Spline models with change-points
Matúš Maciá
mmatthew@matfyz.cz

1. Basis selection
With respect to interpretability and nice properties we have used the third degree B-spline basis.

2. Knots selection
We have used k ∈ N regulärly spaced inner knots for the main-spline basis (with no allowance for change-points) and later on we have implemented an algorithmdriven data-driven procedure proposed by Stone in order to find the optimal locations for change-points based on the chosen B-spline basis. The optimal knots on the main-spline basis and the knots of knots Δ = (ξ1, ..., ξq) were used to fit the final estimate.

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

One can nicely verify that the number of basic functions K is given within the set of knots Δ = (ξ1, ..., ξq), considered for the building of the spline basis (K = n + k + 1) if we consider just a number of inner knots, i.e., the mesh Δ = (ξ1, ..., ξq) considered for the full, saturated model with possible change-points in all knots at all order derivatives. Hence, by using a matrix representation for the basic coefficients Θ ∈ Rk and the set of basic functions fi(x), i = 1, ..., k, defined as

\[ \Theta = \left( \begin{array}{c} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{array} \right) \]

we estimate θ for Θ in some normal equations. However, in case of logistic regression the solution to this equation is not obtained as explicitly as in the case of the canonical parameter ξ ∈ R, also we don’t have any parameters. Therefore, we have proposed a modified penalized likelihood approach based on Newton-Raphson iterations in order to get a final solution while taking into account also a penalty term.

2. Current status data (continuous responses)
We have considered B19 parovirus antibody level in Belgian population, given the age of a patient. Optimal model was taken from the fit of all possible models which considered at least one change-point (selection via BIC criterion).

3. Seroprevalence data (discrete responses)
For fitting a logistic regression model we have assumed a prior knowledge from the epidemiological between the current status data and the seroprevalence data; we have assumed the same change-points behaviour of the same knot-points locations. The final model again considers only one zero-order change-point located at the age of 23 years.

Bibliography


[2] All other bibliography is listed at the end of our report (Maciá (2008)).

Acknowledgment:
We would like to express our thanks to the CSOB bank for supporting our and our presence at the conference ROBUST ’08.