Outlier Detection in General Profiles Using Penalized Regression Method

Changliang Zou[1], Sheng-Tsaing Tseng[2] and Zhaojun Wang[1]
[1]*LPMC and School of Mathematical Sciences, Nankai University, PR China;*
[2]*Institute of Statistics, National Tsing Hua University, Taiwan;*

Abstract

Profile monitoring is a technique for checking the stability of functional relationships between a response variable and one or more explanatory variables over time. The presence of outliers has seriously adverse effects on the modeling, monitoring and diagnosis of profile data. This paper proposes a new outlier detection procedure from the viewpoint of penalized regression, aiming at identifying any abnormal profile observations from a baseline dataset. We treat profiles as high-dimension vectors and re-formulate the model into a specific regression model. We then apply a group-type regularization favoring a sparse vector of mean shift parameters. Using the classical hard-penalty yields a computationally efficient algorithm and delivers robust detection ability. By appropriately choosing the only one tuning parameter in our procedure, it enables us to control the type-I error. Simulation results show that the proposed method has outstanding performance in identifying outliers in various situations compared with other existing approaches. This methodology is also extended to the case that within-profile correlations exist.

Keywords: High-dimensional inference; Masking effect; Outlier detection; Profile monitoring; Sparsity; Statistical process control

1 Introduction

Because of recent progress in sensing and information technologies, automatic data acquisition has become the norm in various industries. Consequently, a large amount of quality-related data of certain processes have become available. Statistical process control (SPC) based on such data is an important component of process monitoring and control. In many applications, the quality of a process is characterized by the relationship between a response variable and one or more explanatory variables. A collection of data points of these variables can be observed at each sampling stage, which can be represented by a curve or profile. In some calibration applications, the profile can be described adequately by a linear regression model. In other applications, however, more flexible and complex models are necessary in order to describe the profiles properly. Extensive discussion of the related research problems can be found in Woodall (2007) and Noorossana et al. (2011).

In the SPC of profile problems, one of the most crucial steps is to identify any outlying profiles among a set of complex profiles and to remove them from the the reference dataset because the presence of outliers has seriously adverse effects on the modeling of functional curve and accordingly on the properties of control charts (Qiu et al. 2010). The outlier detection problem is a fundamental task in Phase-I analysis of profiles and has attracted certain attention in the literature. Among others, Jin and Shi (1999), Lada et al. (2002), Ding et al. (2006) and Paynabar and Jin (2011) investigated a general class of nonlinear profiles, using dimension-reduction techniques, wavelet transformations, independent component analysis and mixed-effect modelling. Mahmoud and Woodall (2004) and Mahmoud (2008) studied simple and multiple linear Phase-I profile monitoring respectively. Williams et al. (2007) suggested three general approaches to nonlinear profile monitoring in Phase-I analysis, based on Hotelling's T^2 statistics and non-linear regression. Colosimo and Pacella (2007) proposed methods for monitoring roundness profiles of manufactured items. Recently, Zou et al. (2010) extended the nonparametric smoothing based method suggested by Zou et al. (2008) to Phase I analysis of general profiles. However, their method relies on a basic assumption that the underlying functional relationship possesses certain smoothness, which is often invalid in applications. For instance, see stamping tonnage profile in Jin and Shi

(1999) and a semiconductor example in Section 4. Alternatively, Zhang and Albin (2009) proposed to treat profiles as vectors and applied a χ^2 control chart to identify outliers. Zhang and Albin (2009) argued that the χ^2 control chart is especially useful, and sometimes the only option, when profiles are highly complex. It is usually hard, if not impossible, to fit a regression function (parametric or nonparametric) to express the complex relationship between the response and explanatory variables. Comparing it with the existing non-linear regression method shows that the χ^2 chart has a better performance for complex profiles.

The χ^2 chart is simple and effective, however, when there are many outliers in baseline dataset, this method may fail to identify true outliers with larger probabilities. This tendency is not surprising since it suffers from the so-called "masking effect". When there are multiple outliers in the sample, estimates of parameters in the χ^2 chart would be contaminated to certain degree, and as a result, outlying profiles may not look like outliers. Therefore, multiple outliers may mask each other and go undetected. Some evidence can be found from the simulation results in Zhang and Albin (2009).

To this end, this paper proposes a new outlier detection procedure from the viewpoint of penalized regression. We start by assuming within-profile observations are independent and follow Zhang and Albin (2009) to treat profiles as high-dimension vectors. The profile model is re-formulated as a specific regression model. We then apply a group-type regularization favoring a sparse vector of mean shift parameters. Using the classical hard-penalty yields a computationally efficient algorithm and delivers a robust detection ability. By appropriately choosing the only one tuning parameter in our procedure, it enables us to control the type-I error. Simulation results show that the proposed method has outstanding performance in identifying outliers in various situations. Compared with the χ^2 chart, the proposed method misidentifies much smaller fractions of outlier profiles as non-outliers when the number of outlying profiles is large. Moreover, in practice, within-profile data are usually spatially or serially correlated. For instance, within-profile data of the vertical-density profiles (VDPs) considered by Walker and Wright (2002) are spatially correlated, since the density measurements are taken in intervals that are close to each other along the vertical depth of a particle board. As another example, within-profile data in the AEC example considered by Qiu et al. (2010) exhibit obvious serial correlation over time. This motivates us to extend the proposed

methodology to the case that within-profile correlations exist.

Our proposed procedure is described in detail in Section 2. Its numerical performance is investigated in Section 3. In Section 4, we apply this method to a dataset from a semiconductor manufacturing process. Several remarks conclude the article in Section 5. Some technical details are provided in an appendix.

2 Methodology

2.1 Problems and existing works

Suppose we have a baseline profile dataset which consists of m profiles. The explanatory variable takes a set of n fixed values $\mathbf{X} = \{x_1, x_2, \ldots, x_n\}$. Assume the m profile observations $\{(y_{i1}, x_1), \ldots, (y_{in}, x_n)\}_{i=1}^m$ are collected from the following model

$$y_{ij} = \begin{cases} g_0(x_j) + \sigma_0 \varepsilon_{ij}, & j = 1, \ldots, n, \quad \text{if} \quad i \notin \mathcal{O} \\ g_i(x_j) + \sigma_i \varepsilon_{ij}, & j = 1, \ldots, n, \quad \text{if} \quad i \in \mathcal{O}, \end{cases} \tag{1}$$

where $\mathcal{O}$ is the outlying profile set (a subset of $\{1, \ldots, m\}$) and y_{ij} is the j-th response variable in the i-th profile. Here g_0 and $g_i, i \in \mathcal{O}$ are the unknown in-control (IC) and out-of-control (OOC; outlying) regression functions, σ_0 and σ_i are the unknown IC and OOC profile variations, and ε_{ij} are independently and identically distributed (i.i.d.) as $N(0, 1)$. Any profile with $g_i \neq g_0$ and/or $\sigma_i^2 > \sigma_0^2$ is considered as an outlier. We assume that the explanatory variables are fixed for different j. This assumption is often valid in calibration applications of manufacturing industry, and is also consistent with the existing literature on profile monitoring. Possible extensions of the proposed method to cases in which the design points are random or unbalanced, are briefly discussed in Section 5.

When g_0 and $g_i, i \in \mathcal{O}$ are complex, it is difficult to fit explicit expressions for them. To this end, Zhang and Albin (2009) took the m profile responses as vectors in n-dimension space, among which $m_o \equiv |\mathcal{O}|$ profiles are outliers. Write $\mathbf{y}_i = (y_{i1}, \ldots, y_{in})^T$. They proposed to use the following χ^2-type statistic for outlier detection

$$\Delta_i = \sum_{j=1}^{n} \frac{(y_{ij} - \hat{y}_j)^2}{((m-1)/m)\tilde{\sigma}_0^2}, \tag{2}$$

where $\hat{y}_j$ and $\tilde{\sigma}_0^2$ are the robust estimators for $E(y_{ij}), i \notin \mathcal{O}$ and σ_0^2 respectively. $\tilde{\boldsymbol{\mu}}_0 = (\hat{y}_1, \ldots, \hat{y}_n)^T$ is defined as the median of the vectors in profile baseline data, say,

$$\hat{y}_j = \text{Median}\{y_{1j}, \ldots, y_{mj}\}.$$

$\tilde{\sigma}_0^2$ is the median of the $m(m-1)/2$ pair-wise estimates, $\sigma_{(i,k)}^2$, defined as

$$\sigma_{(i,k)}^2 = \frac{1}{2n}\sum_{j=1}^{n}(y_{ij} - y_{kj})^2, \ i = 1, \ldots, m-1, \ k = i+1, \ldots, m.$$

The reason for using the estimators $\hat{y}_j$ and $\tilde{\sigma}_0^2$ rather than the regular sample mean and variance is that the median is usually more robust to outlier as well known. Profile i is considered an outlier if $\Delta_i > \chi_{\alpha,n}^2$, where $\chi_{\alpha,n}^2$ is the upper α percentile of the chi-square distribution with n degrees of freedom.

2.2 A framework for detecting outliers based on penalty function

Here we follow Zhang and Albin (2009) to consider the m profile responses as n-dimensional vectors. We firstly focus on identifying outliers with $g_i \neq g_0$, but later we will show our suggested procedure is also effective for the case $\sigma_i^2 \neq \sigma_0^2$. Denote

$$\begin{aligned}
\boldsymbol{\mu}_0 &= (g_0(x_1), \ldots, g_0(x_n))^T, \\
\boldsymbol{\gamma}_i &= (g_i(x_1) - g_0(x_1), \ldots, g_i(x_n) - g_0(x_n))^T, \\
\boldsymbol{\varepsilon}_i &= (\varepsilon_{i1}, \ldots, \varepsilon_{in})^T.
\end{aligned}$$

Note that the model (1) can be reformulated as the following form

$$\mathbf{Y} = \boldsymbol{\theta} + \boldsymbol{\gamma} + \sigma_0\boldsymbol{\varepsilon}, \tag{3}$$

where $\mathbf{Y} = (\mathbf{y}_1^T, \ldots, \mathbf{y}_m^T)^T$, $\boldsymbol{\theta} = (\boldsymbol{\mu}_0^T, \ldots, \boldsymbol{\mu}_0^T)^T$, $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1^T, \ldots, \boldsymbol{\gamma}_m^T)^T$, and $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}_1^T, \ldots, \boldsymbol{\varepsilon}_m^T)^T$. The χ^2 test statistic Δ_i for whether the i-th profile is an outlier is essentially the same as testing whether the parameter $\boldsymbol{\gamma}_i$ is $\mathbf{0}$ in model (3), where $\mathbf{0}$ is the zero vector. Because we do not know which observations might be outliers, the goal is to determine which $\boldsymbol{\gamma}_k$ of $\boldsymbol{\gamma}$ are not zero vector. A natural way to accomplish this task is to obtain an appropriate estimate

of $\boldsymbol{\gamma}$ and to find out which components are non-zero. However, usual estimations suffer from inadequacy because those components usually take non-zero values.

Hawkins (1980) defined an outlier as an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism. Thus, in outlier detection or Phase I study, we often assume that only a small portion of profiles are outliers, i.e., the so-called sparsity characteristic (Wang and Jiang 2009; Zou and Qiu 2009). Accordingly, the outlier detection is essentially analogous to the variable or model selection problem, that is to say, one wishes to select those parameters ($\boldsymbol{\gamma}_k$'s) that deviate significantly from zero. In this spirit, stepwise or all subset selection procedures along with a model selection criterion (e.g., AIC or BIC) may be employed. Such model selection procedures are practically useful; but they have several limitations, including lack of stability and extensive computation (cf., Tibshirani 1996). To overcome these limitations, some penalized least squares methods, such as the LASSO and SCAD, become quite popular over the last two decades (see Zou et al. 2011 for related discission in the context of SPC). To this end, in what follows, we will derive our procedure from the viewpoint of penalized regression.

It is firstly worth noting that although it can be regarded as a specific linear regression model with both the sample size and dimension being mn, the variable selection problem in model (3) is not a typical one because we want to identify which sub-vector $\boldsymbol{\gamma}_i$ of $\boldsymbol{\gamma}$ is not $\mathbf{0}$ rather than which components of $\boldsymbol{\gamma}$ are non-zero. By taking this issue into account, we re-express (3) as the following model

$$\boldsymbol{Y} = \boldsymbol{\theta} + \sum_{i=1}^{m} \boldsymbol{X}_i \boldsymbol{\gamma}_i + \sigma_0 \boldsymbol{\varepsilon}, \tag{4}$$

where

$$\boldsymbol{X}_i = (\underbrace{\mathbf{0}_n, \ldots, \mathbf{0}_n}_{1,\ldots,i-1}, \underbrace{\mathbf{I}_n}_{i}, \underbrace{\mathbf{0}_n, \ldots, \mathbf{0}_n}_{i+1,\ldots,m})^T$$

is a $mn \times n$ pseudo-covariate matrix, and $\mathbf{0}_n$ and $\mathbf{I}_n$ are the n-dimensional zero and identity matrices respectively. (4) is a specific regression model with m factors ($\boldsymbol{X}_i$) considered by Yuan and Lin (2006). Our task now amounts to deciding whether to set the vector $\boldsymbol{\gamma}_i$ to zero vectors for each i. Similar to Yuan and Lin (2006), the assumed sparsity of $\boldsymbol{\gamma}$ motivates

using a grouped penalized regression to minimize

$$f(\boldsymbol{\mu}_0, \boldsymbol{\gamma}; \lambda) \equiv \frac{1}{2}||\boldsymbol{Y} - \boldsymbol{\theta} - \sum_{i=1}^{m} \boldsymbol{X}_i \boldsymbol{\gamma}_i||^2 + \sum_{i=1}^{m} P_\lambda(||\boldsymbol{\gamma}_i||), \tag{5}$$

where $||\cdot||$ denotes the Euclidean norm and $P_\lambda(\cdot)$ is a penalty function with a tuning parameter $\lambda > 0$. This group penalized method encourages sparsity at the factor level which well serves our purpose of selecting $\boldsymbol{\gamma}_i$. Given an appropriate function $P_\lambda(\cdot)$ and a well-chosen λ, the i-th profile is identified as an outlier if the minimizer of $\boldsymbol{\gamma}_i$ is not a zero vector.

By describing the connection between outlier detection based on penalized regression and M-estimator, She and Owen (2011) argued that the usual L_1 penalty (Tibshirani 1996) fails to deliver a robust outlier detection result in regression diagnostic. Following their recommendation, we consider the following hard penalty (Donoho and Johnstone 1994) function

$$P_\lambda(x) = I_{\{|x|<\lambda\}}(\lambda|x| - \frac{x^2}{2}) + I_{\{|x|\geq\lambda\}}\frac{\lambda^2}{2}, \tag{6}$$

where $I_{\{\cdot\}}$ is the indicator function. Using the above $P_\lambda(x)$ would not only result in a simple and intuitively meaningful detection procedure as shown in Section 2.3 but also could correctly identify outliers on some extremely hard test problems as shown in Section 3.

As a convention, λ can be chosen by using cross-validation, Akaike information criterion (AIC), Schwarz's Bayesian information criterion (BIC), or other model selection criteria. In the literature, it is well demonstrated that AIC tends to select the model with the optimal predication performance, while BIC tends to identify the true sparse model well if the true model is included in the candidate set (Yang 2005). As we want to identify the non-zero components in $\boldsymbol{\gamma}$ rather than obtaining an estimate, BIC is more relevant and appealing here. Please refer to Zou et al. (2011) and She and Owen (2011) for related formulations and discussions. However, in the context of SPC, it is more reasonable and typical to use type I error (false alarm rate) as a criterion for measuring the performance of procedures (Montgomery 2001). The type-I error is defined as either the percentage of non-outlier profiles identified as outliers (Zhang and Albin 2009) or the probability of identifying at least one outlier under the null hypothesis of no outliers (Mahmoud and Woodall 2004). The former focuses more on the accuracy of outlier isolation while the latter cares mainly on the overall testing power. Thus, it is more appealing to consider the one used by Zhang and

Albin (2009) here because our aim is to identify the true outliers as accurately as possible. In the next section, λ will be determined in terms of controlling the type-I error.

2.3 An efficient and practical procedure

In what follows, our discussion focuses on using $P_\lambda(x)$ in (6). Given a λ and $\boldsymbol{\mu}_0$, the solution of the objective function (5) is motivated by the following proposition.

Proposition 1 *A necessary condition for* $\boldsymbol{\gamma}$ *to be a solution to* $\min f(\boldsymbol{\mu}_0, \boldsymbol{\gamma}; \lambda)$ *given* $\boldsymbol{\mu}_0$ *and* λ *is*

$$\begin{aligned} \boldsymbol{X}_i^T(\boldsymbol{Y} - \boldsymbol{\mu} - \sum_{i=1}^m \boldsymbol{X}_i\boldsymbol{\gamma}_i) = \mathbf{0},\ \forall ||\boldsymbol{\gamma}_i|| \geq \lambda, \\ -\boldsymbol{X}_i^T(\boldsymbol{Y} - \boldsymbol{\mu} - \sum_{i=1}^m \boldsymbol{X}_i\boldsymbol{\gamma}_i) + (\lambda\boldsymbol{\gamma}_i/||\boldsymbol{\gamma}_i|| - \boldsymbol{\gamma}_i) = \mathbf{0},\ \forall\ 0 \neq ||\boldsymbol{\gamma}_i|| < \lambda, \\ ||\boldsymbol{X}_i^T(\boldsymbol{Y} - \boldsymbol{\mu})|| \leq \lambda,\ ||\boldsymbol{\gamma}_i|| = 0. \end{aligned} \tag{7}$$

This result is a direct consequence of the Karush-Kuhn-Tucker conditions. Note that ${\boldsymbol{X}_i}^T\boldsymbol{X}_i = \mathbf{I}_n$ and ${\boldsymbol{X}_i}^T\boldsymbol{X}_j = \mathbf{0}_n$. It can be easily verified that the solution to the expressions above is (see the proof of Proposition 2)

$$\hat{\boldsymbol{\gamma}}_i = I_{\{||\mathbf{y}_i - \boldsymbol{\mu}_0|| > \lambda\}}(\mathbf{y}_i - \boldsymbol{\mu}_0). \tag{8}$$

In other words, the solution to (5) given $\boldsymbol{\mu}_0$ and λ is simply zero or $(\mathbf{y}_i - \boldsymbol{\mu}_0)$ depending on whether $||\mathbf{y}_i - \boldsymbol{\mu}_0||$ is larger than λ, which coincides with χ^2-chart mentioned above in certain degree. Now, given $\hat{\boldsymbol{\gamma}}_i$, the objective function (in $\boldsymbol{\mu}_0$) becomes

$$\frac{1}{2}\sum_{\{i:\hat{\boldsymbol{\gamma}}_i = \mathbf{0}\}} ||\mathbf{y}_i - \boldsymbol{\mu}_0||^2 + \sum_{i=1}^m P_\lambda(||\hat{\boldsymbol{\gamma}}_i||), \tag{9}$$

whose minimizer is simply $\hat{\boldsymbol{\mu}}_0 = q^{-1}\sum_{\{i:\hat{\boldsymbol{\gamma}}_i = \mathbf{0}\}} \mathbf{y}_i$ with $q = \sum I_{\{\hat{\boldsymbol{\gamma}}_i = \mathbf{0}\}}$. This together with (8) suggest the following iterative algorithm for the proposed procedure which is called penalized profile outlier detection (PPOD).

Algorithm 1 (PPOD)

1. Given $\boldsymbol{Y}$. Specify $\lambda > 0$, $\epsilon > 0$ and an initiate robust estimator $\hat{\boldsymbol{\mu}}_0^{(0)}$;

2. Obtain $\hat{\boldsymbol{\gamma}}_i^{(k)} = I_{\{||\mathbf{y}_i - \hat{\boldsymbol{\mu}}_0^{(k-1)}|| > \lambda\}}(\mathbf{y}_i - \hat{\boldsymbol{\mu}}_0^{(k-1)})$;

3. Compute $\hat{\boldsymbol{\mu}}_0^{(k)} = \frac{1}{q^{(k)}} \sum_{\{i:\hat{\boldsymbol{\gamma}}_i^{(k)} = \mathbf{0}\}} \mathbf{y}_i$ with $q^{(k)} = \sum I_{\{\hat{\boldsymbol{\gamma}}_i^{(k)} = \mathbf{0}\}}$;

4. Repeat 2-3 until $\sum_{i=1}^{m} ||\hat{\boldsymbol{\gamma}}_i^{(k)} - \hat{\boldsymbol{\gamma}}_i^{(k-1)}|| < \epsilon$.

To initialize the algorithm, we must specify $\hat{\boldsymbol{\mu}}_0^{(0)}$. Empirically, the starting point is not crucial and the median of $\boldsymbol{Y}$, $\tilde{\boldsymbol{\mu}}_0$, is recommended. The algorithm is found to be very stable and usually reaches a reasonable convergence tolerance within a few iterations. The following proposition guarantees the PPOD procedure converges.

Proposition 2 *The PPOD iteration sequence* $(\hat{\boldsymbol{\mu}}_0^{(k)}, \hat{\boldsymbol{\gamma}}^{(k)})$ *satisfies*

$$f(\hat{\boldsymbol{\mu}}_0^{(k)}, \hat{\boldsymbol{\gamma}}^{(k)}; \lambda) \geq f(\hat{\boldsymbol{\mu}}_0^{(k)}, \hat{\boldsymbol{\gamma}}^{(k+1)}; \lambda) \geq f(\hat{\boldsymbol{\mu}}_0^{(k+1)}, \hat{\boldsymbol{\gamma}}^{(k+1)}; \lambda), \text{ for any } k \geq 0. \tag{10}$$

Any limit point of $(\hat{\boldsymbol{\mu}}_0^{(k)}, \hat{\boldsymbol{\gamma}}^{(k)})$ must be a stationary point of (5). It is worth pointing out that this suggested iterative procedure is essentially a coordinate descent algorithm (Friedman et al. 2010). See also Breheny and Huang (2011) for some discussions on the convergence of the coordinate descent algorithm in solving non-convex penalized regression problems. It should be emphasized that although designing for detecting the outliers with $g_i \neq g_0$, the PPOD procedure is effective for the case $\sigma_i^2 \geq \sigma_0^2$ in which $||\mathbf{y}_i - \hat{\boldsymbol{\mu}}_0^{(k-1)}||$ would be large as well. See some numerical evidence in Section 3.

Before proceeding, the tuning parameter λ needs to be specified. To control the type-I error (the percentage of non-outlier profiles identified as outliers), a direct way is to find λ so that

$$\Pr(||\mathbf{y}_i - \hat{\boldsymbol{\mu}}_0|| > \lambda \mid \mathbf{y}_i \text{ is non-outlier}) \approx \alpha.$$

Similar to the arguments in Zhang and Albin (2009), we can conclude that

$$\frac{||\mathbf{y}_i - \hat{\boldsymbol{\mu}}_0||^2}{\frac{q-1}{q}\sigma_0^2} \approx \chi_n^2.$$

Accordingly, α can be chosen as

$$\lambda = \left[\frac{q-1}{q}\chi^2_{n,\alpha}\right]^{1/2}\sigma_0$$

As σ_0^2 is unknown, it could be estimated (updated) in the iterative procedure with $\tilde{\sigma}_0^2$ being an initiate one. This suggests the following Algorithm 2, denoted as (PPOD-R), which is a revised (practical) version of Algorithm 1 by updating the estimate of σ_0^2 and specifying λ iteratively.

Algorithm 2 (PPOD-R)

1. Given $\boldsymbol{Y}$. Specify $\lambda > 0$, $\epsilon > 0$. Let $\hat{\boldsymbol{\mu}}_0^{(0)} = \tilde{\boldsymbol{\mu}}_0$, $\hat{\sigma}_0^{(0)} = \tilde{\sigma}_0$, and $q^{(0)} = m$;
2. Let $\lambda^{(k-1)} = \left[\frac{q^{(k-1)}-1}{q^{(k-1)}n}\chi^2_{n,\alpha}\right]^{1/2}\hat{\sigma}_0^{(k-1)}$. Obtain $\hat{\boldsymbol{\gamma}}_i^{(k)} = I_{\{||\mathbf{y}_i - \hat{\boldsymbol{\mu}}_0^{(k-1)}|| > \lambda^{(k-1)}\}}(\mathbf{y}_i - \hat{\boldsymbol{\mu}}_0^{(k-1)})$;
3. Compute $\hat{\boldsymbol{\mu}}_0^{(k)} = \frac{1}{q^{(k)}}\sum_{\{i:\hat{\boldsymbol{\gamma}}_i^{(k)}=\mathbf{0}\}}\mathbf{y}_i$ with $q^{(k)} = \sum I_{\{\hat{\boldsymbol{\gamma}}_i^{(k)}=\mathbf{0}\}}$; Update the estimate of σ_0^2 by
$$\hat{\sigma}_0^{2(k)} = \frac{1}{n}\sum_{j=1}^{n}\frac{1}{q^{(k)}-1}\sum_{\{i:\hat{\boldsymbol{\gamma}}_i^{(k)}=\mathbf{0}\}}(y_{ij} - \hat{\mu}_{0j}^{(k)})^2,$$
where $\hat{\mu}_{0j}^{(k)}$ denotes the j-th component of $\hat{\boldsymbol{\mu}}_0^{(k)}$.
4. Repeat 2-3 until $\sum_{i=1}^{m}||\hat{\boldsymbol{\gamma}}_i^{(k)} - \hat{\boldsymbol{\gamma}}_i^{(k-1)}|| < \epsilon$.

PPOD-R is computationally efficient and usually requires no more than a dozen of iterations to achieve convergence criterion with $\epsilon = 10^{-3}$. Certainly, due to involve estimating λ, there is no theoretical justification on its convergence. However, we never occur any non-convergence in all our simulation studies and real-data analysis with the suggested initial values. It is easy to implement because the iteration does not involve complicated operations like matrix inversion. Also note that the dominant computational cost in each iteration is essentially similar to that of Zhang and Albin's (2009) χ^2-chart. In other words, the total cost of PPOD-R is several times that of the χ^2-chart. However, as we mentioned, the χ^2-chart is more prone to masking effects (see also simulation results in Section 3). In contrast, PPOD-R is derived from a well-formulated model and optimal criterion. In each iteration,

potential outliers have been removed from estimation by a large amount and thus the updated (refined) estimates would be more robust. As a consequence, PPOD-R would be more resistant to masking effects. In spite of more computationally extensive, PPOD-R is still an acceptable candidate, because we consider speed to be secondary compared with robustness as masking causes much harm in Phase I analysis.

Remark 1 The main idea of PPOD-R also differs from the so-called retrospective Phase-I analysis in univariate SPC practice (Montgomery 2001). In retrospective analysis, a control chart is established to find some outlying observations in a baseline dataset. Then, those observations identified as outliers are removed from the baseline dataset and the control chart is is re-designed based on the new baseline dataset. The procedure is repeated until all the observations in the baseline dataset are classified as in-control (fall within the control limits). That is, in retrospective analysis an observation cannot be involved in the analysis (model estimation) once it is identified as an outlier, while in PPOD-R the observations with $\boldsymbol{\gamma}_i \neq \mathbf{0}$ in the k-th iteration may still be useful in the $(k+1)$-th iteration. In other words, the difference between PPOD-R and retrospective Phase-I analysis is analogous to that between the stepwise and backward regression. The benefit of using PPOD-R partly lies that it may be more robust to swamping. In swamping, the effect of outliers is to make the charting statistic large for a nonoutlying case i. Swamping could lead one to delete good observations, and correspondingly the type-I error would be increased. It becomes more serious in the presence of multiple outliers, and is often a matter of lost efficiency. Of course, it should be also pointed out that if the hard penalty is replaced by other appropriate penalty functions, the resulting penalized-based profile outlier detection procedure would be much more different from classical methods used in SPC. The PPOD is advocated here because of its intuitive explanation, fast computation and good detection ability. Similar procedures with other penalty functions definitely deserve future research.

2.4 Extensions to cases that within-profile correlations exist

One limit of the proposed PPOD-R method is that it is based on the assumption of independence among noise terms. Although this assumption is reasonable in many applications and commonly used in the literature, it might still be violated in some applications. See Qiu

et al. (2010) and the references therein for discussions and examples. The PPOD-R may give higher type-I and type-II errors when within-profile correlations are strong. One way to overcome this issue is to replace the Euclidean distance used in (5) with the Mahalanobis distance by plugging a robust estimator of covariance matrix. Although standardizing by the covariance brings benefits for data with a fixed dimension, it becomes a liability for high dimensional data (when n is large). In particular, the sample covariance matrix may not converge to the population covariance when n and m are of the same order (Bai and Yin 1993). Even worse, in many applications, profile may have a huge number of fixed values of the explanatory variable but we only have a few profiles in a baseline dataset. Since the estimator of covariance matrix would often not be invertible when the dimension n is larger than m, the corresponding PPOD-R cannot be used.

Therefore, we still use the proposed PPOD method (5) based on the Euclidean distance but suggest some modifications to account for within-profile correlations. Our main idea is to adjust the value of λ to make the type-I error acceptable. Assume $\varepsilon_i \overset{\text{iid}}{\sim} N_n(\mathbf{0}, \mathbf{\Sigma})$, where $\mathbf{\Sigma}$ is a positive correlation matrix. Let $\eta_1 \leq \eta_2 \leq \cdots \leq \eta_n$ be the eigenvalues of $\mathbf{\Sigma}$. The following result establishes the asymptotic null distribution of $||\mathbf{y}_i - \hat{\boldsymbol{\mu}}_0||^2$.

Proposition 3 *Assume* $\mathbf{y}_i = \boldsymbol{\mu}_0 + \boldsymbol{\varepsilon}_i$. *Under the condition that* $\eta_n^2/\mathrm{tr}(\mathbf{\Sigma}^2) \to 0$ *as* $n \to \infty$, *we have*

$$\frac{||\mathbf{y}_i - \hat{\boldsymbol{\mu}}_0||^2 - n\sigma_0^2}{\sqrt{2\mathrm{tr}(\mathbf{\Sigma}^2)}} \xrightarrow{d} N(0,1).$$

The condition imposed here is quite mild. Note that if all eigenvalues are bounded, then the condition is trivially true. The condition is also valid for the most common case that $\mathbf{\Sigma} = \mathbf{I}_n$. By this result, λ can be modified as

$$\lambda = \left[n + z_\alpha \sqrt{2\mathrm{tr}(\mathbf{\Sigma}^2)} \right]^{1/2} \sigma_0, \tag{11}$$

where z_α is the upper α quantile of standard normal distribution. Our simulation results show that such choice of λ works reasonably well in many cases. The reason for attaining this in the face of high data dimension is because the statistic $||\mathbf{y}_i - \hat{\boldsymbol{\mu}}_0||^2$ is univariate, of which limiting distribution depends only on univariate quantities (σ_0^2 and $\mathrm{tr}(\mathbf{\Sigma}^2)$), despite

the hypothesis is of high dimensional. Thus, it is possible to use asymptotical normality for calibration if proper estimates of unknown quantities are available.

Since profile baseline data may contain multiple outliers, we need to derive robust estimators of $\mathrm{tr}(\boldsymbol{\Sigma}^2)$ which is resistent to the presence of outliers. To remove the adverse effect of outliers, we suggest to implement PPOD-R in Algorithm 2 firstly and then use a revised PPOD-R with adjusted λ again to obtain the final diagnosis result. Denote the corresponding identified outlier subset as $\mathcal{O}_1$, the final estimates of $\boldsymbol{\mu}_0$ and σ_0^2 as $\hat{\boldsymbol{\mu}}_{0c}^{(0)}$ and $\hat{\boldsymbol{\sigma}}_{0c}^{2(0)}$ respectively after a preliminary PPOD-R is completed. Let $q_c^{(0)} = m - |\mathcal{O}_1|$. Similar to the proposal in Chen and Qin (2010), we use the following estimator of $\mathrm{tr}(\boldsymbol{\Sigma}^2)$

$$\widehat{\mathrm{tr}(\boldsymbol{\Sigma}^2)} = \frac{2}{q_c^{(0)}(q_c^{(0)}-1)\hat{\boldsymbol{\sigma}}_{0c}^{4(0)}} \sum_{l\notin\mathcal{O}_1} \sum_{k\notin\mathcal{O}_1, k>l} \left[\left(\mathbf{y}_l - \frac{\sum_{i\notin\mathcal{O}_1, i\neq l,k} \mathbf{y}_i}{q_c^{(0)}-2} \right)^T \left(\mathbf{y}_k - \frac{\sum_{i\notin\mathcal{O}_1, i\neq l,k} \mathbf{y}_i}{q_c^{(0)}-2} \right) \right]^2 .$$

This estimator is similar to the idea of cross-validation, in the sense that when we construct the deviations of $\mathbf{y}_l$ and $\mathbf{y}_k$ from the sample mean, both $\mathbf{y}_l$ and $\mathbf{y}_k$ are excluded from the sample mean calculation. By doing so, the above estimator would be more accurate than traditional sample estimators in high-dimensional cases as argued by Chen and Qin (2010). Theorem 2 of Chen and Qin (2010) reveals the following consistency of $\widehat{\mathrm{tr}(\boldsymbol{\Sigma}^2)}$ under certain conditions

$$\widehat{\mathrm{tr}(\boldsymbol{\Sigma}^2)}/\mathrm{tr}(\boldsymbol{\Sigma}^2) \xrightarrow{p} 1, \text{ as } m \to \infty, \ n \to \infty.$$

Thus, by Slutsky's theorem, plugging this estimate into (11) results in an asymptotically correct λ which is able to control the type-I error of the proposed procedure. To alleviate the computational effort, we do not update this $\widehat{\mathrm{tr}(\boldsymbol{\Sigma}^2)}$ any more in the following steps. Now, with initial values obtained from PPOD-R, we can implement a new PPOD-R again with new values of λ. The whole procedure, called PPOD-C is summarized in Algorithm 3.

Algorithm 3 (PPOD-C)

1. Implement PPOD-R and obtain $\hat{\boldsymbol{\mu}}_{0c}^{(0)}$, $\hat{\boldsymbol{\sigma}}_{0c}^{2(0)}$ and $q_c^{(0)} = m - |\mathcal{O}_1|$.
2. Compute $\widehat{\mathrm{tr}(\boldsymbol{\Sigma}^2)}$ and set

$$\lambda^{(k-1)} = \left[n + z_\alpha \sqrt{2\widehat{\mathrm{tr}(\boldsymbol{\Sigma}^2)}} \right]^{1/2} \hat{\sigma}_{0c}^{(k-1)}. \tag{12}$$

3. Execute the steps 2 and 3 in PPOD-R with updated $\lambda^{(k-1)}$ in (12) until convergence.

We will see in Section 3 that PPOD-C is able to largely reduce type-I errors in comparison with PPOD-R when within-profile correlations cannot be ignored.

3 Numerical comparison

To see the performance of PPOD-R and PPOD-C, we have conducted many simulation studies. Some of the results are reported here. Firstly, we compare the proposed PPOD-R with Zhang and Albin's (2009) χ^2 chart. The non-linear regression method by Williams et al. (2007) is not included here because Zhang and Albin's (2009) has shown that it is outperformed by the χ^2 chart. Choosing group L_1 penalty in (5) seems to be a natural alternative solution. As thoroughly investigated by She and Owen (2011), using the L_1 penalty is unable to obtain satisfactory detection results. By empirical studies, we find that this is still true for the group L_1 in the present problem and its detection performance is worse than the χ^2 chart in many cases. Thus, we choose not to include it as a comparison benchmark.

For a fair and clear comparison, we follow all the settings and scenarios used in Zhang and Albin (2009). To be more specific, we use type-I and type-II detection errors to assess the performance of PPOD-R and compare it with χ^2 chart. Suppose among the $m - m_o$ non-outlier profiles, m_1 profiles are incorrectly identified as outlier profiles, and among the m_o outlier profiles, m_2 profiles are correctly identified, then type-I and type-II errors are defined as $100m_1/(m - m_o)$ and $100(m_o - m_2)/m_o$ respectively. To determine the control limit in the χ^2 chart and the value of λ, we choose $\alpha = 0.05$. We generate profile datasets, each with $m = 200$ profiles, that consist of $m - m_o$ non-outlier and m_o outlying profiles where m_o takes values of 20, 40, 60 or 80. The profiles are generated as follows:

$$\begin{aligned} y_{ij} &= f_a(x) + \varepsilon_{ij},\ \varepsilon_{ij} \sim N(0, \sigma^2), \\ f_a(x) &= 10 - 20ae^{-ax_j} \sin(\sqrt{4 - a^2}x_j)/\sqrt{4 - a^2} + 10e^{-ax_j} \cos(\sqrt{4 - a^2}x_j), \end{aligned} \tag{13}$$

in which there are $n = 100$ covariates, taking values of $0.08, 0.16, \ldots, 8$. The non-outlier profiles have $a = 0.5$ and $\sigma = 1$.

Table 1: Average percentage of type-I and type-II errors for various values of a and m_o when $\sigma = 1$; The standard deviations are given in parentheses; In each category, the type-I and type-II errors are given in the left and right columns respectively

m_o	PPOD-R				χ^2 chart			
	$a = 0.5$		$a = 0.7$		$a = 0.5$		$a = 0.7$	
20	6.9 (2.1)	-	6.1 (2.0)	36.6 (14.1)	6.0 (1.7)	-	4.1 (1.3)	50.2 (11.6)
40	6.9 (2.1)	-	5.0 (2.1)	51.0 (13.4)	6.0 (1.7)	-	3.4 (1.4)	68.5 (7.2)
60	6.9 (2.1)	-	4.1 (1.8)	72.4 (10.5)	6.0 (1.7)	-	3.3 (1.5)	81.5 (4.9)
80	6.9 (2.1)	-	4.7 (2.1)	87.2 (5.2)	6.0 (1.7)	-	4.1 (1.7)	89.7 (3.3)
	$a = 0.9$		$a = 1.1$		$a = 0.9$		$a = 1.1$	
20	6.7 (2.2)	0.1 (0.7)	6.6 (2.3)	0.0 (0.0)	3.4 (1.2)	1.0 (2.3)	3.4 (1.3)	0.0 (0.0)
40	6.7 (2.2)	0.1 (0.5)	6.6 (2.3)	0.0 (0.0)	2.2 (1.1)	8.9 (4.7)	2.2 (1.1)	0.1 (0.4)
60	6.5 (2.5)	0.1 (0.4)	6.5 (2.5)	0.0 (0.0)	1.7 (1.0)	36.6 (6.3)	1.7 (0.9)	2.9 (2.3)
80	6.2 (2.5)	1.4 (9.8)	6.3 (2.6)	0.0 (0.0)	2.4 (1.3)	74.3 (5.0)	1.8 (1.1)	40.7 (6.1)
	$a = 1.3$		$a = 1.5$		$a = 1.3$		$a = 1.5$	
20	6.7 (2.2)	0.0 (0.0)	6.7 (2.2)	0.0 (0.0)	3.5 (1.2)	0.0 (0.0)	3.5 (1.2)	0.0 (0.0)
40	6.6 (2.3)	0.0 (0.0)	6.6 (2.3)	0.0 (0.0)	2.4 (1.1)	0.0 (0.0)	2.5 (1.1)	0.0 (0.0)
60	6.4 (2.4)	0.0 (0.0)	6.5 (2.4)	0.0 (0.0)	2.0 (1.1)	0.0 (0.3)	2.3 (1.1)	0.0 (0.0)
80	6.2 (2.6)	0.0 (0.0)	6.5 (2.6)	0.0 (0.0)	2.4 (1.3)	7.8 (3.4)	3.5 (1.4)	0.5 (0.8)

Tables 1 and 2 give the average type-I and type-II errors of applying two methods on simulated data when a and σ increases respectively. All the results are obtained with 1,000 replications. For clear comparisons, we also list the standard deviations of the type-I and type-II errors in parentheses. Note that in Table 1, when $a = 0.5$, the calculation of type-II error is not applicable ("-") because there are no outlying profiles. In both tables, the realized type-I error of PPOD-R is slightly higher than the nominal one 5% because several approximations are made in deriving λ. It is interesting to see that type-I errors of PPOD-R do not change too much in different settings. In contrast, the χ^2 control chart has quite different type-I errors in different combinations of (m_o, a) or (m_o, σ). This is mainly due to over-estimation of the error variance. Specially, as shown in Table 2, the type-I errors of χ^2 chart tend to be extremely small, and as a result its outlier detection ability (in terms of type-II error) would be largely compromised. In other words, the χ^2 chart cannot attain the targeted type-I error in certain cases and thus pays too much price on reducing swamping

effect. In outlier detection, masking is usually more serious than swamping. The former can cause gross distortions, whereas the latter is often just a matter of lost efficiency. In most cases, our proposed method approximately achieves the designed type-I error (thus swamping effect is not serious) and has a much lower type-II errors (thus alleviates masking effect) when m_o is large.

Table 2: Average percentage of type-I and type-II errors for various values of σ and m_o when $a = 0.5$; The standard deviations are given in parentheses; In each category, the type-I and type-II errors are given in the left and right columns respectively

m_o	PPOD-R				χ^2 chart			
	$\sigma = 1.2$		$\sigma = 1.4$		$\sigma = 1.2$		$\sigma = 1.4$	
20	6.5 (2.2)	14.3 (8.3)	6.7 (2.3)	0.1 (0.8)	3.5 (1.3)	20.9 (8.8)	2.9 (1.2)	0.3 (1.3)
40	5.8 (2.2)	15.8 (6.5)	6.6 (2.4)	0.1 (0.6)	1.9 (1.0)	29.2 (6.8)	0.9 (0.7)	1.0 (1.6)
60	5.2 (2.3)	17.1 (6.3)	6.5 (2.5)	0.1 (0.5)	0.8 (0.7)	39.3 (5.6)	0.1 (0.2)	4.3 (2.4)
80	4.3 (2.2)	18.8 (6.3)	6.1 (2.5)	0.1 (0.4)	0.3 (0.5)	50.5 (4.5)	0.0 (0.0)	16.5 (3.0)
	$\sigma = 1.6$		$\sigma = 1.8$		$\sigma = 1.6$		$\sigma = 1.8$	
20	6.7 (2.2)	0.0 (0.0)	6.6 (2.2)	0.0 (0.0)	2.9 (1.2)	0.0 (0.0)	2.9 (1.2)	0.0 (0.0)
40	6.5 (2.2)	0.0 (0.0)	6.6 (2.3)	0.0 (0.0)	0.7 (0.7)	0.0 (0.0)	0.8 (0.7)	0.0 (0.0)
60	6.5 (2.5)	0.0 (0.0)	6.5 (2.4)	0.0 (0.0)	0.0 (0.0)	0.3 (0.7)	0.0 (0.0)	0.0 (0.3)
80	6.3 (2.6)	0.0 (0.0)	6.1 (2.5)	0.0 (0.0)	0.0 (0.0)	6.1 (2.3)	0.0 (0.0)	2.6 (1.7)

A question naturally arise. If two methods have a similar type-I error, whether PPOD-R still performs better? To get a fairer picture, we perform a type-I-corrected comparison in the sense that the control limit is found through simulations so that the χ^2 chart has the same type-I error as PPOD-R with $\alpha = 0.05$. Note that such a scheme with adjustment is only for comparison use in our simulations but not applicable in practical applications since the type-I error of χ^2 chart depends on both the number of outliers and shift model which are unknown. Figure 1 shows the type-II error comparison between PPOD-R and χ^2 chart under various signal strength (a) with corrected type-I errors. Clearly, the proposed PPOD-R outperforms χ^2 chart by a quite large margin in detecting multiple outliers, which justifies our claim in Section 2 that masking effects can be avoided to certain degree through a penalized-type procedure. We conducted some other simulations with various combinations of m and m_o to check whether the above conclusions are true in other settings. Some representative

results are tabulated in Table 3 which shows type-I and type-II errors of PPOD-R when m is relatively small. The ratio of the outliers in the sample, m_o/m, varies from 0.05 to 0.40 in this table. As we found n does not affect the results much, $n = 50$ is fixed in Table 3. The values given here and some other results show that the proposed PPOD-R chart performs quite satisfactorily under these other settings as well.

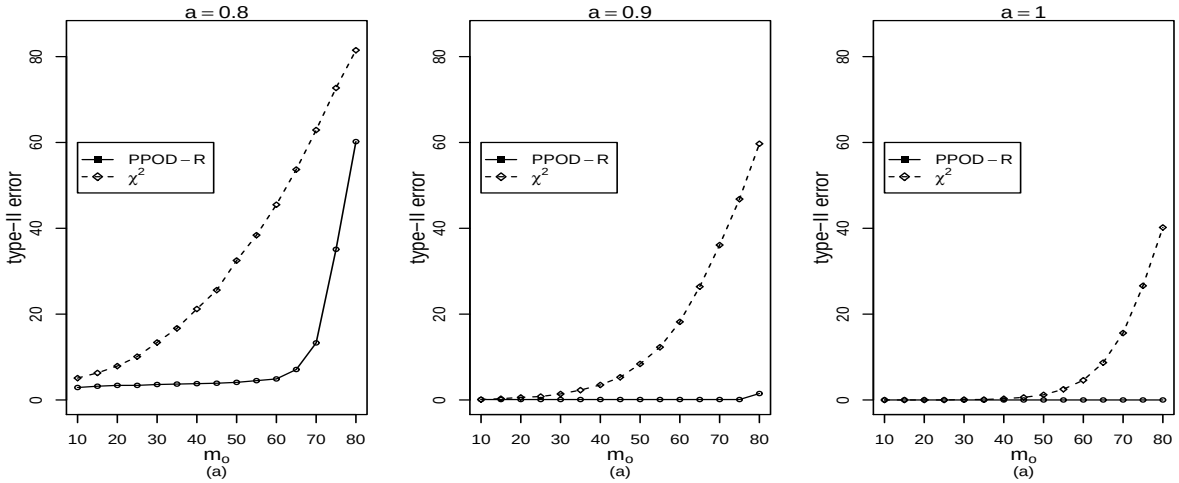

Figure 1: Type-II error comparison between PPOD-R and χ^2 chart under various signal strength (a) when both type-I errors are controlled to be the same. In each plot, the type-II error is plotted against m_o.

Finally, we evaluated the performance of PPOD-C and PPOD-R when within-profile correlations exist. Specifically, the profile model (13) and its associated parameter settings are still considered expect that the following two error structures are used instead. Scenario (I): $\boldsymbol{\varepsilon}_i$'s are distributed as multivariate normal with mean zero vector and covariance matrix $\boldsymbol{\Sigma} = (\sigma_{kl})$ being $\sigma_{kk} = 1$ and $\sigma_{kl} = 0.5^{|k-l|}$, for $k, l = 1, 2, \ldots, n$; Scenario (II): AR(1) process with autocorrelation parameter 0.6. The type-I and type-II errors of PPOD-R and PPOD-C under Scenarios (I) and (II) when a changes are tabulated in Table 4. Again, $\alpha = 0.05$ is chosen. We can observe that the within-profile correlations indeed have adverse effect on PPOD-R in the sense that the type-I error usually has a quite large bias and the degradation becomes more pronounced under Scenario (II). That is to say, PPOD-R would suffer high swamping probabilities. In comparison, by taking the correlation into consideration, the PPOD-C procedure is more robust, giving a reasonably reduction in type-I errors. Certainly, the type-I errors of PPOD-C are still higher than the nominal one due to the use of inaccurate estimation and limiting distribution. However, considering its simplicity

Table 3: Average percentage of type-I and type-II errors of PPOD-R under various combinations of (m_o, m, a); In each category, the type-I and type-II errors are given in the left and right columns respectively

	m_o/m	$a = 0.75$		$a = 1.0$		$a = 1.5$	
	0.05	7.5	8.4	7.3	0.0	7.5	0.0
	0.10	6.1	11.7	6.2	0.0	6.4	0.0
	0.15	5.8	13.4	6.3	0.0	6.2	0.0
	0.20	5.1	18.3	5.8	0.0	5.6	0.0
$m = 30$	0.25	4.9	21.9	5.4	0.0	5.8	0.0
	0.30	4.2	37.4	5.0	0.0	5.9	0.0
	0.35	3.9	50.5	4.8	0.0	5.5	0.0
	0.40	4.3	74.1	4.2	2.0	5.1	0.0
	0.05	6.8	7.8	6.8	0.0	6.8	0.0
	0.10	6.6	9.5	6.6	0.0	6.4	0.0
	0.15	5.9	10.9	6.4	0.0	6.5	0.0
	0.20	5.8	13.0	6.3	0.0	6.4	0.0
$m = 60$	0.25	5.1	18.7	6.2	0.0	6.4	0.0
	0.30	4.6	31.5	6.0	0.0	6.3	0.0
	0.35	3.9	53.2	5.7	0.0	6.5	0.0
	0.40	3.9	76.3	5.5	0.3	6.5	0.0
	0.05	6.8	7.4	6.9	0.0	6.9	0.0
	0.10	6.3	8.8	6.9	0.0	6.8	0.0
	0.15	6.3	9.5	6.6	0.0	6.8	0.0
	0.20	6.0	10.9	6.5	0.0	6.5	0.0
$m = 90$	0.25	5.5	13.7	6.4	0.0	6.4	0.0
	0.30	5.0	25.3	6.2	0.0	6.6	0.0
	0.35	4.2	46.7	6.0	0.0	6.7	0.0
	0.40	3.8	76.2	5.8	0.1	7.0	0.0

and generality, PPOD-C could be an ideal candidate for outlier detection in practice.

4 A real-data example

In this section, we apply the proposed methodology to a real dataset taken from an industrial etching process in semiconductor manufacturing. In this example, the etching chamber is equipped with more than 50 sensors which record the values of several variables with time during a batch. For illustrative purposes, only the variables related to spectral analysis of

Table 4: Performance comparison between PPOD-R and PPOD-C under correlated Scenarios (I) and (II) and $\sigma = 1$; The standard deviations are given in parentheses; In each category, the type-I and type-II errors are given in the left and right columns respectively

m_o	PPOD-C				PPOD-R			
	Scenario (I)							
	$a = 0.9$		$a = 1.1$		$a = 0.9$		$a = 1.1$	
20	10.5 (3.1)	2.5 (3.6)	10.5 (3.0)	0.0 (0.4)	16.2 (3.7)	1.4 (2.6)	16.1 (3.6)	0.0 (0.3)
40	10.6 (3.4)	2.6 (2.8)	10.4 (3.3)	0.0 (0.3)	16.2 (4.0)	1.4 (2.0)	16.0 (4.0)	0.0 (0.2)
60	10.6 (3.6)	2.9 (2.6)	11.0 (3.6)	0.0 (0.3)	15.7 (4.1)	1.8 (2.0)	15.9 (4.0)	0.0 (0.2)
80	10.6 (4.2)	3.4 (5.6)	11.0 (4.0)	0.0 (0.2)	15.5 (4.3)	2.0 (3.4)	15.5 (4.3)	0.0 (0.2)
	Scenario (II)							
	$a = 0.9$		$a = 1.1$		$a = 0.9$		$a = 1.1$	
20	11.0 (3.3)	21.9 (11.6)	11.5 (3.3)	3.6 (4.6)	22.2 (4.6)	11.3 (8.2)	22.9 (4.5)	1.5 (3.0)
40	9.8 (3.3)	26.8 (11.6)	11.3 (3.6)	4.0 (3.6)	21.4 (4.7)	12.9 (7.6)	22.5 (4.8)	1.6 (2.1)
60	8.2 (3.3)	40.6 (16.3)	11.3 (4.0)	4.4 (3.7)	20.3 (5.0)	18.2 (10.0)	22.0 (5.1)	1.7 (1.9)
80	5.8 (2.9)	73.5 (17.5)	11.1 (4.6)	5.6 (7.2)	19.0 (4.9)	36.6 (18.5)	21.6 (5.4)	2.1 (3.5)

chamber gas, is considered. There are many steps involved in the batch operation, but here we only focus on the second step since engineers consider that this step is one of the most crucial steps which usually contains enough critical information to distinguish the out-of-control conditions in the process. In each batch, the observations of chamber gas are collected at many time-points. See Figure 2-(a) for illustration. The engineers are usually concerned about significant changes of those profiles which may indicate some assignable causes in the process have occurred. More detailed discussion about this example can be found in Lee et al. (2011).

The entire dataset contains 364 wafer. The profile observations are synchronized so that the number and positions of covariate points (say, time, in this example) are equal. Figure 2-(a) shows the profiles for 364 wafers. The x-axis represents totally 137 synchronized profile points. Clearly, the profile curves present a similar functional change along with the time point. It is not easy to obtain a parametric or nonparametric model for representing these profiles because it seems that there is a very sharp peak at the beginning of each profile. Such a peak is partly due to the on-off feature in the etching process. Please refer to Lee et al.

(2011) for detailed explantation. We also note that some of the profiles significantly deviate from the population curves and should be regarded as outlying profiles which we would like to automatically screen out with some detection procedures. Thus, in this example, we consider to apply our proposed PPOD method to this dataset. As illustrated by Lee et al. (2011), in this dataset, there is a significant aging trend caused by the change of environmental temperature. Hence, we firstly apply the method suggested by Lee et al. (2011) to remove the aging drifts.

Since the measurements in each profile (wafer) are taken in consecutive time intervals, the data exhibit a considerable amount of positive serial autocorrelation in each profile, which is confirmed by Figure 2-(b). This figure depicts the estimated within-profile correlations $\hat{\rho}(t, 5)$, $\hat{\rho}(t, 35)$, and $\hat{\rho}(t, 65)$, for $t = 1, \ldots, 137$, where $\hat{\rho}(t_1, t_2)$ denotes the estimated correlation between the observations at t_1 and t_2 time points. From the plot, we can see that correlation within profiles is substantial; thus, it should not be ignored. Accordingly, the PPOD-C procedure, which takes the within-profile correlations into consideration, seems more appropriate for this example.

We choose the type-I error as 0.01 and apply PPOD-C described in Algorithm 3 to this dataset. The PPOD-C procedure is completed with four iterations and identifies totally 28 outlying profiles. Figure 3 shows all the identified curves along with the final estimate of $\boldsymbol{\mu}_0$, $\hat{\boldsymbol{\mu}}_{0c}^{(4)}$. The deviation of these curves from the mean function is quite clear. The validity of these detection results needs to be checked by engineers based on physical knowledge and experience in a follow-up analysis.

5 Concluding remarks

The main contribution of this paper is to consider a penalized function, under the framework of treating profiles as vectors, defined by adding a hard-penalty function to the Euclidian distance, which can be directly minimized by a recursive algorithm. With a good design of tuning parameter, the proposed method successfully identified the outliers given a type-I error. This technique is associated with the χ^2 chart, but provides a new characterization in form of penalized regressions and thus is able to alleviate masking effects to a large

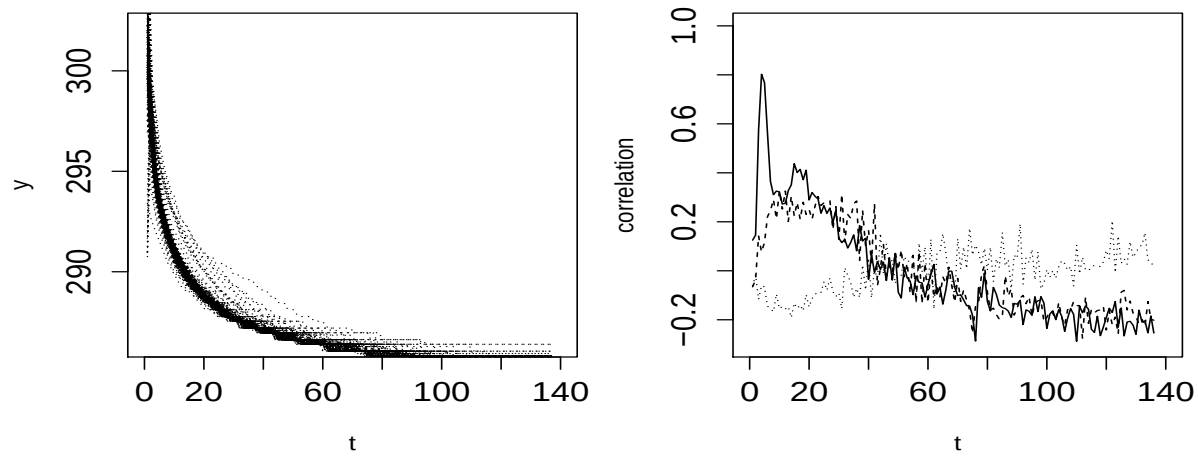

Figure 2: (a): The synchronized profiles of sensor variable chamber gas over a 364 wafer batch; (b): Solid, dashed and dotted curves represent estimated within-profile correlations $\hat{\rho}(t, 5)$, $\hat{\rho}(t, 35)$, and $\hat{\rho}(t, 65)$

extent. Furthermore, we successfully generalized this penalized/thresholding methodology to correlated profile problems to reduce the type-I error by accommodating within-profile correlations.

We assume that there is only one response variable and one explanatory variable. However, the proposed PPOD-R and PPOD-C can be also applied with one response variable and multiple explanatory variables or with multiple response variables (Noorossana et al. 2010). We also assume that the explanatory variable takes a fixed set of values in all profiles. If this is violated in practice, we can use linear interpolation or local smoothing to synchronize the profile observations of so that all profiles share a fixed set of values for the explanatory variable (as done for the real-data example in Section 4). In addition, the standpoint of this paper is that profile data may sometimes be so complex that all parametric or non-parametric modelling methods do not apply. However, in many applications, the functional relationship in profile data is indeed sufficiently smooth which allows us to consider some methods based on functional data analysis (Ramsay and Silverman 2005). Recently, Yu et al. (2012) proposed a new testing procedure for profile outliers based on the functional principal component analysis. Their procedure is essentially a retrospective one and thus also suffers from masking effect much as they have shown. One of our ongoing work is to integrate PPOD with their test to construct some more robust outlier detection methods in detecting outlying profiles with certain smoothness.

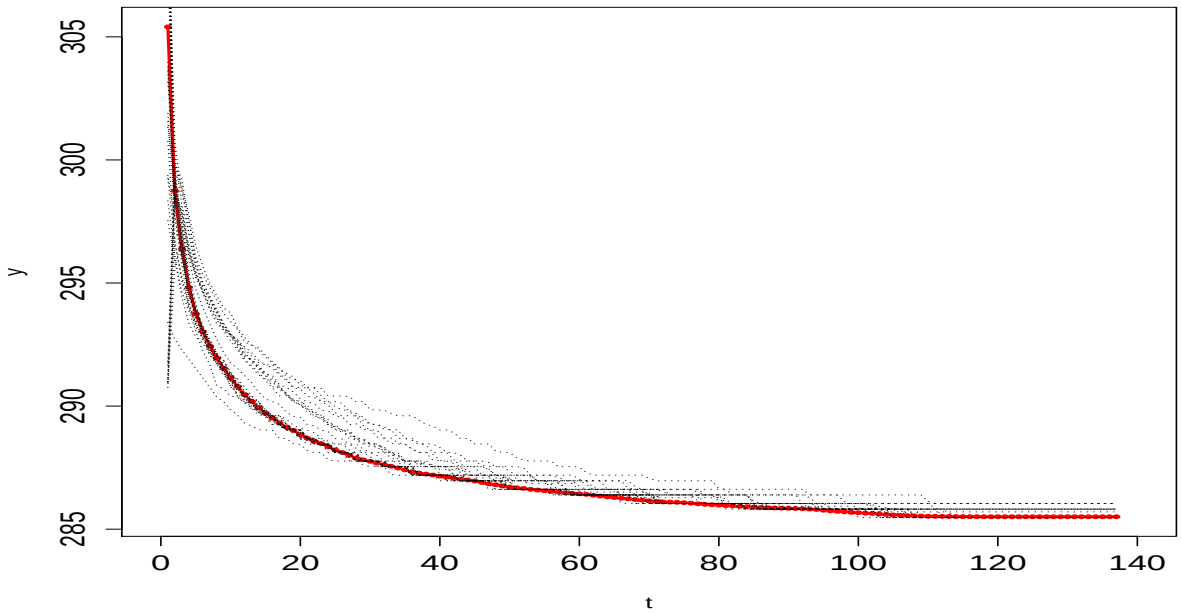

Figure 3: All the identified curves (black dotted line) along with the final estimate of $\boldsymbol{\mu}_0$ (red solid line)

Acknowledgement

The authors would like to thank the Department Editor and two anonymous referees for their many helpful comments that have resulted in significant improvements in the article. This research was supported by the NNSF of China Grants 11001138, 11071128, 11131002, 11101306, the RFDP of China Grant 20110031110002, the Fundamental Research Funds for the Central Universities 65012231 and the PAPD of Jiangsu Higher Education Institutions. Zou and Tseng also thank for the support of the National Center for Theoretical Sciences, Math Division.

Appendix:

Proof of Proposition 2

The second inequality in (10) is straightforward from the algorithm and the fact that $\hat{\mu}_0$ is the minimizer of (9) given $\boldsymbol{\gamma}$. The first inequality is clearly true by Proposition 1, provided $\hat{\boldsymbol{\gamma}}_i = I_{\{||\mathbf{y}_i - \boldsymbol{\mu}_0|| > \lambda\}}(\mathbf{y}_i - \boldsymbol{\mu}_0)$ is the solution to (7). Note that the second equation in

the simultaneous equations (7) can be reduced to

$$\mathbf{y}_i - \boldsymbol{\mu} = \frac{\lambda \boldsymbol{\gamma}_i}{||\boldsymbol{\gamma}_i||}, \ \forall\, 0 \neq ||\boldsymbol{\gamma}_i|| < \lambda,$$

which leads to the fact that for any $0 \neq ||\boldsymbol{\gamma}_i|| < \lambda$, $||\mathbf{y}_i - \boldsymbol{\mu}|| = \lambda$. Thus, if $||\mathbf{y}_i - \boldsymbol{\mu}|| \neq \lambda$, $||\boldsymbol{\gamma}_i|| \geq \lambda$ or $||\boldsymbol{\gamma}_i|| = 0$. On the other hand, when $||\mathbf{y}_i - \boldsymbol{\mu}|| > \lambda$, $\boldsymbol{\gamma}_i = \mathbf{y}_i - \boldsymbol{\mu}$ from the first equation in (7), while $\boldsymbol{\gamma}_i = \mathbf{0}$ if $||\mathbf{y}_i - \boldsymbol{\mu}|| \leq \lambda$ by using the third equation in (7). This completes the proof. □

Proof of Proposition 3 In what follows, we always assume that $\hat{\boldsymbol{\mu}}_0$ is the overall mean of $\mathbf{y}_i, i = 1, \ldots, m$ as in Zhang and Albin (2009), which will facilitate our exposition much. Note that $|||\mathbf{y}_i - \hat{\boldsymbol{\mu}}_0||^2 - ||\mathbf{y}_i - \boldsymbol{\mu}_0||^2| \leq ||\hat{\boldsymbol{\mu}}_0 - \boldsymbol{\mu}_0||^2$ and $m||\hat{\boldsymbol{\mu}}_0 - \boldsymbol{\mu}_0||^2$ is distributed as the same as $||\mathbf{y}_i - \boldsymbol{\mu}_0||^2$. Consequently, $||\mathbf{y}_i - \hat{\boldsymbol{\mu}}_0||^2 = ||\mathbf{y}_i - \boldsymbol{\mu}_0||^2(1 + O_p(m^{-1}))$. It suffices to show the asymptotic distribution of $||\mathbf{y}_i - \boldsymbol{\mu}_0||^2$.

By standard theory in multivariate analysis (Box 1954), the distribution of $||\mathbf{y}_i - \boldsymbol{\mu}_0||^2$ can be shown to be equivalent to the distribution of the linear combination of independent χ^2_1-variates with coefficients given by the eigenvalues of $\boldsymbol{\Sigma}$, say

$$||\mathbf{y}_i - \boldsymbol{\mu}_0||^2 \stackrel{d}{=} \sum_{j=1}^{n} \lambda_j z_j^2,$$

where $z_j \overset{\text{iid}}{\sim} N(0,1)$. Note that $\sum_{j=1}^n \lambda_j z_j^2$ is a weighted sum of i.i.d random variables. By Hajek-Sidak central limit theorem,

$$\frac{\sum_{j=1}^n \lambda_j z_j^2 - E}{\sqrt{V}} \xrightarrow{d} N(0,1),$$

where E and V denote the expectation and variance of $\sum_{j=1}^n \lambda_j z_j^2$ respectively, provided $\eta_n^2/\text{tr}(\boldsymbol{\Sigma}^2) \to 0$ is valid. The proof can be completed by directly calculating the closed-forms of E and V. □

References

Bai, Z., and Yin, Y. Q. (1993) Limit of the Smallest Eigenvalue of a Large Dimensional Sample Covariance Matrix. *The Annals of Probability*, **21**, 1276–94.

Box, G. E. P. (1954) Some Theorems on Quadratic Forms Applied in the Study of Analysis of Variance Problems, I. Effect of Inequality of Variance in the One-Way Classification. *The Annals of Mathematical Statistics*, **25**, 290–302.

Breheny, P., and Huang, J. (2011) Coordinate Descent Algorithms for Nonconvex Penalized Regression, with Application to Biological Feature Selection. *The Annals of Applied Statistics*, **5**, 232–253.

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

Ding, Y., Zeng, L., and Zhou, S. (2006). Phase I Analysis for Monitoring Nonlinear Profiles in Manufacturing Processes. *Journal of Quality Technology*, **38**, 199–216.

Donoho, D., and Johnstone, I. (1994). Ideal Spatial Adaptation via Wavelet Shrinkages. *Biometrika*, **81**, 425–455.

Friedman, J. H., Hastie, T., and Tibshirani, R. (2010) Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, **33**, 1–22.

Hawkins, D. (1980). *Identification of Outliers*, Chapman and Hall, New York.

Jin, J., and Shi, J. (1999). Feature-Preserving Data Compression of Stamping Tonnage Information Using Wavelet. *Technometrics*, **41**, 327–339.

Kang, L., and Albin, S. L. (2000) On-line Monitoring when the Process Yields a Linear Profile. *Journal of Quality Technology*, **32**, 418–426.

Lada, E. K., Lu, J. -C., and Wilson, J. R. (2002). A Wavelet-Based Procedure for Process Fault Detection. *IEEE Transactions on Semiconductor Manufacturing*, **15**, 79–90.

Lee, S. P., Chao, A. G., Tsung, F., Wong, D. S. H., Tseng, S. T., and Jang, S. S. (2011). Monitoring Batch Processes with Multiple On-Off Steps in Semiconductor Manufacturing. *Journal of Quality Technology*, **43**, 142–157.

Mahmoud, M. A. (2008) Phase I Analysis of Multiple Linear Regression Profiles. *Communications in Statistics: Simulation and Computation*, **37**, 2106–2130.

Mahmoud, M. A., and Woodall, W. H. (2004) Phase I Analysis of Linear Profiles with Calibration Applications. *Technometrics*, **46**, 380–391.

Montgomery, D. C. (2001) *Introduction to Statistical Quality Control, 4th edition*, John Wiley & Sons, New York, NY.

Noorossana, R., Eyvazian, M., Amiri, A., and Mahmoud, M. A. (2010) Statistical Monitoring of Multivariate Multiple Linear Regression Profiles in Phase I with Calibration Application. *Quality and Reliability Engineering International*, 26, 291–303.

Noorossana, R., Saghaei, A., Amiri, A. (2011) *Statistical Analysis of Profile Monitoring*, Wiley, New Jersey.

Paynabar, K., and Jin, J. (2011) Characterization of Non-Linear Profiles Variations Using Mixed-Effect Models and Wavelets. *IIE Transactions*, **43**, 275–290.

Qiu, P., Zou, C., and Wang, Z. (2010) Nonparametric Profile Monitoring By Mixed Effect Modeling (with discussions). *Technometrics*, **52**, 265–277.

Ramsay, J. O., and Silverman, B. W. (2005) *Functional Data Analysis*, Springer, New York.

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

Walker, E., and Wright, S. P. (2002). Comparing Curves Using Additive Models. *Journal of Quality Technology*, **34**, 118–129.

Wang, K., and Jiang, W. (2009) High-Dimensional Process Monitoring and Fault Isolation via Variable Selection. *Journal of Quality Technology*, **41**, 247–258.

Williams, J. D., Woodall, W. H., and Birch, J. B. (2007) Statistical Monitoring of Nonlinear Product and Process Quality Profiles. *Quality and Reliability Engineering International*, **23**, 925–941.

Woodall, W. H. (2007) Current Research on Profile Monitoring. *Revista Producão*, **17**, 420–425.

Yang, Y. (2005) Can the Strengths of AIC and BIC Be Shared?–A Conflict between Model Identification and Regression Estimation. *Biometrika*, **92**, 937–950.

Yu, G., Zou, C., and Wang, Z. (2012) Outlier Detection in the Functional Observations with Applications to Profile Monitoring. *Technometrics*, **54**, 308–318.

Yuan, M., and Lin, Y. (2006) Model Selection and Estimation in Regression with Grouped Variables. *Journal of Royal Statistical Society B*, 68, 49–67.

Zhang, H., and Albin, S. (2009) Detecting Outliers in Complex Profiles Using a χ^2 Control Chart Method. *IIE Transactions*, **41**, 335–345.

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

Zou, C., Liu, Y., Wang, Z., and Zhang, R. (2010). Adaptive Nonparametric Comparison of Regression Curves. *Communications in Statistics: Theory and Method*, **39**, 1299–1320.

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

Zou, C., Tsung, F., and Wang, Z. (2008). Monitoring Profiles Based on Nonparametric Regression Methods. *Technometrics*, **50**, 512–526.