Statistica Sinica **28** (2018), 2733-2748
doi:https://doi.org/10.5705/ss.202017.0056

DATA SHARPENING IN LOCAL REGRESSION GUIDED BY GLOBAL CONSTRAINT

W. John Braun, X. Joan Hu and Xiuli Kang

University of British Columbia, Simon Fraser University and University of Kansas

Abstract: Data sharpening for kernel regression and density estimation was introduced by the late Peter Hall. We review briefly his enormous contribution to the literature in this area and then propose a data sharpening procedure arising from imposition of a soft global functional constraint in local regression analysis. Instead of enforcing the constraint everywhere, the procedure *guides* the data in directions which enable satisfaction or near-satisfaction of the given property globally through the use of a penalty. It results in a modified local regression estimator which possesses a closed functional form and which includes a conventional local regression estimator as a special case. The approach can accommodate various constraints, most of which in practice are motivated by expert prior knowledge. We demonstrate theoretically and numerically that the proposed estimator is an improved variant of the corresponding local regression estimator. It achieves a reduction in variance while maintaining the bias at the same level. Although the focus in the paper is on local polynomial regression, the technique can be applied, in principle, to any linear nonparametric estimator, including regression splines, smoothing and penalized splines and other recently proposed kernel estimators. We exhibit usefulness of the proposed approach with an analysis of a collection of temperatures at the airport of Vancouver. The analysis reveals a possible monotonic trend underlying the conventional supposition of a periodic (seasonal) temporal structure.

Key words and phrases: Bias-variance trade-off, functional constraint, kernel smoothing, quadratic penalty.

1. Introduction

Local regression has benefited different fields (Wand and Jones (1995); Fan and Gijbels (1996); Loader (1999)), and its popularity will likely continue to grow because of the relative ease with which it can be applied. Employing local fitting by conventional parametric regression procedures, the approach confers a large degree of robustness to functional misspecification of the systematic component. However, imposing a global constraint on the resulting curve is not always straightforward. This may explain why most published research on nonparametric/semiparametric regression subject to a global (or shape) constraint is usually based on other approaches, such as smoothing splines (Ramsay and Silverman (2005)) or nonparametric regression with Bernstein polynomials (Wang and Ghosh (2012)). Notable exceptions include the weighted kernel estimator (otherwise, known as "tilting") proposed by Hall and Huang (2001) and its extension provided in Du, Parmeter and Racine (2013), where the weights are chosen according to the primary constraint and a few other requirements. Their weights are determined implicitly by certain equations as opposed to closed-form expressions, resulting in the loss of computational simplicity enjoyed by local regression. This motivated our research presented in this paper. We propose an alternative approach to imposing a global constraint on a local regression estimator via a penalized data sharpening procedure.

Data sharpening perturbs observations prior to application of a conventional estimator; the goal is to achieve improved performance of the estimator in terms of some criterion, while retaining most or all of the attractive characteristics of the original estimator. The contributions of the late Peter Hall to the relatively new field of data sharpening are extensive. (Cheng and Fan (2016), Sect. 6) reviews Hall's contributions to the general field of shape-constrained nonparametric estimation. Here, we review briefly his specific contributions to data sharpening.

As far as we can tell, Hall's first contribution appeared in the context of density estimation (Choi and Hall (1999)) where the mean-shift clustering algorithm was cleverly exploited to move points towards local modes in order to reduce bias. Hall and Minnotte (2002) describe a different way to sharpen the data, exploiting earlier transformation approaches of Samiuddin and El-Sayyad (1990). Claeskens and Hall (2002) extend the approaches of both of these papers to the context of hazard estimation, showing among other things that the same data perturbation can be applied to hazard estimation as to density estimation. Choi and Hall (2001) apply data sharpening to nonparametric point process intensity estimation with specific attention to earthquake data. Hall and Kang (2005) provide theoretical support for the constrained sharpening approach to density estimation described in Braun and Hall (2001), which proposed data sharpening to satisfy shape-constraints in both density estimation and nonparametric regression. Choi, Hall and Rousson (2000) presents a simple, but effective, method of sharpening data to reduce bias in nonparametric regression. Doosti and Hall (2016) demonstrate the benefits to be accrued when sharpening and tilting are applied in combination yielding improved "perturbed" density estimates, both qualitatively and in terms of accuracy; theoretical results supplied there show that uniform consistency is attained for a large class of densities.

The data sharpening procedure proposed in the present paper amounts to translating the constraint into an L_2 type of penalty on the data points, perturbing them so that they almost satisfy or completely satisfy the given property, upon application of the local regression estimator. Thus, the estimator is subject to a soft constraint. The proposed penalized data sharpening approach has common features with penalized splines (Eilers and Marx (1996)). The procedure can easily accommodate various functional properties which arise in practice to represent certain qualitative characteristics, including those that can only be handled by the original data sharpening method with a nonlinear programming algorithm. The resulting estimator can be presented analytically as a function of a tuning parameter, explicitly including the conventional local regression estimator as a special case. The analytic form allows the estimator to be easily computable. In fact, it permits exploration of theoretical properties of the proposed estimator and guides selection of the tuning parameter, conditional on the design points. It also leads to insights into the effects of imposing the constraints globally. We also point out that the technique is not only applicable to classical local polynomial regression estimation but also to recently developed extensions such as the double-smoothing estimator of He and Huang (2009), and the approach could be applied to spline smoothers as well although, in that context, other functional data analysis techniques may be preferable.

We organize this paper as follows. After introducing the framework, Section 2 describes the proposed approach to data sharpening in local regression via a quadratic penalty. Some candidate local regression estimators and quadratic penalities are surveyed. We compare the new estimator to its local regression counterpart theoretically in Section 3. We present a systematic procedure for selecting the tuning parameter based on the mean integrated squared error (MISE) of the estimator in Section 4. Section 5 reports two numerical studies, a simulation study to investigate finite sample performance of the proposed procedure and a data analysis to exemplify its applications to studies exploring possible connections to global warming. A few final remarks are given in Section 6.

2. Data Sharpening Subject to Quadratic Penalty

Consider the regression model $Y = g(X)+\epsilon$ with $\mathrm{E}(\epsilon\big|X) = 0$ and $\mathrm{Var}(\epsilon\big|X) = \sigma^2$. Suppose that there is a collection of i.i.d. observations on (Y, X): $\{(y_i, x_i) :$

$i = 1, \ldots, n\}$. Our primary objective is to estimate the function $g(\cdot)$ without (fully) specifying its functional form.

Let $\boldsymbol{x} = (x_1, \ldots, x_n)^\top$ and $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$, and a local estimator of $g(\cdot)$ with the data be $\widetilde{g}(z) = \sum_{i=1}^n a_i(z;h) y_i = \boldsymbol{a}(z;h)^\top \boldsymbol{y}$, where the bandwidth h is constant and $\boldsymbol{a}(z;h)$ is the n-dim column vector with components $a_i(z;h)$. We aim to achieve an improved estimator $\widehat{g}(z)$ using the construction of $\widetilde{g}(z)$ with $\boldsymbol{y}$ replaced by *sharpened* response observations $\boldsymbol{y}^\star = (y_1^\star, \ldots, y_n^\star)^\top$: $\widehat{g}(z) = \boldsymbol{a}(z;h)^\top \boldsymbol{y}^\star$. Rooted in the original response observations $\boldsymbol{y}$, the sharpened response observations $\boldsymbol{y}^\star$ result from imposing a penalty on violation of a global constraint, to encourage certain qualitative characteristics based on possible prior knowledge about $g(\cdot)$.

2.1. Proposed estimator

Assume that the constraint (based on prior knowledge) is in the form of a functional equation of $g(\cdot)$: $\big[b \circ g\big](z) = 0$ for z over an interval $\mathscr{Z}$, where b is a linear operator. Let the components of $\boldsymbol{z} = (z_1, \ldots, z_m)^\top$ be the chosen grid points in $\mathscr{Z}$. We encourage the constraint with $g(\cdot)$ by controlling the length of the vector $b \circ \widehat{\boldsymbol{g}}_{\boldsymbol{z}} = (\big[b \circ \widehat{g}\big](z_1), \ldots, \big[b \circ \widehat{g}\big](z_m))^\top = \mathbf{B}(\boldsymbol{z};h)^\top \boldsymbol{y}^\star$, where $\mathbf{B}(\boldsymbol{z};h)$ is the $n \times m$ matrix with the jth column $\boldsymbol{b}(z_j;h) = \big[b \circ \boldsymbol{a}\big](z;h)|_{z=z_j}$. This yields a proposed objective function adjoining the constraint as a quadratic penalty to the squared Euclidean distance between $\boldsymbol{y}$ and its sharpened version $\boldsymbol{y}^\star$. Specifically, we choose the $\boldsymbol{y}^\star$ to minimize

$$O(\boldsymbol{y}^\star; \lambda, \mathbf{B}, \boldsymbol{y}) = (\boldsymbol{y} - \boldsymbol{y}^\star)^\top (\boldsymbol{y} - \boldsymbol{y}^\star) + \lambda \boldsymbol{y}^{\star\top} \mathbf{B}(\boldsymbol{z};h) \mathbf{B}(\boldsymbol{z};h)^\top \boldsymbol{y}^\star \tag{2.1}$$

with a fixed tuning parameter $\lambda > 0$.

Let $\mathbf{I}$ be the identity matrix with an appropriate order depending on the context. Here $\mathbf{B}(\boldsymbol{z};h)\mathbf{B}(\boldsymbol{z};h)^\top$ is positive semi-definite, and thus $\mathbf{I}+\lambda\mathbf{B}(\boldsymbol{z};h)\mathbf{B}(\boldsymbol{z};h)^\top$ is positive definite when $\lambda \geq 0$. The objective function (2.1) achieves its minimum at the unique point $\boldsymbol{y}^\star = \operatorname{argmin}_{all\ \widetilde{\boldsymbol{y}^\star}} O(\widetilde{\boldsymbol{y}^\star}; \lambda, \mathbf{B}, \boldsymbol{y}) = \big\{\mathbf{I}+\lambda\mathbf{B}(\boldsymbol{z};h)\mathbf{B}(\boldsymbol{z};h)^\top\big\}^{-1}\boldsymbol{y}$. This yields the new estimator of $g(\cdot)$, a penalized local regression estimator,

$$\widehat{g}(z) = \boldsymbol{a}(z;h)^\top \boldsymbol{y}^\star = \boldsymbol{a}(z;h)^\top \big\{\mathbf{I} + \lambda\mathbf{B}(\boldsymbol{z};h)\mathbf{B}(\boldsymbol{z};h)^\top\big\}^{-1}\boldsymbol{y}. \tag{2.2}$$

The new estimator reduces to the local regression estimator $\widetilde{g}(z) = \boldsymbol{a}(z;h)^\top\boldsymbol{y}$ when the tuning parameter $\lambda = 0$. Denote $\big\{\mathbf{I} + \lambda\mathbf{B}(\boldsymbol{z};h)\mathbf{B}(\boldsymbol{z};h)^\top\big\}^{-1}\boldsymbol{a}(z;h)$ by $\boldsymbol{a}^\star(z;h,\lambda)$. Thus $\widehat{g}(z) = \boldsymbol{a}^\star(z;h,\lambda)^\top\boldsymbol{y}$ may also be viewed as a weighted kernel estimator when the matrix $\big\{\mathbf{I} + \lambda\mathbf{B}(\boldsymbol{z};h)\mathbf{B}(\boldsymbol{z};h)^\top\big\}^{-1}$ is diagonal.

The proposed estimation procedure does not strictly enforce the constraint. Instead, it guides the local regression estimator $\widetilde{g}(z)$ through a shape-related penalty to $\widehat{g}(z)$, to better approximate the qualitative feature. We verify this notion theoretically and numerically in the rest of this paper.

2.2. Candidates for the base estimator and functional constraint

Possible candidates for the unsharpened estimator $\widetilde{g}(z)$ to which our proposed penalty method can be applied require the linearity property in the response vector $\boldsymbol{y}$. This is clearly true of spline regression estimators. We demonstrate the property in *Local polynomial regression (LPR)* (Wand and Jones (1995); Fan and Gijbels (1996); Loader (1999)) and *Double-Smoothing* (He and Huang (2009)) with kernel methods in Section S1.1 of the Supplementary Material.

The desired functional constraint on $g(\cdot)$ in many applications may be formulated as a linear transform-based equation, such as the constant coefficient linear homogeneous differential equation $\big[b \circ g\big](z) = 0$, with $b = \sum_{l=L_\star}^{L^\star} \alpha_l D^l$, where α_l are constants, D is the differential operator, and $0 < L_\star < L^\star < \infty$. This class has been employed in the literature (Heckman and Ramsay (2000)). It includes many desirable functional constraints such as the *roughness penality* with $\big[b \circ g\big](z) = D^2 g(z) = 0$ and the *periodicity constraint* with the functional constraint $\big[b \circ g\big](z) = D^4 g(z) + \gamma D^2 g(z) = 0$, which encourages the resulting estimator to pick up both linear and periodic trends. More discussion is provided in Section S1.1 of the Supplementary Material.

3. Theoretical Verification

The sharpened estimator $\widehat{g}(z) = \boldsymbol{a}(z;h)^\intercal \boldsymbol{y}^\star = \boldsymbol{a}^\star(z;h,\lambda)^\intercal \boldsymbol{y}$ depends on the response observations linearly as does the conventional local regression estimator $\widetilde{g}(z) = \boldsymbol{a}(z;h)^\intercal \boldsymbol{y}$, the unsharpened one. In the following, we derive theoretical properties of $\widehat{g}(\cdot)$ based on its analytic form (2.2) and its connection to $\widetilde{g}(\cdot)$. In the rest of this paper, $\mathbf{1}$ and $\mathbf{0}$ are the vectors $(1,\ldots,1)^\intercal$ and $(0,\ldots,0)^\intercal$ with appropriate dimensions, respectively, and $(\boldsymbol{x} - z\mathbf{1})^l$ is the n-dim vector with the components $(x_i - z)$ to the power l.

By a general result for the inverse of a sum of matrices (Henderson and Searle (1981)), we have $\big\{\mathbf{I} + \lambda \mathbf{B}(\boldsymbol{z};h)\mathbf{B}(\boldsymbol{z};h)^\intercal\big\}^{-1} = \mathbf{I} - \mathbf{B}(\boldsymbol{z};h)\big\{\mathbf{I}/\lambda + \mathbf{B}(\boldsymbol{z};h)^\intercal \mathbf{B}(\boldsymbol{z};h)\big\}^{-1}$ $\mathbf{B}(\boldsymbol{z};h)^\intercal$. This yields that

$$\widehat{g}(z) - \widetilde{g}(z) = -\boldsymbol{a}(z;h)^\intercal \mathbf{B}(\boldsymbol{z};h)\left\{\frac{\mathbf{I}}{\lambda} + \mathbf{B}(\boldsymbol{z};h)^\intercal \mathbf{B}(\boldsymbol{z};h)\right\}^{-1} b \circ \widetilde{\boldsymbol{g}}_{\boldsymbol{z}}, \tag{3.1}$$

where $b \circ \widetilde{\boldsymbol{g}}_{\boldsymbol{z}} = \mathbf{B}(\boldsymbol{z};h)^\intercal \boldsymbol{y}$ is the m-dim vector with components $b \circ \boldsymbol{a}(z_j;h)^\intercal \boldsymbol{y} = \big[b \circ \widetilde{g}\big](z_j)$ for $j = 1,\ldots,m$.

The display at (3.1) reveals that, with a fixed value of the tuning parameter λ, the adjustment to achieve $\widehat{g}(\cdot)$ from $\widetilde{g}(\cdot)$ is proportional to $b \circ \widetilde{\boldsymbol{g}}_{\boldsymbol{z}}$, the components of which are the departures of $\widetilde{g}(\cdot)$ at the grid points from the constraint. In the extreme case that $\widetilde{g}(\cdot)$ satisfies the constraint $\big[b \circ \widetilde{g}\big](z) = 0$, and thus $b \circ \widetilde{\boldsymbol{g}}_{\boldsymbol{z}} = \mathbf{0}$, and $\widehat{g}(z)$ is the same as $\widetilde{g}(z)$.

3.1. Conditional variance and bias

Conditional on the design points $\boldsymbol{x} = (x_1, \ldots, x_n)^\intercal$ and with fixed h and λ, the expectation of the sharpened response $\boldsymbol{y}^\star$ is $\{\mathbf{I} + \lambda \mathbf{B}(\boldsymbol{z}; h)\mathbf{B}(\boldsymbol{z}; h)^\intercal\}^{-1}\boldsymbol{g}$ with $\boldsymbol{g} = \big\{g(x_1), \ldots, g(x_n)\big\}^\intercal$. Thus the conditional expectation and conditional variance of $\widehat{g}(z)$ at a fixed z are

$$\mathrm{E}\big(\widehat{g}(z)\big|\boldsymbol{x}, \boldsymbol{z}; h, \lambda\big) = \boldsymbol{a}(z; h)^\intercal \big\{\mathbf{I} + \lambda \mathbf{B}(\boldsymbol{z}; h)\mathbf{B}(\boldsymbol{z}; h)^\intercal\big\}^{-1}\boldsymbol{g} = \boldsymbol{a}^\star(z; h, \lambda)^\intercal \boldsymbol{g}, \quad \text{and} \tag{3.2}$$

$$\begin{aligned}\mathrm{Var}\big(\widehat{g}(z)\big|\boldsymbol{x}, \boldsymbol{z}; h, \lambda\big) &= \sigma^2 \boldsymbol{a}(z; h)^\intercal \big\{\mathbf{I} + \lambda \mathbf{B}(\boldsymbol{z}; h)\mathbf{B}(\boldsymbol{z}; h)^\intercal\big\}^{-2}\boldsymbol{a}(z; h) \\ &= \sigma^2 \boldsymbol{a}^\star(z; h, \lambda)^\intercal \boldsymbol{a}^\star(z; h, \lambda).\end{aligned} \tag{3.3}$$

Recall that $\mathrm{Var}\big(\widetilde{g}(z)\big|\boldsymbol{x}; h\big)\big|_{z=x_i} \leq \sigma^2$ for $i = 1, \ldots, n$ if the kernel function is symmetric and decreasing on $[0, \infty)$ (Loader (1999)). The propositions below establish an improvement of $\widehat{g}(\cdot)$ in terms of variance reduction while its bias remains at the same order as $\widetilde{g}(\cdot)$'s. We outline their proofs in Section S1.2 of the Supplementary Material.

Proposition 1. *Given $\widetilde{g}(z) = \boldsymbol{a}(z; h)^\intercal \boldsymbol{y}$, a local regression estimator with independent observations $\big\{(y_i, x_i) : i = 1, \ldots, n\big\}$ and fixed h, Var$\big(\widehat{g}(z)\big|\boldsymbol{x}, \boldsymbol{z}; h, \lambda\big) \leq$ Var$\big(\widetilde{g}(z)\big|\boldsymbol{x}; h\big)$ with $z \in \mathcal{Z}$ for all $\boldsymbol{z}$, where the equal sign holds only when either $\lambda = 0$ or $\mathbf{B}(\boldsymbol{z}; h)^\intercal \boldsymbol{a}^\star(z; h, \lambda) = \mathbf{0}$.*

Proposition 2. *Suppose $\widetilde{g}(z) = \boldsymbol{a}(z; h)^\intercal \boldsymbol{y}$ in Proposition 1 is the local regression estimator of order q (≥ 0), and that $g(x_i)$ can be expanded in a Taylor series around $z \in \mathcal{Z}$ as $g(x_i) = \sum_{l=0}^{\infty} g^{(l)}(z)(x_i - z)^l/l!$. When the functional constraint is based on a constant coefficient linear homogeneous differential equation, $\big[b \circ g\big](z) = 0$ with $b = \sum_{l=L_\star}^{L^\star} \alpha_l D^l$ with $q < L_\star < L^\star < \infty$, the conditional bias of $\widehat{g}(z)$ is*

$$\boldsymbol{a}(z; h)^\intercal \sum_{l=q+1}^{\infty} \frac{1}{l!}\Big[g^{(l)}(z)(\boldsymbol{x} - z\mathbf{1})^l - \lambda \mathbf{B}(\boldsymbol{z}; h)\big\{\mathbf{I} + \lambda \mathbf{B}(\boldsymbol{z}; h)^\intercal \mathbf{B}(\boldsymbol{z}; h)\big\}^{-1}\widetilde{\boldsymbol{g}}_l(\boldsymbol{x}, \boldsymbol{z}; h)\Big], \tag{3.4}$$

where $\widetilde{\boldsymbol{g}}_l(\boldsymbol{x}, \boldsymbol{z}; h)$ is the m-dim vector with the kth component $g^{(l)}(z_k)\boldsymbol{b}(z_k; h)^\intercal(\boldsymbol{x} - z_k\mathbf{1})^l$.

Proposition 2 shows that, when the constraint is specified as above, the bias of the proposed estimator depends only on the $(q+1)$th or higher order derivatives of $g(\cdot)$, the same order as without penalty. Therefore, it is unbiased when $g(\cdot)$ is polynomial of order q or less. It is easy to see from (3.4) that the difference between the two biases in general reduces to zero with $\lambda = 0$ and converges to $-\boldsymbol{a}(z;h)^\intercal \mathbf{B}(\boldsymbol{z};h)\big\{\mathbf{B}(\boldsymbol{z};h)^\intercal \mathbf{B}(\boldsymbol{z};h)\big\}^{-1}\big\{\sum_{l=q+1}^{\infty}\widetilde{\boldsymbol{g}}_l(\boldsymbol{x},\boldsymbol{z};h)/l!\big\}$ when $\lambda \to \infty$.

3.2. Sum of squared residuals

The variance σ^2 can be estimated using the normalized sum of residual squares with a local regression estimator $\widetilde{g}(z) = \boldsymbol{a}(z;h)^\intercal \boldsymbol{y}$ based on the i.i.d. observations $\big\{(y_i, x_i) : i = 1, \dots\big\}$: $\widetilde{\sigma}^2 = \sum_{i=1}^n \{y_i - \widetilde{g}(x_i)\}^2/(n - 2\nu_1 + \nu_2)$ with $\nu_1 = \text{tr}\{\mathbf{A}(\boldsymbol{x};h)\}$ and $\nu_2 = \text{tr}\big\{\mathbf{A}(\boldsymbol{x};h)\mathbf{A}(\boldsymbol{x};h)^\intercal\big\}$, where ν_1 and ν_2 are two most commonly used generalizations of the degrees of freedom in local regression Loader (1999). A small bandwidth is often used to obtain a local regression estimator with small bias and then an approximately unbiased variance estimator. The sharpened estimator $\widehat{g}(z) = \boldsymbol{a}(z;h)^\intercal \boldsymbol{y}^\star = \boldsymbol{a}^\star(z;h,\lambda)^\intercal \boldsymbol{y}$ has a property analogous to that of the local regression estimator.

Proposition 3. *The expectation of the sum of squared residuals is*

$$E\left[\sum_{i=1}^{n}\big\{Y_i-\widehat{g}(x_i)\big\}^2\Big|\boldsymbol{x};h,\lambda\right] = \sigma^2(n-2\nu_1^\star+\nu_2^\star)+\sum_{i=1}^{n} Bias^2\big\{\widehat{g}(x_i)\big|\boldsymbol{x};h,\lambda\big\}, \quad (3.5)$$

where $\nu_1^\star = tr\{\mathbf{A}^\star(\boldsymbol{x},\boldsymbol{z};h,\lambda)\}$ *and* $\nu_2^\star = tr\big\{\mathbf{A}^\star(\boldsymbol{x},\boldsymbol{z};h,\lambda)\mathbf{A}^\star(\boldsymbol{x},\boldsymbol{z};h,\lambda)^\intercal\big\}$.

A proof for the proposition is outlined in Section S1.2 of the Supplementary Material. The proposition indicates that the normalized sum of residual squares with the sharpened estimator $\widehat{g}(z)$ can also estimate the variance σ^2:

$$\widehat{\sigma}^2 = \frac{1}{n - 2\nu_1^\star + \nu_2^\star}\sum_{i=1}^{n}\{y_i - \widehat{g}(x_i)\}^2. \quad (3.6)$$

The variance estimator $\widehat{\sigma}^2$ is approximately unbiased if the bias of $\widehat{g}(\cdot)$ is small in terms of $1/(n - 2\nu_1^\star + \nu_2^\star)\sum_{i=1}^n \text{Bias}^2\big\{\widehat{g}(x_i)\big|\boldsymbol{x};h,\lambda\big\}$ being close to zero.

4. Selection of Tuning Parameter λ

When evaluating a realization of $\widehat{g}(\cdot)$, the Asymptotic Integrated Squared Error, $\text{AISE}(\widehat{g}) = \sum_{j=1}^m \{\widehat{g}(z_j) - g(z_j)\}^2/m$, is often used to approximate the integrated square error $\int_{-\infty}^{\infty}\{\widehat{g}(z) - g(z)\}^2 p(z)dz$ if the grid points $z_1, \dots, z_m$ are generated from a distribution $p(\cdot)$. We consider evaluation of the estimator $\widehat{g}(\cdot)$'s overall performance using the approximation to the conditional mean integrated squared error (MISE), which is the conditional expectation of AISE($\widehat{g}$):

$$\text{MAISE}_{\boldsymbol{z}}(\widehat{g}|\boldsymbol{x};h,\lambda)=\frac{1}{m}\sum_{j=1}^{m}\text{Var}\big\{\widehat{g}(z_j)\big|\boldsymbol{x};h,\lambda\big\}+\frac{1}{m}\sum_{j=1}^{m}\text{Bias}^2\big\{\widehat{g}(z_j)\big|\boldsymbol{x};h,\lambda\big\} \quad (4.1)$$

Plugging (3.2) and (3.3) in (4.1) gives $\text{MAISE}_{\boldsymbol{z}}(\widehat{g}|\boldsymbol{x};h,\lambda)$ as the sum of two terms:

$$\frac{\sigma^2}{m}\text{tr}\big[\mathbf{A}(\boldsymbol{z};h)^{\intercal}\big\{\mathbf{I}+\lambda\mathbf{B}(\boldsymbol{z};h)\mathbf{B}(\boldsymbol{z};h)^{\intercal}\big\}^{-2}\mathbf{A}(\boldsymbol{z};h)\big],$$

and

$$\frac{1}{m}\big\|\mathbf{A}(\boldsymbol{z};h)^{\intercal}\big\{\mathbf{I}+\lambda\mathbf{B}(\boldsymbol{z};h)\mathbf{B}(\boldsymbol{z};h)^{\intercal}\big\}^{-1}\boldsymbol{g}-\boldsymbol{g}_{\boldsymbol{z}}\big\|^2,$$

where $||\cdot||$ is the l^2-norm, $\mathbf{A}(\boldsymbol{z};h)$ is the $n\times m$ matrix with the jth column $\boldsymbol{a}(z_j;h)$, $\boldsymbol{g}_{\boldsymbol{z}}$ is the m-dim vector with components $g(z_j)$, and $\text{tr}(\cdot)$ is the trace operator. We employ this formula for MAISE to develop algorithms for selecting the tuning parameter λ after examining it in two extreme scenarios.

4.1. Two extreme cases

First $\boldsymbol{y}^{\star}=\boldsymbol{y}$ in case $\lambda=0$. That is, no sharpening takes place. Then

$$\text{MAISE}_{\boldsymbol{z}}(\widehat{g}|\boldsymbol{x};h,\lambda=0)=\frac{1}{m}\Big[\sigma^2\text{tr}\big\{\mathbf{A}(\boldsymbol{z};h)^{\intercal}\mathbf{A}(\boldsymbol{z};h)\big\}+\big\|\mathbf{A}(\boldsymbol{z};h)^{\intercal}\boldsymbol{g}-\boldsymbol{g}_{\boldsymbol{z}}\big\|^2\Big]$$

is in fact $\text{MAISE}_{\boldsymbol{z}}(\widetilde{g}|\boldsymbol{x})$, the corresponding approximate to MISE of the conventional estimator $\widetilde{g}(\cdot)$. If λ is determined independently from $\mathbf{A}(\boldsymbol{z};h)$ and $\mathbf{B}(\boldsymbol{z};h)$, as $\lambda\to\infty$, $\text{MAISE}_{\boldsymbol{z}}(\widehat{g}|\boldsymbol{x};h,\lambda)\to||\boldsymbol{g}_{\boldsymbol{z}}||^2$ and $\mathbf{B}(\boldsymbol{z};h)^{\intercal}\boldsymbol{y}^{\star}=\mathbf{B}^{\intercal}\boldsymbol{y}-\mathbf{B}^{\intercal}\mathbf{B}(\mathbf{I}/\lambda+\mathbf{B}^{\intercal}\mathbf{B})^{-1}\mathbf{B}^{\intercal}\boldsymbol{y}$ converges to $\mathbf{0}$. Thus, when using the roughness penalty, for example, the penalty imposes a linear restriction and results in a reduction in the degrees of freedom of $g(\cdot)$'s estimator to the order 1.

Between these two extremes should lie a value of λ that leads to an estimator as an improved version of $\widetilde{g}(\cdot)$ with respect to the global constraint. It is desirable to choose the tuning parameter $\lambda>0$ according to $\mathbf{A}(\boldsymbol{z};h)$ and $\mathbf{B}(\boldsymbol{z};h)$ to achieve a meaningful $\boldsymbol{y}^{\star}$, a sharpened $\boldsymbol{y}$, and thus to yield a welcome new estimator $\widehat{g}(\cdot)$. This motivates our procedure for determining the tuning parameter λ with a fixed bandwidth h.

4.2. Procedure for determining λ

For a given $\widetilde{g}(\cdot)$ and a predetermined constraint $\big[b\circ g\big](z)=0$, we aim to select the tuning parameter as $\lambda^{\star}=\text{argmin}_{all\ \lambda\geq 0}\text{MAISE}_{\boldsymbol{z}}(\widehat{g}\big|\boldsymbol{x};h,\lambda)$ with fixed h and $\boldsymbol{z}$ based on the data. Theoretically speaking, the choice of $\lambda=\lambda^{\star}$ secures the resulting $\widehat{g}(\cdot)$ a performance better than or at least the same as the conventional estimator $\widetilde{g}(\cdot)$ in the sense of having small conditional MAISE. An algorithm to determine $\lambda^{\star}$ follows.

Algorithm A. Provided an estimate $\widetilde{\sigma}^2$,

Step A.1. Calculate $\mathbf{A}(\boldsymbol{z}; h)$ and $\mathbf{B}(\boldsymbol{z}; h)$.

Step A.2. Plug in $\mathbf{A}(\boldsymbol{z}; h), \mathbf{B}(\boldsymbol{z}; h)$ and the estimate $\widetilde{\sigma}^2$ of σ^2 into (4.1), and substitute $g(\cdot)$ by $\widetilde{g}(\cdot)$ to obtain $\text{MAISE}_{\boldsymbol{z}}(\widehat{g}\big|\boldsymbol{x}; h, \lambda)$ as a function of λ.

Step A.3. Compute $\lambda^{\star} = \text{argmin}_{all\ \lambda \geq 0}\text{MAISE}_{\boldsymbol{z}}(\widehat{g}\big|\boldsymbol{x}; h, \lambda)$.

We list a few remarks on the implementation of the algorithm.

Remark 1. The magnitudes of the two terms in $O(\boldsymbol{y}^{\star}; \lambda, \mathbf{B}, \boldsymbol{y})$ of (2.1) can be rather different in some applications. When that is of concern, we suggest replacing λ in the second term of (2.1) by $\lambda\eta_{ratio}$ with η_{ratio} fixed at the ratio $\sum(y_i - \bar{y})^2 \big/ \boldsymbol{y}^{\intercal}\mathbf{B}(\boldsymbol{z}; h)\mathbf{B}(\boldsymbol{z}; h)^{\intercal}\boldsymbol{y}$. This can narrow the search interval for $\lambda^{\star}$.

Remark 2. When implementing the algorithm, the bandwidth h may be determined with a standard bandwidth selection in local regression such as the direct plug-in method to select the bandwidth of a local kernel regression estimate and a cross-validation selection (Fan and Gijbels (1996)).

Remark 3. The normalized residual sum of squares, the ratio of $\sum_{i=1}^{n}\{y_i - \widetilde{g}(x_i)\}^2$ to $(n - 2\nu_1 + \nu_2)$ with $\nu_1 = \text{tr}\{\mathbf{A}(\boldsymbol{x}; h)\}$ and $\nu_2 = \text{tr}\{\mathbf{A}(\boldsymbol{x}; h)^{\intercal}\mathbf{A}(\boldsymbol{x}; h)\}$, may be used as $\widetilde{\sigma}^2$ (Loader (1999)).

Remark 4. An alternative procedure is to adapt a cross-validation type of procedure to determine h and λ in the proposed approach. For example, consider an extension of the classical cross-validation criterion in local regression: $CV(h, \lambda) = (1/n)\sum_{i=1}^{n}\widehat{e}_{i,-i}^{2}$, where $\widehat{e}_{i,-i} = y_i - \widehat{g}_{-i}(x_i)$ and $\widehat{g}_{-i}(\cdot)$ is the proposed estimator using the available data with (x_i, y_i) excluded.

In Section S2 of the Supplementary Material, we present an iterative algorithm, a naturally refined version of *Algorithm A*. Compared to *Algorithm A*, this algorithm may provide a better selection of λ, but can be computationally more intensive.

5. Numerical Performance

5.1. Simulation

We conducted a simulation study to assess the proposed approach numerically with the software package *R* (R Development Core Team (2015)). The simulations considered the mean function $g(x) = 6x + 3\sin(4\pi x) + 5\cos(4\pi x)$ for $x \in [0,1]$, which has two cycles with period 1/2. We generated observations (y_i, x_i) for $i = 1, \ldots, n$, independently by $y_i = g(x_i) + \epsilon_i$ with $\epsilon_i \sim N(0, \sigma^2)$ and design points x_i from the uniform distribution $U(0,1)$. Grid points were taken as the m equally spaced points over $(0,1)$: $z_j : j = 1, \ldots, m$. The conventional estimator $\widetilde{g}(\cdot)$ was specified as either the local constant or local linear estimator, denoted by $\widetilde{g}_{LC;h}(\cdot)$ or $\widetilde{g}_{LL;h}(\cdot)$ with bandwidth h, respectively.

Penalized local constant and local linear estimators, denoted by $\widehat{g}_{LC;h,\lambda}(\cdot)$ and $\widehat{g}_{LL;h,\lambda}(\cdot)$ with bandwidth h and tuning parameter λ, were evaluated with generated data together with their conventional counterparts. We determined the bandwidth h of $\widetilde{g}_{LC;h}(\cdot)$ and $\widetilde{g}_{LL;h}(\cdot)$ by the plug-in method, cross-validation, and generalized cross-validation at selected simulation settings. The resulting penalized regression estimates appeared not to differ much from each other. We thus focused on using the Gaussian kernel and choosing h by the plug-in method in the simulation through the *dpill* function in the R library *KernSmooth* (Wand (2015)). The global constraint was set as $b = D^4 + (4\pi)^2 D^2$, to allow the true mean function to be a solution of $\big[b \circ g\big](x) = 0$.

We considered settings with $n = 50$ or 100, $m = 50$ or 100, and varied $\sigma = 0.3$, 1, 2, or 3 to generate random errors with small to large (but constant) variance. In each of the simulation settings, the two local regression estimators $\widetilde{g}_{LC;h}(\cdot)$ and $\widetilde{g}_{LL;h}(\cdot)$ and their penalized variations $\widehat{g}_{LC;h,\lambda}(\cdot)$ and $\widehat{g}_{LL;h,\lambda}(\cdot)$ were evaluated with simulated observations $\big\{(y_i, x_i) : i = 1, \ldots, n\big\}$, where the bandwidth was determined as h_0 by the *R*-function *dpill*(). The tuning parameter λ was chosen as (1) $\lambda_0 = \eta_{ratio}$ defined in Remark 4.1, (2) λ^* to minimize the AISE $\sum_{j=1} \big\{\widehat{g}_{h,\lambda}(z_j) - g(z_j)\big\}^2$ with $h = h_0$, or (3) $\lambda^{**} = \lambda^\star$ determined by Algorithm A. The selection of λ^* in (2) requires the true function $g(\cdot)$. It is not practical but it provides the best possible choice of the tuning parameter λ. We used it as a reference to assess the convenient choice in (1) and the selection of λ^{**} by Algorithm A in (3). The AISE values of the resulting estimates were calculated to measure the departure of the estimates from the true $g(\cdot)$. For illustration, we present in Section S3.1 of the Supplementary Material the true mean function and a set of generated observations ($n = 100$, $m = 50$, $\sigma = 2.0$) together with the evaluations of the proposed penalized local constant/linear estimators $\widehat{g}_{LC;h,\lambda}(\cdot)$ and $\widehat{g}_{LL;h,\lambda}(\cdot)$ together with their conventional counterparts $\widetilde{g}_{LC;h}(\cdot)$ and $\widetilde{g}_{LL;h}(\cdot)$.

The discussions below are based on 100 repetitions in each simulation setting. Our findings from the simulation outcomes are rather similar with different combinations of (n, m). The following focuses on the outcomes from the simu-

Table 1. Summary statistics of the Approximate Integrated Squared Errors (AISE) in simulation: *associated with the local linear (LL) estimator.*

	Estimator		$\widehat{g}_{LL;h,\lambda}(\cdot)$		
		$\widetilde{g}_{LL;h}(\cdot)^a$	$(1)^b\ \lambda = \lambda_0$	$(2)^c\ \lambda = \lambda^*$	$(3)^d\ \lambda = \lambda^{**}$
		Case A. $n = 50$, $m = 50$			
$\sigma = .3$	AISE^e_{SM}	0.026	0.015	0.009	0.014
	AISE^f_{SD}	(0.009)	(0.008)	(0.006)	(0.006)
$\sigma = 1$	AISE_{SM}	0.267	0.178	0.103	0.161
	AISE_{SD}	(0.103)	(0.090)	(0.069)	(0.070)
$\sigma = 3$	AISE_{SM}	2.191	1.687	1.004	1.477
	AISE_{SD}	(0.860)	(0.795)	(0.573)	(0.568)
		Case B. $n = 50$, $m = 100$			
$\sigma = 1$	AISE_{SM}	0.186	0.120	0.070	0.112
	AISE_{SD}	(0.062)	(0.047)	(0.033)	(0.035)
		Case C. $n = 100$, $m = 50$			
$\sigma = 1$	AISE_{SM}	0.272	0.181	0.105	0.166
	AISE_{SD}	(0.097)	(0.085)	(0.063)	(0.063)

[a] $h = h_0$ determined by *R*-function *dpill*().
[b] (1) $\lambda_0 = \lambda_{coef}$ as defined in Remark 4.1.
[c] (2) $\lambda^* = \text{argmin}_\lambda MAISE(\widehat{g}_\lambda)$ in §4.2.
[d] (3) λ^{**} determined by Algorithm A in §4.2.
[e,f] AISE_{SM}, AISE_{SD} is the sample mean, sample standard deviation of the evaluations of the approximate integrated squared error.

lation Case A. $n = 50$, $m = 50$, and exemplifies Case B. $n = 50$, $m = 100$ and Case C. $n = 100$, $m = 50$ using the settings with $\sigma = 1.0$. Table 1 presents the sample means and sample standard deviations of the AISE evaluations associated with the procedures based on the local linear (LL) estimator in the different simulation settings. A summary of the simulation outcomes associated with the procedures based on the local constant (LC) estimator is given in Section S3.1 of the Supplementary Material.

The smaller AISE values are in general associated with the estimates using the proposed approach compared to their corresponding conventional local regression estimates. The AISE values of the proposed estimates with the tuning parameter determined by Algorithm A are smaller than the ones with the tuning parameter chosen by simply standardizing the two terms in the objective function (2.1). The similarity in the sample standard derivation of the different variations of the sharpened estimator to their corresponding local estimator in each simulation setting indicates that improvement is rather stable. The sharpened estimates with the convenient choice $\lambda = \lambda_0$ of (1) reduce the AISEs of their

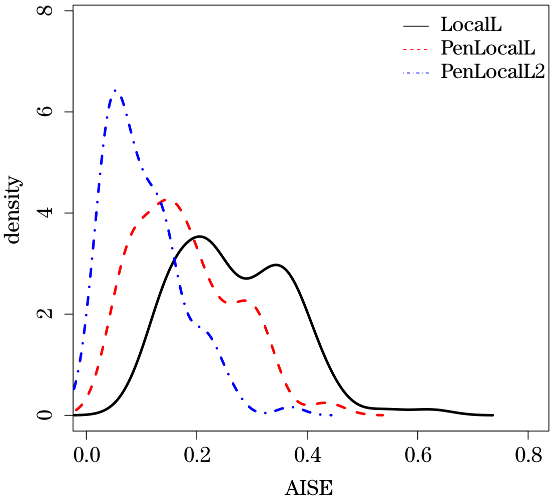

Figure 1. Density curves of the AISE (Asymptotic Integrated Squared Error) realizations in the setting of $n = 50$, $m = 50$, and $\sigma = 1.0$: LocalL, PenLocalL, and PenLocalL2 label the curves associated with $\widetilde{g}_{LL;h}(\cdot)$ (solid), $\widehat{g}_{LL;h,\lambda_0}(\cdot)$ (dashed), $\widehat{g}_{LL;h,\lambda^{**}}(\cdot)$ (dash-dotted), with $h = h_0$ determined by R-function *dpill*(), $\lambda_0 = \eta_{ratio}$ defined in Remark 4.1, and λ^{**} by Algorithm A.

corresponding conventional estimates in all the settings. It gives further reduced AISEs using the tuning parameter λ^{**} determined by Algorithm A in sharpening the estimates. As expected, the optimal tuning parameter λ^* of (2) leads to the sharpened estimates with the lowest AISE sample means. Figure 1 displays the findings graphically by the density curves of the AISE values associated with the local linear ones with $\sigma = 1$ in Case A. $n = 50$, $m = 50$.

5.2. Data example

We downloaded the collection of weekly minimum and maximum temperatures at the Vancouver airport from the official web site of Environment Canada (`http://www.ec.gc.ca/`). See the dotted points in Figure S3 of the Supplementary Material for the recorded weekly min/max-temperatures during the periods 1937-1939, 1967-1969, and 1997-1999.

Assuming the regression model specified in Section 2, the average min/max-temperatures over time in each of the three time periods were estimated by (i) the local linear regression estimator (LocalReg) with the bandwidth h determined by R-function *dpill*(), and (ii) the penalized local linear regression estimator (PenLocalReg) guided by the differential equation $b \circ g = D^4 g + (2\pi/52)^2 D^2 g = 0$ and the tuning parameter λ selected according to Algorithm A. We used (ii) to impose

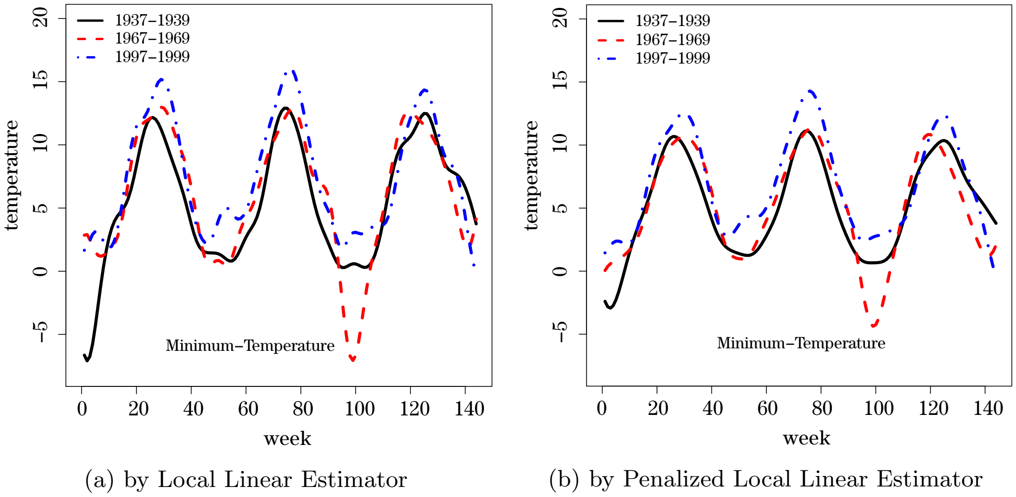

Figure 2. Change in Minimum-Temperature at Vancouver Airport: 1937-1939, 1967-1969 and 1997-1999 label the estimate curves for the periods of 1937-1939 (solid), 1967-1969 (dashed) and 1997-1999 (dash-dotted).

the one-year periodic pattern in temperature. Here the constant $\gamma = (2\pi/52)^2$ was chosen due to the time scale in weeks and each year approximated by 52 weeks. The figure in Section S3.2 of the Supplementary Material displays the local linear and penalized local linear estimate curves in red and blue, respectively. The sharpened estimates appear smoother than their local linear counterparts. We plot separately the estimated average min-temperature functions for the three three-year periods using the two approaches in Figure 2. Both sets of estimates reveal a clear increase in min-temperature over time.

6. Final Remarks

This paper is premised on the situation where an expert has provided advice about a possible functional or differential form that could underlie the given data. An equally important scenario would involve a trial of possible constraints to see how far the data must be perturbed in order for the constraints to be satisfied; thus, the approach could also be useful as an exploratory data analysis tool. We have considered penalties arising from functional equations based on linear transforms, such as homogeneous differential equations with known coefficients. When there are unknown coefficients, following Heckman and Ramsay (2000), we suggest first estimating the parameters using nonlinear least squares techniques. Exceptions to those in this large class include penalties to encourage non-negativity and monotonicity. Ramsay and Silverman (2005) provide presentations of a positive function and a monotone function by differential equations. These equations are not homogeneous and thus not based on linear transforms. The penalty approach proposed here is more general; nonlinear operators can be handled, in principle, but the computations become more complicated, and convenience of closed-form expressions for the sharpened data and the resulting estimators is lost.

Some practical situations involve random errors with non-constant variance. We can adapt the proposed approach using a variance function estimate obtained by, for example, the method presented in Fan and Yao (1998). We can accommodate correlated observations similarly with an estimate of the covariance function. Further, it is straightforward, in principle, to extend the proposed procedure to higher-dimensional data. It would be of interest to see if the form of data sharpening can alleviate the "curse of dimensionality", which makes it difficult to apply kernel estimators without resorting to additive models.

Supplementary Materials

The online supplementary materials provide (i) technical details of Sections 2 and 3, (ii) an alternative algorithm to Algorithm A for selecting the tuning parameter, and (iii) additional numerical results from the simulation study and the data analysis.

Acknowledgment

The authors are thankful to a Co-Editor for his instructive suggestions. We gratefully acknowledge the supports from the Natural Research Council of Canada (NSERC) and the Canadian Statistical Sciences Institute (CANSSI).

References

Braun, W. J. and Hall, P. (2001). Data sharpening for nonparametric inference subject to constraints. *Journal of Computational and Graphical Statistics* **10**, 786–806.

Cheng, M. Y. and Fan, J. (2016). Peter Hall's contributions to nonparametric function estimation and modeling. *The Annals of Statistics* **44**, 1837–1853.

Choi, E. and Hall, P. (1999). Miscellanea. Data sharpening as a prelude to density estimation. *Biometrika* **86**, 941–947.

Choi, E. and Hall, P. (2001). Nonparametric analysis of earthquake point-process data. *Lecture Notes-Monograph Series* **36**, 324–344.

Choi, E., Hall, P. and Rousson, V. (2000). Data sharpening methods for bias reduction in nonparametric regression. *The Annals of Statistics* **28**, 1339–1355.

Claeskens, G., and Hall, P. (2002). Data sharpening for hazard rate estimation. *Australian and New Zealand Journal of Statistics* **44**, 277–283.

Doosti, H. and Hall, P. (2016). Making a non-parametric density estimator more attractive, and more accurate, by data perturbation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **78**, 445–462.

Du, P., Parmeter, C. and Racine, J. S. (2013). Nonparametric kernel regression with multiple predictors and multiple shape constraints. *Statistica Sinica* **23**, 1343–1372.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties (with discussion). *Statistical Science* **11**, 89–121.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*, Chapman and Hall.

Fan, J. and Yao, Q. (1998). Efficient estimation of conditional variance functions in stochastic regression. *Biometrika* **85**, 645–660.

Hall, P. and Huang, H. (2001). Nonparametric kernel regression subject to monotonicity constraints. *The Annals of Statistics* **29**, 624–647.

Hall, P. and Kang, K. H. (2005). Unimodal kernel density estimation by data sharpening. *Statistica Sinica* **15**, 73–98.

Hall, P. and Minnotte, M. C., (2002). High order data sharpening for density estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **64**, 141–157.

He, H. and Huang, L.-S. (2009). Double-smoothing for bias reduction in local linear regression. *Journal of Statistical Planning and Inference* **139**, 1056–1072.

Heckman, N. E. and Ramsay, J. O. (2000). Penalized regression with model-based penalties. *Canadian Journal of Statistics* **28**, 241–258.

Henderson, H. V. and Searle, S. R. (1981). On deriving the inverse of a sum of matrices. *SIAM Review* **23**, 53–60.

Loader, C. (1999). *Local Regression and Likelihood*, Springer.

R Development Core Team (2015). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis* (2nd Edition), Springer.

Samiuddin, M. and El-Sayyad, G. M. (1990). On nonparametric kernel density estimates. *Biometrika* **77**, 865–874.

Wand, M. P. (2015). KernSmooth: Functions for Kernel Smoothing Supporting Wand & Jones (1995). R package version 2. 23–15.

Wand, M. P. and Jones, M. C. (1995). *Kernel Smoothing*, Chapman and Hall.

Wang, J. and Ghosh, S. K. (2012). Shape restricted nonparametric regression with Bernstein polynomials. *Computational Statistics and Data Analysis* **56**, 2729–2741

Irving K. Barber School of Arts and Sciences, University of British Columbia, Kelowna, BC V1V 1V7, Canada.

E-mail: john.braun@ubc.ca

Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, BC V5A 1S6, Canada.

E-mail: joanh@stat.sfu.ca

Department of Mathematics, University of Kansas, Lawrence, KS 66045, USA.
E-mail: x553k951@ku.edu

(Received January 2017; accepted September 2017)