

SUBSPACE METHOD IDENTIFICATION FOR DYNAMIC MULTIVARIATE STATISTICAL PROCESS CONTROL

¹Richard Treasure, ²Uwe Kruger and ¹Victor Sreeram

1. Control Systems Research Group, University of Western Australia, Nedlands, 6009. Australia.
2. Intelligent Systems and Control Research Group, Queens University of Belfast, BT5 9AH, UK

Abstract: Subspace method identification (SMI) and model reduction for Multivariate Statistical Process Control has been proposed as an improvement to dynamic principal component analysis (DPCA). The linear parametric model structure captures both static and dynamic information from the system. In this paper, an analysis of the dimension reduction capabilities of the subspace approach is provided. It is proven that the SMI method yields a parsimonious model structure that requires fewer latent variables and uses fewer process measurements than DPCA. These findings are illustrated by an industrial application study. Copyright © 2005 IFAC.

Keywords: model reduction, fault detection, multivariate quality control, process control, identification, subspace methods.

1. INTRODUCTION

Process condition monitoring is a technique for the detection and diagnosis of abnormal process behavior in a wide range of industries. It has the potential to improve bottom line returns, improve process quality, and reduce the energy required per unit of product (Brisk, 2004). This has led to the application of various statistically based condition monitoring strategies, collectively referred to as multivariate statistical process control (MSPC) (MacGregor and Kourti, 1995; Kourti and MacGregor, 1996). The most commonly used approaches have involved Principal Component Analysis (PCA) (Martin and Morris, 1996).

PCA (Wold *et al.*, 1987; Wise *et al.*, 1990; Kresta *et al.*, 1991) uses a static process model to monitor a reduced set of statistically uncorrelated variables. It is generally assumed that the process is operating at a predefined steady state operating point, however it is not uncommon for process variables to be affected by controller feedback and dynamic transients. Process

variables may then move away from steady state conditions and exhibit some degree of auto-correlation (Ku *et al.*, 1995).

Subspace method identification (SMI) is a dynamic extension for MSPC, which employs a state space model structure to incorporate the dynamic effects of auto- and cross-correlated process variables into the model (Treasure *et al.*, 2004). This technique is an alternative to the more widely known DPCA. The SMI method is based on the N4SID algorithm (Van Overschee and De Moor, 1994), and error-in-variables (EIV), to generate a set of state-variables that can describe dynamic process data. An MSPC framework is used to define Hotelling's T^2 and Q statistics.

In this paper, a review of the SMI method (Treasure *et al.*, 2004) and DPCA (Wise and Ricker, 1992) procedures is provided and then building on from this work, there is an analysis of the respective model structures. It is proven that the SMI method is able to provide a process monitor using fewer latent

variables and fewer measured variables than DPCA. An industrial application study is presented to demonstrate that the SMI method provides a unique analysis, providing the same information as DPCA using fewer latent variables and fewer measured variables.

2. DYNAMIC MONITORING MODELS

Section 2 presents an outline of the SMI method and DPCA as described in Treasure *et al* (2004).

2.1 SMI Algorithm

The SMI algorithm calculates a state sequence $\mathbf{X}_k = (\mathbf{x}_k \ \mathbf{x}_{k+1} \ \mathbf{x}_{k+2} \ \dots \ \mathbf{x}_{k+N}) \in R^{n \times N}$ based on an n^{th} order state space model:

$$\begin{pmatrix} \mathbf{x}_{k+1} \\ \mathbf{y}_k \end{pmatrix} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix} \begin{pmatrix} \mathbf{x}_k \\ \mathbf{u}_k \end{pmatrix} \quad (1)$$

$\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}$ are matrices of appropriate dimension and \mathbf{y}_k and \mathbf{u}_k are the process output/input measurements. An error-in-variables approach is applied to (1) to calculate the system parameters. This is equivalent to finding the total least squares solution to the state equations (1).

Calculating the state sequence. A full description of the algorithm used to calculate a state sequence for the process is given in (Van Overschee and De Moor, 1992). The identification procedure involves block Hankel matrices, $\mathbf{Y}_p, \mathbf{Y}_f, \mathbf{U}_p, \mathbf{U}_f$, constructed as follows:

$$\mathbf{Y}_p = \begin{bmatrix} \mathbf{y}_1 & \dots & \mathbf{y}_N \\ \vdots & & \vdots \\ \mathbf{y}_j & \dots & \mathbf{y}_{N+j-1} \end{bmatrix}; \quad \mathbf{Y}_f = \begin{bmatrix} \mathbf{y}_{j+1} & \dots & \mathbf{y}_{N+j} \\ \vdots & & \vdots \\ \mathbf{y}_{2j} & \dots & \mathbf{y}_{N+2j-1} \end{bmatrix}$$

and similarly for \mathbf{U}_p and \mathbf{U}_f , where the initial prediction of \mathbf{Y}_f is the least squares solution to

$$\mathbf{Y}_f = [\mathbf{R}_1 \ \mathbf{R}_2 \ \mathbf{R}_3] \begin{bmatrix} \mathbf{U}_p \\ \mathbf{Y}_p \\ \mathbf{U}_f \end{bmatrix} + \text{Residuals} \quad (2)$$

The state sequence is then calculated from

$$\Gamma \mathbf{X}_k = \mathbf{R}_1 \mathbf{U}_p + \mathbf{R}_2 \mathbf{Y}_p \quad (3)$$

where $\Gamma = [\mathbf{C}^T \ (\mathbf{CA})^T \ \dots \ (\mathbf{CA}^{j-1})^T]^T$ represents the extended observability matrix and \mathbf{X}_k is the identified states sequence. The final model is estimated by applying SVD to (3), then using

balanced truncation (Moore, 1981), where the estimated states are equivalent to the system states up to within a similarity transformation. \mathbf{X}_{k+1} is determined by adding incoming measurements to $\mathbf{U}_f, \mathbf{U}_p, \mathbf{Y}_f, \mathbf{Y}_p$ and recalculating (3).

2.2 Dynamic Principal Component Analysis(DPCA)

DPCA (Ku *et al.*, 1995) involves PCA of a data matrix that corresponds to an ARX model structure

$$\mathbf{Z}_{DPCA} = \begin{bmatrix} \mathbf{Y}_k^T & \dots & \mathbf{Y}_{k-n}^T & \mathbf{U}_k^T & \dots & \mathbf{U}_{k-n}^T \end{bmatrix}^T \quad (4)$$

$$\mathbf{Y}_k = \mathbf{B}_0 \mathbf{U}_k + \sum_{i=1}^n \mathbf{A}_i \mathbf{Y}_{k-i} + \mathbf{B}_i \mathbf{U}_{k-i} \quad (5)$$

where $\mathbf{B}_0, \mathbf{A}_i$ and \mathbf{B}_i are parameter matrices of appropriate dimension and n is the number of time lags incorporated into the model.

2.3 Statistics for the process condition monitors

Multivariate statistics are developed on the basis of the PCA decomposition

$$\mathbf{Z} = \mathbf{TP}^T + \mathbf{E} \quad (6)$$

$\hat{\mathbf{Z}} = \mathbf{TP}^T$ is the model of significant process variation within the recorded process variables and \mathbf{E} contains the residuals which are statistically independent to $\hat{\mathbf{Z}}$. If \mathbf{P}^T is the ‘‘in control’’ model based on normal process operation, the co-ordinates of the k^{th} score of the monitored process are then determined as:

$$\mathbf{t}_k = \mathbf{P}^T \mathbf{z}_k \quad (7)$$

where \mathbf{z}_k is the k^{th} column vector in \mathbf{Z}^T :

$$\text{SMI } \mathbf{z}_k = \left(\mathbf{y}_k^T \ \mathbf{u}_k^T \ \mathbf{x}_k^T \ \mathbf{x}_{k+1}^T \right)^T \quad (8)$$

$$\text{DPCA } \mathbf{z}_k = \left(\mathbf{y}_k^T \ \dots \ \mathbf{y}_{k-n}^T \ \mathbf{u}_k^T \ \dots \ \mathbf{u}_{k-n}^T \right)^T \quad (9)$$

The residuals of each model are used to calculate the prediction error, or mismatch, between the measured and reconstructed process variables:

$$\mathbf{e}_k = \mathbf{z}_k - \mathbf{P} \mathbf{t}_k \quad (10)$$

Given the above definitions for \mathbf{t}_k and \mathbf{e}_k , Hotelling’s T^2 statistic is calculated as

$$T_k^2 = \mathbf{t}_k^T \mathbf{\Lambda}^{-1} \mathbf{t}_k \quad (11)$$

where $\mathbf{\Lambda}$ is a diagonal matrix storing the variance of each of the columns of \mathbf{T} , as defined by the reference data set. The Q statistic is calculated as

$$Q_k = \mathbf{e}_k^T \mathbf{e}_k \quad (12)$$

The univariate statistics (T^2 and Q) are plotted against time. The confidence limits for the T^2 and Q statistics are calculated as defined in (Jackson and Mudholkar, 1979; Jackson, 1980). A contributions analysis is used to diagnose the likely source of process faults. Contributions of individual process measurements to the T^2 and Q statistics are obtained using the following expressions:

$$C_{T^2} = \mathbf{P}\mathbf{\Lambda}^{-1}\mathbf{P}^T \mathbf{z} \quad (13)$$

$$C_Q = [\mathbf{I} - \mathbf{P}\mathbf{P}^T] \mathbf{z} \quad (14)$$

where \mathbf{I} is an identity matrix of appropriate dimension.

3. MODEL STRUCTURES

In this section, it is assumed that the system under study can be described by a state space model. A proof is provided that the SMI method uses fewer measurement variables and fewer latent variables than DPCA. This is followed by an analysis of the consistency of the state space estimate.

3.1 SMI algorithm compared to DPCA

Subject to the conditions imposed on the measurement uncertainty, the model structures employed by the SMI method and DPCA are equivalent. For simplification, a SISO system in state space form is studied to demonstrate that the SMI model requires fewer latent variables than DPCA. Consider the following system:

$$\begin{aligned} x_{k+1} &= ax_k + bu_k + e_{k+1} \\ y_k &= cx_k + du_k + f_k \end{aligned} \quad (15)$$

where e_k and f_k are independent, zero mean, Gaussian distributed noise terms with variance σ_e and σ_f . The state space model (15) is equivalent to the following ARX model structure (Treasure, 2004):

$$y_k = ay_{k-1} + du_k + (bc - ad)u_{k-1} + g_k \quad (16)$$

where $g_k = f_k - af_{k-1} + ce_k$, and the variance of the error on y_k is $\text{var}(g_k) = (1-a)\sigma_f + c\sigma_e$. Assuming the input variables are also measured inaccurately, then for the SMI/DPCA model structures, the following EIV problems can be formulated:

$$\begin{pmatrix} y_k \\ u_k \\ x_k \\ x_{k+1} \end{pmatrix} = \begin{pmatrix} \hat{y}_k \\ \hat{u}_k \\ \hat{x}_k \\ \hat{x}_{k+1} \end{pmatrix} + \begin{pmatrix} f_k \\ h_k \\ e_k \\ e_{k+1} \end{pmatrix} \quad (17)$$

$$\begin{pmatrix} y_{k+1} \\ y_k \\ u_{k+1} \\ u_k \end{pmatrix} = \begin{pmatrix} \hat{y}_{k+1} \\ \hat{y}_k \\ \hat{u}_{k+1} \\ \hat{u}_k \end{pmatrix} + \begin{pmatrix} g_{k+1} \\ g_k \\ h_{k+1} \\ h_k \end{pmatrix} \quad (18)$$

Modern manufacturing plants routinely monitor hundreds of variables (Martin and Morris, 1996), often with high degrees of correlation, and therefore dimension reduction techniques are an important aspect of MSPC. Equations (17) and (18) can be represented as $\mathbf{Z} = \mathbf{P}\mathbf{T} + \mathbf{E}$, where h_k represents the measurement uncertainty of u_k . Although (17) and (18) contain the same number of process variables, a large-scale process with highly correlated process measurements admits to significant dimension reduction. In the case of large-scale MIMO processes, the number of internal variables used by DPCA is prohibitive. In contrast, the SMI algorithm provides considerable dimension reduction, where relatively few (orthogonal) state sequences are used in the model, in place of the time-lagged process variables used in DPCA.

The SMI method generally provides a condition monitor using fewer principal directions, as demonstrated in the following analysis. Ignoring the noise terms in (17), if the input is uncorrelated with the states, then a stable system is fully described by two principal directions where y_k and x_{k+1} are linear combinations of x_k and u_k

$$\begin{pmatrix} y_k \\ u_k \\ x_k \\ x_{k+1} \end{pmatrix} = \begin{pmatrix} c & d \\ 0 & 1 \\ 1 & 0 \\ a & b \end{pmatrix} \begin{pmatrix} x_k \\ u_k \end{pmatrix} \quad (19)$$

The sequences x_k and u_k lie in the subspace described by the principal directions. Equation 19 corresponds to $\mathbf{Z} = \mathbf{P}\mathbf{T}$, and indicates the maximum number of dimensions required by SMI analysis. For large-scale MIMO processes the maximum number of dimensions required is:

$$\dim_{\max} \leq \dim(\mathbf{u}) + \dim(\mathbf{x}) = m + n \quad (20)$$

Lemma 1

The upper bound for the number of dimensions required for SMI analysis of an n^{th} order MIMO process is $\dim_{\max} \leq m + n$.

Proof

Assuming the system is stable and the input sequence \mathbf{u}_k to be statistically independent of the state sequences, then the SMI method fully describes the process using two principal directions, i.e. both \mathbf{y}_k and \mathbf{x}_{k+1} are linear combinations of \mathbf{x}_k and \mathbf{u}_k :

$$\begin{pmatrix} \mathbf{y}_k \\ \mathbf{u}_k \\ \mathbf{x}_k \\ \mathbf{x}_{k+1} \end{pmatrix} = \begin{bmatrix} \mathbf{C} & \mathbf{D} \\ \mathbf{0} & \mathbf{I}_{m \times m} \\ \mathbf{I}_{n \times n} & \mathbf{0} \\ \mathbf{A} & \mathbf{B} \end{bmatrix} \begin{pmatrix} \mathbf{x}_k \\ \mathbf{u}_k \end{pmatrix} \quad (21)$$

The deterministic part of this system is estimated as $\hat{\mathbf{z}} = \mathbf{P}\mathbf{P}^T \mathbf{z}$, with coefficient matrix

$$\mathbf{P}\mathbf{P}^T = \begin{bmatrix} \mathbf{0} & \mathbf{D} & \mathbf{C} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{m \times m} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I}_{n \times n} & \mathbf{0} \\ \mathbf{0} & \mathbf{B} & \mathbf{A} & \mathbf{0} \end{bmatrix} \quad (22)$$

By inspection, the maximum number of dimensions required is $\dim_{\max} \leq \dim(\mathbf{u}) + \dim(\mathbf{x}) = m + n$, since column blocks 1 and 4 are padded with zeros. ■

Lemma 1 indicates that SMI requires $m + n = 2$ latent variables to fully describe the process (15). In contrast, DPCA requires $(m + p)n + m = 3$ latent variables to describe the same process (16), i.e.

$$\begin{pmatrix} y_{k+1} \\ y_k \\ u_{k+1} \\ u_k \end{pmatrix} = \begin{pmatrix} a & d & q \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} y_k \\ u_{k+1} \\ u_k \end{pmatrix} \quad (23)$$

where $q = bc - ad$, and y_k , u_{k+1} and u_k lie in the subspace defined by the three principal directions

Lemma 2

The upper bound for the number of dimensions required for DPCA of an n^{th} order MIMO process is

$$\dim_{\max} \leq (m + p)n + m$$

Proof

An n^{th} order DPCA model, as described by (9) is based on the ARX model structure

$$\mathbf{y}_k = \sum_{i=1}^n \mathbf{A}_i \mathbf{y}_{k-i} + \sum_{i=0}^n \mathbf{B}_i \mathbf{u}_{k-i} \quad (24)$$

It can be seen directly from (24) that for an n^{th} order ARX model with random input, the upper bound for the number of principal components required by DPCA is equal to

$$\dim_{\max} \leq n * \dim(\mathbf{y}) + (n + 1) * \dim(\mathbf{u}) = (m + p)n + m$$

Lemma 3

SMI uses $(m + p) + 2n$ variables to model the process which is fewer than the $(n + 1)(m + p)$ variables required by an equivalent order DPCA model.

Proof

The n^{th} order state space model (1) requires $(m + p) + 2n$ variables and the SMI method (8) also uses $(m + p) + 2n$ variables. The ARX model (24) uses $(m + p)n + m$ process measurements, and DPCA (9) requires $(n + 1)(m + p)$ variables. For a MIMO (or SISO) system with m inputs and p outputs, $m \geq 1$, $n \geq 1$, it follows directly that:

$$\begin{aligned} (m + p) &\geq 2, \\ \text{i.e. } n(m + p) &\geq 2n \\ \text{i.e. } (n + 1)(m + p) &\geq m + p + 2n \end{aligned}$$

4. APPLICATION STUDY

The distillation process is designed to purify Butane from a fresh feed containing a mixture of hydrocarbons, primarily Butane (C4), Hexane (C5) and some impurities of Propane (C3). A purified C4 stream leaves the distillation process as the top product while C5 and impurities of C3 leave the distillation process with the bottom draw. The product quality is measured by the C5 concentration in the top draw, which should be kept below a predefined limit. Furthermore, to achieve economic operation of the plant, it is desirable to maintain the C4 concentration of the bottom draw below a specified limit. The process is affected by frequent changes in the feed, since no control system is employed to regulate the top and bottom concentrations of C5 and C4. The input and output variables of the process are listed in Table 1. Significant drops in fresh feed leads to violations of the C5 and the C4 concentration limits. The aim of this study is to detect these drops. Note that the SMI method uses considerably fewer variables in the model than DPCA, since the tray temperatures and pressures inside the distillation tower are highly correlated. The state space model is 2nd leading to a total of 20 variables overall; and the DPCA model is first order, requiring 32 variables. The number of principal components required for each of the models was found using parallel analysis (Ku et al., 1995) as 4 principal components (the SMI method) and 6 principal components (DPCA).

Figures 1 and 2 show the T^2 and Q statistics for the SMI method and DPCA. It is clear that the results for

the SMI method compare favorably with those for DPCA. In each case the arrival of abnormal behavior is indicated by a clear violation of the confidence limits. Figures 3 and 4 show the Q statistic contributions for the SMI method and DPCA. In particular the column temperatures (M1, M3) are significantly larger than predicted (as expected from prior knowledge of the process dynamics). In addition, the fresh feed temperature and the fresh feed flow (A1 and A2) are contributing to this event. Other contributions are from the reflux flow (A3), and the concentrations.

Table 1: Process variables (Wang *et al.*, 2003)

VARIABLE	DESCRIPTION	
PREDICTOR VARIABLES	A1	Flow of fresh feed
	A2	Temperature of fresh feed
	A3	Reflux flow
	A4	Reboiler steam flow
RESPONSE VARIABLES	M1	Tray 14 temperature
	M2	Column overhead pressure
	M3	Tray 2 temperature
	M4	Reflux vessel level
	M5	Butane product flow
	M6	Percentage of C3 in C4
	M7	Percentage of C5 in C4
	M8	Tray 31 temperature
	M9	Reboiler vessel level
	M10	Bottom draw
	M11	Percentage of C4 in C5
	M12	Reboiler temperature

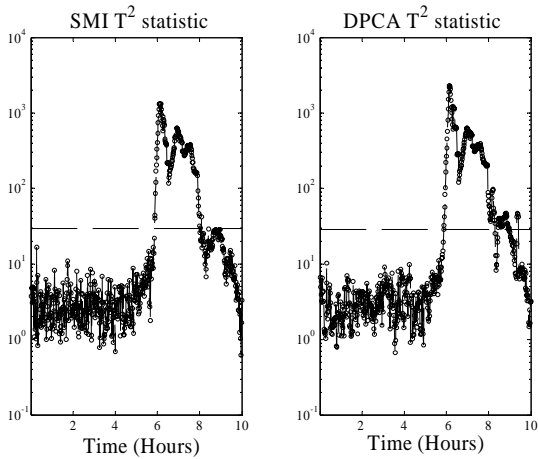


Fig. 1. Industrial Process data comparison of Hotelling's T^2 statistic the SMI method (left) and DPCA (right).

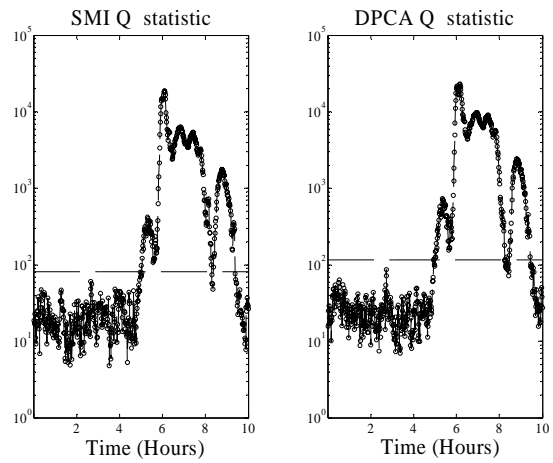


Fig. 2. Industrial process data comparison of Q statistic. The SMI method (left) and DPCA (right).

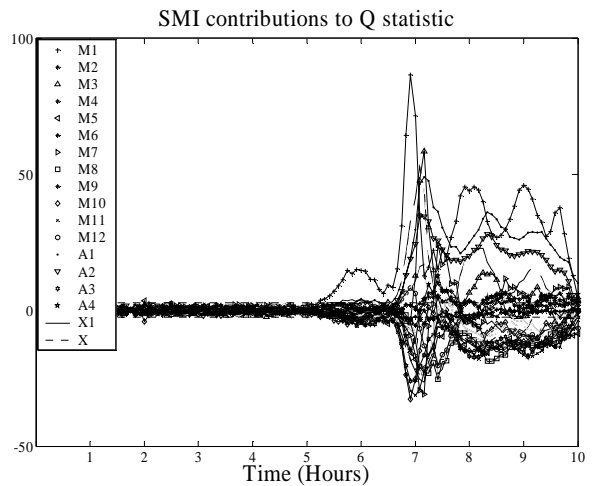


Fig. 3. Contributions to the Q statistic (SMI).

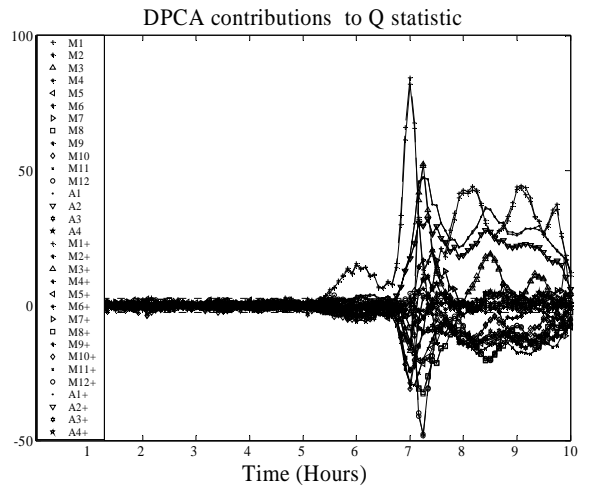


Fig. 4. Contributions to the Q statistic (DPCA).

5. CONCLUSION

It has been proven that the SMI method has the potential to provide a more concise model, using fewer latent variables than DPCA. In contrast, DPCA produces redundant information compared to the SMI method, as it requires more measurement variables and more latent variables. The simulation study demonstrates that the SMI method produces results that are consistent with DPCA, at the same time providing a less cluttered fault analysis than DPCA.

6. REFERENCES

- Brisk, M. L. (2004). Process Control: Potential Benefits and Wasted Opportunities. Plenary Keynote Paper. *The 5th Asian Control Conference*, Melbourne.
- Jackson, J. E. (1980). Principal Components and Factor Analysis: Part 1 - Principal Components. *Journal of Quality Control*, **12**, pp. 201-213.
- Jackson, J. E. and G. S. Mudholkar (1979). Control procedures for residuals associated with principal component analysis. *Technometrics*, **21**, pp. 341-349.
- Kourti, T. and J. F. MacGregor (1996). Multivariate SPC methods for process and product monitoring. *Journal of Quality Technology*, **28**, pp. 409-428.
- Kresta, J. V., J. F. MacGregor and T. E. Marlin (1991). Multivariate statistical monitoring of process operating performance. *Canadian Journal of Chemical Engineering*, **69**, pp. 35-47.
- Ku, W., R. H. Storer and C. Georgakis (1995). Disturbance rejection and isolation by dynamic principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, **30**, pp. 179-196.
- MacGregor, J. F. and T. Kourti (1995). Statistical process control of multivariate processes. *Control Engineering Practice*, **3**, pp. 403-414.
- Martin, E. B. and A. J. Morris (1996). An overview of multivariate statistical process control in continuous and batch process performance monitoring.
- Moore, B. C. (1981). Principal component analysis in linear systems - controllability, observability and model reduction. *IEEE Transactions on Automatic Control*, **AC-26**, pp. 17-32.
- Treasure, R. J. (2004). Subspace System Identification for Multivariate Statistical Process Control. *Engineering Doctorate Thesis*.
- Treasure, R. J., U. Kruger and J. E. Cooper (2004). Dynamic multivariate statistical process control using subspace identification. *Journal of Process Control*, **14**, pp. 279-292.
- Van Overschee, P. and B. De Moor (1992). Two subspace algorithms for the identification of combined deterministic-stochastic systems. *31st Conference on Decision and Control*, Tucson, Arizona.
- Van Overschee, P. and B. De Moor (1994). N4SID: Subspace algorithms for the identification of combined deterministic-stochastic systems. *Automatica*, **30**, pp. 75-93.
- Wang, X., U. Kruger and B. Lennox (2003). Recursive partial least squares algorithms for monitoring complex industrial processes. *Control Engineering Practice*, **11**, pp. 613-632.
- Wise, B. M. and N. L. Ricker (1992). Identification of Finite Impulse Response Models by Principal Components Regression: Frequency-Response Properties. *Process Control and Quality*, **4**, pp. 77-86.
- Wise, B. M., N. L. Ricker, D. F. Veltkamp and B. R. Kowalski (1990). Theoretical basis for the use of principal component models for monitoring multivariate processes. *Process Control and Quality*, **1**, pp. 41-51.
- Wold, S., K. Esbensen and P. Geladi (1987). Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, **2**, pp. 37-52.