Available online at www.sciencedirect.com

SCIENCE DIRECT

Chemometrics and Intelligent Laboratory Systems 80 (2006) 57 – 66

Chemometrics and intelligent laboratory systems

www.elsevier.com/locate/chemolab

Heteroscedastic latent variable modelling with applications to multivariate statistical process control

Marco S. Reis *, Pedro M. Saraiva

GEPSI-PSE Group, Department of Chemical Engineering, University of Coimbra, Pólo II-Pinhal de Marrocos, 3030-290 Coimbra, Portugal

Received 3 March 2005; received in revised form 13 July 2005; accepted 15 July 2005
Available online 18 August 2005

Abstract

We present an approach for conducting multivariate statistical process control (MSPC) in noisy environments, i.e., when the signal to noise ratio is low, and, furthermore, noise standard deviation (uncertainty) affecting each collected value can vary over time, and is assumingly known. This approach is based upon a latent variable model structure, HLV (standing for heteroscedastic latent variable model), that explicitly integrates information regarding data uncertainty. Moderate amounts of missing data can also be handled in a coherent and fully integrated way through HLV. Several examples show the added value achieved under noisy conditions by adopting such an approach and a case study illustrates its application to a real industrial context of pulp and paper product quality data analysis.

Keywords: Multivariate statistical process control; Measurement uncertainty; Latent variable modelling

1. Introduction

Wide streams of data are typically acquired and stored in modern industrial plants containing relevant and updated information about the status of the processes along time. Within the Statistical Process Control (SPC) framework, data are frequently fed to control charts, in order to decide whether it is operating under statistical control or if some special cause has interfered with it [1]. The normal operation conditions (NOC [2]) in control charts are set by analyzing data collected from periods of normal operation using Phase 1 methods [3] and, as long as the process rests within the NOC limits during Phase 2 implementation, no corrective actions should be taken. But, as soon as it moves outside of such boundaries, the root cause of abnormality should be identified and corrected, in such a way that the process is brought back to normal operation.

From what was stated above, we can see that measurements do play a central role in SPC. However, the growing availability and increasing complexity of measurement systems often raises practical problems that require adaptation of the classical SPC frameworks in order to accommodate for them. For instance, there was a certain tradition of using SPC charts in the supervision of single isolated variables, through univariate SPC charts, but it is now widely recognized that such a procedure is not effective when dealing with multivariate data exhibiting correlated behaviour [1]. Therefore, multivariate SPC (MSPC) procedures based on the Hotelling T^2 statistic were developed to redefine the shape of the multivariate NOC regions. As the number of variables to be monitored increases, even MSPC control charts begin to experiment some difficulties, and methodologies based on latent variables models [4], specially suited for environments where the underlying dimension of the process is (much) smaller than the dimensionality of data, were developed [2].

All of the above SPC approaches not only solve a particular problem raised by a given measurement data structure, but also rely on a particular assumed statistical description for the process common cause variability, from which statistical limits that define the NOC regions are derived. In this regard, current SPC methodologies based

* Corresponding author. Tel.: +351 239 798 700; fax: +351 239 798 703.
E-mail address: marco@eq.uc.pt (M.S. Reis).

0169-7439/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2005.07.002

on latent variables do not take explicitly into consideration measurement uncertainty information that is often available. As such, they do not explore this quite valuable knowledge, which is becoming even more so given the current trend towards the explicit consideration of data quality in all data analysis tasks, where "data quality" can be adequately expressed by the uncertainty figures associated with raw data values, a well defined quantity that can be evaluated following standard guidelines [5–7]. Some previous works can be referred regarding efforts undertaken in order to integrate measurement uncertainties in various data analysis contexts. Wentzell et al. [8] developed the so-called maximum likelihood principal component analysis (MLPCA), that estimates a PCA model in an optimal maximum likelihood sense when data are affected by measurement errors exhibiting complex structures, such as cross-correlations along sample or variable dimensions. The reasoning underlying MLPCA was then applied to multivariate calibration [9] extending the consideration of measurement uncertainties to some input/output modelling approaches closely related to PCA. Bro et al. [10] presented a general framework for integrating data uncertainties in the scope of (maximum likelihood) model estimation, comprehending MLPCA as a special case. The issue of (least squares) model estimation is also referred by Lira [7], along with the presentation of general expressions for uncertainty propagation in several input/output model structures. Both multivariate least-squares (MLS) and its univariate version, bivariate least-squares (BLS), were applied in several contexts of linear regression modelling, when all variables are subject to measurement errors with known uncertainties [11–13]. On the other hand, Faber and Kowalski [14] explicitly considered the influence of measurement errors in the calculation of confidence intervals for the parameters and predictions in PCR and PLS, and similar efforts can be found in [15–18]. These techniques provide us with new and more flexible tools, in the sense that they are applicable in more general measurement error structure contexts, including those not covered by the classical approaches. Therefore, as SPC frequently shares the same type of data sets as the above-referred methodologies, it is quite relevant to develop SPC procedures that explicitly take into account data uncertainties. In this paper, we present an approach that enables one to extend the use of well known control chart tools based on the T^2 and Q statistics [19,20] to such contexts, making explicit use of measurement uncertainty information that is available.

We present the statistical model on which our approach for integrating data uncertainties is based in the next section, and show how it can be properly estimated. Then, Section 3 provides a description of our MSPC procedure based on latent variables when measurements have heteroscedastic Gaussian behavior and, furthermore, shows how the proposed approach can easily handle missing data. In the following section, several examples are presented in order to illustrate the various features of the proposed approach, including a case study based upon real industrial data collected from a Pulp and Paper mill.

2. Underlying statistical model

We consider the fairly common situation where a large number of measurements are being collected and stored, coming from different devices and sources within the process and carrying important information about its current state. Quite often the underlying process phenomena, along with existing process constraints, do require a significantly lower dimensionality to be described than that arising from the consideration of all the variables. In fact, for monitoring purposes, we are only interested in following what happens around the subspace where the overall normal process variability is concentrated. Latent variable models do provide useful frameworks for modeling the relationships linking the whole set of measurements, arising from different sources, in terms of a few inner variability sources [4]. Therefore, let us consider the following latent variable multivariate linear relationship:

$$x(k) = \mu_x + Al(k) + \varepsilon_{\mathrm{m}}(k) \tag{2.1}$$

where x is the $n \times 1$ vector of measurements, μ_X is the $n \times 1$ mean vector of x, A is the $n \times p$ matrix of model coefficients, l is the $p \times 1$ vector of latent variables and ε_{m} is the $n \times 1$ vector of measurement noise. This model is completed by specifying the probability density functions relative to each random component:

$$\begin{aligned} &l(k) \sim iid\ N_{\mathrm{p}}(0, \Delta_{\mathrm{l}}) \\ &\varepsilon_{\mathrm{m}}(k) \sim id\ N_{\mathrm{n}}(0, \Delta_{\mathrm{m}}(k)) \\ &l(k) \text{ and } \varepsilon_{\mathrm{m}}(j) \text{ are independent } \forall k, j \end{aligned} \tag{2.2}$$

where N_{p} stands for the p-dimensional multivariate normal distribution, Δ_{l} is the covariance matrix for the latent variables (l), $\Delta_{\mathrm{m}}(k)$ is the covariance matrix of the measurement noise at time k ($\varepsilon_{\mathrm{m}}(k)$), given by $\Delta_{\mathrm{m}}(k) = \mathrm{diag}(\sigma^2_{\mathrm{m}}(k))$ ($\mathrm{diag}(u)$, represents a diagonal matrix with the elements of vector u along the main diagonal and $\sigma^2_{\mathrm{m}}(k)$ is the vector of error variances for all the measurements at time k), 0 is an array of appropriate dimension, with only zeros in its entries. Thus, Eqs. (2.1) and (2.2) basically consider that the multivariate variability of x can be adequately described by the underlying behavior of a smaller number of p latent variables, plus noise added in the full variable space. We can also see that such model essentially consists of two parts: one that captures the variability due to normal process sources ($\mu_X + A \cdot l(k)$), and the other that explicitly describes the characteristics of measurement noise or uncertainties ($\varepsilon_{\mathrm{m}}(k)$), each one with its own independent randomness. In the sequel, we will refer to this model as our Heteroscedastic Latent Variable (HLV) model, to differentiate it from classical latent

variable techniques, where measurement uncertainties features are not explicitly accounted for.

Given the above model structure, parameter estimation is achieved from the probability density function for x under the conditions outlined above, which is a multivariate normal distribution with the following form:

$$x(k)\sim N_{\mathrm{n}}\left(\mu_X, \Sigma_x(k)\right) \tag{2.3}$$

with

$$\Sigma_x(k) = \Sigma_{\mathrm{l}} + \Delta_{\mathrm{m}}(k)$$

$$\Sigma_{\mathrm{l}} = A\Delta_{\mathrm{l}}A^T. \tag{2.4}$$

The likelihood function for a reference data set, composed by n_{obs} multivariate observations, is then given by:

$$L\left(\mu_X, \Sigma_{\mathrm{l}}\right) = \prod_{k=1}^{n_{\mathrm{obs}}} \left\{ \frac{1}{(2\pi)^{n/2}\left|\Sigma_x(k)\right|^{1/2}} \exp \times \left[-\frac{1}{2}(x(k)-\mu_X)^T \Sigma_x^{-1}(k)(x(k)-\mu_X) \right] \right\} \times \Sigma_x(k) = \Sigma_{\mathrm{l}} + \Delta_{\mathrm{m}}(k). \tag{2.5}$$

Therefore, the log-likelihood function, in terms of which calculations are actually conducted, is (C stands for a constant):

$$\begin{aligned} \Lambda\left(\mu_X, \Sigma_{\mathrm{l}}\right) &= \frac{n\cdot n_{\mathrm{obs}}}{2}\ln(2\pi) - \frac{1}{2}\sum_{k=1}^{n_{\mathrm{obs}}} \ln\left|\Sigma_x(k)\right| \\ &\quad - \frac{1}{2}\sum_{k=1}^{n_{\mathrm{obs}}}\left[(x(k)-\mu_X)^T \Sigma_x^{-1}(k)(x(k)-\mu_X)\right] \\ &= C - \frac{1}{2}\sum_{k=1}^{n_{\mathrm{obs}}} \ln\left|\Sigma_x(k)\right| \\ &\quad - \frac{1}{2}\sum_{k=1}^{n_{\mathrm{obs}}}\left[(x(k)-\mu_X)^T \Sigma_x^{-1}(k)(x(k)-\mu_X)\right]. \end{aligned} \tag{2.6}$$

Parameter estimates are then found from those elements of the parameter vector $\theta=[\mu_X^T, \mathrm{vec}(\sum_{\mathrm{l}})^T]^T$ that maximize the log-likelihood function:

$$\hat{\theta}_{\mathrm{ML}} = \max_{\theta} \Lambda\left(\theta | \{x(k), \sigma_{\mathrm{m}}(k)\}_{k=1,n_{\mathrm{obs}}}\right). \tag{2.7}$$

In fact, the situation is more involved, as $\sum_{\mathrm{l}}$ has certain a priori properties that should be satisfied also by its estimate, $\sum_{\mathrm{l}}$, namely that it should be both symmetric and non-negative definite [21]. During the course of our work, several approaches to solve (2.7) were tried out, with different degrees of enforcement of the restrictions arising from symmetry and non-negative definiteness. The one that provided more consistent performance is based upon the (usual) assumption that latent variables have a diagonal covariance matrix, Δ_{l}, being the coefficient matrix A estimated according to a procedure similar to the one adopted in [8]. In this procedure, we start from an initial estimate, A_0, and the numerical optimization algorithm proceeds by finding the optimal rotation matrix R, defined by the angles $\underline{\alpha}=[\alpha_1, \alpha_2, \ldots, \alpha_{n-1}]^T$, that maximizes (along with the reminding parameters, Δ_{l} and μ_{l}) objective function (2.7):

$$\hat{A} = R\left(\underline{\alpha}\right)\hat{A}_0 \tag{2.8}$$

$$R\left(\underline{\alpha}\right) = R_1(\alpha_1)\cdot R_2(\alpha_2)\ldots R_{n-1}(\alpha_{n-1}) \tag{2.9}$$

where,

$$R_1(\alpha_1)\begin{bmatrix} \cos\alpha_1 & -\sin\alpha_1 & 0 & \ldots & 0 \\ \sin\alpha_1 & \cos\alpha_1 & 0 & \ldots & 0 \\ 0 & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & 1 \end{bmatrix},$$

$$R_2(\alpha_2) = \begin{bmatrix} 1 & 0 & 0 & \ldots & 0 \\ 0 & \cos\alpha_2 & -\sin\alpha_2 & \ldots & 0 \\ 0 & \sin\alpha_2 & \cos\alpha_2 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & 1 \end{bmatrix}, \text{ etc.} \tag{2.10}$$

As $\sum_{\mathrm{l}}=\hat{A}\Delta_{\mathrm{l}}\hat{A}^T$ (from the invariance property of the maximum likelihood estimators, [22]), the symmetry property is automatically satisfied. With these considerations, the optimization problem to be solved remains an unconstrained one, and we have used a gradient optimization algorithm to address it. Approaches based on the Alternating Least Squares principle [8] are also worth being explored in future developments.

3. HLV-MSPC statistics

In this section we present the monitoring statistics that we do propose and discuss some issues regarding the implementation of MSPC within the scope of our HLV model, formulated in the previous section. Efforts were directed towards developing statistics that would be analogous to their well known counterparts, i.e., to T^2 and Q for MSPC based on PCA [20].

3.1. Monitoring statistics

The conventional T^2 and Q statistics were designed to follow the behavior of the two random components present

in a PCA model: one reflecting the structured variation arising from latent variables sources, which is "followed" by the T^2 statistic, and the other reflecting the unstructured part, driven by the residuals, followed by the Q statistic. As in our underlying model we also have structured and unstructured components (Section 2), we will pursue the same rational. Regarding the structured or "within" latent variables subspace variability, we will monitor it in the original variable domain, instead of the latent variable domain (as done in PCA-MSPC), in order to account for the effects of the (known) measurement uncertainties. This leads to the definition of the following statistic:

$$T_{\mathrm{w}}^2(k) = (x(k) - \mu_X)^T \Sigma_x^{-1}(k)(x(k) - \mu_X)$$

$$\Sigma_x(k) = \Sigma_{\mathrm{l}} + \Delta_{\mathrm{m}}(k)$$

$$\left(\Sigma_{\mathrm{l}} = A\Delta_{\mathrm{l}}A^T\right) \tag{3.1}$$

where $x(k)$ represents the kth measured multivariate observation, and the other quantities maintain the same meaning as before. It follows a $\chi^2(n)$ distribution, n being the number of variables. $T_{\mathrm{w}}^2(k)$ considers simultaneously the variability arising from both the structured (process) and unstructured (measurement noise) variability. Let us now define the statistic Q_{w}, that considers only the unstructured part of the HLV model, say $r(k)$, associated with measurement noise:

$$Q_{\mathrm{w}} = r^T(k)\Delta_{\mathrm{m}}^{-1}(k)r(k)$$

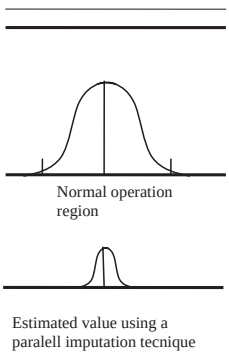
$$r(k) = x(k) - \mu_X - Al(k) = \underline{\varepsilon}_{\mathrm{m}}(k) \tag{3.2}$$

which follows a $\chi^2(n-p)$ distribution, with n and p being the number of variables and latent variables (pseudo-rank), respectively. In practice, we don't know the true values for the above quantities, but will use those that maximize the log-likelihood function as their estimates. Furthermore, $l(k)$ values are calculated using non-orthogonal (maximum likelihood) projections [9], given by:

$$\hat{l}_{\mathrm{ML}}(k) = \left(\hat{A}_{\mathrm{ML}}^T\Delta_{\mathrm{m}}^{-1}(k)\hat{A}_{\mathrm{ML}}\right)^{-1}\hat{A}_{\mathrm{ML}}^T\Delta_{\mathrm{m}}^{-1}(k)\left(x(k) - \hat{\mu}_{X,\mathrm{ML}}\right). \tag{3.3}$$

3.2. Missing data

The incorporation of uncertainty information regarding each measured value in our HLV-MSPC analysis not only adds a new important dimension to it, but also brings some parallel additional advantages. One of them is the inherent ability to handle reasonable amounts of missing data, in a coherent and integrated way. Usually, missing data are replaced by conditional estimates obtained under a set of more or less reasonable assumptions, or through iterative procedures where, in practical terms, the missing values play the role of additional parameters to be estimated. In the proposed procedure, when a datum is missing, we simply have to assign a value to it, together with its associated uncertainty. This assigned datum can simply be the mean of the normal operation data, with the corresponding standard deviation as an adequate uncertainty value. Alternatively, we can also assign the mean value together with a very large score for its associated measurement uncertainty, the rational being that a missing value is virtually given by any value with an "infinite uncertainty". More precise estimates, obtained through data imputation techniques, can also be adopted if they are able to provide us also with the associated uncertainties (Fig. 1).



a) No external knowledge: *use any value + "Infinite" uncertainty*

b) Knowledge of normal operation: *use mean value + uncertainty encompassing normal operation region*

c) Knowledge about an estimate of the missing value: *use estimate + estimate uncertainty*

Fig. 1. Three levels of knowledge incorporation with regard to missing data estimation: (a) no external knowledge; (b) knowledge about the mean and standard deviation under normal operation conditions; (c) imputation of missing data values using a parallel imputation technique.

4. Illustrative applications of HLV-MSPC

In this section we present the main results obtained with the application of our HLV-MSPC procedure to a number of different simulated scenarios where measurement uncertainties are allowed to vary (heteroscedastic noise). A final case study, based upon real industrial pulp quality data covering an extended operation period for a particular Portuguese plant, is also shown, where the purpose regards the extraction of knowledge regarding variability patterns in a real world context.

4.1. Application case studies

Our first four examples are based on data generated by the following latent variable model:

$$\begin{aligned} x(k) &= 5\cdot\begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \\ 1 & -1 \end{bmatrix}\cdot l(k) + \underline{\varepsilon}_{\mathrm{m}}(k) \\ l(k) &\sim iid\ N\left(0, \Sigma_{\mathrm{l}}\right), \Sigma_{\mathrm{l}} = \begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix} \\ \underline{\varepsilon}_{\mathrm{m}}(k) &\sim id\ N(0, \Delta_{\mathrm{m}}(k)) \end{aligned} \tag{4.1}$$

Table 1
Median of the percentages of significant events identified in 100 simulations, for normal and abnormal operation conditions (*Faults F1 and F2*)

Fault	Statistic	Normal operation	Abnormal operation
F1	T^2	2.40	17.80
	Q	31.40	79.70
	T^2_w	1.20	27.80
	Q_w	1.00	25.20
F2	T^2	2.30	1.40
	Q	31.60	45.20
	T^2_w	1.20	4.80
	Q_w	1.00	6.80

The measurement noise covariance can vary along time in various ways, as explained below, and each example covers a different scenario regarding time variation of measurement uncertainty. For comparative purposes, the results obtained using classic PCA are also presented. The statistics for PCA-MSPC are denoted by T^2 and Q, and those for HLV-MSPC as T^2_w and Q_w. All simulations carried out for the different scenarios share a common structure: first, in the training phase, 1024 multivariate observations are generated using model (4.1) in order to estimate the reference PCA and HLV models; then, in the testing phase, 1000 observations of new data are generated, half of which are relative to normal operation (from observations 1 to 500), while the other half correspond to an abnormal operation situation (observations 501 to 1000). For each of these two parts we calculate MSPC statistics, and the percentage of significant events identified (events above statistical limits), for the significance level adopted ($\alpha = 0.01$). In order to enable for a more sound assessment of results, the testing phase was repeated 100 times, and the performance medians over such repetitions computed. Furthermore, two abnormal situations (faults) are explored in each scenario, as follows:

(F1) A step change of magnitude 10 is introduced in all variables.

Table 2
Median of the percentages of significant events identified in 100 simulations, for normal and abnormal operation conditions (*Faults F1 and F2*)

Fault	Statistic	Normal operation	Abnormal operation
F1	T^2	0.40	5.20
	Q	0.00	1.00
	T^2_w	1.00	23.80
	Q_w	1.00	24.00
F2	T^2	0.40	0.20
	Q	0.00	0.00
	T^2_w	0.80	3.80
	Q_w	1.00	6.00

Table 3
Median of the percentages of significant events identified in 100 simulations, for normal and abnormal operation conditions (*Fault F1*)

Fault	Statistic	Normal Operation	Abnormal operation
F1	T^2	0.40	4.80
	Q	0.20	1.60
	T^2_w	1.00	28.00
	Q_w	1.10	30.20

(F2) A structural change in the model is simulated, by modifying one of the entries in the coefficient matrix:

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \\ 1 & -1 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & -0.5 \\ 1 & -1 \end{bmatrix} \tag{4.2}$$

Example 1. Constant uncertainty for the reference data (at minimum level).

In this example, measurement noise standard deviations for the reference data set (used to define control limits) were kept constant and at the minimum values that will be used during the test phase. For the test data, measurement uncertainties are allowed to vary randomly, according to the uniform distribution $\sigma^{Xi}_m(k) \sim U(2, 6)$ (we will refer to this situation as "complete heteroscedasticity"). The corresponding results are presented in Table 1, for the two types of faults mentioned above (*F1* and *F2*).

The PCA Q statistic detects a very large number of false alarms, whereas T^2 detects almost twice the expected rate under the adopted statistical significance level (0.01). The apparently good performance of Q under abnormal conditions is a consequence of the low statistical limits established, which are related with the low noise reference data used. This leads to a sensitive detection of any fault, but at the expense of a very large rate of false alarms under normal operation. HLV-MSPC statistics perform consistently better, particularly when we compare T^2_w and T^2 performances.

Table 4
Median of the percentages of significant events identified in 100 simulations, for normal and abnormal operation conditions (*Faults F1 and F2*)

Fault	Statistic	Normal Operation	Abnormal operation
F1	T^2	1.00	8.80
	Q	1.40	15.40
	T^2_w	1.00	25.20
	Q_w	1.00	25.00
F2	T^2	1.00	0.80
	Q	1.40	3.40
	T^2_w	1.00	4.60
	Q_w	1.00	6.60

Table 5
Results for fault *F1*, with variable uncertainty both in the reference and test data

Fault	Statistic	Normal Operation	Abnormal operation
F1	T^2	0.80	9.90
	Q	1.90	13.40
	T^2_w	1.00	28.50
	Q_w	1.00	29.20

Example 2. Constant uncertainty for the reference data (at maximum level).

Looking now to what happens if uncertainties in the reference data are held constant at the maximum levels used in the test data set (Table 2), we can see that the opposite detection pattern occurs with the T^2 and Q statistics, as expected. In these examples, as the reference data consists of highly noisy measurements, and therefore the control limits are set at higher values, the detection ability for false alarms becomes smaller when noise characteristics change. This also drastically reduces the capability for detecting significant events. Under this situation, HLV-MSPC statistics also outperform their classical counterparts.

In the previous results, measurement uncertainties for each value of each variable in the test set were allowed to change randomly from observation to observation, according to the probability distribution referred. We also tested scenarios where the values for all variables in the same row were assumed to have the same uncertainty, and found out that the same conclusions hold for this situation. For illustrative purposes, we present in Table 3 the results obtained for fault *F1*, when the reference data was generated at maximum uncertainty values.

Example 3. Variable data uncertainty for reference and test sets.

The examples mentioned so far address situations where the training set variables have constant measurement uncertainty, whereas the test set uncertainties have heteroscedastic behavior. This mismatch between training and testing situations has serious consequences in the performance of PCA-based MSPC. The following examples explore situations where both the reference and test data were generated under similar conditions of measurement uncertainty heteroscedasticity. First, we consider the already described situation of complete heteroscedasticity. From Table 4, it is possible to see that HLV-MSPC statistics still seem to present the best performance, although PCA-based MSPC counterparts also achieve good scores for normal operation.

Once again, the above conclusions do not change for the situation where uncertainty for all of the variables does change together, as shown for fault F1 in Table 5.

Example 4. Handling the presence of missing data.

This example explores the capability of the proposed methodology for handling missing data randomly scattered through data sets. The underlying model used to generate noiseless data sets is the same as before (Example 1), but we now removed some data records using an automatic random procedure that approximately eliminates a pre-specified percentage of values (it removes *on average* the chosen percentage), here fixed at 10%. As it happened with our previous examples, results presented below regard testing data performances. For HLV-MSPC we followed two different simple procedures for replacement of missing data: (i) in the first one (*MD I*) we inserted the un-weighted mean for each variable in a missing datum position, and associated to it a high value for the corresponding position in the uncertainty table (e^{10}); (ii) in the second procedure (*MD II*) we refined this estimate, using the available reference data to estimate the mean and standard deviations for each variable, being the former used to replace missing data and the last one to specify the associated uncertainty. For PCA-MSPC we estimated missing data using the reference data unweighted means (*MD*). Table 6 presents the results obtained for fault *F1*, with the values for HLV-MSPC and PCA-MSPC for the original data (i.e., without missing data) also being reported. It is possible to verify that there is a sensible and expected decrease of detection performances for the HLV-MSPC statistics under the more pessimistic imputation method, *MD I*, which are improved by using procedure *MD II*. From these results we can say that it is still advisable to continue with the implementation of HLV-MSPC in the presence of missing data, as the results with missing data are in general superior to those of PCA-MSPC *without* missing data.

Example 5. Analysis of pulp quality data.

A selected subgroup of nine key quality variables relative to the pulp produced in an integrated pulp and paper Portuguese mill (Portucel) was collected during a period of four and a half years, and are to be analyzed in order to identify any relevant variation patterns along time, as well as process upsets and disturbances, so that potential root causes can then be found and worked out, leading to process

Table 6
Median of the percentages of significant events identified in 100 simulations, for normal and abnormal operation conditions (fault F1)

Statistic	Operation	PCA (orig)	PCA (*MD*)	ML-HLV (orig)	ML-HLV (*MD I*)	ML-HLV (*MD II*)
T^2	Normal	1.10	0.80	1.00	0.80	1.20
	Abnormal (F1)	10.80	8.40	31.90	25.80	27.80
Q	Normal	2.80	5.70	1.20	0.80	1.20
	Abnormal (F1)	18.80	24.10	32.20	24.60	28.00

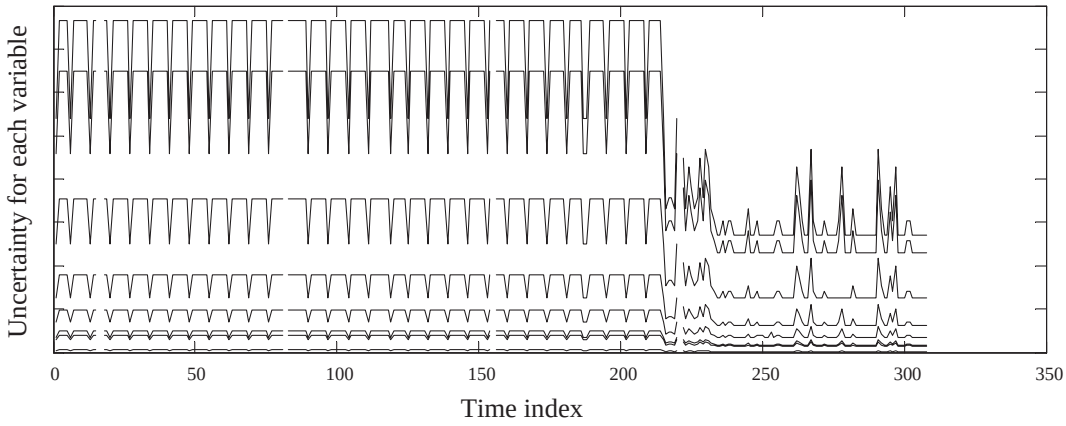


Fig. 2. Patterns of data uncertainty variation along time index for the 9 pulp quality variables analyzed (data is aggregated in periods of 8 days, and such time periods are reflected by the time index shown here).

improvement. These variables are related to paper structure properties, strength and optical properties. The first decision that one has to make concerns the time scale for conducting data analysis. A preliminary study did lead us to choose non-overlapping time windows of 8 days, over which we computed the average value for each variable. The associated uncertainties were initially estimated using a priori knowledge available regarding the measurement devices and the number of significant digits employed in the records (following a Type B procedure for evaluating measurement uncertainty, and assuming constant distributions in ranges defined by the last significant digit, [5]). However, this approach usually tends to provide rather optimistic estimates for the uncertainty figures in industrial settings, since additional noise sources come into place when one is not under standard and well-controlled conditions. Therefore, these estimates were corrected by analyzing noise characteristics of the signals using a wavelet-based approach (noise standard deviation was estimated from the details obtained in the first decomposition [23]), and the corresponding values for the averages over non-overlapping 8 days windows computed using standard uncertainty propagation formulas [5,7]. These uncertainty profiles along time are represented in Fig. 2. Since all of these variables are derived from the plant quality control laboratory, their acquisition periodicity is almost the same, and therefore their profiles do exhibit similar patterns.

We conducted a Phase I study, and calculated the HLV-MSPC statistics in order to analyze the variability structure along time. For setting the pseudo-rank parameter, a first guess can be easily provided by applying classical PCA to our data and then using one of the associated selection procedures available (e.g. [24–30]) for identification of the proper number of PC to retain. This initial guess can then be tested and revised in pilot implementations of the method over real data. A final selection should also be validated against the values of the diagonal matrix, Δ_1, estimated from such implementations, in order to check if they are also consistent with such choice. In the present case study, we did set $p=3$. Fig. 3 illustrates the values obtained for the T_w^2 statistic, where it is possible to identify a process shift after period 240, occasionally spiked with some rare but very significant abnormal events. For comparison purposes, we also present, in Fig. 4, the values obtained for the analogous T^2 statistic, obtained by conducting the same analysis using PCA-MSPC, where the sustained shift in the last period of

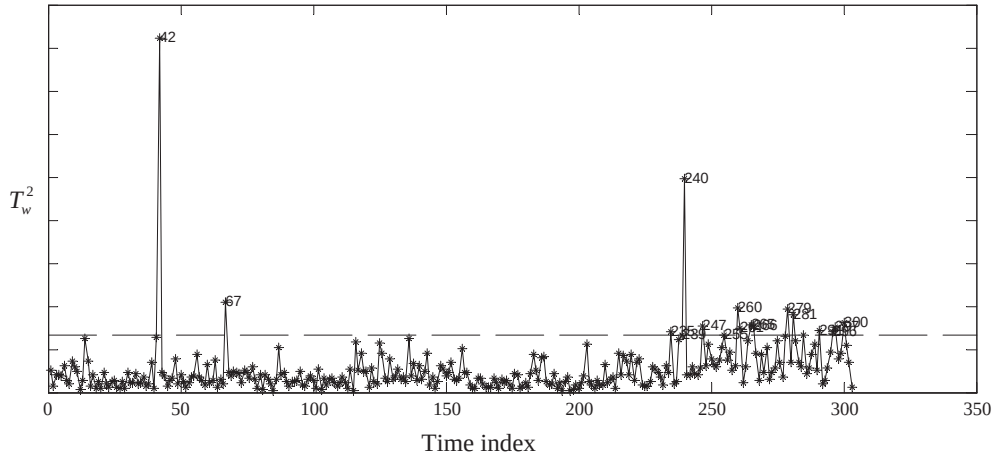


Fig. 3. HLV-MSPC: values for the T_w^2 statistic, in the pulp quality data set.

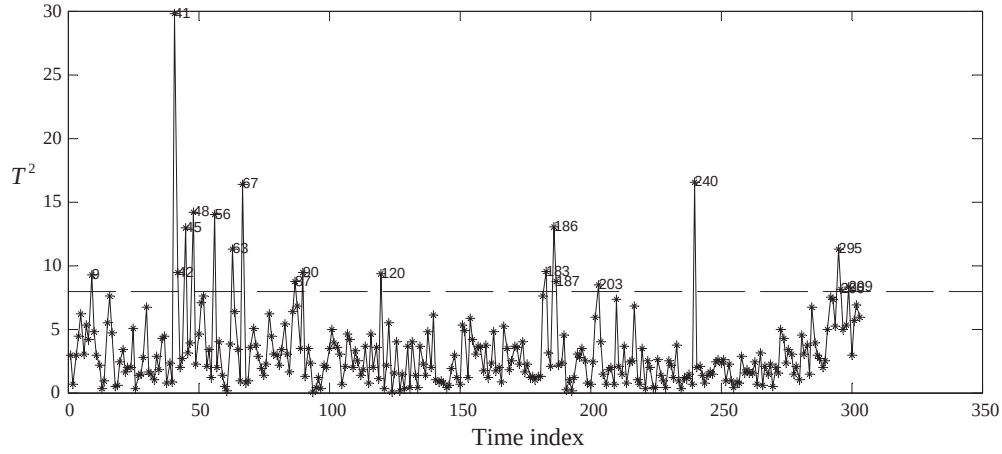


Fig. 4. PCA-MSPC: values for the T^2 statistic, in the pulp quality data set.

time is almost overlooked, whereas high data variability present in the beginning, where uncertainties have higher values, is not properly down-weighted, leading to an inflated variation pattern.

The T_w^2 profile provides a rough vision over the conjoint time behavior, but we can zoom into it (without having to analyze the variables separately, in which case we would be missing any changes in their correlation structure), by looking to what happens to the HLV scores provided by Eq. (3.3), as shown in Fig. 5. It is therefore possible to identify several trends affecting the three scores: a long range oscillatory pattern for the first score, a decreasing trend with shorter cyclic patterns superimposed for the second score, and a stable pattern that begins to oscillate in the final periods of time for the third score. By looking into the variables that are responsible for such behaviors, namely

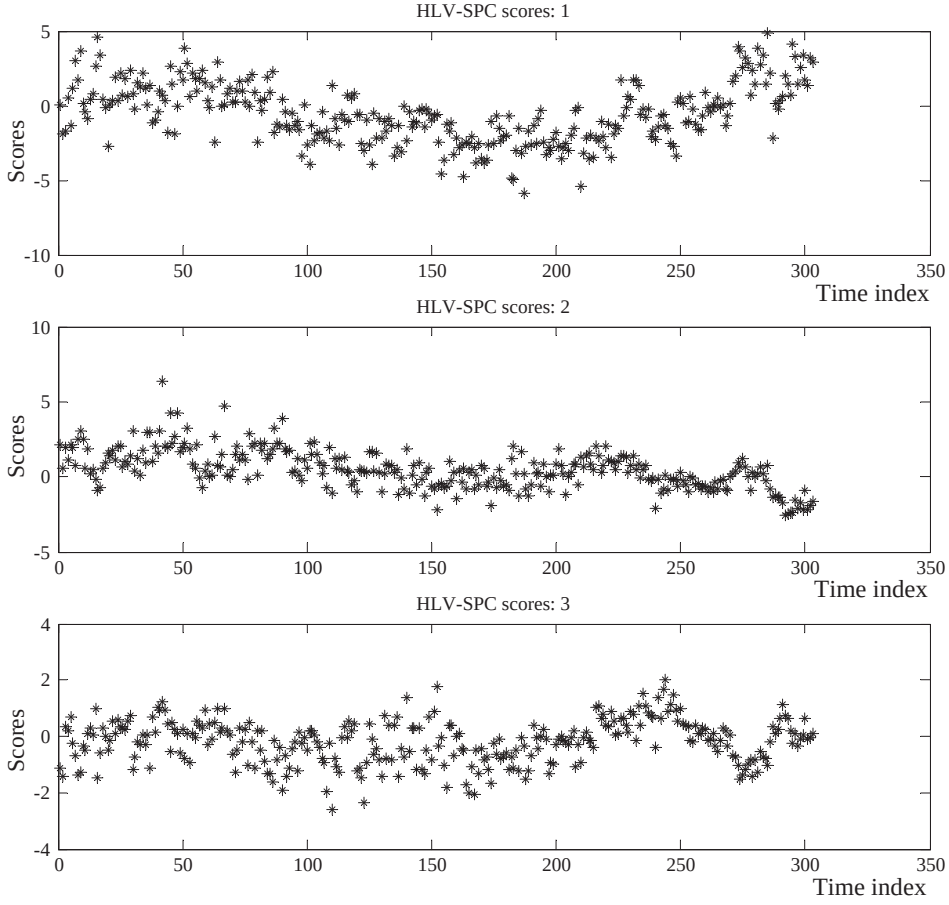


Fig. 5. HLV scores for the pulp quality data set.

Table 7
Mean and standard deviation of the results obtained for the angle, distance and similarity factor between the estimated subspace and the true one, using PCA and HLV (first row); paired t-test statistics for each measure, regarding 100 simulations carried out, along with the respective p-values (second row)

	Ang(PCA) (°)	Ang(HLV) (°)	Dist(PCA)	Dist(HLV)	Simil(PCA)	Simil(HLV)
Mean (Standard dev.)	26.62 (3.58)	17.23 (2.63)	0.42 (0.06)	0.30 (0.04)	0.91 (0.02)	0.95 (0.01)
t Statistic (p-value)	29.84 ($\ll 10^{-5}$)		30.54 ($\ll 10^{-5}$)		-25.89 ($\ll 10^{-5}$)	

through contribution plots for the scores, we can get more insight into the nature of these disturbances, and, eventually, about their root causes. Even though a detailed discussion can not be given here, due to space restrictions, one should notice that these types of trends are common in pulp and paper quality data, and can be due to issues ranging from seasonal wood variability and harvesting cycles to wood supply policies.

4.2. Discussion

The approach proposed in this paper was designed to perform SPC under noisy environments, i.e., scenarios where the signal to noise ratio (or, more adequately, signal to uncertainty ratio) is rather low, and, furthermore, where the magnitude of the uncertainty affecting each collected value can vary across time. Not only standard measurement systems that conform to the underlying statistical model are covered by this approach (e.g. laboratory tests, measurement devices), but also any general procedure for obtaining data values with an associated uncertainty (e.g. computational calculations, raw material quality specifications, etc.). The added value of this approach increases when the signal variation to uncertainty ratio becomes smaller. Therefore, it provides an alternative to PCA-MSPC for applications where low signal to noise ratio tends to happen.

The better capability of our approach to estimate the underlying true data subspace was also analyzed through a simulation study. Noiseless data were generated using the model described in Example 1, and then corrupted with noise, whose measurement uncertainties vary randomly between 2 and 6 (uniform distribution). For each trial, 100 multivariate observations were used to estimate the underlying latent variable subspace using classical PCA and our HLV approach. The angle that these estimates make with the true subspace, as well as the respective distances [31] and the Krzanowski similarity factor [32] between the estimated and the true subspaces, were calculated: ANG(PCA), ANG(HLV), DIST(PCA), DIST(HLV), SIMIL(PCA) and SIMIL(HLV), respectively. The Krzanowski similarity factor is a measure of the similarity between two PCA subspaces, ranging from 0 (no similarity) to 1 (exact similarity). The means and standard deviations for these quantities derived from 100 trials are presented in Table 7, along with the values of the t-statistic for paired t-tests between PCA and HLV results, and the respective p-values. A highly significant better estimation performance in favor of the HLV procedure was thus obtained.

5. Conclusions

We presented and discussed an approach for performing SPC in a multivariate process, explicitly incorporating measurement uncertainty information. It is a generalization of the current latent variable approach to MSPC based on PCA to a more general scenario where measurement uncertainties can vary from observation to observation. A statistical model was defined and statistics analogous to T^2 and Q were derived, that allow one to monitor both the within model variability as well as the variability around the identified model. Furthermore, this approach adequately handles the presence of missing data in a simple and consistent way. Preliminary results point out in the direction of advising the use of this framework when measurement uncertainties are available and significant noise affects process measurement behaviour. So far we have implemented and tested our approach in examples that do cover dozens of variables. In even larger scale problems, we may apply the same methodology over a subset of variables where heteroscedasticity is believed to be more critical.

Acknowledgements

The authors would like to acknowledge Portuguese FCT for financial support through project POCTI/EQU/47638/2002 and Portucel for providing us with sets of real industrial data.

References

[1] D.C. Montgomery, Introduction to Statistical Quality Control, Wiley, New York, 2001.
[2] J.V. Kresta, J.F. MacGregor, T.E. Marlin, Can. J. Chem. Eng. 69 (1991) 35–47.
[3] W.H. Woodall, J. Qual. Technol. 32 (4) (2000) 341–350.
[4] A.J. Burnham, J.F. Macgregor, R. Viveros, Chemom. Intell. Lab. Syst. 48 (1999) 167–180.
[5] ISO, Guide to the Expression of Uncertainty, Geneva, Switzerland, 1993.
[6] S.K. Kimothi, The Uncertainty of Measurements, ASQ, Milwaukee, 2002.
[7] I. Lira, Evaluating the Measurement Uncertainty, Institute of Physics Publishing, Bristol, 2002.
[8] P.D. Wentzell, D.T. Andrews, D.C. Hamilton, K. Faber, B.R. Kowalski, J. Chemom. 11 (1997) 339–366.
[9] P.D. Wentzell, D.T. Andrews, B.R. Kowalski, Anal. Chem. 69 (1997) 2299–2311.
[10] R. Bro, N.D. Sidiropoulos, A.K. Smilde, J. Chemom. 16 (2002) 387–400.

[11] À. Martínez, J. Riu, F.X. Rius, Chemom. Intell. Lab. Syst. 54 (2000) 61–73.
[12] F.J. Río, J. Rio, F.X. Rius, J. Chemom. 15 (2001) 773–788.
[13] J. Riu, F.X. Rius, Anal. Chem. 68 (1996) 1851–1857.
[14] K. Faber, B.R. Kowalski, J. Chemom. 11 (1997) 181–238.
[15] K. Faber, Chemom. Intell. Lab. Syst. 52 (2000) 123–134.
[16] K. Faber, R. Bro, Chemom. Intell. Lab. Syst. 61 (2002) 133–149.
[17] A. Phatak, P.M. Reilly, A. Penlidis, Anal. Chim. Acta 277 (1993) 495–501.
[18] J.A.F. Pierna, L. Jin, F. Wahl, N.M. Faber, D.L. Massart, Chemom. Intell. Lab. Syst. 65 (2003) 281–291.
[19] J.F. MacGregor, T. Kourti, Control Eng. Pract. 3:3 (1995) 403–414.
[20] B.W. Wise, N.B. Gallagher, J. Process Control 6:6 (1996) 329–348.
[21] C.R. Rao, Linear Statistical Inference and its Applications, Wiley, New York, 1973.
[22] D.C. Montgomery, G.C. Runger, Applied Statistics and Probability for Engineers, Wiley, New York, 1999.
[23] S. Mallat, A Wavelet Tour of Signal Processing, Academic Press, San Diego, 1998.
[24] F. Vogt, B. Mizaikoff, J. Chemom. 17 (2003) 346–357.
[25] S. Wold, Technometrics 20:4 (1978) 397–405.
[26] S. Valle, W. Li, S.J. Qin, Ind. Eng. Chem. Res. 38 (1999) 4389–4401.
[27] S.J. Qin, R. Dunia, J. Process Control 10 (2000) 245–250.
[28] M. Meloun, J. Èapek, P. Mikšík, R.G. Brereton, Anal. Chim. Acta 423 (2000) 51–68.
[29] B.K. Dable, K.S. Booksh, J. Chemom. 15 (2001) 591–613.
[30] E.V. Thomas, J. Chemom. 17 (2003) 653–659.
[31] G.H. Golub, C.F. Van Loan, Matrix Computations, The John Hopkins University Press, Baltimore, 1989.
[32] W.J. Krzanowski, J. Am. Stat. Assoc. 74:367 (1979) 703–707.