

## Estimating Multiple Derivatives Simultaneously: What Is Optimal?

Richard Charnigo\* and Cidambi Srinivasan

Department of Statistics, University of Kentucky, Lexington KY 40506-0027

Nonparametric regression techniques including kernel smoothing [1], spline smoothing [2], and local regression [3] are useful for estimating a mean response function  $\mu(x)$  in the statistical model  $Y_i = \mu(x_i) + \epsilon_i$  when one is unwilling to assume that  $\mu(x)$  is linear (or polynomial of higher but known degree) in the covariate  $x$ . These same techniques can also be employed to estimate one or more derivatives of  $\mu(x)$ . While the techniques differ in their details, they have a common underlying theme. One specifies a covariate value  $x_0$  and estimates  $\mu(x)$  or one of its derivatives at  $x_0$  by solving an optimization problem that is localized to a neighborhood of  $x_0$ , in that only observations with covariate values inside the neighborhood contribute substantively to the solution. For example, the simplest incarnation of this theme is to define  $\mu(x_0)$  to be the average of all responses  $Y_i$  for which  $|x_i - x_0|$  is sufficiently small. As one slides  $x_0$  through a continuum of all possible covariate values, an estimated mean response or derivative is then traced out. Selecting the neighborhood size is a crucial implementation decision to which much literature has been devoted [4].

Under mild conditions, including appropriate dependence of the neighborhood size on the sample size  $n$ , Stone [5] established that local regression yields an optimal convergence rate of  $n^{-(J+1-k)/(2J+3)}$  in estimating  $\frac{d^k}{dx^k} \mu(x)$  for  $0 \leq k \leq J$  when  $\mu(x)$  has  $(J+1)$  bounded derivatives. However, optimality may be defined even more stringently than the attainment of a particular convergence rate. For instance, optimality may entail minimizing mean square error or an asymptotic approximation thereto. Yet, kernel and local regression estimators of  $\frac{d^k}{dx^k} \mu(x)$  with minimal mean square error are not the  $k^{\text{th}}$  order derivatives of kernel and local regression estimators of  $\mu(x)$  with minimal mean square error [6,7]. While the existing literature thus provides guidance on the optimal estimation of  $\frac{d^k}{dx^k} \mu(x)$  by itself, or of  $\mu(x)$  by itself, the existing literature does not elucidate what is optimal for the simultaneous estimation of  $\mu(x)$  and  $\frac{d^k}{dx^k} \mu(x)$  or, more generally, the simultaneous estimation of  $\mu(x)$  and all of its derivatives up to order  $J$ . Here we clarify that by simultaneous we refer not merely to the explicit estimation of multiple derivatives in a single data analysis but also to the requirement that  $\widehat{\frac{d^k}{dx^k} \mu(x)}$  and  $\widehat{\mu(x)}$  honor the same functional relationship as  $\frac{d^k}{dx^k} \mu(x)$  and  $\mu(x)$ , namely that  $\widehat{\frac{d^k}{dx^k} \mu(x)} = \frac{d^k}{dx^k} \widehat{\mu(x)}$ . Charnigo and Srinivasan [8] have termed this requirement "self-consistency".

There are several practical applications in which estimating a mean response function and its derivatives may help to address important scientific questions. These applications include the modeling of:

- Human height [9], for which the first derivative is the growth rate and the second derivative can be employed to delineate time intervals over which growth is speeding up or slowing down;
- Kidney function for a lupus nephritis patient [10], for which the first derivative quantifies the progress of the disease and the second derivative can be used to delineate time intervals over which the disease is progressing unstably;
- Scattering profiles of submicroscopic nanoparticles [11], for which

the mean response function and its derivatives may be employed like "fingerprints" to identify nanoparticles of unknown size or structure given existing results for nanoparticles of known size and structure; and,

- Raman spectra of bulk materials [4], for which the mean response function and its derivatives may likewise be used to identify materials of unknown chemical composition. The Raman spectrum application may be particularly interesting to readers of this journal because of its potential to detect impurities as part of a quality control process in pharmaceutical production [12] and its potential to complement existing mammography and ultrasound technology for the noninvasive diagnosis of breast cancer via the detection of calcified lesions [13].

In some of these practical applications, one may reach contradictory scientific conclusions if  $\mu(x)$  and its derivatives are not estimated simultaneously. For example, Charnigo and Srinivasan [8] illustrated the consequences of having inequalities among  $\frac{d^2}{dx^2} \{\widehat{\mu(x)}\}$ ,  $\frac{d}{dx} \left\{ \widehat{\frac{d}{dx} \mu(x)} \right\}$ , and  $\frac{d^2}{dx^2} \mu(x)$  in the human height application. Employing local regression, Charnigo and Srinivasan [8] found that the estimated first derivative for one child had a local maximum at 10.5 years, suggesting that the child's growth spurt peaked at 10.5 years. On the other hand, the estimated second derivative for that same child was nonzero at 10.5 years. The closest zero of the estimated second derivative was at 10.1 years, translating to a discrepancy of five months in pinpointing the peak of the growth spurt. While there is inherently some uncertainty about when the growth spurt peaked, acquiring two different estimates from a single data analysis is unsettling. The preceding illustration thus demonstrates that insisting upon optimal estimation of  $\mu(x)$  by itself, of  $\frac{d^k}{dx^k} \mu(x)$  by itself, of  $\frac{d^2}{dx^2} \mu(x)$  by itself, and similarly for higher order derivatives may lead to incoherent scientific conclusions.

We therefore perceive the need for a new criterion by which optimality may be defined when multiple derivatives are estimated simultaneously. Such a criterion would evaluate a family of self-consistent estimators  $\widehat{\mu(x)}, \widehat{\frac{d}{dx} \mu(x)}, \widehat{\frac{d^2}{dx^2} \mu(x)}$  and so forth rather than evaluating each estimator by itself. Ideally, this criterion would favor good estimation of several derivatives over excellent estimation of one derivative accompanied by poor estimation of the remaining derivatives. Hence, a family of estimators deemed optimal by such a criterion would not be anticipated to include, for example, a  $\widehat{\mu(x)}$  that

\*Corresponding author: Department of Statistics, University of Kentucky, Lexington KY 40506-0027; E-mail: [RJCharn2@aol.com](mailto:RJCharn2@aol.com)

Received March 28, 2011; Accepted March 30, 2011; Published March 31, 2011

Citation: Charnigo R, Srinivasan C (2011) Estimating multiple derivatives simultaneously: What is optimal? J Biomet Biostat 2:102e. doi:10.4172/2155-6180.1000102e

Copyright: © 2011 Charnigo R. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

minimized the mean square error in estimating  $\widehat{\mu(x)}$ ; the derivatives of such a  $\widehat{\mu(x)}$  would likely be too undersmoothed to serve as good proxies for the derivatives of  $\mu(x)$ . Likewise, an optimal family would not be anticipated to include a  $\widehat{\frac{d^J}{dx^J}\mu(x)}$  that minimized the mean square error in estimating  $\widehat{\frac{d^J}{dx^J}\mu(x)}$ ; the antiderivatives of such a  $\widehat{\frac{d^J}{dx^J}\mu(x)}$  would likely be too oversmoothed to serve as good proxies for the antiderivatives of  $\widehat{\frac{d^J}{dx^J}\mu(x)}$ .

What, then, might such a criterion look like? One idea would be to consider the sum of mean square errors,  $\sum_{k=0}^J E \left[ \left( \widehat{\frac{d^k}{dx^k}\mu(x)} - \frac{d^k}{dx^k}\mu(x) \right)^2 \right]$ , with the rationale that the sum could be minimized only if each derivative were well estimated. Yet, because  $E \left[ \left( \widehat{\frac{d^J}{dx^J}\mu(x)} - \frac{d^J}{dx^J}\mu(x) \right)^2 \right]$  might be much larger than any other summand, the sum might overemphasize the estimation of  $\widehat{\frac{d^J}{dx^J}\mu(x)}$  and thereby lead to oversmoothing in the estimation of  $\mu(x)$  and its lower order derivatives. A less naive idea would be to consider a weighted sum of mean square errors,  $\sum_{k=0}^J a_k^2 E \left[ \left( \widehat{\frac{d^k}{dx^k}\mu(x)} - \frac{d^k}{dx^k}\mu(x) \right)^2 \right]$ , or the mean square error of a weighted sum of derivative estimators,  $E \left[ \left( \sum_{k=0}^J a_k \widehat{\frac{d^k}{dx^k}\mu(x)} - \sum_{k=0}^J a_k \frac{d^k}{dx^k}\mu(x) \right)^2 \right]$ . Either way, a sensible specification of  $a_0$  through  $a_J$  would be required for the criterion to serve its intended purpose. In light of Stone's [5] theory, one might think to let  $a_k$  escalate in proportion to  $n^{(J+1-k)/(2J+3)}$  as the sample size  $n$  increased. However, prescribing  $a_k = c_k n^{(J+1-k)/(2J+3)}$  with  $c_k$  not dependent on  $n$  only reduces the question of specifying  $a_0$  through  $a_J$  to the problem of choosing  $c_0$  through  $c_J$ . One might imagine that  $c_0 = c_1 = \dots = c_J = 1$  would be a natural default choice and least vulnerable to criticism for appearing ad-hoc, but whether such a choice would allow the criterion to serve its intended purpose is unclear.

The considerations in the preceding paragraph are, of course, predicated on the belief that  $\mu(x)$  is  $(J+1)$  - times differentiable. A still greater challenge remains in formulating a criterion by which optimality may be defined when  $\mu(x)$  is infinitely differentiable and all of its derivatives are to be estimated simultaneously. Such a criterion is motivated by the recognition that, although one may not envisage practical applications in which estimates of all derivatives are required, there may exist practical applications in which the number of derivative estimates required is not known a priori. For example, in the Raman spectrum application, a researcher may first examine an estimate of  $\mu(x)$ . If the estimate of  $\mu(x)$  reveals the chemical composition of the material, then the researcher may stop. Otherwise, the researcher may examine an estimate of  $\mu'(x)$ . This process may continue, with the researcher subsequently examining estimates of  $\mu''(x)$  and higher order derivatives, until the researcher either knows the chemical composition

of the material or regards the higher order derivative estimates as so noisy that he/she is simply forced to make a guess about the chemical composition. Charnigo, Hall and Srinivasan [4] provide an example in which an estimate of  $\mu'(x)$  leads to successful identification of a sample of cerium bastnasite. In general, then, the number of derivative estimates to be examined may not be known a priori. Unfortunately, since most nonparametric regression techniques make no provision for the estimation of infinitely many derivatives, there is little theory to inform the construction of a criterion by which optimality may be defined when  $\mu(x)$  is infinitely differentiable and all of its derivatives are to be estimated simultaneously. We thus conclude the present editorial by calling for additional research on the simultaneous estimation of a mean response function and its derivatives when the mean response function is infinitely differentiable.

### Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. DMS-0706857. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

### References

1. Hardle W (1990) Applied nonparametric regression. Cambridge University Press, Cambridge.
2. Wahba G (1990) Spline models for observational data. SIAM, Philadelphia.
3. Loader C (1999) Local regression and likelihood. Springer-Verlag, New York.
4. Charnigo R, Hall B and Srinivasan C (2011) A generalized Cp criterion for derivative estimation. Technometrics, tentatively accepted.
5. Stone CJ (1980) Optimal rates of convergence for nonparametric estimators. Annals of Statistics 8: 1348–1360.
6. Muller HG, Stadtmuller U, and Schmitt T (1987) Bandwidth choice and confidence intervals for derivatives of noisy data. Biometrika 74: 743–749.
7. Fan J, Gijbels I (1995) Data-driven bandwidth selection in local polynomial fitting: Variable bandwidth and spatial adaptation. J. R. Statist. Soc. B 57: 371–394.
8. Charnigo R, Srinivasan C (2011) Self-consistent estimation of mean response functions and their derivatives. Canadian Journal of Statistics, in press.
9. Ramsay JO, Silverman BW (2002) Applied functional data analysis: Methods and case studies. Springer-Verlag, New York.
10. Ramsay JO, Silverman BW (1997) Functional data analysis. Springer-Verlag, New York.
11. Charnigo R, Francoeur M, Kenkel P, Menguc MP, Hall B et al. (2011) Estimating quantitative features of nanoparticles using multiple derivatives of scattering profiles. J Quant Spectrosc Radiat Transf. 112: 1369–1382.
12. Matousek P, Parker AW (2006) Bulk Raman analysis of pharmaceutical tablets. Applied Spectroscopy 60: 1353–1357.
13. Matousek P, Stone N (2007) Prospects for the diagnosis of breast cancer by noninvasive probing of calcifications using transmission Raman spectroscopy. J Biomed Opt 12: 024008.