Multi-Criteria Variable Selection for Process Monitoring

Luan Jaupi, Philippe Durand, Dariush Ghorbanzadeh, Dyah E. Herwindiati

Multi-Criteria Variable Selection for Process Monitoring

Luan Jaupi[1,3], Philippe Durand[1], Dariush Ghorbanzadeh[1] and Dyah E. Herwindiati[2]
[1] Conservatoire National des Arts et Métiers, Paris, France
[2] Tarumanagara University, Jakarta, Indonesia
[3] Corresponding author: Luan Jaupi, e-mail: jaupi@cnam.fr

Abstract

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. For many applications, it may be useful for external information to influence the selection process. For example, some variables may be easier and cheaper to carry out then others or they might be very important according to some engineering criteria. Neglecting this information in statistical process control, would be counterproductive. In this article, we propose a statistical methodology to select a reduced number of relevant variables for multivariate statistical process control that makes use of engineering and variability evaluation criteria. A double reduction of dimensionality is applied in conjunction with economic and variability selection criteria. 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 method.

Keywords: process control, dimension reduction, variance efficiency, influence function, measurement cost

1. 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. Woodall et al. (2004) and Colosimo et al. (2008) present examples of situations in which variable selection is necessary. Gonzalez and Sanchez (2010) propose 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. However, in many cases measured variables generally 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 redaction because their measurement are made at earlier points in the process. Neglecting this information in SPC would be counterproductive. 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 a statistical methodology to select a reduced number of

relevant variables for multivariate SPC. 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 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 (1964) approach on principal components, PCs, of instrumental variables.

2. Formulation

In what follows we suppose that $\mathbb{X}=(X_1,X_2,\ldots,X_m)$ is the vector of the 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 redaction because their measurement are made at earlier points in the process. Without loss of generality let $\mathbb{C}_1=(X_1,X_2,\ldots,X_p)$ and $\mathbb{C}_2=(X_{p+1},\ldots,X_m)$ 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 ($c\leq p$), 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 Jolliffe, 1972, 1973, 2002, McCabe, 1984, Krzanowski 1987, Tanaka and Mori 1997, Cadima and Jolliffe 2001, Cumming and Wooff 2007, Gonzalez and Sanchez 2010). 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

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

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:

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

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 $A = (\alpha_1, \alpha_2, ..., \alpha_q)$, Rao (1964).

3. 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 (2001, 2005) 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:

$$PX_1 = \frac{\lambda_1 + \lambda_2 + ... + \lambda_q}{trace(\Sigma)} \quad (5)$$

Classical PCA results guarantee that the maximum value of the right hand of Eq. (5) is attained for $\mathbb{X}_1$=$\mathbb{X}$. The index PX_1 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 $\Sigma_{X_2/Y}$ given by:

$$\Sigma_{X_2/Y} = \Sigma_{22} - A^t \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12} A \quad (6)$$

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

$$P_{X_2/Y} = 1 - \frac{\lambda_1^{'} + \lambda_2^{'} + ... + \lambda_{m-c}^{'}}{trace(\Sigma_{22})} \quad (7)$$

where $\lambda_1^{'}, \lambda_2^{'}, ..., \lambda_{m-c}^{'}$ are eigenvalues of $\Sigma_{X_2/Y}$. The criterion $P_{X_2/Y}$ is similar to index R_{EX} defined in Gonzalez and Sanchez (2010). 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 $P_{X_2/Y}$ 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 $P_{X_2/Y}$ value will guarantee that the selected variables may provide a SPC almost as efficient as if we were monitoring all *m* variables.

4. 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:

Compute dispersion matrix of the whole data set. Based on $\mathbb{C}_1$ calculated PCs that explain well the whole set of original variables $\mathbb{X}$. Looking carefully at eigenvalues and the cumulative proportions, determine the number of PCs to be used. Remove each one among the p variables in $\mathbb{C}_1$ in turn, and solve p eigenvalue problems, Eq.(4), with (p-1) variables. Find the best subset of size (p-1) according to selection criterion that is used and remove the corresponding variable. 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 $\mathbb{X}_1=\mathbb{C}_1$.

5. 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 used to build up control charts to monitor components of process variability. To build up such control charts one may use either the principal components or the influence functions of eigenvalues of dispersion matrix, Jaupi (2001), Jaupi and Saporta (1993). The control limits of the proposed control charts are three sigma control limits as in any Shewhart control chart.

6. Application

The proposed methodology will be illustrated by using data from a real production process. The process 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 10 points. The variables that are measured are holes diameters. To fit well with the automobile's overall holes diameters have tight dimensional tolerances. But no all these variables are equally important according to engineering and economic criteria. Six 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 four variables their deviations from target diameters can be handled easily by operators and lead to designs that fit well. So the number of elements in the sets of primary and secondary variables $\mathbb{C}_1$ and $\mathbb{C}_2$ are 6 and 4 respectively. We applied our proposed selection methodology to bumper cover manufacturing process. The proposed methodology shows that efficient monitoring of this process according to criterion PX_1 could be attained by using only four primary variables. Shewhart control charts of influence function of eigenvalues of covariance matrix are used to monitor components of process variability. Average influence is zero. Graphical displays of these charts will be presented in oral presentation of the paper. But in process logbook there are clear explanations for all assignable causes that are detected by influential charts of eigenvalues**.**

7. Conclusions

This article presents a methodology to select a reduced subset of variables to be used in multivariate SPC that has some advantages with respect to the existing ones, filling a vacancy in the quality-control literature. A double-reduction of dimensionality is applied in conjunction with engineering, economic and variability criteria. 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 Eq(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 matrix.

References

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

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

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

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

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

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

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

[8] Jaupi L and Saporta G. (1993). “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, pp. 147-156.

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

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

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

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

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

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

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

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