TESI DI DOTTORATO

Dipartimento di Scienze Economiche Aziendali e Statistiche

Change-point estimation in piecewise constant regression models and extensions

Stima dei punti di svolta in modelli di regressione costanti a tratti ed estensioni

SALVATORE FASOLA

Tutor: **Vito Muggeo**

Coordinatore Dottorato: **Marcello Chiodi**

Dottorato di Ricerca in "Statistica, Statistica Applicata e Finanza Quantitativa", XXV Ciclo
Settore Scientifico Disciplinare: SECS/S01 - Statistica
Anno conseguimento titolo: 2015

Università degli Studi di Palermo

DSEAS

Dipartimento di Scienze Economiche,
Aziendali e Statistiche

To my family

Acknowledgements

I want to thank Prof. Helmut Küchenhoff for giving me the great opportunity to spend three months in the “Statistischen Beratungslabor” at the University of Munich and his precious support...
Special thanks to Lucie, Pia, and Katja Oberschmidt for being my family in Munich...
Many thanks to Prof. Vito Muggeo, Prof. Massimo Attanasio, the coordinator Prof. Marcello Chiodi and the teaching staff at the department for feeding my passion for Statistics...
Thanks to my sister, for several and long waits in the car (parked in the sun...), whatever month, whatever time, whatever temperature...
Thanks to my officemates Andrea, Antonio, Federico, Marianna and the “acquired colleague” Gianluca for bearing me, for nice time together, for encouraging me...
Thanks to ”the most honest person in the world”, Renato Cataldo, for his friendship (...and large amount of money payed for me!!!)...
Thanks to Rosangela Anzalone, Rita Amari and the administrative staff for taking care of my heavy nuisances...
Thanks to You for making everything possible...

Contents

Chapter 1

Introduction

1.1 Background

The problem of *change-point detection* appears to have been introduced in the *quality control* context (Shewhart, 1925), where it was concerned with the monitoring of the characteristics of products to detect possible significant departures from a specified target level. In such context, data collection and analysis go on until a change is detected, with the main goal of restoring the initial state after the change occurrence. Since the analyst searches for a change at a time, this statistical approach is referred to as *sequential change-point analysis* (Box *et al.*, 2011; Lai and Xing, 2010).

In recent years, due to the widespread occurrence of change-point problems in different areas, such as econometrics (Bai and Perron, 1998), engineering (Blythe *et al.*, 2012) and biology (Siegmund, 2013), research about the topic has developed in the more general context of *regression analysis*; see Khodadadi and Asgharian (2008) for an extensive survey.

A general problem of *change-point regression* for the response variable Y can be stated as follows:

$$\mathbb{E}[Y_i] = \begin{cases} f(\boldsymbol{z}_i, \boldsymbol{\beta}_1) + \eta_i & x_i \leq \psi_1 \\ f(\boldsymbol{z}_i, \boldsymbol{\beta}_2) + \eta_i & x_i \in (\psi_1, \psi_2] \\ \vdots & \vdots \\ f(\boldsymbol{z}_i, \boldsymbol{\beta}_{K+1}) + \eta_i & x_i > \psi_K \end{cases}, \tag{1.1}$$

$i = 1, 2, \ldots, n$. $\boldsymbol{z}_i$ is a vector of covariates which affects the response via the regression coefficients $\boldsymbol{\beta}_k$, $k = 1, 2, \ldots, K+1$, depending on the value of the continuous covariate x.

The shape of the function $f(\cdot)$ changes any time x_i overcomes a change-point ψ_k, $k = 1, 2, \ldots, K$ (we use the convention $\psi_0 = x_{(1)}$ and $\psi_{K+1} = x_{(n)}$), while the function η_i does not depend on x. The purpose is estimating both the K points at which changes occur and the vectors of (varying) parameters $\boldsymbol{\beta}_k$. Since the K change-points are estimated at the same time, these procedures are referred to as *non-sequential change-point analyses* (Frick *et al.*, 2014), and represent a very popular research topic in computational mathematics. Here we provide a quite schematic literature review, without the presumption of being exhaustive.

1.2 Linear change-point regression

In the simplest scenario the function $f(\cdot)$ in (1.1) is linear in $\boldsymbol{\beta}_k$, namely

$$f(\boldsymbol{z}_i, \boldsymbol{\beta}_k) = \boldsymbol{z}_i^T \boldsymbol{\beta}_k + \eta_i \quad x_i \in (\psi_{k-1}, \psi_k], \tag{1.2}$$

$i = 1, 2, \ldots, n$, $k = 1, 2, \ldots, K+1$, where η_i is linear as well.

A considerable amount of research about linear change-point regression is concerned with models in which x_i is simply a discrete indicator, generally $x_i = i$, which does not belong to $\boldsymbol{z}_i^T$, that is

$$\mathbb{E}[Y_i] = \boldsymbol{z}_i^T \boldsymbol{\beta}_k + \eta_i \qquad x_i \in (\psi_{k-1}, \psi_k], \tag{1.3}$$

$i = 1, 2, \ldots, n$, $k = 1, 2, \ldots, K+1$. Since x_i does not enter the model matrix, but only affects the coefficients $\boldsymbol{\beta}_k$, we refer to these models as *varying coefficient models*. For example, in *structural change models* (Perron, 2006) interest lies in detecting single (Bai, 1997a) or multiple (Bai, 1997b) time points at which one or more covariate effects change abruptly. If all the parameters are subject to shift ($\eta_i = 0$) we deal with *pure structural change models*, otherwise we deal with *partial structural change models* (Bai and Perron, 2003).

Another common situation considers $\boldsymbol{z}_i^T = (1, x_i)$, that is

$$\mathbb{E}[Y_i] = \beta_{0k} + \beta_{1k} x_i + \eta_i \qquad x_i \in (\psi_{k-1}, \psi_k], \tag{1.4}$$

$i = 1, 2, \ldots, n$, $k = 1, 2, \ldots, K + 1$. Now x_i enters the model matrix and the resulting regression function is *piecewise linear*. Models allowing discontinuities in the linear function are referred to as *abruptly changing models* (Basseville and Nikiforov, 1993), while when the linear function is assumed to be continuous at the points of change, the model is referred to as *segmented regression model* (Küchenhoff, 1997; Muggeo, 2003).

1.3 Piecewise constant models

In this thesis we focus on a further simplification of model (1.4) assuming

$$\mathbb{E}[Y_i] = \beta_{0k} + \eta_i \qquad x_i \in (\psi_{k-1}, \psi_k], \tag{1.5}$$

$i = 1, 2, \ldots, n$, $k = 1, 2, \ldots, K + 1$. If $\eta_i = 0$, model (1.5) leads to a step function, and is referred to as *piecewise constant regression model*, or *mean shift model.* A recent and important application in the biological area is concerned with *array-based comparative genomic hybridization* (*aCGH*, Pinkel *et al.*, 1998). The goal of such analyses is to identify location of possible damaged genes along a chromosome, involved in cancer or other diseases. In such applications the coefficient β_{0k} represents the expected value of the biological marker in the k-th segment of the chromosome and, usually, no additional linear terms are included in the model.

Several approaches have been implemented for examining the problem of change-point detection in piecewise constant regression models.

Most of the existing techniques to estimate model (1.5) make use of an optimization criterion, such as likelihood (Horváth, 1993), quasi-likelihood (Braun *et al.*, 2000) or least squares (Yao and Au, 1989). However, the main problem arising with such approaches is the non-smoothness of the objective function when the change-point vector is considered as unknown parameter. In fact, due to the non-differentiability with respect to the change point parameters, Newton-Raphson type algorithms result not to be viable. Some authors have tried to circumvent this problem by using *smooth transitions* between the different regimes (Zhou *et al.*, 2008; Pastor-Barriuso *et al.*, 2003; Tishler and Zang, 1981).

The most common approach for point estimation in change-point regression is *grid search* (Bai and Perron, 2003; Hawkins, 2001), namely the final estimate is selected among a grid of candidate values on a given interval.

Computationally efficient grid search approaches have been proposed both for the maximum likelihood (Friedrich *et al.*, 2008) and least squares solutions (Boysen *et al.*, 2009). Such approaches rely on the use of *dynamic programming algorithms* (Jackson *et al.*, 2005), and yield the exact solution reducing the computational cost from $O(n^K)$ to $O(n^2)$ for any K.

More recently, *multiscale penalization methods* have been used for inference about change-points (Frick *et al.*, 2014). In this works, the particular structure of the constraints allows the inclusion of *pruning steps* in the grid-search algorithm, that, under certain conditions, can reduce the computational complexity to $O(n)$ without much affecting the exactness of the resulting segmentation (Killick *et al.*, 2012).

Other optimization approaches aim at giving approximate solutions in a lower computational time. The most widely used alternative search method is *binary segmentation* proposed by Scott and Knott (1974); among other recent techniques we mention the *genetic algorithm* proposed by Jong *et al.* (2003) and the *Smith-Waterman algorithm* adapted by Price *et al.* (2005) to the problem of change-points.

For more general situations, nonparametric methods offer various ways to identify abrupt changes in the regression function, essentially based on the study of first derivatives (Loader *et al.*, 1996). Among recent examples of application of such *data denoising techniques*, we mention the work of Hsu *et al.* (2005), based on the use of *wavelets* (Donoho and Johnstone, 1995).

Yet other approaches rely on *penalized least squares regression*, which makes use of L_1 (Eilers and De Menezes, 2005; Huang *et al.*, 2005) and L_0 (Rippe *et al.*, 2012) penalties, or combined penalties as in the *fused lasso* approach (Tibshirani and Wang, 2008). Note that point estimates for the parameters in model (1.5) are not provided by these methods.

Another important group of methods developed to detect change-points without parametric assumptions concerns *segmentation techniques*, essentially aiming at partitioning observations into contiguous regions of the variable x. Some examples include *circulary binary segmentation* (Venkatraman and Olshen, 2007; Olshen *et al.*, 2004), *hidden Markov models* (Fridlyand *et al.*, 2004; Guha *et al.*, 2008), *resampling techniques* (Lai and Zhao, 2005) and *clustering algorithms* (Wang *et al.*, 2005).

Among other approaches we mention a recent technique based on data transformation and *segmented regression* (Muggeo and Adelfio, 2011), and the area of *Bayesian approaches* (Rigaill *et al.*, 2012; Fearnhead, 2006).

Most of the cited literature is also concerned with the estimation of the optimal number K of change-points. In this case, an additional term penalizing for the model complexity is usually added to the objective function. Because of the likelihood irregularities, the standard Bayes Information Criterion (BIC) does not work well in this framework; in fact, typically it leads to an overestimation of the number of segments (Picard *et al.*, 2005). For this reason, some alternatives have been proposed, such as the modified BIC (Zhang and Siegmund, 2007) or adaptive criteria (Lavielle, 2005).

1.4 Hypothesis testing and interval estimation

Hypothesis testing and interval estimation represents quite a difficult task in change-point models. In fact, most of the usual regularity conditions do not hold in such contexts. As pointed out in Hawkins (1980), the inferential theory depends strongly on whether or not continuity at the change-point is assumed. In particular, the piecewise linear model presents the greatest difficulties; in fact, the maximum likelihood estimate is given by an interval, so that the statistical model is not fully identifiable; besides, the likelihood function is not even once differentiable with respect to the change-points. This creates difficulties with the use of the 'usual' asymptotic chi-squared theory for the test statistics.

Several authors have proposed methods to extend the standard theory to models with change-points, mainly focusing on three different aspects: the development of a statistical test for evaluating the non-existence of change-points, the construction of a confidence interval for the change-point and the construction of a confidence interval for the expected values μ_i.

In the hypothesis testing context, some tests are designed for a specific alternative, usually the existence of a single shift (Worsley, 1983) or multiple shifts (Bai and Perron, 2003), while generalized fluctuation tests (Kuan and Hornik, 1995) do not assume a particular pattern of deviation from the null hypothesis. Zeileis (2005) discusses a unified approach for testing parameter instability. The basic problem with this approaches is that the asymptotic distributions of the usual test statistics are non-standard (Davies, 1987). However, under appropriate conditions, inference can be based on the stochastic process theory (Andrews, 1993).

Different methods have been proposed to derive confidence intervals for a change-point. Several authors have studied the asymptotic behaviour of change-point estimators (Hinkley, 1970; Bai and Perron, 2003) to derive Wald-type confidence sets, also using bootstrapping methods (Hušková and Kirch, 2008). Other approaches use the Likelihood Ratio test evaluated on a grid of candidate values for the change-point, and derive the confidence interval as the set of values that cannot be rejected as the true change-point (Worsley, 1986). The confidence set may well include disconnected regions, and this may indicate the presence of more than one change-point.

A more general problem concerns the construction of confidence intervals for the expected values μ_i. Frick *et al.* (2014) provide quite a detailed discussion about the topic, and derive asymptotic results to determine 'honest' confidence intervals for an unknown step function in exponential family regression.

1.5 Aim and contribution of the thesis

The aim of the thesis is to set up a novel, simple and flexible iterative algorithm for maximum likelihood estimation in change-point models. For sake of simplicity we focus on the piecewise constant model (1.5) with a single change-point ($K = 1$), and ignore possible fixed terms, namely $\eta_i = 0$. We therefore rewrite the model using a single regression equation,

$$\mathbb{E}[Y_i|x_i] = \beta_0 + \beta_1 I(x_i > \psi),$$

where $I(\cdot)$ is the indicator function, and use an iterative algorithm to estimate the model.

As discussed in Lavielle (1999), when the number K of change-points is known, the best partition is given by the global optimization of the objective function; in fact, such approach ensures convergence of the change-point estimator to the true change-points. Besides, as pointed out in Picard *et al.* (2005), the global optimum can only be provided by some exact grid search algorithm, and not by other algorithms.

Grid search approaches based on dynamic programming represent, nowadays, the most powerful and widespread tool for change-point detection; however, there are frameworks in which grid search appears to be unfeasible or difficult to apply.

For example, grid search turns out to be unfeasible in models with subject specific change-points modelled by *random effects*:

$$\mathbb{E}(Y_{ij}|x_{ij}, b_{0i}, b_{1i}, p_i) = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})I(x_{ij} > \{\psi + p_i\}),$$

where b_{0i}, b_{1i} and p_i are zero-mean random effects.
Grid search is also difficult to apply when the change-point is a function of *parameters on unbounded supports*:

$$\mathbb{E}(Y_i|x_i, v_i) = \beta_0 + \beta_1 I(x_i > \theta_0 + \theta_1 v_i);$$

$\psi_i = \theta_0 + \theta_1 v_i$ is a linear change-point, v_i is an additional covariate.
Sometimes, grid search is virtually feasible, but dynamic programming can not help to reduce the computational cost; it is the case of models with *change-points in several covariates*:

$$\mathbb{E}(Y_i|x_{1i}, x_{2i}) = \beta_0 + \beta_1 I(x_{1i} > \psi_1) + \beta_2 I(x_{2i} > \psi_2).$$

Note that, even for the multiple change-point model (1.5) with a single covariate, dynamic grid search has a $O(n^2)$ computational cost, and a *very large sample size* remains a concern.
The idea we develop represents an attempt to handle the change-points like typical parameters in regression models, in order to make estimation feasible in quite general frameworks.
The rest of this thesis is organized as follows. In Chapter 2 we focus on the simple model with a single change-point and illustrate the proposed iterative algorithm. In Chapter 3 the algorithm is extended to estimate random effect models, while in Chapter 4 it is extended to estimate models

with change-point modelled by a linear function. Chapter 5 is devoted to some simulation studies evaluating the properties of the proposed estimators. Extension to multiple change-point models is discussed in Chapter 6, while Chapter 7 is devoted to discussion and future work. The proposed algorithms have been implemented in the *R* environment, and a brief description of function usage is given in the Appendix.

Chapter 2

Single change-point models

In this chapter we focus on the simplest model with a single shift and relevant change-point in the mean level, and illustrate the proposed iterative algorithm to estimate the change-point without grid searching. Page (1955) was among the first authors dealing with the detection of a unique change-point, based on the analysis of sequences of test statistics (Kim and Siegmund, 1989). Most of the literature has then developed for the multiple change-point framework, of which the single change-point scenario represents a special case.

2.1 Model definition

When a unique change-point occurs ($K = 1$) the piecewise constant regression model (1.5) takes the form

$$\mathbb{E}[Y_i] = \beta_{0k} \qquad x_i \in (\psi_{k-1}, \psi_k], \qquad k = 1, 2. \tag{2.1}$$

There is only one point ψ_1 at which the mean level of Y shifts instantaneously from β_{01} to β_{02}, while $\psi_0 = x_{(1)}$ and $\psi_2 = x_{(n)}$; for sake of simplicity we ignore possible fixed terms, namely $\eta_i = 0$.

The main idea developed in this thesis is built from the expedient of rewriting model (1.5) (and its generalizations) using a single regression equation, relying on the use of indicator functions. For example, model (2.1) simply becomes

$$\mu_i = \beta_0 + \beta_1 I(x_i > \psi), \tag{2.2}$$

where $\mu_i = \mathbb{E}[Y_i|x_i]$, $\beta_0 = \beta_{01}$, $\beta_0 + \beta_1 = \beta_{02}$, and $\psi = \psi_1$. All the model parameters, including the change-point, are unknown and have to be estimated.

The main problem is that model (2.2) is clearly non-linear, and the relevant likelihood is non-differentiable. For this reason, maximum likelihood estimation cannot be performed through standard methods such as Newton-Raphson. Figure 2.1 shows a simulated dataset with $n = 100$, $x_i = i, \beta_0 = 2$, $\beta_1 = 1.5$, $\psi = 30$ and $Y_i \sim \mathcal{N}(\mu_i, 1.5^2)$; Figure 2.2 displays the profile log-likelihood with respect to ψ. Note how profiling the log-likelihood yields a highly wiggly step function: if $x_i \neq x_{i+1}$, all the values of ψ between x_i and x_{i+1} (excluded), lead to the same log-likelihood. Therefore, the only feasible approach to find the maximum likelihood solution appears to be grid searching among the values of the explanatory variable; actually, the search task can be reduced to the observed distinct values of x, at which the log-likelihood exhibits a jump. We use notation x_j, $j = 1, 2, \ldots, n_d$, to indicate the n_d distinct observations of x, $[x_{\hat{j}}, x_{\hat{j}+1})$ to indicate the 'optimum interval' associated with the maximum likelihood estimate and, by convention, $\hat{\psi} = x_{\hat{j}}$. For example, estimating model (2.2) via grid search in the simulated dataset yields $\hat{\beta}_0 = 1.54, \hat{\beta}_1 = 2.16$ and $\hat{\psi} = 30$ (Figure 2.1).

In many applications, x simply represents an index variable $i = 1, 2, \ldots, n$, and therefore $x_i = i$ and $n_d = n$; in our framework this restriction is not necessary, and any continuous variable may be employed. In fact, the algorithm we propose yields a point estimate for the change-point, which is not necessarily equal to some observed x_j.

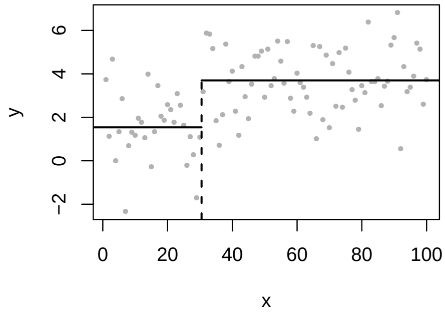

Figure 2.1: *Simulated dataset illustrating model (2.2).* $\hat{\beta}_0 = 1.54$ *(left segment),* $\hat{\beta}_0 + \hat{\beta}_1 = 3.7$ *(right segment) and* $\hat{\psi} = 30$ *(dashed line).*

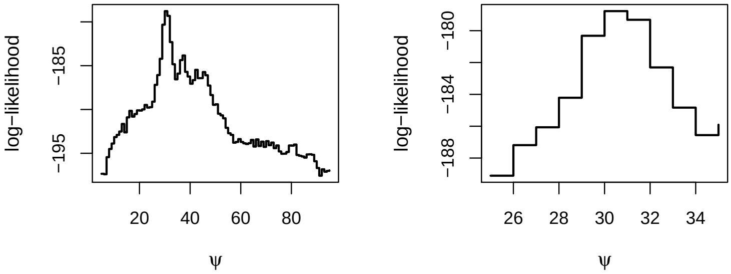

Figure 2.2: *Profile log-likelihood for simulated data. Left panel: whole range of variation. Right panel: neighbourhood of the solution.*

2.2 Methods

In order to estimate all the model parameters (β_0, β_1, ψ) of model (2.2) we write the indicator function $I(x_i > \psi)$ via

$$I(x_i > \psi) = \frac{1}{2}\frac{x_i - \psi}{|x_i - \psi|} + \frac{1}{2}, \tag{2.3}$$

for $x_i \neq \psi$. This identity, when placed in (2.2) gives

$$\begin{aligned} \mu_i &= \beta_0 + \beta_1 \left(\frac{1}{2}\frac{x_i - \psi}{|x_i - \psi|} + \frac{1}{2} \right) \\ &= \beta_0 + \beta_1 \left(\frac{1}{2}\frac{x_i}{|x_i - \psi|} + \frac{1}{2} \right) + (-\beta_1 \psi) \left(\frac{1}{2}\frac{1}{|x_i - \psi|} \right) \\ &= \beta_0 + \beta_1 z_i(\tilde{\psi}) + \gamma w_i(\tilde{\psi}), \end{aligned} \tag{2.4}$$

where

$$\gamma = -\beta_1 \psi, \tag{2.5}$$

and the auxiliary or ‘working’ covariates are

$$z_i(\tilde{\psi}) = \left(\frac{1}{2} + \frac{1}{2}\frac{x_i}{|x_i - \tilde{\psi}|} \right) \quad \text{and} \quad w_i(\tilde{\psi}) = \left(\frac{1}{2}\frac{1}{|x_i - \tilde{\psi}|} \right), \tag{2.6}$$

with $\tilde{\psi}$ meaning an approximate value. Notice model (2.2) has been converted in the simpler linear form (2.4).

For reasons to be clarified later (see Chapter 3), it could be more convenient to express the change-point as a ‘usual’ linear parameter, by slightly modifying the working covariate w:

$$\mu_i = \beta_0 + \beta_1 z_i(\tilde{\psi}) + \psi w_i'(\tilde{\beta}_1, \tilde{\psi}), \tag{2.7}$$

where $w_i'(\tilde{\beta}_1, \tilde{\psi}) = -\tilde{\beta}_1 w_i(\tilde{\psi})$ and, as before, $\tilde{\beta}_1$ is an approximate value. Formulas above suggest the following simple iterative algorithm:

1. choose a *starting value* $\tilde{\psi}$;

2. compute the *working covariates* (2.6);

3. estimate the *working linear model* (2.4) and extract $\hat{\beta}_1$ and $\hat{\gamma}$;

4. *update* the change-point value via

$$\hat{\psi} = -\frac{\hat{\gamma}}{\hat{\beta}_1}; \tag{2.8}$$

5. set $\tilde{\psi} = \hat{\psi}$ and *iterate* 2 to 4 until convergence.

The algorithm outlined above looks quite simple, but unfortunately its plain implementation does not always work in practice. In fact, there are two main pitfalls that should be warned. First, the log-likelihood has typically many local optima, and second, x_i values close to $\tilde{\psi}$ may cause computational troubles, since denominators $|x_i - \tilde{\psi}|$ in (2.6) go to zero.

The next section illustrates how to circumvent both problems by moving the x_is away from the approximate $\tilde{\psi}$.

2.2.1 Rescaling x values

To avoid $|x_i - \tilde{\psi}| \to 0$ in (2.6), the idea is moving the x_is away from the approximate value $\tilde{\psi}$. Consider a standard linear transformation to scale a vector $\boldsymbol{x}$ assuming values on a given interval $[a, b]$ to obtain a new vector

$\boldsymbol{x}'$ into a given interval $[a', b']$. Clearly, such transformation must satisfy

$$\frac{x_i - a}{b - a} = \frac{x_i' - a'}{b' - a'},$$

so that

$$x_i' = a' + (x_i - a)\frac{b' - a'}{b - a}. \tag{2.9}$$

The proposal consists into rescaling the covariate values of the two intervals $[x_{(1)}, \tilde{\psi}]$ and $(\tilde{\psi}, x_{(n)}]$ into new intervals $[x_{(1)}, \tilde{\psi}^-]$ and $(\tilde{\psi}^+, x_{(n)}]$ having extremes moved away from $\tilde{\psi}$. To compute the left ($\tilde{\psi}^-$) and right ($\tilde{\psi}^+$) 'threshold' values, we use a rescaling factor $c \in (0, 1)$ such that

$$\tilde{\psi}^- = \tilde{\psi} - c(\tilde{\psi} - x_{(1)}), \qquad \tilde{\psi}^+ = \tilde{\psi} + c(x_{(n)} - \tilde{\psi}).$$

By noting that

$$1 - c = \frac{\tilde{\psi}^- - x_{(1)}}{\tilde{\psi} - x_{(1)}} = \frac{x_{(n)} - \tilde{\psi}^+}{x_{(n)} - \tilde{\psi}},$$

we use (2.9) to obtain a rescaled variable x':

$$x_i' = x_{(1)} + (x_i - x_{(1)})(1 - c) \tag{2.10}$$

for $x_i \in [x_{(1)}, \tilde{\psi}]$, and

$$x_i' = \tilde{\psi}^+ + (x_i - \tilde{\psi})(1 - c) \tag{2.11}$$

for $x_i \in (\tilde{\psi}, x_{(n)}]$. Figure 2.3 illustrates the rescaling for the simulated data in Figure 2.1. In the left panel the covariate is not rescaled, i.e $c = 0$, while the right panel shows the effect of a rescaling factor $c = 0.1$ when $\tilde{\psi} = 50.5$ (the median of x) is chosen as approximate value.

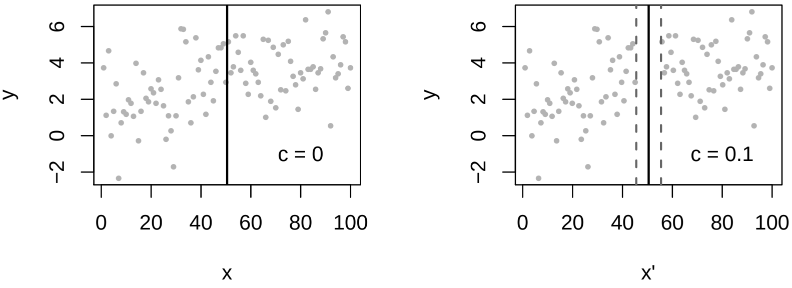

Figure 2.3: *Illustrating the rescaling. Left panel: originary covariate (*$c = 0$*). Right panel: rescaled covariate (*$c = 0.1$*). The rescaling induces a point-free interval (dashed lines) in the neighbourhood of* $\tilde{\psi}$ *(solid line).*

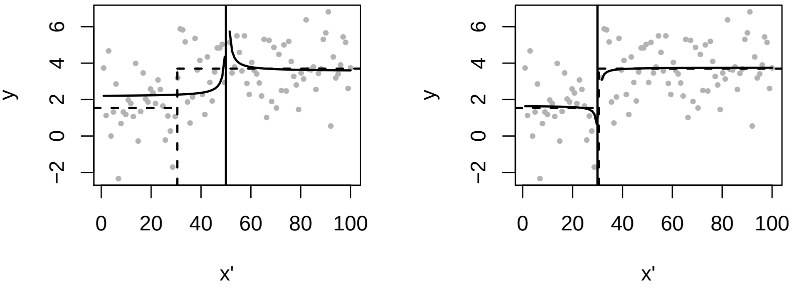

Figure 2.4: *Fitted working regression functions. The rescaling factor is* $c = 0.01$. *The approximate values* $\tilde{\psi}$ *are 50.5 (left panel) and* 30.5 *(right panel). The working model approximates a step function, with a jump in the neighbourhood of the approximate value* $\tilde{\psi}$ *(vertical lines). When* $\tilde{\psi}$ *is close to* $\hat{\psi} = 30$, *the working model approximates the solution (dashed line).*

Note that, in Figure 2.3, the neighbourhood of $\tilde{\psi}$ does not include any rescaled observation x'_i. The rescaled covariate is used to compute the working covariates (2.6), and this avoids computational troubles.

Figure 2.4 shows the fitted working regression function for the simulated data in Figure 2.3; we use $c = 0.01$ and two different starting values $\tilde{\psi}$ (vertical lines). Note that the working model approximates a step function with a jump in the neighbourhood of the approximate value $\tilde{\psi}$. At convergence, when $\tilde{\psi}$ is close to $\hat{\psi} = 30$ (right panel), the working model closely approximates the grid search solution (dashed line). It is also worth noting that the updated estimate $\hat{\psi}$ does not have a graphical feedback, and therefore it is not visualized in Figure 2.4.

2.3 The iterative algorithm in detail

To illustrate how the algorithm works, Figure 2.5 refers to the same dataset illustrated in Figure 2.1. Left panels show the 'true' profile log-likelihood for ψ in model (2.2) and, superimposed, the relevant 'working' profile log-likelihoods in model (2.4) for two different values of c and $\tilde{\psi} = 50$. The black points on the ψ axis indicate the global 'exact' solution obtained by grid search (30), while the grey points indicate the integer part of the updated solutions $\hat{\psi}$ at the iteration.

Given the approximate value $\tilde{\psi}$, the working linear model (2.4) leads to a smooth log-likelihood, with a unique solution $\hat{\psi}$ to be used as starting value in the next iteration.

As pointed out, the scale change from x to x' moves observations away from the updated estimate $\tilde{\psi}$, and prevents computational troubles. Besides, the magnitude of the rescaling affects the variations $|\hat{\psi} - \tilde{\psi}|$ at the iteration,

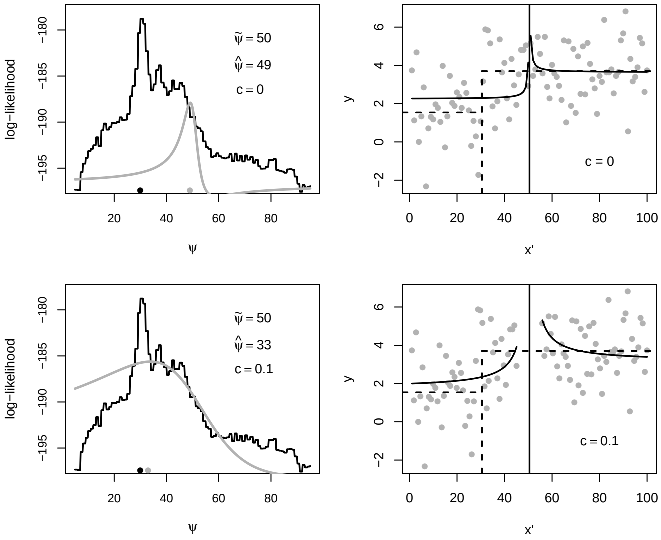

Figure 2.5: *Example of two algorithm iterations. Left panels: true (black lines) and working (grey lines) profile log-likelihoods based on the same approximate $\tilde{\psi}$ = 50 but different values of c. The rescaling affects the updated estimate $\hat{\psi}$ (grey points at the bottom) which moves closer to the grid search solution (black points at the bottom). Right panels: fitted values for the working model (solid lines) and grid search solution (dashed lines).*

which become larger as c increases; for the example data of Figure 2.5 it is $|49-50| = 1$ without rescaling the x_is ($c = 0$), while it is $|33-50| = 17$, much larger, when $c = 0.1$. This helps the algorithm to skip most of the spurious optima and to go towards the global solution. More detailed explanation of the role of c in the estimation procedure and some hints for a proper choice are provided in Chapter 5.

Figures 2.6 and 2.7 illustrate in detail all the iterations of the algorithm until convergence, for the simulated data in Figure 2.1.

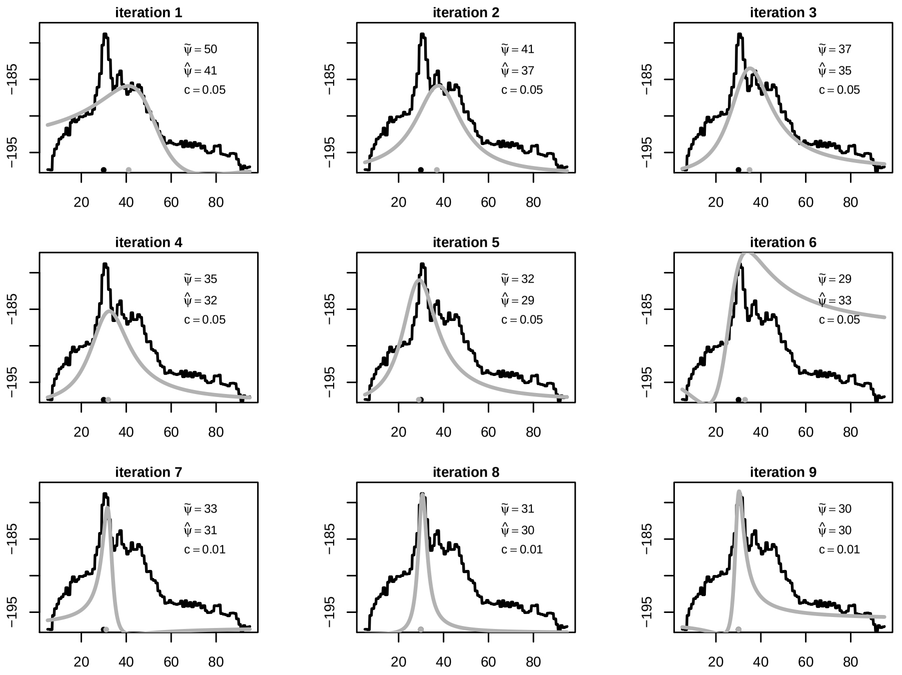

Figure 2.6: *The proposed iterative algorithm in action: each panel represents an iteration. Starting value:* $\tilde{\psi} = 50$*. Black lines: true profile log-likelihood. Grey lines: working profile log-likelihoods. Black points indicate the grid search solution (*$\hat{\psi} = 30$*), grey points indicate the updated solutions coming from the working model.* $c = 0.05$ *until iteration 6, afterwards it is* $c = 0.05 \times 0.2 = 0.01$ *(*$d = 0.2$*).*

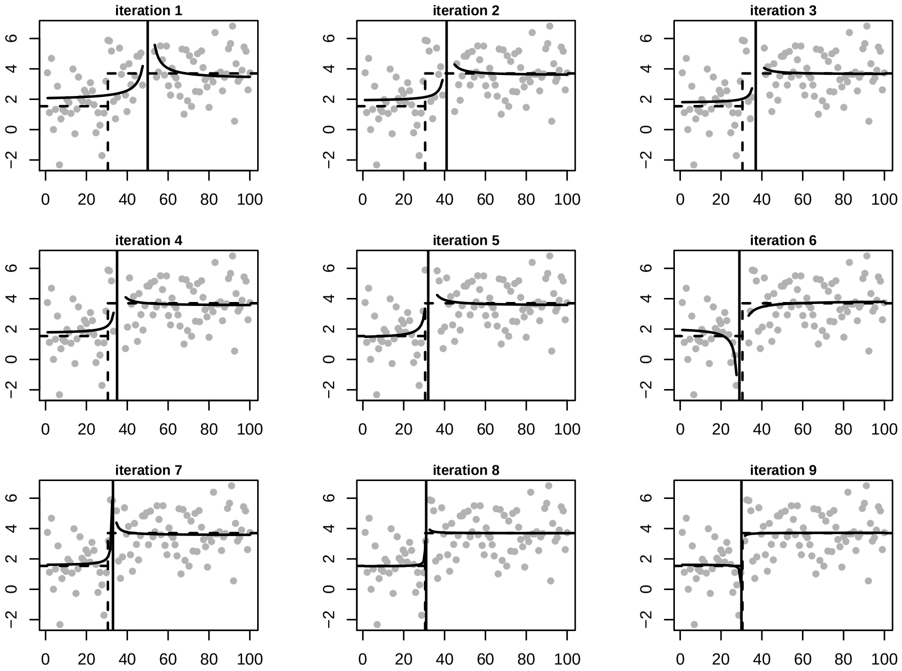

Figure 2.7: *Fitted values for the working models at each iteration of the algorithm in Figure 2.6. The working regression function approximates a step function, with a jump in the neighbourhood of the approximate value $\tilde{\psi}$ (vertical lines). At convergence the working model closely approximates the grid search solution (dashed line).*

Note that the updated estimate goes beyond the solution at iteration 5, moving back to the opposite direction at iteration 6; this suggests to reduce the rescaling factor, for example by multiplying it by a reduction factor, say $d \in (0, 1)$. In most of the examples presented through this thesis we use $d = 0.5$ in the spirit of *step halving* (Jennrich and Sampson, 1968). In some cases we use $d = 0.2$ for a faster but somewhat unstable convergence, or $d = 0.8$ for a more stable but slower algorithm. In this example, c is decreased from 0.05 to 0.01 ($d = 0.2$).

To sum up, the steps of the algorithm are:

1. choose a *starting value* $\tilde{\psi}$;
2. choose a *rescaling factor* c;
3. *rescale* x using (2.10) and (2.11) to obtain x';
4. compute the *working covariates* (2.6) using x';
5. estimate the *working linear model* (2.4) and extract $\hat{\beta}_1$ and $\hat{\gamma}$;
6. use (2.8) to *update* the change-point value;
7. *decrease c* using d if $\tilde{\psi}$ changes direction;
8. set $\tilde{\psi} = \hat{\psi}$ and *iterate* 3 to 7 until convergence.

2.4 Starting value

As in other non-linear models, the choice of a ‘good’ starting value represents a crucial aspect for the algorithm to be successful. Rough indications about the change-point location may be based, for example, on graphical inspections or a priori knowledge. Limited experience suggests to use the intermediate value $\frac{x_{(1)}+x_{(n)}}{2}$ of the range of x, or alternatively, to evaluate the true log-likelihood on a small grid of equally spaced values along the range of x, and choose the value which gives the best model. Ultimately, a small sensitivity analysis can be helpful.

2.5 Convergence

While any value in $[x_{\hat{j}}, x_{\hat{j}+1})$ gives the same likelihood, the proposed algorithm returns a single estimate. As usual, there are two possible criteria to declare convergence: evaluating the magnitude of the absolute variation of the updated estimate, namely $|\hat{\psi} - \tilde{\psi}|$, or evaluating the log-likelihood variation $|\ell(\hat{\psi}) - \ell(\tilde{\psi})|$. Figure 2.8 shows the values of the maximized log-likelihood for the iterative algorithm depicted in Figure 2.6: the solid line represents the value of the true log-likelihood for the updated value $\hat{\psi}$, the dashed line represents the maximized working log-likelihood at the current $\hat{\psi}$. Figure 2.9 shows the values of the variation $\hat{\psi} - \tilde{\psi}$.
Since the model is non-linear, the likelihood does not necessarily increase monotonically through iterations; besides, the objective function is strongly irregular in our framework, and a large absolute variation $|\hat{\psi} - \tilde{\psi}|$ may be associated to a plateau or a decrease in the likelihood also in the very first iterations. On the other hand, the absolute variation $|\hat{\psi} - \tilde{\psi}|$ is higher at the beginning and tends to decrease throughout iterations. Note that the sign of the variation $\hat{\psi} - \tilde{\psi}$ changes only once $\hat{\psi}$ approaches the solution, and this warns that the rescaling factor should be reduced using d.
Repeatedly decreasing the rescaling factor c makes the absolute variation $|\hat{\psi} - \tilde{\psi}|$ decrease until convergence; for some small $\Delta > 0$, we therefore use

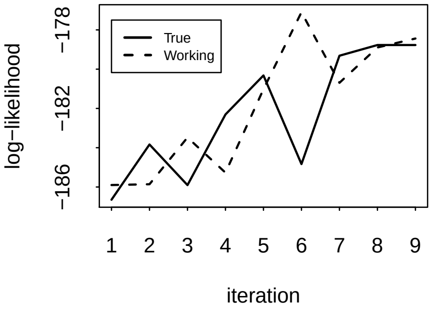

Figure 2.8: *True (solid line) and working (dashed line) maximized log-likelihood throughout iterations. Even in the very first iterations, the likelihood does not increase monotonically, since the model is non-linear and highly irregular.*

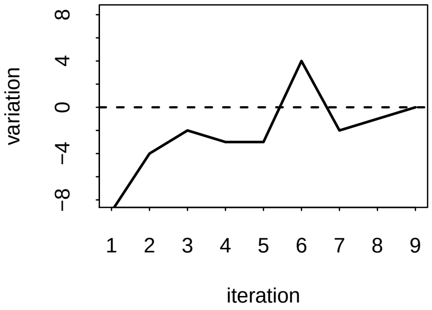

Figure 2.9: *Estimate variation throughout iterations. The absolute variation tends to decrease. The sign of the variation changes at iteration 6 (the solid line crosses the dashed line corresponding to 0) because the updated estimate (*$\tilde{\psi} = 29$*) approaches the solution (*$\hat{\psi} = 30$*); this warns to reduce the rescaling factor, in the spirit of step halving.*

$|\hat{\psi} - \tilde{\psi}| < \Delta$ as a convergence rule. A rule of thumb suggests to fix Δ at some fraction of the smallest distance $x_j - x_{j-1}$, $j = 2, 3, \ldots, n_d$; if we consider the discrete version $x_{\hat{j}}$ (the integer part of $\hat{\psi}$ when $x_i = i$) Δ can be fixed to 0, as in the algorithm in Figure 2.6.

2.6 Remarks

We conclude the description by analyzing the convergence behaviour of the proposed algorithm in practice. We simulate 1000 datasets assuming $x_i = 1, 2, \ldots, 100$ and $\psi = 50$. We use $\tilde{\psi} = 50.5$, $c = 0.02$, $d = 0.5$ and $\Delta = 0.01$ (1% of the distance between the x_is).

Figure 2.10 represents the histogram of the estimates which fall in the interval $[50, 51)$: note that the final solution tends to approach the observed values of the explanatory variable x, 50 or 51.

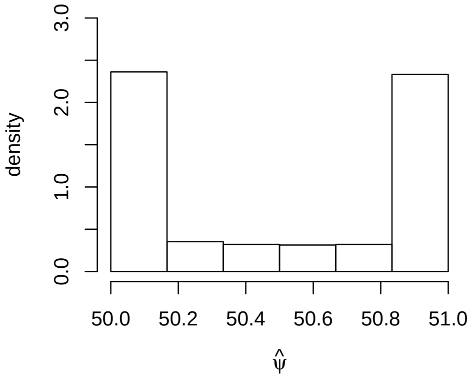

Figure 2.10: *Histogram of the estimates for* 1000 *simulated datasets. Since* $\psi = 50$, *we focus on the interval* $[50, 51)$. *The final solution tends to approach the observed values of the explanatory variable x,* 50 *or* 51.

For this reason, the distribution of the explanatory variable x may influence the performance of the estimator $\hat{\psi}$, and the availability of a uniform and quite densely distributed explanatory variable represents an important regularity condition for accurately estimating ψ.

2.7 Examples

To illustrate the proposed algorithm in practice we apply it to two well known datasets in the ecological and biological literature. Despite the two datasets are very simple, the grid search algorithm results to be computationally more expensive with respect to the proposed algorithm.

2.7.1 River Nile flow data

The first example concerns the annual volume of the Nile River for the years 1871 to 1970 taken from the work of Cobb (1978). This series was examined by several authors, including Balke (1993) and Dumbgen (1991), providing evidence that the Nile River volume experienced a permanent decline in 1899, due to the construction of the first Ashwan dam.

We estimate model (2.2) where $n = 100$ is the length of the time series, x is the year $(1871, 1872, \dots, 1970)$, while Y is the volume of the Nile River (discharge at Ashwan, 10^8 m^3). We assume the Y_is to be independent and Gaussian. Figure 2.11 represents the data and the results yielded by the grid search algorithm (left panel) and the relevant profile log-likelihood for the change-point (right panel).

To estimate the model via the proposed algorithm we set $\tilde{\psi} = 1904$ as the best value among $\{1887, 1904, 1920, 1937, 1953\}$, and also $c = 0.05$, $d = 0.2$ and $\Delta = 0.01$. The algorithms yields $\hat{\beta}_0 = 1097.97$, $\hat{\beta}_1 = -247.94$ and

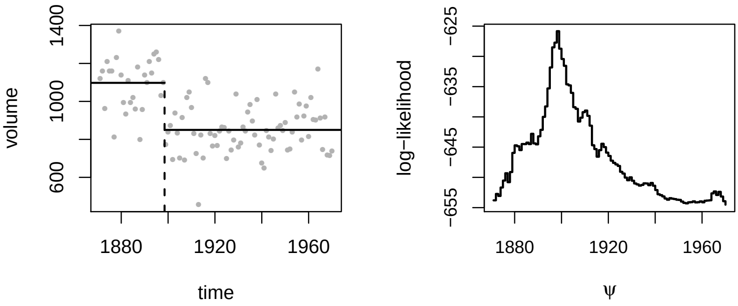

Figure 2.11: *River Nile flow data. Left panel: observed values (points), fitted values (solid lines) and estimated change-point (dashed line) yielded by the grid search algorithm. Right panel: profile log-likelihood.*

$\hat{\psi} = 1898.07$. The grid search algorithm yields $\hat{\beta}_0 = 1097.75, \hat{\beta}_1 = -247.78$ and the 'optimum' interval $[1898, 1899)$ for the change-point. The change-point estimate yielded by the proposed algorithm falls in the optimum interval and requires 4 iterations, while grid search requires the evaluation of $n_d = 100$ candidates models.

2.7.2 Fibroblast cell lines data

We consider another dataset discussed by several authors in the aCGH literature (e.g. Huang *et al.*, 2005). The data consist of single experiments on 15 fibroblast cell lines. The response variable Y is the 'copy number' indicator, namely the normalized average of the log base 2 test over reference fluorescence color ratio, which is 0 in the absence of alterations. By spectral karyotyping, it is known that one alteration is present in chromosomes #9; the explanatory variable x is just a position marker of the cells, and the aim of the analysis is to the detect the damaged genes.

Assuming the Y_is to be independent and Gaussian, the grid search algorithm

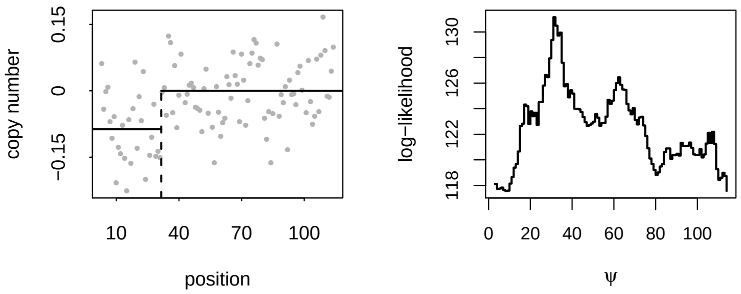

Figure 2.12: *Fibroblast cell lines data. Left panel: observed values (points), fitted values (solid lines) and estimated change-point (dashed line) yielded by the grid search algorithm. Right panel: profile log-likelihood.*

yields the result in Figure 2.12, which represents the data and the fitted values (left panel) and the relevant profile log-likelihood for the change-point (right panel).

As in the previous example, we set $\tilde{\psi} = 40$ as the best candidate value among $\{21, 40, 58, 77, 95\}$, and also $c = 0.05$, $d = 0.2$ and $\Delta = 0.01$. The proposed algorithm yields $\hat{\beta}_0 = -0.09$, $\hat{\beta}_1 = 0.09$ and $\hat{\psi} = 31.95$. The grid search algorithm yields $\hat{\beta}_0 = -0.09$, $\hat{\beta}_1 = 0.09$ and the optimum interval $[31, 32)$ for the change-point. The change-point estimate from the proposed algorithm falls in the optimum interval and requires 9 iterations, while grid search requires the evaluation of $n_d = 107$ candidates models.

2.8 By product goals

The approach we propose in this thesis may result to be helpful, in some circumstances, not only for point estimation; here we consider a simple example concerned with interval estimation for the expected values μ_i.

The simplest approach to determine a 95% confidence interval for μ_i in simple linear regression models with independent and Gaussian responses is the usual

$$[\hat{\mu}_i - t_{n-p,0.975} \times \text{s.e.}(\hat{\mu}_i) \ \ , \ \ \hat{\mu}_i + t_{n-p,0.975} \times \text{s.e.}(\hat{\mu}_i)], \tag{2.12}$$

where $t_{n-p,0.975}$ is the 97.5 percentile point of the Student's t-distribution with $n - p$ degrees of freedom, p is the number of columns of the model matrix and s.e.$(\hat{\mu}_i)$ is the estimated standard error of $\hat{\mu}_i$.

The simplest approach to compute confidence intervals in model (2.2) consists on determining the maximum likelihood estimate $\hat{\psi}$ via grid search, fix ψ at $\hat{\psi}$ in (2.2) as if it was the true value, and use (2.12). Note that the resulting model matrix X_{GS} (grid search) is $n \times 2$, having i-th row $[1, I(x_i > \hat{\psi})]$. Of course, the resulting confidence interval is not any longer exact for at least two reasons: first, we neglect the additional variability induced by $\hat{\psi}$, second, using $\hat{\psi}$ strongly affects the distribution of $\hat{\mu}_i$. This reflects on the actual coverage level of the confidence interval which is, in general, lower than the nominal 0.95.

Since model (2.2) is closely approximated by the working model (2.4) evaluated at convergence of the proposed algorithm, the latter can also be used to construct confidence intervals for the μ_is. The resulting model matrix X_{PA} (proposed approximation) is now $n \times 3$, having i-th row $[1, z_i, w_i]$. Model (2.4) has one additional unknown parameter, and therefore it should better account for the variability of model (2.2), improving the performances of the confidence intervals.

To clarify this aspect, we perform a simple simulation study. We assume $n = 20$, $x_i = i$, $\beta_0 = 2$, $\beta_1 = 0.4$, $\psi = 13$ and $Y_i \sim \mathcal{N}(\mu_i, 0.5^2)$, simulate 100 datasets, and compare the empirical coverage levels of confidence intervals (2.12) for the μ_is with respect to models associated with X_{GS} and X_{PA}. Figure 2.13 displays the results.

Due to the wrong assumptions about the distribution of the fitted values $\hat{\mu}_i$, the empirical coverage levels result to be far from the nominal ones (solid line), especially in the neighbourhood of the true change point ψ. However, due to the additional parameter γ, confidence intervals based on the proposed approximation (dashed line) perform slightly better.

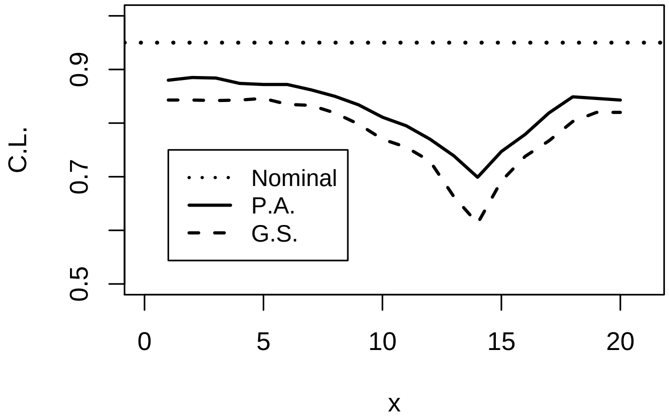

Figure 2.13: *Empirical coverage levels of the* 95% *confidence intervals for μ_i when ψ is fixed at $\hat{\psi}$ (dashed line) and when using the proposed approximation (solid line). The coverage level is closer to the nominal one (dotted line) when using the proposed approximation.*

Chapter 3

Extension 1: *random effects*

One of the most noteworthy advantages of the algorithm proposed in Chapter 2 is that it straightforwardly extends to general change-point models. In this chapter we focus on mixed models with random change-points, as sketched in the introduction.

Several statistical methods for detecting subject specific change-points have been proposed for longitudinal data, and most of them rely on the Bayesian paradigm (see Dominicus *et al.*, 2008; Hall *et al.*, 2003; Kiuchi *et al.*, 1995). Within the likelihood framework, grid search approaches turn out to be unfeasible, unless constancy of the change-point among subjects is assumed (Hall *et al.*, 2000). Related works are developed only for the case in which continuity of the regression function is assumed at the change-point: Muggeo *et al.* (2014) use a linear approximation of the segmented function, while Jacqmin-Gadda *et al.* (2006) rely on smooth transition models.

For the discontinuous case, Jackson and Sharples (2004) use a mixture of hierarchical longitudinal models where a Weibull prior distribution is assumed for the change-points, while the issue of detecting subject specific change-points within the likelihood framework has not been addressed. We discuss how generalization of the iterative algorithm introduced in Chapter 2 allows inclusion of random effects, both for the regression and the change-point parameters.

3.1 Model definition

Consider model (2.2) for a sample of $i = 1, 2, \ldots, n$ subjects, each one with $j = 1, 2, \ldots, n_i$ measurements, and assume all the parameters are given by the sum of fixed and random effects, namely

$$\begin{aligned} Y_{ij} &= \beta_{0i} + \beta_{1i} I(x_{ij} > \psi_i) + \epsilon_{ij} \\ &= (\beta_0 + b_{0i}) + (\beta_1 + b_{1i}) I(x_{ij} > \{\psi + p_i\}) + \epsilon_{ij}. \end{aligned} \tag{3.1}$$

At $\psi_i = \psi + p_i$, the mean level of Y for subject i shifts instantaneously from β_{0i} to $\beta_{0i} + \beta_{1i}$.

The assumptions of normality and independence among the random effects may appear to be restrictive; for example, in the context of segmented mixed models, Muggeo *et al.* (2014) consider a block diagonal covariance matrix, while Jacqmin-Gadda *et al.* (2006) use the Log-normal distribution for the random change-point. However, for simplicity we assume

$$\begin{bmatrix} b_{0i} \\ b_{1i} \\ p_i \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_0^2 & 0 & 0 \\ 0 & \sigma_1^2 & 0 \\ 0 & 0 & \sigma_\psi^2 \end{bmatrix} \right), \tag{3.2}$$

and, as usual, $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$ *i.i.d.* and independent of the random effects.

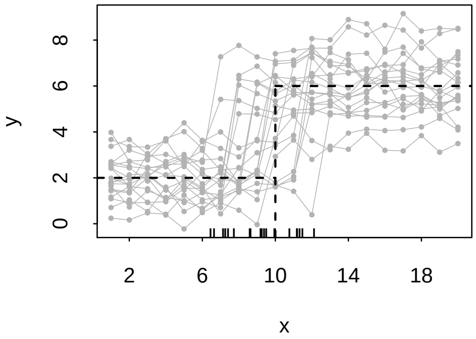

Figure 3.1: *Simulated dataset illustrating the model for $n = 20$ subjects and $n_i = 20$ repeated measurement for each subject. $\beta_0 = 2$ (left segment), $\beta_0 + \beta_1 = 6$ (right segment) and $\psi = 10$ (vertical line). Points at the bottom represent the individual change-points: the variance parameter is $\sigma^2_\psi = 2^2$.*

In this thesis we consider the normality and independence assumptions as valid, while possible generalizations are not discussed. Figure 3.1 shows a simulated dataset with $n = 20$, $n_i = 20$, $x_{ij} = j$, $\beta_0 = 2$, $\beta_1 = 4$, $\psi = 10$, $\sigma^2_0 = 0.7^2$, $\sigma^2_1 = 0.8^2$, $\sigma^2_\psi = 2^2$ and $\sigma^2_\epsilon = 0.4^2$.

3.2 Methods

Likelihood based estimation of model (3.1) represents quite a hard task: (3.1) is apparently non-linear and non-differentiable with respect to the random change-point parameters, and also grid-search here becomes clearly unfeasible. However, it is possible to extend the idea of Chapter 2, with some modifications. In fact, using

$$I(x_{ij} > \psi_i) = \frac{1}{2}\frac{x_{ij} - \psi_i}{|x_{ij} - \psi_i|} + \frac{1}{2}, \tag{3.3}$$

for $x_{ij} \neq \psi_i$, and substituting in (3.1) gives

$$\begin{aligned} Y_{ij} &= \beta_{0i} + \beta_{1i}\left(\frac{1}{2}\frac{x_{ij}-\psi_i}{|x_{ij}-\psi_i|} + \frac{1}{2}\right) + \epsilon_{ij} \\ &= \beta_{0i} + \beta_{1i}\left(\frac{1}{2}\frac{x_{ij}}{|x_{ij}-\psi_i|} + \frac{1}{2}\right) + (-\beta_{1i}\psi_i)\left(\frac{1}{2}\frac{1}{|x_{ij}-\psi_i|}\right) \\ &= \beta_{0i} + \beta_{1i}z_{ij} + \gamma_i w_{ij}, \end{aligned} \tag{3.4}$$

where

$$\gamma_i = -\beta_{1i}\psi_i \tag{3.5}$$

and the auxiliary (or ‘working’) covariates are

$$z_{ij} = \left(\frac{1}{2}\frac{x_{ij}}{|x_{ij}-\tilde{\psi}_i|} + \frac{1}{2}\right) \quad \text{and} \quad w_{ij} = \left(\frac{1}{2}\frac{1}{|x_{ij}-\tilde{\psi}_i|}\right), \tag{3.6}$$

with $\tilde{\psi}_i$ meaning an approximate value of the change-point for subject i; the dependence of the working covariates z_{ij} and w_{ij} on $\tilde{\psi}_i$ has been omitted to simplify the notation.

To better highlight the fixed and the random part of model (3.4) we can write

$$\boldsymbol{Y}_i = [\mathbf{1}_{n_i}, \boldsymbol{z}_i, \boldsymbol{w}_i]\begin{bmatrix}\beta_0 \\ \beta_1 \\ \gamma\end{bmatrix} + [\mathbf{1}_{n_i}, \boldsymbol{z}_i, \boldsymbol{w}_i]\begin{bmatrix}b_{0i} \\ b_{1i} \\ g_i\end{bmatrix} + \boldsymbol{\epsilon}_i. \tag{3.7}$$

To express the change-points directly as linear parameters we write

$$Y_{ij} = \beta_{0i} + \beta_{1i}z_{ij} + \psi_i w'_{ij} + \epsilon_{ij}, \tag{3.8}$$

where

$$w'_{ij} = -\tilde{\beta}_{1i}w_{ij} \tag{3.9}$$

and $\tilde{\beta}_{1i}$ is an approximate value; (3.8) is equivalent to

$$\boldsymbol{Y}_i = [\mathbf{1}_{n_i}, \boldsymbol{z}_i, \boldsymbol{w}_i'] \begin{bmatrix} \beta_0 \\ \beta_1 \\ \psi \end{bmatrix} + [\mathbf{1}_{n_i}, \boldsymbol{z}_i, \boldsymbol{w}_i'] \begin{bmatrix} b_{0i} \\ b_{1i} \\ p_i \end{bmatrix} + \epsilon_i. \tag{3.10}$$

Generalization of the idea proposed in Chapter 2 leads to the following iterative algorithm:

1. choose a vector of (possibly equal) *starting values* $\tilde{\psi}_1, \tilde{\psi}_2, \ldots, \tilde{\psi}_n$;

2. compute the *working covariates* (3.6);

3. *estimate LMM* (3.7) and extract predictions $\hat{\beta}_{1i}$ and $\hat{\gamma}_i$ for each i;

4. *update* the change-point values via

$$\hat{\psi}_i = -\frac{\hat{\gamma}_i}{\hat{\beta}_{1i}}; \tag{3.11}$$

5. set $\tilde{\psi}_i = \hat{\psi}_i$ and *iterate* 2 to 4 until convergence.

Notice model (3.1) is approximated, at each iteration, by the simpler conventional linear mixed model (3.7). To estimate model (3.7) one of the available standard methods can be used. A basic approach is based on the maximization of the marginal likelihood (Pinheiro and Bates, 2000) in which random effects have been integrated out:

$$L(\boldsymbol{\beta}, \gamma, \boldsymbol{\sigma}^2; \boldsymbol{y}) = \prod_{i=1}^{n} \int \int f(\boldsymbol{y}_i \mid \boldsymbol{b_i}, g_i; \boldsymbol{\beta}, \psi) f(\boldsymbol{b_i}, g_i; \boldsymbol{\sigma}^2) d\boldsymbol{b_i} dg_i. \tag{3.12}$$

According to (3.11), random effect predictions are also needed, at each iteration of the algorithm, to update change-point values.

A standard approach looks at this from an empirical Bayes point of view, by considering $f(\boldsymbol{b_i}, g_i; \boldsymbol{\hat{\sigma}^2})$ as the estimated prior density on the random parameters. The estimated posterior density is given by

$$f(\boldsymbol{b_i}, g_i \mid \boldsymbol{y}_i; \boldsymbol{\hat{\sigma}^2}) \propto f(\boldsymbol{y}_i \mid \boldsymbol{b_i}, g_i; \boldsymbol{\hat{\beta}}, \hat{\gamma}) f(\boldsymbol{b_i}, g_i; \boldsymbol{\hat{\sigma}^2}), \tag{3.13}$$

and then we can consider posterior means to get random effect predictions. By virtue of the assumption of independence between random effects, we can write

$$\begin{aligned}
\hat{\beta}_{0i} &= \hat{\beta}_0 + \hat{b}_{0i} = \hat{\beta}_0 + \int b_{0i} f(b_{0i}; \hat{\sigma}_0^2) db_{0i}, \\
\hat{\beta}_{1i} &= \hat{\beta}_1 + \hat{b}_{1i} = \hat{\beta}_1 + \int b_{1i} f(b_{1i}; \hat{\sigma}_1^2) db_{1i}, \\
\hat{\gamma}_i &= \hat{\gamma} + \hat{g}_i = \hat{\gamma} + \int g_i f(g_i; \hat{\sigma}_\gamma^2) dg_i.
\end{aligned} \tag{3.14}$$

3.2.1 Rescaling x values

As discussed in Chapter 2, the x_{ij}s have to be moved away from the change-point values to avoid computational troubles. Extending the idea of Section 2.2.1, we use a vector of rescaling factors $c_1, c_2, \ldots, c_n$, compute

$$\tilde{\psi}_i^- = \tilde{\psi}_i - c_i(\tilde{\psi}_i - x_{i(1)})$$

and

$$\tilde{\psi}_i^+ = \tilde{\psi}_i + c_i(x_{i(n)} - \tilde{\psi}_i),$$

and for each i consider

$$x'_{ij} = x_{i(1)} + (x_{ij} - x_{i(1)})(1 - c_i) \tag{3.15}$$

for $x_{ij} \in [x_{i(1)}, \tilde{\psi}_i]$, and

$$x'_{ij} = \tilde{\psi}_i^+ + (x_{ij} - \tilde{\psi}_i)(1 - c_i) \tag{3.16}$$

for $x_{ij} \in (\tilde{\psi}_i, x_{i(n)}]$, and use x' to compute auxiliary covariates (3.6) and to fit the working linear mixed model (3.7). Once again, decreasing the c_is throughout iterations according to some reduction factor, say $d \in (0, 1)$, is helpful to avoid convergence failures.

The steps of the algorithm are summarized below:

1. choose a vector of (possibly equal) *starting values* $\tilde{\psi}_1, \tilde{\psi}_2, \ldots, \tilde{\psi}_n$;
2. choose a vector of (possibly equal) *rescaling factors* $c_1, c_2, \ldots, c_n$;
3. *rescale* x using (3.15) and (3.16) to obtain x';
4. compute the *working covariates* (3.6) using x';
5. *estimate LMM* (3.7) and extract predictions $\hat{\beta}_{1i}$ and $\hat{\gamma}_i$ for each i;
6. use (3.11) to *update* the change-point predictions;
7. *decrease* c_i, $i = 1, 2, \ldots, n$, using d if $\tilde{\psi}_i$ changes direction;
8. set $\tilde{\psi}_i = \hat{\psi}_i$ and *iterate* 3 to 7 until convergence.

3.3 Starting values

To estimate model (3.1) a vector of approximate change-point values is needed. Basing the choice on the a priori knowledge or graphical inspection may be unfeasible in the presence of many subjects. A simple approach can be considering the same approximate value for each subject, fixed at some intermediate location; for example, a rough approximation can be given by the change-point estimate in model (2.2) where the change-point is assumed to be the same for all subjects. An alternative approach could be selecting $\tilde{\psi}_i$ at random in the neighbourhood of such intermediate location. Of course, a small sensitivity analysis can turn out to be helpful.

3.4 Convergence

In the same spirit of Section 2.5, the variations $\hat{\psi}_i - \tilde{\psi}_i$ are monitored throughout iterations to assess convergence. The sign of the variation for the $i-$th subject changes once $\hat{\psi}_i$ approaches the solution, and warns that the rescaling factor c_i should be reduced. Repeatedly using d to decrease the rescaling factors makes the absolute variations $|\hat{\psi}_i - \tilde{\psi}_i|$ to decrease; therefore, for some small $\Delta > 0$, we use

$$\max_{i=1,2,\dots,n}|\hat{\psi}_i - \tilde{\psi}_i| < \Delta$$

as a convergence rule.

3.5 Remarks

We conclude with some remarks about the presented methodologies.

Note that, despite (3.2) define the distribution of random effects in (3.10), this model is not suitable to use as a working model, since it would require twice as many approximate values as model (3.7).

On the other hand, much attention must be payed in the interpretation of the working model (3.7); in fact, assuming

$$\begin{bmatrix} b_{0i} \\ b_{1i} \\ g_i \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_0^2 & 0 & 0 \\ 0 & \sigma_1^2 & 0 \\ 0 & 0 & \sigma_\gamma^2 \end{bmatrix} \right), \tag{3.17}$$

would be attractive but not completely correct.

Actually, γ_i is the product of two independent Gaussian distributions, and therefore it is Gaussian only approximately; in addition, β_{1i} and γ_i are clearly dependent. However, $\gamma = -\beta_1 \psi$ exactly, so that the assumption may not result to be too restrictive when the goal is just to update the change-point predictions $\hat{\psi}_i$.

Therefore, we propose to use working model (3.4) to update the change-point predictions until convergence, and model (3.10) to get also the variance parameter estimates at convergence. Simulation studies carried out in Section 5.2 show that the proposed approach works correctly in simple scenarios.

Initializing the proposed algorithm requires a vector of approximate values $\tilde{\psi}_i$; therefore, we are implicitly assuming that a change-point exists for all the subject, namely $\beta_{1i} \neq 0 \; \forall i$. The problem of detecting and considering possible subjects without change-point is not addressed in this thesis.

Finally, choosing a vector of rescaling factors c_i may appear to be unsuitable; however, we have experienced that using the same value for each subject works well in practice.

3.6 Example

Jackson and Sharples (2004) analyzed data from 204 patients receiving lung transplant. In the first months after the transplant the patients have an high risk of complications, such as rejection episodes and infections, and thus lung conditions, evaluated via the forced expiratory volume in 1 second (FEV_1), need to be monitored constantly. For each subject, different measurements are available with decline patterns being smooth or changing suddenly.

Unlike Jackson and Sharples (2004) relying on the Bayesian paradigm, we estimate a model with random effects in both the regression and the change-point parameters in a likelihood based framework. We focus on 12 subjects having an apparent abrupt change in their FEV_1 profiles and fit model (3.1), where Y_{ij} is the j-th FEV_1 measurement (baseline percentage) for patient i and x_{ij} is the month at which the measurement is taken after the transplant. To initialize the proposed algorithm we perform a small sensitivity analysis varying the approximate value $\tilde{\psi}_i$ (the same for all the subjects) in $\{20, 30, 40\}$; the approximate value $\tilde{\psi}_i = 40$ leads, at convergence, to the model with the highest marginal likelihood (3.12); the other initialization parameters are $c_i = 0.1\ \forall i$, $d = 0.5$, and $\Delta = 0.1$ and the algorithm converges in 18 iterations.

Table 3.1 shows fixed effect and variance parameter estimates for the fitted model. Fixed parameter estimates indicate that high values of FEV_1 immediately after the transplant ($\hat{\beta}_0 = 98.43$) are followed by an important drop ($\hat{\beta}_1 = -43.14$) occurring, on average, after about 40 months ($\hat{\psi} = 39.82$). However, the variance parameters emphasize considerable heterogeneity among subjects especially in time of occurrence of dropping (via $\hat{\sigma}^2_{\psi} = 18.65$) and relevant amount of dropping (via $\hat{\sigma}^2_1 = 16.64$).

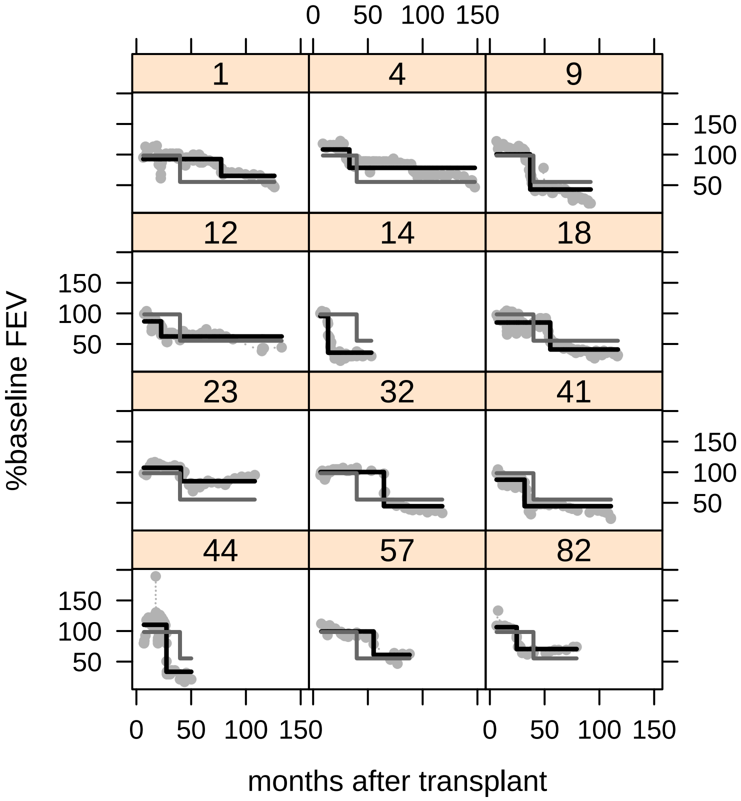

Figure 3.2: *Observed and fitted piecewise constant profiles for some patients under study. Grey lines: fixed effect estimates. Black lines: subject specific estimates. The quite different change-point locations and the different mean levels reflect a considerable heterogeneity among subjects.*

Table 3.1: *Lung transplant data: random effect model estimates.*

Parameter	Estimate
β_0	98.43
β_1	−43.14
ψ	39.82
σ_0^2	9.22
σ_1^2	16.64
σ_ψ^2	18.65

Heterogeneities are well appreciated in Figure 3.2 that illustrates observed trajectories and relevant fitted profiles for the subjects under study: the quite different change-point locations and the different mean levels reflect the high variance estimates reported in Table 3.1.

Chapter 4

Extension 2: *parameters on unbounded supports*

In Chapter 3 we considered a situation in which heterogeneity in the change-point among subjects is modelled through a random effect. However, heterogeneity in ψ could be expressed, more in general, in terms of dependence on a set of additional variables, that is, the change-point could be an unknown function of other covariates.
Parameters involved in the change-point function have not any longer a bounded and discrete support; therefore, conventional grid search techniques result to be difficult to implement for estimating such models. In this chapter we discuss how our proposal is able to work in this framework.

4.1 Model definition

For sake of simplicity, we consider a fixed effect model where the mean level of the response exhibits a shift at a change-point value that is a linear function of a single, additional covariate, namely

$$\mu_i = \beta_0 + \beta_1 I(x_i > \theta_0 + \theta_1 v_i). \tag{4.1}$$

If the point (v, x) lies below the straight line $\psi = \theta_0 + \theta_1 v$, the mean level of Y is β_0, while it shifts instantaneously to $\beta_0 + \beta_1$ above. Figure (4.1) shows an example on a toy dataset of $n = 100$ observations, assuming $v \sim \mathcal{N}(2, 3^2)$, $x \sim \mathcal{N}(5, 6^2)$, $\beta_0 = -0.4$, $\beta_1 = -1.2$, $\theta_0 = -0.5$, $\theta_1 = 2$ and $Y_i \sim \mathcal{N}(\mu_i, 0.5^2)$; estimating (4.1) by searching for $\hat{\theta}_1$ and $\hat{\theta}_2$ on a square uniform grid of 51×51 values (centered at the true pair) yields $\hat{\theta}_0 = -0.58$, $\hat{\theta}_2 = 2$, $\hat{\beta}_0 = -0.44$ and $\hat{\beta}_1 = -1.11$. Left panel displays the v-x scatterplot with point size proportional to y, right panels displays the v-x-y scatterplot. To stress the importance of correctly specifying the change-point as a linear function of covariate v, we set the same parameters of the previous example,

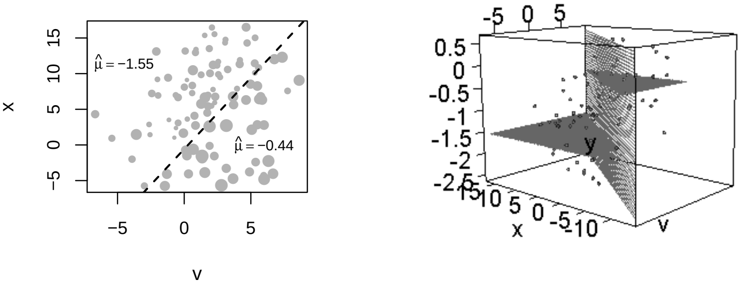

Figure 4.1: *Simulated dataset illustrating model (4.1). Left panel: 2d-scatterplot with point size proportional to y;* $\hat{\psi} = -0.58 + 2v$ *(straight line),* $\hat{\beta}_0 = -0.44$ *(below the straight line),* $\hat{\beta}_0 + \hat{\beta}_1 = -1.55$ *(above the straight line). Right panel: 3d-scatterplot and fitted values (surfaces).*

and generate Y with a negligible standard error, say $Y_i \sim \mathcal{N}(\mu_i, 10^{-6})$, so that model (4.1) holds exactly. Suppose now to ignore v and observe the resulting x-y scatterplot depicted in Figure 4.2.

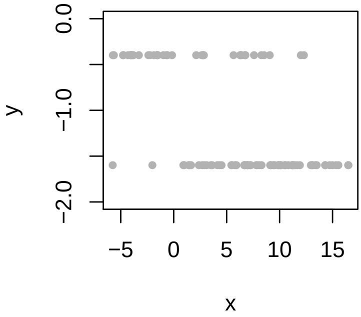

Figure 4.2: *2d scatterplot when ignoring the additional variable. Despite the model holds exactly, the marginal graphical inspection would wrongly exclude the presence of a change-point.*

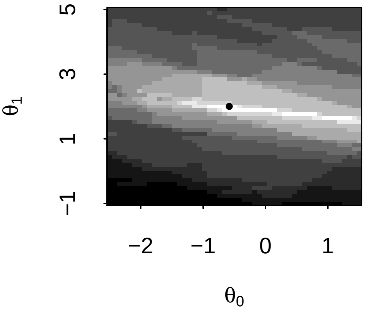

Figure 4.3: *Profile log-likelihood for θ_0 and θ_1 on a grid of values centered at the true pair. Light greys indicate higher values. The black point corresponds to the maximum likelihood estimate.*

Despite the model holds exactly, the marginal graphical inspection does not suggest the presence of a change-point, leading to wrong conclusions.
Figure 4.3 displays the joint profile log-likelihood for θ_0 and θ_1 for the dataset in Figure 4.1; light greys indicate higher values. Note that, despite the solution (black point) is unique in this case, it is difficult, in general, to perform an accurate search on a finite grid; in fact, possibly better solutions may not be included in the grid.
We discuss how to extend the iterative algorithm introduced in Chapter 2 to perform estimation of model (4.1).

4.2 Methods

Extending the idea of Chapter 2, we use the following key identity

$$I(x_i > \theta_0 + \theta_1 v_i) = \frac{1}{2}\frac{x_i - \theta_0 - \theta_1 v_i}{|x_i - \theta_0 - \theta_1 v_i|} + \frac{1}{2}, \tag{4.2}$$

for $x_i \neq \theta_0 + \theta_1 v_i$, which substituted in (4.1) gives

$$\begin{aligned}
\mu_i &= \beta_0 + \beta_1 \left(\frac{1}{2}\frac{x_i - \theta_0 - \theta_1 v_i}{|x_i - \theta_0 - \theta_1 v_i|} + \frac{1}{2} \right) \\
&= \beta_0 + \beta_1 \left(\frac{1}{2}\frac{x_i}{|x_i - \theta_0 - \theta_1 v_i|} + \frac{1}{2} \right) + \\
&+ (-\beta_1\theta_0) \left(\frac{1}{2}\frac{1}{|x_i - \theta_0 - \theta_1 v_i|} \right) + (-\beta_1\theta_1) \left(\frac{1}{2}\frac{v_i}{|x_i - \theta_0 - \theta_1 v_i|} \right) \\
&= \beta_0 + \beta_1 z_i + \gamma_0 w_{i0} + \gamma_1 w_{i1},
\end{aligned} \tag{4.3}$$

where

$$\gamma_0 = -\beta_1\theta_0 \quad \text{and} \quad \gamma_1 = -\beta_1\theta_1. \tag{4.4}$$

Note the auxiliary (or ‘working’) covariates are

$$z_i = \left(\frac{1}{2} \frac{x_i}{|x_i - \tilde{\theta}_0 - \tilde{\theta}_1 v_i|} + \frac{1}{2} \right),$$

$$w_{i0} = \left(\frac{1}{2} \frac{1}{|x_i - \tilde{\theta}_0 - \tilde{\theta}_1 v_i|} \right),$$

and

$$w_{i1} = \left(\frac{1}{2} \frac{v_i}{|x_i - \tilde{\theta}_0 - \tilde{\theta}_1 v_i|} \right), \tag{4.5}$$

with $\tilde{\theta}_0$ and $\tilde{\theta}_1$ meaning approximate values. Notice model (4.1) has been converted in the simple linear model (4.3).
Formulas above suggest the following simple iterative algorithm:

1. choose *starting values* $\tilde{\theta}_0$ and $\tilde{\theta}_1$;

2. compute the *working covariates* (4.5);

3. estimate the *working linear model* (4.3) and extract $\hat{\beta}_1$, $\hat{\gamma}_0$ and $\hat{\gamma}_1$;

4. *update* the parameter values via

$$\tilde{\theta}_0 = -\frac{\tilde{\gamma}_0}{\tilde{\beta}_1} \quad \text{and} \quad \tilde{\theta}_1 = -\frac{\tilde{\gamma}_1}{\tilde{\beta}_1}; \tag{4.6}$$

5. set $\tilde{\theta}_0 = \hat{\theta}_0$ and $\tilde{\theta}_1 = \hat{\theta}_1$ and *iterate* 2 to 4 until convergence.

As in the simpler case discussed in Chapter 2, the likelihood typically exhibits many local optima. Besides, denominators of the working covariates (4.5) go to zero when $x_i \approx \tilde{\theta}_0 + \tilde{\theta}_1 v_i$, namely when points (v_i, x_i) are close to the approximate straight line; therefore, we need some adjustment to favour convergence of the algorithm.

4.2.1 Rescaling v and x and values

Extending the idea of Section 2.2.1 presents an additional issue: moving points (v_i, x_i) away from the straight line $x = \tilde{\theta}_0 + \tilde{\theta}_1 v$. Unless $\tilde{\theta}_1 = 0$, simple rescaling of x is clearly not feasible, because the change-point is not unique. In some way, v_i and x_i have to be rescaled jointly away from the approximate straight line. We illustrate our proposal with an example.

Consider the toy dataset depicted in Figure 4.1, and a starting straight line given by $x_i = 8 + 2v_i$, as depicted in Figure 4.4.

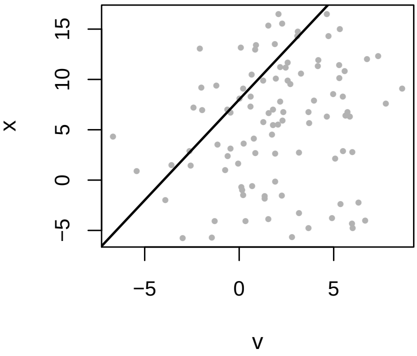

Figure 4.4: *Example of starting straight line in a toy dataset. The starting values for the parameters are* $\tilde{\theta}_0 = 8$ *and* $\tilde{\theta}_1 = 2$.

The idea is to operate a convenient linear transformation on the covariates in order to reduce to a standard situation in which $\tilde{\theta}_1 = 0$. Therefore, we propose to rotate the points (v_i, x_i) according to the angle defined by the approximate straight line, namely

$$\tilde{\rho} = \arctan(\tilde{\theta}_1). \tag{4.7}$$

To rotate each point we use the following transformation:

$$\begin{pmatrix} v_i^* \\ x_i^* \end{pmatrix} = \tilde{\Lambda} \begin{pmatrix} v_i \\ x_i \end{pmatrix}, \tag{4.8}$$

where $\tilde{\Lambda}$ is a standard rotation matrix (Arfken and Weber, 2011):

$$\tilde{\Lambda} = \begin{pmatrix} \tilde{\lambda}_{11} & \tilde{\lambda}_{12} \\ \tilde{\lambda}_{21} & \tilde{\lambda}_{22} \end{pmatrix} = \begin{pmatrix} \cos(\tilde{\rho}) & \sin(\tilde{\rho}) \\ -\sin(\tilde{\rho}) & \cos(\tilde{\rho}) \end{pmatrix}. \tag{4.9}$$

Of course, the approximate straight line $x = \tilde{\theta}_0 + \tilde{\theta}_1 v$ will rotate accordingly, and we denote the rotated straight line with $x^* = \tilde{\theta}_0^*$ (note that $\tilde{\theta}_1^* = 0$ by construction). Rewriting x^* according to (4.8) gives

$$\begin{aligned} x^* &= \tilde{\theta}_0^* \\ \tilde{\lambda}_{21} v + \tilde{\lambda}_{22} x &= \tilde{\theta}_0^* \\ \tilde{\lambda}_{22} x &= \tilde{\theta}_0^* - \tilde{\lambda}_{21} v \end{aligned}$$

and

$$x = \frac{\tilde{\theta}_0^*}{\tilde{\lambda}_{22}} - \frac{\tilde{\lambda}_{21}}{\tilde{\lambda}_{22}} v,$$

so that

$$\tilde{\theta}_0 = \frac{\tilde{\theta}_0^*}{\tilde{\lambda}_{22}} \quad \text{and} \quad \tilde{\theta}_1 = -\frac{\tilde{\lambda}_{21}}{\tilde{\lambda}_{22}}, \tag{4.10}$$

and, conversely,

$$\tilde{\theta}_0^* = \tilde{\lambda}_{22} \tilde{\theta}_0 \quad \text{and} \quad \tilde{\theta}_1^* = 0. \tag{4.11}$$

Rotated points and the relevant straight line are represented in Figure 4.5.

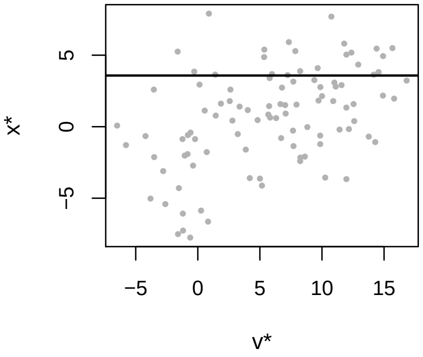

Figure 4.5: *Rotated points and relevant straight line. To rescale the points away from the approximate straight line we reproduce a situation in which the change-point is no longer a linear function of v.*

Note that the slope of the rotated line $\tilde{\theta}_1^*$ is 0 by construction, and this makes easy to operate the rescaling as in Section 2.2.1. As usual, we use a rescaling factor $c \in (0, 1)$ and compute a lower, $\tilde{\theta}_0^{*-}$, and a upper, $\tilde{\theta}_0^{*+}$, 'threshold' value:

$$\tilde{\theta}_0^{*-} = \tilde{\theta}_0^* - c(\tilde{\theta}_0^* - x_{(1)}^*), \qquad \tilde{\theta}_0^{*+} = \tilde{\theta}_0^* + c(x_{(n)}^* - \tilde{\theta}_0^*).$$

We therefore consider

$$x_i' = x_{(1)}^* + (x_i^* - x_{(1)}^*)(1 - c) \tag{4.12}$$

for $x_i^* \in [x_{(1)}^*, \tilde{\theta}_0^*]$, and

$$x_i' = \tilde{\theta}_0^{*+} + (x_i^* - \tilde{\theta}_0^*)(1 - c) \tag{4.13}$$

for $x_i^* \in (\tilde{\theta}_0^*, x_{(n)}^*]$.

The left panel in Figure 4.6 represents the rotated points without rescaling

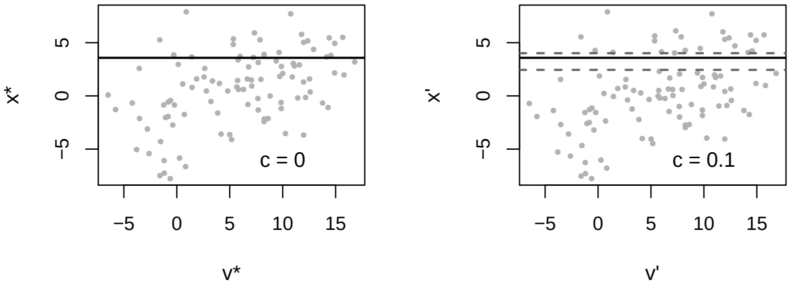

Figure 4.6: *Only rotated (left panel) vs rotated and rescaled data (right panel). The rescaling induced a point-free region in the neighbourhood of the rotated approximate straight line.*

x^*, namely when $c = 0$, while the right panel shows the effect of a rescaling factor $c = 0.1$.

Finally, we use x', $v' = v^*$ and the rotated straight line to compute auxiliary covariates (4.5) and fit the working linear model (4.3). We stress that, despite the multiple transformations induced to the covariates, estimates of the mean levels β_0 and β_1 are substantially unaffected, while the straight line parameter estimates, of course, are. In particular, the rotation only induces a reparametrization according to (4.10) and (4.11), while the rescaling should favour to skip some spurious optima. How to reduce c throughout iterations represents an additional issue.

4.3 The iterative algorithm in detail

To illustrate the algorithm, Figure 4.7 refers to the initialization and the first 8 iterations for the dataset depicted in Figure 4.1.

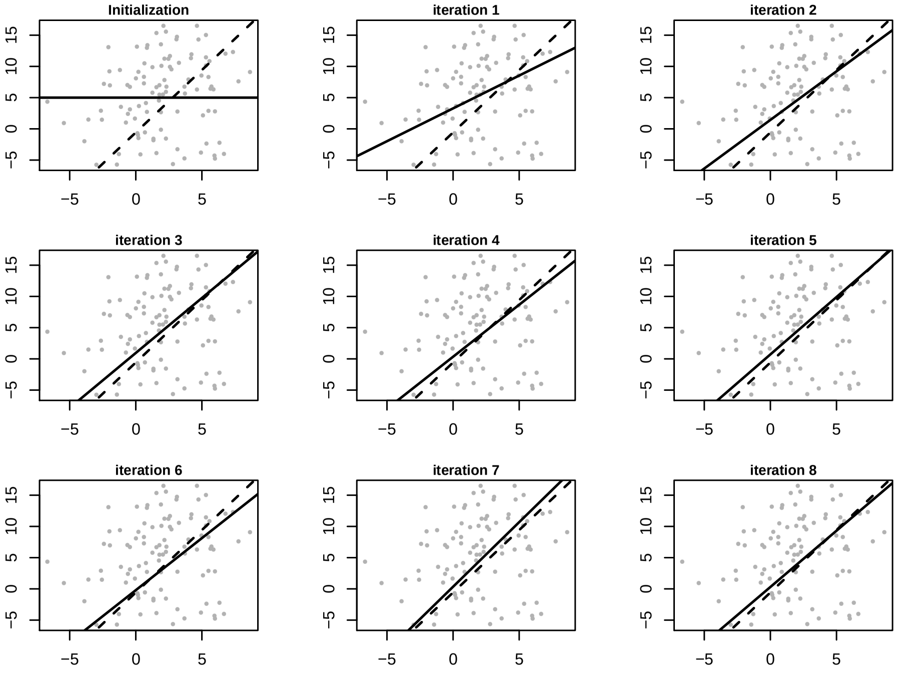

Figure 4.7: *Algorithm initialization (upper left panel) and first 8 iterations. Starting values:* $\tilde{\psi} = 5$ *(the sample mean of x) and* $c = 0.1$. *Dashed lines: final solution. Solid lines: updated solutions. The updated estimates approach the solution after about 3 iterations and stabilize thereafter.*

The dashed lines refer to the final solution $\hat{\psi} = -0.58 + 2v$, while the solid lines indicate the updated solutions $\tilde{\psi} = \tilde{\theta}_0 + \tilde{\theta}_1 v$. Initialization parameters are $\tilde{\theta}_0 = 5$, $\tilde{\theta}_1 = 0$ and $c = 0.1$. Note that the updated estimates approach the solution already after 3 iterations and stabilize thereafter.

Unlike in the single change-point case, monitoring the sign of the variation $\hat{\psi} - \tilde{\psi}$ is no longer viable to figure out when the algorithm approaches the solution, and monitoring the log-likelihood appears to be the only feasible solution. The likelihood value is most fluctuating in the first iterations, while it tends to stabilize when the algorithm approaches the solution (Figure 4.8). In the spirit of *step halving* (Jennrich and Sampson, 1968), we propose to multiply c by a reduction factor, say $d \in (0, 1)$, any time the sign of the variation $\ell(\hat{\psi}) - \ell(\tilde{\psi})$ changes. A rule of thumb suggests to run some preliminary iterations, say p, before starting monitoring the likelihood. Figure 4.8 shows the values of the maximized log-likelihood for the $p = 8$ preliminary iterations depicted in Figure 4.7 and next iterations, using $d = 0.5$; we have referred to the working model to evaluate the variation $\ell(\hat{\psi}) - \ell(\tilde{\psi})$ and reduced c consequently.

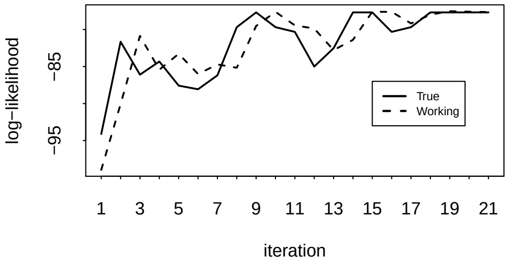

Figure 4.8: *True (solid line) and working (dashed line) log-likelihood throughout iterations. The likelihood tends to stabilize when the algorithm approaches the solution (iteration 3).*

To summarize, the steps of the algorithm are reported below:

1. choose *starting values* $\tilde{\theta}_0$ and $\tilde{\theta}_1$;
2. choose a *rescaling factor* c;
3. *rotate* points (v_i, x_i) according to (4.8) to obtain v^*, x^* and $\tilde{\theta}_0^*$ ($\tilde{\theta}_1^* = 0$);
4. *rescale* x^* using (4.12) and (4.13) to obtain x', while $v' = v^*$;
5. compute the *working covariates* (4.5) using x' and v';
6. estimate the *working linear model* (4.3) and extract $\hat{\beta}_1^*$, $\hat{\gamma}_0^*$ and $\hat{\gamma}_1^*$;
7. use (4.6) to *update* $\tilde{\theta}_0^*$ and $\tilde{\theta}_1^*$ to $\hat{\theta}_0^*$ and $\hat{\theta}_1^*$;
8. use (4.10) to *switch back* to $\hat{\theta}_0$ and $\hat{\theta}_1$;
9. *decrease* c using d if the working likelihood changes direction;
10. set $\tilde{\theta}_0 = \hat{\theta}_0$ and $\tilde{\theta}_1 = \hat{\theta}_1$ and *iterate* 3 to 9 until convergence.

4.4 Starting values

Choosing the starting values by visual inspection may result more difficult in this case. A possible strategy it to fix the starting straight line at $x = \bar{x}$, where $\bar{x}$ is the sample mean of x; in Figure 4.7, we choose $x = 5$, the simulated mean of x.

4.5 Convergence

Due to the strong irregularities of the true likelihood, assessing convergence and correctly stopping the algorithm does not represent a simple task. Despite the working likelihood stabilizes when the algorithm approaches the solution, the relevant absolute variation $|\ell(\hat{\psi}) - \ell(\tilde{\psi})|$ still does not represent a reliable indication of convergence, even after having repeatedly decreased the rescaling factor. We therefore propose to evaluate the joint variation of $\tilde{\theta}_0$ and $\tilde{\theta}_1$. In particular, for some small $\Delta > 0$, we use

$$(\hat{\theta}_0 - \tilde{\theta}_0)^2 + (\hat{\theta}_1 - \tilde{\theta}_1)^2 < \Delta$$

as a convergence rule. Some simulation studies carried out in Section 5.3 show that the proposed approach works correctly in simple scenarios.

4.6 Example

To illustrate the proposed algorithm we apply it to the 'airquality' dataset shipped with the *R* environment. The dataset consists of 154 daily observations concerning some air quality values in New York from May 1, 1973 to September 30, 1973. Tropospheric ozone is an atmospheric pollutant, and its concentration represents a common variable of interest in environmental science. The ozone levels may depend on many factors, among which some atmospheric agents, such as temperature and wind. Therefore, in this simple example we analyze the relationship between ozone (Y, parts per billion), temperature (x, degrees Fahrenheit) and wind (v, average speed in miles per hour). Figure 4.9 displays the wind-temperature scatterplot with point size proportional to the ozone level for the $n = 116$ complete records. The concentration of the pollutant seems to increase abruptly in the top-left region of the plot; in particular, a threshold temperature value appears to be approximately 80 degrees. Estimating model (4.1), assuming the Y_is to be independent and Gaussian, could yield additional information.

To initialize the proposed algorithm we choose $\tilde{\psi} = 80$ ($\tilde{\theta}_1 = 0$) as starting

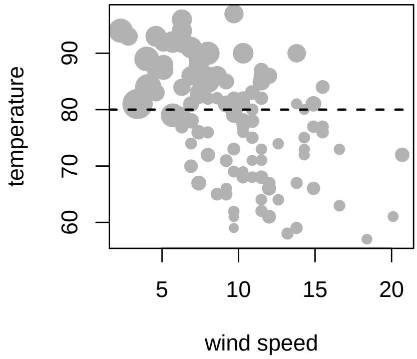

Figure 4.9: *Airquality dataset: wind-temperature scatterplot with point size proportional to the ozone level. Based on a visual inspection, the concentration of the pollutant appears to increase abruptly at about* $\tilde{\psi} = 80$ *degrees of temperature (dashed line).*

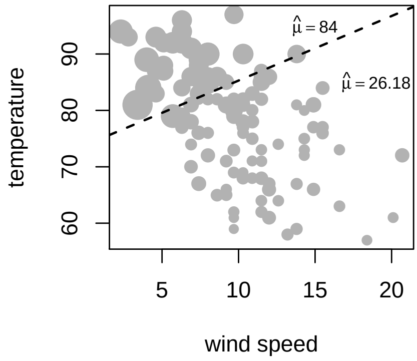

Figure 4.10: *Airquality dataset: wind-temperature scatterplot with point size proportional to the ozone level. The estimated linear change-point is* $\hat{\psi} = 72.83 + 1.24v$ *(dashed line). The mean ozone level appears to increase abruptly as temperature goes beyond a critical value. As the wind speed increases, the critical value increase as well.*

guess; other initialization parameters are $c = 0.03$, $d = 0.5$, and $\Delta = 10^{-6}$. We perform $p = 10$ preliminary iterations before starting evaluating the working likelihood, and get to convergence in 28 iterations. The algorithms yields $\hat{\beta}_0 = 26.18, \hat{\beta}_1 = 57.82$ and

$$\hat{\psi}_i = 72.83 + 1.24 v_i.$$

Figure 4.10 displays the estimated straight line on the 2d-scatterplot. The mean ozone level appears to increase abruptly from about 26 to 84 p.p.b as temperature goes beyond some ‘critical’ value. In the absence of wind, the ‘critical’ value is about 73 degrees. Wind seems to have a ‘positive’ effect in limiting the ozone levels: in fact, as the wind increases, temperature has to increase further to cause a ‘jump’ in the ozone levels. The BIC provides evidence supporting a threshold line rather than a constant change-point: 1046.35 *vs* 1077.54.

Chapter 5

Simulations

In this chapter, some simulation studies are performed to assess the empirical performances of the proposed procedures. Generally, finite sample properties of a change-point estimator depend on the sample size n, the true change-point location ψ, and the 'signal-to-noise ratio' $\frac{\beta_1}{\sigma_\epsilon}$, where σ_ϵ is the residual standard error. Additional issues are related to the starting value $\tilde{\psi}$ and the rescaling factor c.

5.1 Single-Shift Models

In this section we perform two small simulation studies. The first one is aimed at assessing the general behaviour of the estimator. The second one illustrates the effect of the starting point and the rescaling factor selection. Note that we will evaluate the estimator performances as 'good' (in terms of biasedness) when most of the estimates $\hat{\psi}$ fall in the interval $[x_{j^*}, x_{j^*+1})$ such that

$$\psi \in [x_{j^*}, x_{j^*+1}); \tag{5.1}$$

we will refer to (5.1) as the 'true' interval. For sake of simplicity, we simulate reasonable scenarios in which $n_d = n$ and the true value ψ is included among observations.

5.1.1 Changing model parameters

To evaluate the finite sample properties of the proposed estimator we vary n in $\{51, 201\}$, and, accordingly, consider explanatory variables x given by sequences of n evenly spaced values between 0 and 100. We assume model (2.2), where $\beta_0 = 2$, $\beta_1 = 1.5$, $\psi \in \{50, 74\}$, $Y_i \sim \mathcal{N}(\mu_i, \sigma^2)$, and $\sigma \in \{0.5, 1\}$. For each scenario we generate 1000 datasets and perform estimation via the proposed algorithm. To minimize the effect produced by the initialization parameters, the starting value is setted at the true change-point, and the same rescaling factor $c = 0.03$ ($d = 0.5$) is used for each simulated dataset, while the tolerance value for the stopping criterion is $\Delta = 0.01$. Table 5.1 shows the empirical means and standard deviations of the estimates.

Table 5.1: *Sampling distribution of $\hat{\psi}$: means and standard deviations (in brackets) for different σ, n, and ψ. Starting value: true change-point ψ. The estimator is unbiased, since the means fall in the middle of the true intervals. The standard deviation decreases as n increases and σ decreases.*

		ψ	
σ	n	50	74
0.5	51	50.95 (1.23)	74.87 (1.36)
	201	50.25 (0.34)	74.27 (0.36)
1	51	50.95 (3.08)	74.83 (3.42)
	201	50.25 (1.12)	74.48 (1.20)

The estimator appears to be unbiased, since the empirical means tend to falls in the middle of the true intervals, especially when σ is lower: when $n = 51$, the true intervals are $[50, 52)$ and $[74, 76)$, while when $n = 201$ they are $[50, 50.5)$ and $[74, 74.5)$. Besides, as the sample size increases the standard deviations decrease substantially. Standard deviations also decrease as the residual standard error decreases.

5.1.2 Changing initialization parameters

As illustrated in Figure 2.5, the perturbation induced by c on the explanatory variable x affects the step $|\hat{\psi} - \tilde{\psi}|$. In particular, the rescaling helps, to some extent, to skip some spurious maxima, especially when the starting value $\tilde{\psi}$ is far from the final solution $\hat{\psi}$. We here perform a simple simulation study to illustrate this issue.

We vary the true change-point ψ in $\{30, 50, 70\}$ and generate 1000 datasets from model (2.2). We set $n = 100$, $x_i = i$, $i = 1, \ldots, n, \beta_0 = 2, \beta_1 = 1$, and $Y_i \sim \mathcal{N}(\mu_i, 0.4^2)$. We vary the starting point $\tilde{\psi}$ in $\{30, 50, 70\}$, and also evaluate a fourth option given by the best candidate among $\{17, 34, 50, 67, 83\}$. We also vary the rescaling factor c in $\{0.03, 0.05\}$ and use $\Delta = 0.1$.

Table 5.2 shows the empirical means and standard deviations of the estimates. As expected, the performance is good when $\hat{\psi} = \psi$, which would represent the 'gold standard' choice, while it gets worse when the starting point is far from the true value, because the algorithm is more likely to stop at some local solution. Figure 5.1 represents the boxplots of $\hat{\psi}$ when $\tilde{\psi} = 70$ and $\psi = 30$; spurious clusters in the starting value proximity introduce a bias and also affect the variance of the estimator.

Selecting $\tilde{\psi}$ in the middle of the range of x, or among a grid of some candidates, gives results which are comparable with the 'gold standard'.

Table 5.2: *Effect of the starting value $\tilde{\psi}$ and the rescaling factor c on the sampling distribution of $\hat{\psi}$: means and standard deviations (in brackets). The bias and the standard deviation increase when $\tilde{\psi}$ is far from ψ, but increasing the rescaling factor c improves performances.*

		$\tilde{\psi}$			
ψ	c	30	50	70	5 value grid
30	0.03	30.33 (0.93)	30.50 (1.31)	31.47 (5.85)	30.42 (1.22)
	0.05	30.32 (1.07)	30.40 (1.28)	30.54 (2.83)	30.37 (1.21)
50	0.03	50.21 (2.00)	50.41 (0.89)	50.64 (1.32)	50.41 (0.89)
	0.05	50.44 (1.10)	50.47 (0.94)	50.54 (0.99)	50.47 (0.94)
70	0.03	69.01 (7.60)	70.53 (1.37)	70.54 (0.95)	70.58 (1.00)
	0.05	70.23 (4.38)	70.68 (1.09)	70.63 (1.12)	70.67 (1.10)

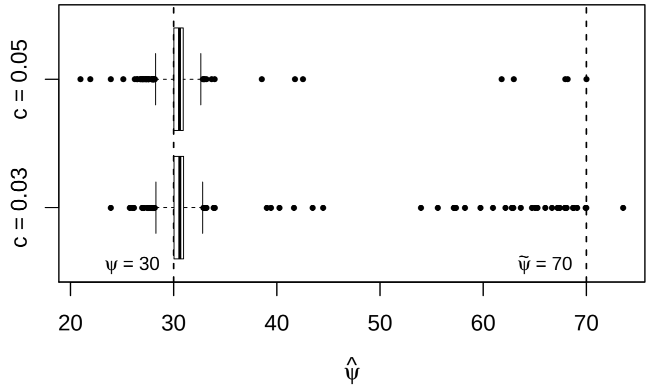

Figure 5.1: *Sampling distributions of $\hat{\psi}$ coming from two values of c. Spurious clusters in the starting value proximity ($\tilde{\psi} = 70$) introduce a bias, but they tend to disappear when c increases (boxplot at the top).*

While increasing the rescaling factor does not affect substantially the performance when selecting reasonable starting values, it improves the performance when $\tilde{\psi}$ is far from the solution: in fact, the empirical mean of $\hat{\psi}$ stabilizes in the true interval $[\psi, \psi + 1)$, with a lower variance as well. Note that, in Figure 5.1, spurious clusters tend to disappear when $c = 0.05$.

We stress that increasing c beyond a certain limit may make $\tilde{\psi}$ fall outside the range of x at some iteration, causing an algorithm failure. In our simulation study, when $c = 0.05$, 4 failures occurred when $\tilde{\psi} = 70$ and $\psi = 30$, another 3 occurred when $\tilde{\psi} = 30$ and $\psi = 70$.

For completeness, Table 5.3 shows the empirical summary measures for the estimators $\hat{\beta}_0$ and $\hat{\beta}_1$, for the same scenarios considered in Figure 5.1 ($\tilde{\psi} = 70$, and $\psi = 30$). The two estimators are unbiased for both values of the rescaling factor; besides, like for the change-point estimator, the standard deviations decrease as the rescaling factor increases.

Table 5.3: *Regression coefficients: empirical summary measures.* $\hat{\beta}_0$ *and* $\hat{\beta}_1$ *are unbiased independently from c. As for the change-point estimator, the standard deviations decrease as c increases.*

Estimator	True	c	Mean	S.D.
$\hat{\beta}_0$	2.00	0.03	2.01	0.11
		0.05	2.00	0.08
$\hat{\beta}_1$	1.00	0.03	0.99	0.12
		0.05	1.00	0.09

5.2 Extension 1: *random effects*

The algorithm is more complex when estimating random effect models. In fact, it performs optimization over a high number of parameters, because also random effect predictions are involved in computations at each step. Besides, the final solution depends strongly on the choice of the initialization parameters, including the rescaling factors. We perform a small simulation study to assess the general behaviour of the estimator.

We set a small residual variance $\sigma_\epsilon^2 = 0.4^2$, locate ψ in the middle of the range of x and choose a moderate σ_ψ^2, so that is difficult for the random change-points to fall outside this range; in fact, at this stage we assume that ψ_i exists for each subject. Setting a high value of β_1 and a small σ_1^2 guarantees not to have profiles with non-identifiable change-point; dealing with possible subjects without change-point represents an additional issue, which is not addressed in this thesis.

We therefore generate 1000 datasets from model (3.1), assuming $n = 20$, $n_i = 20$, $x_{ij} = 1, 2, \ldots, n_i$, and

$$\begin{bmatrix} \beta_{0i} \\ \beta_{1i} \\ \psi_i \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 2 \\ 4 \\ 10 \end{bmatrix}, \begin{bmatrix} 0.7^2 & 0 & 0 \\ 0 & 0.8^2 & 0 \\ 0 & 0 & 2^2 \end{bmatrix} \right), \tag{5.2}$$

and perform estimation via the algorithm discussed in Chapter 3.

We consider three different scenarios. In the first scenario we use, for each simulated dataset, the true random change-point ψ_i as approximate values, and a rescaling factor $c = 0.03$; since we expect to start in the neighbourhood of the solution, this choice would represent the ‘gold standard’. In the second scenario we generate the starting values $\tilde{\psi}_i$ uniformly in the interval $[7, 13]$, located in the middle of the common ranges of x, and use a rescaling factor $c = 0.05$. Finally, in the third scenario we set $\tilde{\psi}_i = 8$ for all the subjects, and $c = 0.05$. Other initialization parameters are $d = 0.8$ and $\Delta = 0.1$, for all the scenarios.

Table 5.4: *Random effect model: means and standard deviations (in brackets) of the estimates. Starting values are fixed at the true occurrences* ψ_i*, selected at random in the interval* $[7, 13]$ *or fixed at* 8*. Estimators are substantially unbiased.* $\hat{\psi}$ *and* $\hat{\sigma}_1^2$ *have the highest variances.*

		$\tilde{\psi}_i$		
Estimator	True	ψ_i	$U(7, 13)$	8
$\hat{\beta}_0$	2.00	1.987 (0.156)	1.999 (0.154)	1.991 (0.158)
$\hat{\beta}_1$	4.00	4.004 (0.177)	3.990 (0.185)	3.971 (0.192)
$\hat{\psi}$	10.00	9.948 (0.465)	9.863 (0.446)	9.830 (0.565)
$\hat{\sigma}_0^2$	0.49	0.483 (0.165)	0.488 (0.164)	0.489 (0.171)
$\hat{\sigma}_1^2$	0.64	0.638 (0.223)	0.665 (0.220)	0.658 (0.245)
$\hat{\sigma}_\psi^2$	4.00	4.050 (1.374)	4.032 (1.346)	4.033 (1.503)

Table 5.4 shows the results. Estimators $\hat{\psi}$ and $\hat{\sigma}_1^2$ appear to have a little bias, especially when the starting values are selected at random or fixed at 8 for all the subjects; these two estimators also have the highest empirical variances. Other estimators appear to be unbiased independently from the starting value selection. The empirical standard deviations do not differ substantially for the three scenarios, with the exception of estimators $\hat{\psi}$ and $\hat{\sigma}_1^2$ which have an higher variance when the algorithm is initializated at $\tilde{\psi}_i = 8$. Even if a deeper investigation would be needed, these results show that the proposed approach works reasonably well in simple scenarios.

5.3 Extension 2: *parameters on unbounded supports*

Finally we perform a small simulation study to evaluate the proposed algorithm when estimating the linear-change-point model (4.1).

We set a moderate sample size $n = 200$ and locate the true linear change-point in an intermediate region of the v-x space. In particular, we generate the variables $v \sim \mathcal{N}(22, 3^2)$ and $x \sim \mathcal{N}(25, 6^2)$, and set $\psi = 5 + v$. The true regression coefficients are $\beta_0 = -0.4$ and $\beta_1 = -1.2$. We therefore generate 1000 datasets from model (4.1), where $Y_i \sim \mathcal{N}(\mu_i, 1)$, and perform estimation via the proposed algorithm.

We consider two different scenarios. In the first scenario we use the true intercept ($\tilde{\theta}_0 = 5$) and the true slope ($\tilde{\theta}_1 = 1$) of the linear change-point as starting values for the iterative algorithm. In the second scenario we set the approximate straight line at the mean value of x, namely $\tilde{\theta}_0 = 25$ and $\tilde{\theta}_1 = 0$. Other initialization parameters are $c = 0.03$, $d = 0.5$ and $\Delta = 10^{-6}$, for all the scenarios. Table 5.5 shows the results.

Table 5.5: *Linear change-point model: means and standard deviations (in brackets) of the estimates. Starting values are fixed at the true linear function and the mean value of x. For the considered scenarios, the intercept estimator appears to be biased with the highest variance.*

		$\tilde{\psi}$	
Estimator	True	ψ	25
$\hat{\beta}_0$	-0.400	-0.393 (0.095)	-0.392 (0.095)
$\hat{\beta}_1$	-1.200	-1.221 (0.138)	-1.217 (0.139)
$\hat{\theta}_0$	5.000	5.223 (3.550)	5.679 (4.178)
$\hat{\theta}_1$	1.000	0.988 (0.160)	0.963 (0.194)

The intercept estimator appears to be biased, with a high variance. Since the regression coefficient estimates are unbiased, we conjecture such bias to be related to the particular data conformation and change-point location.

Figure 5.2 portrays 50 estimated straight lines when the algorithm starts from the true values (left panel) and the mean value 25 (right panel). The estimated linear functions correctly approximate the true one (black line); note that, when using $\tilde{\psi} = 25$ as approximate value (dashed line), the estimated linear functions are more likely to remain close to the starting value.

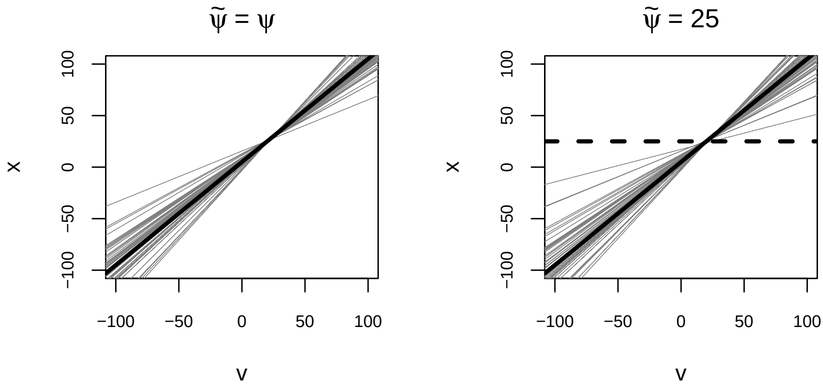

Figure 5.2: *Linear change-point model: estimated straight lines (grey lines) when starting from the true change-point (left panel) and the mean value of x (right panel). The estimated linear functions correctly approximate the true one (black line). When $\tilde{\psi} = 25$ (dashed line), the estimated linear functions could remain close to the starting value.*

Chapter 6

Multiple change-point models

Multiple change-points my arise from two different extensions; in the most common framework, the expected value of the response is assumed to be expressed by a stepwise function with K shifts induced by a single covariate. A typical example concerns *aCGH* analyses (Pinkel *et al.*, 1998), where interest lies in detecting possible aberrations along a chromosome. Alternatively, the change-points can be relevant to more than one covariate, and this issue has not been well addressed in the literature.

When K change-points have to be estimated, efficient grid search algorithms based on dynamic programming are available if the K change-points are relevant to the same covariate (Bai and Perron, 2003). However, dynamic approaches turn out to be unfeasible when the mean shifts are induced by several covariates, because the change-points have different supports. Moreover, large sample sizes still represent a concern also for dynamic approaches.

In this chapter we discuss the possibility to extent the proposed algorithm for the multiple change-point case. Even if preliminary simulations have shown that the proposed algorithms works reasonably well in simple scenarios, we do not study the topic in detail.

6.1 Change-point in several covariates

The issue of estimating change-points in more than one covariate has not been well addressed in the literature. In fact, all the techniques in Chapter 1 are built for detecting one or more change-points in a single covariate. Application of the most efficient techniques based on dynamic grid search (Bai and Perron, 2003) cannot be applied when the change-points are defined different supports, and the computational cost becomes huge as the number of change-points increases.

In this section we propose an extension of the proposed algorithm for estimating such models efficiently.

6.1.1 Model definition

Let's consider a model with K explanatory variables x_k, $k = 1, 2, \ldots, K$, each one with its own change-point ψ_k at which the mean level of Y exhibits a shift given by β_k. The regression function can be written

$$\mu_i = \beta_0 + \sum_{k=1}^{K} \beta_k I(x_{ik} > \psi_k), \tag{6.1}$$

which specifies that the mean level of Y assumes constant values in a set of multidimensional regions of the explanatory variable domain (Figure 6.1). To simplify the notation, we assume that all the explanatory variables consists of n distinct values.

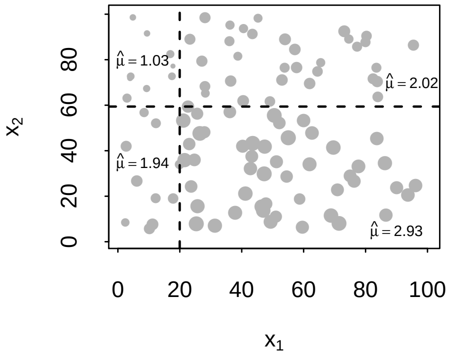

Figure 6.1: *Simulated dataset illustrating the model: 2d-scatterplot with point size proportional to y. The estimated mean level of Y change according to the regions defined by* $\hat{\psi}_1$ = 19.96 *and* $\hat{\psi}_2$ = 59.40 *(dashed lines).*

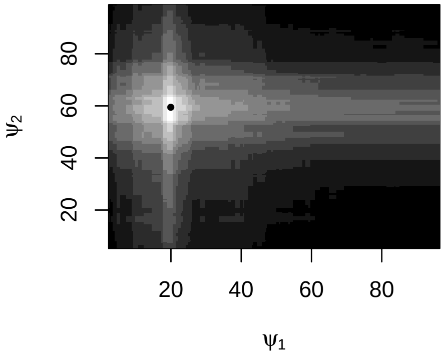

Figure 6.2: *Profile log-likelihood for* ψ_1 *and* ψ_2*. Lighter area indicates higher log-likelihood. The black point corresponds to the maximum likelihood estimate.*

Similarly to the case of the single change-point model discussed in Chapter 2, all the values of ψ_k between x_{ik} and $x_{\{i+1\}k}$, for each k, lead to the same likelihood.

We consider an example on a toy dataset of $n = 100$ observations, where x_1 and x_2 are generated at random from a $U(1, 100)$, and the relevant change-points are $\psi_1 = 20$ and $\psi_2 = 60$. The regression coefficients are $\beta_0 = 2$, $\beta_1 = 1$, and $\beta_2 = -1$, while $Y_i \sim \mathcal{N}(\mu_i, 0.3^2)$; estimating (6.1) by searching for $\hat{\psi}_1$ and $\hat{\psi}_2$ among the n^2 pairs $(x_{i1}, x_{i'2})$ yields $\hat{\psi}_1 = 19.96$, $\hat{\psi}_2 = 59.40$, $\hat{\beta}_0 = 1.94$, $\hat{\beta}_1 = 0.99$ and $\hat{\beta}_2 = -0.91$.

Figure 6.1 displays the scatterplot of points (x_{i1}, x_{i2}) with point size proportional to y_i; in this model, the fitted values change according to the regions defined by $\hat{\psi}_1$ and $\hat{\psi}_2$ (dashed lines). Figure 6.2 displays the joint profile log-likelihood for ψ_1 and ψ_2; light greys indicate higher values.

6.1.2 Methods

To extend our iterative algorithm we use the key identity (2.3) for each covariate. Model (6.1) becomes

$$\begin{aligned}
\mu_i &= \beta_0 + \sum_{k=1}^{K} \beta_k \left(\frac{1}{2} \frac{x_{ik} - \psi_k}{|x_{ik} - \psi_k|} + \frac{1}{2} \right) \\
&= \beta_0 + \sum_{k=1}^{K} \beta_k \left(\frac{1}{2} \frac{x_{ik}}{|x_{ik} - \psi_k|} + \frac{1}{2} \right) + \sum_{k=1}^{K} (-\beta_k \psi_k) \left(\frac{1}{2} \frac{1}{|x_{ik} - \psi_k|} \right) \\
&= \beta_0 + \sum_{k=1}^{K} \beta_k z_{ik} + \sum_{k=1}^{K} \gamma_k w_{ik}, \qquad (6.2)
\end{aligned}$$

where

$$\gamma_k = -\beta_k \psi_k \qquad (6.3)$$

and the auxiliary (or 'working') covariates are

$$z_{ik} = \left(\frac{1}{2} + \frac{1}{2}\frac{x_{ik}}{|x_{ik} - \tilde{\psi}_k|}\right) \quad \text{and} \quad w_{ik} = \left(\frac{1}{2}\frac{1}{|x_{ik} - \tilde{\psi}_k|}\right), \tag{6.4}$$

with $\tilde{\psi}_k$, $k = 1, 2, \dots, K$, meaning approximate values.

Model (6.1) has been converted in the simple linear model (6.2), and formulas above suggest the following simple iterative algorithm:

1. choose a vector of (possibly equal) *starting values* $\tilde{\psi}_1, \tilde{\psi}_2, \dots, \tilde{\psi}_K$;

2. compute the *working covariates* (6.4);

3. estimate the *working LM* (6.2) and extract $\hat{\beta}_k$ and $\hat{\gamma}_k$ for each k;

4. *update* the change-point values via

$$\hat{\psi}_k = -\frac{\hat{\gamma}_k}{\hat{\beta}_k}; \tag{6.5}$$

5. set $\tilde{\psi}_k = \hat{\psi}_k$ and *iterate* 2 to 4 until convergence.

Rescaling x_k values

As widely discussed in Section 2.2.1, the values of each covariate should be moved away from the relevant approximate values $\tilde{\psi}_k$. We therefore choose a vector of rescaling factors $c_1, c_2, \dots, c_K$ and compute

$$\tilde{\psi}_k^- = \tilde{\psi}_k - c_k(\tilde{\psi}_k - x_{(1)k}), \qquad \tilde{\psi}_k^+ = \tilde{\psi}_k + c_k(x_{(n)k} - \tilde{\psi}_k),$$

to obtain the scaled values

$$x'_{ik} = x_{(1)k} + (x_{ik} - x_{(1)k})(1 - c_k) \tag{6.6}$$

for $x_{ik} \in [x_{(1)k}, \tilde{\psi}_k]$, and

$$x'_{ik} = \tilde{\psi}_k^+ + (x_{ik} - \tilde{\psi}_k)(1 - c_k) \tag{6.7}$$

for $x_{ik} \in (\tilde{\psi}_k, x_{(n)k}]$. The rescaled variables x'_k are used to compute auxiliary covariates (6.4) and to fit the working linear model (6.2). As usual, decreasing the c_ks according to some reduction factor, say $d \in (0, 1)$, turns out to be useful to avoid convergence failures.

In summary, these are the steps of the algorithm:

1. choose a vector of (possibly equal) *starting values* $\tilde{\psi}_1, \tilde{\psi}_2, \ldots, \tilde{\psi}_K$;
2. choose a vector of (possibly equal) *rescaling factors* $c_1, c_2, \ldots, c_K$;
3. *rescale* the x_ks using (6.6) and (6.7) to obtain the x'_ks;
4. compute the *working covariates* (6.4) using the x'_ks;
5. estimate the *working LM* (6.2) and extract $\hat{\beta}_k$ and $\hat{\gamma}_k$ for each k;
6. use (6.5) to *update* the change-point values;
7. *decrease* c_k, $k = 1, 2, \ldots, K$, using d if $\tilde{\psi}_k$ changes direction;
8. set $\tilde{\psi}_k = \hat{\psi}_k$ and *iterate* 3 to 7 until convergence.

6.1.3 Starting values

To estimate model (6.1) a vector of approximate change-point values is needed. In general, the same considerations made for the simple model with a single covariate are valid, independently, for each variable x_k.

6.1.4 Convergence

In the same spirit of Section 3.4, we can monitor the maximum absolute variations $|\hat{\psi}_k - \tilde{\psi}_k|$ throughout iterations to assess convergence.

6.2 Several change-points in a covariate

As discussed in Chapter 1, there exist several modern and efficient techniques to detect multiple change-points relevant to the same covariate. In particular, dynamic grid search algorithms are able to obtain exact solutions with a $O(n^2)$ computational cost, for any K. However, also the dynamic approach could become computationally expensive with huge sample sizes; a typical example concerns biological analyses involving thousands of gene expressions. In this section we propose an extension of the proposed algorithm for the multiple change-point case.

6.2.1 Model definition

Consider the more general piecewise constant regression model (1.5) with K change-points

$$E[Y_i] = \beta_{0k} \qquad x_i \in (\psi_{k-1}, \psi_k],$$

$k = 1, 2, \ldots, K+1$. At ψ_k, $k = 1, 2, \ldots, K$, the mean level of Y shifts instantaneously from β_{0k} to $\beta_{0\{k+1\}}$, while $\psi_0 = x_{(1)}$ and $\psi_{K+1} = x_{(n)}$; for sake of simplicity we ignore possible invariant terms ($\eta_i = 0$).

As for the single change-point model (2.1), we reparametrize the model using K indicator functions:

$$\mu_i = \beta_0 + \sum_{k=1}^{K} \beta_k I(x_i > \psi_k), \tag{6.8}$$

where $\sum_{j=0}^{k-1}\beta_k = \beta_{0k}, k = 1, 2, \dots, K + 1.$

Assuming that x consists of n distinct values, any value of ψ_k between x_i and x_{i+1} lead to the same likelihood.

Figure 6.3 displays the scatterplot and fitted values for a toy dataset, in which $x_i = i, i = 1, \dots, 100, \beta_0 = 2, \beta_1 = 1.5, \beta_2 = 1, \psi_1 = 25, \psi_2 = 65$ and $Y_i \sim \mathcal{N}(\mu_i, 1.5^2)$. Grid search provides $\hat{\beta}_0 = 1.79, \hat{\beta}_1 = 1.92, \hat{\beta}_2 = -1.00$, $\hat{\psi}_1 = 30$ and $\hat{\psi}_2 = 65$.

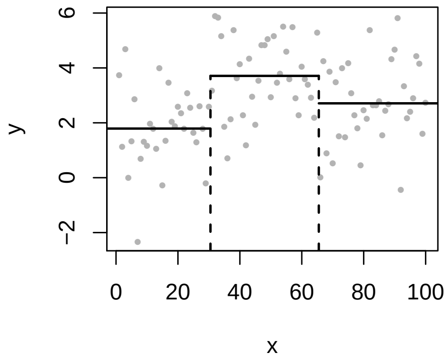

Figure 6.3: *An example of multiple ($K = 2$) change-point model. The mean level of Y is a step function (solid lines), with jumps in correspondence of* $\hat{\psi}_1 = 30$ *and* $\hat{\psi}_2 = 65$ *(dashed lines).*

Figure 6.4 shows the joint profile log-likelihood for ψ_1 and ψ_2, with lighter areas indicating higher likelihood values. Note symmetry with respect to the bisector of the first and third quadrant: the log-likelihood remains the same if we exchange ψ_1 and ψ_2. More in general, the likelihood is invariant to permutations of the ψ_k, so that it suffices to evaluate $\binom{n}{K}$ models. The log-likelihood is highly wiggly and point estimation through an iterative algorithm becomes a challenging task.

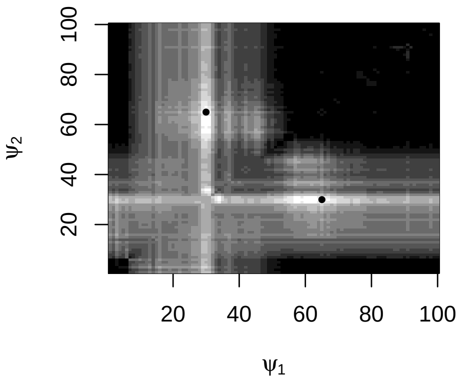

Figure 6.4: *Profile log-likelihood for ψ_1 and ψ_2. Lighter greys indicate higher values. The black points correspond to the maximum likelihood estimate. The likelihood is symmetric with respect to the bisector of the first and third quadrant, because $\ell(\psi_1, \psi_2) = \ell(\psi_2, \psi_1)$.*

6.2.2 Methods

To extend our iterative algorithm we use the key identity (2.3) in model (6.8), which gives

$$\begin{aligned}
\mu_i &= \beta_0 + \sum_{k=1}^{K} \beta_k \left(\frac{1}{2} \frac{x_i - \psi_k}{|x_i - \psi_k|} + \frac{1}{2} \right) \\
&= \beta_0 + \sum_{k=1}^{K} \beta_k \left(\frac{1}{2} \frac{x_i}{|x_i - \psi_k|} + \frac{1}{2} \right) + \sum_{k=1}^{K} (-\beta_k \psi_k) \left(\frac{1}{2} \frac{1}{|x_i - \psi_k|} \right) \\
&= \beta_0 + \sum_{k=1}^{K} \beta_k z_{ik} + \sum_{k=1}^{K} \gamma_k w_{ik}, \qquad (6.9)
\end{aligned}$$

where

$$\gamma_k = -\beta_k \psi_k \tag{6.10}$$

and the auxiliary (or 'working') covariates are

$$z_{ik} = \left(\frac{1}{2} + \frac{1}{2}\frac{x_i}{|x_i - \tilde{\psi}_k|}\right) \quad \text{and} \quad w_{ik} = \left(\frac{1}{2}\frac{1}{|x_i - \tilde{\psi}_k|}\right), \tag{6.11}$$

with $\tilde{\psi}_k$, $k = 1, 2, \ldots, K$, meaning approximate values.
Formulas above suggest the following simple iterative algorithm:

1. choose a vector of *starting values* $\tilde{\psi}_1, \tilde{\psi}_2, \ldots, \tilde{\psi}_K$;

2. compute the *working covariates* (6.11);

3. estimate the *working LM* (6.9) and extract $\hat{\beta}_k$ and $\hat{\gamma}_k$ for each k;

4. *update* the change-point values via

$$\hat{\psi}_k = -\frac{\hat{\gamma}_k}{\hat{\beta}_k}; \tag{6.12}$$

5. set $\tilde{\psi}_k = \hat{\psi}_k$ and *iterate* 2 to 4 until convergence.

Rescaling x values

The case of multiple change-point estimation requires moving the x_is away from the boundaries of the intervals $(\tilde{\psi}_{k-1}, \tilde{\psi}_k]$, $k = 1, 2, \ldots, K+1$. Extending the idea of Section 2.2.1 we compute a left and a right 'threshold' for each interval depending on a vector of rescaling factors $c_1, c_2, \ldots, c_K$,

$$\tilde{\psi}_k^- = \tilde{\psi}_k - c_k(\tilde{\psi}_k - \tilde{\psi}_{k-1}), \qquad \tilde{\psi}_k^+ = \tilde{\psi}_k + c_k(\tilde{\psi}_{k+1} - \tilde{\psi}_k),$$

and then scale covariate values in each interval $(\tilde{\psi}_{k-1}, \tilde{\psi}_k]$ induced by the change-points according to

$$x'_i = \tilde{\psi}^+_{k-1} + (x_i - \tilde{\psi}_{k-1})(1 - c_{k-1} - c_k), \tag{6.13}$$

$k = 1, 2, \ldots, K+1$, where

$$\tilde{\psi}_0 = \tilde{\psi}^+_0 = x_{(1)}, \qquad \tilde{\psi}_{K+1} = \tilde{\psi}^-_{K+1} = x_{(n)}$$

and $c_0 = c_{K+1} = 0$. The rescaled variables x' is used to compute auxiliary covariates (6.11) and to fit the working linear model (6.9).

Left panel in Figure 6.5 represents simulated observations in Figure 6.3 without rescaling x, namely when $c_1 = c_2 = 0$. Right panel shows the effect of two rescaling factors $c_1 = c_2 = 0.1$, when $\tilde{\psi}_1 = 33.5$ and $\tilde{\psi}_2 = 66.5$ are chosen as starting values.

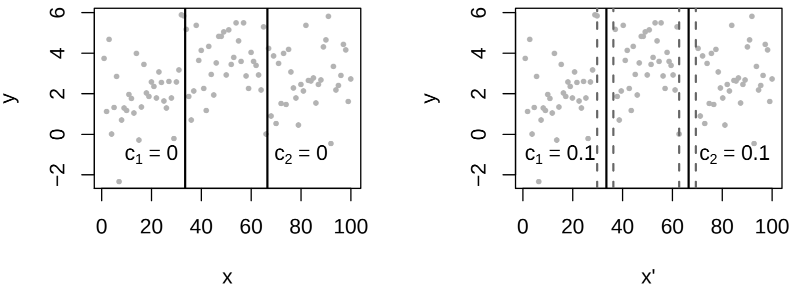

Figure 6.5: *Simulated data. Left panel: originary covariate ($c_1 = c_2 = 0$). Right panel: rescaled covariate ($c_1 = c_2 = 0.1$). The rescaling induces point-free intervals (dashed lines) in the neighbourhood of the approximate values $\tilde{\psi}_1$ and $\tilde{\psi}_2$ (solid lines).*

Notice the rescaling induces point-free intervals in the neighbourhood of the approximate values $\tilde{\psi}_1$ and $\tilde{\psi}_2$. As usual, decreasing the c_ks by some factor $d \in (0,1)$, turns out to be helpful to avoid convergence failures.

In summary, these are the steps of the algorithm:

1. choose a vector of *starting values* $\tilde{\psi}_1, \tilde{\psi}_2, \ldots, \tilde{\psi}_K$;
2. choose a vector of (possibly equal) *rescaling factors* $c_1, c_2, \ldots, c_K$;
3. *rescale* x using (6.13) to obtain x';
4. compute the *working covariates* (6.11) using x';
5. estimate the *working LM* (6.9) and extract $\hat{\beta}_k$ and $\hat{\gamma}_k$ for each k;
6. use (6.12) to *update* the change-point values;
7. *decrease* c_k, $k = 1, 2, \ldots, K$, using d if $\tilde{\psi}_k$ changes direction;
8. set $\tilde{\psi}_k = \hat{\psi}_k$ and *iterate* 3 to 7 until convergence.

6.2.3 Starting values

Good choice of the starting values represents a crucial aspect for the success of the algorithm. Rough indications about the change-point locations may be based on the use of non-parametric techniques, graphical inspections, or a priori knowledge. Without a priori information, intuition suggests to use equally spaced values $\tilde{\psi}_k = x_{(1)} + k\frac{x_{(n)} - x_{(1)}}{K+1}$.

6.2.4 Convergence

The maximum absolute variations $|\hat{\psi}_k - \tilde{\psi}_k|$ can be monitored, throughout iterations, to assess convergence.

Chapter 7

Discussion and future work

7.1 Discussion

We have introduced a novel iterative algorithm for estimation in regression models with piecewise constant relationships, where, to date, the grid search algorithm results to be the only feasible approach.

Our proposal is quite general, and several potential extensions have been discussed. In particular, grid search turns out to be unfeasible when the change-points are subject specific and modelled by random effects, and very difficult to apply when the parameter involved have unbounded supports, for example when the change-point is assumed to be a linear function of another variable. Some simulation studies on simple scenarios have shown that the proposed algorithm works also in such situations, and some motivating examples have been used to illustrate the proposed methodologies in practice.

For a piecewise constant regression model with K change-points, computational efficient grid search algorithms based on dynamic programming have been developed to reduce the computational cost from $O(n^K)$ to $O(n^2)$ for any K (Bai and Perron, 2003). However, when the multiple shifts are induced by several covariates the computational cost can not be reduced, because the change-points have different supports. Moreover, even when the shifts are induced by a single covariate, very large sample sizes still represent a concern for dynamic approaches. We have discussed the possibility to extent the proposed algorithm for the multiple change-point case, although we have not assessed them in practice.

Estimation of a standard linear regression model at each iteration represents the main feature of the proposed approach; in fact, whereas the stepwise function is not differentiable at the change-points, our approximation is linear in the parameters. In particular, we rely on a suitable equivalence for the indicator function which leads to a linear function when substituted into the model equation. Alternative approaches discussed in the literature make the objective function differentiable by using smooth approximations of the indicator function, typically a cumulative distribution function (Zhou *et al.*, 2008); however, the resulting model is non-linear, and approximate values both for the change-points and the regression coefficients are required for estimation. Conversely, we only need approximate values for the change-point parameters to initialize the proposed algorithm.

The profile likelihood with respect to the change-points is a highly wiggly step function, with typically many spurious maxima; for this reason, the choice of suitable starting values represents a crucial aspect for the algorithm to be successful. Rescaling the explanatory variable observations away from the starting values has been shown to reduce the probability of the algorithm to stop at some local solution. More in general, as shown through simulations, the dependency of the final result on the initialization parameters (the approximate change-point values and the rescaling factor) represents the main drawback of the proposed approach.

7.2 Future work

Even if the proposed algorithms appear to work reasonably well in practice, several aspects need to be investigated in more detail.

Some simulation studies have shown the effect of the initialization parameters on the estimator performances, and provided general hints for a suitable choice; however, the possibility to define rules for an optimal choice represents an open problem.

More extensive simulation studies may be helpful to improve the algorithm effectiveness and make it ready for applications; in particular, the algorithm for the multiple change-point case need to be refined and properly checked.

The proposed algorithms have been implemented in the *R* environment, with the aim of organizing them into a new package. A brief description of function usage is given in the Appendix.

We have considered the response variable to be Gaussian distributed, but extension to the exponential class appears to be straightforward, since the proposed algorithm works on the linear predictor.

Possibility to extend the proposed algorithm when using alternative objective functions should also be evaluated: sum of squared errors, sum of absolute errors (in quantile regression), partial likelihood (in Cox regression), quasi-likelihood, penalized likelihood and others.

Finally, the linear approximation used for the step function could be exploited to evaluate the possibility of deriving standard errors for the estimates, or, more in general, to perform interval estimation.

Appendix A

Implementation in *R*

The algorithms we have illustrated throughout this thesis have been implemented in the *R* environment, with the aim of organizing them into a new package. In particular, basic functions have been created to estimate:

1. models with K change-points;
2. mixed models with subject specific change-points;
3. models with a linear change-point;
4. models with K change-points in K covariates.

Such functions have been used for estimation in the examples and simulation studies. Here we provide a brief description of their usage.

A.1 Model with K change-points

Usage

```
PieceLin(y, x, psi, c = 0.05, d = 0.5,
         maxit = 50, tol = 0.01)
```

Arguments

y	a quantitative response variable;
x	a quantitative explanatory variable;
psi	a K-dimensional vector of approximate change-point values;
c	a K-dimensional vector of rescaling factors $c_k \in (0, 1)$; if scalar, the K factors are fixed at such constant value;
d	a reduction factor $d \in (0, 1)$;
maxit	maximum number of iterations admitted;
tol	a tolerance level $\Delta > 0$.

Value

An object of class ”lm” yielded by the working model at convergence.

A.2 Mixed model with subject specific change-points

Usage

```
PieceLinMix(y, x, id, psi, c = 0.05, d = 0.5,
            maxit = 50, tol = 0.01)
```

Arguments

y a quantitative response variable;

x a quantitative explanatory variable;

id a subject identifier;

psi a n-dimensional vector of approximate change-point values;
if scalar, the n values are fixed at such constant value;

c a n-dimensional vector of rescaling factors $c_i \in (0, 1)$;
if scalar, the n factors are fixed at such constant value;

d a reduction factor $d \in (0, 1)$;

maxit maximum number of iterations admitted;

tol a tolerance level $\Delta > 0$.

Value

An object of class ”lme” yielded by the working model at convergence.

A.3 Model with a linear change-point

Usage

```
PieceLinLin(y, x, v, psi, c = 0.05, d = 0.5,
            p, maxit = 50, tol = 0.01)
```

Arguments

y a quantitative response variable;

x a quantitative explanatory variable;

v a quantitative additional variable;

psi approximate intercept and slope of the change-point;

c a rescaling factor $c \in (0, 1)$;

d a reduction factor $d \in (0, 1)$;

p preliminar iterations before evaluating the likelihood;

maxit maximum number of iterations admitted;

tol a tolerance level $\Delta > 0$.

Value

An object of class ”lm” yielded by the working model at convergence.

A.4 Model with *K* change-points in *K* covariates

Usage

```
PieceLinMult(y, x, psi, c = 0.05, d = 0.5,
             maxit = 50, tol = 0.01)
```

Arguments

y	a quantitative response variable;
x	a *K*-dimensional list of quantitative explanatory variables;
psi	a *K*-dimensional list of approximate change-point values; if scalar, the *K* values are fixed at such constant value;
c	a *K*-dimensional list of rescaling factors $c_k \in (0, 1)$; if scalar, the *K* factors are fixed at such constant value;
d	a reduction factor $d \in (0, 1)$;
maxit	maximum number of iterations admitted;
tol	a tolerance level $\Delta > 0$.

Value

An object of class ”lm” yielded by the working model at convergence.

Bibliography

Andrews, D. W. (1993). Tests for parameter instability and structural change with unknown change point. *Econometrica: Journal of the Econometric Society*, pages 821–856.

Arfken, G. B. and Weber, H. J. (2011). *Mathematical methods for physicists: A comprehensive guide*. Academic press.

Bai, J. (1997a). Estimation of a change point in multiple regression models. *Review of Economics and Statistics*, **79**(4), 551–563.

Bai, J. (1997b). Estimatings multiple breaks one at a time. *Econometric Theory*, **13**(03), 315–352.

Bai, J. and Perron, P. (1998). Estimating and testing linear models with multiple structural changes. *Econometrica*, pages 47–78.

Bai, J. and Perron, P. (2003). Computation and analysis of multiple structural change models. *Journal of applied econometrics*, **18**(1), 1–22.

Balke, N. S. (1993). Detecting level shifts in time series. *Journal of Business & Economic Statistics*, **11**(1), 81–92.

Basseville, M. and Nikiforov, I. V. (1993). *Detection of abrupt changes: theory and application*, volume 104. Prentice Hall Englewood Cliffs.

Blythe, D. A., von Bunau, P., Meinecke, F. C. and Muller, K. (2012). Feature extraction for change-point detection using stationary subspace analysis. *Neural Networks and Learning Systems, IEEE Transactions on*, **23**(4), 631–643.

Box, G. E., Luceno, A. and del Carmen Paniagua-Quiñones, M. (2011). *Statistical control by monitoring and adjustment*, volume 898. John Wiley & Sons.

Boysen, L., Kempe, A., Liebscher, V., Munk, A. and Wittich, O. (2009). Consistencies and rates of convergence of jump-penalized least squares estimators. *The Annals of Statistics*, pages 157–183.

Braun, J. V., Braun, R. and Müller, H.-G. (2000). Multiple changepoint fitting via quasilikelihood, with application to dna sequence segmentation. *Biometrika*, **87**(2), 301–314.

Cobb, G. W. (1978). The problem of the nile: conditional solution to a changepoint problem. *Biometrika*, **65**(2), 243–251.

Davies, R. B. (1987). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika*, **74**(1), 33–43.

Dominicus, A., Ripatti, S., Pedersen, N. L. and Palmgren, J. (2008). A random change point model for assessing variability in repeated measures of cognitive function. *Statistics in medicine*, **27**, 5786–5798.

Donoho, D. L. and Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *Journal of the american statistical association*, **90**(432), 1200–1224.

Dumbgen, L. (1991). The asymptotic behavior of some nonparametric change-point estimators. *The Annals of Statistics*, pages 1471–1495.

Eilers, P. H. and De Menezes, R. X. (2005). Quantile smoothing of array cgh data. *Bioinformatics*, **21**(7), 1146–1153.

Fearnhead, P. (2006). Exact and efficient bayesian inference for multiple changepoint problems. *Statistics and computing*, **16**(2), 203–213.

Frick, K., Munk, A. and Sieling, H. (2014). Multiscale change point inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **76**(3), 495–580.

Fridlyand, J., Snijders, A. M., Pinkel, D., Albertson, D. G. and Jain, A. N. (2004). Hidden markov models approach to the analysis of array cgh data. *Journal of multivariate analysis*, **90**(1), 132–153.

Friedrich, F., Kempe, A., Liebscher, V. and Winkler, G. (2008). Complexity penalized m-estimation: fast computation. *Journal of Computational and Graphical Statistics*, **17**(1), 201–224.

Guha, S., Li, Y. and Neuberg, D. (2008). Bayesian hidden markov modeling of array cgh data. *Journal of the American Statistical Association*, **103**(482), 485–497.

Hall, C. B., Lipton, R. B., Sliwinski, M. and Stewart, W. F. (2000). A change point model for estimating the onset of cognitive decline in preclinical alzheimer's disease. *Statistics in medicine*, **19**(11-12), 1555–1566.

Hall, C. B., Ying, J., Kuo, L. and Lipton, R. B. (2003). Bayesian and profile likelihood change point methods for modeling cognitive function over time. *Computational Statistics & Data Analysis*, **42**(1), 91–109.

Hawkins, D. M. (1980). A note on continuous and discontinuous segmented regressions. *Technometrics*, **22**(3), 443–444.

Hawkins, D. M. (2001). Fitting multiple change-point models to data. *Computational Statistics & Data Analysis*, **37**(3), 323–341.

Hinkley, D. V. (1970). Inference about the change-point in a sequence of random variables. *Biometrika*, **57**(1), 1–17.

Horváth, L. (1993). The maximum likelihood method for testing changes in the parameters of normal observations. *The Annals of statistics*, pages 671–680.

Hsu, L., Self, S. G., Grove, D., Randolph, T., Wang, K., Delrow, J. J., Loo, L. and Porter, P. (2005). Denoising array-based comparative genomic hybridization data using wavelets. *Biostatistics*, **6**(2), 211–226.

Huang, T., Wu, B., Lizardi, P. and Zhao, H. (2005). Detection of dna copy number alterations using penalized least squares regression. *Bioinformatics*, **21**(20), 3811–3817.

Hušková, M. and Kirch, C. (2008). Bootstrapping confidence intervals for the change-point of time series. *Journal of Time Series Analysis*, **29**(6), 947–972.

Jackson, B., Scargle, J. D., Barnes, D., Arabhi, S., Alt, A., Gioumousis, P., Gwin, E., Sangtrakulcharoen, P., Tan, L. and Tsai, T. T. (2005). An algorithm for optimal partitioning of data on an interval. *Signal Processing Letters, IEEE*, **12**(2), 105–108.

Jackson, C. H. and Sharples, L. D. (2004). Models for longitudinal data with censored changepoints. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **53**(1), 149–162.

Jacqmin-Gadda, H., Commenges, D. and Dartigues, J.-F. (2006). Random changepoint model for joint modeling of cognitive decline and dementia. *Biometrics*, **62**(1), 254–260.

Jennrich, R. and Sampson, P. (1968). Application of stepwise regression to non-linear estimation. *Technometrics*, **10**(1), 63–72.

Jong, K., Marchiori, E., Van Der Vaart, A., Ylstra, B., Weiss, M. and Meijer, G. (2003). Chromosomal breakpoint detection in human cancer. In *Applications of Evolutionary Computing*, pages 54–65. Springer.

Khodadadi, A. and Asgharian, M. (2008). Change-point problems and regression: An annotated bibliography. *Collection of Biostatistics Research Archive*, **44**.

Killick, R., Fearnhead, P. and Eckley, I. (2012). Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, **107**(500), 1590–1598.

Kim, H.-J. and Siegmund, D. (1989). The likelihood ratio test for a change-point in simple linear regression. *Biometrika*, **76**(3), 409–423.

Kiuchi, A. S., Hartigan, J., Holford, T. R., Rubinstein, P. and Stevens, C. E. (1995). Change points in the series of t4 counts prior to aids. *Biometrics*, pages 236–248.

Kuan, C.-M. and Hornik, K. (1995). The generalized fluctuation test: A unifying view. *Econometric Reviews*, **14**(2), 135–161.

Küchenhoff, H. (1997). An exact algorithm for estimating breakpoints in segmented generalized linear models. *Computational Statistics*, **12**, 235–247.

Lai, T. L. and Xing, H. (2010). Sequential change-point detection when the pre-and post-change parameters are unknown. *Sequential Analysis*, **29**(2), 162–175.

Lai, Y. and Zhao, H. (2005). A statistical method to detect chromosomal regions with dna copy number alterations using snp-array-based cgh data. *Computational Biology and Chemistry*, **29**(1), 47–54.

Lavielle, M. (1999). Detection of multiple changes in a sequence of dependent variables. *Stochastic Processes and their Applications*, **83**(1), 79–102.

Lavielle, M. (2005). Using penalized contrasts for the change-point problem. *Signal Processing*, **85**(8), 1501–1510.

Loader, C. R. *et al.* (1996). Change point estimation using nonparametric regression. *The Annals of Statistics*, **24**(4), 1667–1678.

Muggeo, V. M. (2003). Estimating regression models with unknown breakpoints. *Statistics in medicine*, **22**(19), 3055–3071.

Muggeo, V. M., Atkins, D. C., Gallop, R. J. and Dimidjian, S. (2014). Segmented mixed models with random changepoints: a maximum likelihood approach with application to treatment for depression study. *Statistical Modelling*, page 1471082X13504721.

Muggeo, V. M. R. and Adelfio, G. (2011). Efficient change point detection for genomic sequences of continuous measurements. *Bioinformatics*, **27**(2), 161–166.

Olshen, A. B., Venkatraman, E., Lucito, R. and Wigler, M. (2004). Circular binary segmentation for the analysis of array-based dna copy number data. *Biostatistics*, **5**(4), 557–572.

Page, E. (1955). A test for a change in a parameter occurring at an unknown point. *Biometrika*, pages 523–527.

Pastor-Barriuso, R., Guallar, E. and Coresh, J. (2003). Transition models for change-point estimation in logistic regression. *Statistics in medicine*, **22**(7), 1141–1162.

Perron, P. (2006). Dealing with structural breaks. In *Palgrave handbook of econometrics*, pages 278–352.

Picard, F., Robin, S., Lavielle, M., Vaisse, C. and Daudin, J.-J. (2005). A statistical approach for array cgh data analysis. *BMC bioinformatics*, **6**(1), 27.

Pinheiro, J. C. and Bates, D. M. (2000). *Mixed-effects models in S and S-PLUS*. Springer.

Pinkel, D., Segraves, R., Sudar, D., Clark, S., Poole, I., Kowbel, D., Collins, C., Kuo, W.-L., Chen, C., Zhai, Y. *et al.* (1998). High resolution analysis of dna copy number variation using comparative genomic hybridization to microarrays. *Nature genetics*, **20**(2), 207–211.

Price, T. S., Regan, R., Mott, R., Hedman, Å., Honey, B., Daniels, R. J. *et al.* (2005). Sw-array: a dynamic programming solution for the identification of copy-number changes in genomic dna using array comparative genome hybridization data. *Nucleic acids research*, **33**(11), 3455–3464.

Rigaill, G., Lebarbier, E. and Robin, S. (2012). Exact posterior distributions and model selection criteria for multiple change-point detection problems. *Statistics and computing*, **22**(4), 917–929.

Rippe, R. C., Meulman, J. J. and Eilers, P. H. (2012). Visualization of genomic changes by segmented smoothing using an l0 penalty. *PloS one*, **7**(6), e38230.

Scott, A. and Knott, M. (1974). A cluster analysis method for grouping means in the analysis of variance. *Biometrics*, pages 507–512.

Shewhart, W. (1925). The application of statistics as an aid in maintaining quality of a manufactured product. *Journal of the American Statistical Association*, **20**(152), 546–548.

Siegmund, D. (2013). Change-points: from sequential detection to biology and back. *Sequential Analysis*, **32**(1), 2–14.

Tibshirani, R. and Wang, P. (2008). Spatial smoothing and hot spot detection for cgh data using the fused lasso. *Biostatistics*, **9**(1), 18–29.

Tishler, A. and Zang, I. (1981). A new maximum likelihood algorithm for piecewise regression. *Journal of the American Statistical Association*, **76**(376), 980–987.

Venkatraman, E. and Olshen, A. B. (2007). A faster circular binary segmentation algorithm for the analysis of array cgh data. *Bioinformatics*, **23**(6), 657–663.

Wang, P., Kim, Y., Pollack, J., Narasimhan, B. and Tibshirani, R. (2005). A method for calling gains and losses in array cgh data. *Biostatistics*, **6**(1), 45–58.

Worsley, K. (1983). Testing for a two-phase multiple regression. *Technometrics*, **25**(1), 35–42.

Worsley, K. J. (1986). Confidence regions and tests for a change-point in a sequence of exponential family random variables. *Biometrika*, **73**(1), 91–104.

Yao, Y.-C. and Au, S. (1989). Least-squares estimation of a step function. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 370–381.

Zeileis, A. (2005). A unified approach to structural change tests based on ml scores, f statistics, and ols residuals. *Econometric Reviews*, **24**(4), 445–466.

Zhang, N. R. and Siegmund, D. O. (2007). A modified bayes information criterion with applications to the analysis of comparative genomic hybridization data. *Biometrics*, **63**(1), 22–32.

Zhou, H., Liang, K.-Y. *et al.* (2008). On estimating the change point in generalized linear models. In *Beyond Parametrics in Interdisciplinary Research: Festschrift in Honor of Professor Pranab K. Sen*, pages 305–320. Institute of Mathematical Statistics.