Optimality Aspects of Multivariate Control Charts

Marianne Frisén[1] and Eva Andersson[1, 2]

[1] Statistical Research Unit, Department of Economics, University of Gothenburg,
Box 640, SE40530 Gothenburg, Sweden
(e-mail: marianne.frisen@statistics.gu.se)
[2] Department of Occupational and Environmental medicine, Sahlgrenska University Hospital and Sahlgrenska Academy
University of Gothenburg,
Box 414, SE40530 Gothenburg, Sweden
(e-mail: eva.andersson@statistics.gu.se)

Abstract. Multivariate control charts are of interest in industrial production as they enable the joint monitoring of several components. Recently, there has been an increased interest also in other areas such as the detection of bioterrorism, transaction strategies in finance and the surveillance of outbreaks of epidemic diseases based on spatial information. Multivariate counterparts to the univariate Shewhart, EWMA, and CUSUM methods have earlier been proposed.

General approaches to multivariate surveillance are reviewed. The challenges of evaluating multivariate surveillance methods are of special concern. Optimality is usually hard to derive, and even to define, in multivariate problems. This is true also for multivariate surveillance. Multivariate on-line surveillance problems can be complex. The sufficiency principle can be of great use to clarify the structure of some problems. Here it is used to discuss metrics for evaluation of multivariate surveillance. It is demonstrated that the sufficiency principle allows important reductions of some classes of multivariate surveillance problems. This is used to determine optimal methods.

Keywords. Sequential, Surveillance, Multivariate, Sufficiency

1 Introduction

The continuous observation of time series with the aim of detecting an important change in the underlying process as soon as possible after the change has occurred is of great interest. The first versions of modern control charts (Shewhart (1931)) were made for industrial use. Multivariate surveillance is of interest in industrial production, for example in order to monitor the many sources of variation in assembled products. Wärmefjord (2004) described the multivariate problem for the assembly process of the Saab automobile. In recent years, there has been an increased interest in statistical surveillance also in other areas than industrial production. The increased interest in surveillance methodology in the US following the 9/11 terrorist attack is notable. In the US, as well as in other countries, several new types of data are now being collected. Often the collected data involve several related variables, which calls for multivariate surveillance techniques. Spatial surveillance is useful for the detection of a local change or a spread (for example of a disease or a harmful agent). Spatial surveillance is multivariate since several locations are involved. Recently, there have also been efforts to use multivariate surveillance for financial decision strategies (see for example Okhrin and Schmid, 2007 and Golosnoy, Schmid and Okhrin, 2007) with respect to various assets and their relations and combinations.

Surveillance of several variables involves surveillance of a joint multiparameter distribution. Sometimes the surveillance concerns several parameters in the distribution of one variable. Knoth and Schmid (2002), for example, studied the mean and the variance in the normal distribution. Such surveillance of the multiparameter distribution can involve the same problems as the surveillance of a multiparameter distribution originating from the joint distribution of several variables. Thus the techniques for multivariate surveillance are of interest in both cases.

Reviews on multivariate surveillance methods can be found for example in Basseville and Nikiforov (1993), Lowry and Montgomery (1995), Ryan (2000), Woodall and Amiriparian (2002), Frisén (2003), Sonesson and Frisén (2005), and Bersimis, Psarakis and Panaretos (2007).

Multivariate surveillance can have different aims. Sometimes, the aim is to identify faulty components. In this paper, however, we will not approach that issue. Instead, we will concentrate on the detection of the first defect.

In Section 2 the notations and specifications will be given. In Section 3 different approaches to the construction of multivariate surveillance methods are described and exemplified. In Section 4 we discuss the special challenges of evaluating multivariate surveillance methods. In Section 5 we describe the sufficiency principle and its implications for evaluation measures and optimality in some multivariate situations. Concluding remarks are given in Section 6.

2 Notations

We denote the multivariate process under surveillance by $\mathbf{Y} = \{\mathbf{Y}(t), \quad t = 1,2,...\}$. At each time point, t, a p-variate vector $\mathbf{Y}(t) = \left(Y_1(t) \quad Y_2(t) \quad \cdots \quad Y_p(t)\right)^T$ of variables is observed. The components of the vector may be, for example, a measure of each of p different components of a produced item. As long as the process is in control and no change has occurred, $\mathbf{Y}(t)$ has a certain distribution. The purpose of the surveillance method is to detect a deviation to a changed state as soon as possible in order to give warnings and take corrective actions. We denote the decision time point by s. At each decision time we want to determine whether or not a change in the distribution of $\mathbf{Y}$ has occurred up to now. In a univariate setting this is expressed as discriminating between the events $\{\tau \le s\}$ and $\{\tau > s\}$, where τ denotes the time point of the change. In a multivariate setting, each component can change at different times τ_1, .. τ_p. A natural aim, which we will consider, is to detect the first time that the joint process is no longer in control. Thus we will consider $\tau_{\min} = \min\{\tau_1,...\tau_p\}$. In order to detect the change, we can use all available observations of the process $\mathbf{Y}_s = \{\mathbf{Y}(t), \quad t \le s\}$ to form an alarm statistic.

3 General approaches to multivariate surveillance

3.1 Dimension reduction

One way to reduce the dimensionality is to consider a few of the principal components as proposed for example by Jackson (1985). However, sometimes it is hard to interpret the principal components. Runger et al. (2007) used the U^2 statistics to allow the practitioner to choose the subspace of interest. Another way to reduce the dimensionality is to use projection pursuit, as done by Ngai and Zhang (2001) and Chan and Zhang (2001). Rosolowski and Schmid (2003) use the Mahalanobis distance to reduce the dimensionality of the statistic. After reducing the dimensionality, any of the approaches for multivariate surveillance described below can be used.

3.2 Reduction to one scalar statistic

The reduction of the dimensionality can go as far as to summarize the observations for each time point into one statistic. This is a common way to handle multivariate surveillance problems. We start by transforming the vector from the current time point into a scalar statistic, which we then accumulate over time. The relevant scalar statistics depend on the application. In spatial surveillance it is common to start with a purely spatial analysis for each time point as in Rogerson (1997). One natural reduction is to use the Hotelling T^2 statistic (Hotelling (1947)). The reduction to a univariate variable can be followed by univariate monitoring of any kind. Originally, the Shewhart method was applied to the Hotelling T^2 statistic, and this is often referred to as the Hotelling T^2 control chart. Crosier (1988) suggested that the Hotelling T statistic should be used in a univariate CUSUM method.

3.3 Parallel surveillance

Parallel surveillance is perhaps the most commonly used approach. A univariate surveillance method is used for each of the individual components in parallel. This is referred to as combined univariate methods or parallel methods. There are several ways to combine the univariate methods into a single surveillance procedure. The most common one is to signal an alarm if any of the univariate methods signals. This is an example of using the union-intersection principle for multiple inference problems. General references on parallel methods include Woodall and Ncube (1985) and Yashchin (1994). Par-

allel CUSUM methods were used by Marshall et al. (2004). One advantage with parallel methods is that the interpretation of alarms can be clear. Determining why an alarm was raised is important.

3.4 Vector accumulation

By vector accumulation the accumulated information on each component is utilized by a transformation of the vector of component-wise alarm statistics into a scalar alarm statistic. Lowry et al. (1992) proposed a multivariate extension of the univariate EWMA method, which is referred to as MEWMA. The MEWMA can be seen as the Hotelling T^2 control chart applied to EWMA statistics instead of to the original data. One natural way to construct a multivariate version of the CUSUM method would be to proceed as for EWMA and construct the Hotelling T^2 control chart applied to univariate CUSUM statistics for the individual variables. An important feature of such a method is the lower barrier of each of the univariate CUSUM statistics (assuming we are interested in a positive change). This kind of multivariate CUSUM was suggested by Bodnar and Schmid (2004) and Sonesson and Frisén (2005). Other approaches to construct a multivariate CUSUM have also been suggested. Crosier (1988) suggested the MCUSUM method, and Pignatiello and Runger (1990) suggested another solution. Both these methods use a statistic consisting of univariate CUSUMs for each component and are thus vector accumulation methods. However, the way in which the components are used is different as compared with the MEWMA construction. An important feature of the two latter methods is that the characteristic zero-return of the CUSUM technique is constructed in a way which is suitable when all the components change at the same time point. However, if all components change at the same time, a univariate reduction is preferable, as seen in Section 5.

3.5 Joint solution

A stepwise construction of methods, as above, is often useful for complicated problems. However, the joint likelihood ratio is of interest in order to take advantage of all the joint information. For univariate problems there are sharp optimality results based on the likelihood ratio. However, when there might be changes at different times for the different variables, these results do not automatically apply. Sometimes a sufficient reduction will be possible for a specific situation. Examples are found in Section 5. In such cases there is no loss of information.

4 Evaluation measures

When testing multiple hypotheses, the risk of false rejection is important as it might be interpreted as a proof that the null hypothesis is false. Recently, methods which restrict the probability of any false alarm have been suggested but they have low ability to detect late changes, as described by Bock (2008). There have also been suggestions to use the false discovery rate, FDR, in multivariate surveillance. The problem with adopting FDR is that it uses a probability which is not constant in surveillance. In surveillance, where false alarms appear naturally, it will sometimes be easier to judge the practical burden of a too low alarm limit by the ARL^0 than by the FDR.

The timeliness of detection is of extreme interest in surveillance, and calls for other measures than the ones traditionally used in hypothesis testing. To evaluate the timeliness, measures such as the average run length, the conditional expected delay, and the probability of successful detection (Frisén (1992) are of interest. For multivariate problems, generalizations are necessary. For details see Frisén, Andersson and Schiöler (2009a).

The ARL^1 is the most commonly used measure of the detection ability, also in the multivariate case. Here it is assumed that all variables change immediately. There is no natural generalization of the ARL^1 to cases with different change points. The steady state average run length has been used for multivariate surveillance. When this measure is used it also seems to be assumed that all changes appear simultaneously. When the changes can appear at different times we need other measures.

The detection ability depends on when the changes occur. In the univariate case, the conditional expected delay $CED(t) = \mathrm{E}\left[t_A - \tau \mid t_A \geq \tau = t\right]$ is useful. In the case that we are concentrating on here, where the first change $\tau_{\min} = \min\{\tau_1, \ldots \tau_p\}$ is of interest, a generalization is

$$CED(\tau_1,...\tau_p) = E(t_A - \tau_{\min} \mid t_A \geq \tau_{\min}).$$

The Probability of Successful Detection, suggested by (Frisén (1992), measures the probability of detection with a delay time no longer than a specified value, d. In the multivariate case it can be generalized as

$$PSD(d,\tau_1,...\tau_p) = P(t_A - \tau_{\min} \leq d \mid t_A \geq \tau_{\min}).$$

This measure is a function of both the times of the changes and the length of the interval in which the detection is defined as successful. Also when there is no absolute limit to the detection time, it is often useful to describe the ability to detect the change within a certain time. In such cases it may be useful to calculate the PSD for different time limits d. This has been done for example by Marshall et al. (2004) in connection with the use of the FDR.

5 Sufficient reduction and optimality

Optimality is hard to achieve and even hard to define for all multivariate problems. We will now study some classes of problems where we can get sharp results.

It is important to structure the problem carefully and to determine which type of change to focus on and which not. If we focus on detecting all kinds of changes, the detection ability of the surveillance method for each specific type of change will be inferior. We have a spectrum of problems where one extreme is that there are hardly any relations between the components. The other extreme is that the changes occur simultaneously. Consider, for example, the case when we measure several components of an assembled item. If we restrict our attention to a general change in the factory, changes will be expected to occur at the same time.

We will study how the sufficiency principle can be used to reduce some multivariate problems to simpler ones. For details see Frisén, Andersson and Schiöler (2009c). For a shift at τ in a univariate distribution between two fully specified distributions, the set of likelihood ratios $L(s,t) = f_Y^s(Y^s \mid \tau=t) / f_Y^s(Y^s \mid D)$ is sufficient for the family of distributions of Y^s defined by the time of change τ. In a multivariate situation, a statistic Z is sufficient for a family of distributions $\mathcal{F}$ if and only if $f_{Y|Z}(Y|z)$ is the same for all distributions belonging to the family $\mathcal{F}$. A sequence $Z^1(Y_1)$, $Z^2(Y_2)$,… is a sufficient sequence of statistics for the families $\mathcal{F}^1$, $\mathcal{F}^2$ … of distributions if for all s, $Z^s(Y_s)$ is a sufficient statistic for the family $\mathcal{F}^S$. The use of the sufficient statistic implies that no information is lost. The result by Wessman (1998) is that when all the variables change at the same time, a sufficient reduction to univariate surveillance exists.

The ARL^1 and steady state average run length are usually calculated for the case when all variables change simultaneously. However, for simultaneous changes, as was seen above, there exists a sufficient reduction of the surveillance problem. Thus, the case of simultaneous changes is not a genuinely multivariate situation. The measures which are based on simultaneous changes can thus not be recommended as a base for formal optimality criteria in multivariate surveillance.

Järpe (2001) found the optimal way for detecting an increased radiation level after a nuclear disaster through geographically spaced measurement stations in Sweden. He used a known lag between the changes at the stations due to wind dissemination and demonstrated that a sufficient reduction to univariate surveillance was possible.

Frisén, Andersson and Schiöler (2009b) used a semiparametric method for outbreak detection on the influenza incidence in Sweden. Schiöler (2008) investigated several geographical patterns and found a consistent time lag between two kinds of regions. The sufficient reduction can be used to increase the efficiency of the early detection of influenza outbreaks.

6 Concluding remarks

Multivariate surveillance is challenging in many ways. It involves statistical theory, practical issues concerning the collection of new types of data, and computational issues such as the implementation of automated methods in large scale surveillance data bases. In this paper the focus has been on the

statistical inference aspects and especially the effect of a sufficient reduction of the multivariate surveillance problem.

The question of which multivariate surveillance method is the best has no concise answer. Different methods are suitable for different problems. The more specifically the aim is stated, the greater the possibilities that the surveillance will meet this aim. Multivariate variants of good univariate surveillance methods can be constructed by stepwise reductions of the problem in one way or other. However, the stepwise reduction might lead to a loss of information and suboptimal solutions. The joint likelihood ratio contains all the information and is a promising base for multivariate surveillance. However, the usual optimality theorems for univariate surveillance cannot be applied directly since there can be several change points.

Some causes may lead to a simultaneous increase in all variables, and then one should use a reduction to a univariate surveillance method, as demonstrated by the sufficiency principle. If there is a known lag between the change times, then there also exists a sufficient reduction. Unpublished results show that such a reduction improved the detection of influenza outbreaks in Sweden.

References

Alt, F. B. (1985). Multivariate quality control. In *Encyclopedia of Statistical Science*, N. L. Johnson, and S. Kotz (eds), 110-122. New York: Wiley.

Basseville, M., and Nikiforov, I. (1993). *Detection of abrupt changes- Theory and application*. Englewood Cliffs: Prentice Hall.

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

Bock, D. (2008). Aspects on the control of false alarms in statistical surveillance and the impact on the return of financial decision systems. *Journal of Applied Statistics* **35**, 213-227.

Bodnar, O., and Schmid, W. (2004). CUSUM control schemes for multivariate time series. In *Frontiers in Statistical Quality Control. Intelligent Statistical Quality control*. Warsaw.

Chan, L. K., and Zhang, J. (2001). Cumulative sum control charts for the covariance matrix. *Statistica Sinica* **11**, 767-790.

Crosier, R. B. (1988). Multivariate generalizations of cumulative sum quality-control schemes. *Technometrics* **30**, 291-303.

Fricker, R. D. (2007). Directionally Sensitive Multivariate Statistical Process Control Procedures with Application to Syndromic Surveillance *Advances in Disease Surveillance* **3**, 1-17.

Frisén, M. (1992). Evaluations of Methods for Statistical Surveillance. *Statistics in Medicine* **11**, 1489-1502.

Frisén, M. (2003). Statistical surveillance. Optimality and methods. *International Statistical Review* **71**, 403-434.

Frisén, M., Andersson, E., and Schiöler, L. (2009a). Evaluation of Multivariate Surveillance. *Research Report 2009:1*: Statistical Research Unit, Department of Economics, University of Gothenburg, Sweden.

Frisén, M., Andersson, E., and Schiöler, L. (2009b). Robust outbreak surveillance of epidemics in Sweden *Statistics in Medicine* **28**, 476-493.

Frisén, M., Andersson, E., and Schiöler, L. (2009c). Sufficient reduction in multivariate surveillance. *Research Report 2009:2*: Statistical Research Unit, Department of Economics, University of Gothenburg, Sweden.

Golosnoy, V., Schmid, W., and Okhrin, I. (2007). Sequential Monitoring of Optimal PortfolioWeights. In *Financial surveillance*, M. Frisén (ed), 179-210. Chichester: Wiley.

Hotelling, H. (1947). Multivariate quality control. In *Techniques of statistical analysis*, C. Eisenhart , M. W. Hastay, and W. A. Wallis (eds). New York: McGraw-Hill.

Jackson, J. E. (1985). Multivariate quality control. *Communications in Statistics. Theory and Methods* **14**, 2657-2688.

Järpe, E. (2001). Surveillance, environmental. In *Encyclopedia of Environmetrics*, A. El-Shaarawi, and W. W. Piegorsh (eds). Chichester: Wiley.

Knoth, S., and Schmid, W. (2002). Monitoring the mean and the variance of a stationary process. *Statistica Neerlandica* **56**, 77-100.

Lowry, C. A., and Montgomery, D. C. (1995). A review of multivariate control charts. *IIE Transactions* **27**, 800-810.

Lowry, C. A., Woodall, W. H., Champ, C. W., and Rigdon, S. E. (1992). A multivariate exponentially weighted moving average control chart. *Technometrics* **34**, 46-53.

Marshall, C., Best, N., Bottle, A., and Aylin, P. (2004). Statistical issues in the prospective monitoring of health outcomes across multiple units. *Journal of the Royal Statistical Society A* **167**, 541-559.

Ngai, H. M., and Zhang, J. (2001). Multivariate cumulative sum control charts based on projection pursuit. *Statistica Sinica* **11**, 747-766.

Okhrin, Y., and Schmid, W. (2007). Surveillance of Univariate and Multivariate Nonlinear Time Series. In *Financial surveillance*, M. Frisén (ed), 153-177. Chichester: Wiley.

Pignatiello, J. J., and Runger, G. C. (1990). Comparisons of multivariate CUSUM charts. *Journal of Quality Technology* **22**, 173-186.

Rogerson, P. A. (1997). Surveillance systems for monitoring the development of spatial patterns. *Statistics in Medicine* **16**, 2081-2093.

Rosolowski, M., and Schmid, W. (2003). EWMA charts for monitoring the mean and the autocovariances of stationary Gaussian processes. *Sequential Analysis* **22**, 257-285.

Runger, G. C., Barton, R. R., Del Castillo, E., and Woodall, W. H. (2007). Optimal Monitoring of Multivariate Data for Fault Detection". *Journal of Quality Technology* **39**, 159-172.
Ryan, T. P. (2000). *Statistical methods for quality improvement*, 2nd edition. New York: Wiley.
Schiöler, L. (2008). Explorative analyzis of spatial patterns of influenza incidences in Sweden 1999-2008. In *Research report*: Statistical Research Unit, Department of Economics, Göteborg University, Sweden.
Shewhart, W. A. (1931). *Economic Control of Quality of Manufactured Product*. London: MacMillan and Co.
Sonesson, C., and Frisén, M. (2005). Multivariate surveillance. In *Spatial surveillance for public health*, A. Lawson, and K. Kleinman (eds), 169-186. New York: Wiley.
Wessman, P. (1998). Some Principles for surveillance adopted for multivariate processes with a common change point. *Communications in Statistics. Theory and Methods* **27**, 1143-1161.
Woodall, W. H., and Amiriparian, S. (2002). On the economic design of multivariate control charts. *Communications in Statistics -Theory and Methods* **31**, 1665-1673.
Woodall, W. H., and Ncube, M. M. (1985). Multivariate cusum quality control procedures. *Technometrics* **27**, 285-292.
Wärmefjord, K. (2004). Multivariate quality control and Diagnosis of Sources of Variation in Assembled Products. Licentiat Thesis, Göteborg University, Gbg.
Yashchin, E. (1994). Monitoring Variance Components. *Technometrics* **36**, 379-393.