8-2009

Local Adaptive Smoothing in Kernel Regression Estimation

Qi Zheng
Clemson University, qiz@clemson.edu

Local Adaptive Smoothing in Kernel Regression Estimation

A Master's Thesis
Presented to
the Graduate School of
Clemson University

In Partial Fulfillment
of the Requirements for the Degree
Master of Science
Mathematical Science

---

by
Qi Zheng
August 2009

Accepted by:
Dr. Karunarathna B Kulasekera, Committee Co-Chair
Dr. Colin Gallagher, Committee Co-Chair
Dr. Chanseok Park

Abstract

We consider nonparametric estimation of a smooth regression function of one variable. In practice it is quite popular to use the data to select one global smoothing parameter. Such global selection procedures cannot sufficiently account for local sparseness of the covariate nor can they adapt to local curvature of the regression function. We propose a new method to select local smoothing parameters which takes into account sparseness and adapts to local curvature of the regression function. A Bayesian method allows the smoothing parameter to adapt to the local sparseness of the covariate and provides the basis for a local cross validation procedure which adjusts smoothing according to local curvature of the regression function. Simulation evidence indicates that the method can result in significant reduction of both point-wise mean squared error and integrated mean squared error of the estimators.

Dedication

I dedicate this work to my loving parents. It is their love and support that made this work a complete one.

Acknowledgments

I would like to express my heartiest gratitude to my co-advisors Dr.K.B.Kulasekera and Dr. C. Gallagher for their guidance and support. Their mathematical insights inspired me and helped me making this a success. My sincere thanks to Dr.C. Park for his insightful suggestions.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

The kernel method for regression estimation has become a standard statistical technique in many areas of research. It is well known that the performance of kernel methods depends crucially on the smoothing parameter (the bandwidth). This fact motivated data-driven bandwidth procedures appearing in the literature. Among these, we mention three: least square cross validation (LSCV) proposed by Craven and Wahba(1979); plug-in methods(e.g., see Hardle and Bowman (1988)); and AIC based methods (e.g., see Akaike (1973)).

However, in many situations a single fixed bandwidth does not generally provide a good estimator of a regression function. One reason for this is that the amount of smoothing needed for estimating a function where the data is dense is different from that required to estimate a function where the data is sparse. Besides that, a bandwidth chosen to optimally estimate the function at or near its local mode may differ from the bandwidth that is required to get an accurate estimate of the regression function where the function is flat. In order to overcome the disadvantages of a fixed bandwidth it is necessary to allow the bandwidth to somehow adapt locally. Apart from the direct plug-in method which requires the knowledge of the unknown mean function, very few data based methods to select local bandwidths have been proposed. Fan *et al* (1996) proposed an adaptive procedure to get local bandwidths. Their technique uses a plug-in approach coupled with an arbitrary approximation scheme to develop a bandwidth as a function of the point of estimation. However, as the authors mentioned, their procedure does not work well for moderate sample sizes.

To allow a smoothing method to adapt to sparseness in the data, the choice of a local smoothing parameter should depend on the design density of the covariate. The local sparseness of the covariate would be reflected by a data based local bandwidth for a design density estimator. Gangopadhyay and Cheung (2002) and Kulasekera and Padgett (2006) proposed adaptive Bayesian bandwidth selection for density estimation. They treat the bandwidth as a scale parameter, and use a prior distribution for that to compensate for the lack of information in moderate to small sample sizes. As a result, their approaches perform well for moderate sample sizes.

In this article, we combine Bayesian estimation of a bandwidth for the design density with a local cross-validation to get local bandwidths that enhance nonparametric regression estimator performance. We describe our procedures for local linear regression, but the method can be adapted to other smoothed estimators of regression functions. We show that judicious choices of prior parameters lead to local bandwidths tending to zero as sample size increases, and thus the estimator using bandwidths from our procedure is consistent.

This article is structured as follows. In section 2 we develop the Bayesian cross-validation procedure. Results obtained from a simulation study following by a real data example are provided in Section 3. All the proofs are presented in the Appendix.

Chapter 2

Adaptive bandwidth selection for kernel estimation of regression

We describe the development of the Bayesian cross-validation approach for kernel regression estimators and discuss asymptotic properties of local smoothing parameters from our procedure.

We begin by developing some notation. Let $(X_i, Y_i), i = 1, 2, ..., n$ be n independent and identically distributed (i.i.d.) bivariate observations with a joint density function $f(\cdot, \cdot)$. Let the marginal density of X be $f(\cdot)$. Let K be a classical second order kernel function, that is

(1) $\int_{-\infty}^{\infty} K(u)du = 1$

(2) $\int_{-\infty}^{\infty} uK(u)du = 0$

(3) $M_2 = \int_{-\infty}^{\infty} u^2 K(u)du \neq 0$

(4) $V = \int_{-\infty}^{\infty} K(u)^2 du < \infty$

Our goal is to estimate the regression function, which is the conditional expectation

$$m(x) = E[Y|X = x] \tag{2.1}$$

(assuming $f(x) \neq 0$). Then the model can be written as :

$$Y = m(X) + \epsilon$$

where $E[\epsilon|X]=0, V[\epsilon|X]=\sigma^2$. In this article, we consider the local linear estimator of $m(x)$ given by

$$\hat{m}_h(x)=\frac{1}{nh(x)}\sum_{i=1}^{n}K\left(\frac{x-X_i}{h(x)}\right)\frac{M_{2n}(x)-(\frac{x-X_i}{h(x)})M_{1n}(x)}{M_{2n}(x)M_{0n}(x)-M_{1n}^2(x)}Y_i \tag{2.2}$$

where $M_{jn}(x)=(h(x)n)^{-1}\sum_{i=1}^{n}K\left(\frac{x-X_i}{h(x)}\right)\left(\frac{x-X_i}{h(x)}\right)^j, j=0,1,2$; and $h(x)$ is the bandwidth to be selected at x.

Our aim is to select a bandwidth that depends on the data. Noticing that the amount of smoothing when X is dense is different from that when X is sparse, we initially get an adaptive smoothing window based on the X observations, using a Bayesian approach (see Kulasekera and Padgett, 2006). This window will be used to conduct a local cross-validation resulting in $h(x)$ in (2). The reason for a local cross-validation is to accommodate different mean functions with the same design density for X.

2.1 Bayesian Smoothing window

Here we develop an initial bandwidth which adapts to the local sparseness of data. Define

$$f_h(x)=f*K_h(x)=\int f(u)K_h(x-u)du=E[K_h(X-x)] \tag{2.3}$$

where $K_h(x)=\frac{1}{h}K(\frac{x}{h})$

Now, considering h as a parameter for f_h, for a prior density $\pi(h)$, the posterior distribution of h is

$$\pi(h|x)=\frac{f_h(x)\pi(h)}{\int f_h(x)\pi(h)dh} \tag{2.4}$$

Since $f_h(x)$ is unknown, we can not compute $\pi(h|x)$ directly. However, it is natural to use the sample mean

$$\hat{f}_h(x)=\frac{1}{n}\sum_{i=1}^{n}K_h(X_i-x)$$

to estimate $f_h(x)$. Substituting this in (4), we get

$$\hat{\pi}(h|X_1,...,X_n,x)=\frac{\hat{f}_h(x)\pi(h)}{\int \hat{f}_h(x)\pi(h)dh} \tag{2.5}$$

Then, for the squared-error loss, the best local bandwidth $h = h(x)$ is given by the posterior mean

$$h^*(x) = \int h\hat{\pi}(h|X_1, ..., X_n, x)dh \tag{2.6}$$

Note that with this approach, the posterior is a function of h only and, with a well-selected prior and a kernel, $\pi(h|X_1, ...X_n, x)$ and $h^*(x)$ can be explicitly obtained. Although this approach works in principal for any kernel and the most suitable prior (preferably a conjugate prior), for the remainder of this article we shall use a normal kernel and an inverted-gamma prior due to the algebraic simplicity. In particular for a normal kernel

$$K(u) = \frac{1}{\sqrt{2\pi}}e^{-u^2/2}, -\infty < u < \infty$$

coupled with an inverted-Gamma prior,

$$\pi(h) = \frac{2}{\Gamma(\alpha)\beta^\alpha h^{2\alpha+1}}\exp\{\frac{-1}{\beta h^2}\}, \quad \alpha > 0, \quad \beta > 0, \quad h > 0$$

we get

$$\hat{\pi}(h|X_1, ..., X_n, x) = \frac{\sum_{i=1}^n (1/h^{2\alpha+2})\exp\{-(1/h^2)((X_i - x)^2/2 + 1/\beta)\}}{\sum_{i=1}^n (\Gamma(\alpha + 1/2)/2)\{(X_i - x)^2/2 + 1/\beta\}}$$

resulting in

$$h^*(x) = \frac{\Gamma(\alpha)}{\sqrt{2\beta}\Gamma(\alpha + 1/2)} \frac{\sum_{i=1}^n \{1/(\beta(X_i - x)^2 + 2)\}^\alpha}{\sum_{i=1}^n \{1/(\beta(X_i - x)^2 + 2)\}^{\alpha+1/2}} \tag{2.7}$$

The following theorem shows that if one picks the prior parameters in a suitable manner(fix α and let β diverge with a proper rate as sample size increases) one can get a sequence of $h^*(x)$ that converges to zero almost surely, for every x.

Theorem 2.1.1 *Let $f(\cdot)$ be continuous and bounded away from 0 at x . If $\alpha > 0$ is fixed, $\beta \to \infty$, and $n\beta^{-\frac{1}{2}} \to \infty$, as $n \to \infty$ then*

$$h^*(x) \to 0$$

and

$$nh^*(x) \to \infty$$

with probability one, as $n \to \infty$

The proof of this Theorem is given in the Appendix.

Remark 2.1.1 *The implementations of the Bayesian procedure requires a specification of parameters α, and β for the prior distribution. From the proof of Theorem 2.1 in Appendix A, we can see that the asymptotic rate of the Bayesian bandwidth $h^*(x)$ is between $[\beta^{-\frac{1}{2}}, \beta^{-\frac{\alpha}{4\alpha+2}}]$. Thus, if we let α be sufficiently large and choose β between $n^{\frac{2}{5}}$ and $n^{\frac{4}{5}}$, the Bayesian bandwidths converge to 0 with rate close to that of the mean squared optimal bandwidths.*

2.2 Local cross-validation

The above Bayesian bandwidth only considers the influence from the distribution of X. However, in kernel regression estimation, the bandwidths must reflect the impact of the responses. This motivates the following modifications of the Bayesian bandwidths to obtain the final bandwidths to estimate the regression function.

Since $m(\cdot)$ is continuous and our object is to estimate $m(\cdot)$ locally at x, we argue that only the segment of observations in the neighborhood of x is vitally important. In particular, for estimating m at x, we propose to conduct a cross-validation using observations falling in $I_x = [x - h^*(x), x + h^*(x)]$ only. The version of local CV in this context is developed as follows. First let

$$\hat{m}_{-i}(X_i) = \frac{1}{(n-1)h} \sum_{j \neq i} K\left(\frac{X_i - X_j}{h}\right) \frac{M_{2i}(X_i) - (\frac{X_i - X_j}{h})M_{1i}(X_i)}{M_{2i}(X_i)M_{0i}(X_i) - M_{1i}^2(X_i)} Y_j$$

$$\text{and} \quad M_{ki}(X_i) = ((n-1)h)^{-1} \sum_{j \neq i} K\left(\frac{X_i - X_j}{h}\right)\left(\frac{X_i - X_j}{h}\right)^k, k = 1, 2, 3$$

Now, let $l(x)$ denotes the number of covariate values falling in I_x and let $(X_i^{'}, Y_i^{'}), i = 1, .., l(x)$ denotes the corresponding observations. Now we define the locally cross-validation bandwidth $h^{**}(x)$ by

$$h^{**}(x) = \arg\left\{\min_h \sum_{i=1}^{l(x)} \left(Y_i^{'} - \hat{m}_{-i}(X_i^{'})\right)^2\right\} \tag{2.8}$$

Remark 2.2.1 *Note that the bandwidth $h^{**}(x)$ is selected from a set of candidate bandwidths when minimizing the leave one out cross validation CV above. In global cross validation procedures this set is usually taken as $[n^{-1+\eta}, n^{-\eta}]$ for some small $\eta > 0$. Such cross validations produce bandwidths of order $n^{-1/5}$ with probability tending to 1 when the sample size is n (see Girard, 1998). Since the* *number of observations contained in the interval I_x is of order $nh^*(x)$ under reasonable conditions on the density of X, h^{**} above will be of order $[nh^*(x)]^{-1/5}$. Thus, the proposed adaptive bandwidth h^{**} can achieve near asymptotic optimal rate for suitably chosen β in the prior.*

Chapter 3

Numerical results

To illustrate the improvement of our Bayesian locally adaptive CV procedure over a few existing methods, we conducted a Monte Carlo simulation. Implementation of our technique is straightforward. However, there are two steps that we need to be careful. The first one is how to choose the parameters α, β for the prior distribution. In our numerical work, we chose $\alpha = 100$ and $\beta = n$ for small and moderate sample sizes. The second is the cross validation. Notice that one needs at least three observations to do the local Cross-Validation for local linear estimators. Thus, in practice, if $l(x)$ is greater than 3, we can do the local cross-validation to modify the Bayesian bandwidth $h^*(x)$. For windows with less than 3 observations, we used the Bayesian bandwidth. Also in practice, cross-validation is generally carried over a sequence of candidate bandwidths $(h_1, ..., h_k)$. In our simulations, the grid of h values for each x is given as $\frac{h^*(x)}{\hat{l}(x)}, \frac{2h^*(x)}{\hat{l}(x)}, ..., h^*(x)$, where $\hat{l}(x) = \left\lfloor \frac{l(x)}{2} \right\rfloor$, which equally spaces the initial Bayesian window according to the number of observations falling in the window.

The performance of local linear estimator using different bandwidth selection procedures is related to the sample size, the distribution of X, the true regression function, and the true standard deviation of the errors. Although only few selected results are reported here for the sake of brevity, the following settings of these factors were examined, with 1000 replications for each set of factor combination.

(a) sample size n= 25(small), 50(moderate)

(b) distributions of X: Uniform(0, 1), Normal(0.5, 0.25)

(c) regression functions

(i) $m(x) = 1 - 48x + 218x^2 - 315x^3 + 145x^4$ (a polynomial regression function with a trend)

(ii) $m(x) = sin(5\pi x)$ (a function without much fine structure)

(iii) $m(x) = 10\exp(-10x)$ (a function with a trend but no fine structure)

(iv) $m(x) = x + 0.5\exp\{-50(x - 0.5)^2\}$ (a function with different degrees of curvature)

(v) $m(x) = 0.3\exp\{-64(x - 0.25)^2\} + 0.7\exp\{-256(x - 0.75)^2)\}$ (a function with noticeably different curvature)

(d) error standard deviation σ=0.2, 0.3

Most of the regression functions above were used in earlier studies (Ruppert *et al.* 1995; Hart and Yi, 1996; Herrmann 1997; Eubank, 1999)–. We used R (Ihaka and Gentlemen, 1996) for all computations. We compare the local linear estimators $\hat{m}_l(x)$ using Bayes type local bandwidth to local linear estimators $\hat{m}_g(x)$ using global CV bandwidth by comparing the estimated MSE (ESMSE) of each estimator based on the simulations. Specifically, we consider logged ratio

$$r(x) = \log\left(\frac{ESMSE(\hat{m}_g(x))}{ESMSE(\hat{m}_l(x))}\right)$$

and the estimated integrated MSE

$$EIMSE(\hat{m}) = \frac{1}{p}\sum_{i=1}^{p} ESMSE(\hat{m}(x_i))$$

over $(0, 1)$, since $(0, 1)$ contains more than 95% observations for both design densities in our simulation study. Here, ESMSE($\hat{m}(x)$) $=\sum_{i=1}^{N}(\hat{m}(x) - m(x))^2/N$ at any given x for any estimator $\hat{m}(x)$ of $m(x)$ with N being the number of simulations, x_i's $i = 1, \ldots, p$ are equally spaced points in $(0, 1)$. We chose $p = 100$ for our simulations.

Table 3.1: EIMSE for $m(x) = 1 - 48x + 218x^2 - 315x^3 + 145x^4, \sigma = 0.3$

n	Design Density	Bayesian Local CV	Global CV
25	Uniform	0.1551595	0.4486760
	Normal	0.8738112	4.5184210
50	Uniform	0.0437532	0.0466022
	Normal	0.2342103	1.1904470

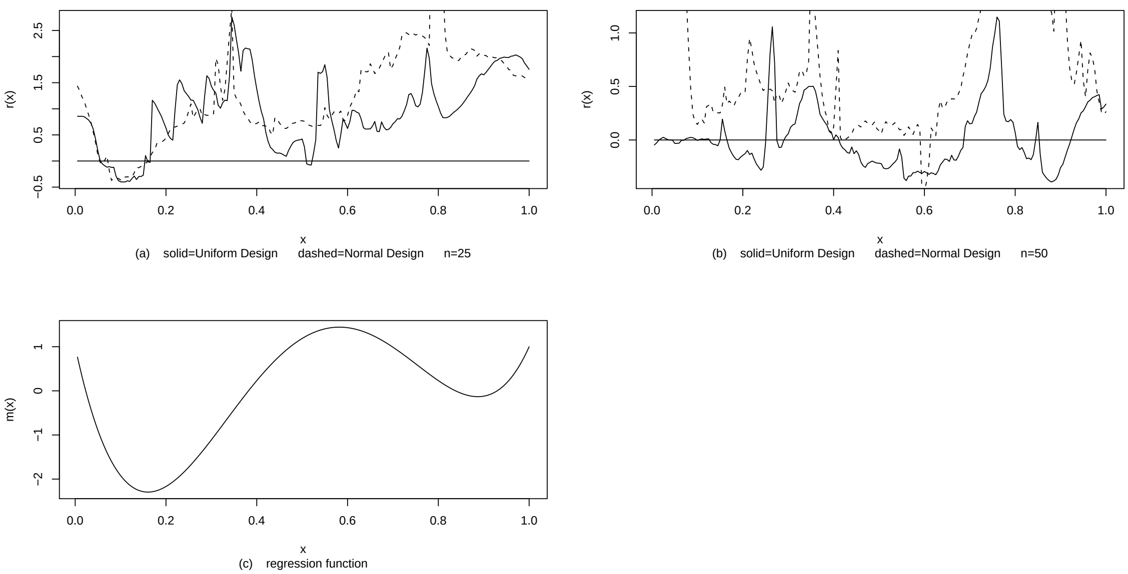

Figure 3.1: $m(x) = 1 - 48x + 218x^2 - 315x^3 + 145x^4, \sigma = 0.3$

Table 3.1-3.2 and Figures 3.1-3.2 give some results of two typical mean functions: (1) $m(x) = 1 - 48x + 218x^2 - 315x^3 + 145x^4$, (2) $m(x) = 0.3\exp\{-64(x - 0.25)^2\} + 0.7\exp\{-256(x - 0.75)^2)\}$. The results with other mean functions and parameter combinations were very similar, and are presented in Appendix B. Figure 3.1 gives the logged ratio $r(x)$ for the regression function $m(x) = 1-48x+218x^2-315x^3+145x^4$ with $\sigma = 0.3$, where the global CV works well for small and moderate sample sizes with the uniform design. As shown in Figure 3.1, our procedure outperforms the global CV over a large proportion of the interval even for the uniform design, when the sample size is small. Although it does not outperform the global CV everywhere for the uniform design when the sample size is moderate, it is still preferable since the integrated MSE (MSE) of $\hat{m}(x)$ was less than that of $\hat{m}_n(x)$ (see Table 3.1). As for the normal design, the Bayesian local CV dominates the global CV nearly everywhere, especially where the data is sparse. Figure 1 has shown that our method not only works well for those regression functions on which the global CV has a good performance; moreover, it gives a much better estimation when the design is not uniform. This phenomenon is simply a manifestation of the fact that our technique takes the sparseness of design into account and adapts to the design density.

Figure 3.2 shows the logged ratio $r(x)$ for the mean function $m(x) = 0.3\exp\{-64(x - 0.25)^2\} + 0.7\exp\{-256(x - 0.75)^2)\}$ with $\sigma = 0.2$. This mean function has two modes and the

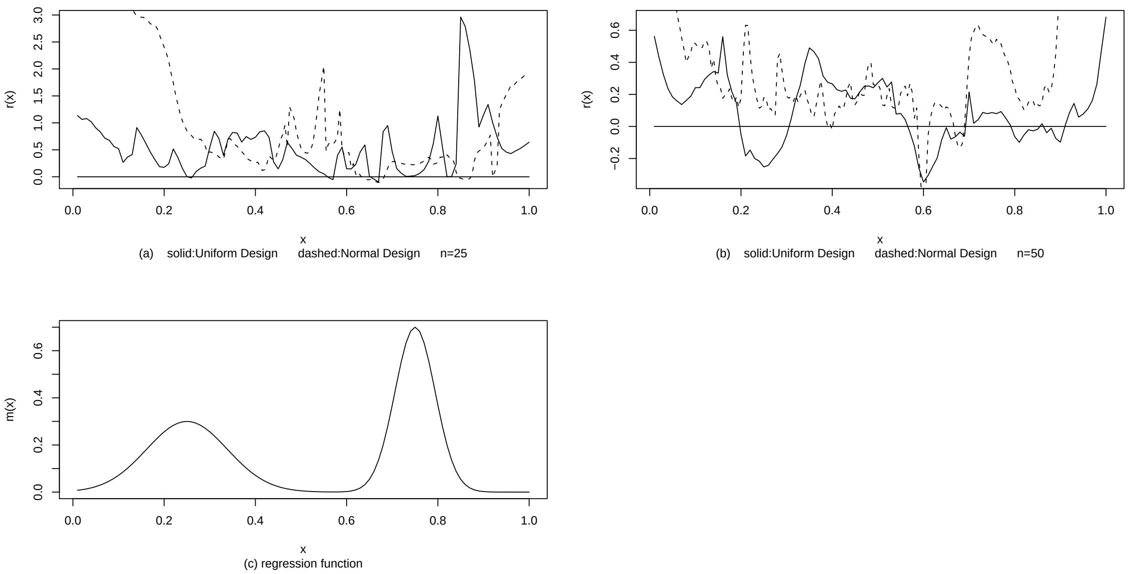

Figure 3.2: $m(x) = 0.3\exp\{-64(x-0.25)^2\} + 0.7\exp\{-256(x-0.75)^2)\}$, $\sigma = 0.2$

Table 3.2: EIMSE for $m(x) = 0.3\exp\{-64(x-0.25)^2\} + 0.7\exp\{-256(x-0.75)^2)\}$, $\sigma = 0.2$

n	Design Density	Bayesian Local CV	Global CV
25	Uniform	0.0283580	0.0521505
	Normal	0.0553061	0.4777479
50	Uniform	0.0135138	0.0153065
	Normal	0.0199792	0.1460045

degrees of its curvature are noticeably different, and hence the global CV performs relatively poorly, especially for areas where design density is sparse. Figure 2 clearly indicates that our Bayesian local cross-validation selection is better than the global CV method for small and moderate sample sizes. This fact is expected since our procedure adapts to the mean function and density design locally. For The improvement of our method is not significant for sample size 200 or larger, since the global CV performs well with a large sample size.

In addition to the above, we examined the performance of our method against the improved AIC method (AIC_c) proposed by Hurvich *et al.* (1998) and the adaptive bandwidth selection mentioned in Fan *et al.* (1996). Fan *et al.* (1996) explicitly discussed the smoothing parameter selection for density estimation only. They however suggested that their method, although not performing well for small to moderate sample sizes, can be extended to regression estimation. Following their approach, we used the same spline approximation to the optimal bandwidth function $h(x)$ where the proposed cross-validation over a large set of cubic spline interpolant of prechosen knot-bandwidth pairs $\{(a_1, h_1), \ldots, (a_p, h_p)\}$(in their notation) was conducted by minimizing $\sum_{i=1}^{n}(\hat{m}_{\hat{h_i}}(X_i) - Y_i)^2$ where $\hat{m}_{\hat{h_i}}(X_i)$ is the estimated regression function for a bandwidth calculated at X_i for each spline function. The knots $(a_1, \ldots a_p)$ for our simulation were chosen to be equispaced on $[0, 1]$ with $p = 6$. We chose four values $\{0.5\hat{h}_G, \hat{h}_G, 1.5\hat{h}_G, 2\hat{h}_G\}$ for each coordinate $h_i, i = 1, \ldots, p$ where $\hat{h}_G$ was the global cross-validation bandwidth.

The MSE performance of local linear estimators with Fan *et al.* type bandwidths and those with AIC$_c$ were both inferior to our proposed method. We just provided the IEMSE for these local linear estimators using these bandwidths in Table 3.

Table 3.3: EIMSE for $m(x) = x + 0.5\exp\{-50(x - 0.5)^2\}$, $\sigma = 0.3$

n	Design Density	Bayesian Local CV	Fan *et al.* type bandwidth	AIC_c
50	Uniform	0.0173432	0.0459627	8.9721330
	Normal	0.0260417	0.0606059	0.8616309

The use of this Bayesian local cross-validation procedure is not restricted to local linear estimator. We examine the Nadaraya-Watson estimator proposed by Nadaraya (1964) and Watson (1964) with our local bandwidths and obtained a similar performance to above.

To illustrate the use of the proposed idea, we now apply the proposed methodology to analyze a real data set. This data set pertains to an exhaust study using ethanol as the fuel in an experimental engine (Cleveland 1993). The experiment is to determine the dependency of the concentration of oxides of nitrogen, the major air pollutants, on various engine settings. In particular, the response variable Y is the concentration of nitric oxide plus the concentration of nitrogen dioxide in the exhaust of an experimental engine when the engine is set at different equivalence ratios (the amount of ethanol) x. We modeled Y with x using the local linear estimator. The Bayesian local CV estimated regrssion function for several pairs of prior parameters along with the global CV type estimator are presented in Figure 3.3.

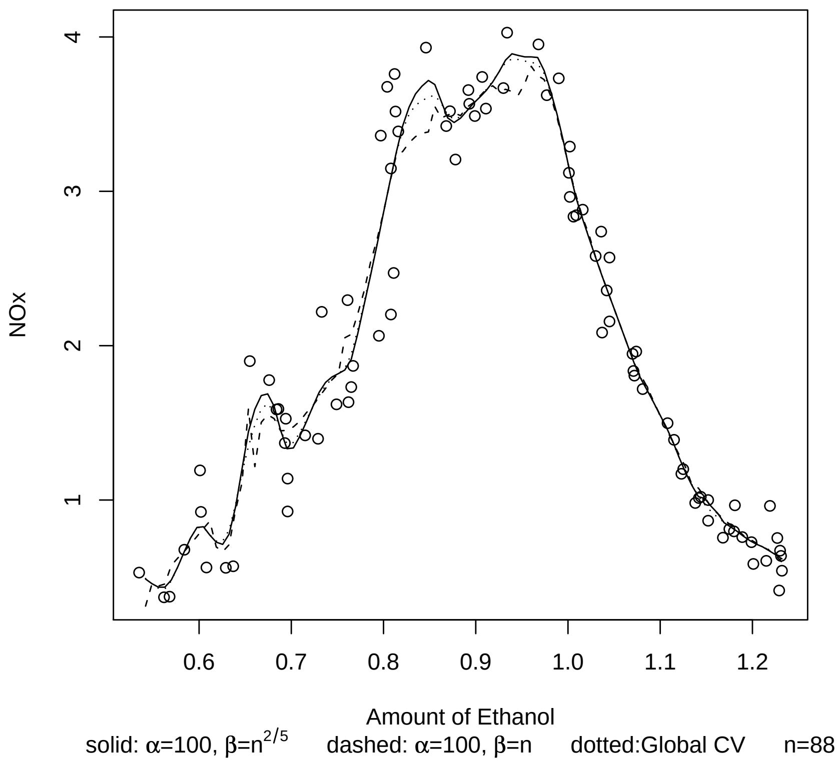

Figure 3.3: regression estimation of Ethanol Data

Chapter 4

Conclusions

The Bayesian local cross-validation procedure for smoothing parameter selection proposed here has several advantages as seen in our simulations. The finite sample dominance of these estimators coupled with reasonable asymptotic properties make such bandwidths highly desirable. However, there are several issues in the Bayesian approach that need further investigation. These include the examination of the Bayes criteria couple with other smoothing parameter selection criterions, such as AIC, AIC_c.

Appendices

Appendix A Proof of Theorem 2.1

Here, we show $h^*(x) \to 0$ and $nh^*(x) \to \infty$, as $n \to \infty$.

Since β diverges as sample size increases, we write β as β_n. Choose ϵ_n, such that (1) $\epsilon_n \to \infty$, and (2) $\frac{\epsilon_n}{\sqrt{\beta_n}} \to 0$, as $n \to \infty$. Since $\frac{\Gamma(\alpha)}{\Gamma(\alpha+\frac{1}{2})}$ is a constant for fixed α, (7) can be rewritten as

$$h^*(x) = C_\alpha \frac{\sum_{i=1}^n \left(\frac{1}{\beta_n (X_i - x)^2 + 2}\right)^\alpha}{\sqrt{\beta_n} \sum_{i=1}^n \left(\frac{1}{\beta_n (X_i - x)^2 + 2}\right)^{\alpha+\frac{1}{2}}} \tag{1}$$

where $C_\alpha = \frac{\Gamma(\alpha)}{\sqrt{2}\Gamma(\alpha+\frac{1}{2})}$. Consider the numerator of (9),

$$\begin{aligned}
\sum_{i=1}^n \left(\frac{1}{\beta_n (X_i - x)^2 + 2}\right)^\alpha &= \sum_{i=1}^n I\left[|X_i - x| \le \frac{\epsilon_n}{\sqrt{\beta_n}}\right] \left(\frac{1}{\beta_n (X_i - x)^2 + 2}\right)^\alpha \\
&+ \sum_{i=1}^n I\left[|X_i - x| \ge \frac{\epsilon_n}{\sqrt{\beta_n}}\right] \left(\frac{1}{\beta_n (X_i - x)^2 + 2}\right)^\alpha \\
&< \sum_{i=1}^n I\left[|X_i - x| \le \frac{\epsilon_n}{\sqrt{\beta_n}}\right] \left(\frac{1}{\beta_n (X_i - x)^2 + 2}\right)^\alpha + n\left(\frac{1}{\epsilon_n^2 + 2}\right)^\alpha \\
&= n\frac{\epsilon_n}{\sqrt{\beta_n}} \sum_{i=1}^n \frac{1}{n\frac{\epsilon_n}{\sqrt{\beta_n}}} I\left[|X_i - x| \le \frac{\epsilon_n}{\sqrt{\beta_n}}\right] \left(\frac{1}{\beta_n(X_i - x)^2 + 2}\right)^\alpha + n\left(\frac{1}{\epsilon_n^2 + 2}\right)^\alpha \\
&\le n\frac{\epsilon_n}{\sqrt{\beta_n}} \sum_{i=1}^n \frac{1}{n\frac{\epsilon_n}{\sqrt{\beta_n}}} I\left[|X_i - x| \le \frac{\epsilon_n}{\sqrt{\beta_n}}\right] \left(\frac{1}{2}\right)^\alpha + n\left(\frac{1}{\epsilon_n^2 + 2}\right)^\alpha
\end{aligned}$$

Consider the denominator of (9)

$$\begin{aligned}
\sqrt{\beta_n} \sum_{i=1}^n \left(\frac{1}{\beta_n (X_i - x)^2 + 2}\right)^{\alpha+\frac{1}{2}} &\ge \sqrt{\beta_n} \sum_{i=1}^n I\left[|X_i - x| \le \frac{\sqrt{\epsilon_n}}{\sqrt{\beta_n}}\right] \left(\frac{1}{\beta_n(X_i - x)^2 + 2}\right)^{\alpha+\frac{1}{2}} \\
&= n\sqrt{\epsilon_n} \sum_{i=1}^n \frac{1}{n\frac{\sqrt{\epsilon_n}}{\sqrt{\beta_n}}} I\left[|X_i - x| \le \frac{\sqrt{\epsilon_n}}{\sqrt{\beta_n}}\right] \left(\frac{1}{\beta_n(X_i - x)^2 + 2}\right)^{\alpha+\frac{1}{2}} \\
&\ge n\sqrt{\epsilon_n} \sum_{i=1}^n \frac{1}{n\frac{\sqrt{\epsilon_n}}{\sqrt{\beta_n}}} I\left[|X_i - x| \le \frac{\sqrt{\epsilon_n}}{\sqrt{\beta_n}}\right] \left(\frac{1}{\epsilon_n + 2}\right)^{\alpha+\frac{1}{2}}
\end{aligned}$$

Let $f_{1n}(x)$ denotes $\sum_{i=1}^n \frac{\sqrt{\beta_n}}{n\epsilon_n} I\left[|X_i - x| \le \frac{\epsilon_n}{\sqrt{\beta_n}}\right]$, and let $f_{2n}(x)$ denotes $\sum_{i=1}^n \frac{\sqrt{\beta_n}}{n\sqrt{\epsilon_n}} I\left[|X_i - x| \le \frac{\sqrt{\epsilon_n}}{\sqrt{\beta_n}}\right]$

then, combing two inequalities above, we get

$$\frac{\sum_{i=1}^{n}\left(\frac{1}{\beta_n(X_i-x)^2+2}\right)^{\alpha}}{\sqrt{\beta_n}\sum_{i=1}^{n}\left(\frac{1}{\beta_n(X_i-x)^2+2}\right)^{\alpha+\frac{1}{2}}} \leq \frac{n\frac{\epsilon_n}{\sqrt{\beta_n}}f_{1n}(x)\left(\frac{1}{2}\right)^{\alpha}+n\left(\frac{1}{\epsilon_n^2+2}\right)^{\alpha}}{n\sqrt{\epsilon_n}f_{2n}(x)\left(\frac{1}{\epsilon_n+2}\right)^{\alpha+\frac{1}{2}}}$$

$$= \frac{\sqrt{\epsilon_n}}{\sqrt{\beta_n}}\frac{f_{1n}(x)\left(\frac{1}{2}\right)^{\alpha}}{f_{2n}(x)\left(\frac{1}{\epsilon_n+2}\right)^{\alpha+\frac{1}{2}}}+\frac{\left(\frac{1}{\epsilon_n^2+2}\right)^{\alpha}}{\sqrt{\epsilon_n}f_{2n}(x)\left(\frac{1}{\epsilon_n+2}\right)^{\alpha+\frac{1}{2}}}$$

Since both $\sum_{n=1}^{\infty}\exp(-\gamma n(\frac{\epsilon_n}{\sqrt{\beta_n}})^2)$ and $\sum_{n=1}^{\infty}\exp(-\gamma n(\frac{\sqrt{\epsilon_n}}{\sqrt{\beta_n}})^2)$ converge for every $\gamma > 0$ with our choices of ϵ_n, β_n, then

$$sup_x|f_{1n}(x)-f(x)| \stackrel{a.s}{\to} 0, \quad sup_x|f_{2n}(x)-f(x)| \stackrel{a.s}{\to} 0$$

as $n\to\infty$; by Theorem 2.1.3 Prakasa Rao (1983).

Notice that $\sqrt{\epsilon_n} \to \infty$ and $\frac{\sqrt{\epsilon_n}}{\sqrt{\beta_n}} \to 0$, as $n \to \infty$, then

$$\frac{\sum_{i=1}^{n}\left(\frac{1}{\beta_n(X_i-x)^2+2}\right)^{\alpha}}{\sqrt{\beta_n}\sum_{i=1}^{n}\left(\frac{1}{\beta_n(X_i-x)^2+2}\right)^{\alpha+\frac{1}{2}}} \leq \left(\frac{1}{2}\right)^{\alpha}\frac{\sqrt{\epsilon_n}}{\sqrt{\beta_n}}C_1\frac{f(x)}{f(x)\left(\frac{1}{\epsilon_n+2}\right)^{\alpha+\frac{1}{2}}}+C_2\frac{\left(\frac{1}{\epsilon_n^2+2}\right)^{\alpha}}{\sqrt{\epsilon_n}f(x)\left(\frac{1}{\epsilon_n+2}\right)^{\alpha+\frac{1}{2}}}$$

$$\leq \left(\frac{1}{2}\right)^{\alpha}C_1\left(\frac{\sqrt{\epsilon_n}(\epsilon_n+2)^{\alpha+\frac{1}{2}}}{\sqrt{\beta_n}}\right)+C_2\left(\frac{(\epsilon_n+2)^{\alpha+\frac{1}{2}}}{f(x)\sqrt{\epsilon_n}(\epsilon_n^2+2)^{\alpha}}\right)$$

$$\leq C_3\left(\frac{\epsilon_n^{\alpha+1}}{\sqrt{\beta_n}}+\frac{1}{\epsilon_n^{\alpha}}\right)$$

with probability one, as $n \to \infty$, where C_1, C_2, C_3 are constant. Hence, for sufficiently large n

$$h^*(x) \leq C\left(\frac{\epsilon_n^{\alpha+1}}{\sqrt{\beta_n}}+\frac{1}{\epsilon_n^{\alpha}}\right)$$

almost surely, where some C. The optimal rate of ϵ_n to minimize the right hand side is $\beta_n^{\frac{1}{2(2\alpha+1)}}$. Then,

$$h^*(x) \leq C\beta_n^{-\frac{\alpha}{2(2\alpha+1)}}$$

Also, since

$$\sum_{i=1}^{n}\left(\frac{1}{\beta_n(X_i-x)^2+2}\right)^{\alpha} \geq \sum_{i=1}^{n}\left(\frac{1}{\beta_n(X_i-x)^2+2}\right)^{\alpha+\frac{1}{2}}$$

we have

$$h^*(x) \geq C_{\alpha}\beta_n^{-\frac{1}{2}}$$

Therefore

$$C_{\alpha}\beta_n^{-\frac{1}{2}} \leq h^*(x) \leq C\beta_n^{-\frac{\alpha}{2(2\alpha+1)}} \tag{2}$$

By the assumptions, $\beta_n \to \infty$ and $n\beta_n^{-\frac{1}{2}} \to \infty$, we get $h^*(x) \to 0$ and $nh^*(x) \to \infty$, as $n \to \infty$. Hence, $h^*(x)$ is a proper bandwidth of regression estimator for every x.

Appendix B More numerical results

More figures about our simulation study are presented here. These figures show the improvement of our procedure convincingly.

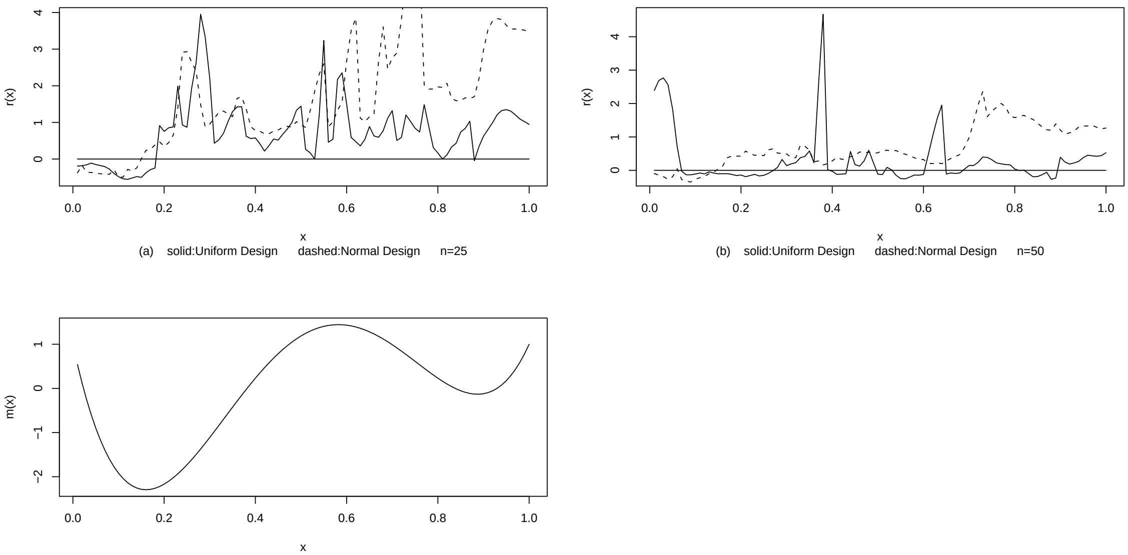

Figure 1: $m(x) = 1 - 48x + 218x^2 - 315x^3 + 145x^4$, $\sigma = 0.2$

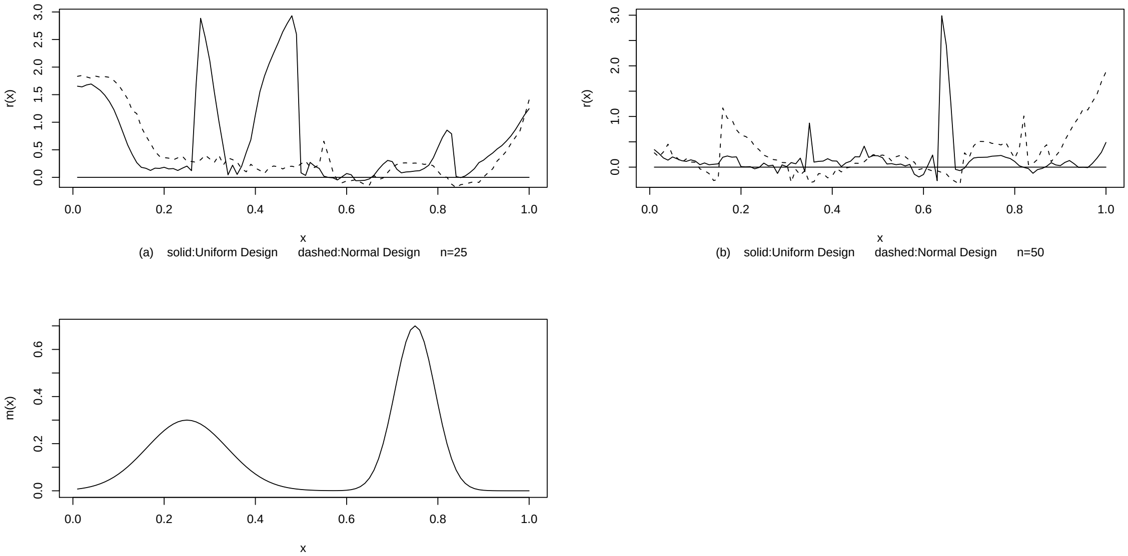

Figure 2: $m(x) = 0.3\exp\{-64(x-0.25)^2\} + 0.7\exp\{-256(x-0.75)^2)\}$, $\sigma = 0.3$

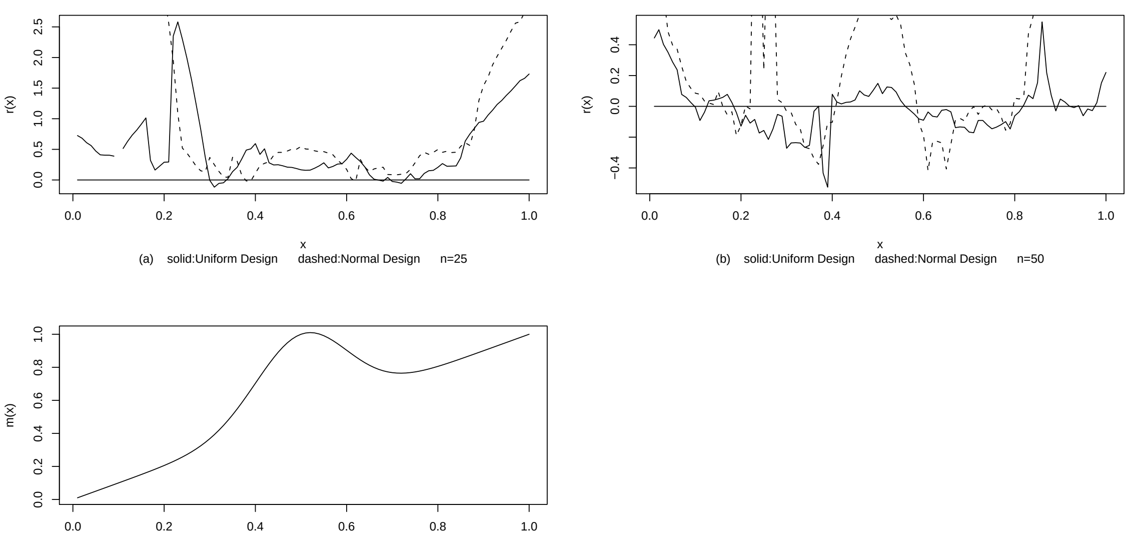

Figure 3: $m(x) = x + 0.5\exp\{-50(x-0.5)^2\}$, $\sigma = 0.2$

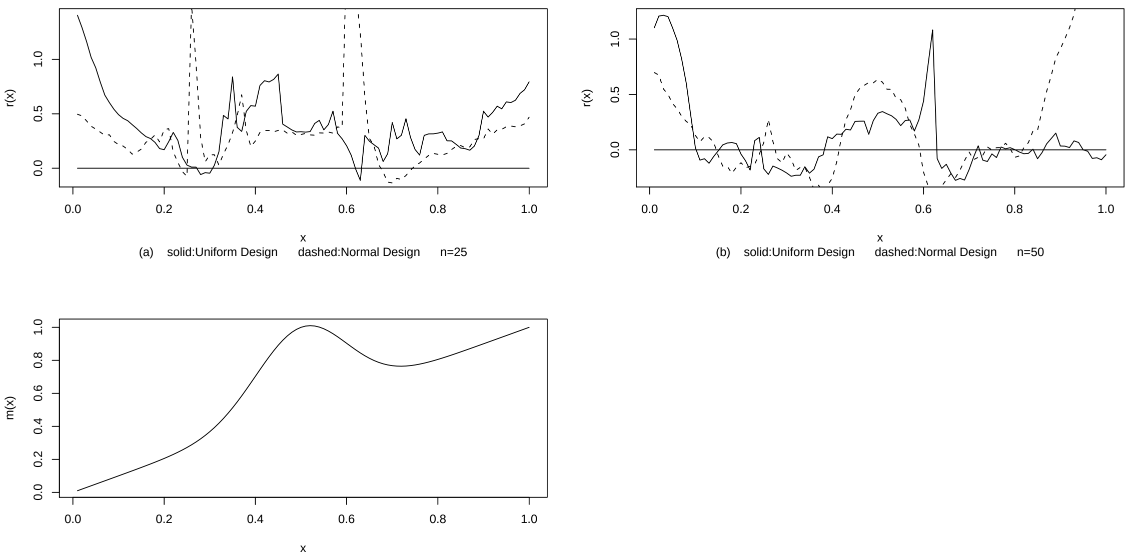

Figure 4: $m(x) = x + 0.5\exp\{-50(x-0.5)^2\}$, $\sigma = 0.3$

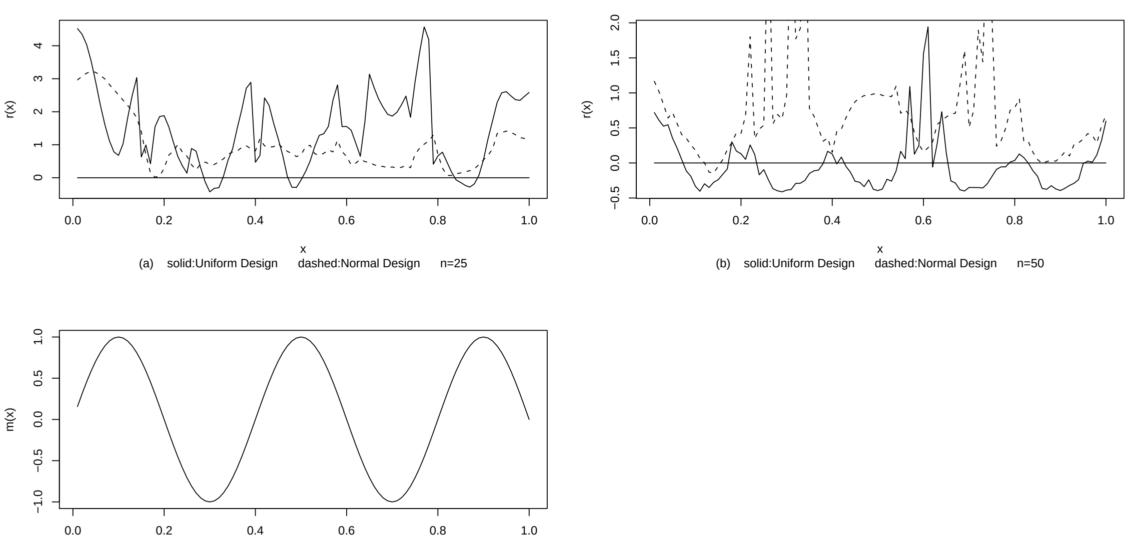

Figure 5: $m(x) = sin(5\pi x)$, $\sigma = 0.2$

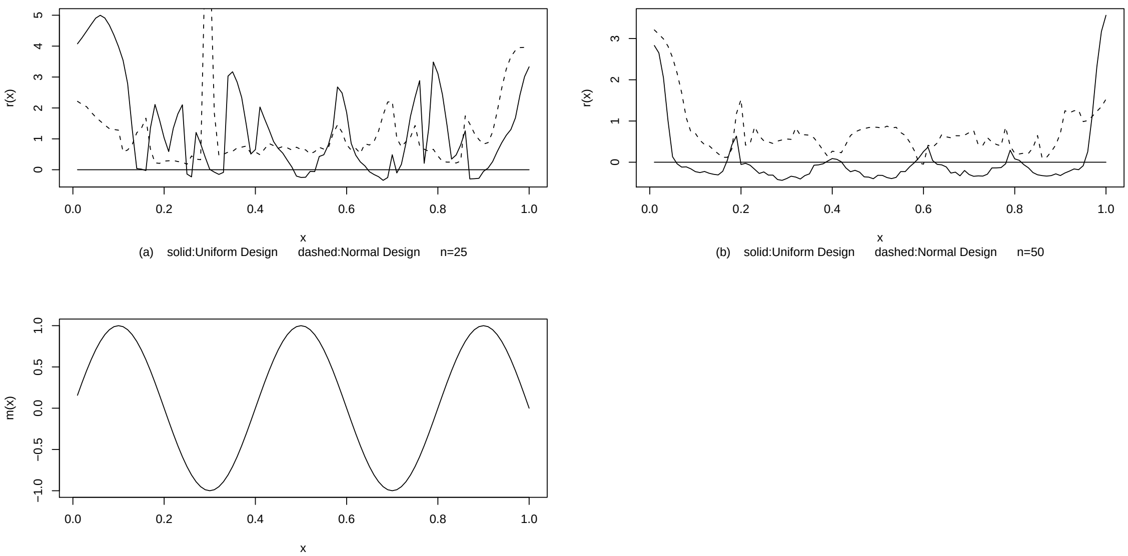

Figure 6: $m(x) = sin(5\pi x)$, $\sigma = 0.3$

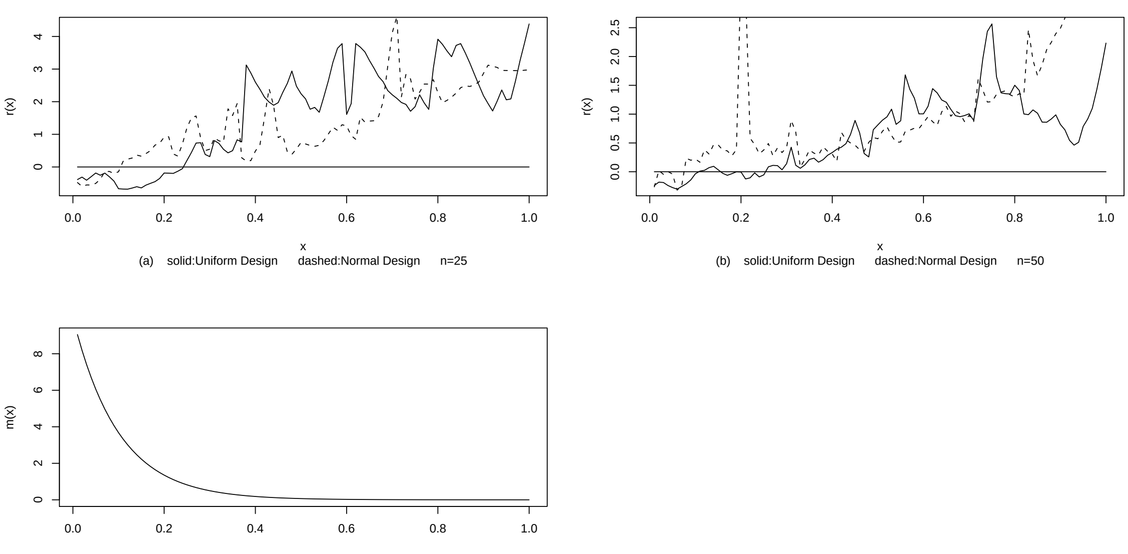

Figure 7: $m(x) = 10\exp(-10x)$, $\sigma = 0.2$

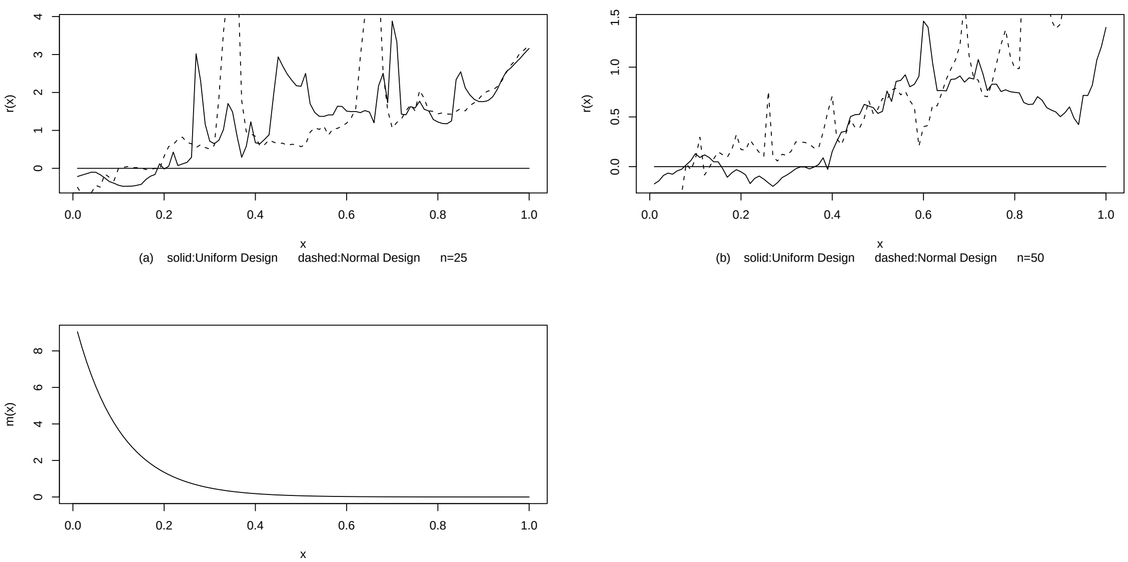

Figure 8: $m(x) = 10\exp(-10x)$, $\sigma = 0.3$

Bibliography

Akaike, H. (1973). "Information theory and an extension of the maximum likelihood principle," In *Proc. 2nd Int. Symp. Information Theory* (eds. B. N. Petrov and F. Csàki), pp. 267-281. Budapest: Akadémiai Kiadó

Cleveland, W. (1993), *Visualizing Data*, Hobart Press, Lafayette.

Craven, P. and Wahba, G. (1979). "Smoothing noisy data with spline functions." in *Numer. Math.*, **31**, 377-403.

Eubank, R. L. (1999), *Nonparametric regression and spline smoothing*, Marcel Dekker, New York.

Fan, J., Hall, P., Martin, M. and Patil, P. (1996). "On local smoothing of nonparametric curve estimators," *Journal of American Statistical Association*, **91**, 258-266.

Gangopadhyay, A. and Cheung, K. (2002). "Bayesian approach to the choice of smoothing parameter in kernel density estimation," *Journal of Nonparametric Statistics*, **14(6)**, 655-664.

Girard D. A. (1998). "Asymptotic comparison of (partial) cross-validation, GCV and randomized GCV in nonparametric regression," *Annals of Statistics*, **26(1)**, 315-334.

Hardle, W. and Bowman, A. G. (1988). "Bootstrapping in nonparametric regression: local adaptive smoothing and confidence bands," *Journal of American Statistical Association*, **83**, 102-110.

Hart, J. D. and Yi, S. "One-sided cross-validation," unpublished

Herrmann, E. (1997). "Local bandwidth choice in kernel regression estimation," *Journal of Computational and Graphical Statistics*, **6**, 35-54.

Hurvich, C. M. Simonoff J. S. and Tsai C. (1998). "Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion," *Journal of Royal Statistical Society*, **60**, Part 2, pp 271-293.

Ihaka, R. and Gentleman, R. (1996). "R: A language for data analysis and graphics," *Journal of Computational and Graphical Statistics*, **5**, 299-314, (*http://www.r-project.org*).

Kulasekera, K.B., and Padgett, W. J (2006), "Bayes bandwidth selection in kernel density estimation with censored data," *Journal of Nonparametric Statistics*, **18**, pp 129-143.

Nadaraya, E. A. (1964), "On Estimating Regression," *Theory of Probability and its Applications* , **9**(1), pp 141-142.

Rao, P. B. (1983), *Nonparametric function estimation*, Academic Press, London.

Ruppert, D., Sheather, S. J. and Wand, M. P. (1995). "An effective bandwidth selector for local least squares regression," *Journal of American Statistical Association*, **90**, 1215-1230.

Watson, G. (1964). "Smooth regression analysis," *Sankhya, Series A*, **26**, 359-372.