Ferrer, A. (2014). Latent Structures based-Multivariate Statistical Process Control: a paradigm shift. Quality Engineering. 26(1):72-91. doi:10.1080/08982112.2013.846093.

Latent Structures-based Multivariate Statistical Process Control: A Paradigm Shift

Alberto Ferrer
Multivariate Statistical Engineering Group,
Department of Applied Statistics, Operations Research and Quality,
Technical University of Valencia, Spain

ABSTRACT

The basic fundamentals of statistical process control (SPC) were proposed by Walter Shewhart for data-starved production environments typical in the 1920´s and 1930´s. In the 21st century the traditional scarcity of data has given way to a data-rich environment typical of highly automated and computerized modern processes. These data often exhibit high correlation, rank deficiency, low signal-to-noise ratio, multi-stage and multi-way structure, and missing values. Conventional univariate and multivariate statistical process control techniques are not suitable in these environments. This talk discusses the paradigm shift to which those working in the quality improvement field should pay keen attention. We advocate the use of latent structured-based multivariate statistical process control methods as efficient quality improvement tools in these massive-data contexts. This is a strategic issue for industrial success in the tremendously competitive global market.

Keywords: Multivariate statistical process control (MSPC); Principal component analysis (PCA); Partial least squares (PLS); Control charts; Latent structures; Quality improvement.

PREFACE

First of all I´d like to thank Geoff Vining and Ronald Does for inviting me to this stimulating conference resembling the famous Gordon Research Conferences from the old days. Being one of the keynote speakers from Europe is an honor and a challenge, and I hope to satisfy the expectations.

As part of the Stu Hunter Conference´s sessions this talk tries to provide a platform for discussion and the free exchange of ideas at the frontiers of statistics and quality research focusing on a particular issue: Statistical Process Control.

The first thing I did when I was invited to prepare this talk was to browse the internet looking for references on discussions, controversies, criticisms, challenges and open research issues in statistical process control (SPC). Out of this material, I strongly recommend the *Journal of Quality Technology* (*JQT*) panel discussion edited by Montgomery and Woodall (1997), the *Technometrics* panel discussion edited by Steinberg (2008) and the following references: MacGregor (1997), Woodall and Montgomery (1999), Stoumbos *et al.* (2000), Woodall (2000) and Bisgaard (2012).

1. INTRODUCTION

Statistical process control (SPC) refers to statistical methods used to monitor and improve the quality and productivity of manufacturing processes and service operations (Stoumbos *et al.* 2000). Featured tools of SPC are control charts.

According to Shewhart (1931), the acknowledged father of control charts: "*A phenomenon will be said to be controlled, when through the use of past experience, we can predict, at least within limits, how the phenomenon may be expected to behave in the future.*" The extent to which this goal is successfully met depends on the understanding of the type of variability affecting the process.

Variation remaining in a stable process reflects “common causes” that are an inherent part of the process and which cannot be removed easily without fundamental changes in the process itself: e.g. the design and operating settings of a polymerization reactor contribute to the common cause variation of the viscosity of a polymer. As long as you keep the process the same, common cause variation is the same from day to day, yielding a process consistent over time. A process affected by only common (random) causes of variation is said to be in “statistical control”. This means that the process is predictable, i.e. you can predict what the process will make in the near future, e.g. the proportion of batches meeting the customer specifications for a particular key property.

Added to the common cause system, other types of variability may occasionally affect the stability of a process: e.g., operator errors, defective raw materials, improperly adjusted or controlled equipment, sudden change in environmental conditions and so on. These sources of variability are not part of the common cause system and, thus, they are

referred as “assignable” or “special causes”. They are sporadic in nature and cannot be predicted. A process affected by special causes of variation is said to be “out of control” in the sense that it is unstable and thus unpredictable.

It is crucial to distinguish between these two types of variation because the responsibility for process improvement depends on what type of variation is present: front-line personnel are responsible to find and make decisions on assignable causes; common cause variation, on the other hand, is management’s responsibility.

Statistical process control (SPC) gathers a powerful collection of problem-solving tools useful in understanding process variability, achieving process stability, and improving process performance through the reduction of avoidable variability (i.e. the removal of special causes). The goal of any SPC scheme is to monitor the performance of a process over time in order to check whether the process behaves as expected (i.e. the predicted behavior from the common cause system), and to detect any unusual (special) event that may occur. By finding assignable causes, significant improvements in process performance can be achieved by eliminating these causes (or implementing them if they are beneficial). Control charts are the essential tools for pursuing this goal.

A control chart is a picture of a process over time that helps identify the magnitude and type of variation present. To implement a control chart, one must register data from a process over time. Decisions are required about the variables to measure, the sample statistics to be monitored, the sample size and the time between samples, the type of control chart, the control limits and the decision rules. All these choices determine the cost of the monitoring scheme and its ability to detect certain out-of-control conditions.

Mathematically, this problem has been formulated as detecting changes over time in the vector of parameters θ of an underlying probability distribution. In the case of standard control charts for univariate variables it is usually assumed that the key characteristic *x* at time *t* is a normally distributed random variable that can be expressed as $x_t = \mu + a_t$, where μ is the process mean, and a_t follows a stochastic “white noise” process, i.e. a series of independent and identically normally distributed (iind) random disturbances with zero mean and variance σ^2 ($a_t \sim N(0,\sigma^2)$). In this case, the distribution of the key characteristic depends on the vector of parameters $\theta^{\mathrm{T}} = \{\mu,\sigma\}$. “There are control charts proposed for monitoring the parameter (or parameters) of all

the standard probability distributions, both discrete and continuous" (Woodall and Montgomery 1999).

Control charts for multivariate variables (used in multivariate statistical process control – MSPC) usually assume that the (K×1) vector of measured variables (i.e. key characteristics) $\mathbf{x}$ is not time dependent and follows a K-dimensional normal distribution, $\mathbf{x} \sim N_K(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. In this case the distribution function of the key characteristics depends on the vector of parameters $\boldsymbol{\theta}^T = \{\mu_1,...,\mu_K,\sigma_1,...,\sigma_K,\sigma_{12},...,\sigma_{(K-1)K}\}$ containing all the marginal means from the mean vector $\boldsymbol{\mu}$ and all the marginal variances and covariances from the covariance matrix $\boldsymbol{\Sigma}$.

Although there are some disagreements regarding the relationships between control charting and repeated hypothesis testing (Woodall 2000), control charts are designed and evaluated under the assumption that if the process is operating with $\boldsymbol{\theta}=\boldsymbol{\theta}_0$ the process is said to be in (statistical) control, while if $\boldsymbol{\theta}\neq\boldsymbol{\theta}_0$ the process is considered out of (statistical) control. From the above relationship regarding the state of the process and the occurrence of special causes of variation, this parametric statistical distribution-based (i.e. theoretical) approach of control charting assumes that assignable causes result in shifts in the values of the parameters vector $\boldsymbol{\theta}_0$.

Although conventional (M)SPC is sound from a statistical point of view, it suffers from lack of applicability in data-rich environments, typical of modern processes. This paper advocates the use of latent structures-based multivariate statistical process control (LSb-MSPC) to deal with this issue.

The remainder of the paper is organized as follows: In Section 2, a critical view of conventional SPC in data-rich environments and the need for a paradigm shift are discussed. An overview of LSb-MSPC and a case study are presented in Section 3. Section 4 discusses the potential and challenges of LSb-MSPC in data-rich environments. Finally, some conclusions are summarized.

2. CRITICAL VIEW OF CONVENTIONAL SPC: THE NEED FOR A PARADIGM SHIFT

The principles (or underlying assumptions) contained in the so-called conventional SPC introduced in Section 1 assumed data sets with relatively low

frequency sampling and a small number of process variables. "*This application environment was typically one in which process data was relatively difficult to collect; parts were sampled and measured manually at relatively long time intervals, and only a few parts were evaluated each day for purposes of process monitoring and control*" (Woodall and Montgomery 1999). In addition, many of these early methods proposed in the first half of the 20th century by Shewhart and others were designed for simplicity and ease of calculation. They seemed to work; otherwise they would have not been so widely applied in so many processes. While many advances in SPC methodology have been achieved, in general, the primary focus of SPC research and available tools still assume relatively low sampling frequency and dimensionality, with the possible exception in the field of chemometrics.

Today´s SPC application environment is often different from the past data-starved production environments. "*On-line measurement, data capture, and analysis through hierarchical, distributed computing systems are becoming the norm in many industrial settings, from discrete parts to the chemical and process industries. This has totally changed the nature of the data available for process monitoring and control*" (Woodall and Montgomery 1999). This dramatic change due to the revolutionary innovations in sensor technology calls for a paradigm shift for quality improvement and control to which those working in the quality improvement field should pay keen attention.

Univariate SPC schemes are totally inadequate in these multivariate contexts. As commented by MacGregor (1997): "*the presence of variable interactions in experimental designs leads to the same difficulties in interpreting the results of one factor at a time experimentation as does the presence of correlation among variables in interpreting univariate SPC charts"*. Applying univariate SPC charts to each individual variable will force the operator to inspect a large number of control charts. When special events occur in a process, they might affect not only the magnitude of the variables but also their relationship to each other. These events are often difficult to detect by charting one variable at a time in the same way as the human eye can not perceive perspective by looking at an object with one-eye-at-a-time (Ferrer 2007).

Multivariate statistical process control (MSPC) schemes that treat all the variables simultaneously are required in these new data-rich environments. In these situations subgroup size (i.e. number of multivariate observations in each sample) is naturally n=1. "*The Shewhart* $\bar{X} - R$ *sampling paradigm (subgrouping) will be taken over by real-time sampling, which favors statistical process control methods based on individual*

observations" (Bisgaard 2012). This occurs frequently in the chemical and process industries, and in automated manufacturing systems with 100% inspection.

At a first glance, it can seem that conventional control charts for multivariate variables can meet the requirements of the paradigm shift but this is not the case as shown in the following.

Conventional MSPC schemes are based on forming a single control statistic (a quadratic form) from the (1×K) vector of measurements $\mathbf{x}$ registered in each sample, assuming that $\mathbf{x} \sim N_K(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. They are designed under the parametric statistical distribution-based (i.e. theoretical) approach of control charting to check for the stability of a subset of the vector of parameters $\boldsymbol{\theta}^{\mathrm{T}} = \{\mu_1,...,\mu_K,\sigma_1,...,\sigma_K,\sigma_{12},...,\sigma_{(K-1)K}\}$, generally either the mean vector $\boldsymbol{\mu}$ or the covariance matrix $\boldsymbol{\Sigma}$.

Hotelling´s T^2 control chart for monitoring process mean

One of the multivariate control charts most used in practice for monitoring the mean vector of a process is the multivariate extension of the univariate Shewhart control chart for monitoring the process mean. The chart is based on monitoring at each time t the Hotelling´s T^2 statistic (Hotelling, 1947; Jackson, 1985; Wierda, 1994; Fuchs and Kenett, 1998; Montgomery, 2005):

$$T_t^2 = (\mathbf{x}_t - \bar{\mathbf{x}})^{\mathrm{T}} \mathbf{S}^{-1} (\mathbf{x}_t - \bar{\mathbf{x}}) \tag{1}$$

where $\mathbf{x}_t$ is the (K×1) vector of measured variables at time t, $\bar{\mathbf{x}}^{\mathrm{T}} = \{\bar{x}_1, \bar{x}_2, ..., \bar{x}_K\}$ is the sample mean vector (estimate of the in-control (K×1) mean vector $\boldsymbol{\mu}$), and $\mathbf{S} = (m-1)^{-1} \sum_{i=1}^{m} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^{\mathrm{T}}$ is the (K×K) sample covariance matrix (estimate of the in-control (K×K) covariance matrix $\boldsymbol{\Sigma}$) from the m samples of size n=1 collected in Phase I. Upper control limits (UCL) at significance level (type I risk) α can be obtained for Phase I and Phase II, based on the multivariate normality assumption (Tracy *et al.* 1992). (For a clarification on the meaning of Phase I and II, see Appendix).

The Hotelling´s T^2 control chart checks if the mean vector $\boldsymbol{\mu}$ of the process remains constant (assuming a constant covariance matrix). This implies that the assignable causes may only affect the mean vector of the process (and not the covariance structure) yielding an extremely simplistic assumption in practice.

Multivariate Shewhart-type control charts for monitoring process variability

Two indices for measuring the overall variability of a set of multivariate data are: (1) the generalized variance $|\boldsymbol{\Sigma}|$, i.e. the determinant of the covariance matrix, and (2) the trace of the covariance matrix $tr\,\boldsymbol{\Sigma}$, i.e. the sum of the variances of the K variables.

Alt (1985) presents control charts for monitoring process variability based on these two indices, assuming that $\mathbf{x}$ follows a $N_K(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution, and that there are m samples (i.e. number of subgroups) of size $n>1$ available. In the first case, the statistic to be charted at time t is

$$W_t = -Kn + Kn\ln(n) - n\ln\left(|\mathbf{A}_t|/|\mathbf{S}|\right) + tr\left(\mathbf{S}^{-1}\mathbf{A}_t\right) \qquad (2)$$

where $\mathbf{A}_t = (n\text{-}1)\mathbf{S}_t$, $\mathbf{S}_t$ is the sample covariance matrix of the t-th subgroup ($n \geq K$), $\mathbf{S}$ is an estimate of the in-control covariance matrix $\boldsymbol{\Sigma}$ from Phase I, and tr is the trace operator. When the process is in control, the W_t statistic follows approximately a χ^2 distribution with $K\times(K+1)/2$ degrees of freedom.

The second chart is based on plotting the sample generalized variance $|\mathbf{S}_t|$, assuming this is approximately normally distributed. Some authors study the distribution of the $|\mathbf{S}_t|$ statistic obtaining control limits for simplistic scenarios (e.g. K =3) (See Aparisi *et al.* 1999).

These charts can be used in combination with the Hotelling´s T^2 charts to check if the process mean μ and covariance matrix $\boldsymbol{\Sigma}$ are constant over time. Although $|\mathbf{S}_t|$ is a widely used measure of multivariate dispersion, it is a simplistic scalar representation of a complex multivariate structure. Therefore, its use can be misleading in the sense that different correlation structures can yield identical generalized variances. Analogously, different changes in marginal variances can yield the identical trace of the covariance matrix. Note that if rational subgroup size n=1 (typical in data-rich environments as commented before) these charts cannot be used.

Multivariate control charts with memory

The previous Shewhart-type control charts are extremely useful in Phase I implementation of SPC (model building), where the process is likely to be out-of-control and experiencing assignable causes that result in large shifts in the monitored parameters. They are also very useful in the diagnostic aspects of bringing a "wild" process into statistical control, because the patterns on these charts often provide

guidance regarding the nature of the assignable cause (Montgomery 2005). This advantage comes from the fact that they are plots of the actual data providing a picture of what the process is doing that makes the interpretation easy. These charts act as global radars, being potentially capable of drawing attention to unusual kinds of behavior and hence to possible signals and causes previously unsuspected (Box and Luceño 1997).

The major disadvantage of Shewhart-type control charts is that they are relatively insensitive to small process shifts. In order to be able to detect small and moderate-sized sustained shifts in the vector of parameters $\boldsymbol{\theta}$, time-weighted control charts that accumulate information across past data have been proposed (Lowry *et al*. 1992).

One approach is to form an exponentially weighted moving average (EWMA) (Hunter 1986) of present and past values of statistics such as Hotelling´s T^2 (Eq. 1):

$$MEWMA_t^2 = \lambda T_t^2 + (1-\lambda) MEWMA_{t-1}^2 \tag{3}$$

where λ is a smoothing constant ($0 < \lambda \leq 1$).

A second approach is to form an EWMA of each individual measured variable and then combine across the different variables to create a new multivariate vector $\mathbf{w}$ defined as

$$\mathbf{w}_t = \lambda \mathbf{x}_t + (1-\lambda)\mathbf{w}_{t-1} \tag{4}$$

Then chart the quadratic form

$$MEWMA_t^2 = \mathbf{w}_t^{\mathrm{T}} \mathbf{S}_{\mathbf{w}_t}^{-1} \mathbf{w}_t \tag{5}$$

where the covariance matrix is $\mathbf{S}_{\mathbf{w}_t} = \frac{\lambda}{2-\lambda}\left[1-\left(1-\lambda\right)^{2t}\right]\mathbf{S}$, and $\mathbf{S}$ is an estimate of the in-control covariance matrix $\boldsymbol{\Sigma}$ from Phase I.

Fault diagnosis

Once the multivariate control chart signals an out-of-control alarm it is necessary to identify an assignable cause. This involves two steps: first (diagnostic) find which measured variable(s) contributes to the out-of-control signal, and second (root cause identification) determine what happened in the process that upset the behavior of these variables. For pursuing the first step (isolation of variables responsible for the out-of-control signal), several approaches have been reported in the literature. Kourti and MacGregor (1996), Mason *et al*. (1997), Bersimis *et al.* (2007) and Vidal-Puig and Ferrer (2013) provide an extensive review and references for diagnostic procedures in

conventional MSPC schemes. Regarding the second step (identifying root causes of the problem), management and operator actions based on technical process knowledge will be required.

Problems of conventional MSPC schemes in data-rich environments

Conventional MSPC control charts suffer from critical problems in data-rich environments.

Many require the assumption of independent multivariate normal distribution. Tests for multivariate normality may be impractical to perform on a regular basis in an industrial setting but these are needed to check this assumption. If this assumption is not reasonable, control limits may be obtained from resampling methods such as bootstrap (Liu and Tang 1996) or other simulation methods. The requirement is to have a sufficiently large historical data set available. Non-parametric multivariate control charts derived from the notion of data depth have been proposed (Liu 1995, Liu *et al.* 2004).

The performance of conventional MSPC control charts in detecting process disturbances and diagnosing their root causes tends to deteriorate as the number of monitored variables increases (Lowry and Montgomery 1995, Vidal-Puig and Ferrer 2013). This is the consequence of the structure of the multivariate control statistics used that relies heavily on the covariance structure. As shown in Eq. (1), (2), (3) and (5), multivariate control charts need the inverse of a covariance matrix. To avoid problems with this inverse, the number of multivariate observations (i.e. samples) (m) has to be larger than the number of variables (K), and the covariance matrix has to be well conditioned (i.e. variables must not be highly correlated). The non-parametric approaches based on the notion of data depth (Liu 1995 and Liu *et al.* 2004) also require that $m>K$, and that K is not too large. Otherwise, computing requirements make the approach unfeasible.

Moreover, complete data (no missing values) are required to calculate Hotelling´s T^2 or $MEWMA^2$ statistics for any particular sample. Multivariate control charts for monitoring process variability based on the generalized variance or the trace of the covariance matrix need subgroup size $n>1$.

These requirements are not met in highly automated processes where one is likely to encounter *large data sets* with not only a few product quality variables but also hundreds (or even thousands) of highly correlated process variables measured at a high

frequency rate, and in Phase I, with (often) *more variables than samples* (i.e. $m<K$). The nature of high-technology data is diverse coming from a huge variety of different kind of sensors providing different type of signals: spectra (chemical signals), pressures, temperatures, flows, … (physical signals), pH, conductivity, dissolved oxygen (biochemical signals), gene expression profiles (genetic signals), electronic eyes (digital images), electronic noses and tongues (potentiometric signal), electronic ears (acoustic signals) and so on. *Multivariate normality* in these cases is an illusion. *Missing data* appear for different reasons: failures in sensors or in the communication between the instrumentation and the digital control system (DCS), sensors taken offline for routine maintenance, manual samples not collected at the required times, data discarded due to gross measurement errors, and sensors with different sampling periods (Arteaga and Ferrer 2002). *Autocorrelation* (i.e. data correlation over time) is also part of the common-cause system affecting this type of processes. Lack of independence between successive observations shows up whenever the interval between samples becomes small relative to the process dynamics (i.e. inertial elements as raw materials flow, storage tanks, reactors' residence times or environmental conditions, defining the settling time of the process). This is more the rule than the exception in modern process environments due to information technologies that allow registering data from every part produced (e.g. in computer integrated manufacturing environments in manufacturing) or at high sampling rates (e.g. in continuous and batch process industries). All these issues advise against using conventional MSPC in these data-rich environments.

The paradigm shift

All the papers mentioned in the Preface agree SPC has to evolve to face the challenges of massive-data environments. Some of the conclusions and future directions (mostly proposed fifteen years ago) follow:

- SPC must be adapted to a changing manufacturing environment, massive data sets, multi-step production processes, better diagnosing methods (traceability), higher quality requirements, and greater computing capability (Woodall and Montgomery 1999).
- "*I cannot believe that there are tests for multivariate normality with sufficient power for practical sample sizes that I would even bother to use them;*

distribution-free multivariate SPC is what we need" (Coleman in Montgomery and Woodall 1997, p. 149).

- "*One important problem is traceability, i.e. rapid identification of the sources of an out-of-control condition by linking the subgroup identifier to information about raw material and earlier steps in the process that is available on-line… Another problem is the management of data and control limits for large number of processes that are continually evolving over time… large-scale applications require software systems that compute, save, update, and display the control limits*" (Palm *et al.* in Montgomery and Woodall 1997, p. 126).
- "*Researchers developing new technologies can help change this situation by emphasizing the gains that can be made using more sophisticated methods alongside the traditional Shewhart charts, by implementing them in easy-to-use software, and by "spreading the word" wherever they can*" (Crowder *et al.* in Montgomery and Woodall 1997, p. 139).
- A problem area in SPC is the detection and subsequent monitoring of multivariate autocorrelated data (Mason *et al.* in Montgomery and Woodall 1997, p. 142).

In 1997 Palm *et al.* (in Montgomery and Woodall 1997, p. 122) stated that there were three groups with distinct perspectives on control charts methods: broad applications (Group 1), technical opportunities (Group 2), and theory and methods (Group 3). For Group 1, the key issues are the breadth and effectiveness of application of control charts, coupled with competitive issues which motivate their use. Group 2 is made up of statisticians, chemical engineers, chemometricians and other scientists who are well trained in SPC and work closely with problem domain specialists who are knowledgeable about the processes. For this group, the future of control charts was moving in the direction of techniques that are appropriate for large multivariate data sets that possess less than full statistical rank: latent structures-based (LSb) techniques for dimension reduction such as principal component analysis (PCA) and partial least squares (PLS). This was also stated by Montgomery and Woodall (1997):*"Areas of future research that have considerable promise include developing methods for statistical monitoring and control with very large data sets, developing distribution-free multivariate methods… developing more insight regarding the performance of methods such as partial least squares and principal components in multivariate process*

monitoring, and investigating approaches for making these techniques easier to use in practice". Some contributors to the panel discussion on the future of industrial statistics edited by Steinberg (2008) stated the same conclusion. For Group 3, made up mainly of statisticians and industrial engineers with an academic interest in control chart methods, there was a challenge: to understand the process generating the data and to work closely with engineers and other scientists and problem domain specialists involved in massive-data environments.

Group 2 has mainly contributed to journals such as *Chemometrics and Intelligent Laboratory Systems*, *Journal of Chemometrics*, *Industrial & Engineering Chemistry Research, AIChE Journal,* and *Computers & Chemical Engineering*. The key journal for Group 3 is *Journal of Quality Technology*, and to some extent *Technometrics*. The papers referred in the Preface belong mainly to the latter group. From the papers published in both group of journals since 1997, I guess that while Group 2 has done a good job in developing useful LSb-methods for MSPC and providing successful case studies mainly in the (bio)chemical industry, Group 3´s challenges posed in 1997 still remain in 2013.

My personal opinion is that a paradigm shift is needed, especially for researchers belong to Group 3 (Theory and Methods). The parametric statistical distribution-based (i.e. theoretical) approach of control charting advocated by Group 3 researchers is no longer appropriate in data-rich environments. The hypothesis that assignable causes of variation result in shifts of a particular subset of the parameter vector is hard to believe in practice. A holistic role of SPC for process improvement, and not only for monitoring, must be considered. The focus must be the process, not the model. Surprisingly, these ideas are not really new; they are rooted in Shewhart´s basic fundamentals (Wheeler in Montgomery and Woodall 1997, pp. 154-155):

- "*Shewhart was not trying to find some exact model to describe the process, but he was trying to determine if a process fits (at least approximately) a very broad model of "random behavior*".
- "*The emphasis is not upon the use of the model to describe the process, but upon the characterization of process behavior as a starting point for process improvement*".
- "*When we focus on the control chart as a modeling procedure we will have, as a consequence, a narrow perception of how to use control charts. This is the origin of the idea that control charts are only useful for process monitoring…*

The point is that monitoring is only a minor part of what control charts can do, rather than being all that they can do". This will be illustrated in the case study in Section 3.

- *The real advantage of Shewhart´s charts is the way they make the data come alive. ... They allow the human mind to immediately focus on the interesting characteristics in the data without being distracted by the unimportant details. Moreover, they provide a uniform method of analysis whereby a group of people can agree on what constitutes a potential signal… The real power of Shewhart´s charts is not quantified by Average Run Length curves. The real power of Shewhart´s charts is their ability to get those with very little training to understand and use data properly.*

Attendees of the 2013 Stu Hunter Conference will recognize themselves as being part of one of Palm *et al.´*s groups. Although I originally belonged to Group 3, since I visited Prof. John F. MacGregor at MacMaster University in 1997, I realized the potential of LSb-methods and started moving towards Group 2. Anyway, no matter which group anyone belongs to, I guess all people interested in SPC should work together to reduce the distance between the different groups and make the future of SPC a successful story. In this paper I will defend my personal view of the paradigm shift needed (rooted in Shewhart´s basic principles), and I will advocate the use of latent structured-based multivariate statistical process control methods as key quality improvement tools in massive-data contexts and a strategic tool for industrial success in the tremendously competitive global market.

3. LATENT STRUCTURES-BASED MSPC

For treating large and ill-conditioned data sets that are not full statistical rank ($m<K$), we advocate the use of latent structures-based (LSb-) MSPC in the way they were proposed by Kourti and MacGregor (1996) and Nomikos and MacGregor (1995). Latent structures methodology exploits the correlation structure of the measured variables by revealing the few independent underlying sources of variation (latent structures) that are driving the process at any time. Multivariate statistical projection methods such as principal component analysis (PCA) (Jackson, 2003) and partial least

squares (PLS) (Geladi and Kowalski 1986, Helland 1988, and Höskuldsson 1988) are used to reduce the dimensionality of the monitoring space by projecting the information in the measured variables down onto low-dimensional subspaces defined by a few latent variables (i.e. scores). The multivariate scores are mathematically orthogonal and optimal summaries of the measured variables, so they are ideally suited for displaying in control charts. The scores are also less noisy than the measured variables, since they are weighted averages (linear combinations) of the measured variables. The process is then monitored in these latent subspaces by using a few multivariate control charts built from multivariate statistics which can be thought of as process performance indices, or *process wellness indices* (Kourti, 2005).

Basically, two multivariate control charts based on T_A^2 (sum of squares of standardized scores) and *SPE* (squared prediction errors) statistics are enough to summarize the information in the measured variables. These control charts are based on a minimal number of assumptions. The only essential one is that the LSb-model has been built on a representative set of data that captures all and only common cause variation (*in-control* or reference data set). The statistical variation about this model that captures the *in-control* behavior provides a reference distribution from which control limits can be derived. Neither the assumption of normality nor independence of consecutive subgroups are needed if one has enough data in Phase I. In this case, control limits for the Phase II multivariate charts can be derived based on reference distributions obtained from Phase I data.

These charts retain all the simplicity of presentation and interpretation of conventional univariate SPC charts. However, by using the information contained in all the measured (process and quality) variables simultaneously, they are much more powerful for detecting out-of-control conditions and diagnosing their root causes. Computing variable contributions eliminates much of the criticism that principal components lack physical interpretation. The contribution plots (see Appendix A.2) will not explicitly reveal the cause of the event but they point to the group of measured variables that are no longer consistent with normal operating conditions, providing good insight into plausible causes to operators and engineers. This is a good starting point for further fault analysis by using process knowledge to deduce possible root causes.

As Kourti (2002) commented, by monitoring only the final quality attributes (e.g. melt index of a polymer) -as frequently done in conventional MSPC- we are in fact

performing statistical quality control (SQC). For true statistical process control (SPC), one must also look at all the process variables involved (e.g. temperature profiles, raw materials properties, mass ratios, and so on). "*With new sensor technologies taking center stage, quality monitoring and control will move upstream and the focus will increasingly shift from monitoring output quality to monitoring and controlling inputs and process parameters*" (Bisgaard 2012).

Monitoring the process variables is expected to provide much more information on the state of the process and to supply it more frequently. Furthermore, any abnormal events that occur will also have their fingerprints in the process data. Thus, once an abnormal situation is detected, it is easier and faster to diagnose the source of the problem, as we are dealing directly with the process variables. On the contrary, control charts on the quality variables will only signal when the product properties are no longer consistent with expected performance, but they will not point to the process variables responsible for the problem, making more difficult and slower the fault diagnostic process.

Another advantage of monitoring process data is that quality data may not be available at certain stages of the process. Sometimes product quality is determined only by the performance of the product later, during another process. For example, if a catalyst is conditioned in a batch process before being used for polymer production, the quality of the catalyst (success of conditioning) is assessed by its performance in the subsequent polymer production. It would be useful to know if the catalyst will perform well before using it; monitoring the batch process variables would detect abnormal situations and would provide an early indication of poor catalyst performance.

In some cases, the scarce properties measured on a product are not enough to define entirely the product quality and performance for different applications. For example, if only the viscosity of a polymer is measured and kept within specifications, any variation in end-use application that arises due to variation in chemical structure (branching, composition, end-group concentration) will not be captured. In these cases, the process data may contain more information about events with special causes that may affect the product structure and thus its performance in different applications. Finally, by monitoring process variables, other abnormal operating conditions may be detected as, for example, a pending equipment failure or a process failure (Kourti 2002).

Finally, by using LSb-MSPC, missing and noisy data are easily handled (Nelson *et al*. 1996, and Arteaga and Ferrer 2002) and predictive models based on partial least

squares (PLS) or principal component regression (PCR) can also be used (Martens and Naes, 1989).

For those interested in the technical details of LSb-MSPC see the the APPENDIX. Kourti and MacGregor (1996) provide an excellent overview and discussion of conventional *vs* LSb-MSPC monitoring, an example in a continuous chemical process, and a list of useful references. For an example of the application of LSb-MSPC in a manufacturing industry, see e.g. Ferrer (2007). For those interested in the particularities of LSb-MSPC for batch processes, I strongly recommend Nomikos and Macgregor (1995). Commercial software implementing LSb-MSPC is available. See, e.g. SIMCA from Umetrics (http://www.umetrics.com/simca) or ProMV from ProSensus (https://prosensus.ca/solution/promv).

Case study

Data from a continuous process owned by a petrochemical company are used to illustrate the potential of LSb-MSPC as an efficient statistical tool for process understanding, monitoring and improvement. The goal is to diagnose the causes of variability of one of the critical to cost characteristics: the yield of the reaction, and to implement a multivariate monitoring scheme. This process takes place in several units. The data base includes: 72 process variables (temperatures, flows, tank levels, etc.) measured every hour from on-line electronic sensors located in the different units, 5 yield variables (%) measured every 8 hours, and 3 yield variables (%) measured every day from off-line laboratory analyses. The data were collected during two campaigns of approximately 4 months each.

Figure 1 shows the evolution of the most critical yield (%) measured every 8 hours in the two campaigns analyzed. Practically, both campaigns have similar average performance in process yield (no statistical significance mean difference, p-value > 0.05).

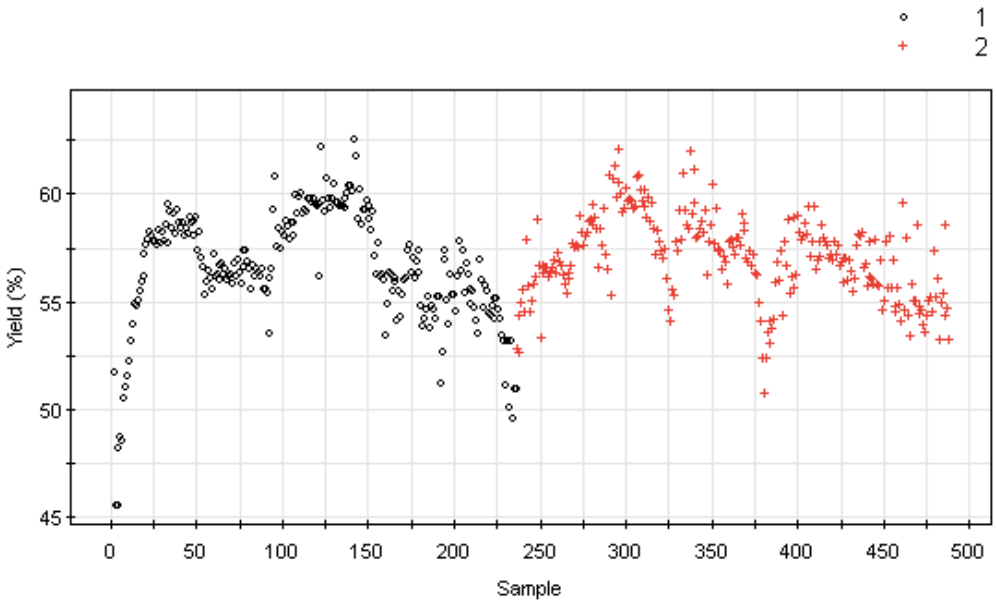

Figure 1: Most critical yield (%) measured every 8 hours in two campaigns: 1 (circles), 2 (crosses).

A question arises: does similar behavior in average output mean similar (i.e., consistent) process? The answer in this case is "no". Figure 2 shows the score scatter plot of the two latent variables of the partial least squares discriminant analysis (PLS-DA) (Sjöström *et al.* 1986) fitted using only data from the 72 process variables (measured every hour). PLS-DA is a multivariate statistical method for classification and discrimination purposes (as e.g. Fisher´s Discriminant Analysis) especially suited for data-rich environments (Prats-Montalbán *et al.* 2006). In this example the original 72-dimensional space has been projected (compressed) into a 2-dimensional subspace that best segregates the two campaigns. Each one of the points shown corresponds to an hour. By looking at the swarm of points corresponding to each campaign it is clear that the campaigns do not overlap, with campaign 1 (circles) being more consistent than campaign 2 (crosses). The different clusters in campaign 2 are the result of discriminating process variables being operated with high variability and around different set points (See also Figures 4 and 7).

To understand which of the process variables has been operated in a different way in the two campaigns, the coefficient plot with 95% jack-knife confidence intervals (Efron and Gong 1983) shown in Figure 3 is a useful tool. In this case almost all the process variables are statistically significant (p-value<0.05) as their corresponding 95% jack-knife confidence intervals do not contain the cero value. This means that their

behavior is statistically different in the two campaigns. Figure 4 shows a time series chart of one of these discriminating variables (signaled with an arrow in Figure 3) as an example.

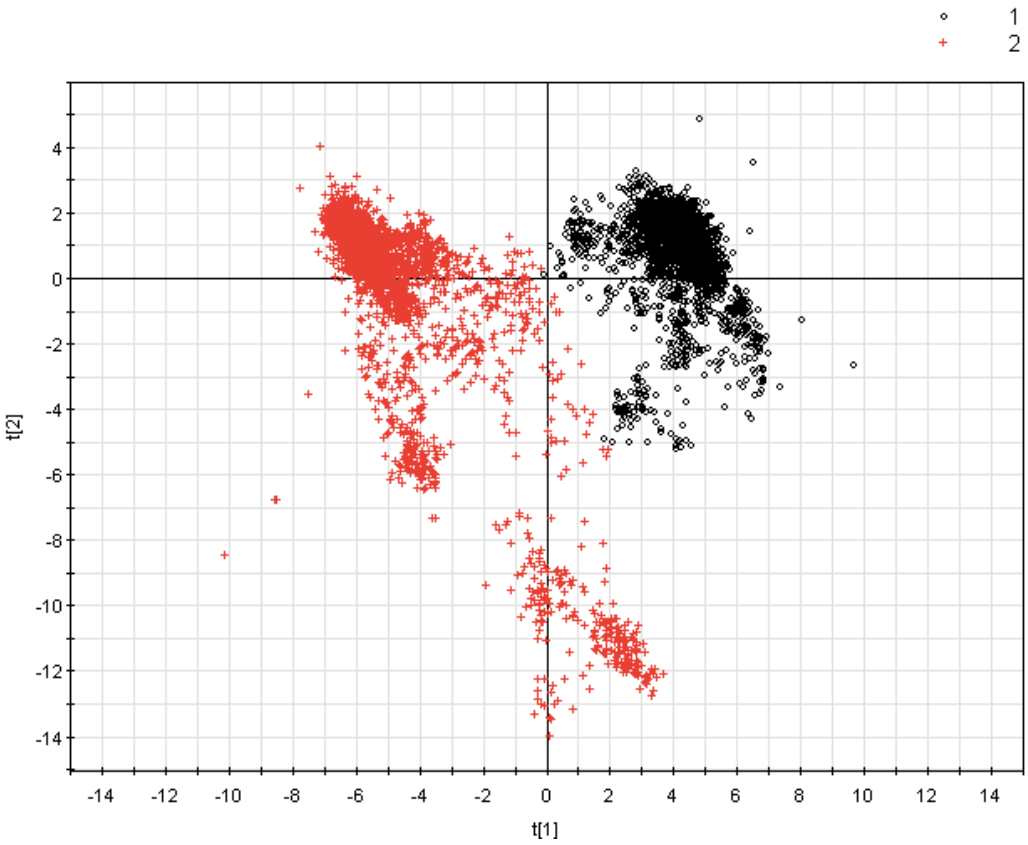

Figure 2: Score plot of the two components of the PLS-DA model: campaign 1 (circles), campaign 2 (crosses).

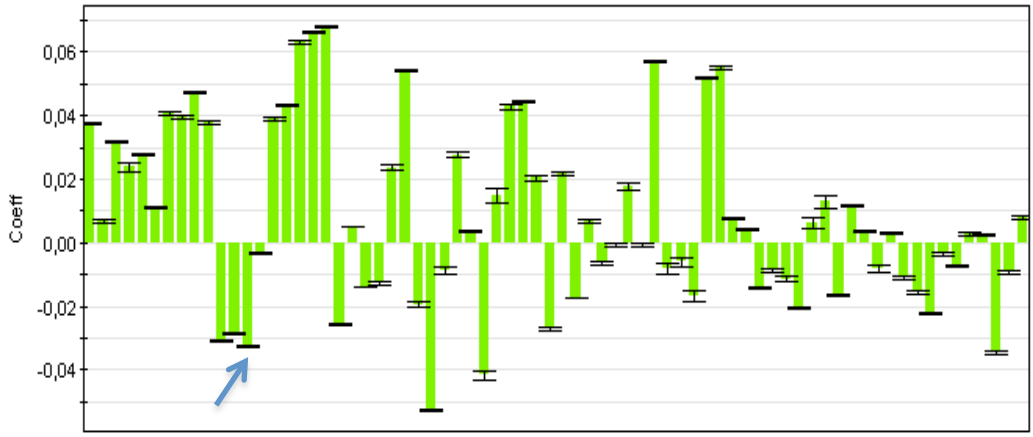

Figure 3: Coefficient plot of the PLS-DA model with 95% jack-knife confidence intervals for each one of the 72 process variables measured.

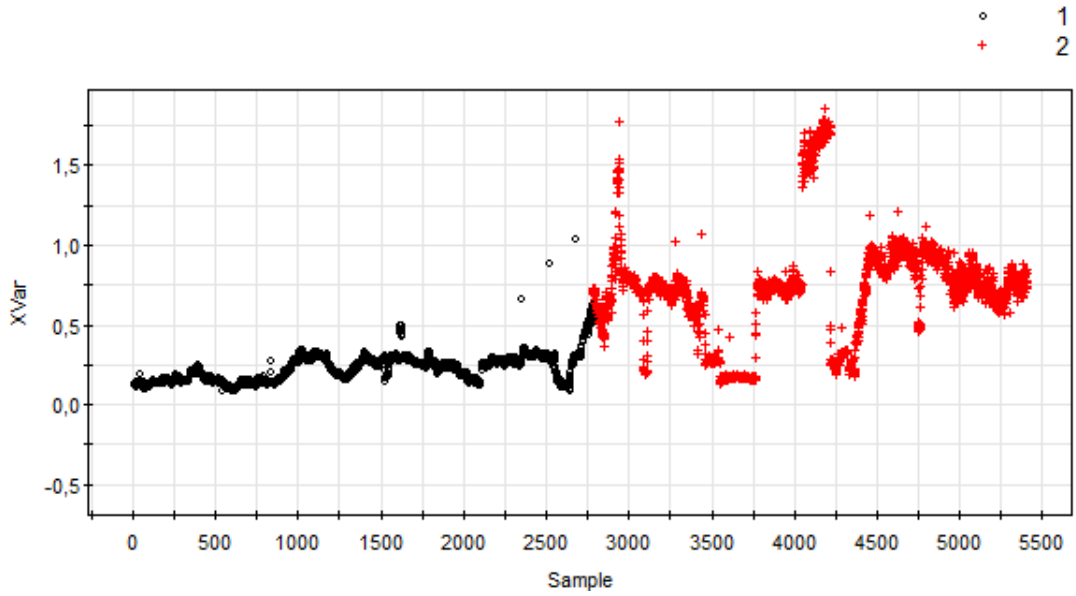

Figure 4: Time series plot of one of the process variables (measured every hour and signaled with an arrow in Figure 3) with different behavior in the two campaigns: 1 (circles), 2 (crosses).

Process engineers realized that both campaigns yielded similar performance in the critical to cost characteristic produced under different operational policies. Therefore, the same performance for the output variable does not necessarily mean a similar process. Contrary to the common belief, this case study revealed that operators were implementing different operational rules, leading to different costs and safety between campaigns.

After comparing both campaigns, the operators considered campaign 1 more appropriate in terms of stability, cost and safety. A latent structured-based MSPC scheme was built using process variables data from this campaign. For technical details, see Phase I Section in Appendix. In order to illustrate the ability of this LSb-MSPC to detect and diagnose future out-of-control events, data from campaign 2 were projected onto the reference model from campaign 1 as if they were future observations from the process. For technical details, see Phase II Section in Appendix.

Figure 5 shows the residual standard deviation (DModX) control chart (see Eq. A.11 in Appendix) with upper control limit (UCL) at 5% significance level (see Eq. A.13 in Appendix) displaying in-control data from campaign 1 (circles) and out-of-control data from campaign 2 (crosses). (Note that although there are some DModX values higher than UCL in campaign 1 (circles) most of them are close to the UCL an obey the in-control average number of false alarm observations -in this case, approx. 0.05 x 2750= 137.5). In order to isolate the measured variables responsible for the out-of-control signals, Figure 6 shows the contribution plot to the DModX (see Eq. A.15 in Appendix) for two out-of-control observations from campaign 2: # 3059 and #3475 (signaled in Figure 5).

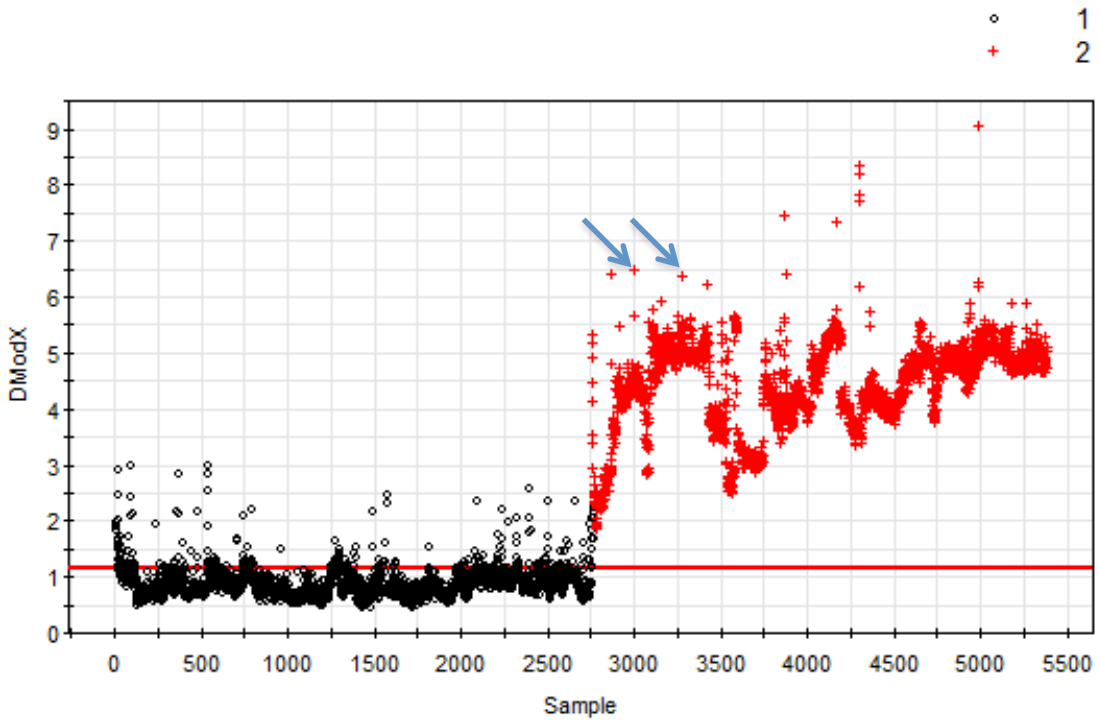

Figure 5: DModX control chart: campaign 1 (circles), campaign 2 (crosses). Upper control limit (UCL) at 5% significance level (red line).

Figure 6: Contribution plot to the DModX for two out-of-control observations from campaign 2: # 3059 (top) and #3475 (bottom).

From Figure 6 it is clear that the same two process variables (signaled with an arrow) have the highest contributions to the DModX in both out-of-control observations. The opposite signs in the contributions mean that their behavior in the two campaigns are opposite as shown in Figure 7.

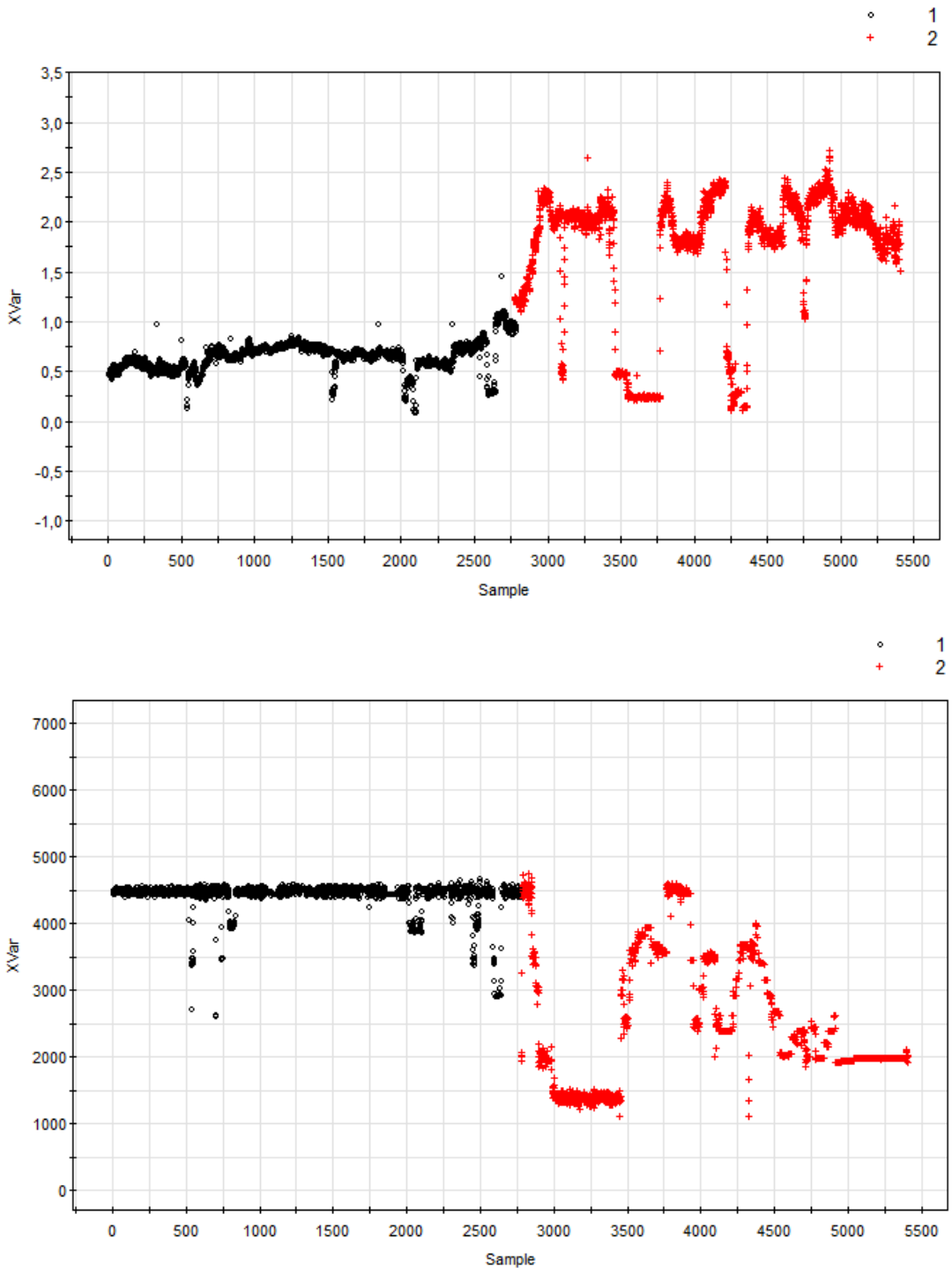

Figure 7: Time series plot of the two process variables (measured every hour) with a high contribution to the DModX (signaled with an arrow in Figure 6): 1 (circles), 2 (crosses).

4. POTENTIAL AND CHALLENGES OF LSb-MSPC IN DATA-RICH ENVIRONMENTS

As commented by Kourti (2005), multivariate monitoring schemes based on projection to latent structures methods have been receiving increasing attention by industrial practitioners in the last 20 years. Several companies have enthusiastically adopted the methods and have reported many success stories. Latent structures-based MSPC has been used for process monitoring, fault detection and diagnosis in continuous and batch industrial processes, and there is a large list of publications reporting successful industrial applications in different fields (petrochemical, polymer, chemical).

In pharmaceutical industry, latent variables-based MSPC is playing a critical role in the successful implementation of process analytical technology (PAT) supported by the United States Food and Drug Administration (FDA). PAT has been defined by the FDA as "*systems for the analysis and control of manufacturing processes based on timely measurements during processes of critical quality parameters and performance attributes of raw and in-process materials and processes to assure acceptable end product quality at the completion of the process"* (http://www.fda.gov/cder/guidance). With this initiative, the FDA tries to motivate the pharmaceutical industry to improve process control strategies for high-quality, cost-effective pharmaceutical products. Kourti (2006) discussed the critical role of latent variables-based MSPC methods for process understanding, abnormal situation detection and fault diagnosis, as linked to PAT.

LSb-MSPC techniques can also be used to extract subtle information from digital images related to product quality. Such information can be used for prediction, monitoring and control. The methodology, based on multivariate image analysis (MIA) (Geladi and Grahn 1996), can be used to monitor solids and other heterogeneous materials (lumber, steel sheets, pulp and paper products, polymer films, multiphase streams, etc.). Image analysis provides non-invasive, informative, inexpensive and robust on-line sensors for the solids and liquids industry. Therefore it opens new ways for the successful monitoring and control of processes that are traditionally difficult due to lack of sensors. Several applications have been reported in the literature: on-line monitoring of time varying images and lumber defects (Bharati and MacGregor 1998,

2003); monitoring and control of the amount of coating applied to the base food product and the distribution of the coating among the individual product pieces in snack food industry (Yu and MacGregor 2003, Yu *et al.* 2003); defect detection and classification in ceramic tiles, artificial stone countertops and orange fruits (Prats-Montalbán and Ferrer 2007). Duchesne *et al.* (2012) provide a highly recommended review of techniques and applications of LSb-MSPC applied to digital images in the process industries. For those interested in using LSb-MSPC with hyperspectral digital images see, e.g. Prats-Montalbán *et al.* (2011).

CONCLUSIONS

As commented by Bisgaard: "*The confluence of the four technological developments of (a) cheap and powerful computing hardware, (b) powerful and easy-to-use statistical software and statistical graphics, (c) easy and cheap transfer and storage of massive amounts of data and (d) the proliferation of sensor technology ... has already transformed industrial statistics, especially in the past 20 years, and will continue to exert dramatic changes to industrial statistics, the applications and the theory and tool developments*" (Steinberg 2008, p. 111).

"The role of industrial statistics is fundamentally as a catalyst to systematic data-driven knowledge generation, learning and innovation in the industrial environment. As such, statistics will remain a powerful general-purpose technology applicable to all sectors of industry, service, and manufacturing…. The quest for quality, whether in services or manufacturing, likely will continue to inspire developments in such areas as ... statistical process monitoring and control" (Bisgaard in Steinberg 2008, pp.106-107).

"Many modern systems generate enormous amounts of useful data during their routine operation, including information about the operating environment and use conditions, logistics, and the state of the system. Often such data can be accessed remotely allowing remote monitoring and diagnostics. ... We have now the capability of, for example, diagnosing (often remotely) the health of products and systems, such as airplanes and locomotives. This provides information for "just-in-time" maintenance or systems shutdown to avoid failures. We foresee such technology extended to humans, equipping people with monitors that provide forewarning of an impending heart attack

even before the patient feels any symptoms. For both products and humans, the goal is the same - recognize a "signal" rapidly, while minimizing the false alarm rate" (Hahn-Doganaksoy in Steinberg 2008, p.114).

"The biggest challenge that I see is to create better statistical methods, especially more intuitive and easier-to-understand multivariate and multivariate time series techniques, software combined with statistical graphics to enable the user to understand the implications of the often complex multivariate statistical analysis and modelling" (Bisgaard in Steinberg 2008, p. 117).

The age of conventional uni- and multivariate SPC has gone. Monitoring one-variable-at-a-time is not only ineffective but also practically unfeasible in data-rich environments. Using conventional MSPC is not suitable due to the instability problems of the Hotelling´s T^2 statistics caused by the ill-conditioning of the covariance matrix of the measured variables. This approach can be mathematically unfeasible if the covariance matrix is singular. Modern manufacturing processes where hundreds of variables are registered through automated in-process sensing calls for a paradigm shift. It is crucial to recover the central role of the process improvement (not the model refinement), rooted in Shewhart´s basic fundamentals. The use of latent structures-based methods for the last 15 years has revolutionized the idea of MSPC. These are flexible tools able to face the challenges stated in Section 2 for both continuous and batch processes: dealing with multivariate autocorrelated processes, multi-step processes, model updating, missing data, traceability and fault diagnosis, among others. This paper intends to generalize the practical use of these powerful tools in massive-data environments typical of modern processes of the 21st century.

Finally, I would like to quote part of a personal communication John Stu Hunter sent me some days before the conference honoring his 90th birthday: "*I fully agree that our modern data sources have long outrun the Shewhart days of univariate and supposed independent observations. Your paper supplies a wonderful stimulus and introduction to the needs of modern QC*".

REFERENCES

Alt, F.B. (1985). Multivariate Quality Control. In *The Encyclopedia of Statistical Sciences*, Kotz S.; Johnson N.L.; Read C.R., Eds.; Wiley: New York, pp 110-122.

Aparisi, F., Jabaloyes, J., Carrión, A. (1999). Statistical properties of the $|\mathbf{S}|$ multivariate control chart. *Communications in Statistics – Theory and Methods*, 28:2671-2686.

Apley, D.W., Shi, J. (2001). A Factor-Analysis Method for Diagnosing Variability in Multivariate Manufacturing Processes. *Technometrics,* 43(1):84-95.

Arteaga, F., Ferrer, A. (2002). Dealing with Missing Data in MSPC: Several Methods, Different Interpretations, Some Examples. *Journal of Chemometrics*, 16:408-418.

Bersimis, S., Psarakis, S., Panaretos, J. (2007) Multivariate Statistical Process Control Charts: An Overview. *Quality and Reliability Engineering International*, 23: 517-543.

Bharati, M.H., MacGregor, J.F. (1998). Multivariate image analysis for real time process monitoring and control. *Industrial and Engineering Chemistry Research*, 37:4715–4724.

Bharati, M., MacGregor, J.F. (2003). Softwood lumber grading through on line multivariate image analysis. *Industrial and Engineering Chemistry Research*, 42:5345–5353.

Bisgaard, S. (2012). The Future of Quality Technology: From a Manufacturing to a Knowledge Economy & From Defects to Innovations. *Quality Engineering*, 24:30–36.

Box, G.E.P. (1954). Some Theorems on Quadratic Forms Applied in the Study of Analysis of Variance Problems: Effect of Inequality of Variance in One-Way Classification. *The Annals of Mathematical Statistics,* 25:290-302.

Box, G.E.P., Hunter, W.G., Hunter, J.S. (2005). *Statistics for Experimenters*: *Design, Innovation, and Discovery*, Wiley: New York.

Box, G.E.P., Luceño, A. (1997). *Statistical Control by Monitoring and Feedback Adjustment*; Wiley: New York.

Camacho, J., Ferrer, A. (2012). Cross-Validation in PCA Models with the Element-Wise *k*-Fold (*ekf*) Algorithm: Theoretical Aspects. *Journal of Chemometrics*, 26(7):361-373.

Duchesne, C., Liu, J.J., MacGregor, J.F. (2012). Multivariate image analysis in the process industries: A review. *Chemometrics and Intelligent Laboratory Systems*, 117:116-128.

Efron, B., Gong, G., (1983). A Leisurely Look at the Bootstrap, the Jack-knife, and Cross-validation. *American Statistician*, 37:36-48.

Eriksson, L., Johansson, E., Kettaneh-Wold, N, Wold, S. (2001). *Multi- and Megavariate data analysis: principles and applications*. Umetrics AB.

Ferrer, A. (2007). Multivariate Statistical Process Control based on Principal Component Analysis (MSPC-PCA): Some Reflections and a Case Study in an Autobody Assembly Process. *Quality Engineering*, 19:311-325.

Fuchs, C., Kenett, R.S. (1998). *Multivariate Quality Control: Theory and Applications* (Quality & Reliability 54). Marcel Dekker: New York.

Geladi, P., Kowalski, B.R. (1986). Partial least-squares regression: a tutorial. *Analytica Chimica Acta*, 185:1–17.

Geladi, P., Grahn, H. (1996). *Multivariate Image Analysis*. Wiley: Chichester, U.K.

Guidance for Industry, PAT A Framework for Innovative Pharmaceutical Development, Manufacturing, and Quality Assurance, U. S. Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research (CDER), Center for Veterinary, Medicine (CVM), Office of Regulatory Affairs (ORA), Pharmaceutical CGMPs, September 2004. (http://www.fda.gov/cder/guidance).

Helland, I.S. (1988). On the structure of partial least squares. *Communications in Statistics—Simulation and Computation*, 17(2):581–607.

Hunter, J. S. (1986). The Exponentially Weighted Moving Average, *Journal of Quality Technology*, 18:203-210.

Höskuldsson, A. (1988). PLS regression methods. *Journal of Chemometrics*, 2:211–228.

Hotelling, H. (1947). Multivariate Quality Control. *Techniques of Statistical Analysis*, C. Eisenhart, M. Hastay, W.A. Wallis (eds.). MacGraw-Hill: New York:111-184.

Jackson, J.E. (1985). Multivariate Quality Control. *Commun. Statistic.-Theor. Meth.,* 14(11):2657-2688.

Jackson, J. E. (2003). *A User´s Guide to Principal Components.* Wiley: New York.

Jackson, J.E., Mudholkar G.S. (1979). Control Procedures for Residuals Associated With Principal Component Analysis. *Technometrics,* 21(3):341-349.

Kourti, T. (2002). Process Analysis and Abnormal Situation Detection: From Theory to Practice. *IEEE Control Systems Magazine*, October:10-25.

Kourti, T. (2005). Application of Latent Variable Methods to Process Control and Multivariate Statistical Process Control in Industry. *Int. J. Adapt. Control Signal Process*, 19:213-246.

Kourti, T. (2006). Process Analytical Technology Beyond Real-Time Analyzers: The Role of Multivariate Analysis. *Critical Reviews in Analytical Chemistry*, 36:257-278.

Kourti, T., MacGregor, J.F. (1996). Multivariate SPC Methods for Process and Product Monitoring. *Journal of Quality Technology,* 28(4):409-428.

Liu, R. (1995). Control Charts for Multivariate Processes. *Journal of the American Statistical Association*, 90:1380-1388.

Liu R.Y., Singh K., Teng J.H. (2004). DDMA-Charts: Nonparametric Multivariate Moving Average Control Charts Based on Data Depth, *Allgemeines Statisches Archiv*, 88:235-258.

Liu, R.Y., Tang, J. (1996). Control charts for dependent and independent measurements based on bootstrap methods. *Journal of the American Statistical Association*, 91:1694-1700.

Lowerse, D.J., Smilde, A. K. (2000). Multivariate Statistical Process Control of Batch Processes Based on Three-Way Models. *Chemical Engineering Science*, 55:1225-1235.

Lowry, C.A., Montgomery, D.C. (1995). A Review of Multivariate Control Charts. *IIE Transactions*, 27:800-810.

Lowry, C.A., Woodall, W.H., Champ, C.W., Rigdon, S.E. (1992). A Multivariate Exponentially Weighted Moving Average Control Chart. *Technometrics,* 34:46-53.

MacGregor, J.F. (1997). Using On-Line Process Data to Improve Quality: Challenges for Statisticians. *International Statistical Review*, 65:309-323.

Martens, H., Naes, T. (1989). *Multivariate Calibration*. Wiley: New York.

Mason, R.L., Champ, C.W., Tracy, N.D., Wierda, S.J., Young, J.C. (1997). Assessment of Multivariate Process Control Techniques. *Journal of Quality Technology,* 29(2):140-143.

Montgomery, D.C. (2005). *Introduction to Statistical Quality Control,* 5th ed. Wiley: New York.

Montgomery, D. C., Woodall, W. H. (eds.) (1997). A Discussion on Statistically-Based Process Monitoring and Control. *Journal of Quality Technology*, 29(2):121–162.

Nelson, P.R.C., Taylor, P.A., MacGregor, J.F. (1996). Missing data methods in PCA and PLS: score calculations with incomplete observations. *Chemometrics Intell. Lab. Syst*., 35:45–65.

Nomikos, P., MacGregor, J.F. (1995). Multivariate SPC Charts for Monitoring Batch Processes. *Technometrics,* 37(1):41-59.

Prats-Montalbán, J.M., Ferrer, A., Malo, J.L., Gorbeña, J. (2006). A comparison of different discriminant analysis techniques in a steel industry welding process. *Chemometrics and Intelligent Laboratory Systems* 80:109 – 119.

Prats-Montalbán, J.M., Ferrer, A. (2007). Integration of colour and textural information in multivariate image analysis: defect detection and classification issues. *Journal of Chemometrics*, 21:10-23.

Prats-Montalbán, J. M., de Juan, A., Ferrer, A. (2011). Multivariate image analysis: a review with applications, *Chemometrics and Intelligent Laboratory Systems*, 107:1–23.

Shewhart, W. A. (1931). *Economic Control of Quality of Manufactured Product*. D. Van Nostrand, New York, NY. (Republished in 1980 by the American Society for Quality Control, Milwaukee, WI).

Sjöström, M., Wold, S., Söderström, B. (1986). PLS discriminant plots, *Proceedings of PARC in Practice*, Amsterdam, June 19– 21, 1985, Elsevier Science Publishers B.V., North-Holland.

Steinberg, D.M. (ed.) (2008). The Future of Industrial Statistics: A Panel Discussion. *Technometrics*, 50(2):103-127.

Stoumbos, Z.G., Reynolds, M.R. Jr., Ryan, T.P., Woodall, W.H. (2000). The State of Statistical Process Control as We Proceed into the 21st Century. *Journal of the American Statistical Association*, 95(45):992-998.

Tracy, N.D., Young, J.C., Mason, R.L. (1992). Multivariate Control Charts for Individual Observations. *Journal of Quality Technology,* 24(2):88-95.

Vidal-Puig, S. Ferrer, A. (2013). A Comparative Study of Different Methodologies for Fault Diagnosis in Multivariate Quality Control. *Communications in Statistics-Simulations and Computations* (in press).

Wierda, S.J. (1994). Multivariate Statistical Process Control – Recent Results and Directions for Future Research. *Statistica Neerlandica,* 48(2):147-168.

Wold, H. (1966). Nonlinear Estimation by Iterative Least Squares Procedures. In *Research papers in Statistics*, David, F. (ed.), Wiley: New York:411-444.

Wold, S. (1978). Cross-Validatory Estimation of the Number of Components in Factor and Principal Component Models. *Technometrics,* 20(4):397-405.

Woodall, W.H. (2000). Controversies and Contradictions in Statistical Process Control. *Journal of Quality Technology,* 32(4):341-350.

Woodall, W.H., Montgomery, D.C. (1999). Research Issues and Ideas in Statistical Process Control. *Journal of Quality Technology,* 31(4):376-386.

Yu, H.; MacGregor, J.F. (2003). Multivariate image analysis and regression for prediction of coating and distribution in the production of snack foods. *Chemometrics and Intelligent Laboratory Systems*, 67:125-144.

Yu, H., MacGregor, J.F., Haarsma, G., Bourg, W. (2003). Digital Imaging for On-line Monitoring and Control of Industrial Snack Food Processes. *Industrial Chemical Engineering Research*, 42:3036-3044.

APPENDIX: OVERVIEW OF LSb-MSPC

The LSb-MSPC scheme, as any SPC scheme, is carried out in two phases. In Phase I (model building) monitoring charts are built according to a set of historical in-control data, once the performance of the process has been understood and modeled, and the assumptions of its behavior and process stability are checked. In Phase II (model exploitation) these charts are used to monitor the process using on-line data, assuming the form of the distribution to be known along with the values of the in-control parameters (Woodall, 2000).

A.1 Phase I (model building): exploratory data analysis and off-line process monitoring

The main goal in Phase I is to model the in-control process performance based on a set of historical in-control (reference) data. This data set is one in which the process has been operating consistently (stable over time) in an acceptable manner.

Occasionally, this historical in-control data set is not directly available, but has to be extracted from historical databases in an iterative fashion as noted below. This explorative analysis of historical databases is a useful technique for improving process understanding and detecting past faults in the process (out-of-control samples). By correctly diagnosing their root causes, some countermeasures can be implemented, improving the future performance of the process.

Suppose the historical database consists of a set of m multivariate observations (objects or samples of size n=1) on K variables (on-line process measurements, dimensional variables or product quality data) arranged in a ($m \times K$) data matrix **Z**. Variables in matrix **Z** are often pre-processed by mean-centering and scaling to unit variance. With mean-centering the average value of each variable is calculated and then subtracted from the data. This usually improves the interpretability of the model because all pre-processed variables will have mean value zero. By scaling to unit variance each measured variable is divided by its standard deviation and will have unit variance. Given that projection methods are sensitive to scaling, this is particularly useful when the variables are measured in different units. After pre-processing, the matrix **Z** is transformed into matrix **X**.

Principal component analysis (PCA) is used to reduce the dimensionality of the process by compressing the high-dimensional measured data matrix **X** into a low-dimensional subspace of dimension A ($A \leq \text{rank}(\mathbf{X})$), in which most of the data variability is explained by a fewer number of latent variables, which are orthogonal and linear combinations of the measured ones. This is done by decomposing **X** into a set of A rank 1 matrices

$$\mathbf{X} = \sum_{a=1}^{A} \mathbf{t}_a \mathbf{p}_a^{\mathrm{T}} + \sum_{a=A+1}^{\mathrm{rank}(\mathbf{X})} \mathbf{t}_a \mathbf{p}_a^{\mathrm{T}} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E} = \mathbf{X}^{*} + \mathbf{E} \qquad \text{(A.1)}$$

P ($K \times A$) is the *loading* matrix containing the *loading* vectors $\mathbf{p}_a$, which are the eigenvectors, corresponding to the A largest eigenvalues of the covariance matrix of the measured pre-treated data set **X**, and define the directions of highest variability of the new latent A-dimensional subspace. **T** ($m \times A$) is the *score* matrix containing the location of the orthogonal projection of the measured observations onto the latent subspace. The columns $\mathbf{t}_a$ of the score matrix **T** ($\mathbf{t}_a = \mathbf{X}\mathbf{p}_a$) represent the new latent variables with variances given by their respective eigenvalues (λ_a). These new latent variables summarize the most important information of the measured K variables, and thus can predict (reconstruct) **X** with minimum mean square error, $\mathbf{X}^{*} = \mathbf{T}\mathbf{P}^{\mathrm{T}}$. Matrix **E** ($m \times K$)

contains the residuals (statistical noise), i.e. the information that is not explained by the PCA model.

The dimension of the latent variable subspace is often quite small compared with the dimension of the measured variable space (i.e., A << rank (**X**)). Several algorithms can be used to extract the principal components. For large ill-conditioned data sets it is recommended to compute the principal components sequentially via the NIPALS (non-iterative partial least squares) algorithm (Wold, 1966) and to stop based on different criteria (Wold, 1978; Jackson, 2003; Camacho and Ferrer, 2012). Another advantage of NIPALS algorithm is that it easily handles missing data (i.e. observation vectors from which some variable measurements are missing). The quality of the fitted PCA model can be evaluated by computing several parameters, such as R^2, that measures the *goodness of fit*, or Q^2 that indicates the predictive capability of the model (Eriksson *et al*., 2001).

Eq. (A.1) shows that the PCA model transforms each K-dimensional measured observation vector $\mathbf{x}_i$ (i^{th} row of matrix **X**) into an A-dimensional score vector $\mathbf{t}_i^{\mathrm{T}} = \{t_{i1}, t_{i2}, \ldots, t_{iA}\}$ (i^{th} row of matrix **T**) and a residual vector $\mathbf{e}_i$ (i^{th} row of matrix **E**). From the scores and the residuals (prediction errors) associated with each observation, two complementary (orthogonal or independent) statistics are derived: the Hotelling´s T_A^2 and the *SPE* (sum of squared prediction errors). The T_A^2 statistic for the i^{th} observation is defined as

$$T_A^2 = \mathbf{t}_i^{\mathrm{T}} \boldsymbol{\Theta}^{-1} \mathbf{t}_i = \sum_{a=1}^{A} \frac{t_a^2}{\lambda_a} \tag{A.2}$$

where $\boldsymbol{\Theta}$ ($A \times A$) is the covariance matrix of **T** (diagonal matrix of the highest A eigenvalues $\{\lambda_1, \ldots, \lambda_A\}$). This statistic is the Hotelling-T^2 statistic when a reduced subspace with A components is used instead of the measured variables space, and it represents the estimated squared Mahalanobis distance from the center of the latent subspace to the projection of an observation onto this subspace. Under the assumption that the scores follow a multivariate normal distribution (they are linear combinations of random variables), it holds (Tracy *et al*., 1992) that in Phase I, T_A^2 (times a constant) follows a beta (B) distribution

$$T_A^2 \sim \frac{(m-1)^2}{m} B_{A/2,(m-A-1)/2} \tag{A.3}$$

while in Phase II, T_A^2 (times a constant) follows an F distribution

$$T_A^2 \sim \frac{A\left(m^2-1\right)}{m(m-A)} F_{A,(m-A)} \tag{A.4}$$

The difference in both distribution comes from the fact that in Phase I the same observation vectors $\mathbf{x}_i$ collected in the reference data set are used for two purposes: (i) to build the PCA model and work out the control limits of the charts, and (ii) to check whether they fall within these control limits. Therefore, observations in the reference data set are not independent of PCA model parameters used to derive the statistics to be monitored. In contrast, in Phase II new observations (not used for model building) are checked against the control limits calculated from the in-control data, and therefore, independence is guaranteed. Anyway, if a large reference data set is available Eq. (A.4) can also be used for approximating the distribution of the T_A^2 statistic in Phase I.

On the other hand, the *SPE* statistic for i^{th} observation $\mathbf{x}_i$ is given by

$$SPE = \mathbf{e}_i^{\mathrm{T}} \mathbf{e}_i = \left(\mathbf{x}_i - \mathbf{x}_i^*\right)^{\mathrm{T}} \left(\mathbf{x}_i - \mathbf{x}_i^*\right) \tag{A.5}$$

where $\mathbf{e}_i$ is the residual vector of i^{th} observation, and $\mathbf{x}_i^*$ is the prediction of the observation vector $\mathbf{x}_i$ from the PCA model. The *SPE* statistic represents the squared Euclidean (perpendicular) distance of an observation from this subspace, and gives a measure of how close the observation is to the A-dimensional subspace. Assuming that residuals follow a multivariate normal distribution, Box (1954), Jackson and Mudholkar (1979) and Eriksson *et al*. (2001) derived approximate distributions for such quadratic forms.

From these two statistics, in LSb-MSPC two complementary multivariate control charts are constructed. Shewhart-type control charts for individual observations are often used in practice. Nevertheless, other types of rational subgrouping or multivariate charts such as MEWMA charts (Lowry *et al.,* 1992) may be used. The latter may be specially suited for autocorrelated processes.

The control limits of the multivariate control charts are calculated following the traditional SPC philosophy. In Phase I, an appropriate historical or reference set of data is chosen which defines the normal or *in-control* operating conditions for a particular process corresponding to common-cause variation. The in-control PCA model is then built on these data. Any periods containing variations arising from special events that one would like to detect in the future are omitted at this stage. The choice of the reference (in-control) data set is critical to the successful application of the procedure (Kourti and MacGregor, 1996). Control limits for good operation on the control charts

are defined based on this reference data set. In Phase II, values of future measurements are compared against these limits.

Upper control limits (UCL) for the Shewhart T_A^2 chart at significance level (type I risk) α can be obtained for Phase I from Eq. (A.3)

$$UCL\left(T_A^2\right)_\alpha = \frac{(m-1)^2}{m} B_{(A/2,(m-A-1)/2),\alpha} \tag{A.6}$$

where $B_{(A/2,(m-A-1)/2),\alpha}$ is the 100(1-α)% percentile of the corresponding beta distribution that can be computed from the 100(1-α)% percentile of the corresponding F distribution by using the relationship (Tracy *et al*., 1992)

$$B_{(A/2,(m-A-1)/2),\alpha} = \left(A/(m-A-1)\right)F_{(A,m-A-1),\alpha} / \left(1+\left(A/(m-A-1)\right)F_{(A,m-A-1),\alpha}\right) \tag{A.7}$$

For Phase II, the corresponding UCL from Eq. (A.4) is given by

$$UCL\left(T_A^2\right)_\alpha = \frac{A\left(m^2-1\right)}{m(m-A)} F_{(A,(m-A)),\alpha} \tag{A.8}$$

Regarding the UCL for the Shewhart *SPE* chart, several procedures can be used. Jackson and Mudholkar (1979) showed that an approximate *SPE* critical value at significance level α is given by

$$UCL(SPE)_\alpha = \theta_1 \left[\frac{z_\alpha \sqrt{2\theta_2 h_0^2}}{\theta_1} + 1 + \frac{\theta_2 h_0 (h_0 - 1)}{\theta_1^2} \right]^{1/h_0} \tag{A.9}$$

where $\theta_k = \sum_{j=A+1}^{\text{rank}(\mathbf{X})} (\lambda_j)^k$, $h_0 = 1 - \frac{2\theta_1\theta_3}{3\theta_2^2}$, λ_j are the eigenvalues of the PCA residual covariance matrix $\mathbf{E}^T\mathbf{E}/(m\text{-}1)$, and z_a is the 100(1$-\alpha$)% standardized normal percentile.

Alternatively, one can use an approximation based on the weighted chi-squared distribution ($g\chi_h^2$) proposed by Box (1954). Nomikos and MacGregor (1995) suggested a simple and fast way to estimate the parameters g and h, which is based on matching moments between a $g\chi_h^2$ distribution and the sample distribution of *SPE*. The mean ($\mu = gh$) and variance ($\sigma^2 = 2g^2h$) of the $g\chi_h^2$ distribution are equated with the sample mean (b) and variance (v) of the *SPE* sample. Hence, the upper *SPE* control limit at significance level α is given by

$$UCL(SPE)_\alpha = \frac{v}{2b} \chi^2_{(2b^2/v),\alpha} \tag{A.10}$$

where $\chi^2_{(2b^2/v),\alpha}$ is the 100(1-α)% percentile of the corresponding chi-squared distribution.

Another method based on the statistical test of equality of variances from normal distributions is proposed by Eriksson *et al*. (2001). Based on the *SPE*, they define the absolute distance to the model (*DModX*) of an observation as its (corrected) residual standard deviation

$$DModX = c\sqrt{\frac{SPE}{(K-A)}} \tag{A.11}$$

where *c* is a correction factor (function of the number of observations and the number of components) to be used in Phase I. This correction factor takes into account that the distance to the model (*DModX*) is expected to be slightly smaller for an observation in the reference set because it has influenced the model. This correction only matters if the number of observations in the reference set is small. In Phase II, *c*=1.

They also define the normalized distance to the model ($DModX_{norm}$) as

$$DModX_{norm} = \frac{DModX}{s_0} \tag{A.12}$$

where $s_0 = \sqrt{\sum_{i=1}^{m}\sum_{k=1}^{K} e_{ik}^2 \Big/ (m-A-1)(K-A)}$ is the pooled residual standard deviation. This is an estimation of the residual variability taking into account all the observations used to build the model (reference data set).

Assuming that the statistic $(DModX_{norm})^2$ has an approximate *F* distribution with *K-A* and (*m-A-1*)(*K-A*) degrees of freedom for the in-control observations, the UCL for the Shewhart *DModX* chart at significance level α is expressed as

$$UCL(DModX)_\alpha = s_o\sqrt{F_{(K-A,(m-A-1)(K-A)),\alpha}} \tag{A.13}$$

where $F_{(K-A,(m-A-1)(K-A)),\alpha}$ is the 100(1-α)% percentile of the corresponding *F* distribution.

The normality assumption on which these calculations are based is usually quite reasonable in practice. Anyway, control limits for multivariate charts can be obtained from distribution-free methods by repeated sampling. The only requirement is to have a large in-control data set from which the external reference distribution (Box *et al.,* 1978) for any statistic can be obtained.

The two multivariate control charts (T_A^2 and *SPE*) differ in their conceptual meaning. They are two complementary indices that provide a picture of the wellness of

the process at a glance (Kourti, 2005). The T_A^2 chart checks if the projection of an observation on the hyperplane defined by the latent subspace is within the limits determined by the reference (in-control) data. Thus, a value of this statistic exceeding the control limits indicate that the corresponding observation presents abnormal extreme values in some (or all) of its measured K variables, even though it maintains the correlation structure between the variables in the model. This observation can be tagged as an abnormal outlier *inside* the PCA model (an *extremist* or *severe* outlier) (Martens and Naes 1989). On the other hand, the *SPE* chart checks if the distance (noise variation) of an observation to the latent hyperplane is inside the control limits. The *SPE* chart values exceeding the control limits are related to observations that do not behave in the same way as the ones used to create the model (in-control data), in the sense that there is a breakage of the correlation structure of the model. This chart will detect the occurrence of any new event that cause the process to move away from the hyperplane defined by the reference model. This kind of observations can be tagged as outliers *outside* the model (an *alien* or *moderate* outlier) (Martens and Naes, 1989).

Severe outliers are influential observations with high leverage on the model, i.e., with strong *power* to pull the principal directions toward themselves, creating fictitious components and misleading the PCA model (Eriksson *et al*., 2001). Therefore, *model validation* is critical in the Phase I stage in order to remove from the data matrix these dangerous outlier (out-of-control) observations and, afterwards, recalculate the PCA model. Before removing any observation from the data matrix, some diagnostics using contribution plots (discussed below) and process insight should be used in order to sort out false alarm outliers from the *real* ones. This process of model building and validation is done iteratively until no multivariate control chart signals any *real* outlier. As a side effect from this debugging procedure, the root causes of the out-of-control observations can be discovered, improving process knowledge and future process performance.

A.2 Phase II (model exploitation): on-line process monitoring

Once the reference PCA model and the control limits for the multivariate control charts are obtained, new process observations can be monitored on-line. When a new observation vector $\mathbf{z}_i$ is available, after pre-processing it is projected onto the PCA

model yielding the scores and the residuals, from which the value of the Hotelling´s T_A^2 and the value of the *SPE* are calculated. This way, the information contained in the measured *K* variables is summarized in these two indices that are plotted in the corresponding multivariate T_A^2 and *SPE* control charts. No matter what the number of the measured variables *K* is, only two points have to be plotted on the charts and checked against the control limits. The SPE chart should be checked first. If the points remain below the control limits in both charts the process is considered to be in-control. If a point is detected to be beyond the limits of one of the charts, then a diagnostic approach to isolate the measured variables responsible for the out-of-control signal is needed. In LSb-MSPC contribution plots (Kourti and MacGregor 1996) are commonly used for this purpose.

Contribution plots can be derived for abnormal points in both charts. If the *SPE* chart signals a new out-of-control observation, the contribution of each measured k^{th} variable to the *SPE* at this new abnormal observation is given by its corresponding squared residual

$$Cont\left(SPE; x_{new,k}\right) = e_{new,k}^2 = \left(x_{new,k} - x_{new,k}^*\right)^2 \tag{A.14}$$

In case of using the distance to the model (*DMODX*) statistic, the contribution of each measured k^{th} variable to the *DMODX* is given by (Eriksson *et al*., 2001)

$$Cont\left(DMODX; x_{new,k}\right) = w_k \, e_{new,k} \tag{A.15}$$

where w_k is the square root of the explained sum of squares for the k^{th} variable. Variables with high contributions in this plot should be investigated.

If the abnormal observation is detected by the T_A^2 chart the diagnosis procedure is carried out in two steps: (i) a bar plot of the normalized scores for that observation $(t_{\text{new},a}/\lambda_a)^2$ is plotted and the a^{th} score with the highest normalized value is selected; (ii) the contribution of each measured k^{th} variable to this a^{th} score at this new abnormal observation is given by

$$Cont\left(t_{new,a}; x_{new,k}\right) = p_{ak} x_{new,k} \tag{A.16}$$

where p_{ak} is the loading of the *k*-th variable at component *a*. A plot of these contributions is created. Variables on this plot with high contributions but with the same sign as the score should be investigated (contributions of the opposite sign, will only make the score smaller). When there are some scores with high normalized values, an

overall average contribution per variable can be calculated, over all the selected scores (Kourti, 2005).

Contribution plots are a powerful tool for fault diagnosis. They provide a list of the process variables that contribute *numerically* to the out-of-control condition (they are no longer consistent with normal operating conditions), but they do not reveal the actual cause of the fault. Those variables and any variables highly correlated with them should be investigated. Incorporation of technical process knowledge is crucial to diagnose the problem and discover the root causes of the fault.

Apart from the T_A^2 and *SPE* control charts, other charts such as the univariate time series plots of the scores, or scatter score plots can be useful (both in Phase I and II) for detecting and diagnosing out-of-control situations and also for improving process understanding.