Technometrics

A Journal of Statistics for the Physical, Chemical, and Engineering Sciences

Technometrics

ISSN: 0040-1706 (Print) 1537-2723 (Online) Journal homepage: https://amstat.tandfonline.com/loi/utch20

Phase I Distribution-Free Analysis of Multivariate Data

Giovanna Capizzi & Guido Masarotto

To cite this article: Giovanna Capizzi & Guido Masarotto (2017) Phase I Distribution-Free Analysis of Multivariate Data, Technometrics, 59:4, 484-495, DOI: 10.1080/00401706.2016.1272494

To link to this article: https://doi.org/10.1080/00401706.2016.1272494

TECHNOMETRICS
2017, VOL. 59, NO. 4, 484–495
http://dx.doi.org/10.1080/00401706.2016.1272494

Taylor & Francis
Taylor & Francis Group

Phase I Distribution-Free Analysis of Multivariate Data

Giovanna Capizzi and Guido Masarotto

Department of Statistical Sciences, University of Padua, Padua, Italy

ABSTRACT
In this study, a new distribution-free Phase I control chart for retrospectively monitoring multivariate data is developed. The suggested approach, based on the multivariate signed ranks, can be applied to individual or subgrouped data for detection of location shifts with an arbitrary pattern (e.g., isolated, transitory, sustained, progressive, etc.). The procedure is complemented with a LASSO-based post-signal diagnostic method for identification of the shifted variables. A simulation study shows that the method compares favorably with parametric control charts when the process is normally distributed, and largely outperforms other multivariate nonparametric control charts when the process distribution is skewed or heavy-tailed. An R package can be found in the supplementary material.

ARTICLE HISTORY
Received November 2015
Revised November 2016

KEYWORDS
Change-point detection; Control charts; LASSO; Multivariate signed ranks; Nonparametric methods; Statistical process control

1. Introduction

Statistical process control (SPC) consists of ideas and methods that are useful for maintaining a process in a stable and, hopefully, satisfactory state. There are generally two phases in the application of these methods. In Phase I, a fixed-size sample of time-ordered data is analyzed to assess the process stability. Specifically, Phase I control charts are statistical and graphical tools used to understand the nature of process variation over time. Using the knowledge of the process gathered during the first phase, a sequential scheme is then designed for the prospective monitoring of the incoming data (Phase II).

Distribution-free techniques for Phase I analysis have received increasing attention in recent literature. Indeed, the ability of parametric Phase I control charts to correctly distinguish between in-control (IC) and out-of-control (OC) observations is connected to the correct specification of the IC probability model. However, during Phase I, little information on IC distribution is available to practitioners. When distributional assumptions underlying a parametric control chart are not satisfied, or cannot be tested, the performance and sensitivity of parametric Phase I methods deteriorate. For example, the real false alarm probability (FAP), that is, the probability of declaring a stable process unstable, may be substantially larger than the desired value. Thus, several researchers (see, e.g., Chakraborti, Human, and Graham 2009; Jones-Farmer et al. 2014; Capizzi 2015) recommend verifying the form of the underlying IC distribution *only after* process stability has been established using a suitable distribution-free control chart.

Several distribution-free methods have been suggested for analyzing Phase I univariate data (e.g., Zou et al. 2007; Jones-Farmer, Jordan, and Champ 2009; Jones-Farmer and Champ 2010; Graham, Human, and Chakraborti 2010; Capizzi and Masarotto 2013; Zou et al. 2014; Capizzi 2015). When the joint distribution of multiple quality characteristics is unknown, some distribution-free control charts have also been proposed for Phase II monitoring (see, e.g., Liu 1995; Qiu and Hawkins 2001, 2003; Qiu 2008; Zou and Tsung 2011; Holland and Hawkins 2014; Li 2015; Chen, Zi, and Zou 2016; Liang, Xiang, and Pu 2016). These proposals can be properly modified for Phase I data. However, statistical methods for these two phases may benefit of being different because of practical and statistical peculiarities of prospective and retrospective monitoring (see Qiu 2013, p. 7). For example, the main aim of Phase II control charts consists in detecting a *single* change-point while Phase I data can be easily contaminated by *multiple* change-points. Unfortunately, although some multivariate distribution-free methods have been proposed for Phase II, "very little published research has considered the issue of robust, distribution-free, or nonparametric multivariate control charts for use in Phase I" (Jones-Farmer et al. 2014, p. 276).

The current state-of-the-art approach consists of the two multivariate Shewhart-type control charts recently proposed by Bell, Jones-Farmer, and Billor (2014) and Cheng and Shiau (2015) for the detection of location shifts for subgrouped data from elliptical distributions. The two control charts are based on the ranks of the Mahalanobis depths and the spatial signs, respectively. However, concerning the practical applicability and efficiency of these proposals, we would point out that: (a) Both control charts require subgrouped observations, notwithstanding the collection of individual data is increasingly common in many applications. (b) These procedures are not completely distribution-free. Indeed, they require to test in Phase I, *before establishing the stability of the process*, the assumption that the IC process distribution is elliptical. (c) The ranks of the Mahalanobis depths only depend on the magnitude of the distances of the observed points from the center of the data cloud. On the contrary, the spatial signs only reflect the directions of the vectors connecting the observed points to the center. It seems

CONTACT Giovanna Capizzi capizzi@stat.unipd.it

Supplementary materials for this article are available online. Please go to http://www.tandfonline.com/r/TECH.

intuitively reasonable that a more efficient scheme could be based on both the distances and directions. (d) Shewhart-type control charts offer a very good performance against isolated shifts. However, since they do not explicitly use the time-order of the data, they are not efficient against other types of shifts (transitory, sustained, progressive, oscillatory, etc.) that are often encountered in Phase I data.

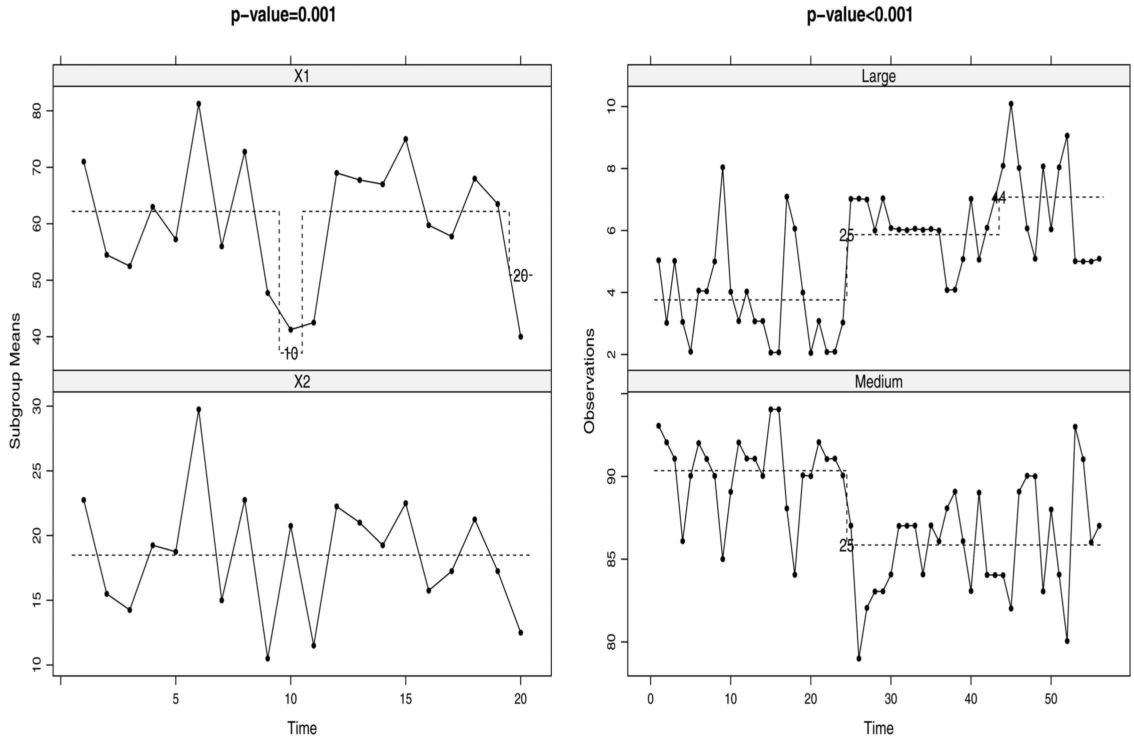
In an effort to overcome these drawbacks, we develop a new distribution-free Phase I control chart for both individual and subgrouped multivariate data. We avoid the need for any distributional assumption using a permutation approach (Pesarin 2001; Good 2005; Lehmann and Romano 2005). The new control chart is based on the multivariate signed ranks that optimally integrate the spatial signs and ranks of the Mahalanobis depths (Hallin and Paindaveine 2002, 2004, 2005). Following an idea developed in the univariate framework by Capizzi and Masarotto (2013), the chart combines different elementary control statistics designed for detecting the presence of one, two, or more, either isolated or step location shifts. As other types of shifts can be approximated using multiple step shifts, the proposed scheme offers good protection against a wide range of shift patterns. Furthermore, since in many practical situations shifts involve only a small number of variables, we complement the procedure with a LASSO-based method for identifying the variables that are likely to be responsible for an OC condition. An easy-to-use R package, available as supplementary material, allows practitioners to perform the proposed Phase I analysis.

Following Bell, Jones-Farmer, and Billor (2014) and Cheng and Shiau (2015), we consider the "standard framework" handled in multivariate statistical process monitoring. In particular, we assume that (i) the number of data points is larger than number of the variables, and (ii) the observation vectors are independent and identically distributed (iid) when the process is IC. Extensions to the high-dimensional and/or time-dependent framework will require further research. Some possible ideas are outlined in Section 5 (and also in Section S1 of the supplementary materials). Notwithstanding these limitations, we believe that the suggested method can be quite useful in many practical situations, and provide a viable and effective alternative to the procedures described by Bell, Jones-Farmer, and Billor (2014) and Cheng and Shiau (2015).

The article is organized as follows. In Section 2, we illustrate two practical applications of the proposed procedure. In Section 3, we describe the suggested Phase I control chart. In Section 4, we compare the new proposal with other methods. Concluding remarks are given in Section 5. Additional comments, performance evaluations, and examples are provided in the online supplementary material.

2. Examples

As a first example, we use the data given in Table 9.2 by Ryan (2011, p. 323). The sample comprises 20 subgroups, each with four observations, on two quality characteristics X_1 and X_2. According to Ryan (2011), the 10th and 20th subgroups are OC. Figure 1(a) illustrates the application of the proposed Phase I analysis. The p-value, shown in the center above the graphics, can be used to assess the stability over time of the process location. In particular, the procedure gives an alarm and the process is declared unstable if the p-value is less than α, where α denotes an acceptable value for the false alarm probability. As previously mentioned, the validity of the p-value does not require any assumption on the IC distribution. In this example, the observed p-value (0.001) is so small that an alarm is given for all



(a) Ryan data

(b) Gravel data

Figure 1. Phase 1 analysis of two datasets.

conventional levels of α, say 0.01, 0.05, and 0.1. Graphics can be used for diagnostic purposes. In particular, the solid lines in the two panels show the sample subgroup means of X_1 and X_2, respectively, and the dashed lines an estimate of the possibly time-varying process means. In this case, the dashed lines correctly suggest that the subgroups 10 and 20 could have a different location. Figure 1(a) points to the OC subgroups but also provides additional information. Indeed, the constant dashed line in the second panel indicates no shift for X_2 and, hence, the procedure suggests that only X_1 is OC. In addition, the estimated mean of X_1 offers information on the direction and magnitude of the shifts. In particular, it suggests that, for both subgroups, the shift was toward lower values and that the process was more severely OC for subgroup 10 than for subgroup 20. In Section 4.3, we show that this diagnostic information is usually quite accurate.

In the second example, we consider a dataset consisting of 56 individual observations from a European gravel-producing plant, which have already been considered by Holmes and Mergen (1993) and Sullivan and Woodall (2000). There are two variables measuring the percentage of the particles (by weight) that are large or medium in size, respectively. Figure 1(b) shows the output of the proposed Phase I procedure. The small p-value points to an unstable process, and, in particular, Figure 1(b) suggests the presence of two step-shifts: the first, at time 25, due to an increase in the percentage of large particles and a simultaneous decrease in the percentage of medium particles; the second, at time 44, due to an increase of the percentage of large particles not involving the other variable.

3. Detecting Location Changes in Multivariate Data

3.1 Overview

Assume that a sample of m time-ordered subgroups of size n, collected on g variables, is available, and denote with $\mathbf{x}_{i,j}$ the g-dimensional vector containing the jth observation from the ith subgroup. Individual observations, that is, process data without subgrouping, can be accommodated by setting $n = 1$. We also assume that (i) $mn > g$, that is, the number of the observed vectors is greater than the number of the variables; (ii) when the process is IC, $\mathbf{x}_{i,j}$ are iid with an unknown but common density function $p_0(\cdot)$; (iii) we are interested in the detection of location shifts, that is, when the process is OC, at least two subgroups have a different location.

As illustrated in the previous section, the proposed Phase I analysis provides a statistical test to verify if the process is in-control or out-of-control and some graphical aids able to suggest times and types of the shifts and responsible variables. It can be divided into four distinct stages.

1. *Preprocessing.* During the first stage, using suitable location and scatter estimates, data are standardized and transformed to the corresponding *multivariate signed ranks*, in the following denoted by $\mathbf{u}_{i,j}$. This transformation has some optimal properties (see Hallin and Paindaveine 2002, 2004, 2005), and we have verified that it improves the performance when the process distribution is heavy-tailed and/or skewed, without substantially affecting the test procedure in other cases.

   Essentially, the next three stages consist in fitting the following multivariate linear regression model

   $$\mathbf{u}_{i,j} = \boldsymbol{\beta}_{\text{common}} + \sum_{\tau=2}^{m-1} \boldsymbol{\beta}_{\text{step},\tau} I\big(i \geq \tau\big) + \sum_{\tau=1}^{m} \boldsymbol{\beta}_{\text{isolated},r} I\big(i = \tau\big) + (\text{residual})_{i,j} \qquad (1)$$

   where the $\boldsymbol{\beta}$'s are unknown g-dimensional vectors of parameters, and $I(C)$ is equal to one if condition C is true, and to zero otherwise. Observe that (i) $\boldsymbol{\beta}_{\text{common}}$ represents the "stable" level of the signed ranks; (ii) $\boldsymbol{\beta}_{\text{step},\tau}$ introduces a level change starting from time τ and affecting all the subsequent observations; (iii) $\boldsymbol{\beta}_{\text{isolated},\tau}$ affects only the observations at time τ and, hence, it corresponds to an isolated outlier. In the framework of model (1), checking process stability is equivalent to test the null hypothesis

   $$\mathrm{H}_0 : \text{all the } \boldsymbol{\beta}_{\text{step},\cdot} \text{ and } \boldsymbol{\beta}_{\text{isolated},\cdot} \text{ are zero.}$$

   Observe that in a "pure" distribution-free framework it is not possible to detect an isolated outlier when $n = 1$. Indeed, a single observation that is far from the others can be the consequence of an isolated shift. However, it can also be due to an extremely long-tailed IC distribution. Since discrimination between the two possibilities is not possible without further information on the shape of the IC distribution, only the presence of (multiple) step shifts is considered for individual observations.

   Note that model (1) is able to represent exactly any possible time-varying location patterns. Since m location vectors are represented using $2m - 1$ vectors of parameters, the model is clearly overparameterized. However, we expect the model to be sparse, that is, that most of the $\boldsymbol{\beta}$'s are zero. Hence, following what have been done for Phase II monitoring by Zou and Qiu (2009); Capizzi and Masarotto (2011); Jiang, Wang, and Tsung (2012); Liang, Xiang, and Pu (2016), the next three stages are based on a variable-selection approach. Observe that the idea of using variable-selection methods for retrospectively detecting outliers and step changes is not new (see Harchaoui and Lévy-Leduc 2010; She and Owen 2011; Ciuperca 2014; Zou, Tseng, and Wang 2014, for some recent examples). However, we believe that our proposal is distinct from earlier works because (i) it addresses the multivariate Phase I framework; (ii) it tries to simultaneously detect multiple isolated and step changes; (iii) it emphasizes the testing phase (see below) and the control of the FAP.
2. *Screening.* During the second stage, we use the popular forward search (FS) algorithm to select, between the $2m - 2$ parameter vectors $\boldsymbol{\beta}_{\text{step},\cdot}$ and $\boldsymbol{\beta}_{\text{isolated},\cdot}$, $K < m$ vectors, which can be viewed as "promising" shifts suggested by the data. Here, K denotes the maximum number of shifts we want to search for. When no a priori information is available, we suggest using $K =$

min(50, integer closest to $\sqrt{m}$), which offers good performances in a variety of OC scenarios (see also Ing and Lai 2011).

We choose the FS algorithm since (i) it is simple, fast, and well-known to practitioners; (ii) it offers a very good *screening* performance, that is, FS can identify all the relevant predictors, together with few irrelevant ones, even if the predictor dimension is larger than the sample size (see Wang 2009; Ing and Lai 2011); (iii) as shown by Capizzi and Masarotto (2015), Phase II FS-based control charts are competitive with schemes based on other variable-selection approaches like LASSO, LAR, etc.

3. *Testing.* During the second stage, K elementary test statistics are computed for detecting the presence of 1, …, K either isolated or step shifts. Then, in the third stage, these statistics are aggregated and a single p-value is computed.
4. *Post-signal diagnostic.* The FS algorithm tends to select the relevant predictors together with some unneeded shifts. Hence, following a suggestion by Wang (2009), when the hypothesis of a stable process is rejected, we *prune* the model using the adaptive LASSO algorithm (Zou 2006) and the information criterion proposed by Chen and Chen (2008). During this stage, the g variables enter into the model independently, that is, we use the adaptive LASSO also for identifying the subset of variables involved in each shift.

3.2 Stage 1: Data Standardization and Computation of the Multivariate Signed Ranks

This stage consists of the following two steps.

1. Compute suitable estimates of the multivariate location vector and dispersion (scatter) matrix. Since an investigation of the properties of our Phase I method, based on different estimates of location and scatter, is beyond the aim of this article, we here consider only the following estimates:
   - *Location.* We use the transformation-retransformation spatial median (see Oja and Randles 2004; Oja 2010) of the subgroup means, that is,

     $$\boldsymbol{\ell} = \mathbf{S}^{1/2}(\text{spatial median of } \mathbf{S}^{-1/2}\overline{\mathbf{x}}_1, \ldots, \mathbf{S}^{-1/2}\overline{\mathbf{x}}_m),$$

     where

     $$\overline{\mathbf{x}}_i = \frac{1}{n}\sum_{j=1}^{n}\mathbf{x}_{i,j} \qquad (2)$$

     and $\mathbf{S}^{1/2}$ is any square-root of the scatter matrix $\mathbf{S}$ given by Equation (3), that is, $\mathbf{S}^{1/2}$ is a $g \times g$ matrix, so that $\mathbf{S} = \mathbf{S}^{1/2}(\mathbf{S}^{1/2})'$. This location estimate is extremely robust, and can be computed using the fast and simple algorithm given by Oja (2010, p. 71).
   - *Dispersion.* The estimate is different in the case of individual ($n = 1$) or subgrouped data ($n > 1$). In particular, we use

     $$\mathbf{S} = \begin{cases} \dfrac{1}{2(m-1)}\displaystyle\sum_{i=2}^{m}(\mathbf{x}_{i,1} - \mathbf{x}_{i-1,1})(\mathbf{x}_{i,1} - \mathbf{x}_{i-1,1})' & \text{if } n = 1; \\ \dfrac{1}{m(n-1)}\displaystyle\sum_{i=1}^{m}\sum_{j=1}^{n}(\mathbf{x}_{i,j} - \overline{\mathbf{x}}_i)(\mathbf{x}_{i,j} - \overline{\mathbf{x}}_i)' & \text{if } n > 1. \end{cases} \qquad (3)$$

     Both estimators have been studied in the SPC literature. See Williams et al. (2007) and Paynabar, Qiu, and Zou (2015) for individual data, and Montgomery (2009), Ryan (2011), and Qiu (2013) for subgrouped data. In particular, note that $\mathbf{S}$ tends to be resistant against location shifts. In the following, we will assume that $\mathbf{S}$ is nonsingular. Otherwise, the monitoring can be restricted to a set of linearly independent variables.
2. Using $\boldsymbol{\ell}$ and $\mathbf{S}$, standardize the observed data obtaining $\mathbf{z}_{i,j} = \mathbf{S}^{-1/2}(\mathbf{x}_{i,j} - \boldsymbol{\ell})$ and then compute the multivariate signed ranks of the standardized data as

   $$\mathbf{u}_{i,j} = \begin{cases} \mathbf{0} & \text{if } \mathbf{z}_{i,j} = \mathbf{0}; \\ \dfrac{\sqrt{F^{-1}_{\chi^2_g}\left(\dfrac{r_{i,j}}{1+mn}\right)}}{||\mathbf{z}_{i,j}||}\mathbf{z}_{i,j} & \text{if } \mathbf{z}_{i,j} \neq \mathbf{0}, \end{cases} \qquad (4)$$

   where $||\mathbf{v}|| = \sqrt{\mathbf{v}'\mathbf{v}}$ denotes the Euclidean norm of $\mathbf{v}$, $r_{i,j}$ is the rank of $||\mathbf{z}_{i,j}||$ among $||\mathbf{z}_{1,1}||, \ldots, ||\mathbf{z}_{m.n}||$, and $F_{\chi^2_g}(\cdot)$ is the cumulative function of a χ^2 random variable with g degrees of freedom. Observe that the multivariate signed ranks, are g-dimensional vectors with the same directions as $\mathbf{z}$'s. However, they are scaled so that the norms $||\mathbf{u}||$'s are those expected for a Gaussian IC process.

3.3 Stage 2: Screening and Computation of the Elementary Tests Statistics

In this stage, a model with an increasing number of parameters (either isolated or step shifts) is fitted to the signed ranks using an FS algorithm. In particular, the fitted values at the kth step are $\hat{\mathbf{u}}_i^{(k)} = \widehat{\boldsymbol{\beta}}_0^{(k)} + \widehat{\boldsymbol{\beta}}_1^{(k)}\xi_i^{(1)} + \cdots + \widehat{\boldsymbol{\beta}}_k^{(k)}\xi_i^{(k)} \quad (i = 1, \ldots, m)$ where $\widehat{\boldsymbol{\beta}}_r^{(k)}$, $r = 0, \ldots, k$, are g-dimensional vectors of parameters and $\xi_i^{(k)}$ is a scalar sequence corresponding either to an isolated or a step shift, that is, $\xi_i^{(k)} = I(i = \tau^{(k)})$ or $\xi_i^{(k)} = I(i \geq \tau^{(k)})$, for some $\tau^{(k)}$. The type (isolated or step) and time ($\tau^{(k)}$) of the shift $\xi_i^{(k)}$ introduced at step k as well as the parameters $\widehat{\boldsymbol{\beta}}_r^{(k)}$, $r = 1, \ldots, k$, are determined by minimizing the residual sum of squares $\sum_{i=1}^{m}\sum_{j=1}^{n}||\mathbf{u}_{i,j} - \hat{\mathbf{u}}_i^{(k)}||^2$ conditionally to the shifts (type and time) identified during the previous $k-1$ steps. After every step, we compute the explained variance

$$T_k = n\sum_{i=1}^{m}||\hat{\mathbf{u}}_i^{(k)}||^2 - mn||\overline{\hat{\mathbf{u}}}||^2,$$

where $\bar{\bar{\mathbf{u}}} = \sum_{i=1}^{m}\sum_{j=1}^{n}\mathbf{u}_{i,j}/mn$. Since we are dealing with a linear model, the FS implementation is straightforward. In particular, in our R package, we use a Gram–Schmidt approach that is optimized by the exclusive presence of dummy regressors in the model.

3.4 Stage 3: Testing

Following the approach proposed for Phase II monitoring by Zou and Qiu (2009) and Capizzi and Masarotto (2011), the idea consists of aggregating the K elementary statistics T_k in the overall test statistic

$$W_{\mathrm{OBS}} = \max_{k=1,\ldots,K} \frac{T_k - \mathrm{E}_0(T_k)}{\sqrt{\mathrm{var}_0(T_k)}}$$

and then computing the p-value as $p\text{-value} = \mathrm{Prob}_0(W > W_{\mathrm{OBS}})$. Here, W denotes the random variable underlying W_{OBS}, and Prob_0, E_0 and var_0 are computed under the IC hypothesis.

The direct application of this idea is unfeasible, since the IC probability distribution of $T_1, \ldots, T_K$ depends on $p_0(\cdot)$, the IC density function of the Phase I data. However, we can use a permutation approach (Pesarin 2001; Good 2005; Lehmann and Romano 2005). In particular, organize the Phase I multivariate observations in the matrix $\mathbf{Y} = (\mathbf{x}_{1,1}, \ldots, \mathbf{x}_{1,n}, \mathbf{x}_{2,1}, \ldots, \mathbf{x}_{2,n}, \ldots, \mathbf{x}_{m,1}, \ldots, \mathbf{x}_{m,n})$, and let $\mathbf{P}_{\mathbf{Y}}$ be the set of all the matrices obtainable by permuting the columns of $\mathbf{Y}$. If we assume that, when the process is IC, the observations are iid, the IC density of $\mathbf{Y}$, that is, $\prod_{i=1}^{m}\prod_{j=1}^{n} p_0(\mathbf{x}_{i,j})$, is constant over $\mathbf{P}_{\mathbf{Y}}$. Hence, the IC conditional distribution of $\mathbf{Y}$ given $\mathbf{P}_{\mathbf{Y}}$ does not depend on the unknown density function $p_0(\cdot)$.

The results suggest to compute the p-value *conditionally* to $\mathbf{P}_{\mathbf{Y}}$. However, the exact computation of the conditional p-value is possible only if m and n are very small, say $m \cdot n < 10$. Therefore, we suggest proceeding as follows:

(a) Compute the statistics $T_1, \ldots, T_K$ for L randomly generated (column) permutations of $\mathbf{Y}$. Let $T^{\star}_{l,k}$ be the value of the kth statistic obtained in the lth replication, $l = 1, \ldots, L$, $k = 1, \ldots, K$.

(b) Compute the p-value as

$$p\text{-value} = \frac{1}{L}\sum_{l=1}^{L} \mathrm{I}\left(\max_{k=1,\ldots,K}\frac{T^{\star}_{l,k} - a_k}{b_k} > \max_{k=1,\ldots,K}\frac{T_k - a_k}{b_k}\right),$$

where $a_k = \sum_{l=1}^{L} T^{\star}_{l,k}/L$ and $b_k^2 = \sum_{l=1}^{L}(T^{\star}_{l,k} - a_k)^2/(L-1)$.

We found that using $L = 1000$ Monte Carlo replications provides sufficient accuracy. An alternative approach based on the assumption that the IC distribution is elliptical is discussed in the supplementary material.

3.5 Stage 4: Post-Signal Diagnostic

In this subsection, we discuss how the process mean can be estimated. When $p\text{-value} \geq \alpha$, with α equal to a desired FAP, the hypothesis that the process is IC is accepted. Therefore, the common mean of each subgroup can be estimated by the overall sample mean

$$\bar{\bar{\mathbf{x}}} = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\mathbf{x}_{i,j}. \tag{5}$$

On the contrary, when $p\text{-value} < \alpha$, the test suggests that the process location is not stable. Hence, it is important, for diagnostic purposes, to identify the times of the location changes as well as the involved variables. With this aim in mind, substitute, in the model obtained at the last step of the FS outlined in Section 3.3, the parameter vectors $\boldsymbol{\beta}_k$ with $\mathbf{S}^{-1/2}\boldsymbol{\delta}_k$, $k = 0, \ldots, K$, obtaining the regression model

$$\mathbf{u}_{i,j} = \mathbf{S}^{-1/2}\boldsymbol{\delta}_0 + \mathbf{S}^{-1/2}\boldsymbol{\delta}_1\xi_i^{(1)} + \cdots + \mathbf{S}^{-1/2}\boldsymbol{\delta}_K\xi_i^{(K)} + (\text{residual})_{i,j}. \tag{6}$$

In Equation (6), $\boldsymbol{\delta}_0$ is the intercept term while $\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_K$ represent the directions of the potential location shifts in the observed data. Indeed, since $\mathbf{u}_{i,j} \propto \mathbf{S}^{-1/2}(\mathbf{x}_{i,j} - \boldsymbol{\ell})$, a shift of $\boldsymbol{\delta}$ in $\mathbf{x}_{i,j}$ results in a shift with direction $\mathbf{S}^{-1/2}\boldsymbol{\delta}$ in the corresponding signed rank $\mathbf{u}_{i,j}$ (see Figure 2). An estimate of these parameters can be obtained by minimizing the sum of squares

$$s^2(\boldsymbol{\delta}_0, \ldots, \boldsymbol{\delta}_K) = \sum_{i=1}^{m}\sum_{j=1}^{n} \|\mathbf{u}_{i,j} - \mathbf{S}^{-1/2}\boldsymbol{\delta}_0 - \sum_{k=1}^{K}\mathbf{S}^{-1/2}\boldsymbol{\delta}_k\xi_i^{(K)}\|^2.$$

However, we expect that part, if not most, of the elements of $\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_K$ are zeros since (i) K can be larger than the true number of change-points, and (ii) only a subset of the variables can be involved in a shift. Therefore, we suggest fitting model (6) using the adaptive LASSO method, that is, to estimate $\boldsymbol{\delta}_0, \ldots, \boldsymbol{\delta}_K$ minimizing the penalized sum of squares

$$s^2(\boldsymbol{\delta}_0, \ldots, \boldsymbol{\delta}_K) + \lambda\sum_{k=1}^{K}\sum_{h=1}^{g}\left|\frac{\delta_{k,h}}{\hat{\delta}^{ls}_{k,h}}\right|, \tag{7}$$

where $\delta_{k,h}$ is the hth element of $\boldsymbol{\delta}_k$ and $\hat{\delta}^{ls}_{k,h}$ is its estimate obtained using the unpenalized least-square method. See Zou (2006) for the definition and motivation of the adaptive LASSO; Tibshirani (1996) for the definition of the original, nonadaptive LASSO; and Zou, Jiang, and Tsung (2011) for an application of a similar idea in Phase II post-signal diagnosis.

In (7), λ is a positive tuning parameter. The unpenalized least-square estimate is obtained when $\lambda = 0$. As λ increases, the estimates are progressively shrunk toward zero, with some (or many) of the parameters becoming exactly zero. The LARS algorithm developed by Efron et al. (2004) can be used to compute the estimates for every $\lambda \geq 0$, with a computational cost equal to that necessary to compute the unpenalized least-square estimate. In selecting λ, we obtain good results using the extended BIC criterion proposed by Chen and Chen (2008):

$$\mathrm{EBIC}_{\gamma}(\lambda) = mng\log\left(\frac{s^2(\hat{\boldsymbol{\delta}}_0(\lambda), \ldots, \hat{\boldsymbol{\delta}}_K(\lambda))}{mng}\right) + \nu(\lambda)\log(mng) + 2\gamma\log\binom{2gm - g}{\nu(\lambda)},$$

where $\hat{\boldsymbol{\delta}}_k(\lambda)$ denotes the estimate of $\boldsymbol{\delta}_k$ obtained by minimizing (7), and $\nu(\lambda)$ is the number of nonzero elements in

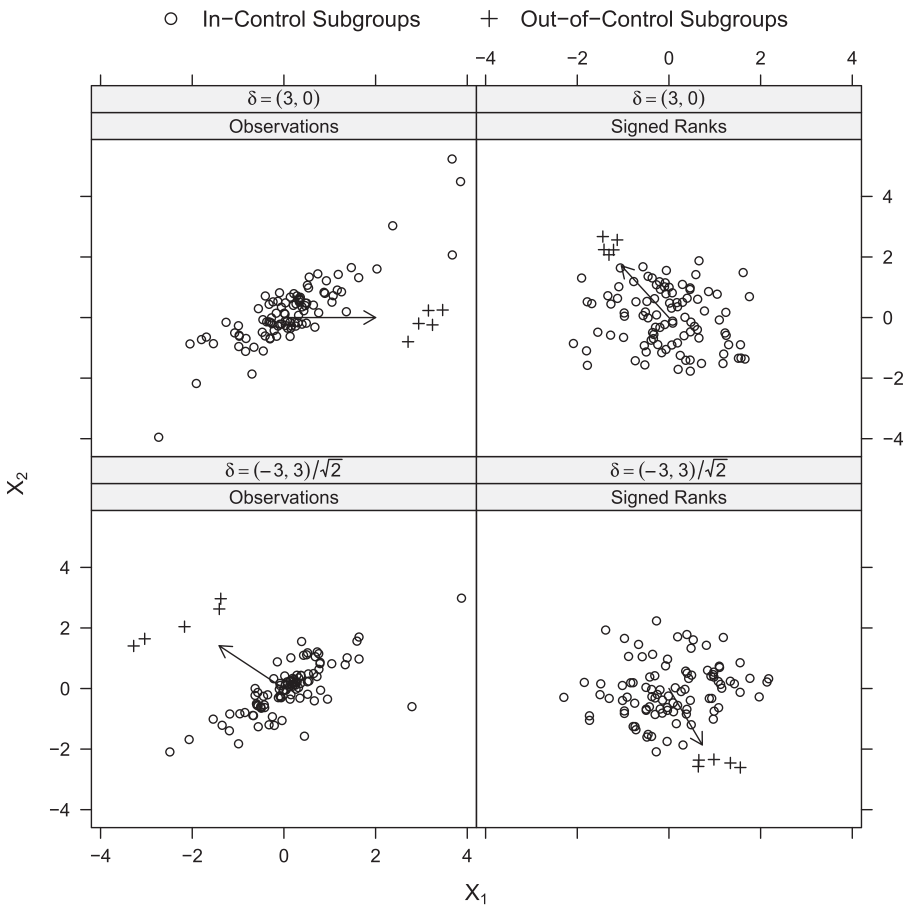


Figure 2. Two simulated datasets ($m = 20$, $n = 5$) and their signed ranks. The IC data have been simulated from a bivariate Student's t distribution with 3 degrees of freedom, so that $\mathrm{E}(X_1) = \mathrm{E}(X_2) = 0$, $\mathrm{var}(X_1) = \mathrm{var}(X_2) = 1$, and $\mathrm{cor}(X_1, X_2) = 0.8$. Each dataset contains one OC subgroup, obtained adding the shift $\boldsymbol{\delta}$. The arrows show the direction of (i) $\boldsymbol{\delta}$ in the left panels; (ii) $\mathbf{S}^{-1/2}\boldsymbol{\delta}$ in the right panels.

$\hat{\boldsymbol{\delta}}_0(\lambda), \ldots, \hat{\boldsymbol{\delta}}_K(\lambda)$. Observe that $2gm - g$ is the dimension of the searched parameter space. Indeed, in addition to $\boldsymbol{\delta}_0$, we have $2m - 2$ possible choices for $\xi_i^{(k)}$, each introducing g parameters in the model. As illustrated in Section 4.3 and in the supplementary material, γ can be used to balance the probability to detect real changes with that of signaling false changes. Typical values are $\gamma = 0$, corresponding to the standard BIC criterion, and $\gamma = 0.5$ or $\gamma = 1$, which offer better protection against false signals.

Notwithstanding that all computation has been performed using the signed ranks, it seems reasonable to present the results to users on the scale of the original, untransformed observations (see, e.g., Figure 1). Therefore, the last step of our procedure consists of refitting the model selected using the adaptive LASSO to the $\mathbf{x}$'s. Numerical details are discussed in the supplementary material.

4. Simulation Study

4.1 Study Design

In this section, we summarize the results of an extensive simulation study. In particular, in Section 4.2, we compare the IC and OC alarm probabilities of the Phase I control charts described in Table 1. Then, in Section 4.3, we study the performance of the post-signal diagnostic method based on the adaptive LASSO. Additional performance results are presented in the supplementary material. All the presented performance measures have been estimated using 10,000 Monte Carlo replications.

The multivariate distributions considered in the study are presented in Table 2. Observe that Normal and Student belong to the family of the multivariate elliptical distributions, Student having heavier tails than Normal. In contrast, Gamma and Poisson are not elliptical. In particular, Poisson is a discrete distribution introduced to show that the suggested method also attains the desired FAP when the distribution is not continuous.

For studying the OC performance of the considered Phase I methods, we assume that the observations are generated by $\mathbf{x}_{i,j} = \boldsymbol{\delta}\xi_i + \boldsymbol{\epsilon}_{i,j}$, where $\boldsymbol{\delta}$ is a g-dimensional vector giving the direction and size of the shift, ξ_i a scalar sequence describing the dynamic pattern and $\boldsymbol{\epsilon}_{i,j}$ are iid drawn from the distributions described in Table 2. The considered patterns, which represent some of the real patterns encountered in Phase I, are summarized in Table 3. Observe that the onset and duration of the shifts are stochastic. Indeed, in a real application, they are unknown and vary from case to case.

4.2 In-Control and Out-of-Control Alarm Probabilities

In this section, we compare the proposed test procedure with that of other Phase I control charts. As performance metrics, we use the FAP, that is, the probability of declaring unstable an IC

Table 1. Six parametric and nonparametric Phase I control charts (CL = control limit).

Chart	Description and references	Applicability
MPhase1	The procedure described in this article ($K =$ (integer closest to $\sqrt{m}$), $L = 1000$, $l\text{min} = 5$).	Individual$^{(*)}$ and subgrouped observations from *any* distribution.
T^2	The Hotelling's T^2 chart that signals if, for some i, $n(\bar{\mathbf{x}}_i - \bar{\bar{\mathbf{x}}})'\mathbf{S}^{-1}(\bar{\mathbf{x}}_i - \bar{\bar{\mathbf{x}}}) > \text{CL}$ (Montgomery 2009; Ryan 2011; Qiu 2013). See (2), (5), and (3) for the definition of $\bar{\mathbf{x}}_i$, $\bar{\bar{\mathbf{x}}}$, and $\mathbf{S}$. We also consider the version with $\bar{\bar{\mathbf{x}}}$ and $\mathbf{S}$ replaced by the minimum covariance determinant (MCD) estimates.	Individual and subgrouped *normally* distributed observations.
GLR	The scheme signals if $\text{GLR} = \max_{\tau=2,\dots,m} \text{lr}_\tau > \text{CL}$ where lr_τ is the likelihood ratio test statistic for verifying, assuming a normal distribution, that the mean of $\mathbf{x}_{i,j}$, $i < \tau$, is equal to the mean of $\mathbf{x}_{i,j}$, $i \geq \tau$ (Zamba and Hawkins 2006; Chen and Gupta 2011; Qiu 2013).	Individual$^{(**)}$ and subgrouped *normally* distributed observations.
LLC	The procedure, proposed by Lung-Yut-Fong, Lévy-Leduc, and Cappé (2011), is a GLR-type control chart based on component-wise ranks.	Individual$^{(*)}$ observations from *any* distribution.
Depth Ranks	The scheme, proposed by Bell, Jones-Farmer, and Billor (2014), signals if, for some i, $\sum_{j=1}^{n}(nm + 1 - \text{MDR}_{i,j})/n > \text{CL}$ where $\text{MDR}_{i,j}$ denote the ranks of the Mahalanobis depth $1/[1 + (\mathbf{x}_{i,j} - \boldsymbol{\ell}_{\text{BACON}})'\mathbf{S}^{-1}(\mathbf{x}_{i,j} - \boldsymbol{\ell}_{\text{BACON}})]$. Here, $\boldsymbol{\ell}_{\text{BACON}}$ is the center of the data cloud computed using the BACON algorithm (Billor, Hadi, and Velleman 2000) and $\mathbf{S}$ the scatter estimate given in Equation (3).	Subgrouped observations from *elliptical* distributions.
Spatial Signs	The scheme, proposed by Cheng and Shiau (2015), signals if, for some i, $n\|\| \sum_{j=1}^{n} \mathbf{v}_{i,j}/n\|\|^2 > \text{CL}$ where $\mathbf{v}_{i,j} = \mathbf{S}_{\text{HR}}^{-1/2}(\mathbf{x}_{i,j} - \boldsymbol{\ell}_{\text{HR}})/\|\|\mathbf{S}_{\text{HR}}^{-1/2}(\mathbf{x}_{i,j} - \boldsymbol{\ell}_{\text{HR}}\|\|$. Here, $\boldsymbol{\ell}_{\text{HR}}$ and $\mathbf{S}_{\text{HR}}$ denote the Hettmansperger–Randles location and scatter estimates, respectively (Hettmansperger and Randles 2002).	Subgrouped observations from *elliptical* distributions.

NOTES: $^{(*)}$ *MPhase1* and *LLC* are not effective against isolated shifts in individual observations (see Section 3.1).
$^{(**)}$ *GLR* has a low power against isolated shifts in individual observations.

Table 2. Four multivariate distributions.

Distribution	Definition	Properties
	Elliptical distributions	
Normal	$\mathbf{x}_{i,j} \sim N_g(\mathbf{0}, \boldsymbol{\Sigma})$.	(i) The univariate marginal distributions are normal; (ii) $\text{E}(\mathbf{x}_{i,j}) = \mathbf{0}$ and $\text{var}(\mathbf{x}_{i,j}) = \boldsymbol{\Sigma}$.
Student	$\mathbf{x}_{i,j} = \widetilde{\mathbf{x}}_{i,j}/\sqrt{w_{i,j}^2/3}$ with $\widetilde{\mathbf{x}}_{i,j,r} \sim N_g(\mathbf{0}, \boldsymbol{\Sigma})$ and $w_{i,j}^2 \sim \chi_3^2$.	(i) The univariate marginal distributions are Student's *t* with three degrees of freedom; (ii) $E(\mathbf{x}_{i,j}) = \mathbf{0}$ and $\text{var}(\mathbf{x}_{i,j}) = 2\boldsymbol{\Sigma}$.
	Nonelliptical distributions	
Gamma	$\mathbf{x}_{\mathbf{i},\mathbf{j}} = \text{diag} \sum_{r=1}^{4} \widetilde{\mathbf{x}}_{i,j,r}\widetilde{\mathbf{x}}'_{i,j,r}/2$ where $\widetilde{\mathbf{x}}_{i,j,r} \sim N_g(\mathbf{0}, \boldsymbol{\Sigma})$ and $\text{diag}(\mathbf{A})$ is the vector of the diagonal entries of matrix $\mathbf{A}$ (see Stoumbous and Sullivan 2002).	(i) The univariate marginal distributions are gamma with shape parameter equal to two; (ii) $E(\mathbf{x}_{i,j}) = 2\text{diag}(\boldsymbol{\Sigma})$ and $\text{var}(\mathbf{x}_{i,j}) = 2\boldsymbol{\Sigma}^{(2)}$, where $\boldsymbol{\Sigma}^{(2)}$ is the matrix obtained by squaring the individual entries of the matrix $\boldsymbol{\Sigma}$.
Poisson	$\mathbf{x}_{i,j} = (r_{i,j,0} + r_{i,j,1}, \dots, r_{i,j,0} + r_{i,j,g})'$ where $r_{i,j,r}$ are iid. Poisson random variables with mean θ if $r = 0$ and $1 - \theta$ otherwise ($0 \leq \theta \leq 1$).	(i) The univariate marginal distributions are Poisson with mean one; (ii) The correlation between two elements of $\mathbf{x}_{i,j}$ is θ.

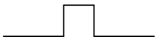
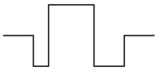
process (first type error), and the detection power, that is, the probability to signal as OC an unstable process. All the considered Phase I methods, with the only exception of *LLC*, are invariant with respect to affine transformations. Hence, if the distribution is elliptical, (i) the IC performance does not depend on $\text{E}(\mathbf{x}_{i,j})$ and $\text{var}(\mathbf{x}_{i,j})$; (ii) the OC alarm probabilities depend only on the noncentrality parameter $\sqrt{\boldsymbol{\delta}'\text{var}^{-1}(\mathbf{x}_{i,j})\boldsymbol{\delta}}$. This is not exactly true for nonelliptical distributions, and, in general, for the LLC chart. However, for all the considered cases, we observe comparable results using different shift directions and covariance matrices. Hence, we will assume that (i) only the first variable shifts, that is, only the first element of $\boldsymbol{\delta}$ is different from zero; and (ii) $\text{var}(\mathbf{x}_{i,j})$ is a matrix with all diagonal elements equal to one and off-diagonal elements equal to 0.6.

Figure 3 shows the real FAPs attained by the proposed method for the IC distributions given in Table 2 when an FAP = 0.05 is desired. Since we obtain comparable results for different numbers of variables, only the case of five variables ($g = 5$) is considered in Figure 3. The figure clearly shows that the suggested Phase I method attains the desired FAP for each sample size and all the considered multivariate distributions.

Figure 3 also presents the real FAPs of two versions of the T^2 control charts based on (i) the classical location and scatter estimates $\bar{\bar{\mathbf{x}}}$ and $\mathbf{S}$, and (ii) the highly robust MCD estimators. The results illustrate the disadvantages of using a Phase I control chart whose distributional assumptions are not satisfied. In particular, observe that the T^2 schemes can be used only under their design condition, that is, if the IC distribution is normal. Indeed, in many nonnormal cases considered in the simulation, the real FAPs are unacceptably higher, sometimes even close to one, than the nominal value. It is also interesting to observe that using a robust estimators like MCD can even worsen the problem.

Figures 4 and 5 compare the OC performance of the suggested method with those of its competitors in the case of normally distributed data. Since we obtain similar results using different numbers of subgroups (or individual observations), only the results obtained for $m = 50$ are reported. Observe that our proposal offers basically the same protection as the Hotelling's T^2 control chart against isolated shifts in subgrouped data, but better protection against the other patterns (both for individual and subgrouped data). The opposite happens with GLR: our

Table 3. Five shift patterns$^{(*)}$.

Pattern	Description
Isolated	A single, isolated shift: ξ_i equal to 1 if $i = \tau$ and to 0 otherwise with τ = integer part of $U(1, m+1)$.
Sustained	A step shift starting at a random instant of time: ξ_i equal to 0 if $i < \tau$, and to $1/\sqrt{n}$ if $i \geq \tau$ with $\tau \sim U(m-20, m-5)$.
Transient 	A transient shift with random onset and duration: ξ_i equal to $1/\sqrt{n}$ if $\tau_1 \leq i \leq \tau_2$ and to 0 otherwise with $\tau_1 \sim U(5, m-15)$ and $\tau_2 - \tau_1 \sim U(5, 12)$.
Mixture 	A process with three operational states with Markovian switching: ξ_i is a Markov chain starting from $\xi_0 = 0$, assuming the values $-1/\sqrt{n}$, 0, and $1/\sqrt{n}$ and transition matrix $P = [p_{r,s}]$ such that $p_{r,s}$ is equal to 0.8 if $r = s$ and to 0.1 otherwise.
Linear	A linear shift starting at a random point: ξ_i equal to zero if $i < \tau$ and to $(i+1-\tau)/\sqrt{n}(m+1-\tau)$ otherwise with $\tau \sim U(m-20, m-5)$.

Note: * $U(f_1, f_2)$ denotes a uniform random variable between f_1 and f_2.

method outperforms this chart in the case of isolated shifts, and globally offers at least a comparable performance for the other patterns. Hence, the suggested procedure can be considered a reasonable alternative to the two normal-based control charts (T^2 and GLR) offering a satisfactory performance for a wide range of shift patterns, even when the distribution is actually normal.

Regarding the comparison with the nonparametric competitors:

- Figure 4 shows that the suggested Phase I method performs considerably better than the two schemes that are based on the Mahalanobis depth ranks and the spatial signs, in the case of subgrouped normally distributed data. We obtain similar results with nonnormal distributions. For example, Figure 6 displays the alarm probabilities in the case of a multivariate Student's t distribution. In this case, the chart based on the depth ranks is never competitive with our scheme, while the chart based on the spatial signs is competitive only in detecting isolated shift when g is equal to 5 or 10, but is substantially inferior in all the other scenarios. Therefore, when compared to these recently proposed Shewhart-type control charts, our procedure not only has a wider applicability (it can also be used for individual observations and/or nonelliptical distributions) but also offers considerably better protection in the situation (subgrouped data from an elliptical distribution) for which the control charts based on the depth ranks and spatial signs have been originally designed.
- In the case of individual normally distributed data, Figure 5 shows that our method outperforms the LLC procedure (Lung-Yut-Fong, Lévy-Leduc, and Cappé 2011), based on the component-wise ranks. Similar results have been obtained when the distribution is not normal (see the supplementary material).

4.3 Efficiency of the Post-Signal Procedure

To illustrate the performance of the post-signal diagnostic method described in Section 3.5, we consider observations $\mathbf{x}_{i,j} = (x_{i,j,1}, \ldots, x_{i,j,g})'$ generated by the model $x_{i,j,r} = \mu_{i,r} + \epsilon_{i,j,r}$, where $\mu_{i,r}$ denotes the mean at time i of the rth variable and $\boldsymbol{\epsilon}_{i,j} = (\epsilon_{i,j,1}, \ldots, \epsilon_{i,j,g})'$ are iid g-variate Student's t random variables with three degrees of freedom, such that $\mathrm{E}(\epsilon_{i,j,r}) = 0$, $\mathrm{var}(\epsilon_{i,j,r}) = 1$, and $\mathrm{cor}(\epsilon_{i,j,r}, \epsilon_{i,j,h}) = 0.6$. We also assume that

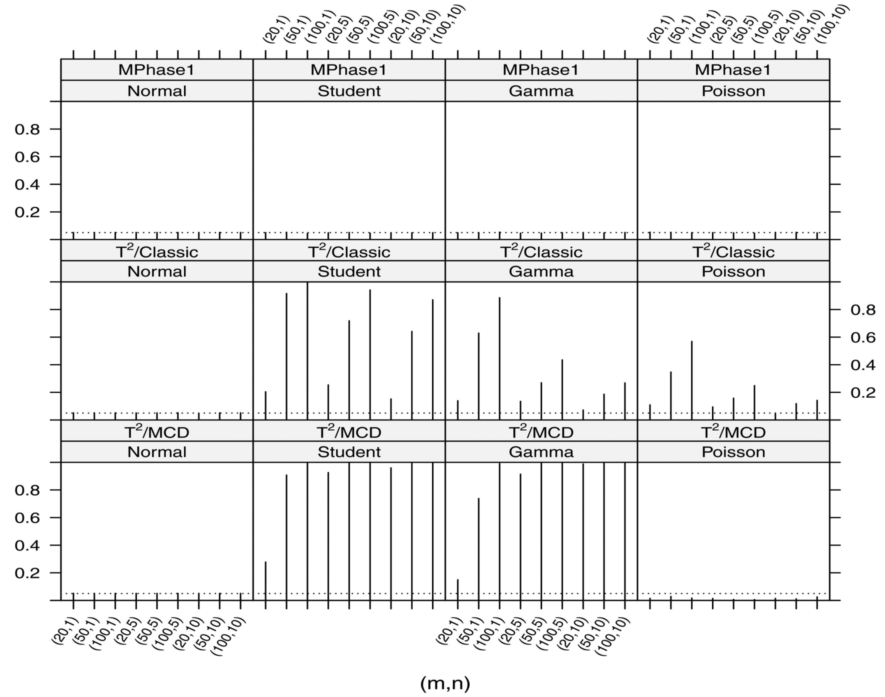


Figure 3. Attained false alarm probability for different distributions, different number of subgroups (m), and different subgroup sizes (n) when the number of variables (g) is five. The nominal false alarm probability $\alpha = 0.05$ is shown by the dotted lines.

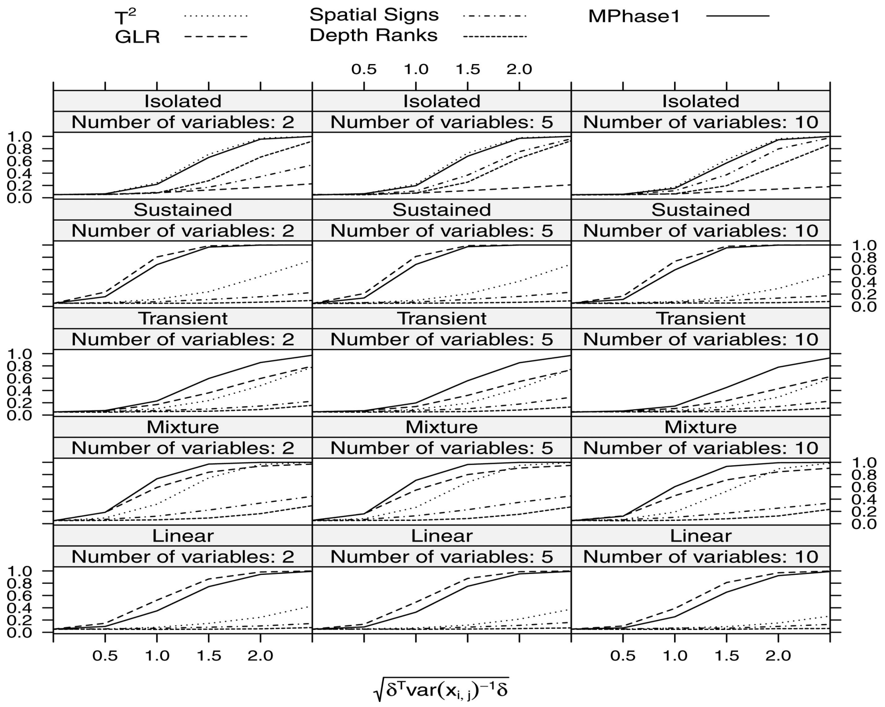


Figure 4. Out-of-control alarm probability of Phase I control charts for subgrouped normal observations ($m = 50, n = 5$).

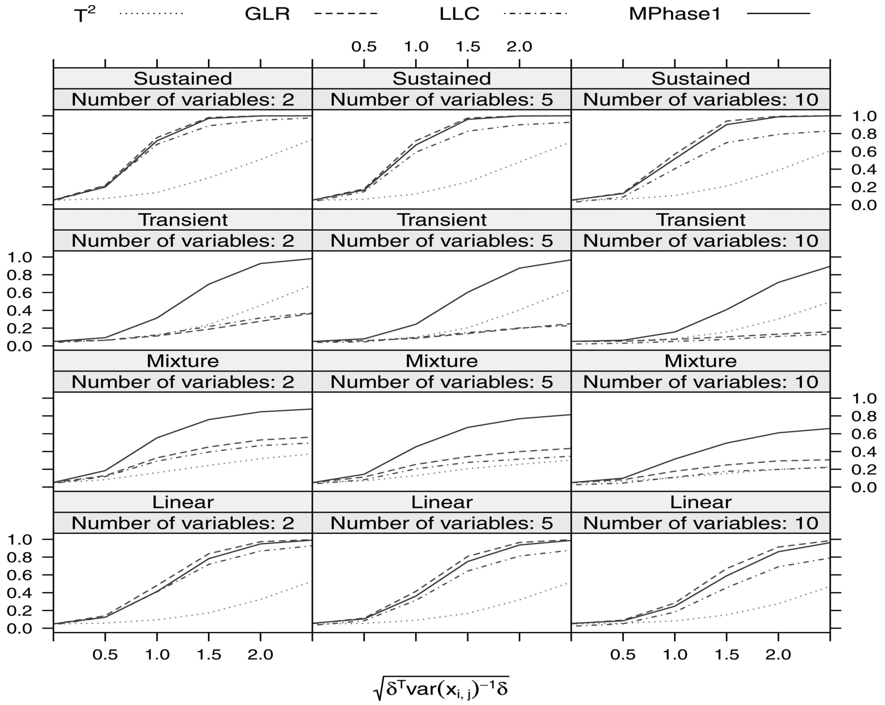


Figure 5. Out-of-control alarm probability of Phase I control charts for individual normal observations ($m = 50, n = 1$).

the Phase I sample is given by $m = 50$ subgroups, each with $n = 5$ observations, and that the number of variables g is either 5 or 10. The following two scenarios for $\mu_{i,r}$ are considered:

$$\text{A. } \mu_{i,r} = \begin{cases} \delta & \text{if } i = 11 \text{ and } r = 1 \\ \dfrac{\delta}{\sqrt{5}} & \text{if } i \geq 31 \text{ and } r = 2 \\ -\dfrac{\delta}{\sqrt{5}} & \text{if } i \geq 31 \text{ and } r = 3 \\ 0 & \text{otherwise} \end{cases};$$

$$\text{B. } \mu_{i,r} = \begin{cases} \dfrac{\delta}{\sqrt{5}} & \begin{array}{l}\text{if } 1 \leq i \leq 10 \text{ or } 21 \leq i \leq 30 \\ \text{or } 41 \leq i \leq 50 \\ \text{and } r = 1 \text{ or } r = 2\end{array} \\ -\dfrac{\delta}{\sqrt{5}} & \begin{array}{l}\text{if } 11 \leq i \leq 20 \text{ or } 31 \leq i \leq 40 \\ \text{and } r = 1 \text{ or } r = 2\end{array} \\ \delta & \text{if } i = 25 \text{ and } r = 3 \text{ or } r = 4 \\ 0 & \text{otherwise} \end{cases}.$$

In the first case, the process experiences an isolated shift in the first variable and a simultaneous step shift in the second and

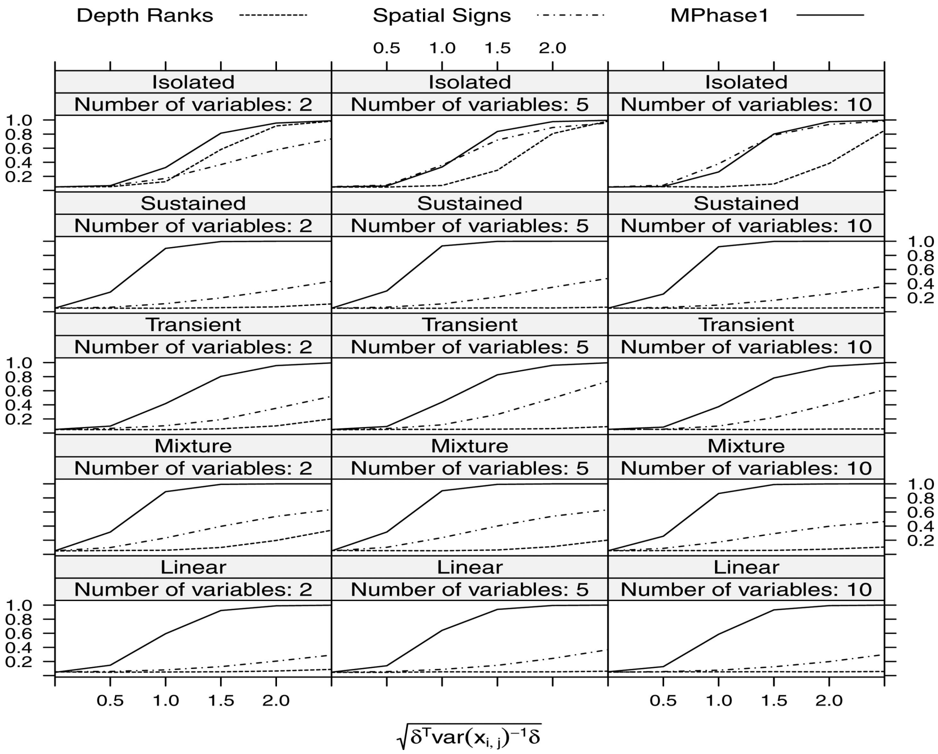


Figure 6. Out-of-control alarm probability of Phase I control charts for subgrouped Student's *t* observations (3 degrees of freedom, $m = 50$, $n = 5$).

third variables. This means that the total number of true mean shifts to detect is three. For the second scenario, the means of the first two variables alternate between two values; the third and fourth variables experience an isolated shift. The total number of true mean shifts to detect is 10 (four step shifts for each of the first two variables; one isolated shift for the third and fourth variables).

Table 4 shows averages (and standard deviations) of the numbers of true and false shifts detected by the diagnostic method based on the adaptive LASSO. Results are shown for δ equal to 1 or 2, and for γ, the tuning parameter of the information criterion $\mathrm{EBIC}_\gamma(\lambda)$, equal to 0, 0.5, and 1. We consider a step shift as approximately detected (third and fourth columns of Table 4) if the procedure signals a step shift in the involved variables and $|(\text{signaled time}) - (\text{true onset})| \leq 5$, that is, for scenario A, the step shift in the second variable is considered exactly detected if it is signaled at $i = 31$, and approximately detected if signaled at $26 \leq i \leq 36$. Table 4 also shows the average number of false shifts detected when the process is IC, that is, when $\delta = 0$.

The suggested approach shows a satisfactory performance since it consistently detects almost all the medium/large shifts ($\delta = 2$), and most of the small/medium shifts ($\delta = 1$) even, when $\gamma = 0.5$ and $\gamma = 1$, maintaining an acceptable number of false detections. Observe that the number of false detections

Table 4. Averages, and, in parentheses, standard deviations, of the number of true or false shifts detected by the post-signal diagnostic method based on the adaptive LASSO ($\alpha = 0.05$).

	Number of true detections						
	Exact		Approximated		Number of false detections		
	$\delta = 1$	$\delta = 2$	$\delta = 1$	$\delta = 2$	$\delta = 0$	$\delta = 1$	$\delta = 2$
	Scenario A / Number of variables: 5 / Number of true shifts: 3						
$\gamma = 0$	2.39 (0.87)	2.98 (0.17)	2.64 (0.62)	2.99 (0.09)	0.17 (1.25)	2.15 (2.58)	2.10 (2.72)
$\gamma = 0.5$	2.12 (0.89)	2.97 (0.22)	2.37 (0.68)	2.98 (0.17)	0.06 (0.47)	0.53 (1.42)	0.52 (1.24)
$\gamma = 1$	1.84 (0.77)	2.95 (0.30)	2.07 (0.63)	2.97 (0.26)	0.05 (0.27)	0.20 (0.74)	0.20 (0.74)
	Scenario A / Number of variables: 10 / Number of true shifts: 3						
$\gamma = 0$	2.33 (0.88)	2.99 (0.17)	2.60 (0.64)	3.00 (0.10)	0.17 (1.04)	2.34 (2.08)	2.11 (2.07)
$\gamma = 0.5$	2.12 (0.90)	2.98 (0.22)	2.36 (0.71)	2.99 (0.16)	0.06 (0.48)	0.51 (1.27)	0.45 (1.21)
$\gamma = 1$	1.84 (0.82)	2.96 (0.30)	2.08 (0.70)	2.97 (0.25)	0.05 (0.30)	0.18 (0.75)	0.14 (0.77)
	Scenario B / Number of variables: 5 / Number of true shifts: 10						
$\gamma = 0$	7.39 (2.49)	9.68 (0.82)	9.02 (2.27)	9.98 (0.31)	0.16 (1.14)	1.85 (2.19)	1.41 (1.32)
$\gamma = 0.5$	6.31 (3.07)	9.66 (0.88)	7.70 (3.32)	9.97 (0.37)	0.06 (0.46)	0.63 (0.67)	0.66 (0.54)
$\gamma = 1$	4.09 (2.00)	9.62 (0.98)	4.88 (2.16)	9.92 (0.64)	0.05 (0.28)	0.12 (0.12)	0.28 (0.24)
	Scenario B / Number of variables: 10 / Number of true shifts: 10						
$\gamma = 0$	7.05 (2.54)	9.70 (0.77)	8.79 (2.43)	9.99 (0.23)	0.17 (1.64)	2.19 (2.17)	1.32 (1.85)
$\gamma = 0.5$	5.98 (2.95)	9.68 (0.81)	7.35 (3.20)	9.97 (0.35)	0.06 (0.58)	0.67 (1.01)	0.54 (1.23)
$\gamma = 1$	3.77 (1.65)	9.64 (0.94)	4.56 (1.77)	9.93 (0.57)	0.05 (0.32)	0.12 (0.43)	0.24 (0.85)

does not essentially depend on either the number of true shifts or the number of variables. As expected, increasing γ results in a decrease of the number of false detections, but also in a reduced ability to detect small/medium shifts. Therefore, γ should be chosen balancing the importance of detecting small/medium shifts with the cost of investigating a larger number of false shifts. In general, our results show that $\gamma = 0.5$ provides a reasonable compromise. Alternatively, when there is an alarm, it is possible to repeat the post-signal procedure with different values of γ. A practitioner can be highly confident about the shifts signaled by $\text{EBIC}_1(\lambda)$, while additional shifts selected by $\text{EBIC}_{0.5}(\lambda)$ and, above all, $\text{EBIC}_0(\lambda)$ should be investigated more carefully.

5. Conclusions

We have suggested a new distribution-free approach for Phase I analysis of multivariate data, which shows wide applicability, and offers a satisfactory performance in a broad class of OC scenarios. The suggested method combines a statistical test, based on the multivariate signed ranks, for verifying the hypothesis that the location of the process is stable, with a LASSO-based post-signal diagnostic procedure for identifying the timing of the shifts and the involved variables. The application of the suggested Phase I procedure is straightforward using the R package available in the supplementary material.

Given the good results, it seems worthwhile to investigate the extension of the proposed approach to the monitoring of the dispersion and dependence structure of multivariate observations. A natural possibility consists of combining the ideas developed in this article with the test based on the multivariate signed ranks for the homogeneity of scatter, as studied by Hallin and Paindaveine (2008). It also seems interesting to consider the extension to high-dimensional data, that is, to situations in which either $g \approx nm$ or $g > mn$. In these cases, it is not possible to estimate the entire covariance matrix. However, a procedure similar to our proposal can be based on suitable tests for high-dimensional data (see, for some examples, Chen and Qin 2010; Ro et al. 2015; Feng, Zou, and Wang 2016). Further, it seems interesting to investigate the use of alternative estimates of the multivariate location and scatter parameters. In addition, as discussed in remarks C and D in Section S1 of the supplementary material, it seems worthwhile to explore the possibility of handling autocorrelated data using some bootstrap for time series techniques and the substitution of the two-stage model identification approach based on the FS and LASSO algorithms with alternative variable selection algorithms.

Supplementary Materials

mphase1-supplementary.pdf:	Additional simulation results and computational details.
mphase1-example.pdf:	A vignette illustrating the use of the R package.
mphase1-package.zip:	An R package implementing the Phase I method described in the article. The archive includes the source package, a version compiled for MS Windows and the manual.

Acknowledgments

The authors thank the editor, the associate editor, and the referees for their timely review and many helpful comments that greatly improved the article. This research was partially funded by UNIPD CPDA128413/12 grant.

ORCID

Giovanna Capizzi http://orcid.org/0000-0002-3187-1365
Guido Masarotto http://orcid.org/0000-0003-4697-1606

References

Bell, R. C., Jones-Farmer, L. A., and Billor, N. (2014), "A Distribution-Free Multivariate Phase I Location Control Chart for Subgrouped Data from Elliptical Distributions," *Technometrics*, 56, 528–538. [484,485,490]

Billor, N., Hadi, A. S., and Velleman, P. F. (2000), "BACON: Blocked Adaptive Computationally Efficient Outlier Nominators," *Computational Statistics & Data Analysis*, 34, 279–298. [490]

Capizzi, G. (2015), "Recent Advances in Process Monitoring: Nonparametric and Variable-Selection Methods for Phase I and Phase II" (with discussion), *Quality Engineering*, 27, 44–80. [484]

Capizzi, G., and Masarotto, G. (2011), "A Least Angle Regression Control Chart for Multidimensional Data," *Technometrics*, 53, 285–296. [488]

——— (2013), "Phase I Distribution-Free Analysis of Univariate Data," *Journal of Quality Technology*, 45, 273–284. [484,485]

——— (2015), "Comparison of Phase II Control Charts Based on Variable Selection Methods," in *Frontiers in Statistical Quality Control* (Vol. 11), eds. S. Knoth, and W. Schmid, New York: Springer, pp. 151–162. [487]

Chakraborti, S., Human, S., and Graham, M. (2009), "Phase I Statistical Process Control Charts: an Overview and Some Results," *Quality Engineering*, 21, 52–62. [484]

Chen, J., and Chen, Z. (2008), "Extended Bayesian Information Criteria for Model Selection with Large Model Spaces," *Biometrika*, 95, 759–771. [488]

Chen, J., and Gupta, A. K. (2011), *Parametric Statistical Change Point Analysis: With Applications in Genetics, Medicine, and Finance* (2nd ed.), New York: Birkhäuser. [490]

Chen, N., Zi, X., and Zou, C. (2016), "A Distribution-Free Multivariate Control Chart," *Technometrics*, 58, 448–459. [484]

Chen, S. X., and Qin, Y. L. (2010), "A Two-Sample Test for High-Dimensional Data With Applications to Gene-Set Testing," *Annals of Statistics*, 38, 808–835. [494]

Cheng, C.-R., and Shiau, J.-J. H. (2015), "A Distribution-Free Multivariate Control Chart for Phase I Applications," *Quality and Reliability Engineering International*, 31, 97–111. [484,485,490]

Ciuperca, G. (2014), "Model Selection by LASSO Methods in a Change-Point Model," *Statistical Papers*, 55, 349–374. [486]

Efron, E., Hastie, T., Johnstone, I., and Tibshirani, R. (2004), "Least Angle Regression," *The Annals of Statistics*, 32, 407–499. [488]

Feng, L., Zou, C., and Wang, Z. (2016), "Multivariate-Sign-Based High-Dimensional Tests for the Two-Sample Location Problem," *Journal of American Statistical Association*, 111, 721–735. [494]

Good, P. (2005), *Permutation, Parametric and Bootstrap Tests of Hypotheses* (3rd ed.), New York: Springer. [485,488]

Graham, M. A., Human, S. W., and Chakraborti, S. (2010), "A Phase I Nonparametric Shewhart-Type Control Chart Based on the Median," *Journal of Applied Statistics*, 37, 1795–1813. [484]

Hallin, M., and Paindaveine, D. (2002), "Optimal Tests for Multivariate Location Based on Interdirections and Pseudo-Mahalanobis Ranks," *The Annals of Statistics*, 30, 1103–1133. [485]

——— (2004), "Multivariate Signed-Rank Tests in Vector Autoregressive Order Identification," *Statistical Science*, 19, 697–711. [485]

——— (2005), "Affine-Invariant Aligned Rank Tests for the Multivariate General Linear Model with VARMA Errors," *Journal of Multivariate Analysis*, 93, 122–163. [485]

——— (2008), "Optimal Rank-Based Tests for Homogeneity of Scatter," *The Annals of Statistics*, 36, 1261–1298. [494]

Harchaoui, Z., and Lévy-Leduc, C. (2010), "Multiple Change-Point Estimation With a Total Variation Penalty," *Journal of the American Statistical Association*, 105, 1480–1493. [486]

Hettmansperger, T. P., and Randles, R. H. (2002), "A Practical Affine Equivariant Multivariate Median," *Biometrika*, 89, 851–860. [490]

Holland, M. D., and Hawkins, D. M. (2014), "A Control Chart Based on a Nonparametric Multivariate Change-Point Model," *Journal of Quality Technology*, 46, 63–77. [484]

Holmes, D. S., and Mergen, A. (1993), "Improving the Performance of the T^2 Control Chart," *Quality Engineering*, 5, 619–625. [486]

Ing, C.-K., and Lai, T. L. (2011), "A Stepwise Regression Method and Consistent Model Selection for High-Dimensional Sparse Linear Models," *Statistica Sinica*, 21, 1473–1513. [487]

Jiang, W., Wang, K., and Tsung, F. (2012), "A Variable-Selection-Based Multivariate EWMA Chart for Process Monitoring and Diagnosis," *Journal of Quality Technology*, 44, 209–230. [486]

Jones-Farmer, L. A., and Champ, C. W. (2010), "A Distribution-Free Phase I Control Chart for Subgroup Scale," *Journal of Quality Technology*, 42, 373–387. [484]

Jones-Farmer, L. A., Jordan, V., and Champ, C. W. (2009), "Distribution-Free Phase I Control Charts for Subgroup Location," *Journal of Quality Technology*, 41, 304–316. [484]

Jones-Farmer, L. A., Woodall, W. H., Steiner, S. H., and Champ, C. W. (2014), "An Overview of Phase I Analysis for Process Improvement and Monitoring," *Journal of Quality Technology*, 46, 265–280. [484]

Lehmann, E. L., and Romano, J. P. (2005), *Testing Statistical Hypotheses* (3rd ed.), New York: Springer. [485,488]

Li, J. (2015), "Nonparametric Multivariate Statistical Process Control Charts: A Hypothesis Testing-Based Approach," *Journal of Nonparametric Statistics*, 27, 384–400. [484]

Liang, W., Xiang, D., and Pu, X. (2016), "A Robust Multivariate EWMA Control Chart for Detecting Sparse Mean Shifts," *Journal of Quality Technology*, 48, 265–283. [484]

Liu, R. Y. (1995), "Control Charts for Multivariate Processes," *Journal of the American Statistical Association*, 90, 1380–1387. [484]

Lung-Yut-Fong, A., Lévy-Leduc, C., and Cappé, O. (2011), "Homogeneity and Change-Point Detection Tests for Multivariate Data Using Rank Statistics," arXiv preprint arXiv:1107.1971. [490]

Montgomery, D. C. (2009), *Introduction to Statistical Quality Control* (6th ed.), New York: Wiley. [490]

Oja, H. (2010), *Multivariate Nonparametric Methods with R. An Approach Based on Spatial Signs and Ranks*, New York: Springer. [487]

Oja, H., and Randles, R. H. (2004), "Multivariate Nonparametric Tests," *Statistical Science*, 19, 598–605. [487]

Paynabar, K., Qiu, P., and Zou, C. (2015), "A Change Point Approach for Phase I Analysis in Multivariate Profile Monitoring and Diagnosis," *Technometrics*, 58, 191–204. [487]

Pesarin, F. (2001), *Multivariate Permutation Tests : With Applications in Biostatistic*, New York: Wiley. [485,488]

Qiu, P. (2008), "Distribution-Free Multivariate Process Control Based on Log-Linear Modeling," *IIE Transactions*, 40, 664–677. [484]

——— (2013), *Introduction to Statistical Process Control*, Boca Raton, FL: Chapman & Hall/CRC Press. [484,490]

Qiu, P., and Hawkins, D. M. (2001), "A Rank-Based Multivariate CUSUM Procedure," *Technometrics*, 43, 120–132. [484]

——— (2003), "A Nonparametric Multivariate CUSUM Procedure for Detecting Shifts in All Directions," *Journal of the Royal Statistical Society*, Series D, 52, 151–164. [484]

Ro, K., Zou, C., Wang, Z., and Yin, G. (2015), "Outlier Detection for High-Dimensional Data," *Biometrika*, 102, 589–599. [494]

Ryan, T. P. (2011), *Statistical Methods for Quality Improvement* (3rd ed.), New York: Wiley. [485,490]

She, Y., and Owen, A. B. (2011), "Outlier Detection Using Nonconvex Penalized Regression," *Journal of the American Statistical Association*, 106, 626–639. [486]

Stoumbous, Z. G., and Sullivan, J. H. (2002), "Robustness to Non-normality of the Multivariate EWMA Control Chart," *Journal of Quality Technology*, 34, 260–276. [490]

Sullivan, J. H., and Woodall, W. H. (2000), "Change-Point Detection of Mean Vector or Covariance Matrix Shifts using Multivariate Individual Observations," *IIE Transactions*, 32, 537–549. [486]

Tibshirani, R. J. (1996), "Regression Shrinkage and Selection via the LASSO," *Journal of the Royal Statistical Society*, Series B, 58, 267–288. [488]

Wang, H. (2009), "Forward Regression for Ultra-High Dimensional Variable Screening," *Journal of the American Statistical Association*, 104, 1512–1524. [487]

Williams, W., Birch, J. B., Woodall, W. H., and Ferry, N. M. (2007), "Statistical Monitoring of Heteroscedastic Dose-Response Profiles From High-Throughput Screening," *Journal of Agricultural, Biological, and Environmental Statistics*, 12, 216–235. [487]

Zamba, K. D., and Hawkins, D. M. (2006), "A Multivariate Change Point Model for Statistical Process Control," *Technometrics*, 48, 539–549. [490]

Zou, C., Jiang, W., and Tsung, F. (2011), "A LASSO-Based Diagnostic Framework for Multivariate Statistical Process Control," *Technometrics*, 53, 297–309. [488]

Zou, C., Liu, Y., Qin, P., and Wang, Z. (2007), "Empirical Likelihood Ratio Test for the Change-Point Problem," *Statistics & Probability Letters*, 77, 374–382. [484]

Zou, C., and Qiu, P. (2009), "Multivariate Statistical Process Control Using LASSO," *Journal of American Statistical Association*, 104, 1586–1596. [488]

Zou, C., Tseng, S.-T., and Wang, Z. (2014), "Outlier Detection in General Profiles using Penalized Regression Method," *IIE Transactions*, 46, 106–117. [486]

Zou, C., and Tsung, F. (2011), "A Multivariate Sign EWMA Control Chart," *Technometrics*, 53, 84–97. [484]

Zou, C., Yin, G., Feng, L., and Wang, Z. (2014), "Nonparametric Maximum Likelihood Approach to Multiple Change-Point Problems," *The Annals of Statistics*, 42, 970–1002. [484]

Zou, H. (2006), "The Adaptive Lasso and Its Oracle Properties," *Journal of American Statistical Association*, 101, 1418–1429. [488]

Supplementary Material to "Phase I Distribution-Free Analysis of Multivariate Data"

Additional Details and Comparisons

Giovanna Capizzi*
Department of Statistical Sciences, University of Padua, Italy
and
Guido Masarotto
Department of Statistical Sciences, University of Padua, Italy

December 5, 2016

Contents

*This research was partially funded by the UNIPD CPDA128413/12 grant.

S1 Some Remarks on the Proposed Phase I Method

A. In our implementation, it is possible to constrain the difference between the onset of two consecutive step shifts (see Subsection 3.3 of the paper) to be greater than an user-controllable constant, say `lmin`. In particular, we found that the choice `lmin` $= 5$ leads to a better performance against several types of location shifts without substantially affecting the possibility of detecting short step shifts or transitory OC periods. Capizzi and Masarotto (2013) obtained similar results in the univariate case.

B. As mentioned in Subsection 3.1 of the paper, the presence of isolated shifts is not considered for individual observations. When $n = 1$, we suggest to plot $G_i = (\mathbf{x}_{i,1} - \boldsymbol{\ell})'\mathbf{S}^{-1}(\mathbf{x}_{i,1} - \boldsymbol{\ell})$ against i, and to further investigate any data point with a value far from the others for a possible OC situation. Unfortunately, it is not possible to provide nonparametric control limits for G_i.

C. The proposed Phase I analysis relies on the assumption that the observations are i.i.d. when the process is IC. Therefore, we recommend complementing the analysis by checking the data for the presence of autocorrelation (see, e.g., Wei, 2006). If this hypothesis cannot be rejected, one possibility consists of computing the p-value replacing the conditional approach, described in Subsection 3.4 of the paper, with some bootstrap methods for time-series (Bühlmann, 2002; Politis, 2015). However, further research work is needed on this topic.

D. The use of the FS algorithm for *screening*, i.e., for identifying a superset of the relevant predictors, followed by the adaptive LASSO for *pruning* have been considered in a slightly different framework by Wang (2009). We decide to consider this approach because of its simplicity and very good performance, but, also because the use of the FS algorithm leads to a very fast computation of the permutation-based p-value. However, observe that our method can also be implemented using alternative variable-selection approaches, including fused-type regularizations (Tibshirani et al., 2005) and bi-level algorithms (Huang et al., 2012). This can be a topic for further research.

E. In the post-signal diagnostic method described in Subsection 3.5 of the paper, it is possible to directly use the observations in place of their signed ranks. However, in a simulation study not reported here, we observed that this substitution leads to a reduced ability to detect the variables responsible for the shifts when the distribution is heavy-tailed and/or skewed.

F. Concerning the multivariate signal ranks, observe that

1. They combine the information on the direction of $\mathbf{z}_{i,j}$, given by its spatial sign $\mathbf{z}_{i,j}/||\mathbf{z}_{i,j}||$, with the information on the magnitude of $||\mathbf{z}_{i,j}||$ given by $r_{i,j}$.

2. They are invariant with respect to an affine transformation of the data, i.e., the signed ranks computed from $\mathbf{x}_{i,j}$ are the same as those obtained from $\mathbf{a} + \mathbf{B}\mathbf{x}_{i,j}$, where $\mathbf{a}$ and $\mathbf{B}$ are a g-dimensional vector and a $g \times g$ non-singular matrix, respectively. This property ensures that results do not depend on the particular scale and coordinate system chosen for the data.

3. Their values are influenced by the choice of $\mathbf{S}^{-1/2}$. However, it can easily be shown that the results of the Phase I procedure, e.g. the p-value, do not depend on this choice. In our implementation, we use the Cholesky factor of $\mathbf{S}$.

4. Differently from the *spatial* signed ranks introduced by Oja (2010), the multivariate signed ranks given in equation (4) of the paper do not require the strong assumption on symmetry of the underlying distribution.

G. Possible alternatives to multivariate signed ranks are provided by *spatial ranks* and *spatial signs* which have been extensively used in the literature regarding multivariate nonparametric tests (Oja, 2010). We have performed a simulation study, with a design similar to that reported in Section 4 of the paper, to investigate the impact of choosing these alternative scores. Due to lack of space, we cannot reproduce all the results but, here, we suggest using the signed ranks for the following reasons:

1. Concerning the comparisons between signed and spatial ranks, our simulation results show that both types of multivariate ranks offer a similar performance in

terms of OC alarm probability. However, the complexity of the computation of signed and spatial ranks are on the order of $gnm \log(nm)$ and $g(nm)^2$, respectively. Hence, our proposal is much faster when based on signed ranks, in particular when nm is not small. Further, it is not easy to understand the relationship between the direction of a shift in the original data, and in the corresponding spatial ranks. Therefore, it is more difficult to use spatial ranks in the LASSO-based post-signal diagnostic method described in Subsection 3.5 of the paper.

2. As shown in Section S6 of this supplementary material, the proposed Phase I method based on multivariate signed ranks outperforms an analogous procedure based on spatial signs.

S2 Additional Details on Stage 4

When the null hypothesis of a stable process is rejected, the model given by equation (6) in the paper can be written in matrix form as

$$\underset{gnm\times 1}{\mathbf{U}} = \underset{gnm\times g(K+1)}{\mathbf{W}} \; \underset{g(K+1)\times 1}{\boldsymbol{\theta}} + \underset{gnm\times 1}{\boldsymbol{\epsilon}} \tag{1}$$

where

$$\mathbf{U} = \begin{pmatrix} \mathbf{u}_{1,1} \\ \vdots \\ \mathbf{u}_{1,n} \\ \vdots \\ \mathbf{u}_{m,1} \\ \vdots \\ \mathbf{u}_{m,n} \end{pmatrix}, \quad \mathbf{W} = \begin{pmatrix} \mathbf{S}^{-1/2} & \mathbf{S}^{-1/2}\xi_1^{(1)} & \cdots & \mathbf{S}^{-1/2}\xi_1^{(K)} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{S}^{-1/2} & \mathbf{S}^{-1/2}\xi_1^{(1)} & \cdots & \mathbf{S}^{-1/2}\xi_1^{(K)} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{S}^{-1/2} & \mathbf{S}^{-1/2}\xi_m^{(1)} & \cdots & \mathbf{S}^{-1/2}\xi_m^{(K)} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{S}^{-1/2} & \mathbf{S}^{-1/2}\xi_m^{(1)} & \cdots & \mathbf{S}^{-1/2}\xi_m^{(K)} \end{pmatrix}, \quad \boldsymbol{\theta} = \begin{pmatrix} \boldsymbol{\delta}_0 \\ \vdots \\ \boldsymbol{\delta}_K \end{pmatrix}$$

and $\boldsymbol{\epsilon}$ is the residual vector.

Model (1) is a standard linear regression model. So, the entire path of the adaptive LASSO can be obtained using the LARS algorithm (Efron et al., 2004) as modified by Zou (2006).

Suppose now that, using the information criterion of Chen and Chen (2008), H elements of $\boldsymbol{\theta}$ are selected as different from zero. Let $\mathbf{W}_{\mathbf{H}}$ be the corresponding $gnm \times H$ submatrix of $\mathbf{W}$. Observe that the selected elements always contain the elements of $\boldsymbol{\delta}_0$ since the intercept is not penalized. To obtain the fitted values, shown for diagnostic purpose as dashed lines in the output of our procedure, we proceed as follows:

1. Regress, using an ordinary least squares approach,

   $$\mathbf{Z} = \left(\mathbf{z}'_{1,1}, \ldots, \mathbf{z}'_{1,n}, \ldots, \mathbf{z}'_{m,1}, \ldots, \mathbf{z}'_{m,n}\right)'$$

   on $\mathbf{W}_H$ and denote by

   $$\widehat{\mathbf{Z}} = (\widehat{\mathbf{z}}'_1, \ldots, \widehat{\mathbf{z}}'_1, \ldots, \widehat{\mathbf{z}}'_m, \ldots, \widehat{\mathbf{z}}'_m)'$$

   the corresponding fitted values. Here, the $\mathbf{z}_{i,j}$'s are the standardized observations described in Subsection 3.2 of the paper. Further, observe that, given the structure of $\mathbf{W}_H$, the fitted values are constant throughout each subgroup.

2. Compute the fitted value for the ith subgroup as

   $$\widehat{\mathbf{x}}_i = \boldsymbol{\ell} + \mathbf{S}^{1/2}\widehat{\mathbf{z}}_{i,1}$$

Remark 1. Some computational saving can be achieved observing that the same fitted values can be obtained applying a similar procedure to the subgroup means of the signed ranks and of the standardized observations.

Remark 2. We regress the standardized observations $\mathbf{z}_{i,j}$, and not the original data $\mathbf{x}_{i,j}$, on $\mathbf{W}_H$, since we are in a Seemingly Unrelated Regression Equations (SURE) framework. Hence, better results can be obtained if heteroscedasticity and correlations between variables are taken into account (see Green, 2012, chapter 10).

S3 Additional Performance Comparisons

Figures 1 and 2 compare, for two non-normal distributions, the OC performance of the suggested Phase I procedure with that of the method based on component-wise ranks

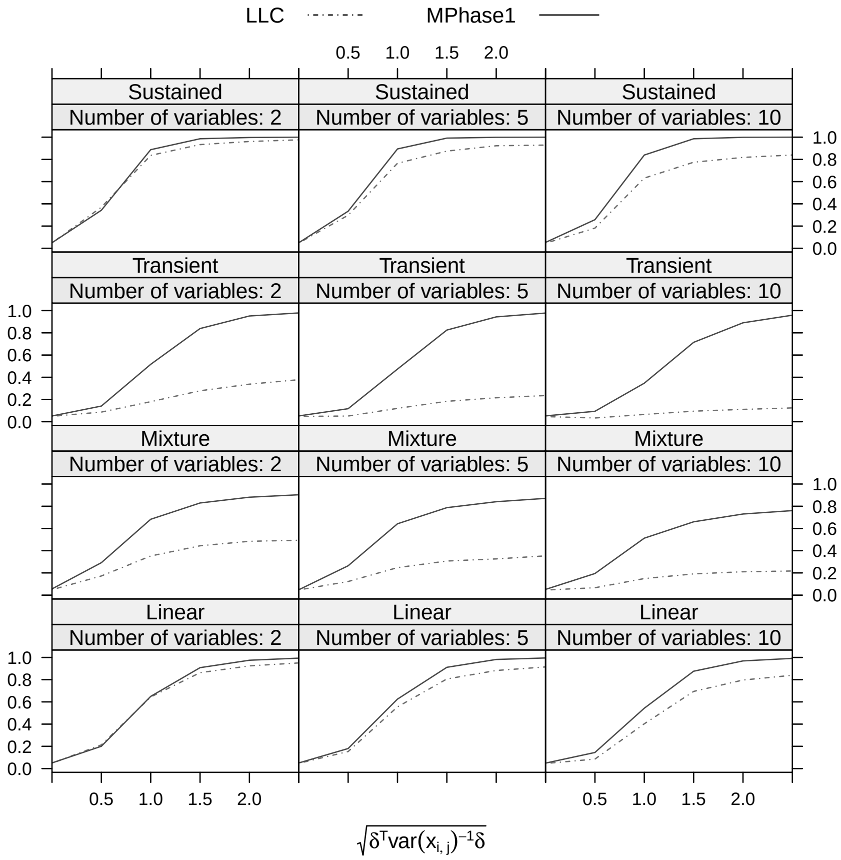


Figure 1: Out-of-control alarm probability of the MPhase1 control chart and of the test suggested by Lung-Yut-Fong et al. (2011) (Student t data with 3 degrees of freedom, $m = 50$, $n = 1$).

proposed by Lung-Yut-Fong et al. (2011). An analogous comparison, in the case of normally distributed data, has been presented in Subsection 4.2 of the paper.

Results depend on the particular distribution, the number of variables and patterns of the location shifts. However, they clearly point to an inferior efficiency of the Lung-Yut-Fong et al. (2011) method. Further, observe that the relative efficiency of this method decreases as the number of variables increases.

We also studied the OC behavior of our method in scenarios for which no competitor is currently available. In general, we observe a similar OC performance for different distributions, as shown in Figure 3 for multivariate normal or gamma data.

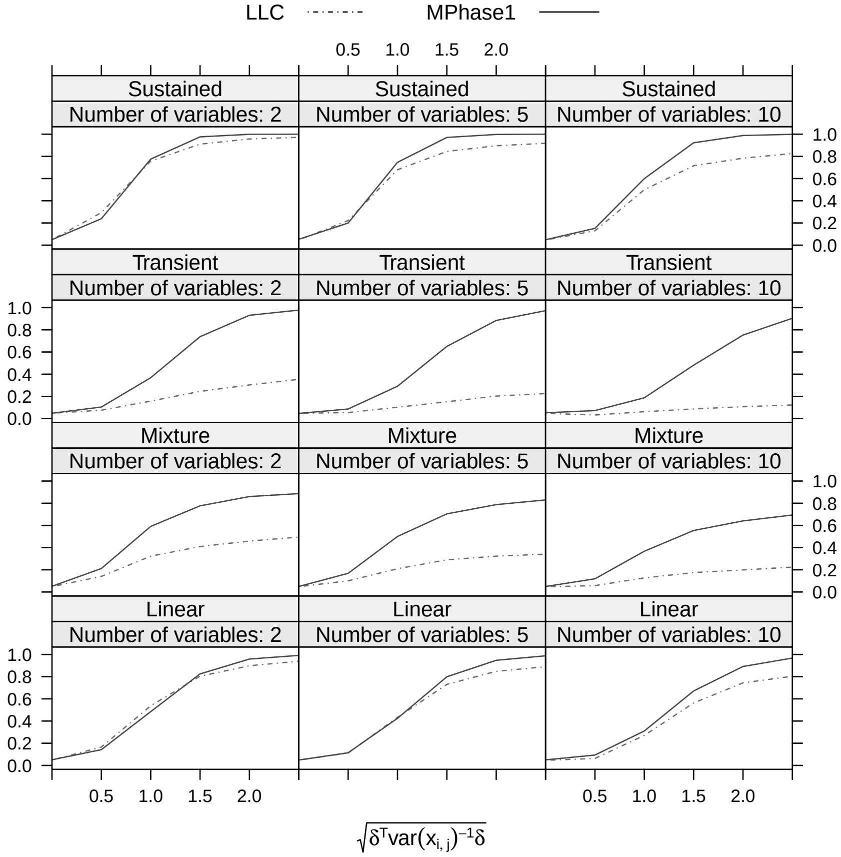


Figure 2: Out-of-control alarm probability of the MPhase1 control chart and of the test suggested by Lung-Yut-Fong et al. (2011) (Multivariate Gamma, $m = 50$, $n = 1$).

S4 Additional Results on the Post-Signal Procedure

Table 1 shows additional simulation results concerning the performance of the post-signal procedure described in Section 3.5 of the paper. In particular, we want to illustrate that, in many practical situations, the ability to correctly classify an observation as IC or OC depends more on γ, the parameter determining the EBIC criteria, than on α, the desired value for the FAP. Complementing the results presented in Table 4 of the paper, Table 1 can also be interesting for choosing a suitable value of γ. Indeed, results show that our procedure can be designed to be

liberal: when $\gamma = 0$ most of the OC data are correctly classified as OC but also a substantial proportion of IC observations are flagged as OC.

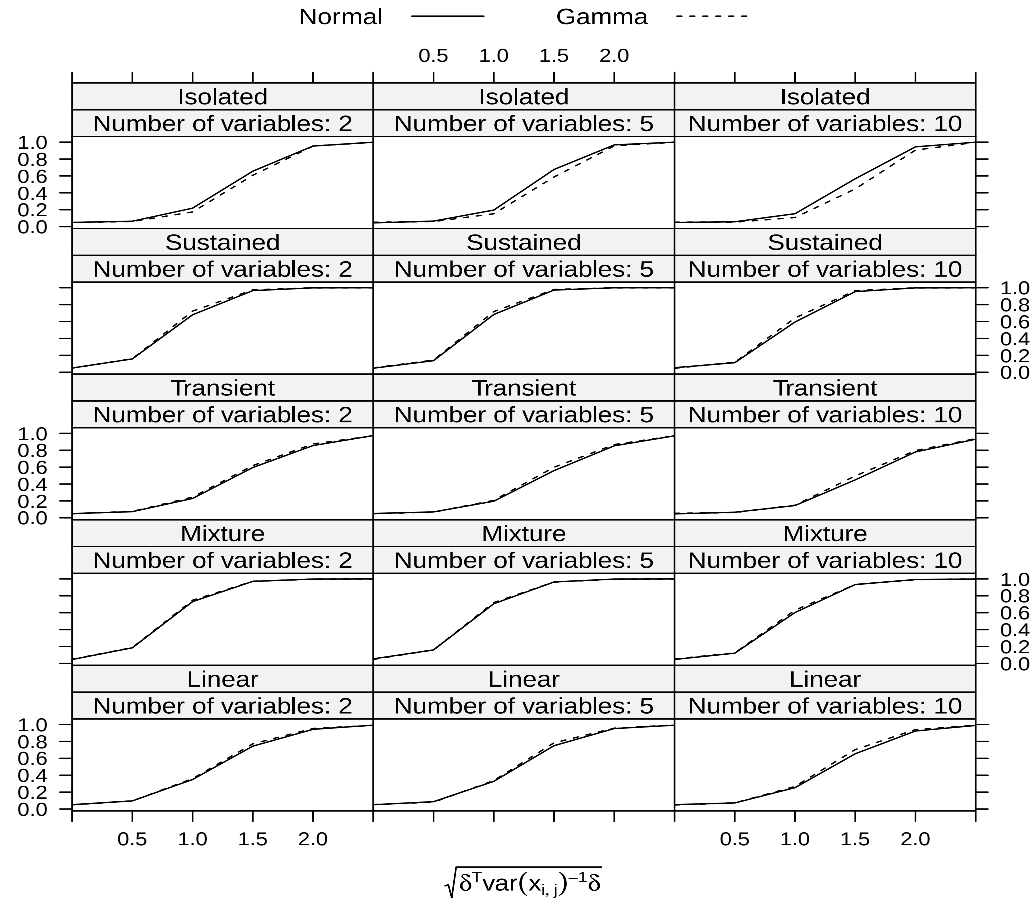


Figure 3: Out-of-control alarm probability of the suggested Phase I procedure for subgrouped normal and gamma observations ($m = 50$, $n = 5$).

conservative: when $\gamma = 1$ few data points are falsely signaled as OC but, if the shifts are small, a not negligible proportion of OC data are not flagged.

balanced: when $\gamma = 0.5$ it is possible to reach a compromise between the number of false and true OC signals.

To illustrate these points, for different values of g, n, m and δ, we proceed as follows:

- m subgroups, each of n observations, are simulated from a g-variate Student's t distribution $(X_1, \ldots, X_g)$ with three degrees of freedom, such that $\mathrm{E}(X_i) = 0$, $\mathrm{var}(X_i) = 1$ and $\mathrm{cov}(X_i, X_j) = 0.6$.

- A shift of size δ is added only to the first variable in 20% randomly chosen subgroups. In the following, these subgroups are considered as OC, the others as IC.

Table 1: Averages, and, in parentheses, standard deviations, of the false in-control and out-of-control rates

		$\gamma = 0.0$		$\gamma = 0.5$		$\gamma = 1.0$	
δ	α	FICR	FOCR	FICR	FOCR	FICR	FOCR
				$g = 5$, $n = 5$, $m = 50$			
1	0.01	0.0470 (0.0478)	0.1878 (0.1316)	0.0994 (0.0519)	0.0428 (0.0918)	0.1557 (0.0410)	0.0132 (0.1001)
	0.05	0.0426 (0.0377)	0.1964 (0.1292)	0.0981 (0.0497)	0.0448 (0.0947)	0.1550 (0.0404)	0.0157 (0.1104)
	0.10	0.0423 (0.0368)	0.1971 (0.1289)	0.0980 (0.0496)	0.0448 (0.0947)	0.1549 (0.0404)	0.0157 (0.1104)
2	0.01	3e-04 (0.0028)	0.1980 (0.1112)	0.0012 (0.0056)	0.0635 (0.0781)	0.0034 (0.0103)	0.0141 (0.0364)
	0.05	3e-04 (0.0028)	0.1980 (0.1112)	0.0012 (0.0056)	0.0635 (0.0781)	0.0034 (0.0103)	0.0141 (0.0364)
	0.01	3e-04 (0.0028)	0.1980 (0.1112)	0.0012 (0.0056)	0.0635 (0.0781)	0.0034 (0.0103)	0.0141 (0.0364)
				$g = 5$, $n = 5$, $m = 100$			
1	0.01	0.0463 (0.0311)	0.1644 (0.0962)	0.1096 (0.0437)	0.0299 (0.0644)	0.1738 (0.029)	0.0156 (0.0846)
	0.05	0.0462 (0.0309)	0.1646 (0.0961)	0.1096 (0.0436)	0.0299 (0.0644)	0.1737 (0.029)	0.0158 (0.0734)
	0.10	0.0462 (0.0309)	0.1646 (0.0961)	0.1096 (0.0436)	0.0299 (0.0644)	0.1737 (0.029)	0.0158 (0.0734)
2	0.01	4e-04 (0.0034)	0.1241 (0.0792)	9e-04 (0.0071)	0.0654 (0.0582)	0.0021 (0.0087)	0.0371 (0.0525)
	0.05	4e-04 (0.0034)	0.1241 (0.0792)	9e-04 (0.0071)	0.0654 (0.0582)	0.0021 (0.0087)	0.0371 (0.0525)
	0.10	4e-04 (0.0034)	0.1241 (0.0792)	9e-04 (0.0071)	0.0654 (0.0582)	0.0021 (0.0087)	0.0371 (0.0525)
				$g = 10$, $n = 5$, $m = 50$			
1	0.01	0.0453 (0.0574)	0.2240 (0.1355)	0.0929 (0.0562)	0.0481 (0.088)	0.1489 (0.0438)	0.0119 (0.0887)
	0.05	0.0369 (0.0420)	0.2400 (0.1278)	0.0888 (0.0507)	0.0511 (0.0924)	0.1475 (0.0425)	0.0131 (0.0942)
	0.10	0.0351 (0.0377)	0.2429 (0.1255)	0.0882 (0.0497)	0.0511 (0.0924)	0.1473 (0.0422)	0.0131 (0.0942)
2	0.01	1e-04 (0.0014)	0.1946 (0.1129)	3e-04 (0.0027)	0.0821 (0.0817)	0.0010 (0.005)	0.0419 (0.0584)
	0.05	1e-04 (0.0014)	0.1946 (0.1129)	3e-04 (0.0027)	0.0821 (0.0817)	0.0010 (0.005)	0.0419 (0.0584)
	0.01	1e-04 (0.0014)	0.1946 (0.1129)	3e-04 (0.0027)	0.0821 (0.0817)	0.0010 (0.005)	0.0419 (0.0584)
				$g = 10$, $n = 5$, $m = 100$			
1	0.01	0.0358 (0.0293)	0.2072 (0.0914)	0.094 (0.0466)	0.0447 (0.0648)	0.1648 (0.0331)	0.0033 (0.0354)
	0.05	0.0353 (0.0274)	0.2080 (0.0907)	0.0938 (0.0463)	0.0452 (0.0663)	0.1648 (0.0331)	0.0043 (0.0474)
	0.01	0.0353 (0.0274)	0.2080 (0.0907)	0.0938 (0.0463)	0.0452 (0.0663)	0.1648 (0.0331)	0.0043 (0.0474)
2	0.01	1e-04 (0.0011)	0.1271 (0.0835)	3e-04 (0.0021)	0.0276 (0.0382)	7e-04 (0.0031)	0.0150 (0.0265)
	0.05	1e-04 (0.0011)	0.1271 (0.0835)	3e-04 (0.0021)	0.0276 (0.0382)	7e-04 (0.0031)	0.0150 (0.0265)
	0.10	1e-04 (0.0011)	0.1271 (0.0835)	3e-04 (0.0021)	0.0276 (0.0382)	7e-04 (0.0031)	0.0150 (0.0265)

– Our procedure is applied to the simulated data using different values of α and γ. Since, we want to study the method's ability to correctly classify subgroups as IC or OC, we only focus on the detection of isolated outliers. Indeed, when a step shift is detected, the automatic classification of an observations as IC or OC can be difficult.

For each simulation run, results can be summarized in the following table

	Number of subgroups classified as	
	in-control	out-of-control
In-control subgroups	A	B
Out-of-control subgroups	C	D

Table 1 shows averages (and standard deviations) of the false IC rate (FICR), and false

OC rate (FOCR) defined as follows

$$\text{FICR} = \frac{C}{A+C} \text{ and } \text{FOCR} = \begin{cases} \dfrac{B}{B+D} & \text{if } B+D > 0 \\ 0 & \text{if } B+D = 0 \end{cases}.$$

The FICR (FOCR) is the proportion of IC (OC) subgroups wrongly classified as OC (IC). Observe that, in the multiple hypothesis testing literature, the average FOCR corresponds to the *false discovery rate*.

Table 1 shows that both rates almost do not depend on α, while they strongly depend on γ. Indeed, in the presence of many OC subgroups, our method signals with a probability close to one also when α is small. Hence, the classification in "good" or "bad" subgroups seems to strictly depends on the information criterion used during Stage 4.

Similarly to Table 4 in the paper, results in Table 1 point to a good performance of the proposed method which correctly detects all the medium/large shifts ($\delta = 2$) and most of the small/medium shifts ($\delta = 1$), even maintaining an acceptable number of false detections when $\gamma = 0.5$ or $\gamma = 1$.

S5 Comparison of Permutation and Elliptical Approaches

An alternative to the use of the permutation-based p-value is possible assuming that the IC distribution is elliptical (Oja, 2010). Indeed, the following result follows from an analogous property of the multivariate signed ranks (Hallin and Paindaveine, 2002).

Proposition. *Assume that $\mathbf{x}_{i,j}$ are i.i.d. with a common elliptical probability distribution. Then,*

$$\text{Prob}_0(T_1 \le t_1, \ldots, T_k \le t_k) = \text{Prob}_{\textit{Standard Normal}}(T_1 \le t_1, \ldots, T_k \le t_k)$$

where $\text{Prob}_{\textit{Standard Normal}}(\cdot)$ is the distribution of $T_1, \ldots, T_k$, computed by assuming that $\mathbf{x}_{i,j}$ have a g-dimensional standard normal distribution.

Hence, we can compute the p-value by assuming a multivariate normal distribution with known mean and dispersion. This is the approach followed for the design of the Shewhart charts by Bell et al. (2014) and Cheng and Shiau (2015).

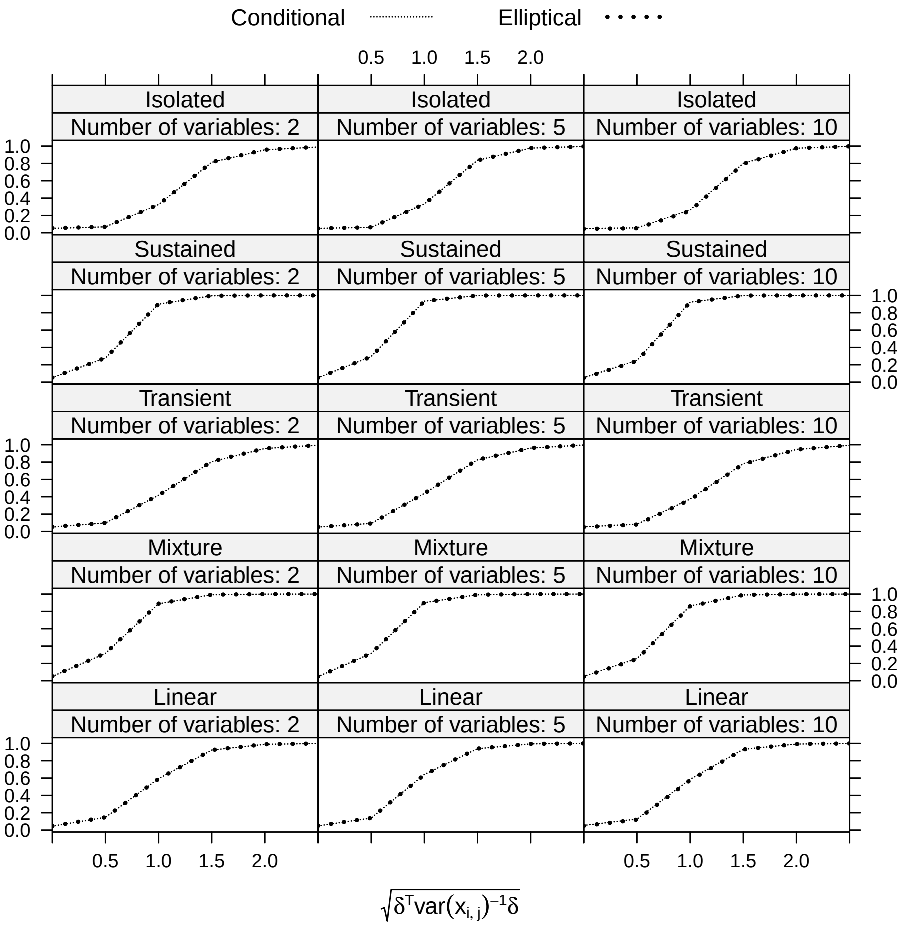


Figure 4: Out-of-control alarm probability of the suggested Phase I procedure with p-values computed using the "permutation" and "elliptical" approaches (subgrouped Student's t observations with 3 degrees of freedom, $m = 50$, $n = 5$).

We have extensively studied this possibility to determine whether it could offer some advantages; however, our conclusion was negative. Indeed, as documented by Figure 4 in the case of a multivariate Student's t distribution with three degrees of freedom, the OC performance of both approaches is essentially the same for every elliptical distribution we have considered. This means that assuming an elliptical IC distribution would restrict the applicability of the proposed method, without providing a better performance when the IC distribution is actually elliptical.

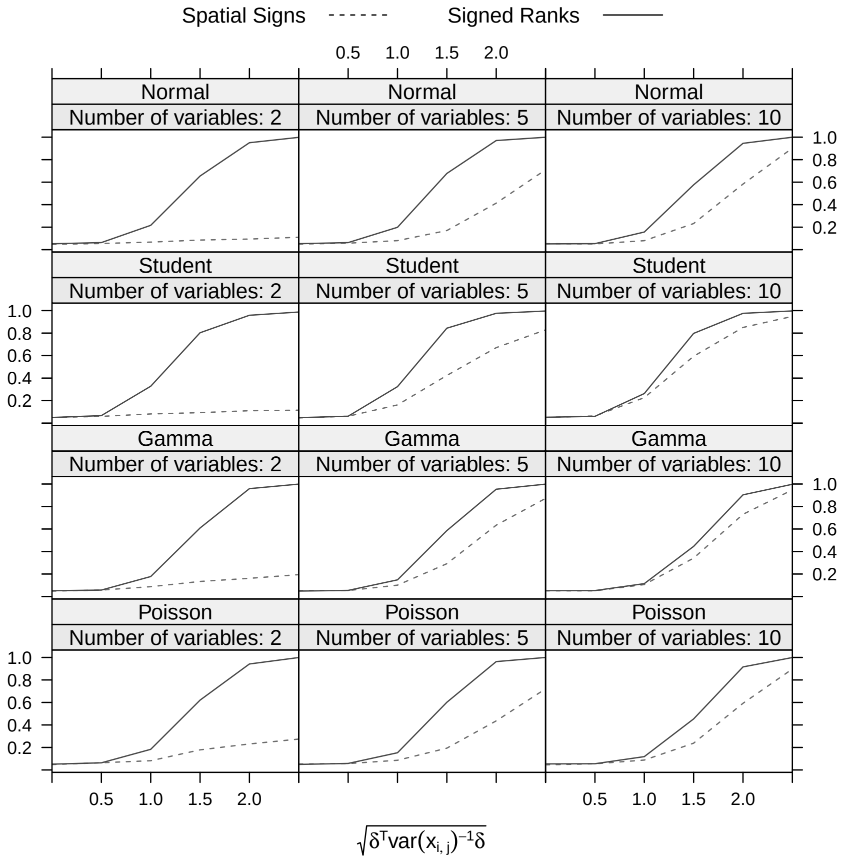


Figure 5: Out-of-control alarm probability of the MPhase1 control chart based on the spatial signs and the multivariate signed ranks (isolated shifts, $m = 50$, $n = 5$).

S6 Spatial Signs vs Signed Ranks

As mentioned in Section S1, we have studied the performance of our method substituting the multivariate signed ranks with the corresponding spatial signs, i.e., replacing equation (4) of the paper with

$$\mathbf{u}_{i,j} = \begin{cases} \mathbf{0} & \text{if } \mathbf{z}_{i,j} = \mathbf{0}; \\ \dfrac{\mathbf{z}_{i,j}}{||\mathbf{z}_{i,j}||} & \text{if } \mathbf{z}_{i,j} \neq \mathbf{0}. \end{cases}.$$

For different distributions and number of variables, Figure 5 shows the detection power of the original and modified methods when there is an isolated location shift. Similar results have been obtained for other shift patterns.

Results show that using the multivariate signed ranks, which combine the information

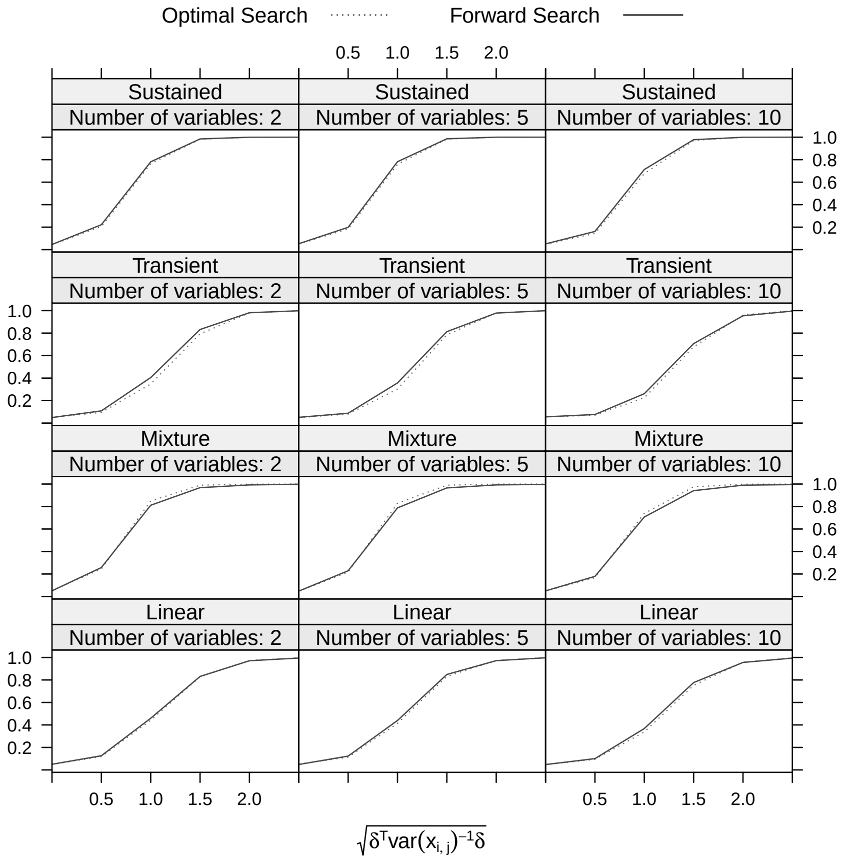


Figure 6: Out-of-control alarm probability of the MPhase1 control chart using the forward and the optimal search (subgrouped data from a Normal distribution, $m = 50$, $n = 5$).

given by the spatial signs with those given by the ranks of the Mahalanobis depths, offers a better performance. However, it should be observed that, consistently with some results regarding the multivariate nonparametric tests (Oja, 2010), the relative efficiency of the method based on the spatial signs increases with the number of variables.

S7 Optimal vs Forward Search

One of the main goals of the proposed Phase I method, consists of the simultaneous detection of multiple isolated and step shifts. However, when the identification of isolated shifts is not deemed important, it is feasible to replace the forward search, which corresponds to a recursive binary partitioning approach, with the search of the opti-

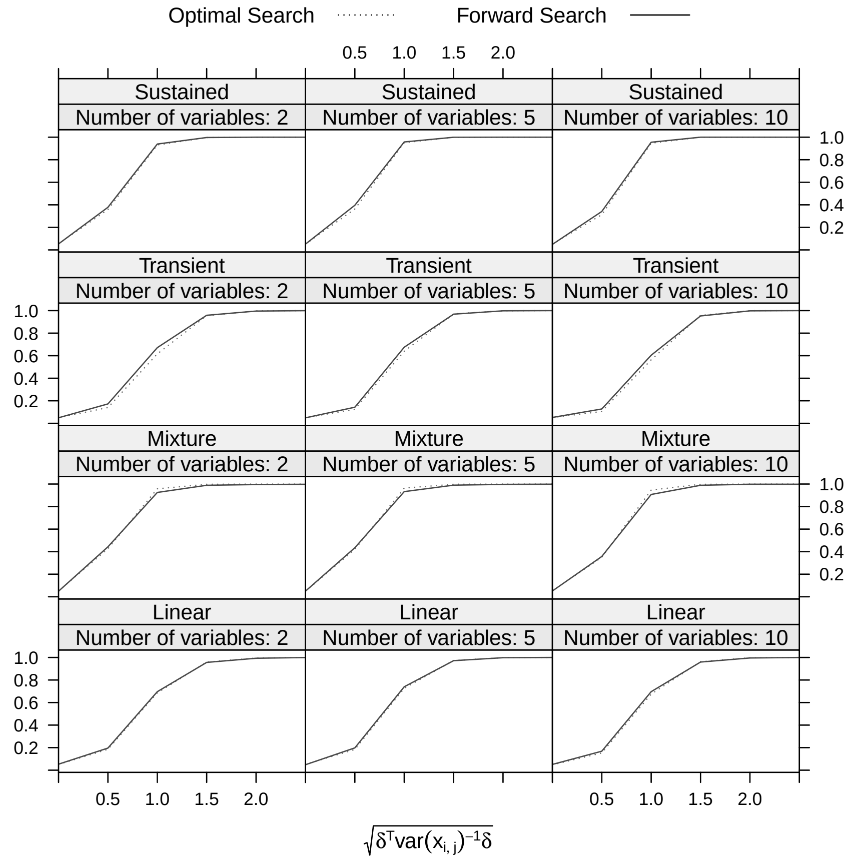


Figure 7: Out-of-control alarm probability of the MPhase1 control chart using the forward and the optimal search (subgrouped data from a Student's t distribution with 3 degrees of freedom, $m = 50$, $n = 5$).

mal time-ordered partition in the least squares sense, for which some relatively fast algorithms exist (see Killick et al., 2012 and Qiu, 2013, Chapter 6).

For this reason, we have investigated if the substitution of the forward with the optimal search leads to a better performance when only step shifts are considered. However, differences in the detection power do not seem of practical importance (see Figures 6 and 7).

Table 2: Performance of the post-signal diagnostic method based on an alternative implementation of the Stage 4. ($\alpha = 0.05$).

	Average Number of True Detections				Average Number of		
	Exact		Approximated		False Detections		
	$\delta = 1$	$\delta = 2$	$\delta = 1$	$\delta = 2$	$\delta = 0$	$\delta = 1$	$\delta = 2$
Scenario A / Number of Variables: 5 / Number of True Shifts: 3							
$\gamma = 0$	1.64	1.99	1.89	2.00	0.16	0.05	0.02
$\gamma = 0.5$	1.54	1.99	1.76	2.00	0.08	0.01	0.00
$\gamma = 1$	1.29	1.99	1.48	2.00	0.06	0.00	0.00
Scenario A / Number of Variables: 10 / Number of True Shifts: 3							
$\gamma = 0$	1.64	1.98	1.89	2.00	0.14	0.05	0.02
$\gamma = 0.5$	1.54	1.98	1.77	2.00	0.08	0.02	0.00
$\gamma = 1$	1.32	1.98	1.51	2.00	0.05	0.01	0.00
Scenario B / Number of Variables: 5 / Number of True Shifts: 10							
$\gamma = 0$	6.01	7.91	7.40	8.21	0.14	0.59	0.25
$\gamma = 0.5$	3.92	7.90	4.74	8.21	0.07	0.16	0.07
$\gamma = 1$	1.74	7.88	2.10	8.17	0.06	0.04	0.02
Scenario B / Number of Variables: 10 / Number of True Shifts: 10							
$\gamma = 0$	5.58	7.82	6.97	8.11	0.18	0.97	0.32
$\gamma = 0.5$	3.38	7.83	4.09	8.12	0.08	0.19	0.10
$\gamma = 1$	1.46	7.83	1.79	8.11	0.06	0.06	0.03

S8 An Alternative Implementation of Stage 4

Given the definition of the overall test statistic

$$W_{OBS} = \max_{k=1,\ldots,K} \frac{T_k - \mathrm{E}_0(T_k)}{\sqrt{\mathrm{var}_0(T_k)}}$$

(see Subsection 3.4 of the paper), the application of the adaptive LASSO, during the post-signal diagnostic stage (Subsection 3.5 of the paper), could be restricted to the first

$$\widehat{K} = \underset{k=1,\ldots,K}{\mathrm{argmax}} \frac{T_k - \mathrm{E}_0(T_k)}{\sqrt{\mathrm{var}_0(T_k)}}$$

shifts selected during the screening stage. Indeed, $T_{\widehat{K}}$ is in some sense the most unlikely elementary statistic among $T_1, \ldots, T_K$. Further, the use of $\widehat{K}$ has been recommended by Capizzi and Masarotto (2013) in the univariate case.

However, at least in the multivariate case, we found that this choice can reduce the efficiency of the post-signal diagnostic stage. For example, Table 2 shows the average number of true and false shifts detected by the modified procedure in the same scenarios considered in Subsection 4.3 of the paper. When compared with the results presented in Table 4 of the paper, Table 2 shows that the use of $\widehat{K}$ decreases the sensitivity of the proposed method.

In particular, we observe this effect when the $\boldsymbol{\delta}$'s are quite sparse, i.e., when many variables do not shift. In these scenarios, $\widehat{K}$ tends to be too small because the explained variances T_k only increase slightly when a non-null but sparse $\boldsymbol{\delta}$ is included.

S9 Attained False Alarm Probabilities

Table 3 shows the attained false alarm probabilities displayed in Figure 4 of the paper.

References

Bell, R. C., Jones-Farmer, L. A., and Billor, N. (2014), "A Distribution-Free Multivariate Phase I Location Control Chart for Subgrouped Data from Elliptical Distributions," *Technometrics*, 56, 528–538.

Bühlmann, P. (2002), "Bootstraps for Time Series," *Statistical Science*, 17, 52–72.

Capizzi, G. and Masarotto, G. (2013), "Phase I Distribution-Free Analysis of Univariate Data," *Journal of Quality Technology*, 45, 273–284.

Chen, J. and Chen, Z. (2008), "Extended Bayesian Information Criteria for Model Selection with Large Model Spaces," *Biometrika*, 95, 759–771.

Cheng, C.-R. and Shiau, J.-J. H. (2015), "A Distribution-Free Multivariate Control Chart for Phase I Applications," *Quality and Reliability Engineering International*, 31, 97–111.

Efron, E., Hastie, T., Johnstone, I., and Tibshirani, R. (2004), "Least Angle Regression," *The Annals of Statistics*, 32, 407–499.

Green, W. H. (2012), *Econometrics Analysis*, Boston, MA: Prentice, 7th ed.

Hallin, M. and Paindaveine, D. (2002), "Optimal Tests for Multivariate Location Based on Interdirections and Pseudo-Mahalanobis Ranks," *The Annals of Statistics*, 30, 1103–1133.

Table 3: Attained false alarm probabilities for different distributions, different number of subgroups (m) and different subgroup sizes (n) when the number of variables (g) is five. The nominal false alarm probability is $\alpha = 0.05$.

Distribution	(m, n)	MPhase1	T2/Classic	T2/MCD
Normal	(20,1)	0.051	0.051	0.052
	(50,1)	0.051	0.052	0.051
	(100,1)	0.051	0.051	0.051
	(20,5)	0.052	0.051	0.052
	(50,5)	0.051	0.052	0.052
	(100,5)	0.051	0.052	0.051
	(20,10)	0.051	0.050	0.052
	(50,10)	0.052	0.051	0.051
	(100,10)	0.051	0.052	0.052
Student	(20,1)	0.051	0.204	0.278
	(50,1)	0.052	0.916	0.908
	(100,1)	0.044	0.997	0.999
	(20,5)	0.049	0.254	0.926
	(50,5)	0.051	0.718	0.997
	(100,5)	0.046	0.942	1.000
	(20,10)	0.049	0.152	0.960
	(50,10)	0.049	0.641	0.998
	(100,10)	0.050	0.870	1.000
Gamma	(20,1)	0.049	0.140	0.150
	(50,1)	0.051	0.629	0.738
	(100,1)	0.054	0.885	0.991
	(20,5)	0.046	0.135	0.915
	(50,5)	0.047	0.270	0.995
	(100,5)	0.049	0.435	1.000
	(20,10)	0.051	0.072	0.988
	(50,10)	0.049	0.187	1.000
	(100,10)	0.051	0.268	1.000
Poisson	(20,1)	0.053	0.109	0.015
	(50,1)	0.052	0.347	0.029
	(100,1)	0.049	0.569	0.017
	(20,5)	0.050	0.095	0.014
	(50,5)	0.049	0.158	0.008
	(100,5)	0.046	0.249	0.007
	(20,10)	0.052	0.046	0.015
	(50,10)	0.052	0.119	0.012
	(100,10)	0.048	0.142	0.026

Huang, J., Breheny, P., and Ma, S. (2012), "A Selective Review of Group Selection in High-Dimensional Models," *Statistical Science*, 27, 481–499.

Killick, R., Fearnhead, P., and Eckley, I. a. (2012), "Optimal Detection of Changepoints with a Linear Computational Cost," *Journal of the American Statistical Association*, 107, 1590–1598.

Lung-Yut-Fong, A., Lévy-Leduc, C., and Cappé, O. (2011), "Homogeneity and change-point detection tests for multivariate data using rank statistics," *arXiv preprint arXiv:1107.1971.*

Oja, H. (2010), *Multivariate Nonparametric Methods with R. An Approach Based on Spatial Signs and Ranks*, New York, NY: Springer.

Politis, D. N. (2015), *Model-Free Prediction and Regression*, New York, NY: Springer.

Qiu, P. (2013), *Introduction to Statistical Process Control*, Boca Raton, FL: Chapman & Hall/CRC Press.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005), "Sparsity and Smoothness via the Fused Lasso," *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 67, 91–108.

Wang, H. (2009), "Forward Regression for Ultra-High Dimensional Variable Screening," *Journal of the American Statistical Association*, 104, 1512–1524.

Wei, W. W. S. (2006), *Time Series Analysis: Univariate and Multivariate Methods*, Redwood City, CA: Addison-Wesley, 2nd ed.

Zou, H. (2006), "The Adaptive Lasso and Its Oracle Properties," *Journal of American Statistical Association*, 101, 1418–1429.

Supplementary material to “Phase I Distribution-Free Analysis of Multivariate Data”

Additional Examples using the R Package mphase1

Giovanna Capizzi
Department of Statistical Sciences, University of Padua, Italy

and

Guido Masarotto
Department of Statistical Sciences, University of Padua, Italy

December 5, 2016

Contents

1 Introduction

This short note illustrates the use of the R package mphase1 accompanying the paper.

The package can be loaded during an R session using

```r
library(mphase1)
```

2 The mphase1 Function

2.1 A simulated dataset

To illustrate the package, we use a simulated dataset included in mphase1, consisting of $m = 50$ subgroups, each with $n = 5$ observations, on $p = 4$ variables.

The observations have been generated by

$$\mathbf{x}_{i,j} = \boldsymbol{\delta}_1 \xi_i^{(1)} + \boldsymbol{\delta}_2 \xi_i^{(2)} + \boldsymbol{\epsilon}_{i,j} \quad (i = 1, \ldots, m; j = 1, \ldots, n)$$

where

(i) $\mathbf{x}_{i,j}$ denotes the jth observation vector for the ith subgroup.

(ii) $\xi_i^{(1)}$ and $\xi_i^{(2)}$ are two scalar sequences describing an isolated and a step shift, respectively. In particular,

$$\xi_i^{(1)} = \begin{cases} 0 & \text{if } i \neq 10 \\ 1 & \text{if } i = 10 \end{cases} \quad \text{and} \quad \xi_i^{(2)} = \begin{cases} 0 & \text{if } i < 31 \\ 1 & \text{if } i \geq 31 \end{cases}.$$

Note that 10 and 31 are the times of the shifts.

(iii) $\boldsymbol{\delta}_1$ and $\boldsymbol{\delta}_2$ are four dimensional vectors giving the directions and sizes of the two shifts. Specifically, we set

$$\boldsymbol{\delta}_1 = (1, 0, 0, 0)' \text{ and } \boldsymbol{\delta}_2 = (0, 0, 0.5, -0.25)'.$$

Note that the first shift involves only the first variable and the second shift only the third and fourth variables.

(iv) $\boldsymbol{\epsilon}_{i,j} = (\epsilon_{i,j,1}, \ldots, \epsilon_{i,j,4})'$ are four dimensional pseudo-random vectors simulated from a Student's t distribution with three degrees of freedom, such that $E(\epsilon_{i,j,r}) = 0$, $\mathrm{var}(\epsilon_{i,j,r}) = 1$ and $\mathrm{cor}(\epsilon_{i,j,r}, \epsilon_{i,j,s}) = 0.8^{|r-s|}$.

The code used to simulate the data is reported in Appendix A.

The dataset can be loaded in the current `R` session using

```
data(Student)
```

Observe that `Student` is a $p \times n \times m$ array

```
is.array(Student)

## [1] TRUE

dim(Student)

## [1]  4  5 50
```

such that `Student[r,j,i]` contains the jth observation on the rth variable belonging to the ith subgroup.

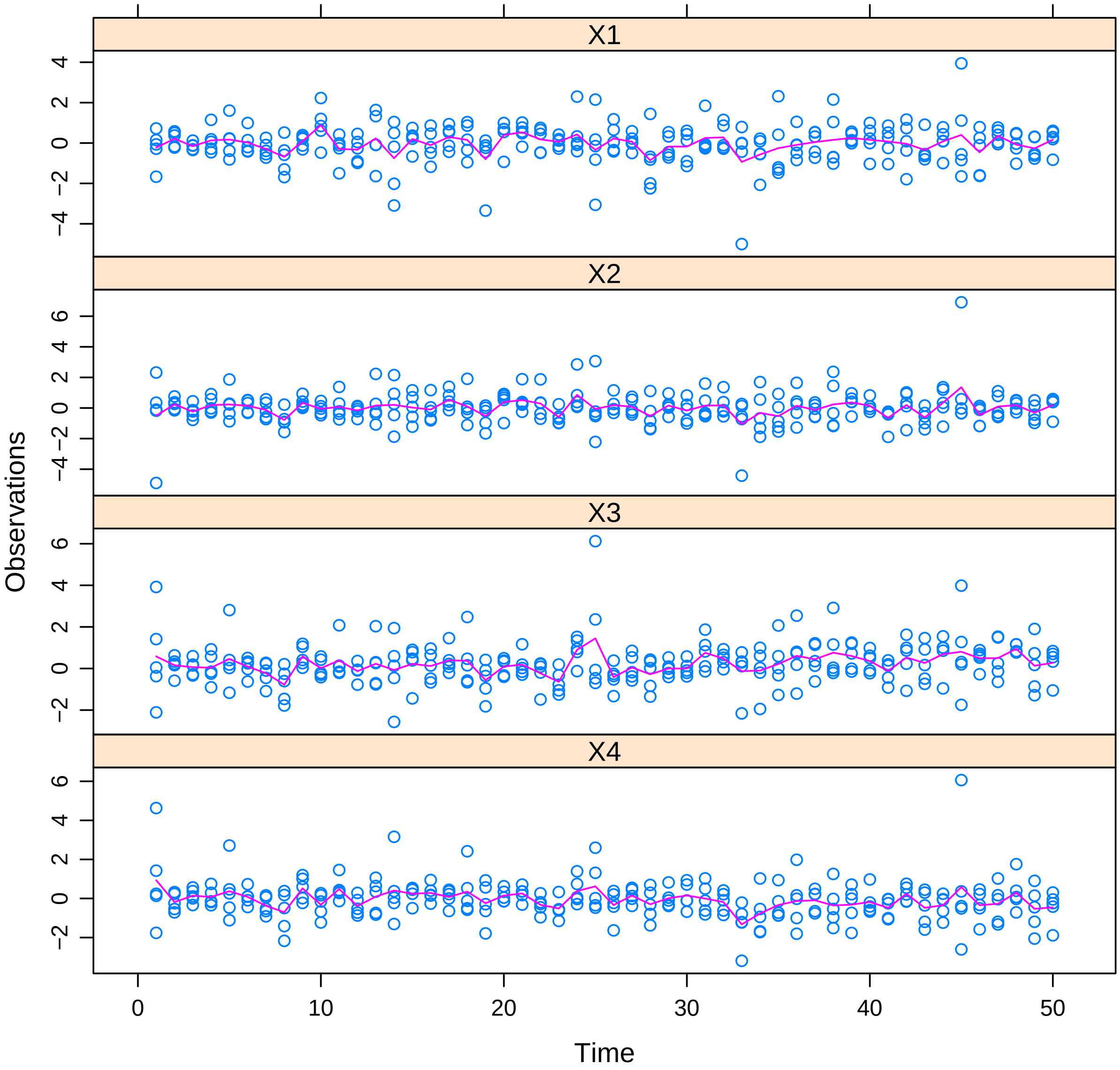
Therefore, for example, we can "extract" the observations of the 10th subgroup with the command

```
Student[,,10]

##          [,1]      [,2]       [,3]        [,4]      [,5]
## X1 -0.4756779 1.1946432  0.8431024  0.61521671 2.2278333
## X2 -0.4730533 0.1201301 -0.3281686 -0.04584001 0.4629794
## X3 -0.2203015 0.4243876 -0.3012886 -0.41782778 0.5809793
## X4 -1.2312955 0.1677948 -0.1530630 -0.66490621 0.2699826
```

The following command can be used to plot the data (the lines show the subgroup means)

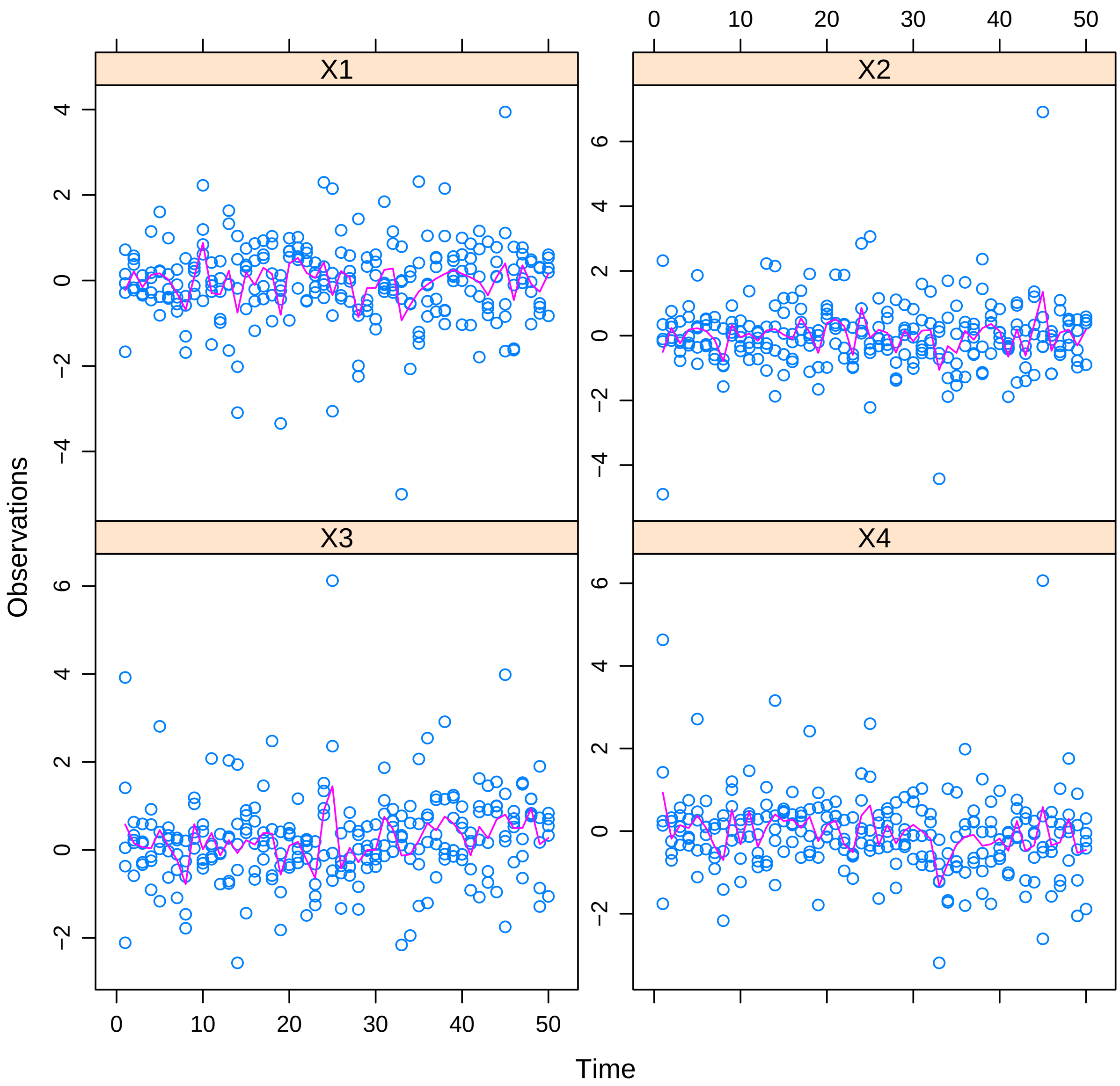
```
mphase1PlotData(Student)
```



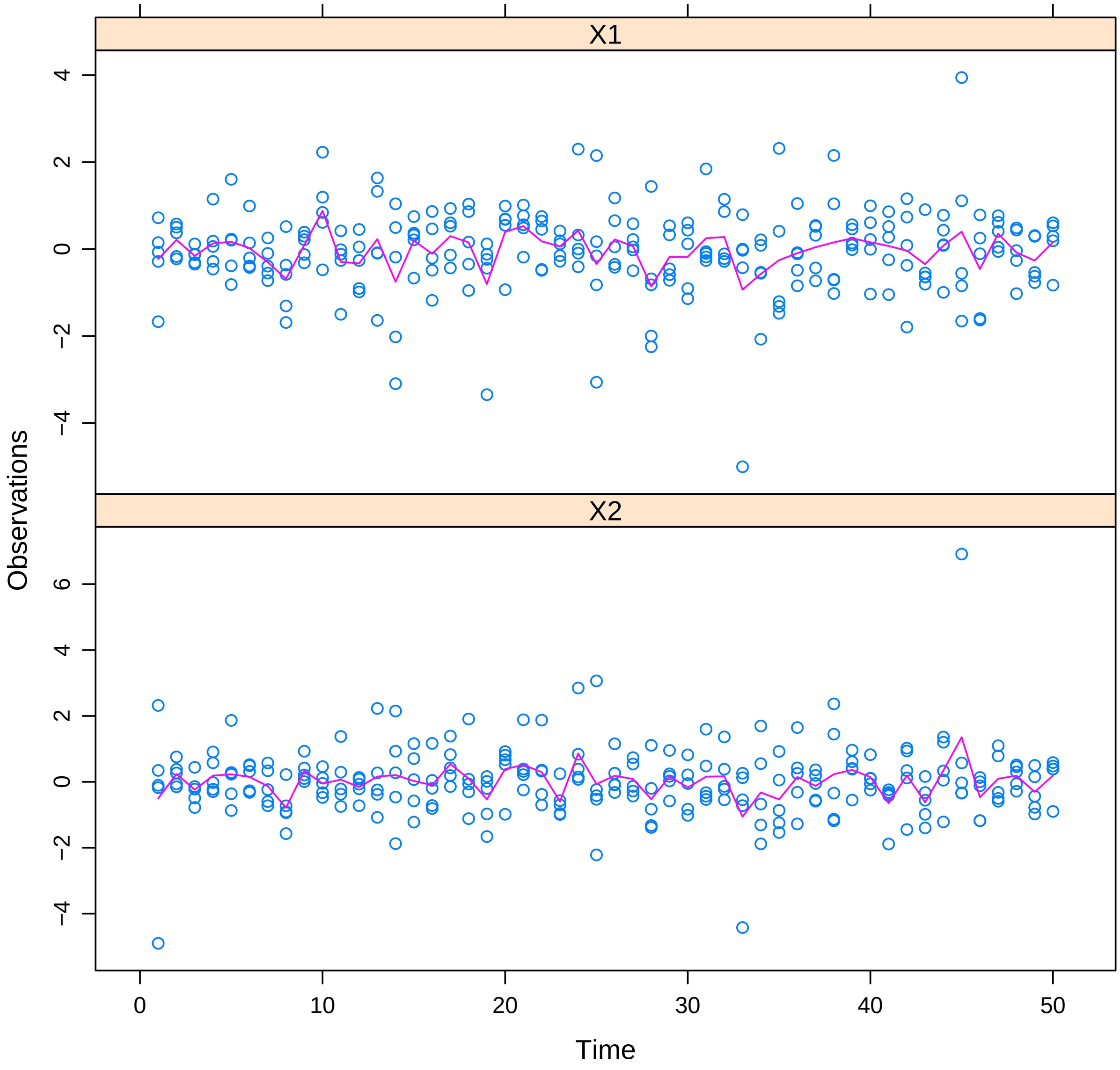
Observe that it is not easy to recognize the shifts in the original data (in particular, the sustained shift).

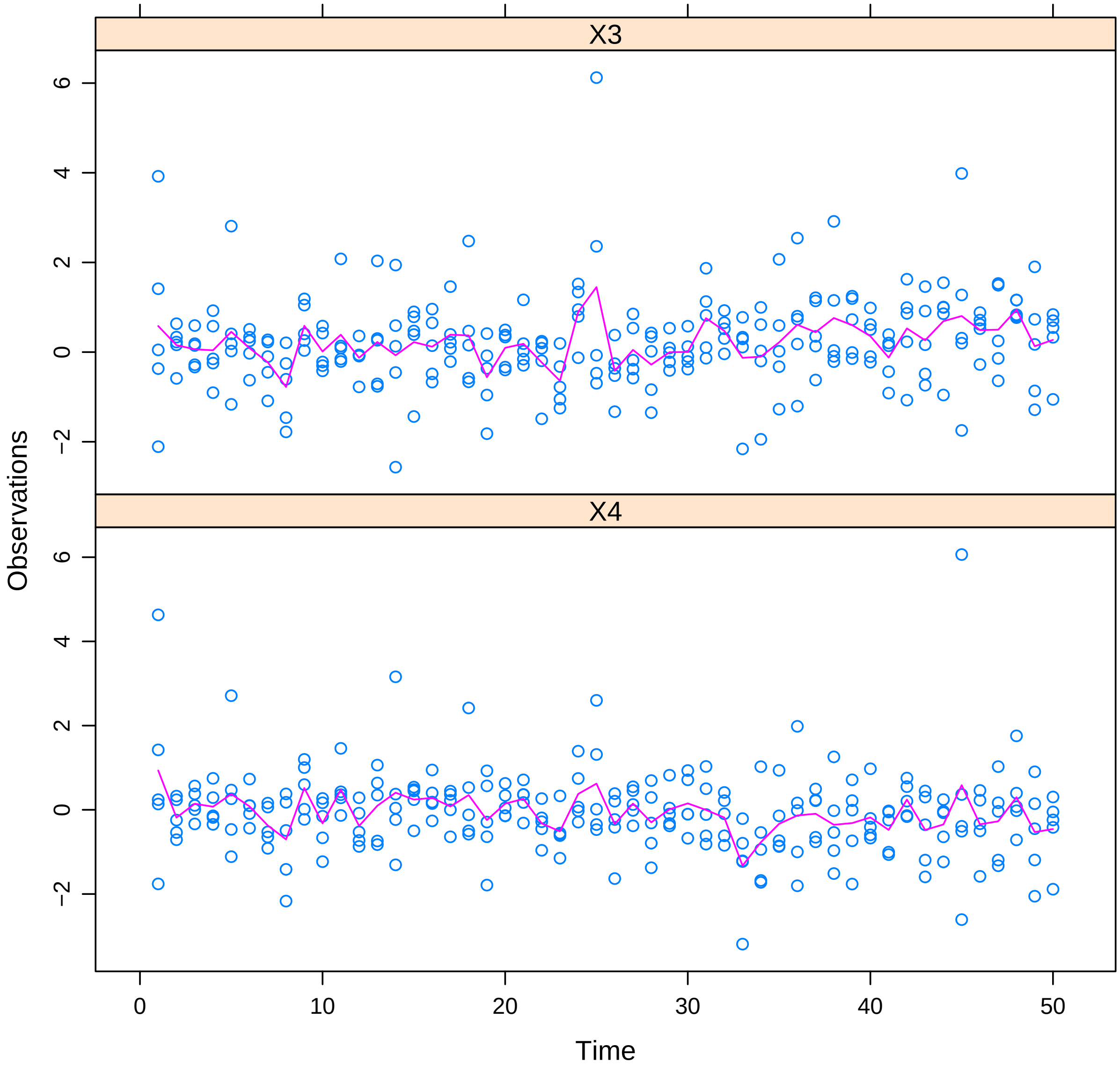
The panels of the previous figure can be rearranged (also in multiple plots) using the optional `layout` parameter. This possibility is particular useful when the number of variables is large. See the following examples.

```
mphase1PlotData(Student,layout=c(2,2))
```



```
mphase1PlotData(Student,layout=c(1,2,2))
```

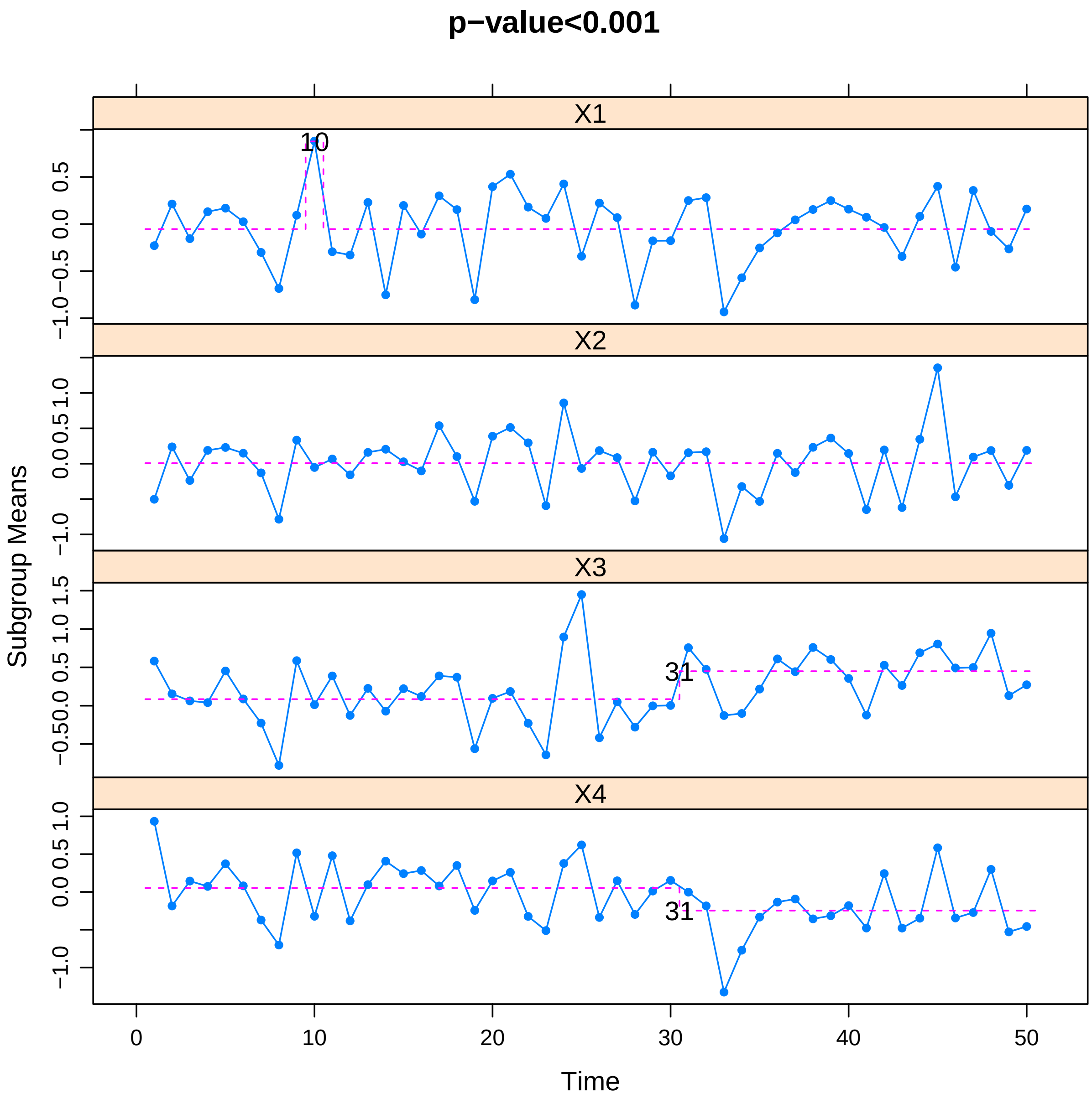

X1
X2
Observations
Time
4
2
0
−2
−4
6
4
2
0
−2
−4
0
10
20
30
40
50



2.2 Basic usage

The main function of the package, `mphase1`, performs the Phase I analysis described in the paper. In many practical situations, it is enough to call this function with only one argument: an array containing the dataset.

```
mphase1(Student)
```



```
## Call:
## mphase1(x = Student, plot = TRUE, post.signal = TRUE, isolated = TRUE,
##     step = TRUE, alpha = 0.05, gamma = 0.5, K = 7, lmin = 5,
##     L = 1000, seed = 11642257)
##
## p-value < 0.001
```

```
##
## Location Shifts:
##       type time variables
## 1     Step   31       3,4
## 2 Isolated   10         1
```

In this case, the small p-value, shown at the top of the figure, points to an unstable process. Further, graphics correctly suggest the presence of (i) one isolated shift at time 10 involving only the first variable, and (ii) a step shift starting from time 31 and regarding the third and fourth variables. The same information (plus the values of the optional parameters that are described in the manual) is provided by the output printed on the `R` console.

2.3 The returned object

Function `mphase1` returns an object of class *mphase1*.

```
system.time(u <- mphase1(Student,plot=FALSE))

##    user  system elapsed
##   0.514   0.000   0.512

class(u)

## [1] "mphase1"
```

Observe that the computation required less than a second.

When the object is assigned to a variable (`u` is this example), we have access to

- the p-value.

  ```
  u$p.value

  ## [1] 0
  ```

- a dataframe containing the results of the LASSO-based diagnostic procedure.

```
u$alasso

##       type time variables
## 1     Step   31       3,4
## 2 Isolated   10         1
```

- the results of the forward search stage.

```
u$forward

##       type time        T        a        b
## 1     Step   31 129.5188 13.85431 3.201762
## 2 Isolated   10 145.4882 25.19917 4.707573
## 3 Isolated   41 156.9932 35.29905 5.892541
## 4 Isolated    1 167.5158 44.47737 6.854161
## 5 Isolated   23 175.9102 52.95266 7.648564
## 6 Isolated   24 182.3908 60.90623 8.334466
## 7 Isolated   33 188.2676 68.41551 8.991980
```

In particular `u$forward` is a dataframe containing the types and times of the shifts identified during the forward search as well as the values of the elementary statistics, T_i, and their estimated in-control mean and standard deviations (a_i and b_i).

- the estimates of the location vector and scatter matrix used to standardize the data.

```
u$center

##           X1           X2           X3           X4
##  0.003218898  0.050398124  0.221409534 -0.035299271

u$scatter
```

```
##           X1        X2        X3        X4
## X1 0.9461620 0.7908112 0.5081340 0.4712398
## X2 0.7908112 1.1107008 0.7538285 0.7381769
## X3 0.5081340 0.7538285 1.0271373 0.8461249
## X4 0.4712398 0.7381769 0.8461249 0.9672659
```

- Three $p \times n \times m$ arrays

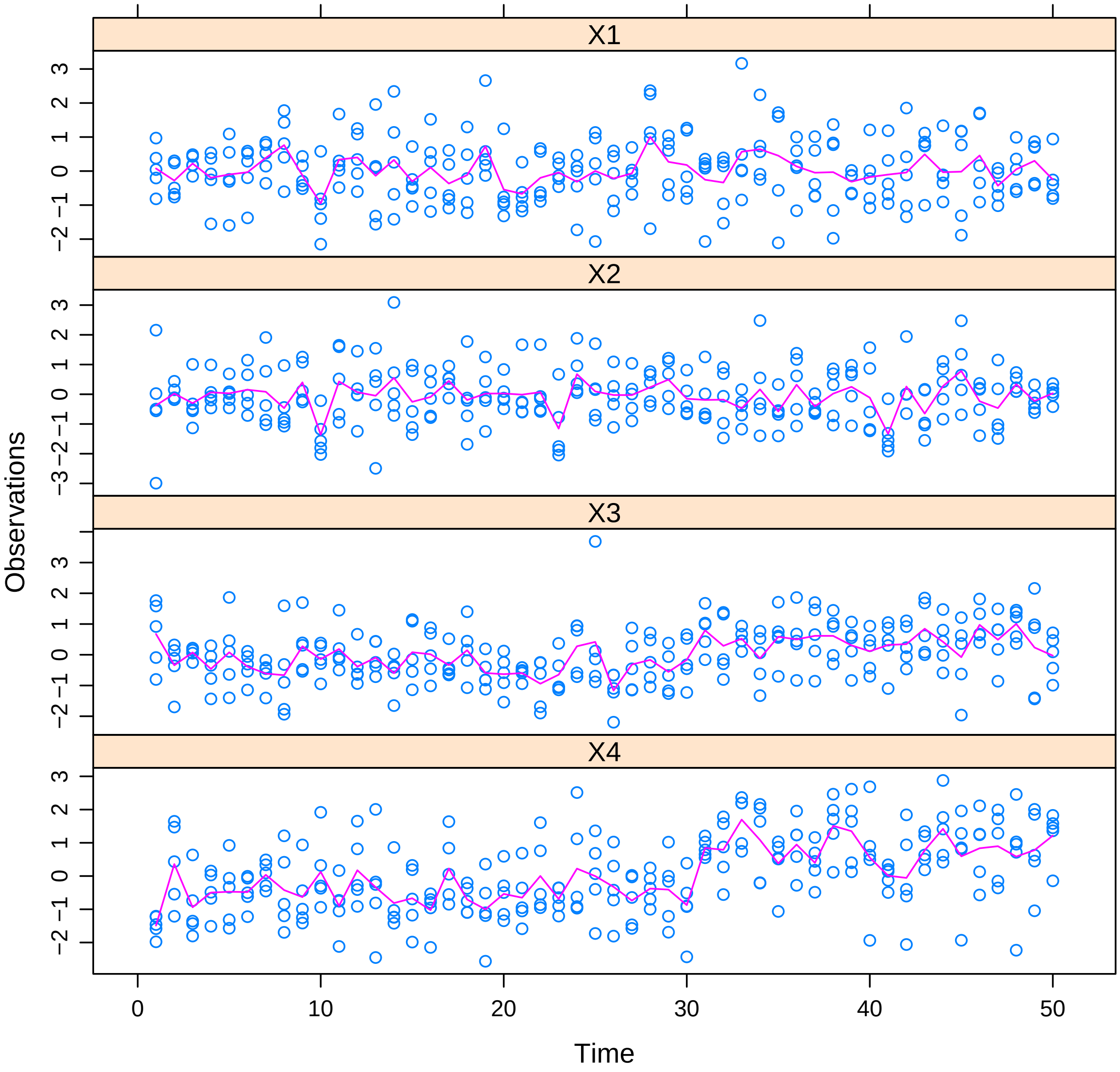
  (i) `signed.ranks` containing the signed ranks of the original data;

  (ii) `fitted` containing the predicted values computed at the last stage of the procedure;

  (iii) `residuals` containing the difference between the data and the fitted values.

  Therefore, for example, the signed ranks can be plotted using the command

```
mphase1PlotData(u$signed.ranks)
```



In addition, we can compute an estimate of the shifts in the four variables at time 10 and 31 using

```
round(u$fitted[,1,10]-u$fitted[,1,9],3)
```

```
##    X1    X2    X3    X4
## 0.931 0.000 0.000 0.000
```

and

```
round(u$fitted[,1,31]-u$fitted[,1,30],3)

##     X1     X2     X3     X4
##  0.000  0.000  0.365 -0.299
```

2.4 print, plot and postsignal

Methods **print**, **plot** and **postsignal** are available for objects of class *mphase1*. The first one allows to print to the console the main results

```
print(u)

## Call:
## mphase1(x = Student, plot = FALSE, post.signal = TRUE, isolated = TRUE,
##     step = TRUE, alpha = 0.05, gamma = 0.5, K = 7, lmin = 5,
##     L = 1000, seed = 11642257)
##
## p-value < 0.001
##
## Location Shifts:
##       type time variables
## 1     Step   31       3,4
## 2 Isolated   10         1
```

The same output can be obtained typing the name of the object

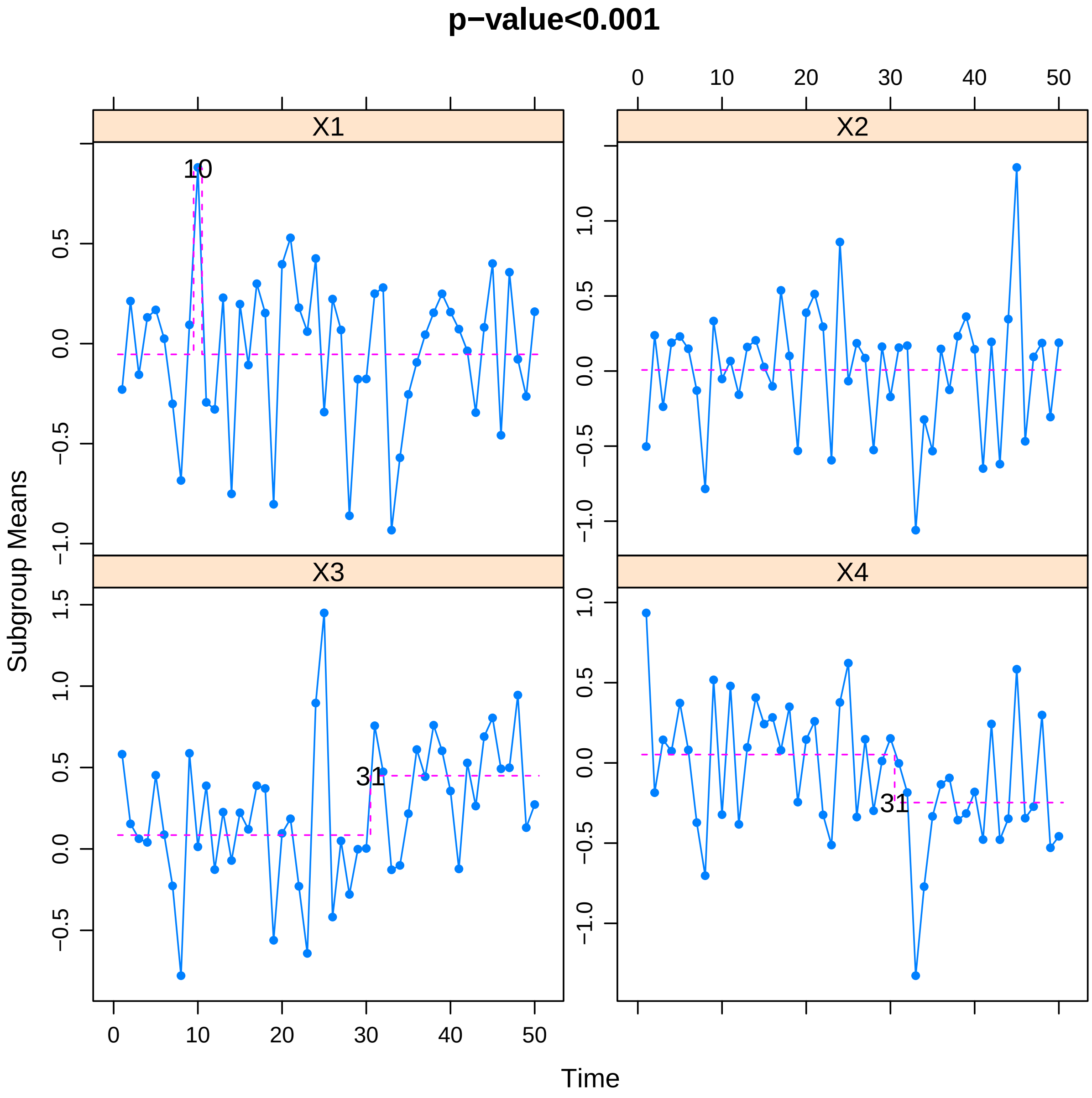
```
u

## Call:
## mphase1(x = Student, plot = FALSE, post.signal = TRUE, isolated = TRUE,
##     step = TRUE, alpha = 0.05, gamma = 0.5, K = 7, lmin = 5,
##     L = 1000, seed = 11642257)
##
```

```
## p-value < 0.001
##
## Location Shifts:
##       type time variables
## 1     Step   31       3,4
## 2 Isolated   10         1
```

The `plot` method can be used to show graphically the results, optionally rearranging the panels, as shown by the following example

```
plot(u,layout=c(2,2))
```



The `postsignal` method allows to re-run the LASSO-based post-signal diagnostic procedure using different values of γ, the extra penalization parameter used in the information criteria EBIC. In this case, the default value, $\gamma = 0.5$, leads to the correct identification of times and types of the two shifts. On the contrary, using $\gamma = 1$, the isolated shift at time 10 is missed

```
postsignal(u,gamma=1,plot=FALSE)

## Call:
```

```
## mphase1(x = Student, plot = FALSE, post.signal = TRUE, isolated = TRUE,
##     step = TRUE, alpha = 0.05, gamma = 1, K = 7, lmin = 5, L = 1000,
##     seed = 11642257)
##
## p-value < 0.001
##
## Location Shifts:
##   type time variables
## 1 Step   31       3,4
```

Finally, using $\gamma = 0$, i.e., the standard BIC criteria, an additional false isolated shift, involving only the fourth variable, is identified.

```
postsignal(u,gamma=0,plot=FALSE)

## Call:
## mphase1(x = Student, plot = FALSE, post.signal = TRUE, isolated = TRUE,
##     step = TRUE, alpha = 0.05, gamma = 0, K = 7, lmin = 5, L = 1000,
##     seed = 11642257)
##
## p-value < 0.001
##
## Location Shifts:
##       type time variables
## 1     Step   31       3,4
## 2 Isolated   10         1
## 3 Isolated    1         4
```

The `postsignal` method can also be used to re-run the diagnostic procedure using a different value of α, the desired false alarm probability. For example, if we force the probability of false detection to be zero, we identify no shift.

```
postsignal(u,alpha=0,plot=FALSE)

## Call:
## mphase1(x = Student, plot = FALSE, post.signal = TRUE, isolated = TRUE,
##     step = TRUE, alpha = 0, gamma = 0.5, K = 7, lmin = 5, L = 1000,
##     seed = 11642257)
##
## p-value < 0.001
##
## Location Shifts:
## [1] "None"
```

3 The Ryan and Gravel Datasets

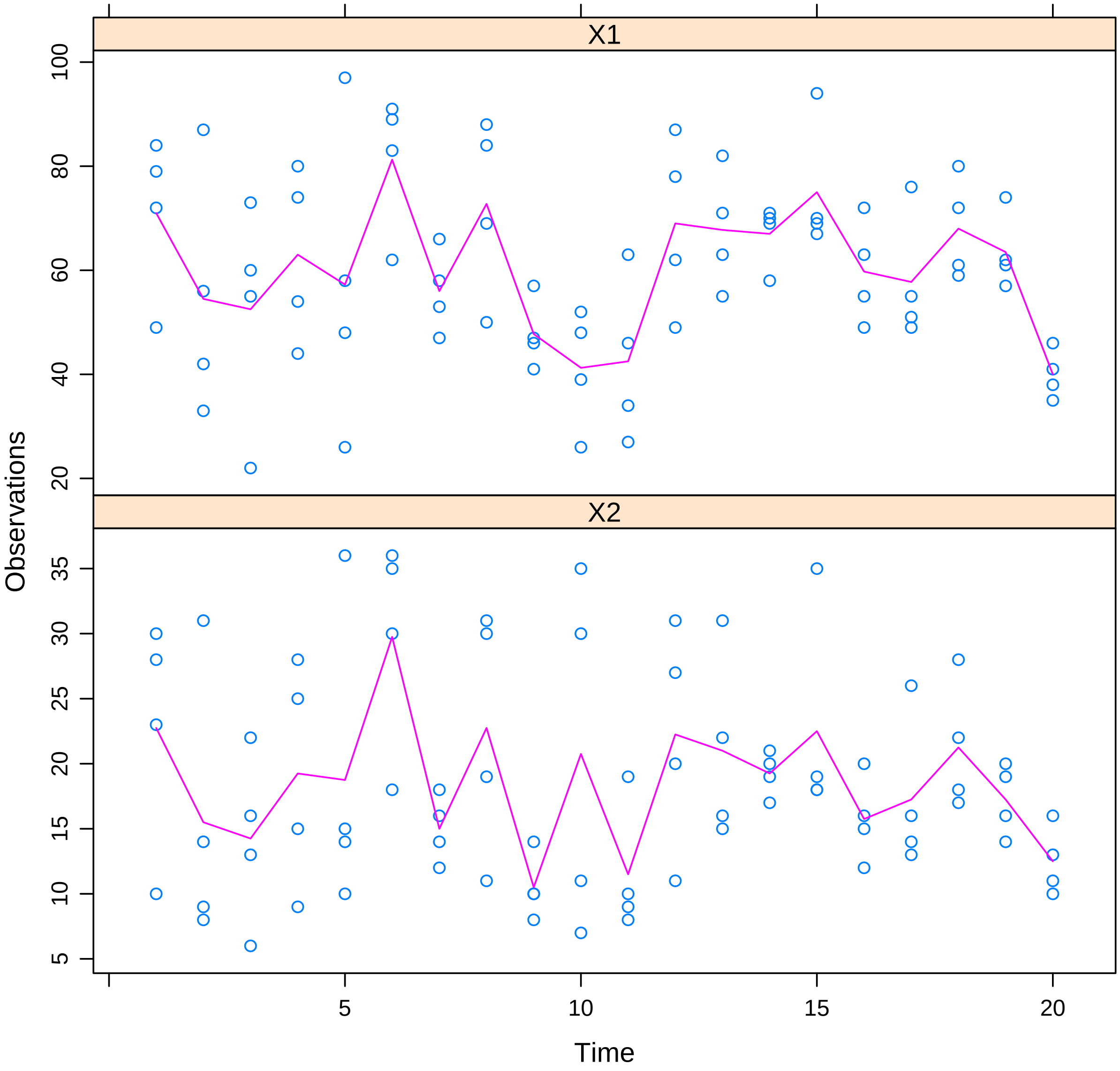
Package mphase1 includes the two datasets used in the paper. Hence, the two examples can be readily reproduced.

3.1 Phase I analysis of the Ryan dataset

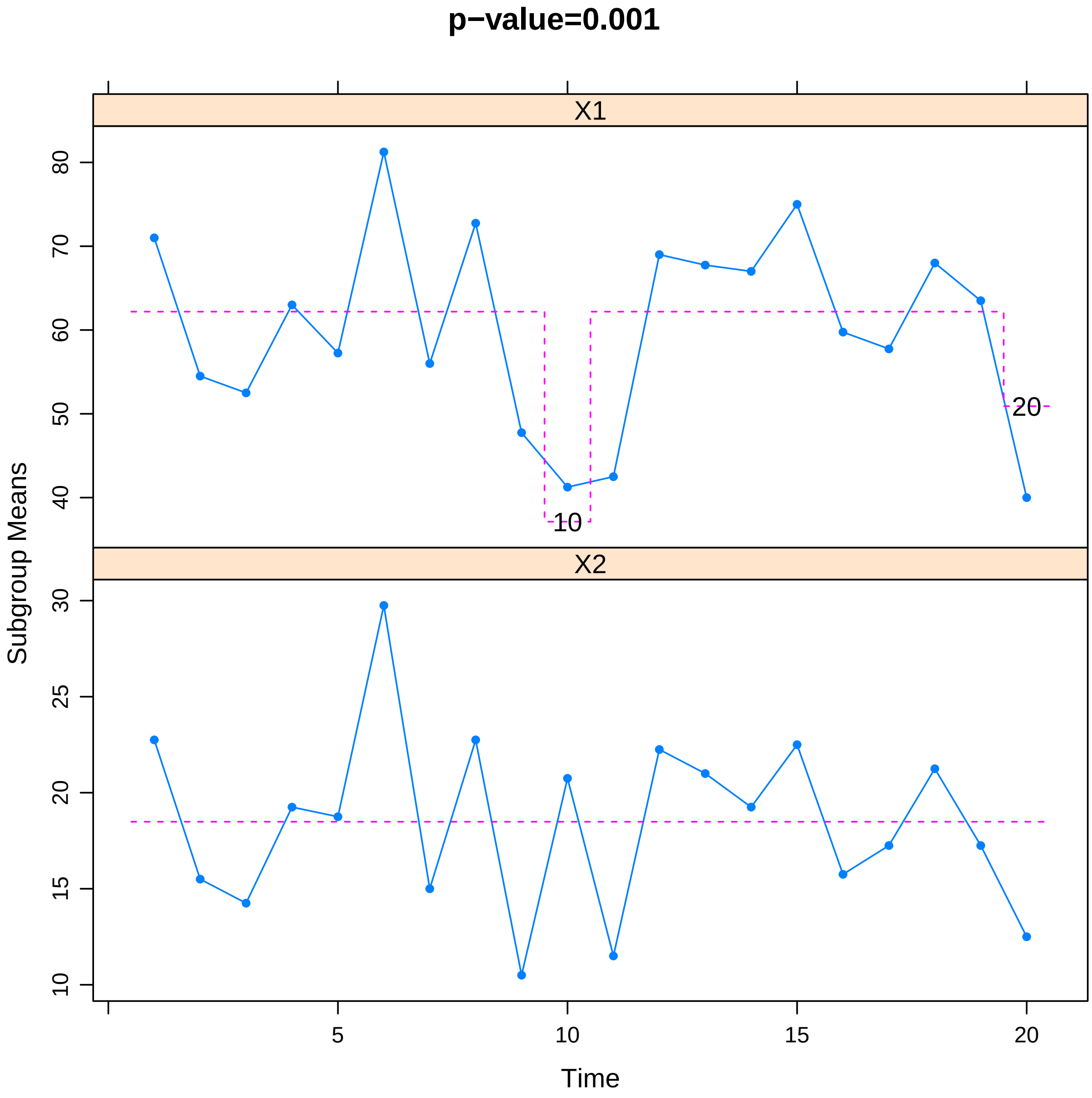
```
data(ryan)
dim(ryan)

## [1]  2  4 20

mphase1PlotData(ryan)
```



```
system.time(u <- mphase1(ryan))
```



```
##    user  system elapsed
##   0.227   0.000   0.226
```

```
u
```

```
## Call:
## mphase1(x = ryan, plot = TRUE, post.signal = TRUE, isolated = TRUE,
##     step = TRUE, alpha = 0.05, gamma = 0.5, K = 4, lmin = 5,
##     L = 1000, seed = 11642257)
```

```
##
## p-value = 0.001
##
## Location Shifts:
##       type time variables
## 1 Isolated   10         1
## 2 Isolated   20         1
```

It is interesting to observe that changing the value of γ does not change the results of the post-signal identification stage.

```
postsignal(u,gamma=0,plot=FALSE)

## Call:
## mphase1(x = ryan, plot = FALSE, post.signal = TRUE, isolated = TRUE,
##     step = TRUE, alpha = 0.05, gamma = 0, K = 4, lmin = 5, L = 1000,
##     seed = 11642257)
##
## p-value = 0.001
##
## Location Shifts:
##       type time variables
## 1 Isolated   10         1
## 2 Isolated   20         1

postsignal(u,gamma=1,plot=FALSE)

## Call:
## mphase1(x = ryan, plot = FALSE, post.signal = TRUE, isolated = TRUE,
##     step = TRUE, alpha = 0.05, gamma = 1, K = 4, lmin = 5, L = 1000,
##     seed = 11642257)
##
```

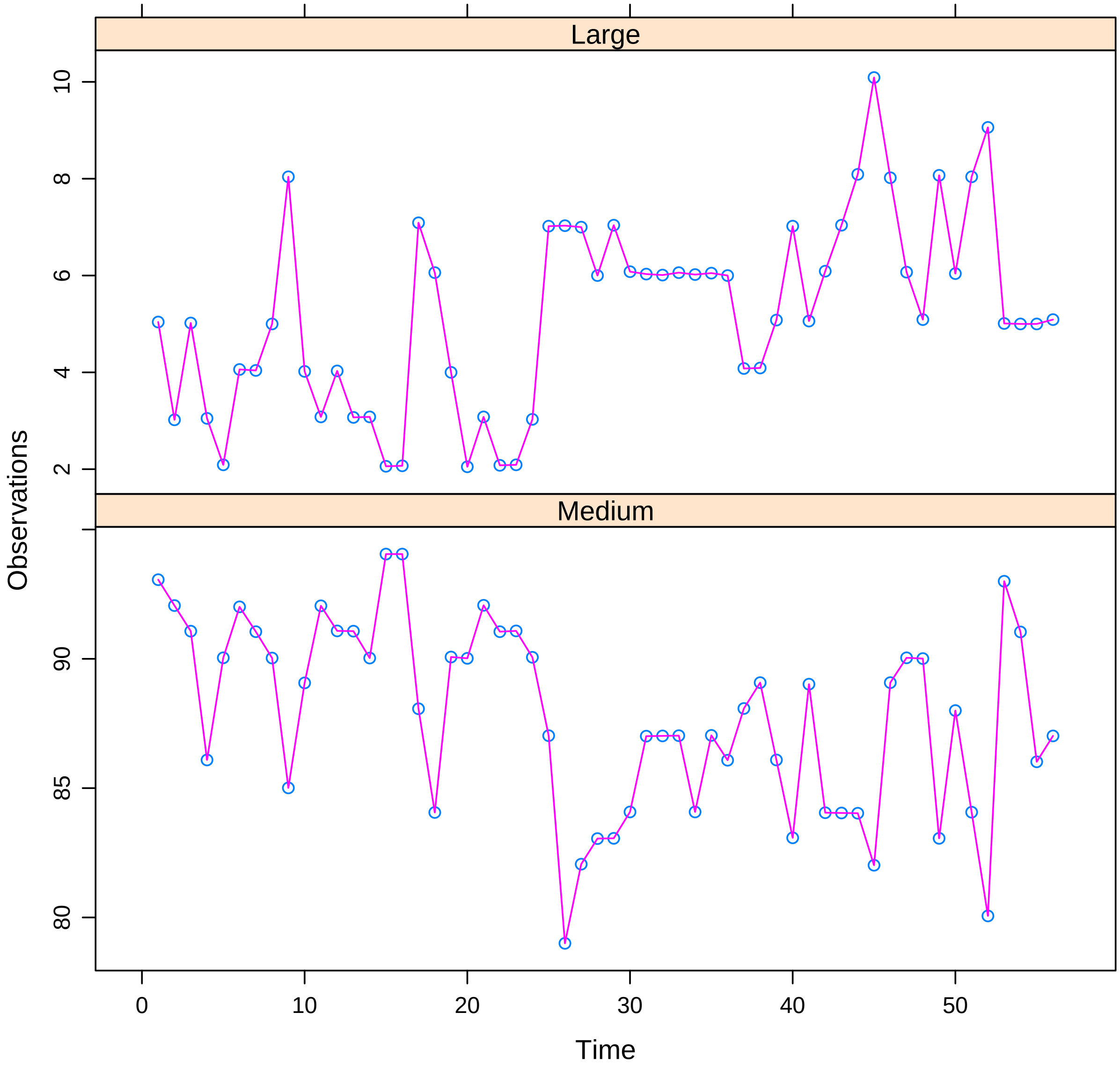
```
## p-value = 0.001
##
## Location Shifts:
##       type time variables
## 1 Isolated   10         1
## 2 Isolated   20         1
```

3.2 Phase I analysis of the gravel dataset

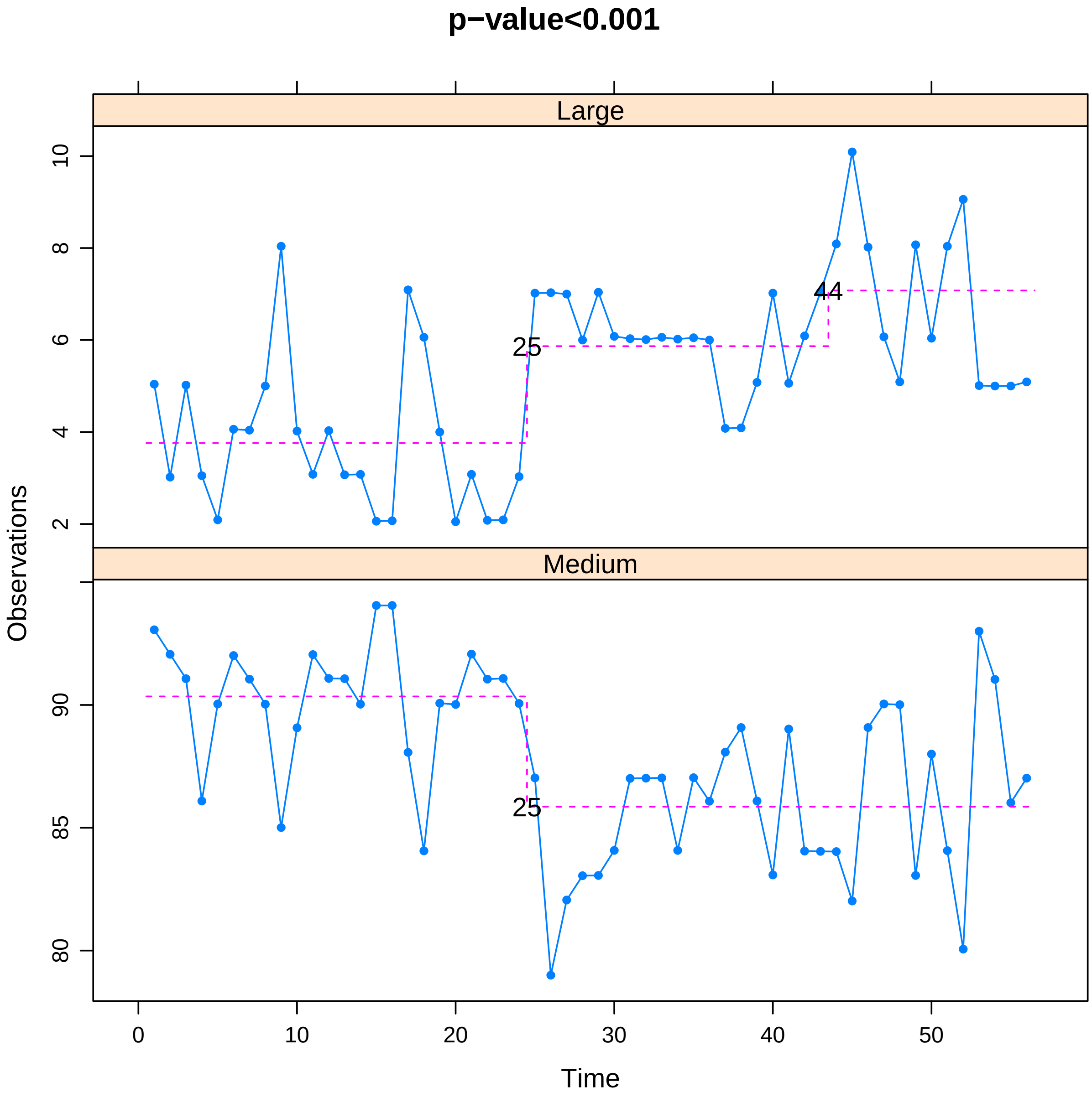
```
data(gravel)
dim(gravel)

## [1]  2  1 56

mphase1PlotData(gravel)
```



```
system.time(u <- mphase1(gravel))
```



```
##    user  system elapsed
##   0.213   0.000   0.213
```

```
u
```

```
## Call:
## mphase1(x = gravel, plot = TRUE, post.signal = TRUE, isolated = FALSE,
##     step = TRUE, alpha = 0.05, gamma = 0.5, K = 7, lmin = 5,
##     L = 1000, seed = 11642257)
```

```
##
## p-value < 0.001
##
## Location Shifts:
##   type time variables
## 1 Step   25       1,2
## 2 Step   44         1
```

No additional shift is identified when $\gamma = 0$.

```
postsignal(u,gamma=0,plot=FALSE)

## Call:
## mphase1(x = gravel, plot = FALSE, post.signal = TRUE, isolated = FALSE,
##     step = TRUE, alpha = 0.05, gamma = 0, K = 7, lmin = 5, L = 1000,
##     seed = 11642257)
##
## p-value < 0.001
##
## Location Shifts:
##   type time variables
## 1 Step   25       1,2
## 2 Step   44         1
```

However, if we set $\gamma = 1$, only the step shift at time 25 involving the first variable is identified.

```
postsignal(u,gamma=1,plot=FALSE)

## Call:
## mphase1(x = gravel, plot = FALSE, post.signal = TRUE, isolated = FALSE,
##     step = TRUE, alpha = 0.05, gamma = 1, K = 7, lmin = 5, L = 1000,
##     seed = 11642257)
##
## p-value < 0.001
##
## Location Shifts:
##   type time variables
## 1 Step   25         1
```

Hence, the additional shifts identified when $\gamma = 0.5$ (the default value) should be analyzed with particular attention.

A Data Simulation

The following code shows how the Student dataset has been simulated:

(a) First, we set the seed of the random number generator to make the example reproducible.

```
set.seed(1)
```

(b) Then, we simulate the in-control data from a multivariate Student's t distribution.

```
m <- 50
n <- 5
p <- 4
df <- 3
Sigma <- outer(1:p,1:p,function(i,j) 0.8^abs(i-j))
Sigma

##       [,1] [,2] [,3]  [,4]
## [1,] 1.000 0.80 0.64 0.512
## [2,] 0.800 1.00 0.80 0.640
## [3,] 0.640 0.80 1.00 0.800
## [4,] 0.512 0.64 0.80 1.000

xnorm <- crossprod(chol(Sigma),matrix(rnorm(p*n*m),p))
xchisq <- sqrt(rchisq(n*m,df)/(df-2))
x <- array(sweep(xnorm,2,xchisq,"/"),c(p,n,m))
dimnames(x)<-list(paste("X",1:4,sep=""),NULL,NULL)
```

(c) Finally, we add an isolated shift at time 10 (involving only the first variable) and a step shift starting at 31 (involving the third and fourth variables)

```
x[1,,10] <- x[1,,10]+1
x[3:4,,31:50] <- x[3:4,,31:50] + c(0.50,-0.25)
```

It is easy to verify that the simulated dataset is identical to that included in the package.

```
identical(x,Student)

## [1] TRUE
```