Some asymptotics for local least-squares regression with regularization

Jürgen Franke and Joseph Tadjuidje Kamgaing
Department of Mathematics, University of Kaiserslautern

Stefan Didas and Joachim Weickert
Department of Mathematics, University of Saarbrücken
P.O.Box 3049, 67653 Kaiserslautern, Germany

Abstract: We derive some asymptotics for a new approach to curve estimation proposed by Mrázek et al. [3] which combines localization and regularization. This methodology has been considered as the basis of a unified framework covering various different smoothing methods in the analogous two-dimensional problem of image denoising. As a first step for understanding this approach theoretically, we restrict our discussion here to the least-squares where we have explicit formulas for the function estimates and where we can derive a rather complete asymptotic theory from known results for the Priestley-Chao curve estimate.

1 Introduction

We consider data generated from the nonparametric regression model

$$f_j = \mu(x_j) + \varepsilon_j,\ \ j = 1, \ldots, N, \tag{1}$$

where $\varepsilon_1, \ldots, \varepsilon_N$ are independent and identically distributed with mean 0 and finite variance σ^2 and $x_j = \frac{j}{N}$, $j = 1, \ldots, N$, form an equidistant grid on the unit interval $[0, 1]$. Mrázek et al. [3] have described a general approach for image denoising which, by combining localization and regularization, includes most of the known image denoising algorithms like local smoothing or nonlinear diffusion filtering. Estimating the function $\mu(x)$ at the grid points $x_1, \ldots, x_N$ from the data $f_1, \ldots, f_N$ generated by (1) is the one-dimensional analogon of the image denoising problem. Mrázek et al. [3] propose to solve this problem by minimizing a target function like

$$Q(\mathbf{u}) = \sum_{i,j=1}^{N} \Psi_D(|u_i - f_j|^2) w_D(|x_i - x_j|^2) + \frac{\lambda}{2} \sum_{i,j=1}^{N} \Psi_S(|u_i - u_j|^2) w_S(|x_i - x_j|^2)$$

w.r.t. $\mathbf{u} = (u_1, \ldots, u_N)^T$.

Let $u(x)$ denote the function estimate of $\mu(x)$ which we first consider only at $x_1, \ldots, x_N$. Ψ_D, Ψ_S are penalizing functions measuring the fit of $u(x)$ to the observations $f_1, \ldots, f_N$ and the smoothness of $u(x)$ resp. The spatial weighting functions w_D, w_S guarantee a localization effect, and $\lambda \geq 0$ balances between fit and smoothness. To get some first insight into the asymptotic properties of the resulting estimates $u_1, \ldots, u_N$ for $\mu(x_1), \ldots, \mu(x_N)$, we investigate the following special case:

$$\Psi_D(s^2) = s^2,\ \ \Psi_S(s^2) = s^2, w_D(x^2) = K_h(x) = \frac{1}{h} K(\frac{x}{h}),\ \ w_S(x^2) = L_g(x) = \frac{1}{g} L(\frac{x}{g})$$

where the kernels K and L are nonnegative, symmetric functions on $\mathbb{R}$ and the bandwidths $h, g > 0$ can be chosen to control the smoothness of the function estimate together with the balancing factor λ. Therefore, we consider minimizing

$$Q(\mathbf{u}) = \sum_{i,j=1}^{N} (u_i - f_j)^2 K_h(x_i - x_j) + \frac{\lambda}{2} \sum_{i,j=1}^{N} (u_i - u_j)^2 L_g(x_i - x_j). \tag{2}$$

We call the resulting estimate $\mathbf{u}$ a *regularized local least-squares estimate.*

For $\lambda = 0$, we immediately get from $\frac{\partial}{\partial u_i} Q(\mathbf{u}) = 0$, $i = 1, \ldots, N$, that $u_i = \tilde{\mu}(x_i, h)$, $i = 1, \ldots, N$, where $\tilde{\mu}(x, h)$ is the classical Nadaraya-Watson kernel regression estimate (compare, e.g., Härdle [2])

$$\begin{aligned} \tilde{\mu}(x, h) &= \frac{\hat{\mu}(x, h)}{\hat{p}_K(x, h)} \qquad \text{with} \qquad\qquad (3) \\ \hat{\mu}(x, h) &= \frac{1}{N} \sum_{j=1}^{N} f_j K_h(x - x_j), \quad \hat{p}_K(x, h) = \frac{1}{N} \sum_{j=1}^{N} K_h(x - x_j). \end{aligned}$$

As $\hat{p}_K(x, h)$ converges to 1 for the equidistant x_i, the Nadaraya-Watson estimate $\tilde{\mu}(x, h)$ and the Priestley-Chao estimate $\hat{\mu}(x, h)$ are asymptotically equivalent.

2 Asymptotic expansion

The main goal of this section is to show that the regularized local least-squares estimate, which we get by minimizing (2), is closely related asymptotically for $N \to \infty$ to the simple Priestley-Chao estimate (Priestley and Chao, [4]) with bandwidth h where, however, the regularization parameter λ provides an additional tool for fine tuning the properties of the estimate. As a first step, we show that the solution of (2) has an explicit representation in terms of the Priestley-Chao estimate. For convenience, we use the following notation for the values of this estimate at the grid points $x_i, i = 1, \ldots, N$:

$$\hat{\mu} = (\hat{\mu}_1, \ldots, \hat{\mu}_N)^T \quad \text{with } \hat{\mu}_i = \hat{\mu}(x_i, h),\ i = 1, \ldots, N.$$

Proposition 1 *Let $\hat{p}_L(x,g)$ be defined analogously to $\hat{p}_K(x,h)$ with L, g replacing K, h, and let $\hat{p}_\lambda(x,h,g) = \hat{p}_K(x,h) + \lambda \hat{p}_L(x,g)$. Let Λ denote the $N \times N$-matrix with entries $\Lambda_{i,j} = \frac{1}{N} L_g(x_i - x_j)$, and let $\hat{P}$ denote the $N \times N$-diagonal matrix with entries $\hat{P}_{ii} = \hat{p}_\lambda(x_i, h, g)$. Then, if $\hat{P} - \lambda\,\Lambda$ is invertible, the solution of (2) is given as*

$$\mathbf{u} = \{\hat{P} - \lambda\Lambda\}^{-1}\hat{\mu}. \tag{4}$$

Proof: Setting the partial derivatives of $Q(\mathbf{u})$ w.r.t. $u_k,\ k = 1, \ldots, N$, to 0, we get

$$0 = 2\sum_{j=1}^{N}(u_k - f_j)K_h(x_k - x_j) + 2\lambda\sum_{j=1}^{N}(u_k - u_j)L_g(x_k - x_j),\ k = 1, \ldots, N,$$

where we have used the symmetry of L_g. Therefore, we have for $k = 1, \ldots, N$,

$$u_k\left\{\sum_{j=1}^{N} K_h(x_k - x_j) + \lambda\sum_{j=1}^{N} L_g(x_k - x_j)\right\}$$

$$= \lambda\sum_{j=1}^{N} u_j L_g(x_k - x_j) + \sum_{j=1}^{N} f_j K_h(x_k - x_j)$$

which, using the definition of $\hat{p}_K, \hat{p}_L$ and $\hat{\mu}$, implies the assertion. ■

Now, we want to prove that the estimate $\mathbf{u}$, given by (4), is consistent in a certain sense if $N \to \infty$. First, we investigate $\hat{p}_K(x,h)$. If the x_j would be i.i.d. random variables, then $\hat{p}_K$ would be the well-known Rosenblatt-Parzen estimate of their common probability density (compare, e.g., Silverman, [6]). In our case, $x_1, \ldots, x_N$ are equidistant and behave similar to uniform random variables, i.e. in particular $\hat{p}_K(x,h) \to 1$ under appropriate assumptions. We assume

(A1) a) K is a nonnegative, symmetric kernel function with compact support $[-1, +1]$.

b) $\int K(y)dy = 1$

c) K is Lipschitz continuous with Lipschitz constant C_K.

We could relax the assumptions of symmetry and compactness of the support of K, but we want to keep our arguments simple in this paper. Due to the same reason, we mainly neglect boundary effects, which could be dealt with as in section 4.4 of Härdle [2], by restricting our attention to $x \in [h, 1-h]$. Asymptotically, this will have no effect as we shall have $h \to 0$ for $N \to \infty$ anyhow. In the following, we will frequently approximate a Riemann sum by the corresponding integral. For reference, we, therefore, state the following Lemma:

Lemma 1 *Let $g(y)$ be Lipschitz continuous on $[0,1]$ with Lipschitz constant C. Then,*

$$\Big|\int_0^1 g(y)dy - \frac{1}{N}\sum_{j=1}^N g(x_j)\Big| \le \frac{C}{N}.$$

Proof: By the mean-value theorem of integration there are $y_j \in [x_{j-1}, x_j],\ j = 1,\ldots,N$, where $x_0 = 0$, such that, using $x_j - x_{j-1} = \frac{1}{N}$ for all j,

$$\begin{aligned}\Big|\int_0^1 g(y)dy - \frac{1}{N}\sum_{j=1}^N g(x_j)\Big| &= \Big|\sum_{j=1}^N \{\int_{x_{j-1}}^{x_j} g(y)dy - \frac{1}{N}g(x_j)\}\Big| \\ &= \Big|\sum_{j=1}^N \{g(y_i) - g(x_j)\}\frac{1}{N}\Big| \\ &\le \frac{1}{N}\sum_{j=1}^N C|y_j - x_j| \le \frac{C}{N}\end{aligned}$$

■

As an immediate consequence, we have with $g(y) = K_h(x-y)$:

Corollary 1 *Assuming (A1) for the kernel K, we have*

$$|1 - \hat{p}_K(x,h)| \le C_K \frac{1}{Nh^2} \quad \textit{for all} \quad x \in [h, 1-h].$$

As a next step, we investigate the asymptotic behaviour of $\hat{\mu}$ as an estimate of $(\mu(x_1),\ldots,\mu(x_N))^T$. We assume

(A2) a) μ is twice continuously differentiable

b) $\mu''(x)$ is Hoelder continuous on $[0,1]$ with exponent β, i.e. for some $\beta > 0, H < \infty$ $|\mu''(x) - \mu''(z)| \le H|x-y|^\beta$ for all $x, y \in [0,1]$

The following asymptotic expansions for bias and variance of the Priestley-Chao estimate is well known. We only use assumption (A2b) to get a slightly more precise assertion about the remainder of the bias which will turn out to be useful later on, and we add a more detailed statement about the asymptotic covariance of estimates at different locations.

Proposition 2 *Assuming (A1) and (A2), we have for the Priestley-Chao estimate $\hat{\mu}(x,h)$, based on (1), for $N \to \infty,\ h \to 0$ such that $Nh^2 \to \infty$:*

i) bias $\hat{\mu}_i = \mathrm{E}\ \hat{\mu}_i - \mu(x_i) = \frac{h^2}{2}\mu''(x_i)V_K + r(x_i,h) + O(\frac{1}{Nh^2})$ *uniformly in $x_i \in [h, 1-h]$, where*

$$V_K = \int z^2 K(z)dz$$

and the main remainder term is

$$r(x,h) = \frac{h^2}{2}\int K(z)z^2\{\mu''(x - \vartheta hz) - \mu''(x)\}dz = O(h^{2+\beta})$$

uniformly in x.

ii) var $\hat{\mu}_i = \mathrm{E}(\hat{\mu}_i - \mathrm{E}\ \hat{\mu}_i)^2 = \frac{\sigma^2}{Nh}Q_K + O(\frac{1}{N^2h^3})$ *uniformly in* $x_i \in [h, 1-h]$*, where*

$$Q_K = \int K^2(z)dz$$

iii) mse $\hat{\mu}_i = \mathrm{E}(\hat{\mu}_i - \mu(x_i))^2 = \frac{\sigma^2}{Nh}Q_K + \frac{h^4}{4}\{\mu''(x_i)\}^2 V_K^2 + O(h^{4+2\beta}) + O(\frac{1}{N^2h^4})$ *uniformly in* $x_i \in [h, 1-h]$*. In particular,*

$$\hat{\mu}_i - \mu(x_i) \longrightarrow 0 \ \ \textit{in probability.}$$

iv) cov $(\hat{\mu}_i, \hat{\mu}_k) = 0$ *if* $|x_i - x_k| > 2h$*, and*
cov $(\hat{\mu}_i, \hat{\mu}_k) = \frac{\sigma^2}{Nh}\int K(z)K(z + \frac{x_i - x_k}{h})dz + O(\frac{1}{N^2h^3})$ *uniformly in* $x_i, x_k \in [h, 1-h]$*, else.*

If μ does not satisfy the smoothness condition (A2) everywhere, then Proposition 1 still holds in every subinterval $[a, b] \subset [0, 1]$ where (A2) is satisfied, as is obvious from the proof. So, if μ jumps in x^*, but otherwise is smooth enough, the assertions of the Proposition hold uniformly in $x_i, x_k \in [h, x^* - h] \cup [x^* + h, 1 - h]$.

Proof:

a) We use the common decomposition of mean-squared error into variance and squared bias

$$\text{mse } \hat{\mu}_i = \mathrm{E}(\hat{\mu}(x_i, h) - \mu(x_i))^2 = \text{var } \hat{\mu}(x_i, h) + \{\text{bias } \hat{\mu}(x_i, h)\}^2.$$

We have uniformly in $x \in [h, 1-h]$

$$\begin{aligned}
\text{var } \hat{\mu}(x, h) &= \frac{1}{N^2}\sum_{j=1}^{N} K_h^2(x - x_j) \text{ var } f_j \\
&= \frac{\sigma^2}{N}\int K_h^2(x - y)dy + O(\frac{1}{N^2h^3}) \\
&= \frac{\sigma^2}{Nh}\int K^2(z)dz + O(\frac{1}{N^2h^3})
\end{aligned}$$

using Lemma 1 with $g(y) = K_h^2(x-y)$. Moreover, using Lemma 1 with $g(y) = K_h(x-y)$,

$$\begin{aligned}
\text{bias } \hat{\mu}(x, h) &= \mathrm{E}\ \hat{\mu}(x, h) - \mu(x) \\
&= \frac{1}{N}\sum_{j=1}^{N} K_h(x - x_j)\mu(x_j) - \mu(x) \\
&= \int K(z)\{\mu(x - hz) - \mu(x)\}dz + O(\frac{1}{Nh^2}) \\
&= \int K(z)(-hz)dz\, \mu'(x) + \int K(z)\frac{h^2z^2}{2}\mu''(x - \vartheta hz)dz \ + O(\frac{1}{Nh^2}) \\
&= \frac{h^2}{2}\mu''(x)\int z^2K(z)dz + r(x, h) + O(\frac{1}{Nh^2})
\end{aligned}$$

again uniformly in $x \in [h, 1-h]$. We have used Assumption (A2) together with a Taylor expansion of m and symmetry of K. Combining the bias and variance expansion, we get for $N \to \infty, h \to 0, Nh^2 \to \infty$

$$\text{mse } \hat{\mu}_i = \frac{\sigma^2}{Nh}Q_K + \frac{h^4}{4}(\mu''(x_i))^2 V_K^2 + O(h^{4+2\beta}) + O(\frac{1}{N^2h^4}).$$

b) Analogously to a), we conclude using the independence of the f_j.

$$\begin{aligned}\operatorname{cov}(\hat{\mu}(x,h),(\hat{\mu}(x',h)) &= \frac{\sigma^2}{N^2}\sum_{j=1}^{N} K_h(x-x_j)K_h(x'-x_j)\\ &= \frac{\sigma^2}{Nh}\int K(z)K(z+\frac{x-x'}{h})dz + O(\frac{1}{N^2h^3}).\end{aligned}$$

By compactness of the support of K, we have for $|x-x'|>2h$ that $K_h(x-x_j)K_h(x'-x_j)=0$ and, therefore, cov $(\hat{\mu}(x,h),\hat{\mu}(x',h))=0$.

■

We need the following generalization of Corollary 1 which takes care of the boundary effects:

Corollary 2 *Assuming (A1) and, additionally, (A1'): K decreasing in $[0,1]$ and $K(1)=0$ we have*

i) $|1-\hat{p}_K(x,h)| \le C_K\frac{1}{Nh^2}$ *for* $h\le x\le 1-h$

ii) $\min\{\frac{1}{2}, 1-\frac{1}{2}C_K(1-\frac{x}{h})^2\} - C_K\frac{1}{Nh^2} \le \hat{p}_K(x,h) \le 1+C_K\cdot\frac{1}{Nh^2}$ *for* $0\le x\le h$

iii) $\min\{\frac{1}{2}, 1-\frac{1}{2}C_K(1-\frac{1-x}{h})^2\} - C_K\frac{1}{Nh^2} \le \hat{p}_K(x,h) \le 1+C_K\cdot\frac{1}{Nh^2}$ *for* $1-h\le x\le 1$

Proof: i) follows from Corollary 1. By Lemma 1, we have for $0\le x\le h:$

$$|\int_0^1 K_h(x-y)dy - \frac{1}{N}\sum_{j=1}^{N} K_h(x-x_j)| \le C_K\cdot\frac{1}{Nh^2}.$$

Now, $\int_0^1 K_h(x-y)dy = \int_{-\frac{x}{h}}^{\frac{1-x}{h}} K(z)dz = 1-\int_{-1}^{-\frac{x}{h}} K(z)dz = 1-\int_{\frac{x}{h}}^{1} K(z)dz$ using the symmetry of K. By symmetry and nonnegativity of K, the right-hand side is in $[\frac{1}{2},1]$. As $K(1)=0$ and K is Lipschitz continuous, $K(z)\le C_K(1-z)$ for $0\le z\le 1$, i.e.

$$\int_0^1 K_h(x-y)dy \ge 1-C_K\int_{\frac{x}{h}}^{1}(1-z)dz = 1-C_K\frac{1}{2}(1-\frac{x}{h})^2$$

ii) follows. iii) can be shown analogously. ■

Proposition 2 describes the asymptotic behaviour of $\hat{\mu}$ which is related to the final estimate $\mathbf{u}$ by (4). To get the asymptotic properties of $\mathbf{u}$, we have to investigate the matrix factor of (4). For $\lambda\to 0$ with $N\to\infty$, we have

$$\begin{aligned}\{\hat{P}-\lambda\Lambda\}^{-1} &= \{I-\lambda\,\hat{P}^{-1}\Lambda\}^{-1}\hat{P}^{-1}\\ &= \{I+\lambda\,\hat{P}^{-1}\Lambda\}\hat{P}^{-1} + O\big(\lambda^2||(\hat{P}^{-1}\Lambda)^2\hat{P}^{-1}||\big).\end{aligned}$$

As, by Corollary 2, $\hat{p}_\lambda(x_j,h,g)=\hat{P}_{jj}$ is bounded away from 0 uniformly in j for large enough N, the remainder term is of order $O(\lambda^2||\Lambda^2||)$. Now, be definition of Λ, we have for each unit vector $e_k\in\mathbb{R}^N,\ k=1,\dots,N,$

$$
\begin{aligned}
||\Lambda^2 e_k||^2 &= \sum_i \Big(\sum_j \Lambda_{ij}\Lambda_{jk} \Big)^2 \\
&= \sum_i \Big(\frac{1}{N^2} \sum_j L_g(x_i - x_j) L_g(x_j - x_k) \Big)^2 \\
&= \sum_i \Big(\frac{1}{N} \Big\{ \int_0^1 L_g(x_i - y) L_g(y - x_k) dy + O(\frac{1}{Ng^2}) \Big\} \Big)^2 \\
&= \frac{1}{N} \int_0^1 \Big(\int_0^1 L_g(z-y) L_g(y - x_k) dy \Big)^2 dz + O\Big(\frac{1}{N^2 g^2}\Big) \\
&\leq \frac{1}{Ng} \int_{-\infty}^{\infty} \Big(\int_{-\infty}^{\infty} L(u) L(v-u) du \Big)^2 dv + O\Big(\frac{1}{N^2 g^2}\Big) \\
&= O\Big(\frac{1}{Ng}\Big)
\end{aligned}
$$

uniformly in k by Lemma 1. Let $y = (y_1, \dots, y_N)^T \in \mathbb{R}^N$ be arbitrary with $||y|| = 1$. Then

$$
||\Lambda^2 y|| \leq \sum_k |y_k| \; ||\Lambda^2 e_k|| \leq O(\frac{1}{\sqrt{Ng}}) \sum_k |y_k| \leq O(\frac{1}{\sqrt{Ng}}) ||y||.
$$

We conclude $||\Lambda^2|| = O(\frac{1}{\sqrt{Ng}})$ for the operator norm of Λ^2. Therefore, we finally have

$$
\mathbf{u} = \Big\{ \hat{P}^{-1} + \lambda \hat{P}^{-1} \Lambda \hat{P}^{-1} + O(\frac{\lambda^2}{\sqrt{Ng}}) \Big\} \hat{\mu}
$$

or, coordinatewise, as $\hat{\mu}_i$ is stochastically bounded by Proposition 2 and Chebyshev's inequality uniformly in $\{i; x_i \in [h, 1-h]\}$

$$
u_i = \frac{1}{\hat{p}_\lambda(x_i, h, g)} \Big\{ \hat{\mu}(x_i, h) + \lambda \frac{1}{N} \sum_j \frac{L_g(x_i - x_j)}{\hat{p}_\lambda(x_j, h, g)} \hat{\mu}(x_j, h) \Big\} + O_p(\frac{\lambda^2}{\sqrt{Ng}}). \tag{5}
$$

We now neglect the boundary and consider only x_i with

$$
\max(h, g) + g \leq x_i \leq 1 - \max(h, g) - g. \tag{6}
$$

Then, as L_g has support $[-g, g]$, we have $\max(h, g) \leq x_j \leq 1 - \max(h, g)$ for all j corresponding to the nonvanishing terms in the sum on the right-hand side of (5). Therefore, for all those j, we have by Corollary 1

$$
|1 - \hat{p}_K(x_j, h)| \leq C_K \frac{1}{Nh^2}, \; |1 - \hat{p}_L(x_j, g)| \leq C_L \frac{1}{Ng^2}.
$$

As, by Corollary 2, $\hat{p}_K, \hat{p}_L$ are uniformly bounded from below by a positive constant for N large enough, we have uniformly for all those j

$$
\Big| \frac{1}{\hat{p}_\lambda(x_j, h, g)} - \frac{1}{1+\lambda} \Big| \leq \; \delta_N \qquad \text{with}
$$

$$\delta_N = O(\frac{1}{Nh^2}) + O(\frac{\lambda}{Ng^2}).$$

As this holds in particular for $j = i$, we have from (5)

$$\begin{aligned} u_i &= \frac{1}{1+\lambda}\Big\{\hat{\mu}(x_i,h) + \frac{\lambda}{1+\lambda}\frac{1}{N}\sum_j L_g(x_i - x_j)\hat{\mu}(x_j,h)\Big\} + R^*_{N,i} \\ &= \frac{1}{1+\lambda}\Big\{\hat{\mu}(x_i,h) + \frac{\lambda}{1+\lambda}\hat{\mu}_2(x_i,h,g)\Big\} + R^*_{N,i} \end{aligned} \tag{7}$$

where the remainder term $R^*_{N,i}$ is $O(\frac{\lambda^2}{\sqrt{Ng}}) + O_p(\delta_N)$ uniformly in all i satisfying (6). As we already know the asymptotic behaviour of $\hat{\mu}$, we have to investigate

$$\begin{aligned} \hat{\mu}_2(x,h,g) &= \frac{1}{N}\sum_j L_g(x - x_j)\hat{\mu}(x_j,h) \\ &= \frac{1}{N^2}\sum_{j,l} L_g(x - x_j)K_h(x_j - x_l)f_l \\ &= \sum_l \gamma_l(x,h,g)f_l + \hat{\mu}(x,h) \end{aligned}$$

with

$$\gamma_l(x,h,g) = \frac{1}{N^2}\sum_j L_g(x - x_j)K_h(x_j - x_l) - \frac{1}{N}K_h(x - x_l).$$

Analogously to Lemma A2 of Franke and Härdle [1], we have

Lemma 2

$$\begin{aligned} \gamma_l(x,h,g) &= O(\frac{g}{Nh^2}) \qquad \textit{uniformly in} \quad h + g \le x \le 1 - (h+g) \\ \gamma_l(x,h,g) &= 0 \qquad\qquad \textit{if } |x - x_l| > h + g \end{aligned}$$

Proposition 3 *Assume (A1), (A2), (A1'). Then, if* $\frac{g}{h} \to 0$ *for* $N \to \infty,\ h \to 0$

i) $\quad \text{var}\,\big(\sum_l \gamma_l(x,h,g)f_l\big) = O(\frac{g^2}{Nh^3})$

ii) $\quad Nh \text{ var } \hat{\mu}_2(x,h,g) = Nh \text{ var } \hat{\mu}(x,h) + o(1)$

iii) $\quad \sqrt{Nh}\,E\big(\sum_l \gamma_l(x,h,g)f_l\big) = \sqrt{Nh}\sum_l \gamma_l(x,h,g)\mu(x_l) = \sqrt{Nh} \text{ bias } \hat{\mu}(x,h) + o(1)$

uniformly in $h + g \le x \le 1 - (h+g)$.

Proof: As in the proof of (A7) of Franke and Härdle [1] we get, using Lemma 2,

$$Nh \text{ var}(\sum_l \gamma_l(x,h,g)f_l) = Nh\ E(\sum_l \gamma_l(x,h,g)\varepsilon_l)^2 \le c\ \frac{(h+g)g^2}{h^3}$$

for some constant $c > 0$. i) follows. As, by Proposition 2, the variance of $\hat{\mu}(x,h)$ is of order $\frac{1}{Nh}$ and as $\frac{g}{h} \to 0$, $\hat{\mu}(x,h)$ is the dominant term in the definition of $\hat{\mu}_2(x,h,g)$, and ii) follows from i). iii) can be shown as in the proof of Theorem 1, part d) of Franke and Härdle [1]. ■

From (7), we get,

$$\begin{aligned} u_i &= \frac{1}{1+\lambda}\left\{\hat{\mu}(x_i,h) + \frac{\lambda}{1+\lambda}\hat{\mu}(x_i,h) + \frac{\lambda}{1+\lambda}\sum_l \gamma_l(x_i,h,g)f_l\right\} + R^*_{N,i} \\ &= \frac{1}{(1+\lambda)^2}\left\{(1+2\lambda)\hat{\mu}(x_i,h) + \lambda\ E\ \sum_l \gamma_l(x_i,h,g)f_l\right\} + R^{**}_{N,i} \end{aligned}$$

with remainder term $R^{**}_{N,i} = R^*_{N,i} + O_p(\frac{g}{h}\frac{\lambda}{\sqrt{Nh}})$, as by Proposition 3, i),

$$\sum_l \gamma_l(x_i,h,g)f_l - E\ \sum_l \gamma_l(x_i,h,g)f_l = O_p(\frac{g}{h}\frac{1}{\sqrt{Nh}}). \tag{8}$$

Using Proposition 3, iii), we get

$$u_i = \frac{1}{(1+\lambda)^2}\{(1+2\lambda)\hat{\mu}(x_i,h) + \lambda\,\text{bias}\ (\hat{\mu}(x_i,h))\} + R_{N,i} \tag{9}$$

with $R_{N,i} = R^{**}_{N,i} + \lambda o_p\big(\frac{1}{\sqrt{Nh}}\big)$. We summarize this to the main result of this section:

Theorem 2.1 *Let (A1), (A2), (A1') hold. Then, for* $N \to \infty,\ h \to 0, g \to 0, \lambda \to 0$, *such that*

$$\frac{g}{h} \to 0, \quad Nh^3\left(\frac{g}{h}\right)^4 \to \infty, \quad \frac{\lambda^2}{min(h, g/h)} \to 0,$$

we have uniformly for all i satisfying (6)

i)

$$u_i = \frac{1}{(1+\lambda)^2}\big\{(1+2\lambda)\hat{\mu}(x_i,h) + \lambda\ \text{bias}\ \ \hat{\mu}(x_i,h)\big\} + o_p\big(\frac{\lambda}{\sqrt{Nh}}\big) + O_p\big(\frac{1}{Nh^2}\big)$$

ii)

$$\begin{aligned} \text{bias}\ u_i &= E\ u_i - \mu(x_i) \\ &= \frac{(1+3\lambda)}{(1+\lambda)^2}\frac{h^2}{2}\mu''(x_i)V_K - \frac{\lambda^2}{(1+\lambda)^2}\mu(x_i) + O(h^{2+\beta}) + o\big(\frac{\lambda}{\sqrt{Nh}}\big) + O\big(\frac{1}{Nh^2}\big) \end{aligned}$$

iii)

$$\begin{aligned} \text{var}\ u_i &= \mathrm{E}(u_i - \mathrm{E}u_i)^2 \\ &= \frac{(1+2\lambda)^2}{(1+\lambda)^4}\frac{\sigma^2}{Nh}Q_K + o(\frac{\lambda}{Nh}) + O(\frac{1}{N\sqrt{Nh^5}}) \end{aligned}$$

iv)

$$\begin{aligned} \text{mse}\ u_i &= E\ (u_i - \mu(x_i))^2 \\ &= \left(\frac{1+3\lambda}{(1+\lambda)^2}\ \frac{h^2}{2}\mu''(x_i)V_K - \frac{\lambda^2}{(1+\lambda)^2}\ \mu(x_i)\right)^2 + \frac{(1+2\lambda)^2}{(1+\lambda)^4}\ \frac{\sigma^2}{Nh}\ Q_K \\ &\quad + o(\frac{1}{Nh}) + O\big(h^{2+\beta}max(h^2,\lambda^2)\big) + o\big(\frac{\lambda}{\sqrt{Nh}}max(h^2,\lambda^2)\big) \end{aligned}$$

Proof: i) follows from looking at the components of the remainder term of (9) and applying our assumptions on the rates of g/h and λ. Then, $R_{N,i}$ is of order $o_p\big(\frac{\lambda}{\sqrt{Nh}}\big) + O_p\big(\frac{1}{Nh^2}\big)$ as all other terms are of smaller order. From i) we get

$$\begin{aligned}
E\ u_i - \mu(x_i) &= \frac{1+3\lambda}{(1+\lambda)^2}\ \text{bias}\,\hat\mu(x_i,h) + \frac{1+2\lambda}{(1+\lambda)^2}\mu(x_i) - \mu(x_i) + o\big(\frac{\lambda}{\sqrt{Nh}}\big) + O\big(\frac{1}{Nh^2}\big)\\
&= \frac{1+3\lambda}{(1+\lambda)^2}\ \text{bias}\,\hat\mu(x_i,h) - \frac{\lambda^2}{(1+\lambda)^2}\mu(x_i) + o\big(\frac{\lambda}{\sqrt{Nh}}\big) + O\big(\frac{1}{Nh^2}\big)\\
&= \frac{1+3\lambda}{(1+\lambda)^2}\ \frac{h^2}{2}\mu''(x_i)V_K - \frac{\lambda^2}{(1+\lambda)^2}\mu(x_i) + r(x_i,h) + o\big(\frac{\lambda}{\sqrt{Nh}}\big) + O\big(\frac{1}{Nh^2}\big)
\end{aligned}$$

with $r(x_i,h) = O(h^{2+\beta})$ by Proposition 2 i).

As the bias of $(\hat\mu(x_i,h))$ is not random and using

$$\hat\mu(x_i,h) - \mathrm{E}\hat\mu(x_i,h) = O_p(\frac{1}{\sqrt{Nh}})$$

by Proposition 2, we have

$$\begin{aligned}
\text{var}\ u_i &= \frac{(1+2\lambda)^2}{(1+\lambda)^4}\ \text{var}\ \hat\mu(x_i,h) + O(\frac{1}{\sqrt{Nh}})\big(o(\frac{\lambda}{\sqrt{Nh}}) + O(\frac{1}{Nh^2})\big)\\
&= \frac{(1+2\lambda)^2}{(1+\lambda)^4}\frac{\sigma^2}{Nh}Q_K + O(\frac{1}{N^2h^3}) + o(\frac{\lambda}{Nh}) + O(\frac{1}{N\sqrt{Nh^5}})\\
&= \frac{(1+2\lambda)^2}{(1+\lambda)^4}\frac{\sigma^2}{Nh}Q_K + o(\frac{\lambda}{Nh}) + O(\frac{1}{N\sqrt{Nh^5}})
\end{aligned}$$

by Proposition 2 ii) and using $Nh \to \infty$.

Finally, we get, using the abbreviations $Q = \sigma^2 Q_K, M_i = \frac{1}{2}\mu''(x_i)V_K, \mu_i = \mu(x_i)$

$$\begin{aligned}
\text{mse}\ u_i &= \text{var}\ u_i + (\text{bias}\ u_i)^2\\
&= \frac{(1+2\lambda)^2}{(1+\lambda)^4}\frac{Q}{Nh} + \frac{1}{(1+\lambda)^4}\Big(M_i(1+3\lambda)h^2 - \mu_i\lambda^2\Big)^2 + \Delta_N
\end{aligned}$$

with remainder

$$\begin{aligned}
\Delta_N &= o(\frac{\lambda}{Nh}) + O(\frac{1}{N\sqrt{Nh^5}})\\
&\quad + O(h^{4+\beta}) + o(\frac{\lambda h^2}{\sqrt{Nh}}) + O(\frac{1}{N}) + O(\lambda^2 h^{2+\beta}) + o(\frac{\lambda^3}{\sqrt{Nh}}) + O(\frac{\lambda^2}{Nh^2})\\
&= o(\frac{1}{Nh}) + O(h^{4+\beta}) + o(\frac{\lambda h^2}{\sqrt{Nh}}) + O(\lambda^2 h^{2+\beta}) + o(\frac{\lambda^3}{\sqrt{Nh}})
\end{aligned}$$

using $\lambda \to 0, Nh^3 \to \infty, \lambda^2 = o(h)$ which follows from our rate assumptions. ∎

3 Applications

Once we have the asymptotic equivalence of the regularized least-squares estimate and the shifted Priestley-Chao estimate given by Theorem 2.1 i), we can exploit it to answer various questions showing up in the practice of curve estimation or, analogously, in the two-dimensional case of image denoising. As an illustration, we discuss here how to choose the order of the tuning parameters h, λ optimally for getting a small mean-squared estimation error, and we derive the asymptotic distribution of the function estimate which allows to quantify the reliability of the estimate or to test if the observed data are a noisy version of a given curve resp. image μ.

First, we use the asymptotic error expansion of Theorem 2.1 iv) for deriving the asymptotically optimal rates for regularization parameter λ and smoothing parameter h and the corresponding optimal rates for the mean-squared error.

Corollary 3 *Under the assumptions of Theorem 2.1, let $\mu''(x_i), \mu(x_i) \neq 0$. Then, the mean-squared error is asymptotically minimized for*

a)

$$h = cN^{-\frac{1}{5+2\beta}} \quad \textit{for some } c > 0, \qquad \frac{\lambda}{h} \to \nu_i > 0 \quad \textit{for } N \to \infty,$$

if $\frac{1}{2} V_K \mu''(x_i)/\mu(x_i) = \nu_i^2 > 0$; in this case, for some $c' > 0$, the minimal asymptotic mean-squared error is for $N \to \infty$ of the order

$$\text{mse } u_i \sim c' \left(\frac{1}{N} \right)^{\frac{4+2\beta}{5+2\beta}}.$$

b)

$$h = cN^{-\frac{1}{5}} \quad \textit{for some } c > 0, \qquad \lambda = o(h),$$

if $\frac{1}{2} V_K \mu''(x_i)/\mu(x_i) < 0$; in this case, for some $c' > 0$, the minimal asymptotic mean-squared error is for $N \to \infty$ of the order

$$\text{mse } u_i \sim c' \left(\frac{1}{N} \right)^{\frac{4}{5}}.$$

Proof: We have to distinguish several cases. Let M_i, μ_i, Q be as in the proof of Theorem 2.1. Then, with remainder Δ_N as in the proof of the theorem, we have

$$\text{mse } u_i = \frac{(1+2\lambda)^2}{(1+\lambda)^4} \frac{Q}{Nh} + \frac{1}{(1+\lambda)^4} \left(M_i (1+3\lambda) h^2 - \mu_i \lambda^2 \right)^2 + \Delta_N$$

The first term on the right-hand side, the asymptotic variance of u_i, decreases in λ and in h. The behaviour of the second term, the squared asymptotic bias of u_i, is not so simple depending on the signs and size of M_i and μ_i.

i) Let us assume $M_i/\mu_i > 0$. We set $\nu_i^2 = M_i/\mu_i$. Then, we can choose λ such that the dominant bias term of Theorem 2.1 ii) vanishes, i.e.

$$(1+3\lambda)M_i h^2 - \lambda^2 \mu_i = 0.$$

We get as the nonnegative solution

$$\lambda_0 = \frac{3}{2}\nu_i^2 h^2 + \frac{1}{2}\sqrt{9\nu_i^4 h^4 + 4\nu_i^2 h^2} = \nu_i h + O(h^2) \tag{10}$$

for $h \to 0$. We first consider the case $\lambda = \lambda_0$ in detail, where we have

$$\text{bias } u_i = r(x_i, h) + o\left(\frac{\lambda_0}{\sqrt{Nh}}\right) + O\left(\frac{1}{Nh^2}\right)$$

with $r(x_i, h) = O(h^{2+\beta})$ as in the proof of Theorem 2.1. Looking again at the remainder term Δ_N in the proof of the Theorem, now using that λ_0 is of the same order as h, we get

$$\text{mse } u_i = \frac{(1+2\lambda_0)^2}{(1+\lambda_0)^4}\,\frac{\sigma^2}{Nh}\,Q_K + r^2(x_i, h) + o\Big(\sqrt{\frac{h^{5+2\beta}}{N}}\Big) + o\left(\frac{1}{Nh}\right), \tag{11}$$

as $Nh^3 \to \infty$. Let us first assume that for some $C > 0$ we have $Nh^{7+2\beta} \le C$ for all N large enough. Then, the third term on the right-hand side of (11) is $o(\frac{1}{Nh})$, and mse u_i is asymptotically of order $c_1 \frac{1}{Nh} + c_2 h^{4+2\beta}$ for suitable constant c_1, c_2. This is minimized if, asymptotically, for some constant c

$$h = cN^{-\frac{1}{5+2\beta}}. \tag{12}$$

The corresponding asymptotic mean-squared error is, for some constant c',

$$\text{mse } u_i = c'\left(\frac{1}{N}\right)^{\frac{4+2\beta}{5+2\beta}}. \tag{13}$$

If, on the other hand, $Nh^{7+2\beta} \ge C$ for all N large enough, then the variance part of (11), which is of order $\frac{1}{Nh} = O(h^{6+2\beta})$, is neglible compared to $r^2(x_i, h) = O(h^{4+2\beta})$, i.e. to the squared bias, which is decreasing in h. Therefore, for minimizing mse u_i, h should be chosen as small as possible, i.e. $Nh^{7+2\beta} = C$. From considering the first case, we already know that this is suboptimal compared to (12).

ii) Let us still assume $M_i/\mu_i > 0$. For $\lambda \le \lambda_0$, the asymptotic mean-squared error increases, as the dominant asymptotic bias term does not vanish any longer and as the asymptotic variance increases for decreasing λ. Therefore, only $\lambda \ge \lambda_0$ can be mse-optimal.

We first consider $\lambda \gg \lambda_0 \sim \nu_i h$ asymptotically, i.e. $h = o(\lambda)$. Then,

$$\begin{aligned}
\text{mse } u_i &= \frac{(1+2\lambda)^2}{(1+\lambda)^4}\frac{Q}{Nh} + \frac{\lambda^4}{(1+\lambda)^4}\mu_i^2 + o\left(\frac{1}{Nh}\right) + o\big(\lambda^2 h^2\big) + o\left(\frac{\lambda^3}{\sqrt{Nh}}\right) \\
&= \frac{(1+2\lambda)^2}{(1+\lambda)^4}\frac{Q}{Nh} + \frac{\lambda^4}{(1+\lambda)^4}\mu_i^2 + o\left(\frac{1}{Nh}\right) + o\big(\lambda^4\big)
\end{aligned}$$

as h is of smaller order than λ and $Nh^3 \to \infty$. The remainder terms are asymptotically neglible to the first two positive terms. So, for minimizing mse u_i asymptotically, we have to minimize

$$\frac{(1+2\lambda)^2}{(1+\lambda)^4}\frac{Q}{Nh} + \frac{\lambda^4}{(1+\lambda)^4}\mu_i^2.$$

As this is decreasing in h, for any λ we have to choose h as large as possible considering the constraint $h = o(\lambda)$. The optimal limiting case is a bandwidth h which is of the same order as the regularization parameter λ, i.e. $\frac{\lambda}{h} \to \gamma$ for $N \to \infty$. From the remark at the beginning of this part of the proof, we only have to consider $\gamma > \nu_i$.

For $\lambda \sim \gamma h$, we get from Theorem 2.1 iv), using that $Nh^3 \to \infty$,

$$\begin{aligned} \text{mse } u_i &= \frac{(1+2\gamma h)^2}{(1+\gamma h)^4}\frac{Q}{Nh} + d_i\frac{h^4}{(1+\gamma h)^4} + o(\frac{1}{Nh}) + o\big(h^4\big) \\ &= \frac{Q}{Nh} + d_i h^4 + o(\frac{1}{Nh}) + o\big(h^4\big) \end{aligned} \tag{14}$$

where $d_i = (M_i - \mu_i\gamma^2)^2 > 0$. Setting the derivative of the nonneglible part of (14) w.r.t. h to 0, we get as the asymptotically optimal rate $h^5 \sim \frac{Q}{4d_iN}$, and as the corresponding asymptotically minimal error mse $u_i = c'' N^{-4/5}$ for some constant $c'' > 0$. This converges slower to 0 than the mean-squared error (13) for $\lambda = \lambda_0$.

For $\mu_i \neq 0$ and $M_i/\mu_i > 0$, we have shown that the asymptotically optimal choice of the regularization parameter is $\lambda = \lambda_0 \sim \nu_i h$ where the dominating part of the asymptotic bias given by Theorem 2.1 ii) vanishes. The corresponding optimal rate for the smoothing parameter h is given by (12).

iii) Now we assume $M_i/\mu_i < 0$. In that case, the dominating part of the bias cannot vanish, and we have

$$\text{mse } u_i = \frac{(1+2\lambda)^2}{(1+\lambda)^4}\frac{Q}{Nh} + \frac{1}{(1+\lambda)^4}\Big(|M_i|(1+3\lambda)h^2 + |\mu_i|\lambda^2\Big)^2 + \Delta_N.$$

Using $(1+\lambda)^{-4} = 1 - 4\lambda + 10\lambda^2 + O(\lambda^3)$ we get

$$\begin{aligned} \text{mse } u_i &= \frac{Q}{Nh} + \Big(|M_i|(1+3\lambda)h^2 + |\mu_i|\lambda^2\Big)^2 \\ &\quad + O(\lambda^5) + o(\frac{1}{Nh}) + o(h^4) + o(h^2\lambda^2) + max(h^2,\lambda^2)o(\frac{\lambda}{\sqrt{Nh}}). \end{aligned}$$

If $\lambda = o(h)$, this reduces to

$$\text{mse } u_i = \frac{Q}{Nh} + M_i^2 h^4 + o(\frac{1}{Nh}) + o(h^4)$$

using again $Nh^3 \to \infty$. This is asymptotically minimized for

$$h^5 = \frac{Q}{4M_i^2 N}, \tag{15}$$

and the corresponding optimal rate of mean-squared error is

$$\text{mse } u_i = \frac{5}{4}\big(4Q^4 M_i^2\big)^{1/5} N^{-4/5} \tag{16}$$

If, on the other hand, $\lambda > 0, h = O(\lambda)$, then the asymptotic mean-squared error reduces to

$$\text{mse } u_i = \frac{Q}{Nh} + c_i^2 \lambda^4 + o(\frac{1}{Nh}) + o(\lambda^4), \tag{17}$$

where $c_i^2 = \mu_i^2$ if $h = o(\lambda)$, and $c_i^2 = (\gamma^2 |M_i| + |\mu_i|)^2$ if $h \sim \gamma\lambda$ for some $\gamma > 0$. As the dominant variance term of (17) decreases in h for given λ, we have to choose h as large as possible to minimize mse u_i asymptotically under the constraint $h = O(\lambda)$. Therefore, the optimal rate of bandwidth is $h \sim \gamma\lambda$ for some $\gamma > 0$. Then, (17) is minimized for $h^5 = \frac{Q\gamma^4}{4c_i^2 N}$, and the corresponding optimal rate of mse is

$$\text{mse } u_i = \frac{5}{4}(4Q^4 N^{-4} \gamma^{-4} c_i^2)^{1/5} = \frac{5}{4}\left(4Q^4 (|M_i| + \frac{1}{\gamma^2}|\mu_i|)^2\right)^{1/5} N^{-4/5}$$

This decreases for $\gamma \to \infty$, and the limit is given by (16). Therefore, for $\mu_i \neq 0$ and $M_i/\mu_i < 0$, the mean-squared error mse u_i is asymptotically minimized for h given by (15) and for $\lambda = o(h)$. ■

The corollary implies that we have to distinguish two cases depending on the signs of $\mu''(x_i)$ and $\mu(x_i)$. If both signs coincide, we can choose the tuning parameters h, λ such that the estimation error $u_i - \mu(x_i)$ converges to 0 with rate $N^{-(2+\beta)/(5+2\beta)}$ for $N \to \infty$. By the results of Stone [7], [8], this is the best possible rate which any nonparametric estimate can achieve for functions of a one-dimensional argument x satisfying the regularity assumptions (A2), provided that the distribution of errors ε_i is, e.g., Gaussian. In this case, the additional regularization improves the error rate similar to replacing the nonnegative kernel K by a higher-order kernel (compare, e.g., section 4.6 of Härdle [2]). On the other hand, if $\mu''(x_i), \mu(x_i)$ have different signs, the estimation error converges to 0 only like $N^{-2/5}$ which would be the best possible rate for functions μ satisfying only assumption (A2), a).

Compared to the pure smoothing estimate with kernel K, the additional regularization serves in the first case mainly as a device for reducing the bias. For appropriately chosen λ, the dominating bias term can be made to vanish. In the second case, we have to require $\lambda = o(h)$ to avoid an unnecessarily large estimation error. From a purely asymptotic point of view, we can omit regularization at all, i.e. choose $\lambda = 0$. For finite sample size N, however, choosing $\lambda > 0$ may be useful, as regularization does not only influence the bias, but it also has a variance reducing effect which only turns out to be asymptotically of smaller order. This can be seen from the factor of the dominant variance term in Theorem 2.1 iii) which is a decreasing function of λ and, therefore, always less than 1 except for the boundary case $\lambda = 0$.

If we are not only interested in point estimates u_i for the function values $\mu(x_i)$, but also in confidence intervals which provide information about the reliability of the estimates u_i, we need an asymptotically valid approximation of the probability distribution of u_i. For that purpose, we prove asymptotic normality of our function estimates which may also be used for constructing hypothesis test for comparing functions. As the grid points x_i, where we observe data and where we calculate estimates u_i, depend on N, we have to extend our method to estimates $u(x)$ of the function of interest $\mu(x)$ at an arbitrary fixed location x. We may interpolate the estimates u_i at $x_i, i = 1, \ldots, N$, smoothly by, e.g., splines, but let us first consider the case where we just define

$$u(x) = u_i \quad \text{for } x_{i-1} < x \leq x_i, i = 1, \ldots, N,$$

with $x_0 = 0$.

Theorem 3.1 *Under the assumptions of Theorem 2.1 we have*

i)

$$\frac{(1+\lambda)^2}{1+2\lambda}\sqrt{Nh}\Big(u(x) - \mathrm{E}\ u(x)\Big) \to_d \mathcal{N}(0, \sigma^2 Q_K) \quad \textit{for } N \to \infty.$$

ii)

$$\text{bias } u(x) = \mathrm{E}\ u(x) - \mu(x) = \frac{(1+3\lambda)}{(1+\lambda)^2}\frac{h^2}{2}\mu''(x)V_K - \frac{\lambda^2}{(1+\lambda)^2}\mu(x) + R_N$$

with remainder $R_N = O(h^{2+\beta}) + o\big(\frac{\lambda}{\sqrt{Nh}}\big) + O\big(\frac{1}{Nh^2}\big)$ *uniformly in* $\max(h, g) + g \leq x \leq 1 - \max(h, g) - g - \frac{1}{N}$.

Proof: For given x, we choose the sequence $i(N)$ depending on N such that $x_{i(N)-1} < x \leq x_{i(N)}$ for all N. Then, $u(x) = u_{i(N)}$. We remark that $x_{i(N)}$ satisfies (6) if $\max(h, g) + g \leq x \leq 1 - \max(h, g) - g - \frac{1}{N}$.

i) Using (1), the centered Priestley-Chao estimate $\hat{\mu}(x_{i(N)}, h) - E\ \hat{\mu}(x_{i(N)}, h)$ is the sample mean of the zero mean random variables $Y_{Nj} = \varepsilon_j K_h(x_{i(N)} - x_j), j = 1, \ldots, N$. Using assumption (A1) on the kernel K, it is easy to check that these random variables satisfy the Lindeberg condition, and we get by the Lindeberg-Feller central limit theorem for triangular arrays of random variables (compare section 1.9.3 of Serfling [5]) that

$$\hat{\mu}(x_{i(N)}, h) - E\ \hat{\mu}(x_{i(N)}, h) = \frac{1}{N}\sum_{j=1}^{N} Y_{Nj} \tag{18}$$

is asymptotically normal with mean 0 and variance

$$B_N^2 = \mathrm{var}\big(\ \hat{\mu}(x_{i(N)}, h)\ \big) = \frac{\sigma^2}{Nh}Q_K + O(\frac{1}{N^2h^3})$$

uniformly in $x \in [h, 1 - h - \frac{1}{N}]$ by part a) of the proof of Proposition 2. Multiplying the left hand-side of (19) with $\sqrt{Nh}$ and using Slutsky's Theorem (compare section 1.5.4 of Serfling [5]), we get a sequence of random variables with the non-degenerate limit distribution $\mathcal{N}(0, \sigma^2 Q_K)$. By Theorem 2.1 i), using $\lambda \to 0$ and, again, Slutsky's Theorem, we get

$$\begin{aligned}
\frac{(1+\lambda)^2}{1+2\lambda}\sqrt{Nh}\big(u(x) - E\ u(x)\big) &= \frac{(1+\lambda)^2}{1+2\lambda}\sqrt{Nh}\big(u_{i(N)} - E\ u_{i(N)}\big) \\
&= \sqrt{Nh}\big(\hat{\mu}(x_{i(N)}, h) - E\ \hat{\mu}(x_{i(N)}, h)\big) + o_p(\lambda) + O\big(\frac{1}{\sqrt{Nh^3}}\big) \\
&\to_d \mathcal{N}(0, \sigma^2 Q_K),
\end{aligned}$$

ii) By assumption (A2), we have $\mu(x_{i(N)}) - \mu(x) = O(\frac{1}{N}), \mu''(x_{i(N)}) - \mu''(x) = O(\frac{1}{N^\beta})$ uniformly in $x \in (0,1)$. Therefore, by Theorem 2.1 ii)

$$\begin{aligned}
\text{bias } u(x) &= \text{bias } u_{i(N)} + \mu(x_{i(N)}) - \mu(x) \\
&= \text{bias } u_{i(N)} + O(\frac{1}{N}) \\
&= \frac{1+3\lambda}{(1+\lambda)^2}\frac{h^2}{2}\mu''(x_{i(N)})V_K - \frac{\lambda^2}{(1+\lambda)^2}\mu(x_{i(N)}) + O(h^{2+\beta}) + o(\frac{\lambda}{\sqrt{Nh}}) + O(\frac{1}{Nh^2}) \\
&= \frac{1+3\lambda}{(1+\lambda)^2}\frac{h^2}{2}\mu''(x)V_K - \frac{\lambda^2}{(1+\lambda)^2}\mu(x) \\
&\quad + O(\frac{h^2}{N^\beta}) + O(\frac{\lambda^2}{N}) + O(h^{2+\beta}) + o(\frac{\lambda}{\sqrt{Nh}}) + O(\frac{1}{Nh^2}),
\end{aligned}$$

and ii) follows. ■

Let x_i satisfy (6). If $\mu''(x_i)/\mu(x_i) < 0$, we have from Corollary 3 a) that the asymptotically optimal rates of tuning parameters are $h = cN^{-\frac{1}{5}}, \lambda = o(h)$. In this case, in the limit for $N \to \infty$, the error $u_i - \mu(x_i)$ is distributed as

$$\mathcal{N}\left(\frac{(1+3\lambda)}{(1+\lambda)^2}\frac{h^2}{2}\mu''(x_i)V_K - \frac{\lambda^2}{(1+\lambda)^2}\mu(x_i), \frac{(1+2\lambda)^2}{(1+\lambda^2)^2}\frac{\sigma^2 Q_K}{Nh}\right) \sim \mathcal{N}\left(\frac{h^2}{2}\mu''(x_i)V_K, \frac{\sigma^2 Q_K}{Nh}\right).$$

This is to first order of approximation the same asymptotic behaviour as for the Priestley-Chao estimate without regularization. If, on the other hand, $\frac{1}{2}V_K\mu''(x)/\mu(x) = \nu^2(x) > 0$, then the optimal rates are $h = cN^{-\frac{1}{5+2\beta}}, \lambda = \lambda_0 \sim \nu_i h$. For that choice of order, we get as approximative error distribution

$$\mathcal{N}\left(R_N, \frac{(1+2\lambda_0)^2}{(1+\lambda_0^2)^2}\frac{\sigma^2 Q_K}{Nh}\right) \sim \mathcal{N}\left(O(h^{2+\beta}), \frac{\sigma^2 Q_K}{Nh}\right).$$

The squared bias and the variance are of the same order $h^{4+2\beta} = \frac{1}{Nh}$ here. If, on the other hand, we choose $h = cN^{-\frac{1}{5}}, \lambda = \lambda_0 \sim \nu_i h$ in that case, the bias is asymptotically neglible, and we get as approximative error distribution

$$\mathcal{N}\left(0, \frac{\sigma^2 Q_K}{Nh}\right).$$

Finally, we consider linear interpolation of the estimates u_i at $x_i, i = 1, \ldots, N$, as an alternative. Analogously, we could as well use, e.g., higher-order splines for an even more smooth interpolation. Here, we define for $x \geq x_1$

$$u(x_i) = u_i,\ i = 1, \ldots, N, \qquad u(x) = (1-\theta)u_{i-1} + \theta u_i \quad \text{for } x_{i-1} \leq x \leq x_i,\ i = 1, \ldots, N,$$

with $0 \leq \theta = N(x - x_{i-1}) \leq 1$ depending on x and N. Then, we get the following result analogous to Theorem 3.1:

Theorem 3.2 *Under the assumptions of Theorem 2.1 we have*

i)

$$\frac{(1+\lambda)^2}{1+2\lambda}\sqrt{Nh}\left(u(x) - \mathrm{E}\ u(x)\right) \to_d \mathcal{N}(0, \sigma^2 Q_K) \quad \textit{for } N \to \infty.$$

ii)

$$\text{bias } u(x) = \text{E } u(x) - \mu(x) = \frac{(1+3\lambda)}{(1+\lambda)^2}\frac{h^2}{2}\mu''(x)V_K - \frac{\lambda^2}{(1+\lambda)^2}\mu(x) + R_N$$

with remainder $R_N = O(h^{2+\beta}) + o\big(\frac{\lambda}{\sqrt{Nh}}\big) + O\big(\frac{1}{Nh^2}\big)$ *uniformly in* $\max(h,g) + g \le x \le 1 - \max(h,g) - g - \frac{1}{N}$.

Proof: For given x, we choose the sequence $i(N)$ depending on N such that $x_{i(N)-1} < x \le x_{i(N)}$ for all N. Then, $u(x) = (1-\theta)u_{i(N)-1} + \theta u_{i(N)}$. We remark that $x_{i(N)-1}, x_{i(N)}$ satisfy (6) if $\max(h,g) + g \le x \le 1 - \max(h,g) - g - \frac{1}{N}$.

a) Using (1), the centered Priestley-Chao estimate $\hat{\mu}(x_{i(N)}, h) - E\ \hat{\mu}(x_{i(N)}, h)$ is the sample mean of the zero mean random variables $Y_{Nj} = \varepsilon_j K_h(x_{i(N)} - x_j), j = 1, \ldots, N$. Using assumption (A1) on the kernel K, it is easy to check that these random variables satisfy the Lindeberg condition, and we get by the Lindeberg-Feller central limit theorem for triangular arrays of random variables (compare section 1.9.3 of Serfling [5]) that

$$\hat{\mu}(x_{i(N)}, h) - E\ \hat{\mu}(x_{i(N)}, h) = \frac{1}{N}\sum_{j=1}^{N} Y_{Nj} \tag{19}$$

is asymptotically normal with mean 0 and variance

$$B_N^2 = \text{var}\big(\ \hat{\mu}(x_{i(N)}, h)\ \big) = \frac{\sigma^2}{Nh}Q_K + O(\frac{1}{N^2h^3})$$

uniformly in $x \in [h, 1 - h - \frac{1}{N}]$ by part a) of the proof of Proposition 2. Multiplying the left hand-side of (19) with $\sqrt{Nh}$ and using Slutsky's Theorem (compare section 1.5.4 of Serfling [5]), we get a sequence of random variables with the non-degenerate limit distribution $\mathcal{N}(0, \sigma^2 Q_K)$. The same result, of course, holds for $x_{i(N)-1}$ instead of $x_{i(N)}$.

b) By Theorem 2.1 i), using $\lambda \to 0$ and, again, Slutsky's Theorem, we get

$$\begin{aligned}
&\frac{(1+\lambda)^2}{1+2\lambda}\sqrt{Nh}\,\big(u(x) - E\ u(x)\big) \\
= \;& \frac{(1+\lambda)^2}{1+2\lambda}\sqrt{Nh}\,\big((1-\theta)(u_{i(N)-1} - E\ u_{i(N)-1}) + \theta(u_{i(N)} - E\ u_{i(N)})\big) \\
= \;& \sqrt{Nh}\big((1-\theta)(\hat{\mu}(x_{i(N)-1}, h) - E\ \hat{\mu}(x_{i(N)-1}, h)) + \theta(\hat{\mu}(x_{i(N)}, h) - E\ \hat{\mu}(x_{i(N)}, h))\big) \\
& + o_p(\lambda) + O(\frac{1}{\sqrt{Nh^3}}) \\
\to_d \;& \mathcal{N}(0, v),
\end{aligned}$$

as the sum of two asymptotically normal random variables is again asymptotically normal. We only have to calculate the asymptotic variance v, i.e. the limit of

$$\begin{aligned}
& Nh\ \text{var}\big((1-\theta)\hat{\mu}(x_{i(N)-1}, h) + \theta\hat{\mu}(x_{i(N)}, h)\big) \\
= \;& Nh\big((1-\theta)^2\text{var}\ \hat{\mu}(x_{i(N)-1}, h) + \theta^2\text{var}\ \hat{\mu}(x_{i(N)}, h) + 2\ \text{cov}(\hat{\mu}(x_{i(N)-1}, h), \hat{\mu}(x_{i(N)}, h))\big)
\end{aligned}$$

The first two terms coincide asymptotically with $(1-\theta)^2\sigma^2 Q_K + \theta^2\sigma^2 Q_K$ by a). For the third term, we use that by Proposition 2 iv) and as $x_{i(N)} - x_{i(N)-1} = \frac{1}{N}$

$$\frac{\sigma^2}{Nh}\int K(z)K(z + \frac{1}{Nh})dz + O(\frac{1}{N^2h^3}) = \frac{\sigma^2}{Nh}\ Q_K + O(\frac{1}{N^2h^2}) + O(\frac{1}{N^2h^3})$$

by assumption (A1). We conclude

$$v = \lim_{N\to\infty} \big((1-\theta)^2\sigma^2 Q_K + \theta^2\sigma^2 Q_K + 2(1-\theta)\theta\sigma^2 Q_K\big) = \sigma^2 Q_K,$$

and i) follows.

c) By assumption (A2), we have $\mu(x_{i(N)}) - \mu(x) = O(\frac{1}{N}), \mu''(x_{i(N)}) - \mu''(x) = O(\frac{1}{N^\beta})$ and correspondingly for $x_{i(N)-1}$ uniformly in $x \in (0,1)$. Therefore,

$$\begin{aligned} \text{bias } u(x) &= (1-\theta) \text{ bias } u_{i(N)-1} + \theta \text{ bias } u_{i(N)} + (1-\theta)\ \mu(x_{i(N)-1}) + \theta\ \mu(x_{i(N)}) - \mu(x) \\ &= (1-\theta) \text{ bias } u_{i(N)-1} + \theta \text{ bias } u_{i(N)} + O(\frac{1}{N}) \end{aligned}$$

Moreover, by Theorem 2.1 ii),

$$\begin{aligned} \text{bias } u_{i(N)} &= \frac{1+3\lambda}{(1+\lambda)^2}\,\frac{h^2}{2}\mu''(x_{i(N)})V_K - \frac{\lambda^2}{(1+\lambda)^2}\mu(x_{i(N)}) + O(h^{2+\beta}) + o\big(\frac{\lambda}{\sqrt{Nh}}\big) + O\big(\frac{1}{Nh^2}\big) \\ &= \frac{1+3\lambda}{(1+\lambda)^2}\,\frac{h^2}{2}\mu''(x)V_K - \frac{\lambda^2}{(1+\lambda)^2}\mu(x) \\ &\quad + O(\frac{h^2}{N^\beta}) + O(\frac{\lambda^2}{N}) + O(h^{2+\beta}) + o\big(\frac{\lambda}{\sqrt{Nh}}\big) + O\big(\frac{1}{Nh^2}\big), \end{aligned}$$

and analogously for $x_{i(N)-1}$. We conclude

$$\text{bias } u(x) = \frac{1+3\lambda}{(1+\lambda)^2}\,\frac{h^2}{2}\mu''(x)V_K - \frac{\lambda^2}{(1+\lambda)^2}\mu(x) + \ O(h^{2+\beta}) + o\big(\frac{\lambda}{\sqrt{Nh}}\big) + O\big(\frac{1}{Nh^2}\big).$$

■

Acknowledgements
The work was supported by the Deutsche Forschungsgemeinschaft (DFG) as part of the priority research program 1114 *Mathematical Methods of Time Series and Digital Image Analysis* as well as by the center of excellence *Dependable Adaptive Systems and Mathematical Modelling* funded by the state of Rhineland-Palatinate.

References

[1] J. Franke and W. Härdle. On bootstrapping kernel spectrum estimates. *Ann. Statist.* 20: 121–145, 1992.

[2] W. Härdle. *Applied Nonparametric Regression.* Cambridge University Press, Cambridge, 1990.

[3] P. Mrázek, J. Weickert and A. Bruhn. On robust estimation and smoothing with spatial and tonal kernels. In R. Klette, R. Kozera, L. Noakes, J. Weickert eds., *Geometric Properties from Incomplete Data.* Springer, Berlin-Heidelberg-New York, 2006.

[4] M. B. Priestley and M. T. Chao. Nonparametric function fitting. *J. Royal Stat. Soc., Ser. B,* 34: 385–392, 1972.

[5] R. J. Serfling. *Approximation Theorems of Mathematical Statistics*. Wiley, New York, 1980.

[6] B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London, 1986.

[7] C. J. Stone. Optimal global rates of convergence for nonparametric estimators. *Ann. Statist.* 8: 1348–1360, 1980.

[8] C. J. Stone. Optimal rates of convergence for nonparametric regression. *Ann. Statist.* 10: 1040–1053, 1982.