DATA SHARPENING IN LOCAL REGRESSION GUIDED BY GLOBAL CONSTRAINT

W. John Braun[1], X. Joan Hu[2], and Xiuli Kang[3]

[1] *University of British Columbia,* [2] *Simon Fraser University,*

and [3] *University of Kansas*

Supplementary Material

This document presents (i) technical details of Sections 2 and 3 of the paper, (ii) an alternative algorithm to Algorithm A for selecting the tuning parameter, and (iii) additional numerical results from the simulation study and the real data analysis.

S1 Technical Details

S1.1 Technical Details of Section 2

We provide more discussions below to support Section 2.2 of the paper.

S1.1.1 Candidates for the conventional estimator $\widetilde{g}(z)$

We demonstrate the property in two cases with kernel methods below.

Local polynomial regression (LPR) (cf. Wand and Jones, 1995; Fan and Gijbels, 1996; Loader, 1999). This is a widely used procedure. Let $\mathbf{Q}_l(z)$ be the $n \times (l+1)$ matrix with the ith row vector of $\big(1, (x_i - z), \ldots, (x_i - z)^l\big)$ with $l \geq 0$; $\mathbf{K}_h(z)$, the diagonal matrix with the elements of $\big\{K_h(x_1 - z), \ldots, K_h(x_n - z)\big\}$, where $K_h(u) = K(u/h)/h$ with $K(\cdot)$ a kernel function and $h > 0$ a chosen bandwidth. An LPR estimator of order q (≥ 0) can be expressed as $\widetilde{g}(z) = \boldsymbol{a}(z;h)^\intercal \boldsymbol{y}$, where $\boldsymbol{a}(z;h)^\intercal$ is the first row vector of the matrix $\mathbf{S}(z) = \big\{\mathbf{Q}_q(z)^\intercal \mathbf{K}_h(z) \mathbf{Q}_q(z)\big\}^{-1} \mathbf{Q}_q(z)^\intercal \mathbf{K}_h(z)$.

Double-Smoothing (cf. He and Huang, 2009). This procedure enjoys a considerable reduction in bias over local linear regression while retaining favourable variance properties. The double-smoothed estimator is given by $\widetilde{g}_{DS}(x) = \int \{\widehat{\beta}_0(z) + \widehat{\beta}_1(z)(x-z)\} K_h(x-z)dz$, where $\widehat{\beta}_0(z)$ and $\widehat{\beta}_1(z)$ are the components of the vector $\mathbf{S}(z)\boldsymbol{y}$. In terms of vector-matrix multiplication, we can thus rewrite the double-smoothed estimator as $\widetilde{g}_{DS}(x) = \boldsymbol{a}(z;h)^\intercal \boldsymbol{y}$, where $\boldsymbol{a}(z;h)$ is the transpose of the $n \times n$ matrix $\int [1 \;\; (x-z)] \mathbf{S}(z) K_h(x - z)dz$.

S1.1.2 Functional equation penalties via linear transform

The desired functional constraint on $g(\cdot)$ in many applications may be formulated into a linear transform based equation. Such a functional constraint is exemplified below together with the induced expression for the matrix $\mathbf{B}(\boldsymbol{z};h)$ of the quadratic penalty in (2.1) of the paper.

Roughness Penality. Often used in nonparametric regression is the roughness penality to control the roughness of the resulting regression estimator. This may be achieved to keep small in a certain sense the second derivative of $\widehat{g}(z)$, which can be realized by imposing the quadratic penalty with $\big[b \circ g\big](z) = D^2 g(z) = 0$, where D is the differential operator. Thus the jth column of the $n \times m$ matrix $\mathbf{B}(\boldsymbol{z}; h)$ is $D^2 \boldsymbol{a}(z; h) = \big(D^2 a_1(z; h), \ldots, D^2 a_n(z; h)\big)^\intercal$ evaluated at $z = z_j$. One may wish to consider a slightly more general differential equation $\big[b \circ g\big](z) = D^2 g(z) + \gamma D g(z) = 0$ to capture the potentially nonlinear trajectory of the response in the form of $c_0 + c_1 \exp(-\gamma z)$.

Periodicity Constraint. There are many practical situations where the response exhibits cyclic variation according to the values of the explanatory variable, such as seasonality. The phenomenon shown in the data, together with prior knowledge about the response process, may suggest the use of a periodic constraint on $g(\cdot)$. In order for the resulting estimator to pick up both linear and periodic trends, Heckman and Ramsay (2000) consider the functional constraint $\big[b \circ g\big](z) = D^4 g(z) + \gamma D^2 g(z) = 0$; such a constraint can be imposed in our context as well. We consider this in the numerical studies in Section 5 of the paper. Note that the solution to the differential equation is $g(z) = a_0 + a_1 z + a_2 \sin(\sqrt{\gamma} z) + a_3 \cos(\sqrt{\gamma} z)$ for any constants

a_0, a_1, a_2 and a_3. The quadratic penalty $\boldsymbol{y}^{\star\intercal}\mathbf{B}(\boldsymbol{z};h)\mathbf{B}(\boldsymbol{z};h)^{\intercal}\boldsymbol{y}^{\star}$ of (2.1) of the paper is determined by $\mathbf{B}(\boldsymbol{z};h)$ with the jth column of $D^4\boldsymbol{a}(z;h) + \gamma D^2\boldsymbol{a}(z;h)$ evaluated at $z = z_j$.

The two examples above are special cases of the functional constraint based on a constant coefficient linear homogeneous differential equation: $\big[b \circ g\big](z) = 0$ with $b = \sum_{l=L_\star}^{L^\star} \alpha_l D^l$, where α_l are constants and $0 < L_\star < L^\star < \infty$. This class includes many desirable functional constraints and has been employed in the literature (e.g. Heckman and Ramsay, 2000). Our approach can also handle certain integral equation constraints. Examples include equations constructed via the Laplace transformation of $g(\cdot)$.

Symmetry Constraint. Symmetry is among the common geometric features of a function. If $g(x)$ is symmetric about $x = c$, $\int_{-z+c}^{z+c}(u-c)^k g(u)du = 0$ for $\forall z$ when k is an odd integer. We may choose to consider the functional constraint $\big[b \circ g\big](z) = \int_{-z+c}^{z+c}(u - c)g(u)du = 0$ in the proposed procedure to obtain $\widehat{g}(\cdot)$, closer to be symmetric about $x = c$ compared to $\widetilde{g}(z)$. The matrix $\mathbf{B}(\boldsymbol{z};h)$ in the quadratic penalty of (2.1) in the paper is determined with the jth column of $\big[b \circ \boldsymbol{a}(\cdot;h)\big](z) = \int_{-z+c}^{z+c}(u - c)\boldsymbol{a}(u;h)du$ evaluated at $z = z_j$. If $g(x)$ is symmetric about the point (c, d), $\int_{-z+c}^{z+c}(u - c)^k\{g(u) - d\}du = 0$ for $\forall z$ when k is even. The functional constraint $\big[b \circ g\big](z) = \int_{-z+c}^{z+c}(u - c)^2\{g(u) - d\}du = 0$ may serve the purpose.

S1.2 Technical Details of Section 3

We outline proofs for the propositions in Section 3 of the paper.

Proposition S1 *Given $\widetilde{g}(z) = \boldsymbol{a}(z;h)^\intercal \boldsymbol{y}$, a local regression estimator with independent observations $\{(y_i, x_i) : i = 1, \ldots, n\}$ 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 $\forall \boldsymbol{z}$, where the equal sign holds only when either $\lambda = 0$ or $\mathbf{B}(\boldsymbol{z}; h)^\intercal \boldsymbol{a}^\star(z; h, \lambda) = \mathbf{0}$.*

Proof. From (3.3) in the paper, we see that $\mathrm{Var}\big(\widetilde{g}(z)\big|\boldsymbol{x}, \boldsymbol{z}; h, \lambda\big) = \sigma^2 \boldsymbol{a}(z;h)^\intercal \boldsymbol{a}(z;h) = \sigma^2 \boldsymbol{a}^\star(z;h,\lambda)^\intercal \big(\mathbf{I} + \lambda \mathbf{B}(\boldsymbol{z};h)\mathbf{B}(\boldsymbol{z};h)^\intercal\big)^2 \boldsymbol{a}^\star(z;h,\lambda)$. Thus the difference of the two variances $\mathrm{Var}\big(\widehat{g}(z)\big|\boldsymbol{x}, \boldsymbol{z}; h, \lambda\big) - \mathrm{Var}\big(\widetilde{g}(z)\big|\boldsymbol{x}; h\big)$ is

$$\sigma^2 \boldsymbol{a}^\star(z;h,\lambda)^\intercal \Big[\mathbf{I} - \big(\mathbf{I} + \lambda \mathbf{B}(\boldsymbol{z};h)\mathbf{B}(\boldsymbol{z};h)^\intercal\big)^2\Big] \boldsymbol{a}^\star(z;h,\lambda).$$

Note further that $\Big[\mathbf{I} - \big(\mathbf{I} + \lambda \mathbf{B}\mathbf{B}^\intercal\big)^2\Big] = -2\lambda \mathbf{B}\big[\mathbf{I} + \frac{\lambda}{2}\mathbf{B}^\intercal\mathbf{B}\big]\mathbf{B}^\intercal$. The proposition follows since $\big[\mathbf{I} + \frac{\lambda}{2}\mathbf{B}\mathbf{B}^\intercal\big]$ is positive definite.

Proposition S2 *Suppose $\widetilde{g}(z) = \boldsymbol{a}(z;h)^\intercal \boldsymbol{y}$ in Proposition S1 is the local regression estimator of order q (≥ 0) and $g(x_i)$ 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*

$$\boldsymbol{a}(z;h)^{\intercal}\sum_{l=q+1}^{\infty}\frac{1}{l!}\Big\{g^{(l)}(z)(\boldsymbol{x}-z\mathbf{1})^{l}-\lambda\mathbf{B}(\boldsymbol{z};h)\big(\mathbf{I}+\lambda\mathbf{B}(\boldsymbol{z};h)^{\intercal}\mathbf{B}(\boldsymbol{z};h)\big)^{-1}\widetilde{\boldsymbol{g}}_{l}(\boldsymbol{x},\boldsymbol{z};h)\Big\}, \tag{S1.1}$$

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

PROOF. By (3.2) in the paper and

$$\boldsymbol{a}^{\star}(z;h,\lambda)=\big\{\mathbf{I}-\lambda\mathbf{B}(\boldsymbol{z};h)(\mathbf{I}+\lambda\mathbf{B}(\boldsymbol{z};h)^{\intercal}\mathbf{B}(\boldsymbol{z};h)\big)^{-1}\mathbf{B}(\boldsymbol{z};h)^{\intercal}\big\}\boldsymbol{a}(z;h),$$

$\mathrm{Bias}\big(\widehat{g}(z)\big|\boldsymbol{x},\boldsymbol{z};h,\lambda\big)$ equals

$$\big[\boldsymbol{a}(z;h)^{\intercal}\boldsymbol{g}-g(z)\big]-\lambda\boldsymbol{a}(z;h)^{\intercal}\mathbf{B}(\boldsymbol{z};h)(\mathbf{I}+\lambda\mathbf{B}(\boldsymbol{z};h)^{\intercal}\mathbf{B}(\boldsymbol{z};h)\big)^{-1}\mathbf{B}(\boldsymbol{z};h)^{\intercal}\boldsymbol{g}.$$

Here $\big[\boldsymbol{a}(z;h)^{\intercal}\boldsymbol{g}-g(z)\big]$ is

$$\mathrm{Bias}\big(\widetilde{g}(z)\big|\boldsymbol{x};h\big)=\boldsymbol{a}(z;h)^{\intercal}\sum_{l=q+1}^{\infty}\frac{g^{(l)}(z)}{l!}(\boldsymbol{x}-z\mathbf{1})^{l}=\sum_{l=q+1}^{\infty}\frac{g^{(l)}(z)}{l!}\sum_{i=1}^{n}a_i(z;h)(x_i-z)^{l} \tag{S1.2}$$

Note that $\boldsymbol{g}=g(u)\mathbf{1}+\sum_{l=1}^{\infty}g^{(l)}(u)(\boldsymbol{x}-u\mathbf{1})^{l}/l!$ for $u\in\mathcal{Z}$. Recall $\boldsymbol{a}(z;h)^{\intercal}\mathbf{1}=1$ and $\boldsymbol{a}(z;h)^{\intercal}(\boldsymbol{x}-z\mathbf{1})^{l}=1$ for $0<l\leq q$, and thus (S1.2) holds. We then see that $b\circ\big[\boldsymbol{a}(z;h)^{\intercal}\mathbf{1}\big]=\boldsymbol{b}(z;h)^{\intercal}\mathbf{1}=0$ and $b\circ\big[\boldsymbol{a}(z;h)^{\intercal}(\boldsymbol{x}-z\mathbf{1})^{l}\big]=\boldsymbol{b}(z;h)^{\intercal}(\boldsymbol{x}-$

$z\mathbf{1})^l = 0$ for $0 < l \le q$ because $b \circ (\boldsymbol{x} - z\mathbf{1})^l = \mathbf{0}$ with $L_\star > q \ge 0$. Therefore, $\mathbf{B}(\boldsymbol{z}; h)^\intercal \boldsymbol{g} = \sum_{l=q+1}^{\infty} \widetilde{\boldsymbol{g}}_l(\boldsymbol{x}, \boldsymbol{z}; h)/l!$ and thus prove the proposition.

Proposition S3 *The mean of the sum of squared residuals is*

$$E\Big\{\sum_{i=1}^{n}\big(Y_i - \widehat{g}(x_i)\big)^2\Big|\boldsymbol{x}; h, \lambda\Big\} = \sigma^2(n - 2\nu_1^\star + \nu_2^\star) + \sum_{i=1}^{n} Bias^2\big(\widehat{g}(x_i)\big|\boldsymbol{x}; h, \lambda\big), \tag{S1.3}$$

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

Proof. The proposition follows from $\mathrm{E}\Big\{\big(Y_i - \widehat{g}(x_i)\big)^2\Big|\boldsymbol{x}; h, \lambda\Big\} = \mathrm{Var}\big(Y_i - \widehat{g}(x_i)\big|\boldsymbol{x}; h, \lambda\big) + \mathrm{Bias}^2\big(\widehat{g}(x_i)\big|\boldsymbol{x}; h, \lambda\big)$, where $\mathrm{Var}\big(Y_i - \widehat{g}(x_i)\big|\boldsymbol{x}; h, \lambda\big) = \sigma^2\big\{1 - 2a_i^\star(x_i; h, \lambda) + \boldsymbol{a}^\star(x_i; h, \lambda)^\intercal \boldsymbol{a}^\star(x_i; h, \lambda)\big\}$.

S2 An Algorithm Alternative to *Algorithm A*

We outline in this section an algorithm for selecting the tuning parameter λ. It is a refined version of *Algorithm A*.

Denote the value of λ determined in the kth run by $\lambda^{(k)}$ for $k \ge 0$; the updated penalized local regression estimator with the tuning parameter $\lambda^{(k)}$, $\widehat{g}^{(k)}$; the updated variance estimate, $\widehat{\sigma^2}^{(k)}$.

Algorithm B. Choose the initial value $\lambda^{(0)}$ via *Algorithm A* with $g(\cdot)$ estimated by $\widetilde{g}(\cdot)$ with chosen h and the given estimate $\widetilde{\sigma}^2$, and update

$\widetilde{g}(\cdot)$ by the penalized local regression estimator evaluated with the tuning parameter $\lambda^{(0)}$, denoted by $\widehat{g}^{(0)}$, and the given estimate $\widetilde{\sigma}^2$ by $\widehat{\sigma^2}^{(0)} = \sum_{i=1}^{n}(y_i - \widehat{g}^{(0)}(x_i))^2 \big/ (n - 2\nu_1^\star + \nu_2^\star)$ as given in (3.5) of the paper with the sharpened estimator $\widehat{g}^{(0)}$.

Step B.1. Provided with $\lambda^{(k)}$, $\widehat{g}^{(k)}$ and $\widehat{\sigma^2}^{(k)}$ from the kth run for $k \geq 0$,

Step B.1.1. Replace $\widetilde{g}$ used in Steps A.2 and A.3 of *Algorithm A.* with $\widehat{g}^{(k)}$ to obtain an updated $\text{MAISE}_{\boldsymbol{z}}(\widehat{g}^{(k)} \big| \boldsymbol{x}; h, \lambda)$.

Step B.1.2. Determine $\lambda^{(k+1)}$ as $\arg\min_{all\ \lambda \geq 0} \text{MAISE}_{\boldsymbol{z}}(\widehat{g}^{(k)} \big| \boldsymbol{x}; h, \lambda)$.

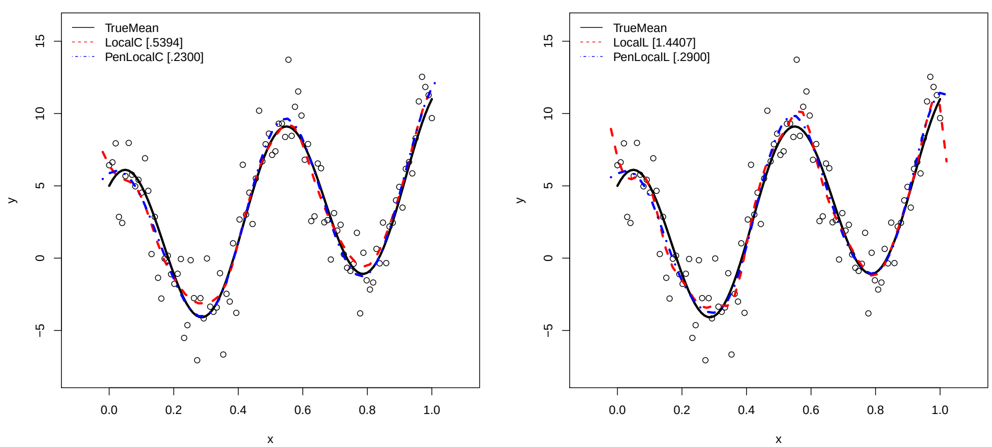
Step B.1.3. Compute $\widehat{g}^{(k+1)}$, the updated penalized local regression estimator evaluated with the tuning parameter $\lambda^{(k+1)}$, and the updated variance estimate $\widehat{\sigma^2}^{(k+1)} = \sum_{i=1}^{n}(y_i - \widehat{g}^{(k+1)}(x_i))^2 \big/ (n - 2\nu_1^\star + \nu_2^\star)$ as given in (3.6) of the paper with the sharpened estimator $\widehat{g}^{(k+1)}$.

Step B.2. Repeat Step B.1. until convergence and choose $\lambda^\star$ to be the limit of the sequence $\lambda^{(k)}, k = 1, \ldots$.

S3 Additional Numerical Results

S3.1 Simulation

This section presents summary of the simulation outcomes associated with the estimation procedures based on the local constant (LC) estimator.



(a) by Local Constant Estimator

(b) by Local Linear Estimator

Figure S1: Penalized local linear/constant estimates and their local regression correspondences in the setting of $n = 100$, $m = 50$, and $\sigma = 2.0$: the dot points are the generated observations; TrueMean, LocalC/LocalL, and PenLocalC/PenLocalL label the true mean $g(\cdot)$ (solid), and estimate curves $\widetilde{g}_{LC;h}(\cdot)/\widetilde{g}_{LL;h}(\cdot)$ (dashed) and $\widehat{g}_{LC;h,\lambda^\star}(\cdot)/\widehat{g}_{LL;h,\lambda^\star}(\cdot)$ (dash-dotted) with $h = h_0$ determined by R-function $dpill()$, and $\lambda^\star$ by Algorithm A; the values in brackets are the corresponding AISE($\widehat{g}$).

Table S1: Summary Statistics of the Approximate Integrated Squared Errors (AISE) in Simulation: *associated with the local constant (LC) estimator*

	Estimator		$\widehat{g}_{LC;h,\lambda}(\cdot)$		
		$\widetilde{g}_{LC;h}(\cdot)^a$	$(1)^b\ \lambda = \lambda_0$	$(2)^c\ \lambda = \lambda^*$	$(3)^d\ \lambda = \lambda^{**}$
		Case A. $n = 50$, $m = 50$			
$\sigma = .3$	AISE^e_{SM}	0.245	0.247	0.243	0.244
	AISE^f_{SD}	(.041)	(.042)	(.041)	(.039)
$\sigma = 1$	AISE_{SM}	0.394	0.360	0.324	0.345
	AISE_{SD}	(.163)	(.164)	(.159)	(.147)
$\sigma = 3$	AISE_{SM}	1.767	1.572	1.112	1.388
	AISE_{SD}	(.764)	(.730)	(.645)	(.529)
		Case B. $n = 50$, $m = 100$			
$\sigma = 1$	AISE_{SM}	0.201	0.172	0.136	0.154
	AISE_{SD}	(.067)	(.063)	(.059)	(.049)
		Case C. $n = 100$, $m = 50$			
$\sigma = 1$	AISE_{SM}	0.401	0.370	0.329	0.353
	AISE_{SD}	(.177)	(.177)	(.166)	(.155)

[a] $h = h_0$ determined by R-function $dpill()$.
[b] (1) $\lambda_0 = \lambda_{coef}$ as defined in Remark 4.1.
[c] (2) $\lambda^* = \arg\min_\lambda MAISE(\widehat{g}_\lambda)$ 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.

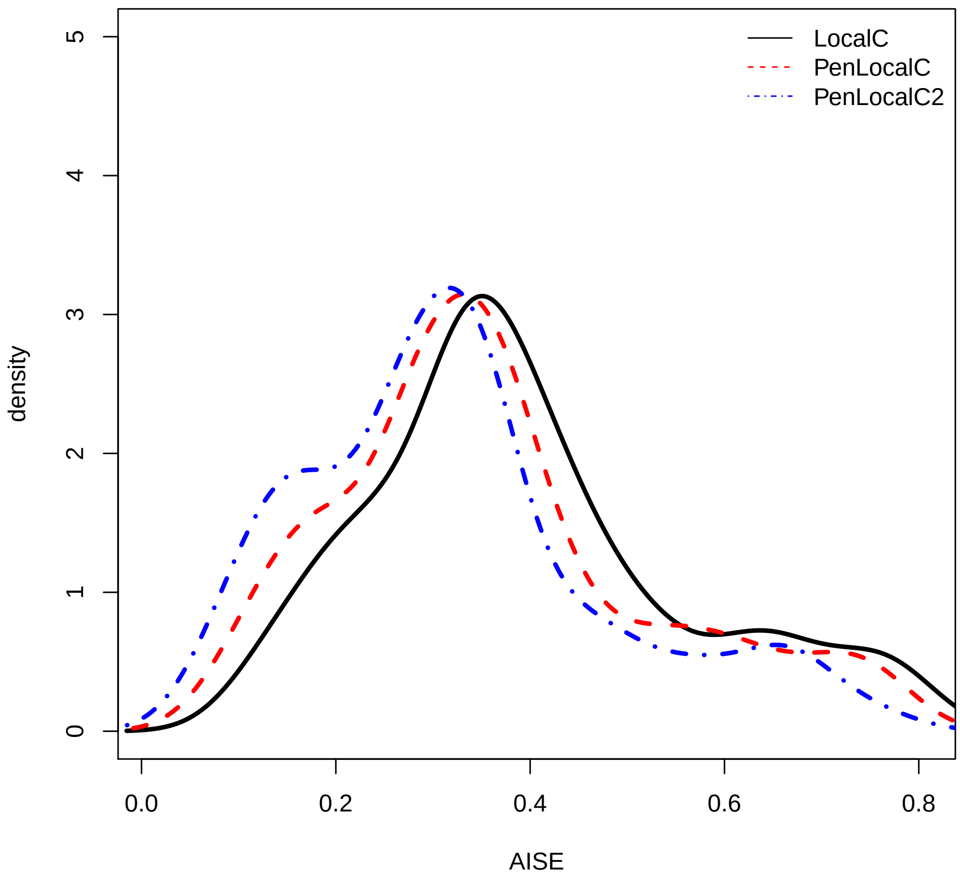


Figure S2: Density curves of the AISE values in the setting of $n = 50$, $m = 50$, and $\sigma = 1.0$: LocalC, PenLocalC, and PenLocalC2 label the curves associated with $\widetilde{g}_{LC;h}(\cdot)$ (solid), $\widehat{g}_{LC;h,\lambda_0}(\cdot)$ (dashed), $\widehat{g}_{LC;h,\lambda^{**}}(\cdot)$ (dash-dotted), with $h = h_0$ determined by R-function $dpill()$, $\lambda_0 = \eta_{ratio}$ defined in Remark 4.1, and λ^{**} by Algorithm A.

S3.2 Real data example

Figure S3 presents the original records of both the minimum and maximum temperatures at the Vancouver airport during the selected three three-year periods together with the local linear estimates and the sharpened local linear estimates.

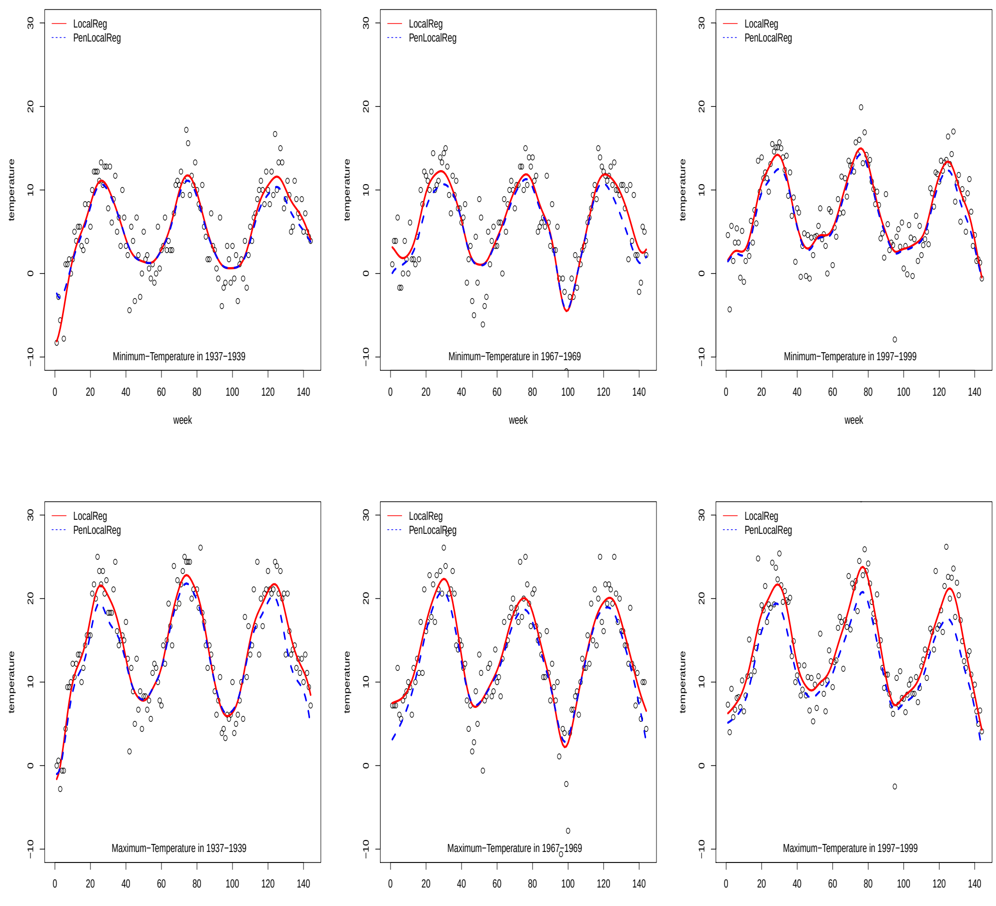


Figure S3: Temperature at Vancouver Airport: dot points are the recorded temperatures; LocalReg and PenLocalReg label the estimate curves by the local linear (solid) and sharpened local linear (dashed) estimators.