Derivative Estimation, B-Splines and P-Splines in R

Andrew Simpkin[1], Alberto Alvarez[1], Jochen Einbeck[2] and John Newell[1,3]

[1] Department of Mathematics, Statistics and Applied Math., National University of Ireland, Galway, Ireland
[2] Department of Math. Sciences, Durham University, Durham DH1 3LE, UK
[3] Clinical Research Facility, National University of Ireland, Galway, Ireland

Derivative estimation is important in a wide range of disciplines. It is often the case that when analysing data, the derivative, or rate of change, of the observed data is of primary interest. In particular, consider the regression model $y = m(x_i)+e_i$, one is often not interested in $m(\cdot)$ itself, but rather in the relative change dm/dx of m when increasing or decreasing x by a small value dx. In the situation where we have data observed over time, the first derivative will correspond to velocity, the second to acceleration. Derivatives are also used in asymptotic approximations to get confidence intervals, optimal bandwidths etc. A further field of application for derivative estimators are change point problems. For instance, when analyzing blood lactate data of elite athletes, one is interested in the workload at which the lactate level suddenly rises, which can be detected by finding the maximum of the 2nd derivative (Newell et al., 2005).

A popular tool for derivative estimation is spline smoothing, with a large number of variants (B-Splines, P-Splines) being available. Choice of smoothing parameter when the derivative is of primary concern is an area on which there seems to be no consensus for the optimal choice. Generally the smoothing parameter is selected based on optimising $\hat{m}$ and not the derivative, which may lead to considerable undersmoothing (Jarrow et al., 2004). In this poster we present the results of a small simulation study to compare classical choices for smoothing parameter selection such as GCV, minimising the AIC criterion, etc. when the derivative is of primary concern.

We present a comparative overview of derivative estimation routines in R; we explore the features of each, how the smoothing parameter is selected within the routines, their advantages and disadvantages. The spline based methods; in particular **pspline** and **SemiPar** work quite well and offer several interesting options. The function `D1D2` in package **sfsmisc** addresses the undersmoothing problem mentioned above by adding a "fudge" offset to the GCV-selected smothing parameter.

An attempt will be made to clear up confusion surrounding B-Splines and P-Splines in R. For example the function `smooth.Pspline` in package **pspline** does not actually use B-Splines, but penalized smoothing splines. The function `smooth.spline` in package **stats**, however, uses B-Splines if the number of knots specified is smaller than n, and it also features penalization.

References

[Jarrow, R., Ruppert, D., and Yu, Y.] 2004). Estimating the term structure of corporate debt with a semiparametric penalized spline model, *JASA*, **99**, 57-66.

[Newell, J., Einbeck, J., Madden, N., and McMillan, K.] 2005). Model free endurance markers based on the second derivative of blood lactate curves. In: Francis et al. (Eds), *Proceedings of the 20th IWSM*, 357-364, Sydney.