Automatic Numerical Differentiation of Noisy, Time-Ordered Data in R

Ravi Varadhan[1,*] **, Ganesh Subramaniam**[2]

1. Center on Aging and Health, Johns Hopkins University
2. AT&T Labs - Research

* Contact author: rvaradhan@jhmi.edu

Keywords: Numerical derivative, Smoothing, FDA, Time Series Data Mining‘

In nonparametric regression, it is often of interest to estimate some functionals of a regression function, such as its derivatives. For example, in biomechanics, the estimation of derivatives of displacement data (e.g. velocity and acceleration of body segments) is a common task (Pezzak 1977). In the study of growth curves, the first (velocity) and second (spurt) derivatives of the height as a function of age are important parameters for study (Ramsay & Silverman 1997).

In this work, we study the estimation of derivatives of regression functions using nonparametric regression approaches. Suppose we observe

$$y_j = g(x_j) + \epsilon_j, j = 1, ..., n,$$

where the ϵ_js are uncorrelated with $E(\epsilon_j) = 0$ and $E(\epsilon_j^2) = \sigma^2 > 0$, and the design points $x_j{}_{j=1}^n$ are either equally-spaced or randomly distributed. The main interest lies in the estimation of $g'(x_j)$ and $g''(x_j)$.

This investigation on derivative estimation was motivated by a time series data mining problem in telecommunication network data, where given a large number of time series the objective was to indetify time series that have potentially ‘interesting behavior’ (Subramaniam and Varadhan 2007 and 2008). Ramsay and Silverman (1997) demonstrated several examples where features that are not visible in the original data are detected in the first and second derivatives. A critical parameter in nonprametric regression is the bandwidth for smoothing noisy data. With the large number of curves, optimal bandwidth selection on a curve-by-curve basis is not practical, but some reasonable approximation of optimal bandwidth using, for example, generalized cross-validation or plug-in-bandwidth, might be acceptable. However, the main problem is that the optimal bandwidth for estimating regression function is generally smaller than the optimal bandwidth for estimating the derivatives (Hardle 1985). Thus automatic smoothing of optimal derivative estimation in a data mining setting presents a unique challenge.

Various nonparametric regression approaches are available in R: kernel regressions, smoothing splines, penalized splines and local polynomial. Our objective here is two-fold: (1) to compare the different smoothing techniques for their accuracy in estimating the first and second derivatives of the regression function using the automatic bandwidth selection techniques that are applied to the noisy data itself, and (2) to propose and evaluate some approaches for approximating optimal bandwidth for derivative estimation. We conduct a comprehensive simulation study involving: (a) various nonparametric regression methods (smoothing splines, penalized splines, kernel smoothing and local polynomial), (b) different regression functions, (c) two types of designs - equally spaced or randomly distributed time points, (d) different signal-to-noise rations, and (e) homoscedastic and heteroscedastic errors. We evaluate the performance in terms of mean integrated squared error, integrated squared bias, and integrated variance. In addition to the simulation results from function and derivative estimation, we will also use simulated telecommunications network data to demonstrate how these techniques can be used in anomaly detection.

References

J. O. Ramsay and B. W. Silverman (1997). Functional Data Analysis. *Cambridge University Press.*

Subramaniam, G. and Varadhan, R (2007). Feature Extraction using FDA, JSM 2007 (Salt Lake City, UT,USA), Aug. 2007

Subramaniam, G. and Varadhan, R (2008). Borrowing Strength In Time Series Data Mining, JSM 2008 (Denver, CO,USA), Aug. 2008

Pezzak, JC, Norman RW, and Winter DA (1977). An assessment of derivative determining techniques used for motion analysis, *J of Biomechanics*, 10: 377–382.