Abstract

Principal Component Analysis (PCA) is widely used in the modern process industry in order to develop models from data sets with large numbers of highly correlated variables, registered on-line by means of sensors hooked up to continuous and batch processes.

Once a PCA model has been built from a set of observations obtained in normal operating conditions, it can be used in Multivariate Statistical Process Control (MSPC) schemes to monitor and diagnose future process operating performance.

In this context missing measurements are a common occurrence due to several causes: sensor failure, sensor routine maintenance, samples not collected at the required times, data discarded by gross measurement errors, sensors with different sampling periods and network transmission failures in highly automated environments. In batch process monitoring at each time t some method is needed to fill in the future (unknown) data corresponding to the period between time t and the end of the batch.

In MSPC two problems related with missing data appears: building PCA models from data sets with missing measurements, and using PCA models for monitoring future observations with missing information, assuming the estimated model to be fixed and known. The first problem has been broadly treated in the literature, this is not the case for the second problem which has been barely developed thoroughly in the literature.

This Thesis faces the second problem, studying two different and fundamental MSPC issues when future observations to be monitored have missing data: how to estimate scores from the PCA model for the new incomplete observation vector, and how to adapt the PCA model-based multivariate monitoring schemes.

Several methods for estimating scores from new data with missing measurements are presented. First of all, the simplest method is introduced, which consists of substituting the missing data for its unconditional mean value (the average is zero for all the variables, because it is assumed that the data have already been centred and scaled to unit variance). This is a simple imputation method and the estimation obtained is the so-called trimmed scores vector (TRI). The second method presented consists of adapting the NIPALS algorithm, used in the construction of PCA models, when there are missing data, to the case when model already exists. The method consists of calculating sequentially the scores, as it is done in the NIPALS algorithm, but not in an iterative way, and it is called Single Component Projection method (SCP). The third method analysed is the Projection to the Model Plane (PMP). Rather that obtaining the scores sequentially (as SCP does) the PMP

is a projection method for obtaining all the scores estimates at once. It is shown that this method is equivalent to two new methods: the first one is a simple imputation method that allows to obtain the scores as a result of the convergence of an iterative process, and the second one is based on replacing the missing data by the values having the smallest squared prediction error (*SPE*) when the pre-built PCA model is applied.

Afterwards, several methods based on regression models are introduced. The first one consists of using the already known data to predict the missing ones by using a regression model on the observations used in the construction of the PCA model (training data set). This is the Known Data Regression method (KDR) and it is shown to be equivalent to the so-called Conditional Mean Replacement method (CMR) that consists of replacing the missing data by their conditional means, given the measured variables and the estimate of the means and covariance matrix of the training data set, assuming multivariate normality. The second method based on regression models is similar to the KDR method, although the regression is done on the trimmed score estimators (TRI). This method is called Trimmed Scores Regression (TSR).

KDR and TSR methods are both particular cases of a more general method that includes, also, two approaches to the KDR method that are the outcome of replacing the least squares regression by Principal Component Regression (PCR) and Partial Least Squares Regression (PLS), respectively.

Some of the methods discussed in this Thesis have already been proposed by other authors in the bibliography (SCP, PMP, and CMR), some of them are original (TRI and TSR) and others are shown to be equivalent to methods already developed by other authors: iterative imputation methods and minimizing the SPE are equivalent to the PMP method; KDR is equivalent to CMR.

The basis for each method and the expressions for the score estimators, estimation errors and their covariance matrices are developed. The efficiency of the methods is studied through simulations based on several industrial and simulated data sets. Squared estimation errors are used as the basis for comparisons.

KDR and TSR methods are shown to be statistically superior to the other ones studied. Although TSR is slightly less efficient than KDR, it shows remarkable algorithmic advantages leading to lower computational costs (the matrix to be inverted is of a much smaller size). This is specially convenient when dealing with data sets with large number of highly correlated variables.

Finally, in this Thesis two different approaches for adapting PCA model-based MSPC techniques and tools when future observations have missing data are studied and compared: the first one, based on the CMR method, has already been proposed in the literature; the second one, based on the TSR method, is a novel method. In both cases the uncertainty that missing data add to every monitoring statistic: residuals, SPE, scores, Hotelling T^2 , and also the process variables contributions to SPE and scores, and the scores contributions on Hotelling T^2, is studied. Each one of these statistics is characterised by a probability distribution, from which the associated uncertainty region is estimated. Useful diagnostic procedures for determining which are the missing variables that most contribute to that uncertainty are defined.

The performance of both approaches is studied through simulations based on industrial and simulated data sets. Both methods give practically identical results, the novel approach (based on TSR estimation method) being computationally simpler.