Spline models with change-points

Matúš Maciak - *mmatthew@matfyz.cz*

Supervisor: *Marc Aerts - University of Hasselt, Belgium (aerts@uhasselt.be)*

Co-supervisor: *Niel Hens - University of Hasselt, Belgium (hens@uhasselt.be)*

Faculty of Mathematics and Physics, CHARLES UNIVERSITY IN PRAGUE
Department of Probability and Mathematical Statistics

Abstract: In this work we will focus on a problem of estimating an unknown regression function based on spline approach, while taking into account also some sudden jumps - change-points. We will propose an algorithm for both, continuous and discrete response variable which allows for possible change-points not only in the regression function itself, but also in all order derivatives.

We will discuss different initial settings required for spline fitting problem, like the degree and type of spline basis, the number and positions of knot points and also the way how to reveal jumps in all level derivatives. The proposed methods are applied to seroprevalence data of B19 parvovirus in Belgium population. Different models have been fitted and compared with respect to several statistical criteria. To have a better view on some aspects of this problem we have also proposed some simulation study.

Splines in statistics

are quite common modelling approach used nowadays as they can offer nice statistical properties together with an easy implementation and a straightforward interpretation.

Even if splines are not parametric in functional form, in most cases they may be written as a linear combination of some basis functions which usually have a polynomial representation. Locally they are defined by parameters and thus there is a certain parametric flavour. However, the set of all admissible functions that may be splines has a cardinality qual to $\mathbb{R}^{\mathbb{R}}$ which gives them enough flexibility to model almost any possible regression function. Moreover, being piecewise polynomial causes that the spline behaviour in one region may be totally unrelated to the bahavior in another region. This is not the case in some other nonparametric approaches. Splines are everywhere represented by simple polynomials, therefore they are easy to handle and their integrals and derivatives are also spline functions of degree higher or lower respectively.

Finaly, there is no need to have the whole data set available in order to reconstruct the estimate or to predict new observations. The only thing one really needs is a set of basis functions and the corresponding coefficients estimates.

Let $\{(X_i, Y_i);\ i = 1, \ldots, N\}$ be a random sample given from a 2-dimensional space $\mathbb{R}^2$ where the distribution of each couple (X_i, Y_i) is described by a joint desity function $f(x, y)$ and the relation between the dependent variable Y and the regressor variable X can be in general described as

$$\mathsf{E}[Y|X = x] = m(x), \quad \text{and} \quad m^{(p)}(x) = m_0^{(p)}(x) + \sum_{t=1}^{T_p} \alpha_{pt} \mathbb{I}_{\{x > x_{pt}\}},$$

where the function m_0 is considered to be smooth up to the order $n \in \mathbb{N}$, $p < n$ stands for the order of derivative we consider and x_{pt} are corresponding change-points locations with the size of a jump equal to α_{pt}.

In case of continuous response variable the estimate is simply defined as a linear combination of some well-defined spline basis functions, where the coefficients of this linear combination are given as a solution to the following minimization problem

$$\widehat{\Theta} = \underset{\Theta \in \mathbb{R}^K}{\text{Argmin}} \left[\sum_{i=1}^{N} \left(Y_i - \sum_{j=1}^{K} \theta_j \psi_j(X_i) \right)^2 + \lambda\, \Theta^\top \mathbf{D} \Theta \right],$$

where λ stays for smoothing parameter and the matrix $\mathbf{D}$ is some diagonal matrix with zeros and ones on its diagonal.

In case of logistic regression where the parameter of interest ζ is modeled via a link function by $m(\cdot)$ the estimates for basis coefficients are given by the maximization problem

$$\widehat{\Theta} = \underset{\Theta \in \mathbb{R}^K}{\text{Argmax}} \left[\sum_{i=1}^{N} \left\{ Y_i \zeta_i - \log(1 + e^{\zeta_i}) \right\} - \frac{1}{2} \lambda\, \Theta^\top \mathbf{D} \Theta \right],$$

where we consider an exponential density distribution family for Y_i, given by $f(y) = \exp\{\frac{y\zeta + b(\zeta)}{a(\phi)} + c(y, \phi)\}$.

1. Basis selection

With respect to interpretability and nice properties we have used the third order ($n = 3$) truncated power spline basis. The actual choice of type of basis (B-splines, modified B-splines or truncated power splines) is not so much important as one can easily show an equivalence even with respect to smoothing parameter λ involved in this problem.

2. Knots selection

We have used $k \in \mathbb{N}$ equidistantly spaced inner knot points for the main spline basis (basis with no allowance for change-points) and later on we have implemented an automatical data-driven procedure proposed by Stone in order to find the optimal locations for change-points knots via minimizing of the GCV criterion. Such a mesh of knots $\Delta = \{\xi_1, \ldots, \xi_k\}$ was used to fit the final estimate.

3. Model selection

We have proposed to use the GVC criterion for model selection with respect to knots locations and smoothing parameter value but as far as the GCV can lead to undersmoothing for large sample sizes we have based our final decision on the BIC criterion which can slightly avoid this unconvenient property.

One can nicely verify that the number of basis functions K is given within the set $\{n + k + 1, \ldots, (n + 1) \times (k + 1)\}$, where k stand for the number of inner knot points, the mesh $\Delta = \{\xi_1, \ldots, \xi_k\}$, considered for the building of the spline basis ($K = n + k + 1$ if we consider just a simple spline model with no change-points, $K = (n + 1) \times (k + 1)$ if we consider the full, saturated model with possible change-points in all knot points at all order derivatives).

By using a matrix representation for the basis coefficients $\Theta \in \mathbb{R}^K$ and the set of basis functions $\{\psi_{11}(x), \ldots, \psi_{(n+1)(k+1)}\}$ defined as

$$\Theta = (\theta_1, \ldots, \theta_{n+1}, \theta_{n+2}, \ldots, \theta_{2(n+1)}, \ldots\ldots, \theta_{(k+1)(n+1)})^\top,$$

$$\mathbb{X}_T = \begin{pmatrix} \psi_{11}(X_1) & \ldots & \psi_{1(n+1)}(X_1) & \psi_{21}(X_1) & \ldots & \psi_{2(n+1)}(X_1) & \ldots & \psi_{(k+1)(n+1)}(X_1) \\ \psi_{11}(X_2) & \ldots & \psi_{1(n+1)}(X_2) & \psi_{21}(X_2) & \ldots & \psi_{2(n+1)}(X_2) & \ldots & \psi_{(k+1)(n+1)}(X_2) \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \ldots & \vdots \\ \psi_{11}(X_N) & \ldots & \psi_{1(n+1)}(X_N) & \psi_{21}(X_N) & \ldots & \psi_{2(n+1)}(X_N) & \ldots & \psi_{(k+1)(n+1)}(X_N) \end{pmatrix},$$

the estimate for Θ can be written in sence of normal equations. However, in case of logistic regression the solution to this equation is not given explicitly as the value of the canonical parameter ζ_i depends via $m(\cdot)$ also on the unknown parameter Θ. Therefore, we have proposed a modified penalized quasilikelihood approach based on Newton-Raphson iterations in order to get a final solution while taking into account also a penalty term.

Fitting algorithm

Iterate: $\Theta^{\text{new}} = \Theta^* - (\mathsf{H}(\Theta^*))^{-1} \cdot \mathsf{S}(\Theta^*)$, where

1. $S(\Theta^*) = \mathbb{X}_T^\top \left(\mathbf{Y} - \frac{e^{\zeta}}{1 + e^{\zeta}} \right) = \mathbb{X}^\top \left(\mathbf{Y} - \frac{\exp\{\mathbb{X}_T^\top \Theta^*\}}{1 + \exp\{\mathbb{X}_T^\top \Theta^*\}} \right) = 0;$

2. $H(\Theta^*) = (-1) \cdot \left(\mathbb{X}_T^\top \mathbf{W} \mathbb{X}_T + \lambda \mathbf{D} \right);$
$H(\Theta^*) = (-1) \cdot \left(\mathbb{X}_T^\top \mathbf{W} \mathbb{X}_T + \lambda_1 \mathbf{D}_1 + \lambda_2 \mathbf{D}_2 \right)$ respectively.

where $\mathbf{W} = diag(b''(\zeta_1, \ldots, \zeta_N))$. The iterative procedure is repetead until convergence, defined as $|\Theta^{\text{new}} - \Theta^*| \le \epsilon$ for some some small $\epsilon > 0$, where

$$|\Theta^{\text{new}} - \Theta^*| \le \epsilon \iff \forall_{j=1,\ldots(k+1)\times(n+1)}\, |\theta_j^{\text{new}} - \theta_i^*| \le \epsilon \ \text{ and } \ \exists_j\, |\theta_j^{\text{new}} - \theta_i^*| < \epsilon.$$

1. Simulation study

We have proposed an extensive simulation study to see how the algorithm works in case of more suitable data sets and to reveal an importance of different initial setting for the final spline estimate.

1. *The original data set were not shown to be fully appropriate for change-points implementation as the variability in data was much higher than the possiblejump sizes in the original regression function $m(\cdot)$. Therefore, the proposed algorithms do not have enough power to fully reveal all change-points, which could be present.*
2. *We want to see the performace of multiple penalization method which we have proposed, where one penalizes for each level of basis coefficients separately. One can easily separate coefficiets which are responsible for a regular spline estimate and coefficients which corresponds with change-points occurences and to impose a double penalization method instead. This means that $\lambda\ \Theta^\top \mathbf{D} \Theta$ is replaced by $(\lambda_1\ \Theta^\top \mathbf{D}_1 \Theta + \lambda_2\ \Theta^\top \mathbf{D}_2 \Theta)$, where both parameters λ_1 and λ_2 need to be estimated via some model selection method and $\mathbf{D}_1$ and $\mathbf{D}_2$ are diagonal matrixes with zeros and ones on their diagonals. One can even extend double penalization and to impose a multiple penalization however, this becomes too much intensive for higher order derivatives.*

All important details coming from the simulation study are stated in our report. However, mainly we can state that:

- **the multiple penalization performs much betther than the simple penalization, with respect to both criteria, GCV and BIC;**
- **the proposed jump detection algorithm was able to reveal all important jumps in the simulated data sets, resulting in a quite reliable model;**
- **the optimal model selection was in correspondence with the real regression function considered for simulations;**

Lasso selection approach

We have also considered the lasso method proposed by Tibshirany, where all important coefficients were picked up directly from the saturated model. This has performed quite nice for simulated data however, in case of real application we have obtained better results with the jump detection algorithm.

2. Current status data (continuous response)

We have considered B19 parvovirus antibody level in Belgium population, given the age of a patient. Optimal model was taken from the class of all plausible models which considered at least one change-point (selection via BIC criterion).

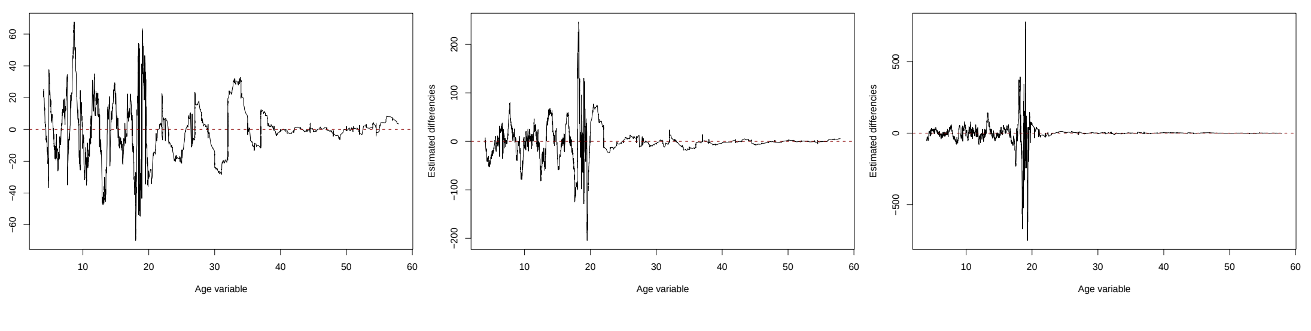

Fig.1 Jump detection algorithm for the function itself and the first two derivatives. However, the boundary limits for this algorithm vere much higher than the ploted values.

Based on a simulation study, we have fitted double penalized spline estimate which can provide better with respect to selection criteria (GCV and BIC). The final model considers one zero-order change-point located at the age of 23 years.

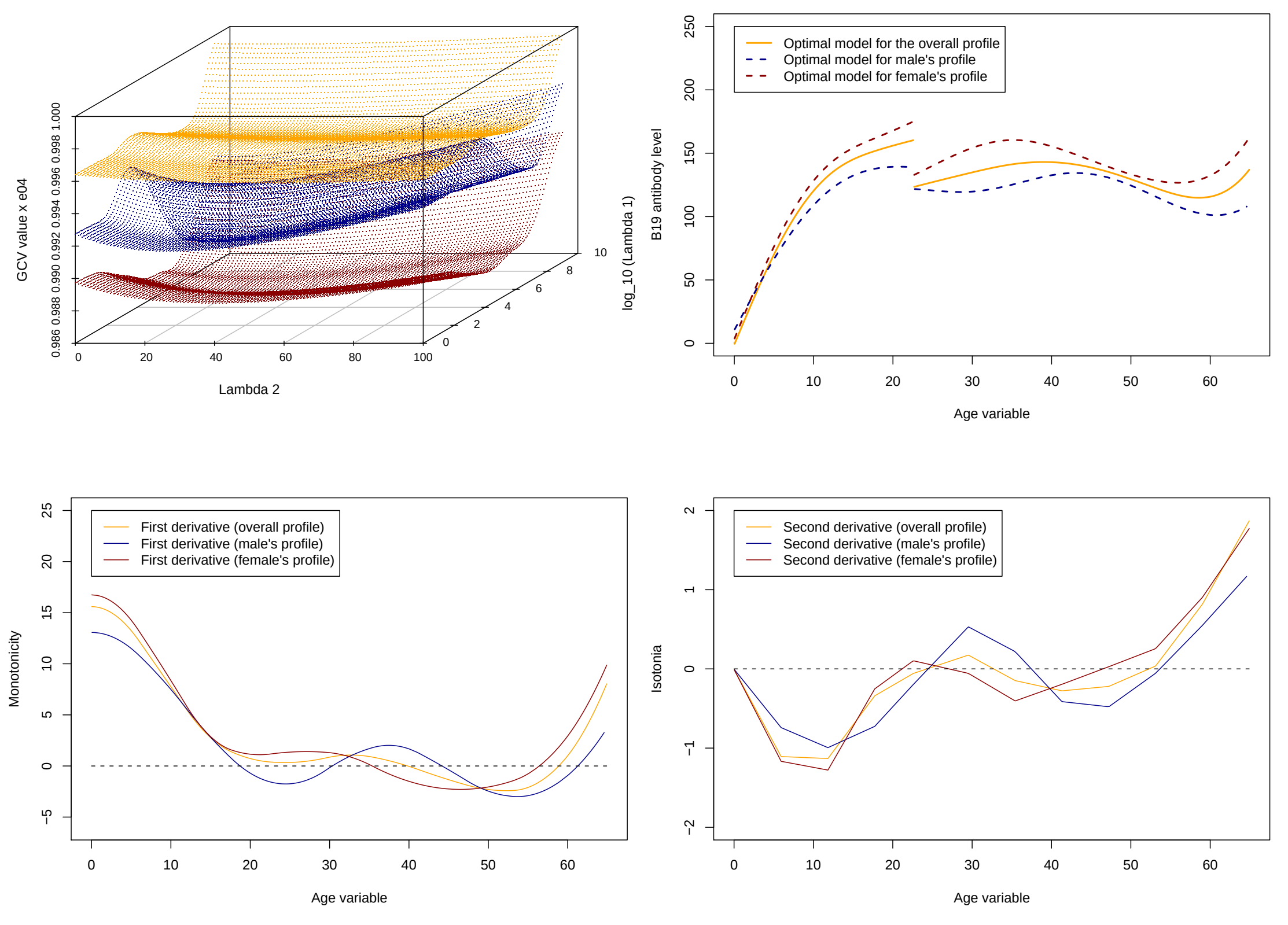

Fig.2 2D generalized cross-validation function with the corresponding final spline model for the overall mean profile, male and female profiles given together with the first two derivatives.

Model	order	inner knots	d.f.	GCV	BIC	$log_{10}\lambda_1$	$log_{10}\lambda_2$
Overall model	3	10/equidistant	7.939	37 079 980	0.3626	1.8808	1.4623
Males only	3	10/equidistant	7.154	16 179 478	0.5766	1.8808	1.4623
Females only	3	10/equidistant	7.214	20 588 471	0.5735	1.8808	1.4623

The zero order jump	Overall model	Model for males	Model for females
Size of a jump	$\alpha_{\text{overall}} = 36.7683$	$\alpha_{\text{male}} = 17.2469$	$\alpha_{\text{female}} = 41.9551$

3. Seroprevalence data (discrete response)

For fitting a logistic regression model we have assumed a prior knowledge: from the specific relation between the current status data and the seroprevalence data we have assumed the same change-points behaviour at the same knot-points locations. The final model again consideres only one zero-order change-point located at the age of 23 years.

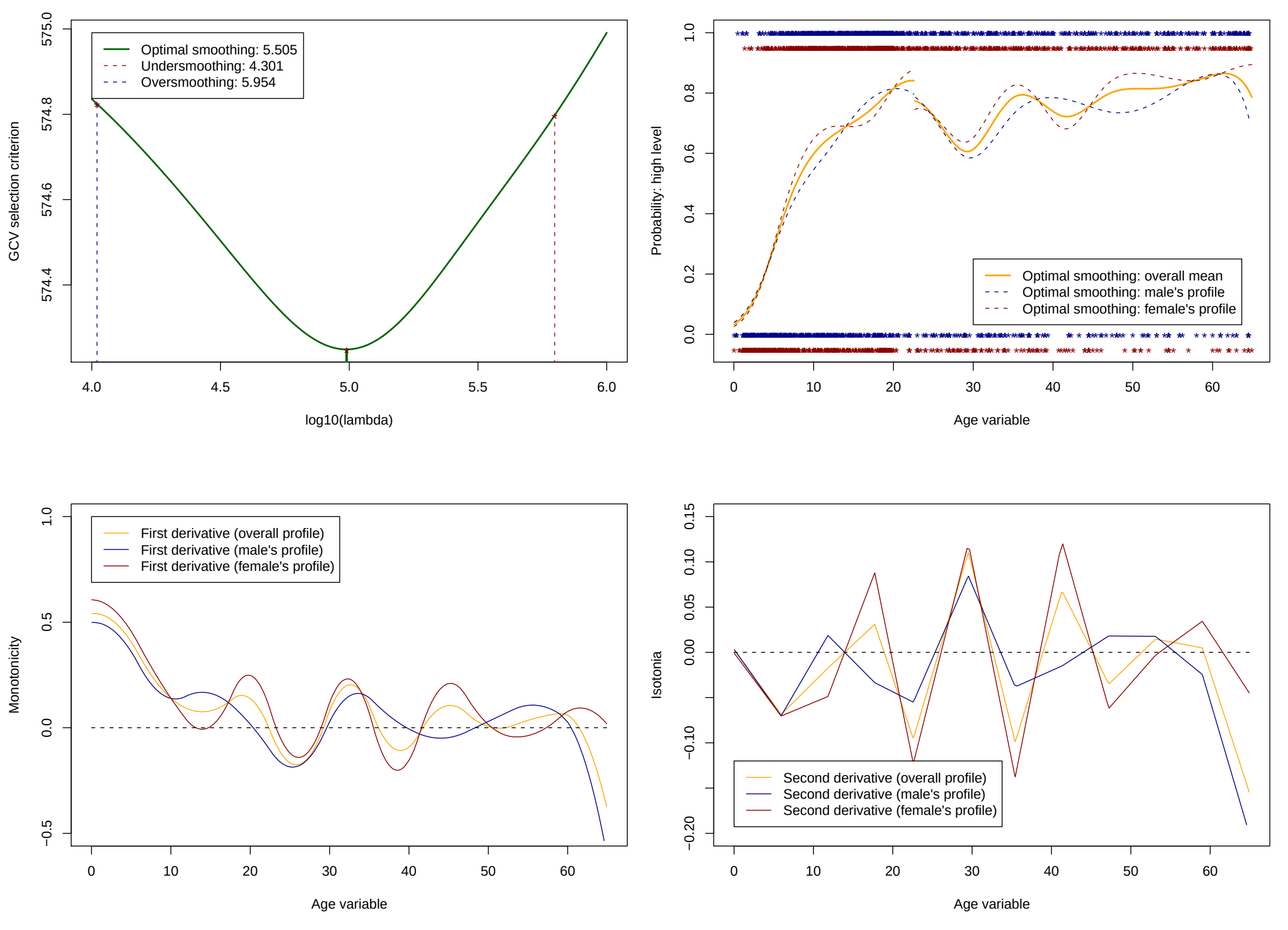

Fig.3 Generalized cross-validation function with the corresponding logistic spline model for the overall mean profile, male and female profiles given together with the first two derivatives.

Model	order	inner knots	d.f.	GCV	BIC	$log_{10}\lambda$	Size of jump
Overall model	3	10/equidistant^4	10.405	574.25	0.1731	4.9899	0.0678
Males only	3	10/equidistant^4	9.782	290.59	0.2689	4.9899	0.0108
Females only	3	10/equidistant^4	9.705	284.09	0.2584	4.9899	0.1311

Bibliography

[1] M.Maciak. (2008). *Spline Models with Change-points, Applied to Seroprevalence Data*, Universiteit Hasselt Press, Diepenbeek

Note: All other bibliography is listed at the end of our report (Maciak (2008)).

Acknowledgement:
We would like to express our thanks to the ČSOB bank for supporting our and our presence at the conference ROBUST '08.