Variable Selection Methods for Multivariate Process Monitoring

Luan Jaupi

Abstract—**In the first stage of a manufacturing process a large number of variables might be available. Then, a smaller number of measurements should be selected for process monitoring. At this point in time, variable selection methods for process monitoring have focused mainly on explained variance performance criteria. However, explained variance efficiency is a minimal notion of optimality and does not necessarily result in an economically desirable selected subset, as it makes no statement about the measurement cost or other engineering criteria. Without measuring cost many decisions will be impossible to make. In this article, we propose two new methods to select a reduced number of relevant variables for multivariate statistical process control that makes use of engineering, cost and variability evaluation criteria. In the first method we assume that a two-class system is used to classify the variables as primary and secondary based on different criteria. Then a double reduction of dimensionality is applied to select relevant primary variables that represent well the whole set of variables. In the second methodology a cost-utility analysis is used to compare different variable subsets that may be used for process monitoring. The objective of carrying out a cost–utility analysis is to compare one use of resources with other possible uses. To do this, to any process monitoring procedure is assigned a score calculated as ratio of the cost at which it might be obtained to explained variance that it might provide. The subset of relevant variables is selected in a manner that retains, to some extent, the structure and information carried by the full set of original variables. A real application from automotive industry will be used to illustrate the proposed methods.**

Index Terms—**Process control, dimension reduction, variance efficiency, cost-utility analysis, influence function.**

I. Introduction

THE aim of Statistical Process Control, SPC, is to bring a production process under control and keep it in stable condition to ensure that all process output is conforming. This under control state is achieved by monitoring process through measurements of selected variables. When large number of variables are available, it is natural to enquire whether they could be replaced by a fewer number of measurements without loss of much information.
Examples of situations in which variable selection is necessary can be found in [4], [19].

A two stage methodology to select a subset of variables that retains as much information on the full set of variables as possible, assuming that all variables are equally important according to engineering and economic criteria is given in [6]. However, in many cases measured variables are not equally important according to given criteria. For example, according to some engineering criteria some variables may be very important for the functionality of the part and others less important, or some variables may be easier and cheaper to carry out then others or some variables may be more efficient in waste reduction because their measurement are made at earlier points in the process. Neglecting this information in SPC would be counterproductive. At this point in time, there is a gap in the SPC literature devoted to statistical selection of variables in conjunction with given engineering or economic criteria.

In this article, we propose two new methods to select a reduced number of relevant variables for multivariate SPC that makes use of engineering, cost and variability evaluation criteria. In the first method we assume that a two-class system is used to classify the variables as primary and secondary based on different criteria. Then a double reduction of dimensionality is applied to select relevant primary variables that represent well the whole set of variables. The selection methodology uses external information to influence the selection process. The subset of relevant variables is selected in a manner that retains, to some extent, the structure and information carried by the full set of original variables, thereby providing a SPC almost as efficient as we were monitoring all original variables. The proposed method is a stepwise procedure. Various variable selection procedures might be used to select relevant primary variables. In this article we propose a backward elimination scheme, which at each step eliminates the less informative variable among the primary variables that have not yet been eliminated. The new variable is eliminated by its inability to supply complementary information for the whole set of variables. To achieve this we propose the use of Principal Components, PCs, which are computed using only the selected subset of primary variables, but represent well the whole set of variables. This strategy mitigates the risk that an assignable cause inducing a shift, that lies entirely in the discarded variables, will go undetected. To find such PCs we use Rao’s approach on principal components of instrumental variables [17].

In the second methodology a cost-utility analysis is used to compare different variable subsets that might be used for process monitoring. The objective of carrying out a cost–utility analysis is to compare one use of resources with other possible uses. To do this, to any process monitoring procedure is assigned a score calculated as ratio of the cost at which it might be obtained to explained variance that it might emanate. The subset of relevant variables is selected in a manner that retains, to some extent, the structure and information carried by the full set of original variables.

L. Jaupi is with Conservatoire National des Arts et Métiers, 292 rue Saint Martin, 75003 Paris, France (corresponding author phone: +33 1 40 27 23 73; fax: +33 1 40 27 27 46; e-mail: jaupi@cnam.fr).

II. METHOD 1: VARIABLE SELECTION WITH PRE-ASSIGED ROLES

A. Formulation

In the first stage of a manufacturing process a large number of variables might be available. Then, a smaller number of measurements should be selected for process monitoring. In what follows we suppose that $\mathbb{X}$ =(X_1,X_2,…,X_m) is the vector of initial stage measured variables, with mean μ and covariance matrix Σ. We collect n observations and let X be the n×m matrix of in-control data. When a large number of measurements are available, it is natural to investigate whether they could be replaced by a fewer number of variables. In the proposed methodology we assume that a two-class system is used to classify the variables as *primary* and *secondary* based on different criteria. For example according to some measurement cost criteria some variables may be easier and cheaper to carry out then others or some variables may be more efficient in waste reduction because their measurement are made at earlier points in the process. Without loss of generality let $\mathbb{C}_1$=(X_1,X_2,…,X_p) and $\mathbb{C}_2$=(X_{p+1},…,X_m) be the sets of primary and secondary variables respectively. We may write $\mathbb{X}$=($\mathbb{C}_1$,$\mathbb{C}_2$). Our goal is to find a subset $\mathbb{X}_1$ of c primary variables (c≤p), which best in some sense represents the whole set of original variables $\mathbb{X}$. PCs that are based on the selected subset of primary variables are suggested for this purpose as an appropriate tool for deriving low-dimension subspaces which capture most of the information of the whole data set. For the case $\mathbb{C}_1$=$\mathbb{X}$, several selection methods have been suggested in different contexts (see for example [3], [5], [6], [8], [13], [14], [15], [16], [18]). Suppose that $\mathbb{X}_1$ is the selected subset of primary variables and similarly $\mathbb{X}_2$ the subset of remaining variables. We may write $\mathbb{X}$=($\mathbb{X}_1$,$\mathbb{X}_2$). Let (μ_1, Σ_{11}) and (μ_2, Σ_{22}) denote the location scale parameters of $\mathbb{X}_1$, and $\mathbb{X}_2$ respectively. We have the following expressions for μ and Σ

$$\mu = (\mu_1, \mu_2) \qquad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \tag{1}$$

Consider a transformation:

$$Y = \mathbb{X}_1 A \tag{2}$$

where A is a matrix of rank q. The residual dispersion matrix of X after subtracting its best linear predictor in terms of Y is

$$\Sigma_{res} = \Sigma - \Theta_1^t A (A^t \Sigma_{11} A)^{-1} A^t \Theta_1 \tag{3}$$

where $\Theta = (\Sigma_{11}, \Sigma_{12})$.

In this article we propose a variable selection procedure based on PCs, which are computed as linear combinations of selected subset, but are optimal with respect to a given criterion measuring how well each subset approximates all variables including those that are not selected. For a given q we wish to determine A such that the predictive efficiency of Y for X is maximum. Using as overall measure of predictive efficiency the trace operator we have the following solution: the columns of matrix A consist of q first eigenvectors of the following determinant equation:

$$\left| (\Sigma_{11}^2 + \Sigma_{12}\Sigma_{21}) - \lambda \Sigma_{11} \right| = 0 \tag{4}$$

Assuming that $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_c$ are the ordered eigenvalues and denoting by $\alpha_1, \alpha_2, ..., \alpha_c$ the associated eigenvectors, the matrix A is given as following $A = (\alpha_1, \alpha_2, ..., \alpha_q)$, [17].

B. Variability Evaluation Criteria

There are several measures to summarize the overall multivariate variability of a set of variables. The choice of indices will depend on the nature and goals of specific aspect of data analysis but the most popular ones are based on trace operator, generalized variance and squared norm of the dispersion matrix. Al-Kandari and Jolliffe [1], [2], have investigated and compared the performance of several selection methods and their results showed that the efficiency of selection methods is dependent on the performance criterion. Furthermore they noted that it may be not wise to rely on a single method for variable selection. In practice it is necessary to know how well Y approximates the whole data set X. A suitable criterion for this purpose is the proportion of variability explained by the best q space spanned by the selected subset $\mathbb{X}_1$ given by:

$$RX = \frac{\lambda_1 + \lambda_2 + ... + \lambda_q}{trace(\Sigma)} \tag{5}$$

Classical Principal Components Analysis, PCA, results guarantee that the maximum value of the right hand of (5) is attained for $\mathbb{X}_1$=$\mathbb{X}$. The index RX is useful to quantify how much information the selected variables have about the whole set of variables. However, it does not tell us how much information the selected variables have about the unselected ones. This information cannot be found in Σ_{res} but it can be found in conditional covariance matrix of subset $\mathbb{X}_2$ given Y, denoted as $\Sigma_{\mathbb{X}_2/Y}$ given by:

$$\Sigma_{\mathbb{X}_2/Y} = \Sigma_{22} - A^t \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12} A \tag{6}$$

We then propose the use of a second variability evaluation criterion defined as:

$$RX_2 = 1 - \frac{\lambda_1^{'} + \lambda_2^{'} + ... + \lambda_{m-c}^{'}}{trace(\Sigma_{22})} \tag{7}$$

where $\lambda_1^{'}, \lambda_2^{'}, ..., \lambda_{m-c}^{'}$ are eigenvalues of $\Sigma_{\mathbb{X}_2/Y}$. The criterion RX_2 is similar to index R_{EX} defined in [6]. It grows both with the variance of the selected variables as well as with the variance of the unselected ones explained by the selected variables. If RX_2 is near zero it shows that the subspaces spanned by $\mathbb{X}_1$ and $\mathbb{X}_2$ are almost orthogonal and the sets of variables $\mathbb{X}_1$ and $\mathbb{X}_2$ describe different phenomena of the same process. Therefore a shift in the unselected variables could not be detected by the selected subset. Conversely, a high RX_2 value will guarantee that the selected variables may provide a SPC almost as efficient as if we were monitoring all *m* variables.

C. Variable Selection Algorithm

Various variable selection procedures might be applied to select relevant primary variables and then find PCs which are based on them but represent well the whole set of variables. Here we propose a backward elimination scheme:

1. Compute dispersion matrix of the whole data set.
2. Based on $\mathbb{C}_1$ calculate PCs that explain well the whole set of original variables $\mathbb{X}$, [17].
3. Looking carefully at eigenvalues and the cumulative proportions, determine the number of PCs to be used.
4. Remove each one among the p variables in $\mathbb{C}_1$ in turn, and solve p eigenvalue problems, (4), with (p-1) variables.
5. Find the best subset of size (p-1) according to selection criterion that is used and remove the corresponding variable.
6. Put p=(p-1) and continue backward elimination till stopping criteria are satisfied.

When selection procedure is stopped we have obtained the selected subset of primary variables $\mathbb{X}_1$.

III. METHOD 2: VARIABLE SELECTION WITH COST-UTILITY ANALYSIS

A. Variance Recovery Cost Index

At this point in time, variable selection methods for process monitoring have focused mainly on the explained variance performance criteria. However, explained variance efficiency is a minimal notion of optimality and does not necessarily result in an economically desirable selected subset, as it makes no statement about the measurement cost or other engineering criteria. Without measuring cost many decisions will be impossible to make. The objective of carrying out a cost–utility analysis is to compare one use of resources with other possible uses. To do this, to any process monitoring procedure is assigned a score calculated as ratio of the cost at which it might be obtained to explained variance that it might provide. Then, the ratio scores are compared to define the best economically desirable selected subset.

Let $\mathbb{X}_1$ be the selected subset under conditions examined and $F(\mathbb{X}_1)$ their associated cost, given by

$$F(\mathbb{X}_1) = \sum_{X_j \in \mathbb{X}_1} f_j \qquad (8)$$

where f_j is the cost for measurement X_j.

To compare one use of resources with other possible uses, we propose variance recovery cost index, noted CR. The equation for CR is

$$CR = F(\mathbb{X}_1) / R \qquad (9)$$

where $F(\mathbb{X}_1)$ is the cost and R is the variance recovery across conditions examined, for example given by (5) or (7). CR score attempts to define, how much, each unit of explained variance costs. Variable subsets for process monitoring can be ranked according to CR values. This allows easy comparison across different selected variable subsets, but still requires value judgments to be made about the quality of explained variance across the structure and information carried by the full set of original variables.

IV. CONTROL CHARTS BASED ON INFLUENCE FUNCTION

A. Influence Function

We assume that under a stable process the distribution of $\mathbb{X}$ is F, ideally multivariate normal. When special causes are present in the process $\mathbb{X}$ has an arbitrary distribution noted G. A distribution function which describes the two sources of variation in a process is the contaminated model, [7], given by:

$$F_{\varepsilon H} = (1 - \varepsilon)\, F + \varepsilon G \qquad (10)$$

with $0 \leq \varepsilon \leq 1$.

If process is under control we have ε=0. When process is not stable, roughly a proportion ε of output subgroups will be contaminants.

Let T=T(F) be a statistical functional. The influence function $IF(x,T,F)$ of the statistical functional T at F is defined as the limit as $\varepsilon \to 0$ of

$$\left\{T\left[(1-\varepsilon)F + \varepsilon\delta_x\right] - T(F)\right\} / \varepsilon \qquad (11)$$

where δ_x denotes the distribution giving unit mass to the point $x \in R^p$. The perturbation of F by δ_x is denoted as

$$F_{\varepsilon x} = (1-\varepsilon)F + \varepsilon\delta_x \quad (0 \leq \varepsilon \leq 1) \qquad (12)$$

As such the influence function measures the rate of change of T as F is shifted infinitesimally in the direction of δ_x, [7]. The importance about the influence function lies in its heuristic interpretation: it describes the effect of an infinitesimal contamination at point x on the estimate. Our idea is that output segments that have a large influence on monitored parameters show up the time when special causes are present in a manufacturing process. The influence functions may be calculated for almost all process parameters. Therefore, based on influential measures derived from them, multivariate control charts for different process parameters and with different sensitivities are be set up, [9], [10], [11], [12].

B. Control Charts

Assignable causes that affect the variability of the output do not increase significantly each component of total variace of $\mathbb{X}$. Instead, they may have a large influence in the variability of some components and small effect in the remaining directions. Therefore an approach to design control charts for variability consists to detect any significant departure from the stable level of the variability of each component. Based on $\mathbb{X}_1$ PCs that represent well the whole set of variables are derived, [17]. To build up control charts one may use either the principal components or the influence functions of eigenvalues of dispersion matrix. The control limits of the proposed control charts are three sigma control limits as in any Shewhart control chart, (for details see [9], [10], [11], [12]).

V. Application

A. Case Study

The proposed methods will be illustrated by using data from a real production process, which manufactures bumper covers for vehicles. Bumper covers are molded pieces made of durable plastic designed to enhance the look and shape of the vehicle while hiding the real bumper. They are attached to the vehicle with fasteners. The current inspection procedure consists of measurements taken at 24 points. The variables that are measured are holes diameters. To fit well with the automobile's overall holes diameters have tight dimensional tolerances. But not all these variables are equally important according to engineering and economic criteria. Ten among them are very important because their deviations from target values lead to designs with less aesthetic fit of automobile's overall and they are very awkward to handle. Meanwhile for the remaining variables their deviations from target diameters can be handled easily by operators and lead to designs that fit well.

B. Variable Selection with Pre-Assigned Roles

We applied our proposed variable selection methodology with pre-assigned roles to bumper cover manufacturing process. The number of elements in the sets of primary and secondary variables $\mathbb{C}_1$ and $\mathbb{C}_2$ are 10 and 14 respectively. In this article we used a backward elimination scheme, which at each step eliminates the less informative variable among the primary variables that have not yet been eliminated. The new variable was eliminated by its inability to supply complementary information for the whole set of variables. The subset of relevant variables was selected in a manner that retains, to some extent, the structure and information carried by the full set of original variables, thereby providing a SPC almost as efficient as we were monitoring all original variables. The results showed that efficient monitoring of this process according to criterion RX in (5) could be attained by using only six primary variables. Shewhart control charts of influence functions of eigenvalues for the covariance matrix were used to monitor components of process variability. These influential control charts, accompanied with process logbook gave clear indications for all known assignable causes present in the process.

C. Variable Selection with Cost-Utility Analysis

To illustrate variable selection with cost-utility analysis we used only the measurements of ten primary variables. The objective of carrying out a cost–utility analysis is to compare one use of resources with other possible uses. Variable subsets for process monitoring were ranked according to CR values in (9). Based on data from the cover bumper process, a surface plot with Cartesian coordinates (c, CR, RX) is displayed in Fig. 1. A subset of relevant variables that retains, to some extent, the structure and information carried by the full set of original variables should have high RX values and low CR and c values. This allows easy comparison across different selected variable subsets, but still requires value judgments to be made about the quality of explained variance across the structure and information carried by the full set of original variables. An inspection of surface plot in Fig. 1 shows that an effective use of resources is obtained for a selected subset with six primary variables.

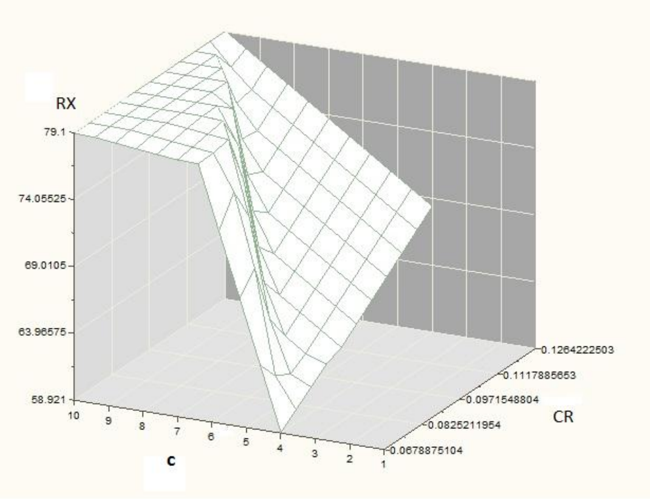

Fig. 1. Surface plot with Cartesian coordinates (c, CR, RX).

VI. Conclusion

This article proposes two new methods to select a reduced number of relevant variables for multivariate statistical process control that makes use of engineering, cost and variability evaluation criteria. In the first method a double reduction of dimensionality is applied to select relevant primary variables that represent well the whole set of variables. In the second methodology a cost-utility analysis is proposed to compare different variable subsets that may be used for process monitoring. The objective of carrying out a cost–utility analysis is to compare one use of resources with other possible uses. The subset of relevant variables is selected in a manner that retains, to some extent, the structure and information carried by the full set of original variables. This strategy mitigates the risk that an assignable cause inducing a shift, that lies entirely in the discarded variables, will go undetected. Just like ordinary PCA the solution of the eigenvalue problem in (4) is not scale invariant, and therefore sometimes it is better to apply the above method to standardized data rather than raw data. In such cases the covariance matrices in their formulation are replaced by the corresponding correlation matrices.

References

[1] N.M. Al-Kandari, and I.T. Jolliffe, “Variable selection and interpretation of covariance principal components”, *Commun. Stat.-Simul. Comput*, vol. 30, 2001, pp 339-354.

[2] N.M. Al-Kandari, and I.T. Jolliffe, “Variable selection and interpretation in correlation principal components”, *Environmetrics*, vol. 16, 2005, pp 659-672.

[3] J.F.C.L. Cadima and I.T. Jolliffe, “Variable selection and the interpretation of principal subspaces”**,** *J. Agric. Biol. Environ. Stat.*, vol. 6, 2001, pp. 62-79.

[4] B. M. Colosimo, Q. Semeraro, and M. Pacella, “Statistical Process Control for Geometric Specifications: On the Monitoring of Roundness Profiles”, *Journal of Quality Technology* , vol. 40, 2008, pp. 1–18.

[5] J. A. Cumming and D. A. Wooff, “Dimension Reduction Via Principal Variables”, *Computational Statistics and Data Analysis* 52, 2007, pp. 550–565.

[6] I. Gonzalez and I. Sanchez, “Variable Selection for Multivariate Statistical Process Control », *Journal of Quality Technology*, vol. 42, n°. 3, 2010, pp. 242-259.

[7] F.R. Hampel, E.M. Ronchetti, P.J. Rousseew, W.A. Stahel, “Robust Statistics - The Approach Based on Influence Functions”, Wiley, 1986, New-York.

[8] W. Krzanowski, “Selection of Variables to Preserve Multivariate Data Structure, Using Principal Components”, *Applied Statistics* 26, 1987, pp. 22–33.

[9] L. Jaupi and G. Saporta, “Using the Influence Function in Robust Principal Components Analysis”. In S. Morgenthaler, E. Ronchetti and W.A. Stahel, eds., *New Directions in Statistical Data Analysis and Robustness*, Birkhäuser Verlag Basel, 1993, pp. 147-156.

[10] L. Jaupi, “Multivariate Control Charts for Complex Processes”. In C. Lauro, J. Antoch, V. Esposito, G. Saporta eds., *Multivariate Total Quality Control*, Springer, 2001, pp. 125-136.

[11] L. Jaupi , D. Herwindiati, P. Durand and D. Ghorbanzadeh, “Short Run Multivariate Control Charts for Process Mean and Variability”, *Proc. World Congress on Engineering,* 2013, Vol. I, pp.670-674.

[12] L. Jaupi , P. Durand , D. Ghorbanzadeh and D. E. Herwindiati, “Multi-Criteria Variable Selection for Process Monitoring”, *59th World Statistical Congress,* August 2013.

[13] I. T. Jolliffe, “Discarding Variables in a Principal Component Analysis I: Artificial Data”. *Applied Statistics* 21, 1972, pp. 160–173.

[14] I. T. Jolliffe, “Discarding Variables in a Principal Component Analysis II: Real Data”. *Applied Statistics* 22, 1973, pp. 21–31.

[15] I. T. Jolliffe, *Principal Components Analysis*, 2nd edition. New York, NY: Springer, 2002.

[16] G. P. McCabe, “Principal Variables”. *Technometrics* 26, 1984, pp. 137–144.

[17] C.R. RAO, “The Use and Interpretation of Principal Components in Applied Research”. *Sankhya,* A, 26, 1964, pp. 329-358.

[18] Y. Tanaka, and Y. Mori, “Principal component analysis based on a subset of variables: Variable selection and sensitivity analysis”. *Amer. J. Mathematical and Management Sciences*, 17, 1 & 2, 1997, pp. 61-89.

[19] W. H. Woodall, D. J. Spitzner, D. C. Montgomery and S. Gupta, “Using Control Charts to Monitor Process and Product Quality Profiles”. *Journal of Quality Technology,* vol. 36, 2004, pp. 309–320.