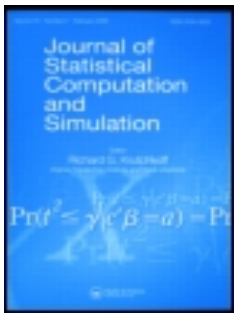

Journal of Statistical Computation and Simulation

ISSN: 0094-9655 (Print) 1563-5163 (Online) Journal homepage: http://www.tandfonline.com/loi/gscs20

Semi-parametric segmentation of multiple series using a DP-Lasso strategy

K. Bertin, X. Collilieux, E. Lebarbier & C. Meza

To cite this article: K. Bertin, X. Collilieux, E. Lebarbier & C. Meza (2016): Semi-parametric segmentation of multiple series using a DP-Lasso strategy, Journal of Statistical Computation and Simulation, DOI: 10.1080/00949655.2016.1260726

To link to this article: http://dx.doi.org/10.1080/00949655.2016.1260726

Published online: 30 Nov 2016.

JOURNAL OF STATISTICAL COMPUTATION AND SIMULATION, 2016
http://dx.doi.org/10.1080/00949655.2016.1260726

Taylor & Francis
Taylor & Francis Group

Semi-parametric segmentation of multiple series using a DP-Lasso strategy

K. Bertin[a], X. Collilieux[b], E. Lebarbier[c] and C. Meza[a]

[a]CIMFAV-Facultad de Ingeniería, Universidad de Valparaíso, Valparaíso, Chile; [b]IGN LAREG, Université Paris Diderot Sorbonne Paris Cité, Paris, France; [c]AgroParisTech/INRA UMR518, Paris 5e, France

ABSTRACT

We consider a semi-parametric approach to perform the joint segmentation of multiple series sharing a common functional part. We propose an iterative procedure based on Dynamic Programming for the segmentation part and Lasso estimators for the functional part. Our Lasso procedure, based on the dictionary approach, allows us to both estimate smooth functions and functions with local irregularity, which permits more flexibility than previous proposed methods. This yields to a better estimation of the functional part and improvements in the segmentation. The performance of our method is assessed using simulated data and real data from agriculture and geodetic studies. Our estimation procedure results to be a reliable tool to detect changes and to obtain an interpretable estimation of the functional part of the model in terms of known functions.

ARTICLE HISTORY
Received 3 February 2016
Accepted 10 November 2016

KEYWORDS
Semi-parametric estimation; multiple series; segmentation; Lasso; dynamic programming; geodetic data

1. Introduction

The aim of segmentation methods is to detect abrupt changes or breakpoints in signals. Segmentation problems arise in many areas: the detection of chromosomal aberrations in biology [1,2], temperature and precipitation series homogenization in meteorology [3] or GPS coordinates change detection in geodesy [4], among others. In many situations, multiple series (corresponding to several patients, meteorological stations or GPS receivers, in the previous examples) are observed and a common functional part, describing a global trend or an underlying signal, is shared by all the series. This is the case of, for instance, the 'wave' effect in the genomic context (see [5] and references therein), climatic effects in agriculture series (see [6]) or geophysical signals in geodesy [7]. A common approach in this setting (usually adopted e.g. in climatic studies, see [3] or [8]) is to avoid the estimation of the functional part by working with series of differences. However, taking into account the functional part in the segmentation model can be crucial for an accurate breakpoint detection, and is also relevant to provide a more complete and reliable interpretation of the observed phenomena. The importance of a simultaneous treatment of breakpoints and functional part can be illustrated by GPS coordinate series, where both components are useful to determine accurate station velocities in tectonic or earth mantles studies, and to infer information about sea level variations, ice melting or climate changes [9–11]. In this context, the best performing methods are based on visual inspection of the time series [12], followed by the estimation and interpretation of periodic components in a second run [13].

In order to integrate some time-dependency, a Minimal Description Length approach has been proposed for the segmentation of climatic time series involving autoregressive processes, see [14,15].

CONTACT C. Meza cristian.meza@uv.cl CIMFAV-Facultad de Ingeniería, Universidad de Valparaíso, General Cruz 222, Valparaíso, Chile

However, those works do not take into account complex functional parts (only a linear trend is considered in [15]) and deal only with one single series. The need for automatic methods to simultaneously model breakpoints and general common functional parts in multiple series arises in different application areas, and it is one of the main motivations of the present work.

We consider here a semi-parametric model where all series are analysed simultaneously, and which can deal with covariates and series with observations at different times. The parametric part of the model deals with the segmentation, which is supposed to be series-specific (i.e. each series has its own segmentation). On the other hand, the functional part, which is supposed to be shared by all the series, corresponds to the non-parametric component. Such a model has been proposed in [5] for a genomic application where, in addition to the changes (the chromosomal aberrations), the aforementioned 'wave' effect is observed. The authors considered a penalized least-squares framework, where the segmentation parameters and the non-parametric part are first estimated for a fixed number of segments, and then this number is estimated using a model selection approach. Their best results were obtained when the functional part was viewed as a fixed effect; that approach did however not furnish an interpretable estimation of the functional part and tended to catch too much noise in the fixed effect. They also proposed to estimate the functional part using splines, but showed that this approach performs well only when the signal is smooth enough. In the present work, we propose a new and more flexible way to estimate the functional part, using a Lasso approach, which can adapt to both smooth and irregular signals and provides an interpretable estimation of the functional part.

For the estimation of the breakpoints, we will use a Dynamic Programming (DP) algorithm. It is by now well known that this algorithm is the only one that retrieves the exact solution in a fast way. However, DP can only be applied when the contrast (e.g. the log-likelihood) to be optimized is additive with respect to the segments (see [2,3,16]). This is not the case in our model since the functional part is a global parameter (compared to the means, which are segment-specific). For that reason, our method consists in an iterative two-steps procedure which alternates the segmentation issue with the estimation of the functional part (as in [5] or [16]). Indeed, in the segmentation of the corrected series with respect to the estimated functional part, the associated contrast is additive with respect to the segments and DP can then be applied. To choose the number of breakpoints, we use a criterion proposed by Picard et al. [6] for segmentation of multiple series. This criterion is an adaptation of the modified BIC criterion of Zhang and Siegmund [17] and it is shown in [6], by a simulation study, to perform better than usual model selection criteria.

The main contribution of our method is to improve the estimation of the functional part and consequently the estimation of the segmentation part, as it will be observed in our simulation study. We propose to that end a flexible estimation of the functional part by using a dictionary approach. More precisely, it is estimated by linear combinations of elements of a set called dictionary, which can contain functions with different regularities: smooth functions, for example splines or Fourier functions, and more irregular ones, such as spiky functions. To select the relevant functions, we use a Lasso-type strategy, introduced in [18] and recently applied in a semi-parametric framework in [19,20], resulting in an estimation procedure with good practical and theoretical performance. Lasso non-parametric estimators have the advantage that the size of the dictionary can be large and even much larger than the data size. In our segmentation model, the use of Lasso-type methods allows us to moreover obtain an interpretable estimation of the functional part, in terms of known functions; this is a very relevant feature in applications, such as the geodesy problem mentioned before. Note that Lasso-type methods have been used in segmentation models [21,22] before, but only for the detection of breakpoints, leading to faster algorithms but tending to overestimate the number of breakpoints.

We apply our procedure to simulated data where the functional part is represented as a mixture of smooth and irregular functions. We obtain very good estimation results for both the segmentation and functional parts. In particular, our method outperforms the methods proposed in [5], where the functional part was treated as a fixed effect or smoothed using wavelets or splines. Moreover, we apply our method to agricultural data from French stations and GPS data from Australian stations and we find several breakpoints of interest. In the case of the agriculture data, the flexible modelling of the

non-parametric part allows us to avoid false detection in the segmentation. As regards the GPS data, we retrieve in the estimated functional part several periodic functions that have been observed in previous studies. Since all periodic components are all together estimated simultaneously with the segmentation part, the associated amplitudes and phases are more reliable for geophysical purposes (see, e.g. [7,13] for their interpretation in that context).

This article is organized as follows. In Section 2, we present the model and our estimation procedure. In Section 3, the comparative simulation study is carried out to assess the performance of our method. In Section 4, we apply our method to the agriculture and geodetic data mentioned before. Finally, conclusions are given in Section 5.

2. Semi-parametric model and estimation procedure

2.1. The model

We observe M series. For each $m \in \{1, \dots, M\}$, the mth series is observed at n_m times $t_{im} \in \mathbb{N}$, $i \in \{1, \dots, n_m\}$, so that the total number of observations is $N = \sum_{m=1}^{M} n_m$. We denote by $y_m(t)$ the signal of the series m at time $t \in \mathbb{N}$ and assume that

$$y_m(t_{im}) = \mu_m(t_{im}) + f(x_m(t_{im})) + e_{im}, \quad \text{for } i = 1, \dots, n_m, \ \ m \in \{1, \dots, M\}, \tag{1}$$

where $\mu_m(t) = \mu_k^m$ if $t \in I_k^m = (\tau_{k-1}^m, \tau_k^m]$, x_m for $m \in \{1, \dots, M\}$ represent covariates, f is an unknown function to be estimated, $\tau_k^m \in \mathbb{N}$ is the kth breakpoint of the series m, μ_k^m is the mean of the series m on the segment I_k^m and the e_{im} are i.i.d centred Gaussian variables with variance σ^2. We denote by K_m the number of segments of the mth series and by $K = \sum_{m=1}^{M} K_m$ the total number of segments.

Note that the segmentation is specific to each series whereas the function f is common to all series. Moreover, since the M series are not necessarily observed at the same times and take into account covariates, approaches using difference between series cannot be applied.

We define the $[n_m \times 1]$ vectors $y_m := (y_m(t_{im}))_i$ and $x_m := (x_m(t_{im}))_i$, and the $[K_m \times 1]$ vector $\mu_m := (\mu_k^m)_k$. We, respectively, concatenate the vectors y_m and x_m into $([N \times 1])$ vectors Y and X, and the mean vectors $\mu_m = (\mu_k^m)_{k=1,\dots,K_m}$ into a vector μ of size $K \times 1$. We denote by T the $[N \times K]$ block diagonal matrix with blocks T_m given by the $[n_m \times K_m]$ block diagonal matrices of blocks $\mathbb{1}_{n_k^m}$. Here, $\mathbb{1}_\ell$ stands for the $[\ell \times 1]$ vector with all coordinates equal to 1 and n_k^m is the number of observations in the kth segment for series m.

The parameters of the model are the segmentation part $(T\mu)$ (or, equivalently, the means μ_k^m and the breakpoints τ_k^m), the function f, the variance σ^2 and the number of segments K.

2.2. The estimation procedure

We consider a penalized least-squares framework to estimate all the parameters of the model. In this context, when the penalty only depends on the segmentation through the number of segments, one can usually proceed in two steps: estimate first the parameters for a fixed number of segments and then choose that number using a model selection strategy.

In the first step, fixing a number K of segments, the function f and the segmentation $(T\mu)$ are estimated. It is now well known that the only efficient algorithm that allows to retrieve the optimal segmentation is the dynamic programming (DP) algorithm. However, as mentioned in [16] or [5], the presence of a global parameter, here f, hampers the use of this algorithm. This is why we consider an iterative procedure that alternates the estimation of the segmentation part and the function f.

Here we want to estimate the function f in a non-parametric fashion considering a Lasso-type method based on a dictionary approach. More specifically, we consider a collection of functions $\{\phi_1, \dots, \phi_J\}$, called dictionary, and we propose to estimate f by a linear combination of the

functions ϕ_j,

$$f_\lambda = \sum_{j=1}^{J} \lambda_j \phi_j, \quad \lambda = (\lambda_1, \ldots, \lambda_J) \in \mathbb{R}^J.$$

Note that the form of the dictionary is very flexible. It can be a base or the union of several bases and, as we said before, the functions ϕ_j can have different regularities: smooth functions or more irregular ones. The estimation of f is then reduced to the estimation of λ that will be performed by a Lasso-type procedure.

Our proposed procedure, described in the next section, is then called DP-Lasso.

2.2.1. A DP-Lasso estimation procedure

The parameters to be estimated in the DP-Lasso procedure are $(T\mu)$, λ and σ and we denote by $(T\mu)^{(h)}$, $\lambda^{(h)}$ and $\sigma^{(h)}$ the estimation of these three parameters obtained at the iteration (h). At iteration $(h+1)$, our DP-Lasso procedure alternates the three following steps.

(1) Given $\lambda^{(h)}$, the segmentation parameters $T\mu^{(h+1)}$ are estimated by:

$$T\mu^{(h+1)} = \underset{T\mu}{\operatorname{argmin}} \|Y - T\mu - F\lambda^{(h)}\|^2,$$

where $\|\cdot\|$ stands for the L_2 norm in $\mathbb{R}^N$ and F is the $[N \times J]$ matrix $F = (f_{ij})$ where $f_{ij} = \phi_j(X_i)$. The problem is then reduced to segment $Y - F\lambda^{(h)}$ into K segments, which is a classical segmentation problem that can be solved using DP. In particular, we use here the two-stages DP proposed by Picard et al. [5,6] which is a fast version of DP designed for the joint segmentation of multiple series.

(2) Given $T\mu^{(h+1)}$ and $\sigma^{(h)}$, the function f is estimated using a Lasso-type strategy by:

$$f^{(h+1)} = f_{\lambda^{(h+1)}},$$

where $\lambda^{(h+1)}$ minimizes

$$\|Y - T\mu^{(h+1)} - F\lambda\|^2 + 2\sum_{j=1}^{J} r_{N,j}|\lambda_j|,$$

where $r_{N,j} = \sigma^{(h)}\|\phi_j\|_N\sqrt{\gamma \log J}$ with $\gamma > 2$ and $\|\phi_j\|_N = \sqrt{\sum_{l=1}^{N} \phi_j^2(X_l)}$. Note that from a theoretical point of view, the condition $\gamma > 2$ ensures that the resulting estimator of f has good properties (see [20]).

(3) Given $T\mu^{(h+1)}$ and $f^{(h+1)}$, the variance σ^2 is estimated by

$$(\sigma^{(h+1)})^2 = \frac{1}{N}\|Y - T\mu^{(h+1)} - F\lambda^{(h+1)}\|^2.$$

The algorithm stops when the difference between parameters of two successive iterations is smaller than ϵ (10^{-3} in practice). The initial values at step 1 are as follows: $(T\mu)^{(1)}$ is given by step (1) with $\lambda^{(0)} = 0$ and $\sigma^{(1)}$ is given by step (3) with $\lambda^{(1)} = 0$.

The final estimators are denoted $\hat{\tau}_k^m$, $\hat{\mu}_k^m$, $\widehat{T\mu}$, $\hat{\sigma}^2$, $\hat{\lambda}$ and $\hat{f} = f_{\hat{\lambda}}$.

2.2.2. Model selection

The last issue is the choice of the number of segments K. We propose here to use the modified BIC criterion proposed in [17] and successfully adapted to the joint segmentation in [6]:

$$\mathrm{mBIC}_{\mathrm{JointSeg}}(K) = \log\left[\Gamma\left(\frac{N-K+1}{2}\right)\right] - \left(\frac{N-K+1}{2}\right)\log\left(\frac{\|Y-\widehat{T\mu}-F\hat{\lambda}\|^2}{N}\right) + \left[\frac{1}{2}-(K-M)\right]\log(N) - \frac{1}{2}\sum_{m=1}^{M}\sum_{k=1}^{K_m}\log\hat{n}_k^m,$$

where $\hat{n}_k^m$ is the number of observations on the estimated segment k of the mth series. Note that, in [6], a simulation study showed that this criterion performs better than other usually used model selection criteria.

3. Study of the performance of the method

In order to assess the performance of our procedure, called here *Lasso* for the sake of simplicity, we conduct the simulation study described below. We also propose to compare our method to the work [5], where either the function f is estimated using splines or f is viewed as a fixed effect depending on the time t, that is, $f(t) = \beta_t$. We call these two approaches *Spline* and *Position*, respectively, and we perform them on the simulated data using the `cghseg` R package, in particular the `multiseg` R function. For our procedure, we develop our own functions in R using the `lars` R package to perform the Lasso estimation of f. Note that we do not consider the wavelet estimation of f since the method of Picard et al. [5] does not yield good results in that case.

3.1. Simulation design and quality criteria

3.1.1. Simulation design

We consider the model (1) with $x_m(t) = t$. Thus the model is rewritten for $m \in \{1, \ldots, M\}$ and $t \in \{1, \ldots, n\}$ as follows:

$$y_m(t) = \mu_m(t) + f(t) + e_{tm}, \tag{2}$$

where the e_{tm} are i.i.d. Gaussian variable $\mathcal{N}(0, \sigma^2)$.

The length n of the series is fixed and equal to 100. We consider two different numbers of series, $M \in \{10, 50\}$, and five values for error variance, $\sigma^2 \in \{0.1, 0.2, 0.5, 1.0, 1.5\}$. For each series, the number of segments K follows a Poisson distribution with mean 3 and their positions are uniformly distributed. The mean value within each segment alternates between 0 and a value in $\{-2, -1, +1, +2\}$ with probability $\{0.2, 0.3, 0.3, 0.2\}$, respectively. The function f is generated as a mixture of a sine function with three peaks (see Figure 1):

$$f(t) = 0.3 \times \sin\left(2\pi\frac{t}{20}\right) + 0.51\mathrm{I}_{t=0.1\times n} - 1\mathrm{I}_{t=0.5\times n} + 21\mathrm{I}_{t=0.6\times n}. \tag{3}$$

Each configuration, that is, specific values of M and σ^2, is simulated 100 times.

For the Lasso strategy, we use a dictionary with 150 functions: 128 Haar functions $(t \mapsto 2^{7/2}1\mathrm{I}_{[0,1]}(2^7 t/100 - k),\ k = 0, \ldots, 2^7 - 1)$, the Fourier functions $(t \mapsto \sin(2\pi jt/100), t \mapsto \cos(2\pi jt/100),\ j = 1, \ldots, 10)$ and the functions $t \mapsto t$ and $t \mapsto t^2$. We applied this procedure with $\gamma = 2.1$.

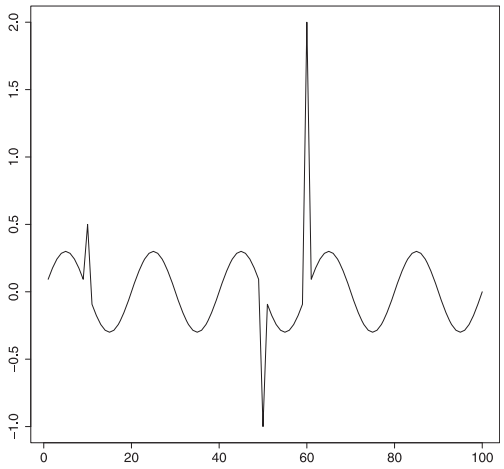

Figure 1. Simulated function *f*.

3.1.2. Quality criteria

To study the quality of the estimation, for each configuration of (M, σ^2), we consider several criteria.

- For the segmentation parameters, in order to study the global quality of the estimation, we consider the root-mean-square distance between the true mean and its estimate: $\text{RMSE}(\mu) = [(1/N)\sum_{m=1}^{M}\sum_{t=1}^{n}\{\mu_m(t) - \hat{\mu}_m(t)\}^2]^{1/2}$ where $N = M \times n$. Moreover, to study the performance of the estimation of the breakpoint positioning, we consider both the proportion of erroneously detected breakpoints among detected breakpoints (false discovery rate, FDR) and the proportion of undetected true breakpoints among true breakpoints (false negative rate, FNR).
- For the function *f*, the root-mean-square distance between f and its estimate: $\text{RMSE}(f) = [(1/n)\sum_{t=1}^{n}\{f(t) - \hat{f}(t)\}^2]^{1/2}$is also considered. In addition, for the particular study of the performance of the Lasso strategy, a FDR criterion is considered, corresponding to the number of false selected functions among the selected ones.

To distinguish between the two FDR criteria, we denote by breakpoint-FDR and dictionary-FDR, the FDR of the breakpoints and *f*, respectively. For each configuration, we consider the average of these criteria over the 100 simulations.

We also compare the algorithms in terms of runtime. Note that, as in the `cghseg` R package, the search of the best segmentation is performed through the faster version of the Dynamic Programming algorithm (PDPA) proposed by Cleynen et al. [23] (using the `Segmentor3IsBack` R package). Recall that in practice, we have to segment all the series for $K = 1, \ldots, K^{\max}$, where $K^{\max}$ is the maximal total number of segments, before selecting the number of segments. In order to make a fair comparison, we impose to the different methods the same maximal number of segments per series, the same maximal total number of segments K and the same simulations.

3.2. Results

3.2.1. Comparison between Lasso, Spline and Position

Only the results with $M = 10$ are presented since the results for $M = 50$ lead to the same conclusions.

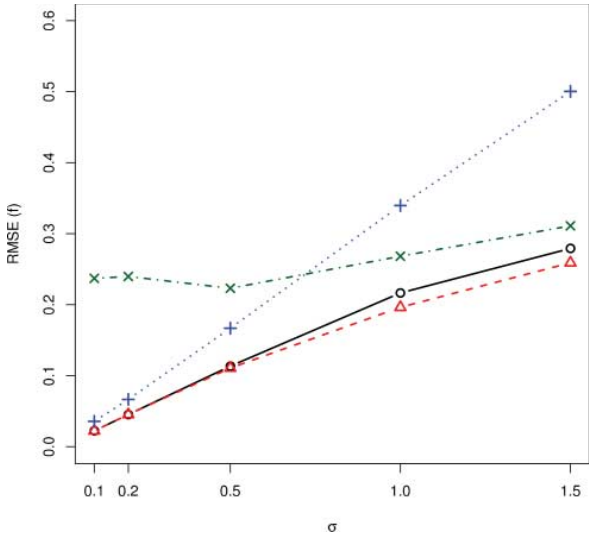

Figure 2. RMSE of f with respect to σ for Lasso $\triangle$, Position $+$, Spline $\times$ and Lasso with the true number of segments $\circ$ for $M = 10$.

Figure 2 presents the RMSE(f) obtained with the three methods as a function of σ. We observe that the larger is the noise, the worst is the estimation of f due to the confusion between the signal and the noise. Whatever the level of noise, *Lasso* outperforms *Position* and *Spline* in terms of the non-parametric part estimation. However, the behaviour of *Position* and *Spline* is opposite with respect to σ. For small σ, *Spline* leads to bad performances compared to *Lasso* and *Position*. Indeed, as expected, *Spline* tends to capture the smooth part of the signal, that is, the sinusoidal trend only, whereas the two other methods catch both the peaks and the trend. However, for large σ, it is more difficult to detect the peaks of the true function, resulting in closest results for *Lasso* and *Spline*. As mentioned in [5], *Position* tends to catch the trend but also the noise, resulting in an erratic estimation of f and in a large RMSE compared to the others. The bad estimation of f can have consequences on the segmentation estimation. Figure 3 summarizes the results for the segmentation estimation obtained with the different methods, as functions of σ. In general, *Lasso* is slightly better than the two other methods. For σ larger than 0.5, the results are similar, even for *Position* for which f is not well estimated. However for small values of σ, since *Spline* does not detect the peaks, they are considered as breakpoints in the segmentation, leading to bad results: more segments are then detected (see $\hat{K} - K$) and these false breakpoints then increase the breakpoint-FDR and so the RMSE(μ).

As a conclusion, *Position* and *Lasso* behave similarly for the estimation of the segmentation part. The main difference concerns the estimation of f which is less reliable for *Position*. An important advantage of *Lasso* is its flexibility in the sense that functions of different regularities can be included in the dictionary and, in particular, some functions chosen according to the knowledge of the expert. The final form of the estimator $\hat{f}$ is a sparse linear combination of the dictionary functions that allows a possible interpretation of f, contrary to *Position* (see Section 4).

Finally, Table 1 gives the mean runtime in seconds for each method and different values of σ. Our method is longer compared to the two others (followed by *Spline* then *Position*). Note that our algorithm is probably not optimized as well as the cghseg package. We can also observe that the mean runtime of all methods decreases with the increasing of the noise. That is explained by the fact that the larger is σ, more difficult is the detection of the breakpoints and the selection of functions. Indeed in this case, the segmentation has less breakpoints, the number of selected functions is small, and the (iterative) algorithms converge faster.

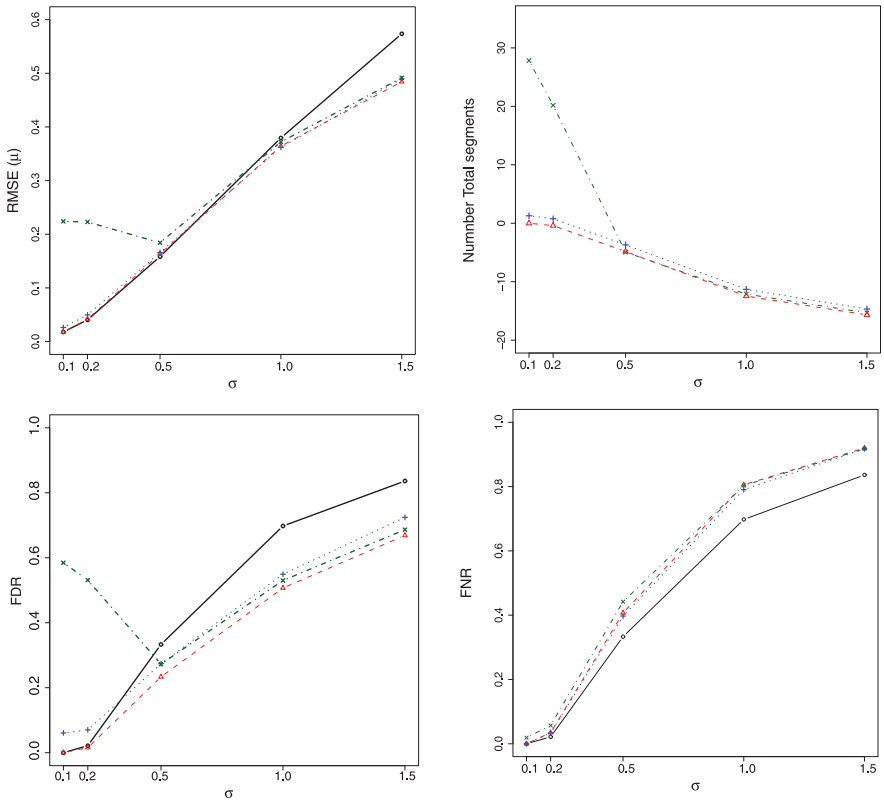

Figure 3. Results for $M = 10$ with respect to σ. Top: RMSE of μ on the left, $\hat{K} - K$ on the right. Bottom: breakpoint-FDR on the left and breakpoint-FNR on the right. Lasso $\triangle$, Position $+$, Spline $\times$ and Lasso with the true number of segments $\circ$.

Table 1. Mean runtime in seconds.

σ	0.1	0.2	0.5	1	1.5
Position	2.6	2.29	1.76	1.38	1.17
Spline	21.5	17.22	1.86	1.40	1.27
Lasso	103.30	74.20	62.25	48.31	37.30

3.2.2. Comparison between Lasso with the true and the estimated number of segments

We first compare the results obtained with the true and estimated number of segments. In Figure 3, we observe that the more difficult is the detection (i.e. the higher is σ), the more under-estimated is the number of segments. This result was expected and is now classical in the study of model selection for segmentation. Indeed, the procedure tends to reduce the number of segments in order to avoid false detection. This is illustrated by a smaller breakpoint-FDR (Figure 3) when the number of segments is estimated. In Figure 2, we see that this leads to a better estimation in terms of segmentation (small RMSE(μ)) and consequently to a slightly better estimation of the function f (small RMSE(f)).

3.2.3. Quality of Lasso as a function of the number of series

Table 2 summarizes the relative differences between $M = 10$ and $M = 50$ for two criteria, the FDR and the root-mean-square of f, for several values of σ using:

$$\mathrm{FDR}^{\sigma} = \frac{\mathrm{FDR}^{\sigma}_{10} - \mathrm{FDR}^{\sigma}_{50}}{\mathrm{FDR}^{\sigma}_{10}},$$

Table 2. Comparison between $M = 10$ and $M = 50$ series for breakpoint – FDR and RMSE(f) criteria.

	Relative differences	
σ	FDR$^\sigma$	RMSE(f)$^\sigma$
0.1	–	57.46
0.2	42.15	57.97
0.5	9.40	55.58
1.0	7.00	50.47
1.5	5.64	47.47

Table 3. Percentage, dictionary-FDR and mean of number of functions selected by Lasso.

		ID function					
	σ	13	64	77	137	Dictionary-FDR function	Mean length
	0.1	100	100	100	100	0.052	4.27
	0.2	100	100	100	100	0.055	4.29
$M = 10$	0.5	26	99	100	100	0.064	3.53
	1.0	5	28	99	99	0.114	2.13
	1.5	0	12	73	76	0.137	1.9
	0.1	100	100	100	100	0.059	4.31
	0.2	100	100	100	100	0.059	4.31
$M = 50$	0.5	100	100	100	100	0.068	4.36
	1.0	53	100	100	100	0.084	3.95
	1.5	18	92	100	100	0.108	3.6

$$\text{RMSE}(f)^\sigma = \frac{\text{RMSE}(f)^\sigma_{10} - \text{RMSE}(f)^\sigma_{50}}{\text{RMSE}(f)^\sigma_{10}},$$

where, for example, FDR^σ_{10} and $\text{RMSE}(f)^\sigma_{10}$ denote, respectively, the breakpoint-FDR and the root-mean-square of f for $M = 10$ series and a specific value of σ. Table 3 shows the percentage of true functions of the simulated function f, selected in the estimator $\hat{f}$, against different values of σ, with $M = 10$ and $M = 50$ series. The ID function corresponds to the position of the true functions in the dictionary with size 150. Specifically, the first three functions (labels 13, 64 and 77) are Haar functions centred at, respectively, 10, 50 and 60 and the function 137 is the function $x \mapsto \sin(2\pi(5t/100))$. As expected, the increase in the number of series improves the estimation of f. For small values of σ, the *Lasso* procedure leads to a good performance in terms of selected functions whatever the number of series: the number of selected functions is close to the true one, and among them all the true functions of the simulated function are retrieved with less false selection (small dictionary-FDR function). That leads logically to an accurate estimation of f (small RMSE(f) Figure 2). For noisy configurations (large σ), fewer functions are selected, which was expected. Indeed, in this case, there is more confusion between noise and signal and the small peaks (in particular ID 13 and 64) are more difficult to detect. This is particularly true for a small number of series. Remark that for $M = 50$, the ID 77 and 137 are always selected. Moreover, the better accuracy of the estimation of f observed for $M = 50$ leads to a better positioning of the breakpoints (see FDR$^\sigma$). This is less marked when σ is large.

4. Application

4.1. Agriculture data

In this section, we want to illustrate the importance of the correct modelling of the function f in order to avoid false breakpoint detection. To this end, we apply our method on harvest dates studied in [6]. In this application, the purpose is to detect changes in the agricultural practices by detecting, in the grape harvest dates, changes which are not due to a climatic effect. The data are harvest dates and mean

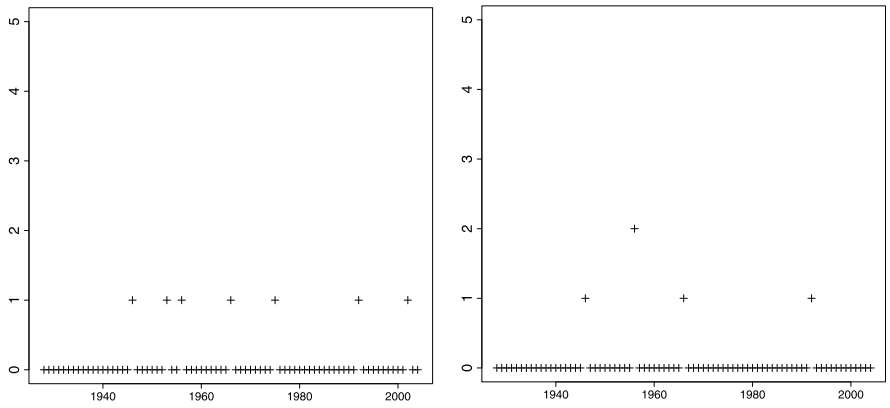

Figure 4. Number of the detected breakpoints over all the stations obtained in case (1) on the left and case (2) on the right.

temperature from April to August obtained at 10 stations from several French regions between 1928 and 2004 (see [3] for more details on the data). In model (1), $y_m(t)$ is the grape harvest date and the covariate $x_m(t)$ is the mean temperature of the year $t \in \{1928, \ldots, 2004\}$ for series $m \in \{1, \ldots, 10\}$. In [6], a climatic effect f of the form $f(x_m(t)) = bx_m(t) + cx_m^2(t)$ was proposed (we call it case (1)). In case (2), no assumptions are made on the function f and it is estimated using our proposed procedure, for which we consider a dictionary combining very different orthonormal families. More precisely, our dictionary contains 36 functions that are Haar wavelets, Fourier functions, and the polynomial functions x, x^2 and x^3. In the resulting estimator of f obtained with $\gamma = 2.1$, five functions are selected. Figure 4 represents the number of detected breakpoints per year over all the series for the two cases.

The most important difference between the two considered cases concerns the year 2003: that year is considered as a breakpoint in case (1) and not in case (2). This breakpoint appears in the series 6. Figure 5 represents, respectively, the harvest dates of the series 6 and its segmentation after correction in case (1) (segmentation of $y_m(t) - \hat{b}x_m(t) - \hat{c}x_m^2(t)$). As shown in Figure 6, the correction of the harvest date at this year by $\hat{b}x_m(t) + \hat{c}x_m^2(t)$ is stronger compared to $\hat{f}(x_m(t))$ obtained in case (2), which is why a breakpoint is added (see Figure 5 right). This year, with a temperature equals to 32.15, corresponds to a very hot summer. This could suggest that this point may be taken into account in the climatic effect and not detected as a practical change. Related to this assumption, our approach (the case (2)) seems then to be flexible enough to fit in a better way the climatic effect and as a consequence avoid false detection.

4.2. Geodesy data

In this section, we present the results obtained with our estimation procedure for a specific GPS data set. GPS permanent stations that continuously monitor their coordinates have been deployed all over the world for more than 20 years. Their three-dimensional coordinate series are usually post-processed by scientists from raw code and phase observations at a daily or weekly basis, yielding series up to 1000 or 7000 records long with a typical precision of a few millimeters. Such series are used to determine accurate station velocities for tectonic and Earth's mantle studies, with a typical magnitude of a few millimeters per year to about 10 centimetres per year [24].

The GPS coordinate series show breakpoints due to instrumental changes (documented or not), earthquakes or changes in the raw data processing strategy. These abrupt changes are superimposed to several types of variations. The observed coordinate variations reflect the ground deformations at the station including tectonic signals (generally a trend, mostly in the horizontal components) as well as

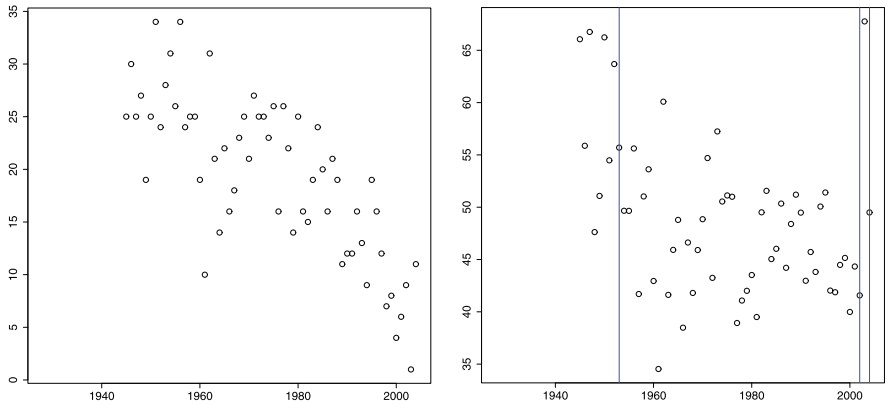

Figure 5. Left: harvest dates of the series 6. Right: obtained segmentation of the corrected series in case (1) (on $y_{mt} - \hat{b}x_{mt} - \hat{c}x_{mt}^2$).

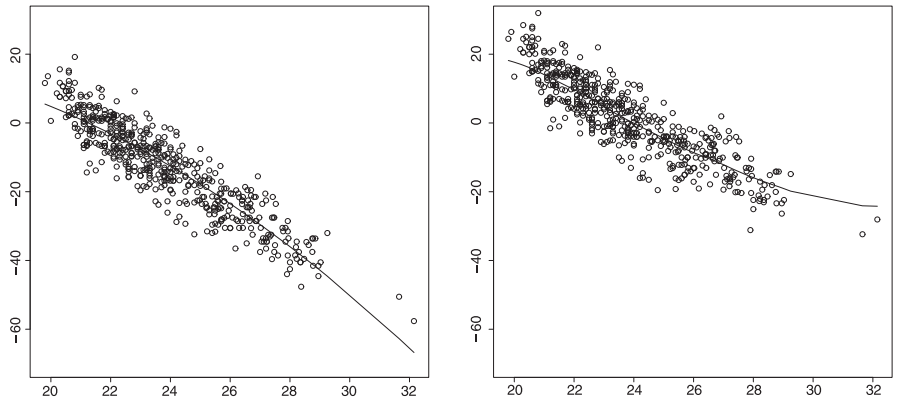

Figure 6. Fit of f in case (1) on the left and case (2) on the right.

environmental signals from the vicinity of the station, such as soil moisture or atmospheric pressure changes. The latter could be approximated by periodic signals with dominant annual and semi-annual periods [7]. The observational noise exhibits more autocorrelation at long periods [25] but specific systematic errors of small magnitudes also show up at some well known periods, which are either submultiples of 350.5 days [26] or annual, like thermal deformation of the station monumentation and the ground.

For geographically close series, one may suppose that these periodic variations are common. We apply our segmentation strategy to the height coordinate series of four GPS stations in Australia located in Yarragadee (YAR1, YAR2, YAR3 and YARR). They were computed by the Jet Propulsion Laboratory (JPL) and can be downloaded at ftp://sideshow.jpl.nasa.gov/pub/JPL_GPS_Timeseries/repro2011b/post/point/. We use the series from their first observations to the 22nd of June 2013 – series provided online are updated everyday. Then, the model (1) is considered with $M = 4$ and $n_1 = 2862$, $n_2 = 5209$, $n_3 = 1443$ and $n_4 = 2197$, the respective lengths of the series. Here they have been averaged at weekly scale. For all these series, the ground motion is assumed to be identically observed and is described by function $f(t)$. Thus, equipment changes or malfunction at individual station should show up in the segmentation. For those series, JPL provides a list of changes, namely those

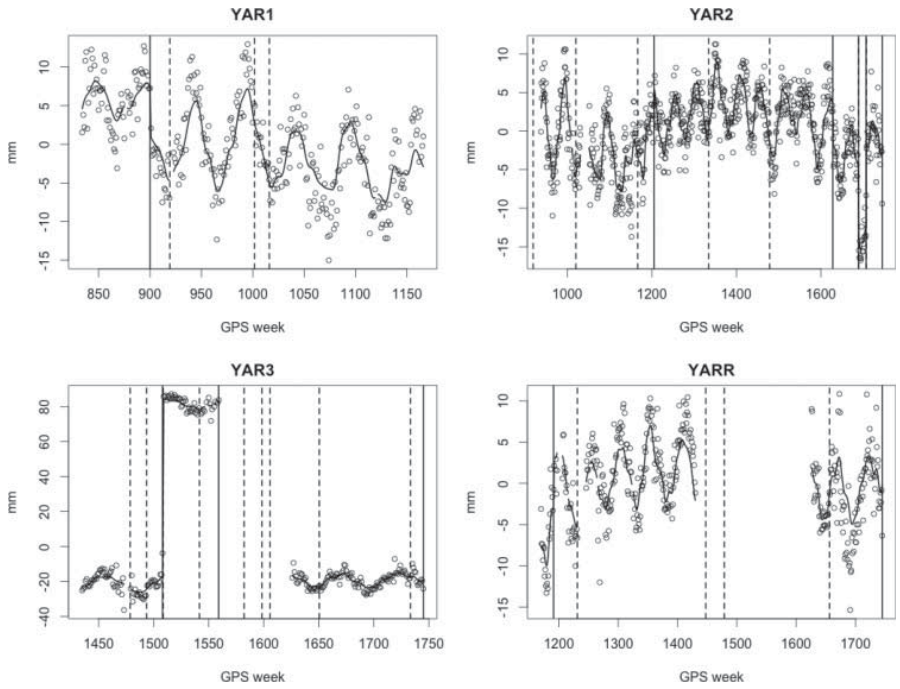

Figure 7. Results for height coordinate series of four GPS stations (YAR1, YAR2, YAR3 and YARR): obtained breakpoints in solid vertical lines; known equipment changes in dashed vertical line; estimated function f in solid line.

being detected using a procedure based on sequential F-test applied to the tridimensional coordinate series (M. Heflin, personal communication, 2014).

We apply our proposed procedure to these series with a dictionary consisting of 226 functions, which are Fourier functions only: $t \mapsto \sin(2\pi w_i t), t \mapsto \cos(2\pi w_i t)$ where $w_i = i/T$, $T = \max(t) - \min(t)$ and T/i is larger than 8 weeks (since smaller period amplitudes are generally negligible (see in [26])). Figure 7 shows the results for the four series: the obtained breakpoints in solid vertical lines, the known equipment changes in dashed vertical line and the estimated function f in solid line.

Eight breakpoints are detected. Four (GPS week 1689 and 1707 of the series YAR2 and 1508 and 1559 of the series YAR3) correspond exactly to receiver and antenna changes. The changes at time 1205 of the series YAR2 is likely to be related to the equipment change at time 1166. In the same series, a change at time 1628 is detected. This change is not related to a known instrumental change, however, it is also proposed by JPL. Compared to the JPL official list of changes, we found three additional changes for YAR2 at GPS week 1205 and the two validated changes at 1689 and 1707. Our two other additional changes at time 900 of the series YAR1 and at time 1191 of the series YARR are not reported by JPL. Up to now, no explanation has been supplied for those.

With respect to the estimation of f, a total of 50 periods (62 functions) has been selected, among them the ones close to the well-known frequencies mentioned above (annual and semi-annual) and submultiples of the draconitic periods. Amplitude and phase of the annual and semi-annual signals are of outmost importance for geophysicists. The 12 long periods – larger than 1 year – reflect well-known GPS low-frequency noise as already noticed by Amiri-Simkooei et al. [27].

As a conclusion, our method found the same known breakpoints as JPL official list, but includes new validated ones. Moreover the 62 functions selected in the Lasso procedure furnish relevant geodetic information.

Note that the cghseg R package cannot be used in the presence of missing data, which is the case in this real data set. To circumvent this problem and to make some comparison of the three methods (*Lasso*, *Spline* and *Position*) for these geodetic data, a way would be to keep only the common support. However, it is empty for these four series. We propose here to consider only two series YAR2 and YARR that share more points and apply the cghseg package to obtain the results of *Spline* and *Position* approaches. For YAR2, *Position* selects 12 breakpoints at positions 1291, 1346, 1349, 1360, 1368, 1413, 1430, 1638, 1654, 1689, 1693 and 1707 and *Spline* 3 at positions 1628, 1689 and 1707. These latter 3 breakpoints are detected by the three methods. However, our approach allows to detect another breakpoint at position 1205 which corresponds to an equipment change. Concerning YARR, *Position* selects 17 breakpoints and *Spline* 10. A common detected breakpoint at position 1184 is close to the only one we detect with our method (at position 1191). The other detected breakpoints are not related to known changes. To conclude in the case of these two series, our method detects a validated breakpoint not detected by *Spline* and *Position* approaches. The two other procedures tend to detect much more breakpoints, and above all more undocumented breakpoints than our method. Moreover, they do not allow to give an interpretation of the functional part in terms of geophysical components.

5. Conclusion

We propose a new semi-parametric approach for the segmentation of single or multiple series using a DP-Lasso strategy. Our method provides a valuable and reliable tool to assess changes and functional variations in series, as illustrated with real and simulated data. More specifically, in comparison with methodologies based on spline and fixed effect, the dictionary approach using LASSO algorithms shows improvements on the estimation of the functional part, thanks to the use of a dictionary combining different bases or orthonormal families. As a consequence, this also enhances the estimation of the segmentation parameters. This point is also observed in the application on agriculture data, where the flexible modelling of the non-parametric part allows us to avoid false breakpoint detection.

Moreover our DP-Lasso strategy furnishes an interpretable estimator of the functional part in terms of functions of the dictionary, which is very relevant for some applications, as in the geodetic study. Indeed, for GPS data, the Lasso functional estimator can be used to better interpret ground deformation observations or to enhance the piece-wise linear coordinate model of the Terrestrial Reference Frame [28,29], widely used for geosciences and mapping applications. This should provide a significant improvement for the users, since such coordinates are aimed to be extrapolated in the future (up to 5 years). In the geodetic context, we also furnish an automatic segmentation procedure that complements the already used methods based on visual inspection.

An interesting extension to be considered is the inclusion of random effects in our model. Indeed, as pointed out by Picard et al. [6], the use of a mixed model would allow us to take into account correlations that could exist across series. This and other extensions will be addressed in future works.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by FONDECYT [grant numbers 1141256 and 1141258]; ANILLO [grant number ACT-1112]; mathamsud [grant number 16-MATH-03] SIDRE and CONICYT [grant number 870100003].

References

[1] Lai WR, Johnson MD, Kucherlapati R, et al. Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. Bioinformatics. 2005;21(19):3763–3770.

[2] Picard F, Robin S, Lavielle M, et al. A statistical approach for CGH microarray data analysis. BMC Bioinformatics. 2005;6:27.

[3] Caussinus H, Mestre O. Detection and correction of artificial shifts in climate series. Appl Stat. 2004;53:405–425.
[4] Williams S. Offsets in global positioning system time series. J Geophy Res Solid Earth. 2003;108(19):2310.
[5] Picard F, Lebarbier E, Hoebeke M, et al. Joint segmentation, calling, and normalization of multiple CGH profiles. Biostatistics. 2011;12(3):413–428.
[6] Picard F, Lebarbier E, Budinska E, et al. Joint segmentation of multivariate gaussian processes using mixed linear models. Comput Statist Data Anal. 2011;55:1160–1170.
[7] Dong D, Fang P, Bock Y, et al. Anatomy of apparent seasonal variations from GPS-derived site position time series. J Geophy Res Solid Earth. 2002;107(B4):ETG 9–1.
[8] Mestre O, Domonkos P, Picard F, et al. HOMER: a homogenization software – methods and applications. Quarterly J Hungarian Meteorol Serv. 2013;1:47–67.
[9] Wahr J, Khan SA, van Dam T, et al. The use of gps horizontals for loading studies, with applications to Northern California and Southeast Greenland. J Geophy Res Solid Earth. 2013;118(4):1795–1806.
[10] Wu X, Collilieux X, Altamimi Z, et al. Accuracy of the international terrestrial reference frame origin and earth expansion. Geophy Res Lett. 2012;38:L13304.
[11] Wu X, Heflin MB, Schotman H, et al. Simultaneous estimation of global present-day water transport and glacial isostatic adjustment. Nature Geoscience. 2011;3(9):642–646.
[12] Gazeaux J, Williams S, King M, et al. Detecting offsets in gps time series: first results from the detection of offsets in gps experiment. J Geophy Res Solid Earth. 2013;118(5):2397–2407.
[13] van Dam T, Collilieux X, Wuite J, et al. Nontidal ocean loading: amplitudes and potential effects in gps height time series. J Geod. 2012;86(11):1043–1057.
[14] Li S, Lund R, Lee T. Multiple changepoint detection via genetic algorithms. J Clim. 2012;25:674–686.
[15] Lu Q, Lund R, Lee TCM. An MDL approach to the climate segmentation problem. Ann Appl Stat. 2010;4(1):299–319.
[16] Bai J, Perron P. Computation and analysis of multiple structural change models. J Appl Econ. 2003;18:1–22.
[17] Zhang NR, Siegmund DO. A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data. Biometrics. 2007;63(1):22–32.
[18] Tibshirani R. Regression shrinkage and selection via the Lasso. J R Stat Soc Ser B. 1996;58:267–288.
[19] Arribas-Gil A, De la Cruz R, Lebarbier E, et al. Classification of longitudinal data through a semiparametric mixed-effects model based on Lasso-type estimators. Biometrics. 2015;71(2):333–343.
[20] Arribas-Gil A, Bertin K, Meza C, et al. Lasso-type estimators for semiparametric nonlinear mixed-effects models estimation. Stat Comput. 2014;24(3):443–460.
[21] Harchaoui Z, Lévy-Leduc C. Multiple change-point estimation with a total variation penalty. J Amer Statist Assoc. 2010;105:1480–1493.
[22] Shen J, Gallagher CM, Lu Q. Detection of multiple undocumented change-points using adaptive Lasso. J Appl Stat. 2014;41:1161–1173.
[23] Cleynen A, Koskas M, Lebarbier E, et al. Segmentor3IsBack: an R package for the fast and exact segmentation of Seq-data. Algorithms Mol Biol. 2014;9(1):6.
[24] King MA, Altamimi Z, Boehm J, et al. Improved constraints on models of glacial isostatic adjustment. A review of the contribution of ground-based geodetic observations. Surv. Geophys. 2010;31(5):465–507.
[25] Williams S, Bock Y, Fang P, et al. Error analysis of continuous GPS position time series. J Geophy Res. 2004;109(B18):B03412.
[26] Ray J, Altamimi Z, Collilieux X, et al. Anomalous harmonics in the spectra of GPS position estimates. GPS Solut.. 2008;12(1):55–64.
[27] Amiri-Simkooei AR, Tiberius CCJM, Teunissen PJG. Assessment of noise in GPS coordinate time series: methodology and results. J Geophy Res Solid Earth. 2007;112:B07413.
[28] Altamimi Z, Dermanis A. The choice of reference systm in itrf formulation: In: Sneeuw N, Novák P, Crespi M, Sansò F, editors. VII Hotine-Marussi symposium on mathematical geodesy, Vol. 137 of International Association of Geodesy Symposia. Berlin, Heidelberg: Springer; 2012. p. 329–334.
[29] Petit G, Luzum B. Iers conventions (2010). (Iers technical note ; 36). Technical report, Frankfurt am Main: Verlag des Bundesamts für Kartographie und Geodäsie, inprint; 2010.