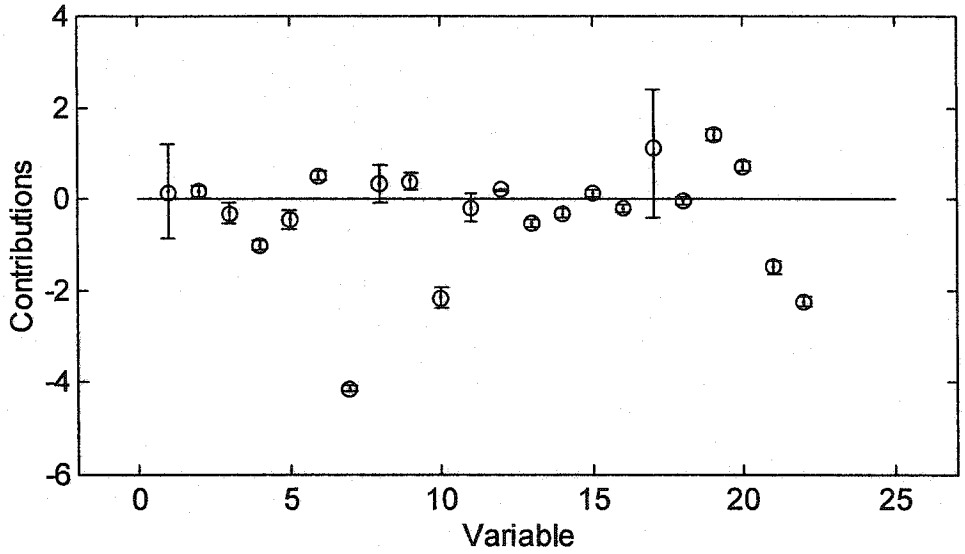


Figure 3-9: SPE Contribution Plot for object 26 with uncertainty intervals due to missing measurements. Missing Measurements are 1 and 17.

Object 26 in Figure 3-8 has an uncertainty interval that straddles the critical value so we would like to clarify whether an alarm condition exists. This requires a diagnosis of which of the missing measurements are responsible for the uncertainty and obtaining values for them. Figure 3-9 shows that the uncertainty in the contributions for the measured variables are negligible compared to the uncertainty in the unmeasured variables. Thus we do not have to take into account the uncertainty propagating through the score calculation to the residuals of the present measurements and need only look at uncertainty arising directly from the missing measurements. The covariance between the unmeasured variables does not have to be taken into account with only two missing variables so it can be concluded that variable 17 is causing more of the uncertainty in the SPE than variable 1. This is shown to be true in Table 3-1 where the length of the SPE

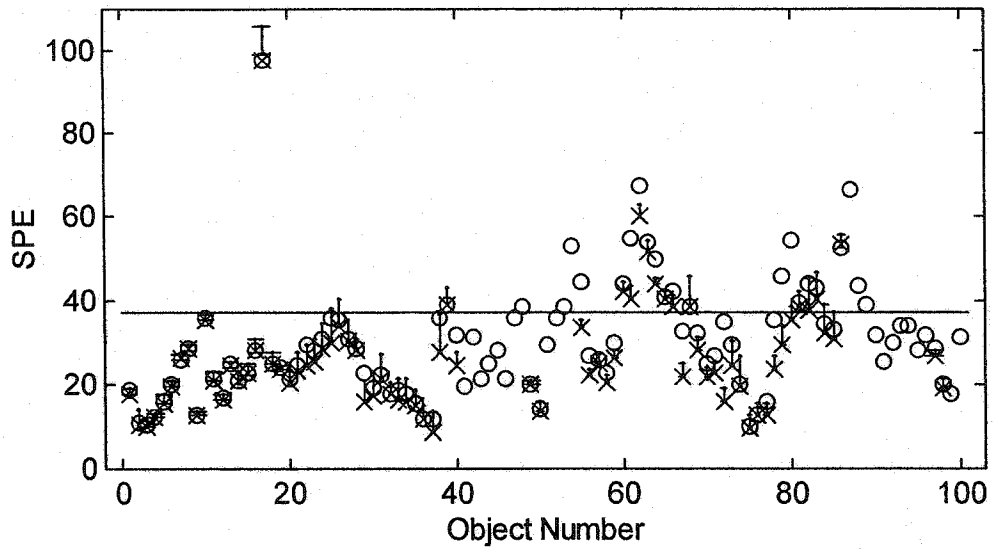


Figure 3-10: SPE Plot with Various Missing Measurements. The bars on some objects are 95% uncertainty intervals for the SPE arising from missing measurements, the 'x' is the SCP value and the 'o' is the full data value of the SPE. The solid horizontal line is the 99% control limit.

Variables Missing	Recovered Measurement	Interval Lower Limit	Interval Upper Limit	Length of Interval
1 and 17	-	33.86	40.49	6.63
1	17	35.04	37.70	2.66
17	1	33.83	39.62	5.79

Table 3-1: 95% Uncertainty intervals for the SPE arising from missing measurements for Object 26.

uncertainty interval with variable 1 missing and 17 recovered is smaller than the interval with variable 17 missing and 1 recovered.

In Figure 3-10, the SCP calculated values of the SPE are plotted along with the intervals and full data SPE's from Figure 3-8. Note that SCP value does not necessarily fall within the missing measurement uncertainty interval since it does not use the assumption about the data distribution that is used to calculate the interval. The SCP value tends to underestimate the SPE and fall in the lower part

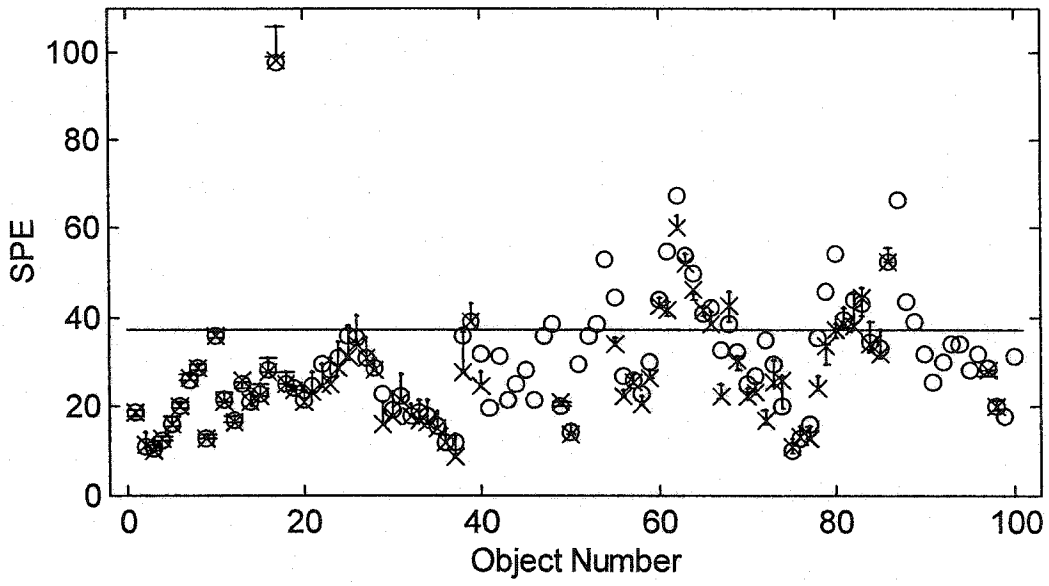


Figure 3-11: SPE Plot with Various Missing Measurements. The bars on some objects are 95% uncertainty intervals for the SPE arising from missing measurements, the 'x' is the mean replacement value and the 'o' is the full data value of the SPE. The solid horizontal line is the 99% control limit.

of the uncertainty interval. This could be due to the zero residuals assigned to the missing measurements, the scores incorrectly using variance that should go to the residuals or variance going to the wrong score.

Figure 3-11 has the full data SPE and uncertainty intervals of Figure 3-8 with the mean replacement values of the SPE plotted. These SPE values do not necessarily fall within the missing measurement uncertainty interval for the SPE for the same reason as the SCP values. The mean replacement value also tends to underestimate the SPE and falls in the lower part of the uncertainty interval.

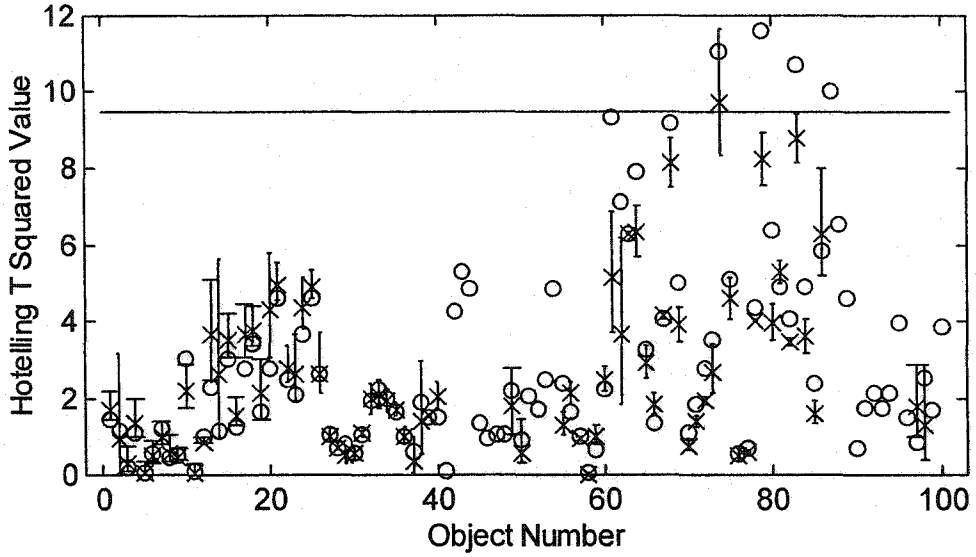


Figure 3-12: Hotelling T^2 Plot with Various Measurements Missing. The bars on some objects are 95% uncertainty intervals arising from missing measurements. The symbol 'o' is the full data value and 'x' is the CMR value. The solid horizontal line is the 99% control limit.

3.4.3.2 Score Monitoring

In this section, objects have been included in the plots that have SPE values that are greater than the critical value. In a monitoring application, analysis would normally stop with the SPE since the PCA or PLS model's validity is questionable once the variance matrix has shifted significantly. These objects have been included to test the robustness of the intervals and missing data score calculation methods to the distribution of the data.

The Hotelling T^2 value from the CMR scores, the uncertainty interval limits due to missing measurements calculated for the Hotelling T^2 and the full data Hotelling T^2 value are plotted in Figure 3-12. In the objects before 50, the intervals are small enough and the SPE values in Figure 3-8 low enough that there

is no doubt that the process is in control according to the Hotelling T^2 statistic. When the full data values are examined this is confirmed. Note that like the SPE, there are many instances where the full data value lies outside the uncertainty interval for objects 55-80 but it was seen in Figure 3-8 that the process is disturbed in this time period so the assumptions on which the model and uncertainty regions are based are likely to be invalid. The objects with a large distance between the interval and the full data value (61, 79, 80 and 83) in particular have SPE's above the critical value.

Object 74, with measurements 1 and 7 missing, has a CMR Hotelling T^2 value slightly above the critical value but an uncertainty interval that straddles the critical value. We would like to know which of the missing measurements are most responsible for the uncertainty. The SPE for this object is below the critical value so it is valid to examine the Hotelling T^2 value. The CMR score value and a 95% uncertainty region derived from the variance matrix in Equation 3-2 are plotted in Figure 3-13 along with the 99% control limit for the scores. The uncertainty region for the scores is not aligned with either axis so the uncertainty is present in both scores and we must examine the contributions to both. Both of the score contribution plots in Figure 3-14 and Figure 3-15 clearly show that variable 1 is mainly responsible for the uncertainty in the scores and this is confirmed by Table 3-2.

The Hotelling T^2 values calculated from scores calculated by SCP and mean replacement are plotted in Figure 3-16 and Figure 3-17. The Hotelling T^2

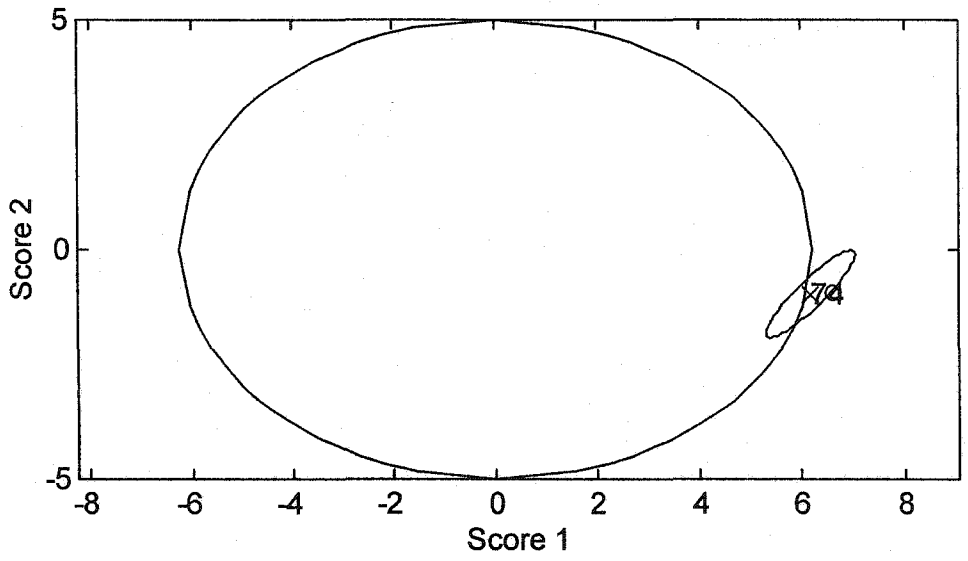


Figure 3-13: Score-Score plot for object 74 with uncertainty region arising from missing measurements for the unknown score. The missing measurements are 1 and 7.

values from SCP and mean replacement do not consistently over- or underestimate the full data values of the statistic. The values tend to fall within the uncertainty interval so the objects for which the methods perform poorly are also the ones where the assumption of a normal distribution following the model data set statistics is not a good one. On the whole, the conclusions about the CMR and uncertainty interval robustness apply to the SPE values calculated from SCP and mean replacement scores. There are two objects, 25 and 74, for which the SCP method produces values that are outside the uncertainty interval but this behaviour is consistent with that seen in section 2.5.2.1 where some objects or missing data combinations can have large errors depending on the alignment of the object and model vectors.

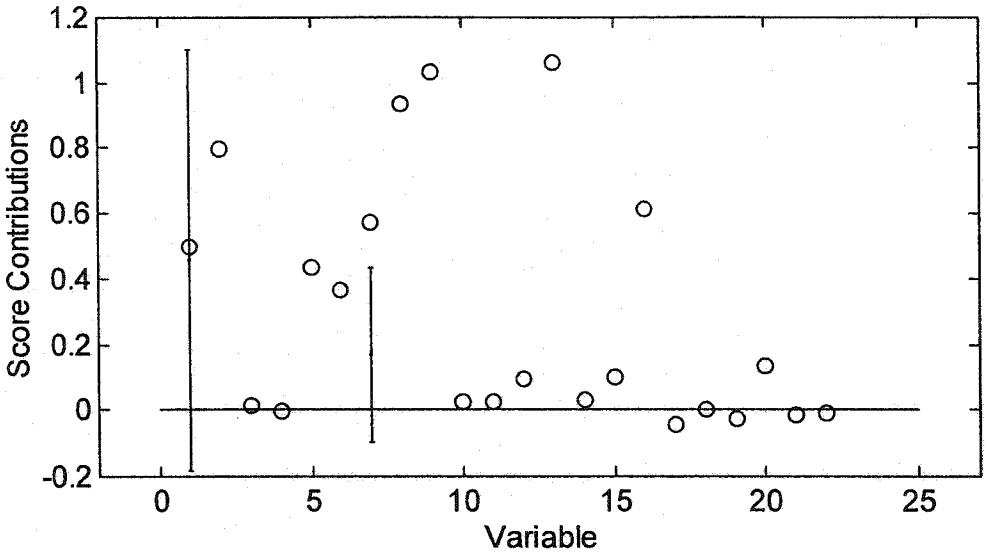


Figure 3-14: First Score Contribution plot for object 74 with 95% Uncertainty Intervals due to Missing Measurements. Missing Measurements are 1 and 7.

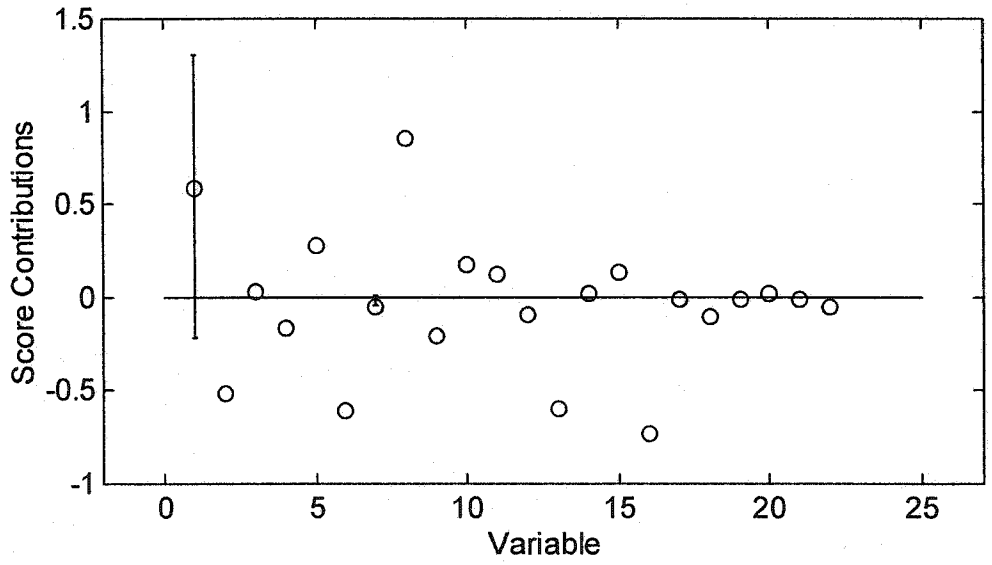


Figure 3-15: Second Score Contribution Plot for object 74 with 95% Uncertainty Intervals due to Missing Measurements. Missing Measurements are 1 and 7.

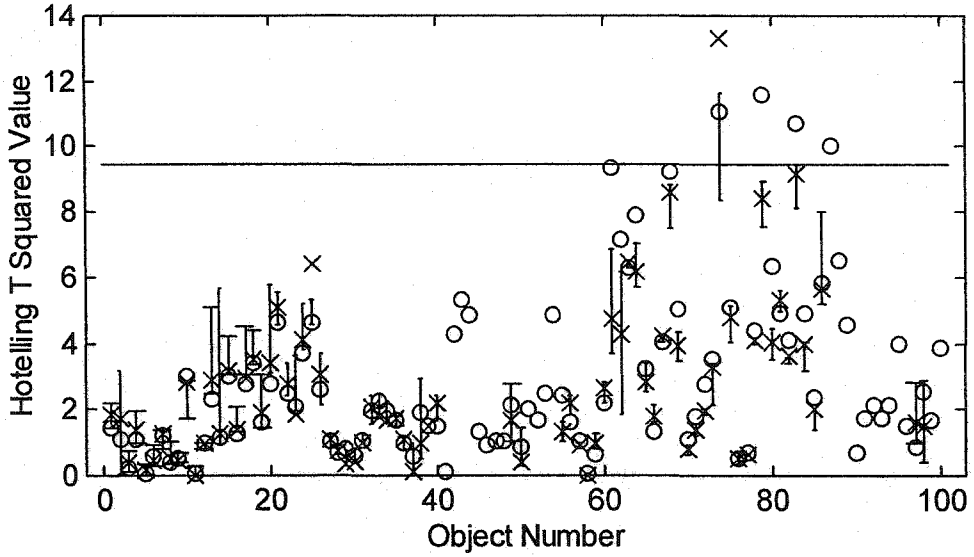


Figure 3-16: Hotelling T^2 Plot with Various Measurements Missing. The bars on some objects are 95% uncertainty intervals arising from the missing measurements. The symbol 'o' is the nominal value and 'x' is the SCP value. The solid horizontal line is the 99% control limit.

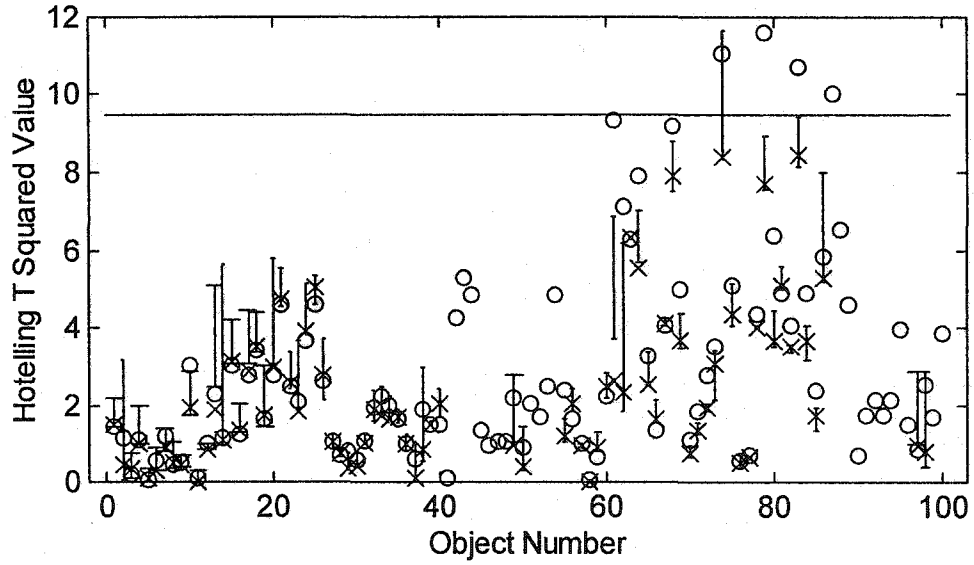


Figure 3-17: Hotelling T^2 Plot with Various Measurements Missing. The bars on some objects are 95% uncertainty intervals arising from the missing measurements. The symbol 'o' is the nominal value and 'x' is the mean replacement value. The solid horizontal line is the 99% control limit.

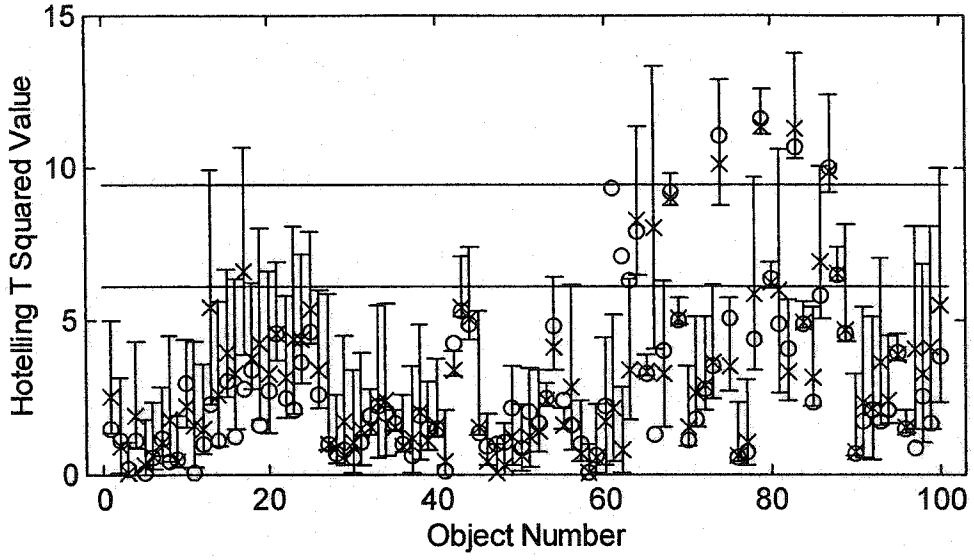


Figure 3-18: Hotelling T^2 Plot with Variables 1 and 8 Missing. The bars on the objects are 95% uncertainty intervals arising from the missing measurements. The symbol 'o' is the nominal value and 'x' is the CMR value. The horizontal lines are critical values of the Hotelling T^2 statistic; the 99% value is the top line and the 95% value is the lower line.

Variables Missing	Recovered Measurement	Interval Lower Limit	Interval Upper Limit	Length of Interval
1, 7	-	8.36	11.64	3.28
1	7	9.87	12.79	2.92
7	1	8.98	10.63	1.65

Table 3-2: Object 74 Hotelling T^2 95% uncertainty intervals caused by missing measurements showing the effect of recovering measurements.

The Hotelling T^2 has been plotted in Figure 3-18 when measurements 1 and 8 are missing. This particular combination of missing variables was previously identified as causing poor performance (section 2.5.2). The large uncertainty intervals on the Hotelling T^2 value due to missing measurements greatly reduce the utility and confidence in the test; especially if the 95% control limit level is used. The monitoring application would be ignored or turned off.

The length of the uncertainty intervals on the calculated T^2 values vary greatly over time. A simple off-line calculation of uncertainty would be of much less utility than the individual limits on the objects as plotted in the figure.

If variable 4 is missing in addition to 1 and 8 we see missed alarms in the SPE and false alarms in the Hotelling T^2 in another section of the kamyr digester data set shown in Figure 3-19 when the scores are calculated by SCP. The missed SPE alarms are in objects 45 to 47 and 94 to 95 and the false Hotelling T^2 alarms in objects 48, 49, 99 and 100. The false Hotelling T^2 alarm at object 95 occurs when the SPE is above the critical limit and so is not included. The uncertainty intervals for the false SPE alarms in objects 45, 94 and 95 straddle the critical value and so indicate the possibility of an alarm condition. For objects 46 and 47 the uncertainty interval stops just short of the critical value and shows that a missed alarm is also possible for the uncertainty interval. The Hotelling T^2 uncertainty intervals for the objects with false alarms show that there is substantial uncertainty in the Hotelling T^2 for those objects and would indicate to an operator that expending effort investigating those alarms would not be productive.

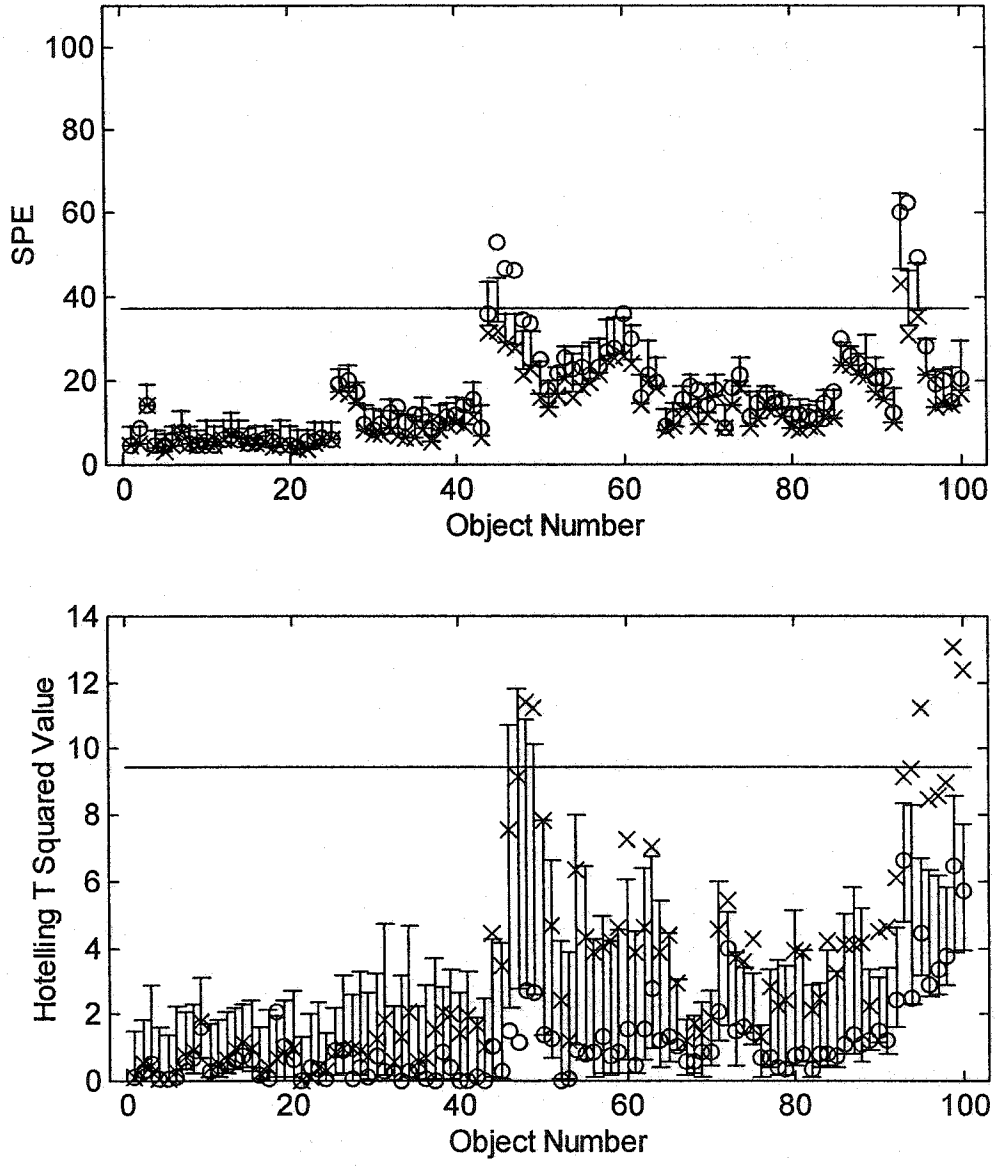


Figure 3-19: SPE and Hotelling T^2 plots for another section of the Kamyr Digester data set illustrating missed and false alarms due to missing measurements. The missing measurements are 1, 4 and 8. The bars on the objects are 95% uncertainty intervals arising from the missing measurements. The symbol 'o' is the nominal value and 'x' is the SCP value. The horizontal lines are the 99% control limits of the statistics.

3.5 Conclusions

The impact that the uncertainty arising from the missing measurements will have on the scores, predictions, SPE, Hotelling T^2 and contributions to these quantities in PCA and PLS applications is quantified in this chapter so appropriate action can be taken. The performance measures developed distinguish between situations where model performance with missing measurements will continue to be acceptable and situations where it will become unacceptable. In the latter case, the measurements must be recovered or the application shut down. Uncertainty intervals arising from missing measurements have been derived for the scores, predictions, SPE, Hotelling T^2 and contributions to these quantities.

A PLS application to an industrial data set has been used to show that these uncertainty intervals can be used as performance measures and diagnostics in monitoring and prediction when there are missing measurements. A specific combination of missing measurements was shown to cause uncertainty intervals in the Hotelling T^2 large enough to justify that the monitoring application be shut down. Another combination of missing measurements caused false Hotelling T^2 and missed SPE alarms with the single component projection method for some objects. These objects had uncertainty intervals that cast doubt on or indicated against the erroneous results.

The approximate prediction variance matrix considering a single output derived in this chapter has two terms: one involving the variance of the scores

conditional on the present measurements and the other the variance of the **Q** loading vector. The scores must be independent of **Q** for the approximation to hold. The development does not restrict the conditional distribution of the scores or PLS loading vector **Q** variance chosen. For this work the conditional variance of the scores was considered only to arise from missing measurements and the **Q** variance from measurement errors in the dependent variable y. The combined estimation and missing measurement prediction variance was shown to be much greater than the estimation variance alone for the kamyr digester industrial data. This shows that the current practice of ignoring the uncertainty introduced by the missing measurements can be unacceptable. The uncertainty regions produced for the kamyr digester data matched the full data predictions well when the SPE was less than the critical value, showing that the assumptions made in producing the uncertainty interval are reasonable in practice.

The uncertainty intervals for the SPE, Hotelling T^2 and the contributions to these statistics that have been derived here bring more information about the state of the process and our knowledge of that state. This increases the confidence in the application of PCA and PLS models when measurements are missing. The process must be 'in control' for the uncertainty intervals to be valid. Quantifying the uncertainty allows alarms to be acted on with confidence if the uncertainty interval is clearly above the critical value of the statistic. It also reveals when the uncertainty in the statistic is so large that alarm conditions can not be detected. An uncertainty interval entirely below the critical value indicates that either an alarm

condition is not present or that the missing measurements are key to detecting that alarm. The uncertainty intervals on the contributions give an indication of which measurements must be recovered to reduce the uncertainty in the test statistic so decisions can be made which balance the cost of recovering a measurement with the benefit to the application. The effect of correlation between missing measurements is not considered when determining the effect of individual missing measurements on the magnitude of the uncertainty interval. This may result in a sub-optimal choice being made for which variables to recover.

The application of the derived uncertainty intervals to an industrial data set showed that the assumptions on the data distribution and the way that the uncertainty intervals were defined were reasonable in practice. The uncertainty intervals were consistent with the values calculated from all of the measurements (full data) when the SPE showed that the process was in control. The uncertainty intervals were clearly above the critical value of the SPE or Hotelling T^2 for many of the objects for which the full data statistic indicated an alarm but there were also objects where this was not so. These cases occurred in time ranges where the full data SPE indicated the process was disturbed so it is likely the uncertainty intervals were invalidated by a shift in the data distribution. Another possibility is that the assumption of a normal distribution for the data was not a good one in that region. The uncertainty intervals on the contributions to the SPE and Hotelling T^2 were examined for objects where the interval was close to or

straddled the control limit. The analysis correctly determined which measurement's recovery would reduce the uncertainty the most.

The three score calculation methods that were used for the example all performed well on the part of the data set where the process was not disturbed. SCP and mean replacement can produce values that are outside the uncertainty intervals because they do not use the same assumptions as those used in deriving the intervals. In the second half of the process, there were objects for which all methods performed poorly which could be due to the measurements that were missing or the process disturbance or both. CMR and SCP were not more sensitive to the data distribution than mean replacement. The SCP and mean replacement methods did tend to under-estimate the SPE, which, it was shown, would cause missed alarms, and the SCP method had some isolated cases where the errors were high, which it was shown would cause false alarms.

Chapter 4: Missing Measurements and Model Building

4.1 Introduction

Two issues will be explored in this chapter: model building with missing measurements and model building for applications that need to be robust to missing measurements. Being able to exploit objects with missing measurements in model building can lead to more accurate and robust models with a given data set or a reduced effort and cost to get a model of a prescribed quality (Ho et al., 2001 and Andrews and Wentzell, 1997). The minimum requirements to enable the use of objects with missing measurements are algorithms that make efficient use of all available objects and measures of model accuracy that consider the effects of the missing measurements. Diagnostics that indicate which sets of missing measurements and individual objects are problematic are also required in order for missing data modelling to enter widespread use. In chapter 2, it was shown to be advantageous to consider robustness to missing measurements in applications when building a model. Removing 'unnecessary' measurements can lead to a model that performs well in model building but poorly in applications when missing measurements are encountered.

Factors that affect the NIPALS, EM, MLPCA and iterative replacement model building algorithms are developed, and improvements to the MLPCA

algorithm proposed, in this chapter. A guide for pruning variables during model building is given which considers the resulting model's robustness to missing measurements, and a procedure proposed to apply the analysis of chapters 2 and 3 to improve model building with missing measurements. Also, the issues unique to model building are reviewed and the way in which they apply to the analysis in the rest of this chapter discussed.

Analysis of missing measurement model building methods is much more challenging than analysis of score calculation algorithms (chapter 2) and model applications (chapter 3) because we cannot assume that the distribution of the data is known. This chapter will first enumerate the issues unique to model building and discuss how they apply to the analysis in this chapter. Each model building algorithm described in chapter 1 will then be discussed in light of the analysis in the rest of the thesis. Building models that are robust to missing measurements in application will then be treated with the view that the model building data set is complete. The final section in this chapter will propose how the results in chapter 2 and chapter 3 can be applied directly to the analysis of building models with missing measurements.

4.2 Issues Unique to Model Building with Missing Measurements

In the previous chapters of this thesis it is assumed that a PCA or PLS model has already been built. The information used in building this model is then available to analyse the performance of the models in applications with missing

measurements. It is also possible that the covariance matrix of the variables has been estimated. When a model has not been built, this information is not available; all of the information about the distribution of the data comes from the data matrices themselves and is thus of questionable quality. The issues which are raised in model building that are not applicable to score calculation and model application are how representative the data set is of the population, possible bias in the calculated model due to the mechanism that causes missing measurements, the lack of centring and scaling weights for the pre-processing of the data and the analysis of iterative algorithms.

As stated in chapter 1, PCA and PLS decompose the covariance matrix or matrices of the variables. Even with complete data there are issues with how well the data set represents the population and with missing measurements these issues become much more important. The possibility of much more serious model bias exists because any bias that is present in the information content of the present measurements will be reinforced when that information is used to calculate replacement values for the missing measurements. An additional concern is the location of the missing data in the data matrix. For any possible subset of variables, there must be sufficient objects with measurements present to adequately represent the relationships between those variables. These issues will not be addressed further in this chapter.

The mechanism by which measurements become missing can bias the model. This is true even if only complete objects are used in model calculation. If

this mechanism does not have to be accounted for in the model building algorithm then it is said to be ignorable (Little and Rubin, 1987). An ignorable missing measurement mechanism is desirable because the model building algorithm will be simpler and the parameters of the missing measurement mechanism do not need to be known. Measurements are said to be missing completely at random (MCAR) when there is no relationship between the values of the variables and their probability of being missing. Examples of measurements that are missing completely at random are those due to randomly missed manual samples and regularly scheduled sensor maintenance that does not affect the operation of the process. This strict condition is sufficient but not necessary to have an ignorable missing measurement mechanism (Little and Rubin, 1987; Schafer, 1997). A necessary and sufficient condition for an ignorable missing measurement mechanism is that the parameters of the function that determines the probability that a measurement is missing must be measured and included in the data set. This is known as missing at random (MAR) and is much less strict. Samples that are not taken when certain process measurements are outside safe limits and maintenance procedures with associated process operating procedures can be examples of MAR mechanisms if certain variables are included in the data set. These variables are the ones checked against safe limits for the samples only taken under safe conditions and all quantities set by the operating procedure for the maintenance procedure. All of the model building methods discussed in this thesis require an ignorable missing measurement mechanism.

Centring and scaling the data before building a PCA or PLS model are not required but are commonly practiced. Centring on the variable means and scaling each variable to unit variance are the norm. This requires that the means of the variables and the diagonal of the covariance matrix be available. Since PCA and PLS are mean and scale dependent, the determination of the centring and scaling parameters must either be performed as a pre-processing step or included as a part of the missing measurement model building algorithm.

The algorithms for model building outlined in chapter 1 are iterative with the operations specified repeated until some measure of change falls below a threshold. In chapter 2, the algorithms for score calculation with missing measurements were analysed to determine the source and magnitude of score calculation errors. Individual steps of the model building algorithms are similar to the score calculation algorithms but while analysis of each step can be useful (Andrews and Wentzell, 1997), serious problems can be missed because the difficulty may lie in how the errors influence the algorithm over several iterations. This is a major limitation in this chapter on the information that can be deduced about the quality of a model from data with missing measurements.

4.3 Extending Analysis of Score Calculation and Model Application with Missing Measurements to Model Building

In chapter 1, the missing measurement score calculation methods were introduced and related to model building algorithms. In chapter 2, the score

calculation methods were analysed and the sources and magnitudes of the score calculation error determined on the assumption that the model has already been built. In this section of the thesis, it will be shown how the results of the score calculation analysis apply to the related model building algorithms. In addition, the issue of centring and scaling will be discussed for each algorithm.

4.3.1 NIPALS Algorithm

One of the steps in the NIPALS algorithm laid out in Equation 1-13 is the calculation of the scores with what is referred to in this thesis as SCP. The NIPALS calculation of the loading vector $\mathbf{p}_j$ for PCA and PLS is similar to SCP with the columns of $\mathbf{X}$ replacing the rows and the scores $\mathbf{t}_j$ replacing the loading vectors $\mathbf{p}_j$. This allows the score calculation error analysis to be extended to the calculation of the loading vectors.

For this analysis, we need to define $\mathbf{t}_j^*$, $\mathbf{u}_j^*$, $\mathbf{x}_k^*$, $\mathbf{e}_k^*$, $\mathbf{f}_k^*$ and $\mathbf{y}_k^*$ to be column vectors from PCA or PLS matrices with the objects removed which have variable k missing. This is analogous to the loading vectors $\mathbf{p}_j^*$ that have the variables which have missing measurements removed. For the data vector structure, we have $\mathbf{x}_k^*(j) = \sum_{m=1}^{K} p_{km}\mathbf{t}_m^* + \mathbf{e}_k^* - \sum_{m=1}^{j-1} \hat{p}_{km}\hat{\mathbf{t}}_m^*$. Once again as in section 2.2.1, the structure for the data is not restricted to the same dimension as the model (A latent vectors) but has K dimensions with some of these dimensions

possibly having scores with zero variance. The equation for the PCA or PLS loading matrix **P** calculation error analogous to Equation 2-2 is

$$\begin{aligned} p_{kj} - \hat{p}_{kj} &= p_{kj} - \left[\mathbf{t}_j^{*T}\mathbf{t}_j^*\right]^{-1}\mathbf{t}_j^{*T}\mathbf{x}_k^*(j) \\ &= p_{kj} - \left[\mathbf{t}_j^{*T}\mathbf{t}_j^*\right]^{-1}\mathbf{t}_j^{*T}\left[\sum_{m=1}^{K} p_{km}\mathbf{t}_m^* + \mathbf{e}_k^* - \sum_{m=1}^{j-1}\hat{p}_{km}\hat{\mathbf{t}}_m^*\right] \\ &= p_{kj} - \left[\mathbf{t}_j^{*T}\mathbf{t}_j^*\right]^{-1}\mathbf{t}_j^{*T}\sum_{m=1}^{K} p_{km}\mathbf{t}_m^* - \left[\mathbf{t}_j^{*T}\mathbf{t}_j^*\right]^{-1}\mathbf{t}_j^{*T}\mathbf{e}_k^* \\ &\quad + \left[\mathbf{t}_j^{*T}\mathbf{t}_j^*\right]^{-1}\mathbf{t}_j^{*T}\sum_{m=1}^{j-1}\hat{p}_{km}\hat{\mathbf{t}}_m^* \\ &= -\sum_{m=j+1}^{K}\left[\mathbf{t}_j^{*T}\mathbf{t}_j^*\right]^{-1}\mathbf{t}_j^{*T}\mathbf{t}_m^* p_{km} - \left[\mathbf{t}_j^{*T}\mathbf{t}_j^*\right]^{-1}\mathbf{t}_j^{*T}\mathbf{e}_k^* \\ &\quad - \left[\mathbf{t}_j^{*T}\mathbf{t}_j^*\right]^{-1}\sum_{m=1}^{j-1}\mathbf{t}_j^{*T}\left[p_{km}\mathbf{t}_m^* - \hat{p}_{km}\hat{\mathbf{t}}_m^*\right] \end{aligned} \qquad \text{Equation 4-1}$$

With no missing measurements, the error will only depend on the measurement error since $\mathbf{t}_j^T\mathbf{t}_m = 0 \quad j \neq m$. Thus the important quantities for the calculation of the loading matrix **P** for either PCA or PLS are the magnitude of the scores for the objects with missing measurements $\mathbf{t}_j^{*T}\mathbf{t}_j^*$, how the objects with missing measurements make $\mathbf{t}_j^*$ and $\mathbf{t}_m^*$ collinear, the magnitude of p_{km} (the importance of this measurement to this component), measurement error and propagation of previous component error. As objects corresponding to large scores are removed, $\mathbf{t}_j^{*T}\mathbf{t}_j^*$ gets smaller and its inverse larger which makes all of the error terms larger. The error term involving previous components can be seen to be affected both by loading and score estimation error.

A similar analysis for the PLS loading matrix **W** error yields

$$
\begin{aligned}
w_{kj} - \hat{w}_{kj} &= w_{kj} - \left[\mathbf{u}_j^{*T}\mathbf{u}_j^*\right]^{-1}\mathbf{u}_j^{*T}\mathbf{x}_k^*(j) \\
&= w_{kj} - \left[\mathbf{u}_j^{*T}\mathbf{u}_j^*\right]^{-1}\mathbf{u}_j^{*T}\left[\sum_{m=1}^{K} p_{km}\mathbf{t}_m^* + \mathbf{e}_k^* - \sum_{m=1}^{j-1}\hat{p}_{km}\hat{\mathbf{t}}_m^*\right] \\
&= w_{kj} - \left[\mathbf{u}_j^{*T}\mathbf{u}_j^*\right]^{-1}\mathbf{u}_j^{*T}\sum_{m=1}^{K} p_{km}\mathbf{t}_m^* - \left[\mathbf{u}_j^{*T}\mathbf{u}_j^*\right]^{-1}\mathbf{u}_j^{*T}\mathbf{e}_k^* \\
&\quad + \left[\mathbf{u}_j^{*T}\mathbf{u}_j^*\right]^{-1}\mathbf{u}_j^{*T}\sum_{m=1}^{j-1}\hat{p}_{km}\hat{\mathbf{t}}_m^* \\
&= \left[w_{kj} - \left[\mathbf{u}_j^{*T}\mathbf{u}_j^*\right]^{-1}\mathbf{u}_j^{*T}\mathbf{t}_j^* p_{kj}\right] - \sum_{m=j+1}^{K}\left[\mathbf{u}_j^{*T}\mathbf{u}_j^*\right]^{-1}\mathbf{u}_j^{*T}\mathbf{t}_m^* p_{km} \\
&\quad - \left[\mathbf{u}_j^{*T}\mathbf{u}_j^*\right]^{-1}\mathbf{u}_j^{*T}\mathbf{e}_k^* - \left[\mathbf{u}_j^{*T}\mathbf{u}_j^*\right]^{-1}\sum_{m=1}^{j-1}\mathbf{u}_j^{*T}\left[p_{km}\mathbf{t}_m^* - \hat{p}_{km}\hat{\mathbf{t}}_m^*\right]
\end{aligned}
\qquad \text{Equation 4-2}
$$

which does not give an equation that can be seen to directly yield loading vector error depending only on measurement error with no missing measurements. The first two terms must cancel with no missing measurements so further work should yield such an equation.

The structure for column k of $\mathbf{Y}$ is $\mathbf{y}_k^*(j) = \sum_{m=1}^{K} q_{km}\mathbf{t}_m^* + \mathbf{f}_k^* - \sum_{m=1}^{j-1}\hat{q}_{km}\hat{\mathbf{t}}_m^*$. The error in the output loading matrix $\mathbf{Q}$ is then

$$
\begin{aligned}
q_{kj} - \hat{q}_{kj} &= q_{kj} - \left[\mathbf{t}_j^{*T}\mathbf{t}_j^*\right]^{-1}\mathbf{t}_j^{*T}\mathbf{y}_k^*(j) \\
&= q_{kj} - \left[\mathbf{t}_j^{*T}\mathbf{t}_j^*\right]^{-1}\mathbf{t}_j^{*T}\left[\sum_{m=1}^{K} q_{km}\mathbf{t}_m^* + \mathbf{f}^* - \sum_{m=1}^{j-1}\hat{q}_{km}\hat{\mathbf{t}}_m^*\right] \\
&= q_{kj} - \left[\mathbf{t}_j^{*T}\mathbf{t}_j^*\right]^{-1}\mathbf{t}_j^{*T}\sum_{m=1}^{K} q_{km}\mathbf{t}_m^* - \left[\mathbf{t}_j^{*T}\mathbf{t}_j^*\right]^{-1}\mathbf{t}_j^{*T}\mathbf{f}^* \\
&\quad + \left[\mathbf{t}_j^{*T}\mathbf{t}_j^*\right]^{-1}\mathbf{t}_j^{*T}\sum_{m=1}^{j-1}\hat{q}_{km}\hat{\mathbf{t}}_m^* \\
&= -\sum_{m=j+1}^{K}\left[\mathbf{t}_j^{*T}\mathbf{t}_j^*\right]^{-1}\mathbf{t}_j^{*T}\mathbf{t}_m^* q_{km} - \left[\mathbf{t}_j^{*T}\mathbf{t}_j^*\right]^{-1}\mathbf{t}_j^{*T}\mathbf{f}^* \\
&\quad - \left[\mathbf{t}_j^{*T}\mathbf{t}_j^*\right]^{-1}\sum_{m=1}^{j-1}\mathbf{t}_j^{*T}\left[q_{km}\mathbf{t}_m^* - \hat{q}_{km}\hat{\mathbf{t}}_m^*\right]
\end{aligned}
$$

Equation 4-3

The result is identical to the **P** matrix result with the **P** matrix elements replaced by **Q** matrix elements. The dependence is therefore the same with the exception that it is a **Y** matrix variable loading that determines the importance of the variable to the component.

Finally, the analysis of the error in the elements of the output score matrix **U** requires a structure for a row of the **Y** matrix $\boldsymbol{\psi}_i$. This structure is

$$\boldsymbol{\psi}_i^* = \sum_{m=1}^{K}\mathbf{q}_m^* t_{im} + \boldsymbol{\phi}_i^* - \sum_{m=1}^{j-1}\hat{\mathbf{q}}_m^*\hat{t}_{im}$$

$$
\begin{aligned}
u_{ij} - \hat{u}_{ij} &= u_{ij} - \left[\mathbf{q}_j^{*T}\mathbf{q}_j^{*}\right]^{-1}\mathbf{q}_j^{*T}\boldsymbol{\psi}_i^{*} \\
&= u_{ij} - \left[\mathbf{q}_j^{*T}\mathbf{q}_j^{*}\right]^{-1}\mathbf{q}_j^{*T}\left[\sum_{m=1}^{K}\mathbf{q}_m^{*}t_{im} + \boldsymbol{\phi}_i^{*} - \sum_{m=1}^{j-1}\hat{\mathbf{q}}_m^{*}\hat{t}_{im}\right] \\
&= u_{ij} - \left[\mathbf{q}_j^{*T}\mathbf{q}_j^{*}\right]^{-1}\mathbf{q}_j^{*T}\sum_{m=1}^{K}\mathbf{q}_m^{*}t_{im} - \left[\mathbf{q}_j^{*T}\mathbf{q}_j^{*}\right]^{-1}\mathbf{q}_j^{*T}\boldsymbol{\phi}_i^{*} \\
&\quad + \left[\mathbf{q}_j^{*T}\mathbf{q}_j^{*}\right]^{-1}\mathbf{q}_j^{*T}\sum_{m=1}^{j-1}\hat{\mathbf{q}}_m^{*}\hat{t}_{im} \\
&= u_{ij} - \left[\mathbf{q}_j^{*T}\mathbf{q}_j^{*}\right]^{-1}\mathbf{q}_j^{*T}\mathbf{q}_j^{*}t_{ij} - \sum_{m=j+1}^{K}\left[\mathbf{q}_j^{*T}\mathbf{q}_j^{*}\right]^{-1}\mathbf{q}_j^{*T}\mathbf{q}_m^{*}t_{im} \\
&\quad - \left[\mathbf{q}_j^{*T}\mathbf{q}_j^{*}\right]^{-1}\mathbf{q}_j^{*T}\boldsymbol{\phi}_i^{*} - \left[\mathbf{q}_j^{*T}\mathbf{q}_j^{*}\right]^{-1}\sum_{m=1}^{j-1}\mathbf{q}_j^{*T}\left[\mathbf{q}_m^{*}t_{im} - \hat{\mathbf{q}}_m^{*}\hat{t}_{im}\right]
\end{aligned}
\qquad \text{Equation 4-4}
$$

The error in the **Y** matrix scores **U** is similar in form to the error for the **X** matrix scores **T**, but the **Y** matrix loading vectors **q** are not required to be orthogonal so it is not clear how the error will be zero with no missing measurements. Future work to yield this understanding will reveal the factors that affect the error.

Finally, the NIPALS algorithm requires that any centring or scaling weights be calculated before the modelling is done. These must be calculated from the available data by an independent method such as using only the complete objects.

4.3.2 EM Algorithm

The expectation step of the EM algorithm computes values for the missing measurements which are used in CMR score calculation. It was shown in section 2.4 that the expectation when calculated by Equation 1-31 could be vulnerable to ill-conditioning. The conditioning of present variable covariance matrices is

handled by latent variable regression methods in chapter 2 but another approach is given in Schafer (1997) where a uniform prior is used to reduce the influence of the ill-conditioning. The speed of convergence of EM is related to the number of missing measurements (Schafer, 1997) and can be quite slow. A number of acceleration methods have been devised (Schafer, 1997).

4.3.3 Analysis of MLPCA for Missing Measurement Model Calculation

As mentioned in chapter 1, in standard PCA notation $\mathbf{P} = \hat{\mathbf{V}}$ so from Equation 1-20 the MLPCA calculation for the scores is

$$\hat{\boldsymbol{\tau}}_i = \left[\hat{\mathbf{V}}^T \boldsymbol{\Sigma}_i^{-1} \hat{\mathbf{V}}\right]^{-1} \hat{\mathbf{V}}^T \boldsymbol{\Sigma}_i^{-1} \mathbf{z}_i = \left[\mathbf{P}^T \boldsymbol{\Sigma}_i^{-1} \mathbf{P}\right]^{-1} \mathbf{P}^T \boldsymbol{\Sigma}_i^{-1} \mathbf{z}_i \qquad \text{Equation 4-5}$$

The inverse of a diagonal matrix is also diagonal with the diagonal elements of one matrix the inverse of the other. If the inverse of the large error variance on the missing measurements is taken to be approximately zero and the variables are re-arranged so the missing measurements occur at the beginning of the object we have $\boldsymbol{\Sigma}_i^{-1} \approx \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}$ where $\mathbf{I}$ is the identity matrix with dimension equal to the number of present measurements. Using the superscript notation for missing and present measurements, this approximation to $\boldsymbol{\Sigma}_i^{-1}$ in Equation 4-5 produces

$$\begin{aligned} \hat{\boldsymbol{\tau}}_i &\approx \left[\begin{bmatrix} \mathbf{P}^{\#} \\ \mathbf{P}^{*} \end{bmatrix}^T \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{P}^{\#} \\ \mathbf{P}^{*} \end{bmatrix}\right]^{-1} \begin{bmatrix} \mathbf{P}^{\#} \\ \mathbf{P}^{*} \end{bmatrix}^T \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{z}_i^{\#} \\ \mathbf{z}_i^{*} \end{bmatrix} \\ &\approx \left[\begin{bmatrix} \mathbf{P}^{\#T} & \mathbf{P}^{*T} \end{bmatrix} \begin{bmatrix} \mathbf{0} \\ \mathbf{P}^{*} \end{bmatrix}\right]^{-1} \begin{bmatrix} \mathbf{P}^{\#T} & \mathbf{P}^{*T} \end{bmatrix} \begin{bmatrix} \mathbf{0} \\ \mathbf{z}_i^{*} \end{bmatrix} \\ &\approx \left[\mathbf{P}^{*T} \mathbf{P}^{*}\right]^{-1} \mathbf{P}^{*T} \mathbf{z}_i^{*} \end{aligned} \qquad \text{Equation 4-6}$$

which is the formula for score calculation by projection to the model plane in Nelson et al. (1996) and chapter 2.

Similarly for row k of the **P** matrix (denoted π_k) from Equation 1-21,

$$\hat{\boldsymbol{\pi}}_k = \mathbf{S}^{-1}\left[\hat{\mathbf{U}}^T\boldsymbol{\Psi}_k^{-1}\hat{\mathbf{U}}\right]^{-1}\hat{\mathbf{U}}^T\boldsymbol{\Psi}_k^{-1}\mathbf{x}_k \qquad \text{Equation 4-7}$$

If the inverse of the large error variance on the missing measurements is taken to be approximately zero and the objects are re-arranged so the missing measurements occur at the beginning of the vector $\mathbf{x}_k$ we have $\boldsymbol{\Psi}_k^{-1} \approx \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}$

where **I** is the identity matrix with dimension equal to the number of objects with variable k present. Using the superscript notation for missing and present measurements, this approximation to $\boldsymbol{\Psi}_i^{-1}$ in Equation 4-5 produces

$$\hat{\boldsymbol{\pi}}_k \approx \mathbf{S}^{-1}\left[\begin{bmatrix}\hat{\mathbf{U}}^{\#}\\ \hat{\mathbf{U}}^{*}\end{bmatrix}^T\begin{bmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{I}\end{bmatrix}\begin{bmatrix}\hat{\mathbf{U}}^{\#}\\ \hat{\mathbf{U}}^{*}\end{bmatrix}\right]^{-1}\begin{bmatrix}\hat{\mathbf{U}}^{\#}\\ \hat{\mathbf{U}}^{*}\end{bmatrix}^T\begin{bmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{I}\end{bmatrix}\mathbf{x}_k \qquad \text{Equation 4-8}$$

$$\approx \mathbf{S}^{-1}\left[\hat{\mathbf{U}}^{*T}\hat{\mathbf{U}}^{*}\right]^{-1}\hat{\mathbf{U}}^{*T}\mathbf{x}_k^{*}$$

which is directly analogous to the result in Equation 4-6 except for the scaling provided by **S**.

It was shown in chapter 2 that when using projection to the model plane it was important to include all model dimensions associated with significant scores and that ill-conditioning of the projection matrices could increase error. It can now be seen that this will also apply to MLPCA with missing measurements since it uses effectively identical projections to calculate its model. Andrews and Wentzell (1997) examine an approximation to the error covariance matrix for the

scores which assumes the loading vectors are fixed. It was suggested that this be used to detect outliers and construct confidence intervals. Recognising that the approximate score error covariance developed is a term from Equation 4-6, we have $\mathrm{cov}(\tau) \approx \left[\mathbf{P}^{*T}\mathbf{P}^{*}\right]^{-1}$. Examining error contours from this matrix will give similar information to that which would be obtained by looking at the condition number.

The model that is calculated using MLPCA is only optimal in a maximum likelihood sense for the set of measurement errors and missing measurement replacement values that are chosen. It was shown in chapter 2 that using the conditional mean for missing measurements can give superior results to projection to the model plane in score calculation. CMR uses the information about the missing measurements in the present measurements to improve the score estimate. The variance of the missing measurements conditional on the present measurements and the data distribution that was used in chapter 3 could be applied here to provide a measurement error variance that is related to the actual degree of uncertainty. EM could be used to provide both the conditional mean values and the covariance matrix used for this purpose. In addition, EM would provide mean and variance values for centring and scaling of the variables that are lacking in the Andrews and Wentzell (1997) paper. EM alone could be used to produce the PCA model as stated in section 1.6.1, but the MLPCA algorithm includes the uncertainty due to missing measurements more explicitly and might prove to be more robust.

4.3.4 Iterative Replacement

The Iterative Replacement algorithm is related to projection to the model plane in that the missing measurements are replaced at each iteration with the estimates from the current model. Iteration continues until the difference between the replacement values and the new values from the model drop below a threshold. The missing measurements will have approximately zero residuals in the model when the model has converged in the same way that the residuals of the missing measurements are set to zero in score calculation with projection to a model plane. It is therefore important to include all model dimensions associated with significant scores. Centring and scaling are a part of the algorithm.

Rannar et al. (1994b) refer to this method as EM because the algorithm alternates between calculating new values of the missing measurements and updating the covariance matrices but it more closely resembles Buck's method as outlined in Little and Rubin (1987). Buck's method uses linear regression on the present variables to calculate replacements for the missing variables. The variances and covariances of variables with missing measurements tend to be underestimated by Buck's method because it does not account for the variance between pairs of variables where both are missing which cannot be predicted from the present variables. Iterative replacement is similar in that it does not correct the updated covariance matrix for variance in the missing measurements that can not be predicted from the measured variables. This may make the iterative replacement algorithm more stable when this correction term is large, but will

cause the variances and covariances of variables with missing measurements to be biased towards zero.

EM uses only one data matrix where iterative replacement separates the data into a dependent and independent data matrix. EM can use covariance between any variables to calculate conditional means and variances whereas iterative replacement can only use covariance from the dependent variables.

4.4 Building Models that are Robust to Missing Measurements

When PCA or PLS models are being built from highly correlated process data, and are to be used for predicting responses or for monitoring the process in the future, there is often an incentive to reduce the number of variables. This is termed pruning the model. This can often be accomplished at the model building stage with little loss in the predictive power of the model. There can be a saving in sensor or analytical measurement costs by reducing the number of variables used and there is a lower probability of having missing data in any new object with fewer variables. However, with less redundancy in the variables, the model can also be much more sensitive to missing data when it does occur. The error analysis in chapter 2 and 3 can be used to provide some useful guidelines for variable selection and for testing the resulting pruned model for robustness in the presence of missing data.

The steps to be followed are:

I. Prune the variables; an iterative procedure may be necessary.

A. Discard:

1. variables that are strongly correlated to others.

2. variables which do not enter the model strongly.

B. Do not discard:

1. all of a group of variables that define a latent direction and could go missing simultaneously.

2. a measurement or group of measurements that can make different latent vectors nearly collinear with one another when using the NIPALS algorithm, that is, make $\mathbf{w}_j^*$ or $\mathbf{p}_j^*$ nearly collinear with $\mathbf{p}_m^*$ $(j \neq m)$. This is especially important among the dominant latent variables that explain a large amount of the variance in the data.

II. Determine the allowable levels of score, prediction, Hotelling T^2 or SPE uncertainty based on the end application of the model.

III. Determine which combinations of variables are to be tested as missing simultaneously in the score calculation procedure. Sources of these combinations can be:

A. the groupings determined by the pruning method.

B. variables with a large (greater than 0.5 or so) weighting in the **p** loading vector for PCA or w loading vector for PLS.

C. variables simultaneously missing due to scheduled shift changes or maintenance.

D. all possible individual, paired or other sets of measurements.

IV. Use the training set data used to build the model and the diagnostics from chapter 3 to evaluate the performance of the desired application.

V. If a set of variables that produces an unacceptably large error is found:

A. try to reduce the set to the smallest number of variables that will produce an unacceptable level of performance to limit the number of variables that will be considered essential for good performance. Use uncertainty intervals in contributions to determine which variables are causing the problem.

B. seek an alternate or alternates for the variables in the reduced set from the grouping determined by the pruning method or mechanistic knowledge. The alternate or alternates should be correlated to the missing variable or variables and enter the loading vectors in the same way.

C. re-calculate the model and go back to step III.

4.5 Applying Results from this Thesis to Model Building with Missing Measurements

This section will propose how the results in chapter 2 and chapter 3 can be applied directly to building models with missing measurements. Each missing

measurement model building method has been associated with a score calculation method which will be used to provide information about missing measurement combinations that will cause problems and give an indication of the quality of model that can be expected. In addition, the results of chapter 3 can be applied.

I. The major hurdle to this analysis is the assumption that a model has been calculated and, for chapter 3, that a covariance matrix can also be supplied. Two alternatives exist to fulfil this need. A model can be calculated with the complete objects only or analysis can be applied to a missing measurement model as a post-model building application, as in Andrews and Wentzell (1997). In either case more latent vectors should be calculated than are necessary for modelling in order to establish score magnitudes and latent vectors. A covariance matrix can also be obtained from complete objects or when EM or iterative replacement are used.

II. One goal of the analysis is to identify which of the missing measurement combinations present in the data are problematic.

   A. Once a model has been obtained, the complete objects can be used with the model and the relevant score calculation method to evaluate each missing measurement combination by comparing the complete data scores to the scores calculated by the missing measurement algorithm.

B. The issues raised in section 4.3 should also be addressed, such as the relative magnitudes of the scores and the degree of collinearity of the loading and score vectors under missing measurements.

C. Objects which have missing measurement combinations identified by the above analysis as causing large score calculation errors should be removed from the data set. Deleting these objects will require the model to be re-calculated if a model from data with missing measurements was used for the analysis. The analysis in steps A and B is then repeated with the remaining objects with missing measurements until no objects are eliminated. This is not necessary with a complete data model since the model does not change with the deletion of objects with missing measurements.

III. The analysis of chapter 3 can then be used with the remaining objects with missing measurements to test individual objects for those with large score uncertainty. The objects with large score uncertainty should be deleted which will require the model to be re-calculated if a model from data with missing measurements was used. The score error analysis of step II and III should then be repeated with the new model.

IV. Finally, there is an additional step if the above analysis was done with a complete data model. The objects with missing measurements would not be included in the data set if it was not thought that they would influence the model. Since the model including the remaining objects with missing

measurements should be different from the complete object model then the results of the above analysis may change. The analysis should therefore be repeated with the model from all remaining objects as a final check.

4.6 Conclusions and Future Work

Factors that affect the NIPALS, EM, MLPCA and iterative replacement model building algorithms have been developed and improvements to the MLPCA algorithm proposed in this chapter. A guide for pruning variables during model building has been given which considers the resulting model's robustness to missing measurements, and a procedure proposed to apply the analysis of chapters 2 and 3 to improve model building with missing measurements. Also, the issues unique to model building have been reviewed and the way in which they apply to the analysis in the rest of this chapter discussed.

The challenges in analysing model building with missing measurements that are not present in analysing score calculation and model application with missing measurements have been reviewed in this chapter. The lack of a prior model or distribution information to provide a basis for analysis and the need to provide centring and scaling factors are addressed in later sections of the chapter. How well the data set represents the true variation of the variables and the iterative nature of the model building algorithms are not addressed. The importance of the mechanism by which the measurements become missing needs to be considered for any of the model building methods discussed.

The NIPALS algorithm, EM, MLPCA and iterative replacement were each examined to bring out how they are affected by the issues identified in chapter 2 and 3 and how centring and scaling of the variables is accomplished.

In the NIPALS algorithm, centring and scaling must be provided separately. The following factors have been shown to cause errors in calculation of **P** and **Q** in individual iterations of the algorithm's steps: (i) large sum of squared scores in the objects that are missing, (ii) the degree to which missing measurements make score vectors collinear and (iii) the magnitude of the loading on the missing measurement. The factors that affect **T** were identified in chapter 2. The analysis of the **W** and **U** elements requires further insight to extract the relevant factors. This analysis does not account for interactions between the steps and iterations of the algorithm.

The EM algorithm does provide mean and variance estimates for centring and scaling of the data. Ill-conditioning of the present variable sub-matrices of the variance matrix was identified as a factor causing error.

The individual projection steps in the MLPCA algorithm were shown to be related to projection to the model plane score calculation. This indicates that it is important to include all model dimensions associated with significant scores and that ill-conditioning of the projection matrices can cause error. MLPCA does not have a methodology for providing mean and variance estimates for centring and scaling and it was proposed that these could be provided by EM. The means and variances would also provide better estimates of the missing measurement values

and error variances than were used in Andrews and Wentzell (1997). The improved missing measurement replacement values in CMR were shown to reduce score error over projection to the model plane. The investigation of combining EM and MLPCA, and the advantage of this new method over just using EM, should be the subject of future work.

Iterative Replacement is also related to projection to the model plane so that all model dimensions associated with significant scores should be included. Unlike MLPCA, centring and scaling is a part of the iterative replacement algorithm. The differences between the iterative replacement model building algorithm and a model building algorithm using EM as outlined in section 1.6.1 were examined. These differences include EM using conditional variance to prevent bias of variances from missing measurements and the usage by EM of information in the Y matrix of PLS on a basis equal to the X matrix.

When reducing the number of variables in a PCA or PLS model it is possible to reduce the robustness of the model to missing measurements in applications. This was demonstrated in chapter 2 with the kamyr digester example and applies even when the model building data set is complete. A procedure has been proposed to use the analysis of chapter 2 and 3 to determine the impact of pruning on model robustness to missing measurements and guide the selection of the variables to be pruned to maximise robustness to missing measurements in the resulting model. Additional work is required to determine the utility of this approach when the data set for model building is incomplete.

Finally, a procedure for applying the analysis of chapter 2 and chapter 3 to model building with missing measurements has been proposed. This procedure has been designed to provide an indication of which variables and objects with missing measurements can have a major negative impact on the calculated model. The procedure does not account for the effects of multiple iterations in an algorithm and interactions between individual algorithm steps. Application to modelling of both complete data sets with measurements deliberately removed and incomplete data sets is required to determine the utility of this approach.

Chapter 5: Conclusions and Future Work

In applications of Principal Components Analysis (PCA) and Projection to Latent Structures (PLS) it is very important to be able to handle observations with some of the measurements missing. The measurements can be missing in a data set for building a model or a new object to which a pre-existing model is applied in an application. A novel missing measurement score calculation algorithm is proposed in chapter 2 of this thesis, and the factors affecting the performance of both the novel and existing algorithms are identified and illustrated by designed and industrial process data sets. Performance measures for the application of PCA and PLS models are derived in chapter 3 and illustrated on an industrial data set. These measures distinguish between situations where model performance with missing measurements will continue to be acceptable and situations where it will become unacceptable. In the latter case, the measurements must be recovered or the application shut down. Diagnostics are developed to aid in determining which missing measurements are causing the greatest uncertainty in the application and will yield the greatest reduction in uncertainty by their recovery. Factors that affect the NIPALS, EM, MLPCA and iterative replacement model building algorithms are developed, and improvements to the MLPCA algorithm proposed in chapter 4. A guide for pruning variables during model building is given which considers the resulting model's robustness to missing measurements and a

procedure proposed to apply the analysis of chapters 2 and 3 to improve model building with missing measurements.

Chapter 2 analyses the algorithms that are used for handling missing measurements when the underlying PCA or PLS model is assumed to be fixed and known. This is the situation when a PCA or PLS model has been built from a large amount of plant data and it is to be applied using new observations that have missing measurements. A novel algorithm for score calculation with missing measurements is developed called conditional mean replacement and its properties analysed and compared with those of existing algorithms. The sources of error for each algorithm are laid out and illustrated using designed data sets. This gives specific quantities calculated from information available from model building to determine score calculation performance with each missing measurement algorithm. Recommendations on pruning variables and on the number of components to use in a model for good performance with missing measurements are made. An analysis of two process data sets from the literature, one simulated and the other industrial, is presented to show application to realistic situations. While all methods perform well in most cases, it is shown that the novel method is best in certain critical situations identified by the analysis.

An expression for the score estimation error from the single component projection missing data algorithm shows that the error increases with: (i) collinearity of the loading vectors, after loadings for the missing measurements are removed, and (ii) similarity in the magnitudes of the scores. An increase in the

noise variance in the measurements will also increase the error. Large errors in a PLS score calculation using the single component projection method are shown to arise when the **w** and **p** loading vectors are significantly different. This is caused by the presence of variables in the independent data block which do not explain a significant amount of the dependent data. Errors are shown to propagate from estimated scores to subsequently calculated ones through deflation. The analysis is demonstrated on industrial process data and a simulated process taken from the literature.

Improvement over single component projection PCA score calculation is possible by fitting all loading vectors at once in a procedure called Projection to the Model Plane. The score estimation error equation developed for this method showed the sources of the error to be: (i) under-estimating the dimensionality of the data, (ii) noise and (iii) ill-conditioning in the least squares projection. A biased regression algorithm is recommended for the projection and was shown to combat ill-conditioning.

Another, generally better, method to calculate scores when there are missing measurements, which can be applied to both PCA and PLS, is to replace a missing variable by its conditional mean given the variables that are observed and then to use standard score estimation routines. This is termed Conditional Mean Replacement, and it had the lowest mean squared score estimation error in all examples evaluated here. This method has information about the expected squared magnitude of the scores, therefore the error is due only to lack of

information, numerical ill-conditioning, noise, or violations of the assumptions made in the least squares regression between the measured and unmeasured variables. Identical score estimates are obtained if it is assumed either that all of the variables follow a multivariate normal distribution and the unmeasured variables are replaced by their conditional means (maximum likelihood estimates), or that the assumptions of least squares regression for a linear model between the measured variables and the scores hold. A biased regression method can be used in place of least squares to combat any ill-conditioning in the latter approach; this was not necessary in any of the examples in this chapter.

An example given in the literature as an illustration of PLS robustness to missing measurements was analysed. It was shown that the factors that cause score calculation error with missing measurements identified in this chapter were small which agrees with the conclusions in the literature.

The analysis of the industrial data set in this chapter agrees with the guideline that most of the missing data algorithms perform quite well with up to 20% of the measurements missing. However, the theoretical score error analysis shows that certain critical combinations of missing measurements can give rise to large errors. This is illustrated by the pathological simulated data sets and the industrial example. In these situations, the conditional mean replacement method is shown to be superior to the single component projection method and the simultaneous projection to the plane method. Pruning of variables during model building was shown to have potential to reduce the ability of the model to be used

with missing measurements. Leaving in some redundant variables was shown to increase robustness while leaving potential for a reduction in the number of variables.

In applications of PCA and PLS models when there are missing measurements the quality of the conclusions made based on the outputs of the model are in doubt. The impact that the uncertainty arising from the missing measurements will have on the predictions, SPE, Hotelling T^2 and contributions to these quantities is quantified in chapter 3 so appropriate action can be taken. This action may be to use the results as presented, to attempt to recover key measurements or to shut the application down. Current practice is to use the scores calculated using the missing measurement algorithms in the same manner as scores from complete objects. Uncertainty intervals arising from missing measurements have been derived for the predictions, SPE, Hotelling T^2 and contributions to these quantities. The prediction uncertainty interval combines the variance from parameter estimation together with the uncertainty arising from the missing measurements, with the missing measurement uncertainty dominating in the example. All other uncertainty intervals developed are due to missing measurement uncertainty alone. Uncertainty intervals for the contributions to these quantities are developed to aid in determining which missing measurements play the largest role in the uncertainty interval.

A PLS application to an industrial data set has been used to show that these uncertainty intervals can be used as performance measures and diagnostics

in monitoring and prediction when there are missing measurements. A specific combination of missing measurements was shown to cause uncertainty intervals in the Hotelling T^2 large enough to justify that the monitoring application be shut down. Another combination of missing measurements caused false Hotelling T^2 and missed SPE alarms with the single component projection method for some objects. These objects had uncertainty intervals that cast doubt on or indicated against the erroneous results.

The approximate prediction variance matrix considering a single output derived in this chapter has two terms: one involving the variance of the scores conditional on the present measurements and the other the variance of the **Q** loading vector. The development does not restrict the conditional distribution of the scores or PLS loading vector **Q** variance chosen but the scores must be independent of **Q** for the approximation to hold. For this work the conditional variance of the scores was considered only to arise from missing measurements and the **Q** variance from measurement errors in the dependent variable y. The combined estimation and missing measurement prediction variance was shown to be much greater than the estimation variance alone for the kamyr digester industrial data. This shows that the current practice of ignoring the uncertainty introduced by the missing measurements can be unacceptable. The uncertainty regions produced for the kamyr digester data matched the full data predictions well when the SPE was less than the critical value, showing that the assumptions made in producing the uncertainty interval are reasonable in practice.

The uncertainty intervals for the SPE, Hotelling T^2 and the contributions to these statistics that have been derived here bring more information about the state of the process and our knowledge of that state. This increases the confidence in the application of PCA and PLS models when measurements are missing. The process must be 'in control' for the uncertainty intervals to be valid. Quantifying the uncertainty allows alarms to be acted on with confidence if the uncertainty interval is clearly above the critical value of the statistic. It also reveals when the uncertainty in the statistic is so large that alarm conditions can not be detected. An uncertainty interval entirely below the critical value indicates that either an alarm condition is not present or that the missing measurements are key to detecting that alarm. The uncertainty intervals on the contributions give an indication of which measurements must be recovered to reduce the uncertainty in the test statistic so decisions can be made which balance the cost of recovering a measurement with the benefit to the application. The effect of correlation between missing measurements was not considered in developing the uncertainty intervals on the contributions which may result in a sub-optimal choice being made for which variables to recover.

The application of the derived uncertainty intervals to an industrial data set showed that the assumptions on the data distribution and the way that the uncertainty intervals were defined were reasonable in practice. The uncertainty intervals were consistent with the values calculated from all of the measurements (full data) when the SPE showed that the process was in control. The uncertainty

intervals were clearly above the critical value of the SPE or Hotelling T^2 for many of the objects for which the full data statistic indicated an alarm but there were also objects where this was not so. These cases occurred in time ranges where the full data SPE indicated the process was disturbed so it is likely the uncertainty intervals were invalidated by a shift in the data distribution. The uncertainty intervals on the contributions to the SPE and Hotelling T^2 were examined for objects where the interval was close to or straddled the control limit. The analysis correctly determined which measurement's recovery would reduce the uncertainty the most.

The three score calculation methods that were used for the example, SCP, CMR and mean replacement, all performed well on the part of the data set where the process was not disturbed. In the second half of the process, there were objects for which they all performed poorly which could be due to the measurements that were missing, or the process disturbance, or both. CMR and SCP were not more sensitive to the data distribution than mean replacement. The SCP and mean replacement methods did tend to under-estimate the SPE, which it was shown would cause missed alarms, and the SCP method had some isolated cases where the errors were high, which it was shown would cause false alarms.

Factors that affect the NIPALS, EM, MLPCA and iterative replacement model building algorithms have been developed, and improvements to the MLPCA algorithm proposed in chapter 4. A guide for pruning variables during model building has been given which considers the resulting model's robustness

to missing measurements and a procedure proposed to apply the analysis of chapters 2 and 3 to improve model building with missing measurements. Also, the issues unique to model building have been reviewed and the way in which they apply to the analysis in the rest of this chapter discussed.

The challenges in analysing model building with missing measurements that are not present in analysing score calculation and model application with missing measurements include the lack of a prior model or distribution information to provide a basis for analysis, the need to provide centring and scaling factors, how well the data set represents the true variation of the variables, the iterative nature of the model building algorithms and the importance of the mechanism by which the measurements become missing. The first two challenges are addressed in this chapter; the next two are not. The importance of the mechanism by which the measurements become missing needs to be considered for any of the model building methods discussed.

The NIPALS algorithm, EM, MLPCA and iterative replacement were each examined to bring out how they are affected by the issues identified in chapters 2 and 3 and how centring and scaling of the variables is accomplished.

In the NIPALS algorithm, centring and scaling must be provided separately and the following factors have been shown to cause errors in individual iterations of the algorithm's steps to calculate the **P** and **Q** matrices: (i) large sum of squared scores in the objects that are missing, (ii) the degree to which missing measurements make score vectors collinear and (iii) the magnitude of the loading

on the missing measurement. These factors are in addition to the factors for calculating **T** using SCP developed in chapter 2 and also do not account for interactions between the steps and iterations of the algorithm. The analysis of the **W** and **U** elements requires further insight to extract the relevant factors.

EM does provide mean and variance estimates for centring and scaling but ill-conditioning of the present variable sub-matrices of the variance matrix can cause problems with ill-conditioning and convergence speed can be a problem.

The individual projection steps in the MLPCA algorithm were shown to be related to projection to the model plane score calculation. This indicates that it is important to include all model dimensions associated with significant scores and that ill-conditioning of the projection matrices can cause error. MLPCA does not have a methodology for providing mean and variance estimates for centring and scaling and it was proposed that these could be provided by EM. The means and variances would also provide better estimates of the missing measurement values and variances than were used in Andrews and Wentzell (1997) which was shown to reduce score error in CMR over projection to the model plane. The investigation of this new method and the advantage of combining EM and MLPCA over just using EM should be the subject of future work.

Iterative Replacement is also related to projection to the model plane so that all model dimensions associated with significant scores should be included. Unlike MLPCA, centring and scaling is a part of the iterative replacement algorithm. The differences between the iterative replacement model building

algorithm and a model building algorithm using EM as outlined in section 1.6.1 were examined. These differences include EM using conditional variance to prevent bias of variances from missing measurements and the usage by EM of information in the Y matrix of PLS on a basis equal to the X matrix.

When reducing the number of variables in a PCA or PLS model it is possible to reduce the robustness of the model to missing measurements in applications. This was demonstrated in chapter 2 with the kamyr digester example and applies even when the model building data set is complete. A procedure has been proposed to use the analysis of chapter 2 and 3 to determine the impact of pruning on model robustness to missing measurements and guide the selection of the variables to be pruned to maximise robustness to missing measurements in the resulting model. Additional work is required to determine the utility of this approach when the data set for model building is incomplete.

Finally, a procedure for applying the analysis of chapter 2 and chapter 3 to model building with missing measurements has been proposed. This procedure has been designed to provide an indication of which variables and objects with missing measurements can have a major negative impact on the calculated model. The procedure does not account for the effects of multiple iterations in an algorithm and interactions between individual algorithm steps. Application to modelling of both complete data sets with measurements deliberately removed and incomplete data sets is required to determine the utility of this approach.

References

Andrews, Darren T. and Peter D. Wentzell. (1997) "Applications of maximum likelihood principal component analysis: incomplete data sets and calibration transfer". Analytica Chimica Acta. Vol. 350. pp. 341-352.

Burnham, A. J., R. Viveros-Aguilera and J.F. MacGregor. (1996) "Frameworks for Latent Variable Multivariate Regression". Journal of Chemometrics. Vol. 10. pp. 31-45.

Dayal, B. (1992) Feedforward Neural Networks for Process Modelling and Control. M. Eng. Thesis. Department of Chemical Engineering, McMaster University. Hamilton, Ontario, Canada.

Dayal, B., J.F. MacGregor, P.A. Taylor, R. Kildaw and S. Marcikic. (1994) "Application of Feedforward Neural Networks and Partial Least Squares Regression for Modelling Kappa Number in a Continuous Kamyr Digester". Pulp and Paper Canada. Vol. 95. No. 1. pp. T7-T13.

Dayal, B.S. and J.F. MacGregor (1997). "Improved PLS Algorithms". Journal of Chemometrics. Vol.11. pp. 73-85.

de Jong, Sijmen and Cajo J.F. Ter Braak. (1994) "Comments on the PLS Kernel Algorithm". Journal of Chemometrics. Vol. 8. pp. 169-174.

De Vries, S. and C.J. Ter Braak. (1995) "Prediction error in partial least squares regression: a critique on the deviation used in the Unscrambler". Chemometrics and Intelligent Laboratory Systems. Vol. 30. pp. 239-245.

Faber, Nicolaas. (2000) "Comparison of two recently proposed expressions for partial least squares regression prediction error". Chemometrics and Intelligent Laboratory Systems. Vol. 52. pp. 123-134.

Geladi, P. and B.R. Kowalski. (1986) "Partial Least Squares Regression: A Tutorial". Analytica Chemica Acta. Vol. 85. pp. 1-17.

Ho, P., M.C.M Silva and T.A. Hogg. (2001) "Multiple imputation and maximum likelihood principal component analysis of incomplete multivariate data from a study of the ageing of port". Chemometrics and Intelligent Laboratory Systems. Vol. 55. pp. 1-11.

Hoskuldsson, Agnar. (1988) "PLS Regression Methods". Journal of Chemometrics. Vol. 2. pp. 211-228.

Hoy, Martin, Kay Steen and Harald Martens. (1998) "Review of partial least squares regression prediction error in Unscrambler". Chemometrics and Intelligent Laboratory Systems. Vol. 44. pp. 123-133.

Imhof, J.P. (1961) "Computing the distribution of quadratic forms in normal variables". Biometrika. Vol. 48. Nos. 3 and 4. pp. 419-426.

Jackson, J. Edward and Govind S. Mudholkar. (1979) "Control Procedures for Residuals Associated With Principal Component Analysis". Technometrics. Vol. 21. No. 3. pp. 341-349.

Jackson, J. Edward. (1991) A User's Guide to Principal Components. John Wiley and Sons. Toronto, Ontario.

Jensen, D.R. and Herbert Solomon. (1972) "A Gaussian Approximation to the Distribution of a Definite Quadratic Form". Journal of the American Statistical Association. Vol. 67. No. 340. pp. 898-902.

Johnson, R.A. and D.W. Wichern. (1988) Applied Multivariate Statistical Analysis. Prentice Hall. Englewood Cliffs, New Jersey.

Kotz, Samuel, N.L. Johnson and D.W. Boyd. (1967) "Series Representations of Distributions of Quadratic Forms in Normal Variables. I. Central Case II. Non-Central Case". Annals of Mathematical Statistics. Vol. 38. pp. 823-848.

Kourti, Theodora and John F. MacGregor. (1995) "Process analysis, monitoring and diagnosis, using multivariate projection methods". Chemometrics and Intelligent Laboratory Systems. Vol. 28. pp 3-21.

Kourti, Theodora and John F. MacGregor. (1996) "Mutivariate SPC Methods for Process and Product Monitoring". Journal of Quality Technology. Vol. 28. No. 4. pp. 409-420.

Kresta, J.V., J.F. MacGregor and T.E. Marlin. (1991) "Multivariate Statistical Monitoring of Process Operating Performance". Canadian Journal of Chemical Engineering. Vol. 69. pp 35-47.

Kresta, J.V., T.E. Marlin and J.F. MacGregor. (1994) "Development of Inferential Process Models using Partial Least Squares". Computers and Chemical Engineering. Vol. 18. No. 7. pp. 597-611

Little, R.J.A. and D.B. Rubin. (1987) Statistical Analysis with Missing Data. John Wiley and Sons. New York.

MACSTAT. (1995) McMaster Advanced Control Consortium. McMaster University. Hamilton, Ontario, Canada. Version 4.3.

Martens, H., and T. Naes. (1989) Multivariate Calibration, John Wiley and Sons. New York. pp. 159.

Medjell, T. and S. Skogestad. (1991) "Estimation of Distillation Compositions from Multiple Temperature Measurements Using Partial-Least-Squares Regression". Ind. Eng. Chem. Res. Vol. 30. pp. 2543-2555.

Nelson, P.R.C., P.A. Taylor and J.F. MacGregor. (1996) "Missing Data Methods in PCA and PLS: Score Calculations with Incomplete Observations". Journal of Chemometrics and Intelligent Laboratory Systems. Vol. 35. No.1. pp. 45-65.

Nomikos, P. and J.F. MacGregor. (1994) "Monitoring of Batch Processes using Multiway Principal Components Analysis". AIChE Journal. Vol. 40. pp. 1361-1375.

Nomikos, P. and J.F. MacGregor. (1995) "Multivariate SPC Charts for Monitoring Batch Processes" Technometrics. Vol. 37, no. 1. pp. 41-59.

Phatak, A., P. M. Reilly, and A. Penlidis. (1993) "An Approach to Interval Estimation in Partial Least Squares Regression". Analytica Chimica Acta. Vol. 277. pp. 495-501.

Rannar, Stefan, Fredrik Lindgren, Paul Geladi and Svante Wold. (1994a) "A PLS Kernel Algorithm for Data Sets with Many Variables and Fewer Objects. Part I: Theory and Algorithm". Journal of Chemometrics. Vol. 8. pp. 111-125.

Rannar, Stefan, Paul Geladi, Fredrik Lindgren and Svante Wold. (1994b) "A PLS Kernel Algorithm for Data Sets with Many Variables and Fewer Objects. Part II: Cross-Validation, Missing Data and Examples". Journal of Chemometrics. Vol. 9. pp. 459-470.

Schafer, J. L. (1997) Analysis of Incomplete Multivariate Data. Chapman and Hall. London, England.

Skagerberg, B., J.F. MacGregor and C. Kiparissides. (1992) "Multivariate Data Analysis Applied to Low-Density Polyethylene Reactors". Chemometrics and Intelligent Laboratory Systems. Vol. 14. pp. 341-356.

Wentzell, Peter D., Darren T. Andrews, David C. Hamilton, Klaas Faber and Bruce R. Kowalski. (1997) "Maximum Likelihood Principal Components Analysis". Journal of Chemometrics. Vol. 11. pp 339-366.

Wentzell, Peter D. and Mitchell T. Lohnes. (1999) "Maximum likelihood principal component analysis with correlated measurement errors: theoretical and practical considerations". Chemometrics and Intelligent Laboratory Systems. Vol. 45. pp. 65-85.

Wise, B.M., D. J. Veltkamp, N. L. Ricker, B. R. Kowalski, S. M. Barnes and V. Arakali. (1991) "Application of Multivariate Statistical Process Control (MSPC) to the West Valley Slurry-Fed Ceramic Melter Process". Waste Management '91 Proceedings. Tucson, Arizona.

Wise, B.M. and N.L. Ricker. (1991) "Recent Advances in Multivariate Statistical Process Control: Improving Robustness and Sensitivity". ADCHEM '91, IFAC International Symposium. Toulouse, France. Oct 14-16, 1991. pp. 125-130.

Wold, Herman. (1966) "Estimation of Principal Components and Related Models by Iterative Least Squares" in Multivariate Analysis. P.R. Krishnaiah (Editor). Academic Press. New York. pp. 391-420.

Wold, S., C. Albano, W.J. Dunn III, K. Esbensen, S. Hellberg, E. Johansson and M. Sjostrom. (1983) "Pattern Recognition: Finding and Using Regularities in Multivariate Data" in Food Research and Data Analysis, H. Martens and H. Russwurm, Jr. (Editors). Applied Science Publishers. London and New York. pp. 183-185.

Wold, S., K. Esbensen and P. Geladi. (1987) "Principal Component Analysis". Chemometrics and Intelligent Laboratory Systems. Vol. 2. pp. 37-52.

Wold, S. (1995) Personal communication.

Appendix 1: Transformation of the General Quadratic Form to Standard Form

The general quadratic form $\mathbf{x}^T\mathbf{A}\mathbf{x}$ is used in the thesis where $\mathbf{x} \sim \mathrm{N}(\boldsymbol{\delta}, \boldsymbol{\Sigma})$ and $\mathbf{A}$ is symmetric but both A and $\boldsymbol{\Sigma}$ may be rank deficient. It is desired to have the quadratic in the standard form $Q = \sum_{k=1}^{K} c_k \{x_k + a_k\}^2$ with all $c_k > 0$ and $\mathbf{x} \sim N(\mathbf{0}, \mathbf{I})$. This is equivalent to a general form with $\mathbf{A}$ diagonal and positive definite and $\boldsymbol{\Sigma}$ equal to the identity matrix. This appendix will show how to transform a general quadratic into this form in two steps. The first step is necessary only if $\mathbf{A}$ is not full rank and involves finding an equivalent, lower dimensional quadratic with a full-rank $\mathbf{A}$ matrix. The second step obtains a quadratic in standard form from a general quadratic with a full-rank $\mathbf{A}$ matrix.

Step 1

Since A is symmetric, there is a $\mathbf{U}$ with $\mathbf{U}^T\mathbf{U} = \mathbf{U}\mathbf{U}^T = \mathbf{I}$ such that $\mathbf{A} = \mathbf{U}\mathbf{S}\mathbf{U}^T$ with $\mathbf{S}$ being diagonal with the eigenvalues of $\mathbf{A}$ in descending order on the diagonal. If $\mathbf{A}$ is not full rank then $\mathbf{S} = \begin{bmatrix} \mathbf{S}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$ with $\mathbf{S}_{11}$ having dimension equal to the number of non-zero eigenvalues of $\mathbf{A}$ and thus being full rank. The matrix $\mathbf{U}$ can be partitioned $\mathbf{U} = [\mathbf{U}_1, \ \mathbf{U}_2]$ with $\mathbf{U}_1$ having the same number of columns as $\mathbf{S}_{11}$. Substituting this into the general form of the quadratic

$$\begin{aligned}\mathbf{x}^T\mathbf{A}\mathbf{x} &= \mathbf{x}^T\mathbf{U}\mathbf{S}\mathbf{U}^T\mathbf{x} \\ &= \mathbf{x}^T[\mathbf{U}_1,\ \mathbf{U}_2]\begin{bmatrix}\mathbf{S}_{11} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}[\mathbf{U}_1,\ \mathbf{U}_2]^T\mathbf{x} \\ &= \mathbf{x}^T\mathbf{U}_1\mathbf{S}_{11}\mathbf{U}_1^T\mathbf{x} \\ &= \mathbf{y}^T\mathbf{H}\mathbf{y}\end{aligned}$$

an equivalent lower dimensional quadratic with $\mathbf{y} = \mathbf{U}_1^T\mathbf{x}$, $mean(\mathbf{y}) = \mathbf{U}_1\boldsymbol{\delta}$, $\text{var}(\mathbf{y}) = \mathbf{U}_1\boldsymbol{\Sigma}\mathbf{U}_1$ and $\mathbf{H} = \mathbf{S}_{11}$ is obtained.

Step 2

First the covariance matrix must be diagonalized in a similar manner to the weighting matrix, **A**, above so $\boldsymbol{\Sigma} = \mathbf{U}\mathbf{S}\mathbf{U}^T$ with $\mathbf{U}^T\mathbf{U} = \mathbf{U}\mathbf{U}^T = \mathbf{I}$ and **S** being diagonal with the eigenvalues of $\boldsymbol{\Sigma}$ in descending order on the diagonal. Now create **D** with $d_{ik} = s_{ik}^{1/2}$ so that $\mathbf{D}^2 = \mathbf{S}$. Partition **D** with $\mathbf{D} = \begin{bmatrix}\mathbf{D}_{11} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}$ so $\mathbf{D}_{11}$ has dimension equal to the number of non-zero eigenvalues of **D** and is thus full rank. Set $\mathbf{y} = \mathbf{D}_{11}^{-1}\mathbf{U}_1^T[\mathbf{x} - \boldsymbol{\delta}]$ so that $mean(\mathbf{y}) = \mathbf{0}$, $\text{var}(\mathbf{y}) = \mathbf{D}_{11}^{-1}\mathbf{U}_1\boldsymbol{\Sigma}\mathbf{U}_1^T\mathbf{D}_{11}^{-1} = \mathbf{D}_{11}^{-1}\mathbf{S}\mathbf{D}_{11}^{-1} = \mathbf{I}$ and $\mathbf{x} - \boldsymbol{\delta} = \mathbf{U}_1\mathbf{D}_{11}\mathbf{y} + \mathbf{U}_2\mathbf{0} = [\mathbf{U}_1\mathbf{D}_{11},\ \mathbf{U}_2]\begin{bmatrix}\mathbf{y}\\ \mathbf{0}\end{bmatrix}$. So

$$\begin{aligned}\mathbf{x} &= \mathbf{x} - \boldsymbol{\delta} + \boldsymbol{\delta} \\ &= [\mathbf{U}_1\mathbf{D}_{11},\ \mathbf{U}_2]\begin{bmatrix}\mathbf{y}\\ \mathbf{0}\end{bmatrix} + [\mathbf{U}_1\mathbf{D}_{11},\ \mathbf{U}_2]\begin{bmatrix}\mathbf{D}_{11}^{-1}\mathbf{U}_1^T\\ \mathbf{U}_2^T\end{bmatrix}\boldsymbol{\delta} \\ &= [\mathbf{U}_1\mathbf{D}_{11},\ \mathbf{U}_2]\left[\begin{bmatrix}\mathbf{y}\\ \mathbf{0}\end{bmatrix} + \begin{bmatrix}\mathbf{D}_{11}^{-1}\mathbf{U}_1^T\\ \mathbf{U}_2^T\end{bmatrix}\boldsymbol{\delta}\right] \\ &= [\mathbf{U}_1\mathbf{D}_{11},\ \mathbf{U}_2]\left[\begin{bmatrix}\mathbf{y}\\ \mathbf{0}\end{bmatrix} + \boldsymbol{\xi}\right]\end{aligned}$$

with $\xi=\begin{bmatrix} \mathbf{D}_{11}^{-1}\mathbf{U}_1^T \\ \mathbf{U}_2^T \end{bmatrix}\delta$. Now set $\mathbf{B}=[\mathbf{U}_1\mathbf{D}_{11},\ \mathbf{U}_2]^T\mathbf{A}[\mathbf{U}_1\mathbf{D}_{11},\ \mathbf{U}_2]=\begin{bmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{21} & \mathbf{B}_{22} \end{bmatrix}$ with $\mathbf{B}_{11}=\mathbf{D}_{11}^T\mathbf{U}_1\mathbf{A}\mathbf{U}_1^T\mathbf{D}_{11}$, substitute into the quadratic and complete the square.

$$
\begin{aligned}
\mathbf{x}^T\mathbf{A}\mathbf{x} &= \left[\begin{bmatrix}\mathbf{y}\\ \mathbf{0}\end{bmatrix}+\xi\right]^T[\mathbf{U}_1\mathbf{D}_{11},\ \mathbf{U}_2]^T\mathbf{A}[\mathbf{U}_1\mathbf{D}_{11},\ \mathbf{U}_2]\left[\begin{bmatrix}\mathbf{y}\\ \mathbf{0}\end{bmatrix}+\xi\right] \\
&= \left[\begin{bmatrix}\mathbf{y}\\ \mathbf{0}\end{bmatrix}+\xi\right]^T\mathbf{B}\left[\begin{bmatrix}\mathbf{y}\\ \mathbf{0}\end{bmatrix}+\xi\right] \\
&= \left[\begin{bmatrix}\mathbf{y}\\ \mathbf{0}\end{bmatrix}+\begin{bmatrix}\xi_1\\ \xi_2\end{bmatrix}\right]^T\begin{bmatrix}\mathbf{B}_{11} & \mathbf{B}_{12}\\ \mathbf{B}_{21} & \mathbf{B}_{22}\end{bmatrix}\left[\begin{bmatrix}\mathbf{y}\\ \mathbf{0}\end{bmatrix}+\begin{bmatrix}\xi_1\\ \xi_2\end{bmatrix}\right] \\
&= \left[[\mathbf{y}+\xi_1]^T,\ \xi_2^T\right]\begin{bmatrix}\mathbf{B}_{11} & \mathbf{B}_{12}\\ \mathbf{B}_{21} & \mathbf{B}_{22}\end{bmatrix}\begin{bmatrix}\mathbf{y}+\xi_1\\ \xi_2\end{bmatrix} \\
&= [\mathbf{y}+\xi_1]^T\mathbf{B}_{11}[\mathbf{y}+\xi_1]+2[\mathbf{y}+\xi_1]^T\mathbf{B}_{12}\xi_2+\xi_2^T\mathbf{B}_{22}\xi_2 \\
&= [\mathbf{y}+\xi_1]^T\mathbf{B}_{11}[\mathbf{y}+\xi_1]+2[\mathbf{y}+\xi_1]^T\mathbf{B}_{11}\mathbf{B}_{11}^{-1}\mathbf{B}_{12}\xi_2+\xi_2^T\mathbf{B}_{22}\xi_2 \\
&\quad +\xi_2^T\mathbf{B}_{12}^T\left(\mathbf{B}_{11}^{-1}\right)^T\mathbf{B}_{11}\mathbf{B}_{11}^{-1}\mathbf{B}_{12}\xi_2-\xi_2^T\mathbf{B}_{12}^T\mathbf{B}_{11}^{-1}\mathbf{B}_{11}\mathbf{B}_{11}^{-1}\mathbf{B}_{12}\xi_2+\xi_2^T\mathbf{B}_{22}\xi_2 \\
&= [\mathbf{y}+\xi_1+\mathbf{B}_{11}^{-1}\mathbf{B}_{12}\xi_2]^T\mathbf{B}_{11}[\mathbf{y}+\xi_1+\mathbf{B}_{11}^{-1}\mathbf{B}_{12}\xi_2] \\
&\quad -\xi_2^T\mathbf{B}_{12}^T\mathbf{B}_{11}^{-1}\mathbf{B}_{12}\xi_2+\xi_2^T\mathbf{B}_{22}\xi_2 \\
&= [\mathbf{y}+\xi_1+\mathbf{B}_{11}^{-1}\mathbf{B}_{12}\xi_2]^T\mathbf{H}\mathbf{G}\mathbf{H}^T[\mathbf{y}+\xi_1+\mathbf{B}_{11}^{-1}\mathbf{B}_{12}\xi_2] \\
&\quad -\xi_2^T\mathbf{B}_{12}^T\mathbf{B}_{11}^{-1}\mathbf{B}_{12}\xi_2+\xi_2^T\mathbf{B}_{22}\xi_2 \\
&= [\mathbf{H}^T\mathbf{y}+\mathbf{H}^T\xi_1+\mathbf{H}^T\mathbf{B}_{11}^{-1}\mathbf{B}_{12}\xi_2]^T\mathbf{G}[\mathbf{H}^T\mathbf{y}+\mathbf{H}^T\xi_1+\mathbf{H}^T\mathbf{B}_{11}^{-1}\mathbf{B}_{12}\xi_2] \\
&\quad -\xi_2^T\mathbf{B}_{12}^T\mathbf{B}_{11}^{-1}\mathbf{B}_{12}\xi_2+\xi_2^T\mathbf{B}_{22}\xi_2 \\
&= [\mathbf{z}+\mathbf{c}]^T\mathbf{G}[\mathbf{z}+\mathbf{c}]-\xi_2^T\mathbf{B}_{12}^T\mathbf{B}_{11}^{-1}\mathbf{B}_{12}\xi_2+\xi_2^T\mathbf{B}_{22}\xi_2
\end{aligned}
$$

with $\mathbf{B}_{11}$ diagonalized as $\mathbf{B}_{11}=\mathbf{H}\mathbf{G}\mathbf{H}^T$ and $\mathbf{z}=\mathbf{H}^T\mathbf{y}$. The term $[\mathbf{z}+\mathbf{c}]^T\mathbf{G}[\mathbf{z}+\mathbf{c}]$ is a quadratic in standard form since $\mathbf{G}$ is diagonal, $mean(\mathbf{z})=\mathbf{H}^T mean(\mathbf{y})=\mathbf{0}$ and $\mathrm{var}(\mathbf{z})=\mathbf{H}^T\mathrm{var}(\mathbf{y})\mathbf{H}=\mathbf{H}^T\mathbf{H}=\mathbf{I}$. The second and third terms are constants.

Appendix 2: Glossary

Term	Section of definition or first use	Definition (if an acronym)
CMR	2.4	conditional mean replacement
contribution	1.4.2	
EM	1.6	expectation maximisation
Hotelling T^2	1.4.2.2	
loading	1.3	
MAW	2.5.1	methanol- acetone-water
MLPCA	1.5.1.4	maximum likelihood principal components analysis
NIPALS	1.5.1.2	nonlinear iterative partial least squares
OLS	2.3.1	ordinary least squares
quadratic form	1.7	
PCA	1.3	principal components analysis
PCR	2.3.1	principal components regression
PLS	1.3	projection to latent

Term	Section of definition or first use	Definition (if an acronym)
		structures (also known as partial least squares)
score	1.3	
SCP	1.5.2.1	single component projection
SPE	1.4.2.1	squared prediction error
UCZAA	2.5.2	upper cook zone active alkali