Research Report
Statistical Research Unit
Department of Economics
University of Gothenburg
Sweden

Sufficient reduction in multivariate surveillance

**M. Frisén, E. Andersson &
L. Schiöler**

**Research Report
2009:2
ISSN 0349-8034**

Mailing address:	Fax	Phone	Home Page:
Statistical Research Unit	Nat: 031-786 12 74	Nat: 031-786 00 00	http://www.statistics.gu.se/
P.O. Box 640	Int: +46 31 786 12 74	Int: +46 31 786 00 00	
SE 405 30 Göteborg			
Sweden			

Sufficient reduction in multivariate surveillance

MARIANNE FRISÉN, EVA ANDERSSON, AND LINUS SCHIÖLER
Statistical Research Unit, University of Gothenburg, Göteborg, Sweden

The relation between change points in multivariate surveillance is important but seldom considered. The sufficiency principle is here used to clarify the structure of some problems, to find efficient methods, and to determine appropriate evaluation metrics. We study processes where the changes occur simultaneously or with known time lags. The surveillance of spatial data is one example where known time lags can be of interest. A general version of a theorem for the sufficient reduction of processes that change with known time lags is given. A simulation study illustrates the benefits or the methods based on the sufficient statistics.

Keywords change-points, exponential family, MEWMA, monitoring, inference principles

1. Introduction

In society there is a great need for continuous surveillance of processes with the aim of detecting an important change in the underlying process as soon as possible after the change has occurred. The inference is quite different in on-line surveillance as compared to hypothesis testing. In surveillance there are no fixed hypotheses. Even if the situation is stable at the current time, a change can happen later. Timeliness is important in surveillance. Since the probability of a false alarm increases with time and tends to one for most surveillance methods, evaluation by significance level, power, and other well-known metrics is not useful for ordinary surveillance problems. Some surveillance methods have been constructed to resemble methods for hypothesis testing, see for example Chu, Stinchcombe, & White, (1996). These methods are constructed to have a false alarm probability less than one. This could be an advantage, since it allows statements like those in hypothesis testing to be made. However, Frisén, (2003), Aue & Horvath, (2004), and Bock, (2008) demonstrated that the detection ability of these methods declines rapidly for late changes. These methods are suitable only for applications where a possible change appears at or soon after the start. Sometimes methods like the CUSUM method by Page, (1954) or the Shiryaev-Roberts method by Shiryaev, (1963), which were constructed to be optimal for on-line surveillance, are demonstrated to be useful also for retrospective hypothesis testing, as in Lee, Ha, Na, & Na, (2003) and Vexler & Wu, (2009). There are problems situated between hypothesis testing and surveillance, but in this paper we will deal only with inference suitable for on-line surveillance.

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 multiple 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. Since the collected data involve several related variables, this calls for multivariate surveillance techniques. The surveillance of several parameters of one distribution (such as the mean and the variance of a normal distribution), see for example Knoth & Schmid, (2002), can involve the same problems as the surveillance of a multidimensional distribution originating from the observation of different variables. Spatial surveillance is useful for the detection of a local change or a spread. One example is the

Address correspondence to Marianne Frisén, Statistical Research Unit, University of Gothenburg, PO Box 640, SE 40530 Gothenburg, Sweden; E-mail: marianne.frisen@statistics.gu.se.

spread of a disease such as influenza, as in Schiöler, (2008) and Frisén, Andersson, & Schiöler, (2009). Another example is the spread of a harmful agent such as nuclear radiation, as in Järpe, (2001). 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 & Schmid, (2007) and Golosnoy, Schmid, & Okhrin, (2007)) with respect to various assets.

Reviews on multivariate surveillance methods can be found for example in Basseville & Nikiforov, (1993), Ryan, (2000), Frisén, (2003), Sonesson & Frisén, (2005a), Bersimis, Psarakis, & Panaretos, (2007), and Frisén, (2009). Optimality is hard to derive and sometimes even hard to define in multivariate problems. However, we will demonstrate how the structure of some multivariate surveillance problems can be simplified by the sufficiency principle and how this will lead to more efficient methods than those suggested earlier.

At each time point, a new observation is made on the process. The p-variate process under surveillance is denoted by $\mathbf{Y} = \{\mathbf{Y}(t),\ t = 1,2,...\}$, where $Y(t) = \{Y_1(t), Y_2(t),..., Y_p(t)\}$. We aim to detect the change from a stable state D to a harmful state C as soon as possible after the change has occurred, in order to give warnings and take corrective actions. At decision time s we base the decision on the available information $Y^s = \{Y(1), Y(2)... Y(s)\}$ and use the observation vector y^s to form an alarm statistic. An alarm is called the first time that the statistic exceeds an alarm limit. In the univariate case, the change happens at the unknown time point τ. In the multivariate case, we observe p processes which can change at different times $\tau_1, .. \tau_p$. Here an important aim is to detect the first time that not all processes is are in control, that is, we want to make inference about $\tau_{\min} = \min\{\tau_1,...,\tau_p\}$. If no change ever occurs in process i, we denote this by "$\tau_i = \infty$". We consider models where the observations $Y_i(t)$ and $Y_i(t+j)$ are independent, given the values of the change points, and for each variable, i, there is one distribution, $f_i^0(t)$, for $t < \tau_i$ and another, $f_i^1(t)$, for $t \geq \tau_i$. In this paper we concentrate on the one-parameter exponential family.

In Section 2 different approaches to the construction of multivariate surveillance methods are described and exemplified. Theoretical results on sufficient reduction are given in Section 3. In Section 4 we discuss the challenges of evaluating multivariate surveillance methods with special focus on how the structure of the multivariate problem is clarified by the sufficiency principle. In Section 5 we illustrate the theory by a simulation study. Concluding remarks are made in Section 6.

2. Approaches to multivariate surveillance

Some commonly used general approaches for adapting univariate methods to multivariate surveillance will be described and exemplified. Principal differences between approaches for handling multivariate data in surveillance will be demonstrated.

2.1. *Dimension reduction*

In Statistical Process Control (SPC) it is practical to use only one control chart instead of several. Thus many suggestions have been made on reduction to one chart (see e.g. Cheng & Thaga, (2006)). A stepwise reduction of the multivariate surveillance problem is natural. An easy way to simplify the situation is to reduce the p-variate vector at each time point into one statistic and then use a system for univariate surveillance on this statistic. One example is the suggestion by Crosier, (1988) to summarize data by the Hotelling T variable and then apply the univariate CUSUM method to the T variable, making it a scalar accumulation method. As

we will describe in Section 3, a sufficient dimension reduction can be found for some situations.

2.2. *Parallel surveillance*

A stepwise solution of the multivariate surveillance problem can alternatively be accomplished by monitoring each variable separately. The approach with parallel systems is often called "combined univariate" methods or "parallel" methods. The most common way to combine the information from several univariate methods is to signal an alarm at the first time that any of the univariate methods gives an alarm. This is a special case of the union-intersection technique suggested by Roy, (1953).

2.3. *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. Thus a surveillance method is applied to each of the p processes, resulting in p-variate alarm statistics at each decision time s. This p-variate statistic is then transformed into a scalar, which is the alarm statistic for the whole system at time s. An alarm is triggered if this statistic exceeds a limit. As an example, Lowry, Woodall, Champ, & Rigdon, (1992) proposed a multivariate extension, MEWMA, of the univariate EWMA. The MEWMA method uses a vector of univariate EWMA statistics. For each variable Y_j and each time t, we have the EWMA statistic $Z(t)=\lambda_j Y_j(t)+(1-\lambda_j)Z_j(t-1)$ where Z(0)=0. At decision time s we have $\mathbf{Z}(s)=\{Z(\mathbf{y}_1^s), Z(\mathbf{y}_2^s),..., Z(\mathbf{y}_p^s)\}$. An alarm is triggered at $t_A = \min\{s;\quad \mathbf{Z}(s)^T \boldsymbol{\Sigma}_{\mathbf{Z}(s)}^{-1} \mathbf{Z}(s) > L\}$. The properties of the method are described in Section 5.2.2. Vector accumulation methods based on CUSUM have also been proposed, but there are several possibilities of how to handle the characteristic barrier of the CUSUM methods (see Sonesson & Frisén, (2005b)).

2.4. *Joint solution*

A joint solution of the original full problem, without stepwise solutions, is preferred when possible. In general, the likelihood ratios are sufficient for the problem and so is the set of partial likelihoods for surveillance problems:

$$L(s,m_1,...m_p)=\frac{f(Y_1^s,...Y_p^s \,|\, \tau_1 = m_1,...\tau_p = m_p)}{f(Y_1^s,...Y_p^s \,|\, \tau_1 > s,...\tau_p > s)} \ .$$

The full likelihood ratio method for the multivariate problem (see for example Andersson, (2009)) requires knowledge of the distribution of the change times. When the full likelihood for $\mathbf{Y}^{\mathbf{s}} = \{\mathbf{Y}(1), \mathbf{Y}(2)... \mathbf{Y}(s)\}$ is available, it provides a good basis for surveillance since optimal methods are mostly constructed based on the likelihood. However, the full likelihood can be complicated for some problems, and therefore a reduction may be considered. A sufficient reduction will not reduce the information, but other reductions will. A jointly optimal solution can be constructed by a sufficient reduction (where no information is lost in the reduction step), followed by an optimal surveillance method applied to the reduced statistic. Stepwise approaches which start with a reduction (either in time or in the variables) and then use a possibly optimal univariate method can be suspected to be suboptimal. Only reductions which are sufficient can be expected to result in jointly optimal solutions, since no information is lost.

3. Sufficient reduction

A statistic T is sufficient for a family of distributions if and only if $f_{Y|T}(y|t)$ is the same for all distributions belonging to the family $\mathcal{F}$ (see for example Cox & Hinkley, (1974)). A sequence $T^1(Y_1)$, $T^2(Y_2)$,… is a sufficient sequence of statistics for the distributional families $\mathcal{F}^1$, $\mathcal{F}^2$ … if for all s, $T^s(Y_s)$ is a sufficient statistic for the family $\mathcal{F}^s$.

For a shift at τ in a univariate distribution between two fully specified distributions, the set of likelihood ratios $L(s,t) = f_{Y_s}(Y^s|\tau=t) / f_{Y_s}(Y^s|D)$ is sufficient for the distributional family of Y^s defined by the time of change τ.

According to the sufficiency principle, all conclusions to be drawn should depend on one sufficient statistic only.

3.1. *Simultaneous changes*

Consider the case where all processes have the same change point so that $\tau_1 = \tau_2 = \ldots \tau_p = \tau$. For a change at τ between the distributions f^0 and f^1 we have the distribution for the s observations

$$f(Y_s \mid \tau = m) = \prod_{t=1}^{m-1} f^0(Y(t)) \prod_{t=m}^{s} f^1(Y(t)) = \prod_{t=1}^{s} f^0(Y(t)) \prod_{t=m}^{s} \frac{f^1(Y(t))}{f^0(Y(t))}.$$

It now becomes possible to identify the separate factors: the part that depends on the data (but not the value of τ) as well as the part that depends on the s-dimentional vector $L^s(\mathbf{Y}^s)=\{L(s,m), t=1,...s\}$, where m is the common change time. Thus $L^s(\mathbf{Y}^s)$ is sufficient for the distributional family for each s. From this it follows that the sequence of s likelihood ratios is a sufficient sequence. This was proven by Wessman, (1998) both for a fixed unknown value of τ and for a stochastic time of change. When the aim is to detect a fully specified, simultaneous change in a multivariate process and the distributions before and after the change are fully specified, it is possible to construct a univariate surveillance procedure based on the sufficient sequence of likelihood ratios. Examples will be given in the next section as special cases of the general theorem in the next section.

3.2. *Changes with time lags*

We will now consider the case where there are known time lags between the changes of the p processes. In the context of nuclear incidents, Järpe, (2000) studied measurements at different geographical locations in Sweden. Several models for the spread of radioactive material by the wind were studied. At each location, the radioactivity increased with a time lag which was assumed to be proportional to the distance from the source (a nuclear plant). For the situation with a shift of equal size in the expected value of Gaussian processes, when the shifts occur with known lags and where we have independent (given the change points) normal distributions with the same variances, Järpe, (2000) demonstrated that a sufficient reduction to univariate surveillance exists. Here we will prove that a sufficient reduction to a univariate statistic exists as long as the processes belong to the one-parameter exponential family.

Theorem

For p processes which all belong to the one-parameter exponential family and which are independent (conditional on the change points), there exists a sufficient reduction of the observation vectors {y_1, y_2, ...y_p} to a univariate statistic for the detection of shifts in the parameter vector when the changes occur with known time lags (q_2, q_3,... q_p) where $q_i=\tau_i - \tau_{i-1}$. A sufficient statistic for the detection of shifts of sizes $\delta_1, \delta_2, ...\delta_s$ is the set

$\delta_1 y_1(t) + \delta_2 y_2(t+q_2) + ... + \delta_p y_p(t+q_p)$, for $1 \le t \le s\text{-}q_2\text{ -}q_3\text{ -...-}q_p$

$\delta_1 y_1(t) + \delta_2 y_2(t+q_2) + ... + \delta_{p-1} y_{p-1}(t+q_{p-1})$, for $s\text{-}q_2\text{ -}q_3\text{ -...-}q_p < t \le s\text{-}q_2\text{ -}q_3\text{ -...-}q_{p-2}$,

…

$\delta_1 y_1(t) + \delta_2 y_2(t+q_2)$, for $s\text{-}q_2\text{ -}q_3 < t \le s\text{-}q_2$,

$y_1(t)$, for $s\text{-}q_2 < t \le s$.

Proof

Since the observations are independent given the values of the change points, the distribution can be written as a product. The likelihood expressions for the exponential family can be written as

$$f(Y \mid \tau_{\min} \le s) =$$

$$\exp\left\{\sum_{t=1}^{\tau_1-1}\sum_{j=1}^{p}\left[y_j(t)\cdot(\varphi_j) + g(\varphi_j) + h(y_j(t))\right]\right\} +$$

$$\exp\left\{\sum_{t=\tau_1}^{\tau_2-1}\left[\sum_{j=1}^{1}\left[y_j(t)\cdot(\varphi_j+\delta_j) + g(\varphi_j+\delta_j) + h(y_j(t))\right] + \sum_{j=2}^{p}\left[y_j(t)\cdot(\varphi_j) + g(\varphi_j) + h(y_j(t))\right]\right]\right\} +$$

$$+ ... + \exp\left\{\sum_{t=\tau_{p-1}}^{\tau_p-1}\left[\sum_{j=1}^{p-1}\left[y_j(t)\cdot(\varphi_j+\delta_j) + g(\varphi_j+\delta_j) + h(y_j(t))\right] + \sum_{j=p}^{p}\left[y_j(t)\cdot(\varphi_j) + g(\varphi_j) + h(y_j(t))\right]\right]\right\} +$$

$$\exp\left\{\sum_{t=\tau_p}^{s}\sum_{j=1}^{p}\left[y_j(t)\cdot(\varphi_j+\delta_j) + g(\varphi_j+\delta_j) + h(y_j(t))\right]\right\}$$

and

$$f(Y \mid \tau_{\min} > s) =$$

$$\exp\left\{\sum_{t=1}^{s}\sum_{j=1}^{p}\left[y_j(t)\cdot(\varphi_j) + g(\varphi_j) + h(y_j(t))\right]\right\}$$

The likelihood ratio, conditional on $\tau_{\min}$=m, equals $L(s,m) = \dfrac{f(Y \mid \tau_{\min} = m \le s)}{f(Y \mid \tau_{\min} > s)}$ and thus the log likelihood ratio is

$$\sum_{t=1}^{m-1}\sum_{j=1}^{p}\left[y_j(t)\cdot(\varphi_j)+g(\varphi_j)+h(y_j(t))\right]+$$

$$\sum_{t=m}^{m+q_2-1}\left[\sum_{j=1}^{1}\left[y_j(t)\cdot(\varphi_j+\delta_j)+g(\varphi_j+\delta_j)+h(y_j(t))\right]+\sum_{j=2}^{p}\left[y_j(t)\cdot(\varphi_j)+g(\varphi_j)+h(y_j(t))\right]\right]+$$

$$\sum_{t=m+q_2}^{m+q_2+q_3-1}\left[\sum_{j=1}^{2}\left[y_j(t)\cdot(\varphi_j+\delta_j)+g(\varphi_j+\delta_j)+h(y_j(t))\right]+\sum_{j=3}^{p}\left[y_j(t)\cdot(\varphi_j)+g(\varphi_j)+h(y_j(t))\right]\right]+$$

$$+\ldots+$$

$$+\sum_{t=m+q_2+\ldots+q_{p-1}}^{m+q_2+\ldots+q_p-1}\left[\sum_{j=1}^{p-1}\left[y_j(t)\cdot(\varphi_j+\delta_j)+g(\varphi_j+\delta_j)+h(y_j(t))\right]+\sum_{j=p}^{p}\left[y_j(t)\cdot(\varphi_j)+g(\varphi_j)+h(y_j(t))\right]\right]+$$

$$+\sum_{t=m+q_2+\ldots+q_p}^{s}\sum_{j=1}^{p}\left[y_j(t)\cdot(\varphi_j+\delta_j)+g(\varphi_j+\delta_j)+h(y_j(t))\right]-$$

$$-\sum_{t=1}^{s}\sum_{j=1}^{p}\left[y_j(t)\cdot(\varphi_j)+g(\varphi_j)+h(y_j(t))\right]$$

This can be arranged into

$$\sum_{t=m}^{s-q_2-\ldots-q_p}y_1(t)\cdot\delta_1+\sum_{t=s-q_2-\ldots-q_p+1}^{s-q_2-\ldots-q_{p-1}}y_1(t)\cdot\delta_1+\ldots+\sum_{t=s-q_2-q_3+1}^{s-q_2}y_1(t)\cdot\delta_1+\sum_{t=s-q_2+1}^{s}y_1(t)\cdot\delta_1$$

$$+\sum_{t=m+q_2}^{s-q_3-\ldots-q_p}y_2(t)\cdot\delta_2+\sum_{t=s-q_3-\ldots-q_p+1}^{s-q_3-\ldots-q_{p-1}}y_2(t)\cdot\delta_2+\ldots+\sum_{t=s-q_3+1}^{s}y_2(t)\cdot\delta_2$$

$$+\ldots+\sum_{t=m+q_2+\ldots+q_{p-1}}^{s-q_p}y_{p-1}(t)\cdot\delta_{p-1}+\sum_{t=s-q_p+1}^{s}y_{p-1}(t)\cdot\delta_{p-1}$$

$$+\sum_{t=m+q_2+\ldots+q_p}^{s}y_p(t)\cdot\delta_p$$

$$+z(\delta_1,\delta_2,\ldots,\delta_p,\varphi_1,\varphi_2,\ldots,\varphi_p)$$

where $z(\delta_1, \delta_2,\ldots \delta_p, \varphi_1, \varphi_2,\ldots \varphi_p) =$

$$\sum_{t=m}^{s}\left(g(\varphi_1+\delta_1)-g(\varphi_1)\right)$$

$$+\sum_{t=m+q_2}^{s}\left(g(\varphi_2+\delta_2)-g(\varphi_2)\right)$$

$$+\ldots+\sum_{t=m+q_2+\ldots+q_{p-1}}^{s}\left(g(\varphi_{p-1}+\delta_{p-1})-g(\varphi_{p-1})\right)$$

$$+\sum_{t=m+q_2+\ldots+q_p}^{s}\left(g(\varphi_p+\delta_p)-g(\varphi_p)\right)$$

is independent of the observations. The expression above can be rewritten as

$$\sum_{t=m}^{s-q_2-\ldots-q_p}\left[\left(y_1(t)\cdot\delta_1\right)+\left(y_2(t+q_2)\cdot\delta_2\right)+\ldots+\left(y_{p-1}(t+q_2+\ldots+q_{p-1})\cdot\delta_{p-1}\right)+y_p(t+q_2+\ldots+q_p)\cdot\delta_p\right]$$

$$+\sum_{t=s-q_2-\ldots-q_p+1}^{s-q_2-\ldots-q_{p-1}}\left[y_1(t)\cdot\delta_1+y_2(t+q_2)\cdot\delta_1+\ldots+y_{p-1}(t+q_2+\ldots+q_{p-1})\cdot\delta_{p-1}\right]$$

$$+\ldots+\sum_{t=s-q_2-q_3+1}^{s-q_2}\left[y_1(t)\cdot\delta_1+y_2(t+q_2)\cdot\delta_2\right]$$

$$+\sum_{t=s-q_2+1}^{s}y_1(t)\cdot\delta_1$$

$$+z(\delta_1,\delta_2,\ldots,\delta_p,\varphi_1,\varphi_2,\ldots,\varphi_p)$$

Thus logL(s,m) is a one-one function of the statistic in the Theorem, and thus it is a sufficient statistic for L(s,m) and thus for the problem.

The Theorem is general and thus has many parameters. In order to illustrate the idea we will now look at some special cases. The performance for these special cases will be illustrated in Section 5.

Corollary 1

A special case of the Theorem concerns two processes (p=2) when the changes occur at the same time (q=0). In this situation we have by the Theorem that $\delta_1 Y_1(t)+\delta_2 Y_2(t)$ for $1\le t\le s$ is sufficient. If, for example, $\delta_1=2\delta_2$ we have that $2\delta_2 Y_1(t)+\delta_2 Y_2(t)$ is sufficient. From this it follows that the statistic $\frac{2}{3}Y_1(t)+\frac{1}{3}Y_2(t)$ is sufficient. If we have equal shifts in the parameter vector ($\delta_1=\delta_2=\delta$), then $\delta Y_1(t)+\delta Y_2(t)$ is a sufficient statistic. From this it follows that the set of means of the observations

$$\text{SuffR}^0(t)=\frac{Y_1(t)+Y_2(t)}{2}$$

is sufficient.

Corollary 2

Another special case of the Theorem concerns two processes (p=2) which have equal shifts in the parameter vector ($\delta_1=\delta_2=\delta$) and where the changes occur with a known time lag q. In this situation we have, by the Theorem, that a sufficient statistic is the set

$\delta\left(Y_1(t)+Y_2(t+q)\right)$ for t=1,…s-q,

$\delta Y_1(t)$ for t=s-q+1, …s

We need two arguments to specify the statistic when q>0, since the series changes when s increases. For q=1 a sufficient statistic is the set

$\{\text{SuffR}^1(s, t)\}$, for t=1, 2, … s.

Thus, for s=1, the sufficient set is

$\{\text{SuffR}^1(1, 1)=Y_1(1)\}$.

For s=2, the sufficient set is

$\{\text{SuffR}^1(2, 1)=\left(Y_1(1)+Y_2(2)\right)/2, \text{SuffR}^1(2, 2)=Y_1(2)\}$.

For s=3, the sufficient set is

$\{\text{SuffR}^1(3, 1)=\left(Y_1(1)+Y_2(2)\right)/2, \text{SuffR}^1(3, 2)=\left(Y_1(2)+Y_2(3)\right)/2, \text{SuffR}^1(3, 3)=Y_1(3)\}$.

For q=5, a sufficient statistic is the set

$\{\text{SuffR}^5(s, t)=\{\left(Y_1(t)+Y_2(t+5)\right)/2,...\left(Y_1(s-5)+Y_2(s)\right)/2, \quad Y_1(s-4),...Y_1(s)\}$, for t=1, 2…s.

The main theory of statistical surveillance is constructed for a change between two distributions – one for $t<\tau_i$ and another for $t\geq\tau_i$,. The $\text{SuffR}^q(s,t)$ statistic does not necessarily change between two distributions for $q>0$. For iid Gaussian distributions (conditional on τ_i) with expected values μ^0 for $t<\tau_i$ and μ^1 for $t\geq\tau_i$., and constant variance σ^2, the distributions of the sufficient $\text{SuffR}^q(s,t)$ statistics have the expected value μ^0 for $t< \tau_{min}$ and μ^1 for $t\geq \tau_{min}$.. However, the variance is not the same for $t>q$ as for $t\leq q$. For example, for a lag of 1, the variance for $\text{SuffR}^1(2,1)$ equals $\sigma^2/2$, whereas the variance for $\text{SuffR}^1(2,2)$ equals σ^2. Other transformations, which are also sufficient, could be considered. One alternative is to divide the sums in the sufficient statistic SuffR^q with $\sqrt{2}$ instead of 2. This results in a constant variance for all components but not constant expected values. For $t\geq \tau_{min}$ the expected value shifts from $\sqrt{2}\mu^1$ for the first components of the series to μ^1 (for the last components). This seems like a larger drawback, and we will thus study the $\text{SuffR}^q(s,t)$ statistic in the examples in Section 5.4. In spite of the fact that we cannot rely on theoretical optimality (since the SuffR^q statistic does change between more than two distributions), we will see that the statistic works well.

4. Evaluation

4.1. *Optimality*

It can be difficult to find a definition of optimality that holds for all different aspects of multivariate problems in surveillance, see Frisén, (2003). In multivariate problems there are always many dimensions to consider. In surveillance there is the additional complexity of the different relations between the change points, ranging from simultaneous changes to independent changes. Nevertheless, sufficient reductions make it possible to find optimal solutions for at least one important situation.

After sufficient reduction to a univariate statistic, we can use earlier optimality results of univariate surveillance. Different combinations of the partial likelihood ratios are known to have different optimality properties, as described by Frisén, (2003). In Frisén & de Maré, (1991) it is shown that the full likelihood ratio method, which is a weighted sum of L(s,t), with the weights proportional to $P(\tau=t)$, yields a minimal expected delay in univariate surveillance. This follows from the results by Shiryaev, (1963), where optimality is shown when the change point follows a geometric distribution. Another function of the partial likelihood ratios is the maximum likelihood ratio component L(s,t) with respect to t. This alarm statistic is mini-max optimal, as proved by Moustakides, (1986). The EWMA method was demonstrated by Frisén, (2003) and Frisén & Sonesson, (2006) to be an approximation of the full likelihood ratio method.

For simultaneous changes, it was demonstrated in Section 3 that the multivariate problem can be reduced to a univariate problem of a change between two distributions: one for $t<\tau$ and another for $t\geq\tau$. Thus, the ordinary theory of optimal surveillance can be applied. Surveillance

of the sufficient statistic by an optimal univariate method is thus optimal for the multivariate problem.

In the multivariate setting with different change points, the full likelihood ratio equals the joint solution. We may be able to find the full likelihood ratio, weighted by the geometric distribution of τ, which in the univariate case guarantees a minimal delay. Sun & Basu, (1995) studied multivariate surveillance with p=2 and used the assumption that (τ_1,τ_2) follows a bivariate geometric distribution. This means that also τ_{min} follows a geometric distribution. If τ_{min} is considered as the change point, then the requirement of a geometric distribution is satisfied. However, in proofs for optimality such as those of Shiryaev, (1963) and Moustakides, (1986), it is also required that $\mathbf{Y}(t)$ is independently and identically distributed before as well as after the change point. The requirement of identical distributions is not satisfied for $\mathbf{Y}(t)$ for all t after τ_{min}, for the situation when there are several change points. Nevertheless, the different types of combinations of partial likelihood expressions (as described above) can be assumed to be suitable for different types of (approximate) optimality. In Section 5, examples will be used to demonstrate that the methods based on the sufficient statistic work well also for situations where optimality cannot be proven.

4.2. *Evaluation measures in multivariate surveillance*

The most commonly used measure of delay of the time, t_A, of the alarm is $ARL^1 = E[t_A \mid \tau = 1]$ which is also called the zero state ARL since it is a measure of the delay when the change happens immediately. A measure for the opposite situation, when the change time tends to infinity, is the steady state ARL (see for example Lu & Reynolds Jr, (1999) and Reynolds & Kim, (2007)). In univariate surveillance this measure is unique for specified distributions and a specified method. In a multivariate setting, however, this measure is not unique but depends on the relation between the change points when they tend to infinity. It is common to calculate the measure for the situation of simultaneous changes even if the assumption of simultaneous changes is only implicit. However, as was pointed out in Section 3.1, the situation with simultaneous changes is not a genuine multivariate problem since it can be reduced to a univariate one. As was seen in Section 3.1, there are optimal methods for this situation.

In {Frisén, 2009 #382} the conditional expected delay was recommended for situations with different relations between the τ-values

$$\mathrm{CED}(\tau_1, \tau_{2\ldots} \tau_p) = E[t_A - \tau_{\min} \mid \tau_{\min} \geq t_A].$$

This measure will be used in the next section to evaluate methods for different situations.

5. Examples

In order to illustrate the performance of different multivariate methods, especially those based on reduction, we apply them to a number of different situations. We will concentrate on the way in which the relations between the change times, τ_1, τ_2, ..., τ_p, influence the properties of different surveillance methods. In Section 5.1 we give a simple model which will be used in the simulation study, in Section 5.2 we describe the methods which are compared, and in Sections 5.3 and 5.4, respectively, we report the results for simultaneous changes and changes with different change points.

5.1. *Simple model*

A very simple example with two processes will be used. The two processes, Y_1 and Y_2, are assumed to be independent (conditional on the change times)

$$Y_1(t) \sim \begin{cases} N(0,1) & t < \tau_1 \\ N(2,1) & t \geq \tau_1 \end{cases}$$

$$Y_2(t) \sim \begin{cases} N(0,1) & t < \tau_2 \\ N(2,1) & t \geq \tau_2 \end{cases}$$

5.2. *Methods*

In Section 2 we described how univariate techniques can be generalized to handle multivariate situations. We have chosen the EWMA method as the method for accumulating the information over time, since it is commonly used also in multivariate situations. The EWMA method was introduced in the quality control literature by Roberts, (1959) and has received much attention. As regards the variance of the EWMA statistic there are two versions: the exact and the asymptotic variance. We will use the asymptotic variance, both for simplicity and on the basis of the arguments given in Frisén & Sonesson, (2006) concerning properties. At time s the statistic of the EWMA method for the univariate surveillance of Y is

$$Z(s)=\lambda(1-\lambda)^s\sum_{t=1}^{s}(1-\lambda)^{-t}Y(t),$$

where $0<\lambda\leq 1$ and Z_0 is the target value, which is zero in the examples. The EWMA statistic is a weighted sum of all observations available at the decision time s. Here we choose the value $\lambda=0.35$. For the comparisons we set alarm limits to ensure the same median run length to a false alarm ($MRL^0=100$). We will compare the results of several approaches to multivariate surveillance: i) the EWMA method applied to a sufficient reduction of data, ii) the MEWMA method, iii) a system based on two parallel EMWA methods, and iv) the EWMA method applied to the univariate process that changes first. These methods will now be described.

5.2.1 EWMA based on reduction

If the two processes in Section 5.1 have simultaneous change points ($\tau_1=\tau_2$), then the reduction to the statistic $SuffR^0(t)=(Y_1(t)+Y_2(t))/2$ is sufficient. The EWMA method can then be applied to this statistic. This reduction method is labeled $SuffR^0$ in the figures.

We will also study the reduction $SuffR^5(s,t)$ for the case of a lag of 5 ($\tau_2=\tau_1+5$). In the surveillance process the EWMA is applied to the sufficient statistics, and the time of alarm for the reduction methods is the first time when the EWMA statistic exceeds a constant alarm limit. Note that the recursive formula $Z(s) = (1-\lambda)Z(s-1)+\lambda Y(s)$, for s=1, 2,... , which can be used for a univariate statistic Y, is not always valid here. The whole $SuffR^q(t)$ series is revised at each decision time (except for q=0). Thus the original EWMA $Z(s) = \lambda(1-\lambda)^s\sum_{t=1}^{s}(1-\lambda)^{-t}SuffR^q(t)$ should be used. For lag 5 we have

$$Z(s)=(1-\lambda)^sZ_0+(1-\lambda)^{s-1}\lambda(Y_1(1)+Y_2(6))/2+ (1-\lambda)^{s-2}\lambda(Y_1(2)+Y_2(7))/2 +...$$
$$+ (1-\lambda)^2\lambda(Y_1(s-5)+Y_2(s))/2 +...+ (1-\lambda)^1\lambda Y_1(s-1) + \lambda Y_1(s).$$

5.2.2 MEWMA

MEWMA can be described as a Hotelling T^2 control chart applied to univariate EWMA statistics instead of to the original data and is thus a vector accumulation method. For our simple example and with the value of λ equal for both processes it is

$$EWMA(s) = \frac{Z_1(s)^2 + Z_2(s)^2}{\lambda/(2-\lambda)}.$$

5.2.3 Parallel EWMA

The parallel approach means that the EWMA method is applied to $Y_1(t)$ and $Y_2(t)$ separately. The time of alarm for the Parallel method is the first of either of the alarm times.

5.2.4 Univariate

For comparison we also have the results from the EWMA method applied to only one process. This corresponds to the situation when there is prior knowledge about which process will change first and therefore efficient to monitor only this one. This method is labeled "Univariate" in the diagrams.

5.3. *Results for simultaneous changes*

Below we present the results of the delay curve for the methods described above and the model in Section 5.1. First we study the situation when $\tau_1=\tau_2=\tau$. By Corollary 1, a method based on the sufficient reduction to the $SuffR^0$ statistic should be used. We compare the EWMA method based on $SuffR^0$ with the MEWMA method and the Parallel method.

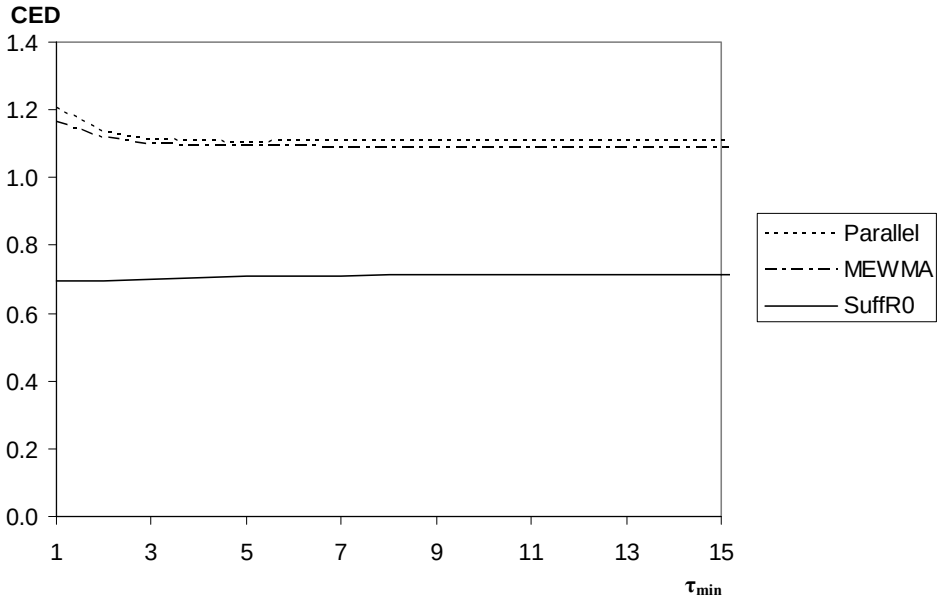

Figure 1. CED(τ_1, τ_2) vs τ_{min} for EWMA based on $SuffR^0$, EWMA Parallel, and MEWMA, when $\tau_1=\tau_2=\tau_{min}$.

In Figure 1 we see that for simultaneous changes, the EWMA method based on reduction to the statistic $SuffR^0(t)= (Y_1(t)+Y_2(t))/2$ gives the shortest delay. This is in accordance with theory, as described in Section 3.1. It may be surprising that the popular MEWMA method

gives the worst result. In this simple example, however, the flexibility of the MEWMA method does not constitute an advantage. When using the other methods it is advantageous to know the direction of the change. By contrast, the MEWMA method based on Hotelling T^2 is directionally invariant. There are suggestions of one-sided versions of MEWMA, but they were not used here.

5.4. *Results for changes with a time lag*

We now study the two variables Y_1 and Y_2 in the situation when they change with a known time lag. For the time lag of 1 unit, we find from Corollary 2 that the reduction $SuffR^1$ should be used. Correspondingly, for a known lag of 5 time units, the $SuffR^5$ should be used. In Figure 2 we examine the situation when $\tau_2=\tau_1+1$ and in Figure 3 we examine $\tau_2=\tau_1+5$. We compare the EWMA method based on the sufficient statistic for the specific situation (lag 1 or lag 5) with MEWMA, a parallel EWMA system, and EWMA based on $SuffR^0$.

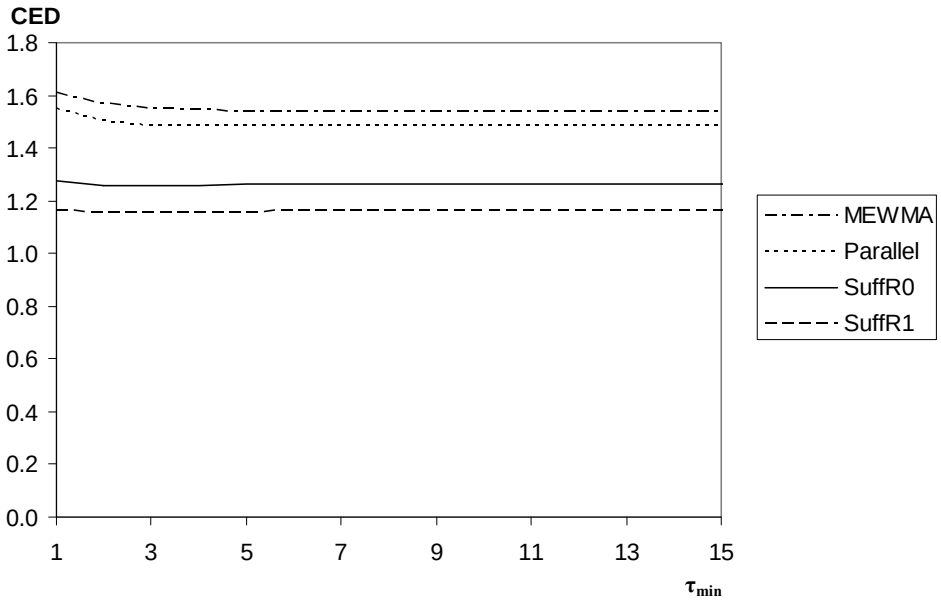

Figure 2. CED(τ_1, τ_2) vs $\tau_{min=}\tau_1$, for $\tau_2=(\tau_1+1)$ for MEWMA, EWMA Parallel, EWMA based on $SuffR^0$, and EWMA based on $SuffR^1$.

In Figure 2, we can see that EWMA based on the $SuffR^1$ reduction gives a shorter CED than the other methods for the case when $\tau_2=(\tau_1+1)$.

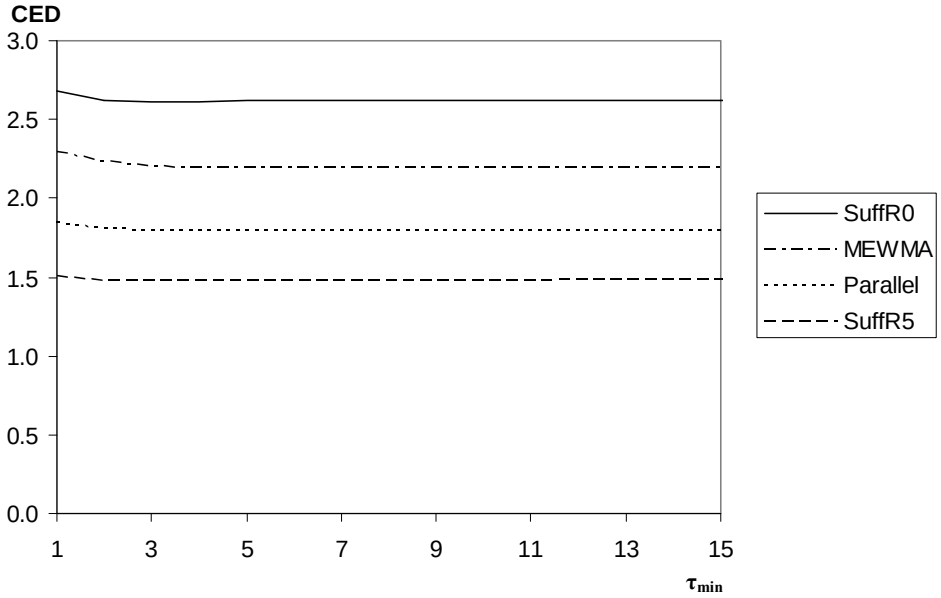

Figure 3. CED(τ_1, τ_2) vs $\tau_{min}=\tau_1$, for $\tau_2=(\tau_1+5)$ for EWMA based on SuffR^0, MEWMA, EWMA Parallel, and EWMA based on SuffR^5.

In Figure 3 we can see that EWMA based on the SuffR^5 reduction has the shortest expected delay.

If we know that only the Y_1 variable can change ($\tau_2=\infty$), then it makes sense to base the surveillance on this variable only, i.e. monitor Y_1 by univariate surveillance.

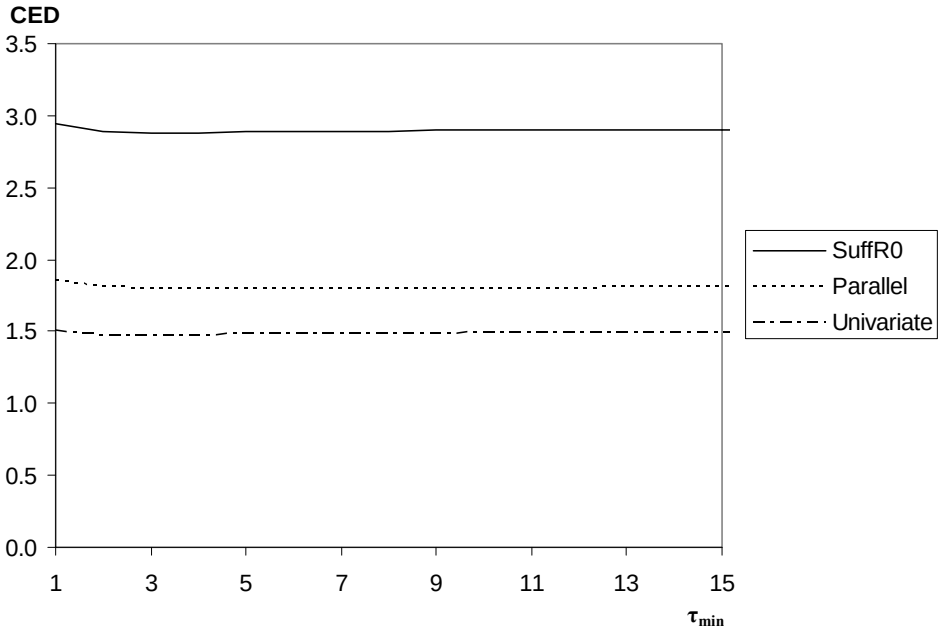

Figure 4. CED(τ_1) vs $\tau_{min} = \tau_1$ for $\tau_2=\infty$, presented for EWMA Parallel, EWMA Univariate, and EWMA based on SuffR^0.

In Figure 4 we see that for $\tau_2=\infty$ the univariate EWMA based on Y_1 is clearly the best alternative. Thus, knowledge considerably improves the CED of the surveillance.

The conclusion is that for simultaneous changes ($\tau_1=\tau_2$), EWMA based on the $SuffR^0$ reduction gives the shortest delay. This is in accordance with theory, see Wessman, (1998). However, if there is a long time interval between the changes as in Figure 3, or if only one process changes as in Figure 4, the reduction to $SuffR^0$ is not favorable.

6. Discussion

Since many important problems involve several data sources, multivariate surveillance has attracted much interest. It is challenging in many ways. Multivariate surveillance 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 impact of the relation between the change points is seldom considered. However, here it was demonstrated that the relations between the change points do have a great impact and can be utilized to find efficient methods.

Evaluations are often made by the ARL^1 or the steady state ARL, together with an implicit assumption that all processes change simultaneously. However, if the processes do change simultaneously, there exists a sufficient reduction to a univariate statistic which should be the base for optimal surveillance. Genuinely multivariate problems with different change points should be evaluated by generalized metrics, as suggested in this paper.

According to the sufficiency principle, all conclusions to be drawn should depend only on a sufficient statistic. We have demonstrated that a considerable improvement can be made by basing the surveillance on the suggested $SuffR^0$ statistic instead of using the Parallel method or the MEWMA.

In the Theorem it is demonstrated, for the exponential family, that a known time lag allows a sufficient reduction. In this situation (i.e. with different change times), the sufficient statistic does not change between two distributions only, and therefore previous optimality results on how to aggregate the information over time cannot be used directly. However, we have demonstrated that for some situations, the method based on a sufficient reduction for the known lag gives the shortest delay to detection compared to a parallel approach or the MEWMA method.

It was also demonstrated – as expected – that in a situation where only one process changes, the performance is considerably improved if this knowledge is utilized in the surveillance procedure.

Acknowledgements

Kjell Pettersson has given constructive comments. The work was supported by the Swedish Emergency Management Agency (grant 0314/206).

References

Andersson, E. (2009). Effect of dependency in systems for multivariate surveillance. *Communications in Statistics. Simulation and Computation* 38:454-472.

Aue, A., Horvath, L. (2004). Delay time in sequential detection of change. *Statistics & Probability Letters* 67:221-231.

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

Bersimis, S., Psarakis, S., 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.

Cheng, S.W., Thaga, K. (2006). Single Variables Control Charts: an Overview. *Quality and Reliability Engineering International* 22:811-820.

Chu, C.-S.J., Stinchcombe, M., White, H. (1996). Monitoring structural change. *Econometrica* 64:1045-1065.

Cox, D.R., Hinkley, D.V. (1974). *Theoretical statistics*.

Crosier, R.B. (1988). Multivariate Generalizations of Cumulative Sum Quality-Control Schemes. *Technometrics* 30:291-303.

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

Frisén, M. (2009). Principles for Multivariate Surveillance. In: Lenz, H.-J., Wilrich, P.-T., eds. *Frontiers in Statistical Quality Control*.

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

Frisén, M., de Maré, J. (1991). Optimal surveillance. *Biometrika* 78:271-280.

Frisén, M., Sonesson, C. (2006). Optimal surveillance based on exponentially weighted moving averages. *Sequential Analysis* 25:379-403.

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

Järpe, E. (2000). *On univariate and spatial surveillance. Ph.D Thesis.* Göteborg University, Göteborg.

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

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

Lee, S., Ha, J., Na, O., Na, S. (2003). The cusum test for parameter change in time series models. *Scandinavian Journal of Statistics* 30:651-739.

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.

Lu, C.W., Reynolds Jr, M.R. (1999). EWMA control charts for monitoring the mean of autocorrelated processes. *Journal of Quality Technology* 31:166-188.

Moustakides, G.V. (1986). Optimal stopping times for detecting changes in distributions. *The Annals of Statistics* 14:1379-1387.

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

Page, E.S. (1954). Continuous inspection schemes. *Biometrika* 41:100-114.

Reynolds, M.R., Jr, Kim, K. (2007). Multivariate Control Charts for Monitoring the Process Mean and Variability Using Sequential Sampling. *Sequential Analysis* 26:283-315.

Roberts, S.W. (1959). Control Chart Tests Based on Geometric Moving Averages. *Technometrics* 1:239-250.

Roy, S. (1953). On a heuristic method of test construction and ties use in multivariate analysis. *Annals of Mathematical Statistics* 24:220-238.

Ryan, T.P. (2000). *Statistical methods for quality improvement*. 2nd ed. New York: Wiley.

Schiöler, L. (2008). *Explorative analyzis of spatial patterns of influenza incidences in Sweden 1999-2008* (No. 2008:5): Statistical Research Unit, Department of Economics, Göteborg University, Sweden.

Shewhart, W.A. (1931). *Economic Control of Quality of Manufactured Product*. London: MacMillan and Co.

Shiryaev, A.N. (1963). On optimum methods in quickest detection problems. *Theory of Probability and its Applications*. 8:22-46.

Sonesson, C., Frisén, M. (2005a). Multivariate surveillance. In: Lawson, A., Kleinman, K., eds. *Spatial surveillance for public health*. New York: Wiley.

Sonesson, C., Frisén, M. (2005b). Multivariate surveillance. In: Lawson, A., Kleinman, K., eds. *Spatial surveillance for public health*. New York: Wiley.

Sun, K., Basu, A.P. (1995). A characterization of a bivariate geometric distribution. *Statistics & Probability Letters* 23:307-311.

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.

Vexler, A., Wu, C. (2009). An optimal retrospective change point detection policy. *Scandinavian Journal of Statistics*.

Wärmefjord, K. (2004). *Multivariate quality control and Diagnosis of Sources of Variation in Assembled Products.* Licentiat Thesis, Göteborg University, Göteborg.

Research Report

2007:9	Bock, D.	Evaluations of likelihood based surveillance of volatility.
2007:10	Bock, D. & Pettersson, K.	Explorative analysis of spatial aspects on the Swedish influenza data.
2007:11	Frisén, M. & Andersson, E.	Semiparametric surveillance of outbreaks.
2007:12	Frisén, M., Andersson, E. & Schiöler, L.	Robust outbreak surveillance of epidemics in Sweden.
2007:13	Frisén, M., Andersson, E. & Pettersson, K.	Semiparametric estimation of outbreak regression.
2007:14	Pettersson, K.	Unimodal regression in the two-parameter exponential family with constant or known dispersion parameter.
2007:15	Pettersson, K.	On curve estimation under order restrictions.
2008:1	Frisén, M.	Introduction to financial surveillance.
2008:2	Jonsson, R.	When does Heckman's two-step procedure for censored data work and when does it not?
2008:3	Andersson, E.	Hotelling´s T2 Method in Multivariate On-Line Surveillance. On the Delay of an Alarm.
2008:4	Schiöler, L. & Frisén, M.	On statistical surveillance of the performance of fund managers.
2008:5	Schiöler, L.	Explorative analysis of spatial patterns of influenza incidences in Sweden 1999—2008.
2008:6	Schiöler, L.	Aspects of Surveillance of Outbreaks.
2008:7	Andersson, E & Frisén, M.	Statistiska varningssystem för hälsorisker
2009:1	Frisén, M., Andersson, E. & Schiöler, L.	Evaluation of Multivariate Surveillance