Monitoring the covariance matrix of a multivariate skew normal distribution

Adelaide M. Figueiredo[1] and Fernanda O. Figueiredo[2]

[1] Faculdade de Economia da Universidade do Porto and LIAAD-INESC Porto
(E-mail: adelaide@fep.up.pt)
[2] Faculdade de Economia da Universidade do Porto and Centro de Estatística e Aplicações, Universidade de Lisboa
(E-mail: otilia@fep.up.pt)

Abstract. The multivariate skew normal distribution is very useful for modeling asymmetric data in many practical applications, and in particular in Statistical Quality Control for monitoring several quality characteristics. In this study in order to monitor the covariance matrix of a multivariate skew normal process, we consider a control chart based on the Statis methodology. More precisely, the chart is based on a similarity measure between two data tables, the RV coefficient. The performance of this chart is evaluated for several skew-normal processes.

Keywords: Control chart, Monte Carlo simulation, Multivariate skew normal distribution, Process monitoring, RV coefficient, STATIS, Statistical Quality Control.

1 Introduction

In Statistical Quality Control it is crucial to monitor simultaneously several quality characteristics. Often these characteristics are correlated and thus, multivariate techniques of quality control are more appropriate than univariate methods for monitoring the individual characteristics. Many multivariate techniques of quality control have been proposed in the literature, in particular many control charts have appeared for monitoring processes.
Control charts are the tools most used for process monitoring in Statistical Quality Control (SQC) and were introduced by Shewhart at Bell Laboratories in 1924. Control charts help us to decide if the process that is being monitored is in-control or out-of-control. When a control chart triggers an out-of-control signal, which may be eventually a false alarm, it is important to investigate what are the causes responsible for the emission of such signal, so that appropriate actions may be taken.

Several multivariate schemes have been proposed for monitoring the mean vector or the covariance matrix of a multivariate process. In particular, control charts based on the Hotelling T^2 statistic, among others, have been implemented for monitoring the mean vector, and control charts based on the generalised variance (Alt, 1985) and based on the maximum of the sample variances

Stochastic Modeling, Data Analysis and Statistical Applications (*pp.* 87-94)
Lidia Filus - Teresa Oliveira - Christos H Skiadas (Eds)

or on the maximum of the ranges (Costa and Machado, 2008a, 2008b), among other charts have been proposed for monitoring the covariance matrix.
Additionally, several control schemes have appeared in the literature to monitor simultaneously the mean vector and the covariance matrix of a process (Chen et al., 2005, Zhang and Chang, 2008, etc).

Figueiredo and Figueiredo (2014) proposed a control scheme for controlling the variability of a multivariate process based on Statis methodology. More precisely, this scheme is based on a similarity measure between two positive semi-definite matrices, the RV coefficient proposed by Escoufier (1973).
In this study we consider the previous control scheme for monitoring the covariance matrix of a multivariate skew normal process.

The STATIS (Structuration des Tableaux a Trois Indices de la Statistique) methodology was introduced by L′Hermier des Plantes (1976) and later developed by Lavit (1988) and Lavit et al. (1994). This methodology enables us to analyse simultaneously several data tables measured on the same individuals or variables for different circumstances or time instants.
We'll use this methodology for comparing several data tables. More precisely, we'll compare the relations between the variables along the data tables through the covariance matrices and we'll determine the compromise covariance matrix. Statis methodology has been applied in Statistical Quality Control to monitor batch processes (see for instance, Scepi, 2002, Gourvénec et al., 2005 and Niang et al., 2009).

The multivariate skew normal distribution was proposed by Azzalini and Dalla Valle (1996), and further discussed by Azzalini and Capitanio (1999) and others. This distribution is an extension of the univariate skew normal distribution, such that the marginal densities are scalar skew-normal. It also extends the multivariate normal distribution, by the addition of a shape parameter.

In Section 2 we briefly refer the multivariate skew normal distribution, in Section 3 we describe the control chart based on RV coefficient between the compromise covariance matrix obtained from a set of reference samples and the covariance matrix of a new sample. In Section 4 we evaluate the performance of the chart for monitoring the covariance matrix of a multivariate skew normal process.

2 The multivariate skew normal distribution

A k-dimensional random variable $\mathbf{Z}$ is said to have a multivariate skew normal distribution if it has density function

$$f(\mathbf{z}) = 2\ \phi_k(\mathbf{z}\ ;\Omega_{\mathbf{z}})\ \Phi\left(\boldsymbol{\alpha}'\mathbf{z}\right), \quad \mathbf{z} \in \mathbb{R}^k, \tag{1}$$

where $\phi_k(\mathbf{z};\Omega_{\mathbf{z}})$ is k-dimensional normal density with zero mean and correlation matrix $\Omega_{\mathbf{z}}$, $\Phi\left(.\right)$ is the $N(0,1)$ distribution function and $\boldsymbol{\alpha}$ is a k-dimensional vector.

When $\boldsymbol{\alpha} = \mathbf{0}$, density (1) reduces to the multivariate normal distribution $N_k(\mathbf{0},\Omega_{\mathbf{z}})$ density. The parameter $\boldsymbol{\alpha}$ is then referred as a shape parameter.

Next, we introduce location and scale parameters, which are not allowed in density (1). Let

$$\mathbf{Y} = \boldsymbol{\xi} + \omega\ \mathbf{z},$$

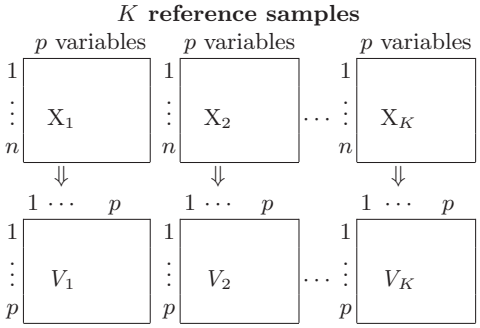
where $\boldsymbol{\xi} = (\xi_1, ..., \xi_k)'$ and $\omega = diag\,(w_1, ..., w_k)$ are location and scale parameters respectively, being $w_i > 0, i = 1, ..., k$. The density function of $\mathbf{Y}$ is

$$g(\mathbf{y}) = 2\ \phi_k(\mathbf{y} - \boldsymbol{\xi};\ \Omega)\ \Phi\left(\boldsymbol{\alpha}'\omega^{-1}\,(\mathbf{y} - \boldsymbol{\xi})\right), \quad \mathbf{y} \in \mathbb{R}^k, \tag{2}$$

where $\Omega = \omega\ \Omega_{\mathbf{z}}\ \omega$ is a covariance matrix. We will use the notation $\mathbf{Y} \sim SN_k\,(\boldsymbol{\xi}, \Omega, \boldsymbol{\alpha})$ to indicate that $\mathbf{Y}$ has density function (2).
For more details about this distribution, see Azzalini and Dalla Valle (1996) and Azzalini and Capitanio (1999).

3 Control chart for monitoring the covariance matrix

We consider K reference samples of size n measured on p variables taken in K different time instants, when the process is in the in-control state, and we represent these matrices by their covariance matrices V_k 's. See the following scheme.



We determine the compromise covariance matrix, V, as defined in the Statis methodology, a weighted mean of the K covariance matrices V_k 's:

$$V = \sum_{k=1}^{K} \alpha_k V_k,$$

where the weights α_k represent the agreement between the K tables and the compromise, and are obtained from the RV coefficients.
The RV coefficient (Escoufier, 1973) between V_k and $V_{k'}$ is defined by

$$RV(V_k, V_{k'}) = \frac{Tr(V_k Q V_{k'} Q)}{\sqrt{Tr\,(V_k Q)^2\, Tr\,(V_{k'} Q)^2}},$$

where Tr denotes the trace operator of a matrix and Q is the metric in the individuals space, defined by the identity matrix or by a diagonal matrix whose main elements are equal to the reciprocal of the variances of the variables. The

RV coefficient varies between 0 and 1. The closer the RV coefficient is to 1, the more similar the two covariance matrices V_k and $V_{k'}$ are.
More precisely, the weights α_k are the elements of the eigenvector associated with the largest eigenvalue of the following matrix Z containing the RV coefficients between the V_k 's:

$$Z = \begin{pmatrix} 1 & RV(V_1, V_2) & \cdots & RV(V_1, V_K) \\ RV(V_2, V_1) & 1 & \cdots & RV(V_2, V_K) \\ \vdots & \vdots & \ddots & \vdots \\ RV(V_K, V_1) & RV(V_K, V_2) & \cdots & 1 \end{pmatrix}$$

The control chart, which we denote RV-chart, is implemented as follows. For a new time instant $k+1$, we compare its covariance matrix V_{k+1} with the compromise covariance matrix V through the RV coefficient. Denoting CL the control limit of the chart, we consider the following decision criterion:

- If $RV(V, V_{k+1}) \geq$ CL we consider that the process is in-control.
- Otherwise, we decide that the process is out-of-control. In this case it is important to identify which variables are responsible for this situation.

The exact distribution of the RV coefficient is unknown, and thus we fix CL at an empirical percentile of the sampling distribution of the RV coefficient.

4 Performance of the control chart for a skew normal process

For evaluating the efficiency of the RV-chart, we computed by simulation the Average Run Length (ARL), the most commonly used measure of performance of control charts.
We generated multivariate skew normal processes $SN_p(\boldsymbol{\xi}, \Omega, \boldsymbol{\alpha})$, for p=2,3 assuming different structures for the covariance matrices when the process is in-control and out-of-control and different shape parameters. In each case, we obtained the compromise covariance matrix based on 4 reference samples generated when the process is in-control. For a false alarm rate α=0.005, we determined the control limit of the chart, i.e., the percentile 0.5% of the distribution of the RV coefficient, obtained through a Monte Carlo simulation experiment of size 100000 and we calculated the in-control and out-of-control ARL values through 10000 replicates for different shape parameters and structures of the covariance matrix.
More precisely, we generated samples from a bivariate skew normal distribution $SN_2(\boldsymbol{\xi}, \Omega, \boldsymbol{\alpha})$ with location vector $\boldsymbol{\xi} = (0,0)'$, covariance matrix $\Omega = \begin{pmatrix} 1 & \sigma_{12} \\ \sigma_{12} & 1 \end{pmatrix}$ and shape parameter $\boldsymbol{\alpha}$. Note that we could consider another location vector because we will work with centered data. The unit variances in Ω imply that the covariance is equal to the linear correlation coefficient. Some

obtained results are presented in Tables 1, 2 and 3. We also generated samples from a multivariate skew normal distribution $SN_3(\boldsymbol{\xi}, \Omega, \boldsymbol{\alpha})$ with location vector $\boldsymbol{\xi} = (0,0,0)'$, covariance matrix $\Omega = \begin{pmatrix} 1 & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & 1 & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & 1 \end{pmatrix}$ and shape parameter $\boldsymbol{\alpha}$. As previously we could use another location vector and the unit variances imply covariances equal to the correlation coefficients. Some obtained results are indicated in Tables 4 and 5.

σ_{12}= 0 in-control										
$\boldsymbol{\alpha}'$	(0,0)	(2,2)	(6,6)	(-2,-2)	(-6,-6)	(0,0)	(2,2)	(6,6)	(-2,-2)	(-6,-6)
n			5					15		
CL	0.359	0.331	0.324	0.333	0.323	0.696	0.704	0.707	0.698	0.708
σ_{12}			ARL					ARL		
0	**198.1**	**206.2**	**201.8**	**193.4**	**203.3**	**196.3**	**186.2**	**204.1**	**209.3**	**197.4**
0.4	89.0	60.4	54.1	58.3	54.5	39.5	12.3	11.3	13.0	10.9
0.75	69.8	15.2	12.9	15.0	13.0	6.2	1.7	1.6	1.8	1.6
0.95	32.1	4.3	3.5	4.3	3.5	2.5	1.0	1.0	1.0	1.0

Table 1. Control limit and ARL for several shape parameters $\boldsymbol{\alpha}$ and n=5,15, being σ_{12}= 0 when the process is in-control. The in-control ARL values are in bold.

σ_{12}= 0 in-control										
$\boldsymbol{\alpha}'$	(0,0)	(-2,6)	(2,-6)	(0,2)	(0,-2)	(0,0)	(-2,6)	(2,-6)	(0,2)	(0,-2)
n			5					15		
CL	0.359	0.325	0.323	0.337	0.340	0.696	0.705	0.705	0.698	0.696
σ_{12}			ARL					ARL		
0	**198.1**	**198.4**	**205.6**	**205.2**	**196.3**	**196.3**	**201.9**	**204.1**	**193.1**	**201.5**
-0.4	**143.7**	84.7	**88.8**	**181.1**	**170.7**	40.2	22.7	23.1	7.9	82.0
-0.75	69.1	29.7	29.6	114.9	109.4	6.1	2.7	2.7	11.2	11.5
-0.95	32.3	9.3	9.5	57.2	53.7	2.5	1.1	1.1	2.3	2.3

Table 2. Control limit and ARL for several shape parameters $\boldsymbol{\alpha}$ and n=5,15, being σ_{12}= 0 when the process is in-control. The in-control ARL values are in bold.

The control limit and the ARL depend on the sample size (see Tables 1-3) and in general, depend on the shape parameter and on the structure of covariance of covariance matrix. See Tables 1-5.

From these tables, we observe that the in-control ARL is large and approximately equal to the expected value 200. When the process is out-of-control, the ARL quickly decreases as the sample size increases.

For a bivariate process with correlation matrix equal to the identity matrix, the chart detects easily a positive correlation when both components of the shape vector are null or have the same sign (positive or negative). Moreover, the detection is as fast as larger is the value of the correlation. See Table 1.

σ_{12}= 0.9 in-control										
$\boldsymbol{\alpha}'$	(0,0)	(2,2)	(6,6)	(-2,-2)	(-6,-6)	(0,0)	(2,2)	(6,6)	(-2,-2)	(-6,-6)
n			5					15		
CL	0.596	0.357	0.328	0.368	0.336	0.949	0.855	0.836	0.855	0.835
σ_{12}			ARL					ARL		
0.9	**191.6**	**202.6**	**208.5**	**184.1**	**189.2**	**204.2**	**201.9**	**190.6**	**201.0**	**197.9**
0.75	32.0	36.3	37.9	32.7	35.2	9.6	7.6	7.3	7.4	7.5
0.5	8.5	10.2	10.8	8	10.5	1.9	1.7	1.7	1.7	1.7
0	2.4	3.1	3.3	2.9	3.1	1.0	1.0	1.0	1.0	1.0
-0.9	1.0	1.1	1.1	1.1	1.1	1.0	1.0	1.0	1.0	1.0

Table 3. Control limit and ARL for several shape parameters $\boldsymbol{\alpha}$ and n=5,15, being σ_{12}= 0.9 when the process is in-control. The in-control ARL values are in bold.

$\sigma_{ij}= 0, i \neq j$, in-control					
$\boldsymbol{\alpha}'$	(0,0,0)	(2,2,2)	(6,6,6)	(-2,-2,-2)	(-6,-6,-6)
CL	0.671	0.676	0.678	0.676	0.678
$\sigma_{12}, \sigma_{13}, \sigma_{23}$			ARL		
0,0,0	**201.4**	**201.2**	**195.8**	**195.8**	**194.1**
0.4,0.4,0.4	13.9	10.4	10.1	10.6	9.8
0.75,0.75,0.75	1.9	1.3	1.3	1.3	1.3
0.95,0.95,0.95	1.2	1.0	1.0	1.0	1.0
0.5,0.2,0.9	3.5	1.8	1.7	1.8	1.7
0.9,0.75,0.9	1.4	1.0	1.0	1.0	1.0

Table 4. Control limit and ARL for several shape parameters $\boldsymbol{\alpha}$ and n=15, being $\sigma_{ij}= 0, i \neq j$ when the process is in-control. The in-control ARL values are in bold.

$\sigma_{ij}= 0.9, i \neq j$, in-control					
$\boldsymbol{\alpha}'$	(0,0,0)	(2,2,2)	(6,6,6)	(-2,-2,-2)	(-6,-6,-6)
CL	0.951	0.845	0.833	0.844	0.834
$\sigma_{12}, \sigma_{13}, \sigma_{23}$			ARL		
0.9,0.9,0.9	**200.0**	**201.1**	**202.4**	**202.7**	**196.6**
0.75,0.75,0.75	6.7	5.8	5.9	6.0	5.9
0.5,0.5,0.5	1.4	1.3	1.3	1.3	1.3
0,0,0	1.0	1.0	1.0	1.0	1.0
0.9,0.5,0.1	1.2	1.2	1.2	1.2	1.2
0.1,0.5,0.3	1.0	1.0	1.0	1.0	1.0

Table 5. Control limit and ARL for several shape parameters $\boldsymbol{\alpha}$ and n=15, being $\sigma_{ij}= 0.9, i \neq j$ when the process is in-control. The in-control ARL values are in bold.

If the process is normal ($\boldsymbol{\alpha} = \mathbf{0}$) or when one component of the shape vector is positive and the other is negative, the chart easily detects a negative correlation, being more sensitive to large negative correlations. See Table 2.
In a 3-dimensional framework with data from a normal or a skew normal process with a shape parameter, having all components positive or negative and when the correlation matrix is equal to the identity matrix or has all off-diagonal elements equal to 0.9, the chart detects changes in the correlations as fast as we move away from the in-control correlation structure. See Tables 4 and 5.

To conclude, the analysed cases suggest that the RV-control chart enables us to detect easily changes in the correlations between variables when the process has a normal or a skew normal multivariate distribution, being therefore a very useful monitoring tool in a large variety of industrial applications.

Acknowledgement

This work is financed by the ERDF – European Regional Development Fund through the COMPETE Programme (Operational Programme for Competitiveness) and by National Funds through the FCT – Fundação para a Ciência e Tecnologia (Portuguese Foundation for Science and Technology) within the project FCOMP-01-0124-FEDER-037281 and the project PEst-OE/MAT/UI0 006/2014 (CEA/UL).

References

1. F. B. Alt. Multivariate quality control. In *Encyclopedia of Statistical Sciences, S. Kotz and N. L. Johnson (eds)*, Wiley, New York, 1985.
2. A. Azzalini and A. Capitanio. Statistical applications of the multivariate skew normal distribution. *J. R. Statist. Soc. B*, 61, part 3, pp. 579-602, 1999.
3. A. Azzalini and A. Dalla Valle. The multivariate skew-normal distribution. *Biometrica*, 83, no.4, pp. 715-726, 1996.
4. G. Chen, S. W. Cheng and H. Xie. A new multivariate control chart for monitoring both location and dispersion. *Communications in Statistics: Simulation and Computation*, 34, 203–217, 2005.
5. A. F. B. Costa and M. A. G. Machado. A new chart based on sample variances for monitoring the covariance matrix of multivariate processes. *The International Journal of Advanced Manufacturing Technology*, 41, 770–779, 2008a.
6. A. F. B. Costa and M. A. G. Machado. A new multivariate control chart for monitoring the covariance matrix of bivariate processes. *Communications in Statistics: Simulation and Computation*, 37, 1453–1465, 2008b.
7. Y. Escoufier. Le traitement des variables vectorielles.*Biometrics*, 29, 751–760, 1973.
8. A. Figueiredo and F. Figueiredo. Monitoring the variability of a multivariate normal process using STATIS. In *Proceedings of COMPSTAT 2014*, 2014.
9. S. Gourvénec, I. Stanimirova and O.A. Saby. Monitoring batch process with the STATIS approach. *Journal of Chemometrics*, 19, 288–300, 2005.
10. D. M. Hawkins and E. M. Maboudou-Tchao. Multivariate exponentially weighted moving covariance matrix. *Technometrics*, 50, 155–166, 2008.
11. H. Hotelling. Multivariate quality control, illustrated by the air testing of sample bombsights. *Techniques of Statistical Analysis, McGraw Hill*, New York, pp. 111-184, 1947.
12. C. Lavit. *Analyse Conjointe de Tableaux Quantitatives.* Collection Méthodes+Programmes, Masson, 1988.
13. C. Lavit, Y. Escoufier, R. Sabatier and P. Traissac. The ACT (Statis method). *Computational Statistics and Data Analysis*, 18, 97–119, 1994.
14. H. L′Hermier Des Plantes. *Structuration des Tableaux a Trois Indices de la Statistique.* Thèse de 3 ème cycle. Université de Montpellier II, 1976.

15. N. Niang, F.S. Fogliatto and G. Saporta. *Batch Process Monitoring by three-way data analysis approach.* The XIII International Conference Applied Stochastic Models and Data Analysis (ASMDA 2009), June 30-July 3, Vilnius, Lithuania, 2009.
16. G. Scepi. *Parametric and non parametric multivariate quality control charts.* IN Lauro, C. et al. (Eds) Multivariate Total Quality Control, 163–189. Heidelberg: Physica-Verlag, 2002.
17. G. Zhang and S. I. Chang. Multivariate EWMA control charts using individual observations for process mean and variance monitoring and diagnosis. *International Journal of Production Research*, 46, pp. 6855-6881, 2008.