

DEPARTMENT OF STATISTICS

University of Wisconsin

1210 West Dayton St.

Madison, WI 53706

TECHNICAL REPORT NO. 1084rr

February 17, 2004

Doubly Penalized Likelihood Estimator in Heteroscedastic Regression¹

Ming Yuan² and Grace Wahba³

Department of Statistics, University of Wisconsin, Madison WI

Key words and phrases: Nonparametric regression, heteroscedasticity, doubly penalized likelihood estimator (DPLE), GCV, GACV

¹Research supported in part by NSF Grant DMS-0772292 and NASA Grant NAG5-1073

²Email: yuanm@stat.wisc.edu

³Email: wahba@stat.wisc.edu

Doubly Penalized Likelihood Estimator in Heteroscedastic Regression

Ming Yuan and Grace Wahba

Abstract

A penalized likelihood estimation procedure is developed for heteroscedastic regression. A distinguishing feature of the new methodology is that it estimates both the mean and variance functions simultaneously without parametric assumption for either. An efficient implementation of the estimating procedure is also provided. The procedure is illustrated by a Monte Carlo example. A potential generalization, and application to the covariance modeling problem in numerical weather prediction is noted.

1 Introduction

Consider the following heteroscedastic regression model

$$y_i = \mu(x_i) + \sigma(x_i)\varepsilon_i, \quad i = 1, 2, \dots, n, \quad (1.1)$$

where μ and σ are unknown functions to be estimated, $x_i, i = 1, \dots, n$ are univariate or multivariate covariates and ε 's are i.i.d. noise with mean 0 and variance 1. The focus of most nonparametric regression methodologies centers around the conditional mean. However, the understanding of the local variability of the data, measured by the conditional variance is very important in many scientific studies (e.g, Andersen and Lund, 1997; Gallant and Tauchen, 1997). These applications necessitate the development of new nonparametric techniques that allow the modeling of varying variance.

Two step procedures are commonly used. First, the conditional mean is estimated. Then, an estimate of the conditional variance is constructed based on the regression residuals. Earlier proposals include Carroll (1982), Muller and Stadtmuller (1987), Hall and Carroll (1989), Ruppert *et. al.* (1997) and Fan and Yao (1998) among many others. More recently, a Bayesian treatment to (1.1) has also been introduced by Yau and Kohn (2003).

In this paper, an alternative approach is proposed. We introduce a penalized likelihood based procedure which can take the heteroscedasticity into account and estimate both mean function and variance function simultaneously.

The remaining note is organized as follows. The DPLE is introduced in the next section. Then it is expressed as a nonlinear optimization problem and a practical estimating procedure

is given in Section 3. In Section 4, we construct the Bayesian confidence interval for the DPLE. The paper is concluded by a Monte Carlo example.

2 Doubly Penalized Likelihood Estimator

By assuming that $\varepsilon_i \sim_{iid} N(0, 1)$, we can write down the average negative log likelihood of (1.1), which, up to a constant not depending on μ or σ^2 , is

$$L(\mu, \sigma) = \frac{1}{n} \sum_{i=1}^n \left(\frac{(y_i - \mu(x_i))^2}{2\sigma^2(x_i)} + \frac{1}{2} \log \sigma^2(x_i) \right)$$

Like other nonparametric settings, instead of assuming a parametric form of μ and σ^2 , we allow them to reside in some Hilbert spaces $\mathcal{H}_\mu, \mathcal{H}_{\sigma^2}$ of smooth functions. Using the penalized likelihood strategy, we may define our estimators $(\hat{\mu}, \hat{\sigma}^2)$ to be the minimizer of the following penalized negative log likelihood function over the corresponding functional spaces:

$$L(\mu, \sigma) + \frac{\lambda_\mu}{2} J_\mu(\mu) + \frac{\lambda_{\sigma^2}}{2} J_{\sigma^2}(\sigma^2), \quad (2.1)$$

where both penalty functions J_μ and J_{σ^2} are quadratic forms defined on $\mathcal{H}_\mu, \mathcal{H}_{\sigma^2}$ respectively and smoothing parameters λ_μ and λ_{σ^2} are nonnegative constants which control the tradeoff between L , the goodness-of-fit on the data, and the smoothness penalties J_μ and J_{σ^2} .

To work around the positivity constraint of σ^2 , write

$$\sigma^2(x) = e^{g(x)}. \quad (2.2)$$

Then we assume that g lies in a Hilbert space \mathcal{H}_g and re-express the penalized negative likelihood as

$$T_n(\mu, g; y, \lambda_\mu, \lambda_g) \equiv \frac{1}{n} \sum_{i=1}^n \left((y_i - \mu(x_i))^2 e^{-g(x_i)} + g(x_i) \right) + \lambda_\mu J_\mu(\mu) + \lambda_g J_g(g). \quad (2.3)$$

If both smoothing parameters are big enough, then it is not hard to see that (2.3) is convex in (μ, g) jointly. But if the λ 's are small enough it can be seen, by looking at the Hessian, that bivariate convexity cannot be guaranteed. A practically useful statement concerning the exact nature of “big enough” is a subject of ongoing research. However, in a number of realistic simulations, for example, if the actual variance varies slowly compared to the data density, then our simulations suggest that λ 's of interest are “big enough”, and possible lack of bivariate convexity is not a problem.

It is interesting to consider some special cases of the proposed DPLE. If we choose $\lambda_\mu = \infty$, then we end up with a model having a parametric conditional mean and a nonparametric conditional variance. A special case of this where $J_\mu(\mu) = f(\mu'')^2$ has been treated by Carroll (1982) using the kernel smoother.

If we choose $\lambda_g = \infty$, then the conditional variance of model (1.1) is forced to take a parametric form. This includes the usual nonparametric regression model with constant variances in which our DPLE becomes the well-known smoothing spline estimator.

If we choose $\lambda_\mu = \lambda_g = \infty$, then our DPLE boils down to a maximum likelihood estimate for a parametric model where both μ and g lie in the null spaces in the sense that $J_\mu(\mu) = J_g(g) = 0$, for some penalty functions J_μ and J_g . A case of common interest is $J_\mu(\mu) = f(\mu'')^2$. The null space corresponding to this penalty function is the usual linear model for μ . Jobson and Fuller (1980) argued that the maximum likelihood estimator dominates the simple least squares based estimators in this setting.

3 Algorithm

A multiple functional extension of the representer theorem of Kimeldorf and Wahba (1971) ensures that the minimizer of (2.3) lies in a finite dimensional space, even when the minimization is carried out in infinite dimensional Hilbert spaces.

For brevity, let $\mathcal{H} \equiv \mathcal{H}_\mu = \mathcal{H}_g$ and $J \equiv J_\mu = J_g$. However, it is worth noting that in some applications, penalty functions J_μ and J_g could be different. Actually, μ and g could even be defined on entirely different sets of covariates.

Penalty J induces a decomposition of $\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1$ where \mathcal{H}_0 is the null space of J . Assume that $\{\phi_\nu : 1 \leq \nu \leq M\}$ are the basis functions of \mathcal{H}_0 and $K(\cdot, \cdot)$ is the reproducing kernel of \mathcal{H}_1 . Then the representer theorem guarantees the minimizer of (2.3) to be

$$\mu(x) = \sum_{i=1}^M d_i^\mu \phi_i(x) + \sum_{i=1}^n c_i^\mu K(x_i, x), \quad (3.1)$$

$$g(x) = \sum_{i=1}^M d_i^g \phi_i(x) + \sum_{i=1}^n c_i^g K(x_i, x), \quad (3.2)$$

for some vectors c_μ, d_μ, c_g, d_g , where $d_\mu = (d_1^\mu, \dots, d_M^\mu)'$, $c_\mu = (c_1^\mu, \dots, c_n^\mu)'$ and c_g, d_g are defined in the same fashion.

For the ease of exposition, we shall use the matrix notation: Let T be a $n \times M$ matrix with the (i, ν) th entry $\phi_\nu(x_i)$ and $y = (y_1, \dots, y_n)'$. With some abuse of notation, let $\mu = (\mu_1, \dots, \mu_n)' \equiv (\mu(x_1), \dots, \mu(x_n))'$, $g = (g_1, \dots, g_n)' \equiv (g(x_1), \dots, g(x_n))'$, and K be the

$n \times n$ matrix with (i, j) th entry $K(x_i, x_j)$. Rewrite (3.1) and (3.2) as

$$\mu = Kc_\mu + Td_\mu, \quad g = Kc_g + Td_g. \quad (3.3)$$

The variational problem (2.3) now becomes finding (c_μ, d_μ, c_g, d_g) to minimize

$$\Phi(d_\mu, c_\mu, d_g, c_g) \equiv (y - Td_\mu - Kc_\mu)'D(y - Td_\mu - Kc_\mu) + e'g + n\lambda_\mu c_\mu'Kc_\mu + n\lambda_g c_g'Kc_g, \quad (3.4)$$

where D is a diagonal matrix with the i th diagonal element $e^{-g(x_i)}$ and $e = (1, \dots, 1)'$. A clue for minimizing (3.4) is that it is a convex function of each of μ and g by fixing the other. Thus, we can minimize (3.4) with respect to μ and g in an iterative fashion.

Fixing g , (3.4) reduces to a penalized weighted least squares functional

$$\frac{1}{n} \sum_{i=1}^n d_{ii}(y_i - \mu(x_i))^2 + \lambda_\mu J(\mu), \quad (3.5)$$

where d_{ii} is the (i, i) entry of diagonal matrix D . The solution to (3.5) is

$$\hat{\mu} = A_\mu(\lambda_\mu)y, \quad (3.6)$$

where A_μ is the so-called hat matrix (see Wahba, 1990). In this step, only smoothing parameter λ_μ is involved. A commonly used smoothing parameter tuning technique is to choose a λ_μ which minimizes the GCV score:

$$V(\lambda_\mu) = \frac{n^{-1}y'D^{1/2}(I - A_\mu(\lambda_\mu))^2 D^{1/2}y}{[n^{-1}tr(I - A_\mu(\lambda_\mu))]^2}. \quad (3.7)$$

Fixing μ , the objective functional becomes

$$\frac{1}{n} \sum_{i=1}^n (z_i e^{-g(x_i)} + g(x_i)) + \lambda_g J(g), \quad (3.8)$$

where $z_i = (y_i - \tilde{\mu}(x_i))^2$ and $\tilde{\mu}$ is the current estimate of μ . It is worth noting that (3.8) has the form of a penalized Gamma likelihood as if $z_i, i = 1, \dots, n$ were independent samples from Gamma distributions with shape parameter 1 and scale parameters $e^{g(x_i)}, i = 1, \dots, n$. This connection makes it possible to apply the general methodology for solving penalized likelihood problems with responses from exponential family. For a fixed smoothing parameter λ_g , (3.8) is strictly convex in g and thus can be minimized using the Newton iteration. In this step, the objective functional only contains λ_g . We tune λ_g by minimizing the generalized approximate cross validation (GACV) technique developed by Xiang and Wahba (1996):

$$GACV(\lambda_g) = \frac{1}{n} \sum_{i=1}^n (z_i e^{-g(x_i)} + g(x_i)) + \frac{tr(A_g(\lambda_g))}{n} \sum_{i=1}^n \frac{z_i e^{-\hat{g}(x_i)} (z_i - e^{-\hat{g}(x_i)})}{1 + tr(HA_g(\lambda_g))}, \quad (3.9)$$

where H is a diagonal matrix with (i, i) entry $e^{-\widehat{g}(x_i)}$ and $A_g(\lambda_g)$ is the influence matrix of (3.8) (see Xiang and Wahba, 1996).

Summing up, an algorithm which chooses the smoothing parameters automatically is as follows.

-
1. Initialize c_g, d_g .
 2. (i) Given the current c_g, d_g , choose λ_μ such that $V(\lambda_\mu)$ is minimized.
(ii) Given the current c_g, d_g and λ_μ , update c_μ, d_μ by minimizing (3.4) with respect to c_μ and d_μ .
 3. (i) Given the current c_μ, d_μ , choose λ_g such that $GACV(\lambda_g)$ is minimized.
(ii) Given the current c_μ, d_μ and λ_g , update c_g, d_g by minimizing (3.4) with respect to c_g and d_g .
 4. Iterate Step 2 and Step 3 till convergence.
-

In our simulations, we found that extremely large negative intermediate estimates of g in the iteration often lead to overfitting and numerical instability. To avoid this problem, one could slightly perturb those near zero residuals. For example, in our simulation, we used $z_i = \max(0.00001, (y_i - \tilde{\mu}(x_i))^2)$.

The algorithm converges fairly fast according to our experience. Usually, it converges after 4-5 iterations between Step 2 and 3 if one starts from $c_g = 0$ and $d_g = 0$. It is also worth noting that the estimate of μ after the first iteration is just the usual unweighted μ estimation if we choose $c_g = 0$ and $d_g = 0$ as the initial value.

4 Bayesian Confidence Interval

Approximating Bayesian inference is often used to construct a confidence interval for penalized likelihood estimators. Usually, the penalty term can be interpreted as a prior and the penalized likelihood estimator is then equivalent to a maximum posterior estimator. Applying the Laplace approximation to the corresponding posterior distribution of the unknown function around this posterior mode, we can get an approximating Bayesian confidence interval. For details, see Wahba (1990) and Gu (2002).

Similarly, here we can regard penalty terms as priors for μ and g respectively. More precisely, let the prior distributions be

$$\begin{aligned} d_\mu &\sim N(0, \tau_1^2 I_{(M)}), & c_\mu &\sim N(0, b_1^2 K) \\ d_g &\sim N(0, \tau_2^2 I_{(M)}), & c_g &\sim N(0, b_2^2 K) \end{aligned}$$

Using the argument for smoothing splines (Wahba, 1978), we can characterize our DPLE as a maximum a posteriori estimator.

Lemma 4.1 *Let $\tau_1^2, \tau_2^2 \rightarrow \infty$, the DPLE defined in the last section is a posterior mode of (μ, g) given y with*

$$\lambda_\mu = 1/nb_1^2 \quad \text{and} \quad \lambda_g = 1/nb_2^2.$$

To construct the approximate Bayesian confidence intervals for each of μ and g , we will let the other be fixed. Let us first consider the Bayesian formulation for μ . For fixed smoothing parameters, the approximate solution for μ is

$$\mu(\cdot) = \phi(\cdot)'d_\mu + \xi(\cdot)'c_\mu$$

where $\phi = (\phi_1, \dots, \phi_M)'$ and $\xi(\cdot) = (K(x_1, \cdot), \dots, K(x_n, \cdot))'$. Using the improper prior for (c, d) as in Lemma 4.1, we have

$$p(c_\mu, d_\mu) \propto \exp\left(-\frac{n\lambda_\mu}{2}c_\mu'Kc_\mu\right).$$

Thus, given g , the posterior distribution of c_μ, d_μ would be

$$p(c_\mu, d_\mu|y) \propto \exp\left(-\frac{1}{2}(y - Td_\mu - Kc_\mu)'D(y - Td_\mu - Kc_\mu) - \frac{n\lambda_\mu}{2}c_\mu'Kc_\mu\right).$$

This suggest that the posterior distribution of (c_μ, d_μ) given y is a multivariate normal distribution. Hence

$$E(\mu(x)|y, \lambda_\mu) = (\xi(x)', \phi(x)')Q_\mu^{-1} \begin{pmatrix} K \\ T \end{pmatrix} Dy; \quad (4.1)$$

$$Cov(\mu(x), \mu(x^1)|y, \lambda_\mu) = (\xi(x)', \phi(x)')Q_\mu^{-1} \begin{pmatrix} \xi(x^1) \\ \phi(x^1) \end{pmatrix} \quad (4.2)$$

where

$$Q_\mu = \begin{pmatrix} KDK + n\lambda_\mu K & KDT \\ T'DK & T'DT \end{pmatrix}.$$

Similarly, for fixed smoothing parameters, the approximate solution for g is

$$g(\cdot) = \phi(\cdot)'d_g + \xi(\cdot)'c_g.$$

Given μ , the posterior distribution of (c_g, d_g) could be approximated by

$$p(c_\mu, d_\mu|y) \propto \exp\left(-\frac{1}{2}(\tilde{z} - Td_g - Kc_g)'W(\tilde{z} - Td_g - Kc_g) - \frac{n\lambda_g}{2}c_g'Kc_g\right),$$

where \tilde{z} is the response used for the last Newton iteration to minimize (3.8) (see Gu, 2002) and W is a diagonal matrix with the (i, i) th entry $\tilde{z}_i e^{-g(x_i)}$. Thus,

$$E(g(x)|y, \lambda_g) \approx (\xi(x)', \phi(x)')Q_g^{-1} \begin{pmatrix} K \\ T \end{pmatrix} W\tilde{z}; \quad (4.3)$$

$$Cov(g(x), g(x^1)|y, \lambda_g) \approx (\xi(x)', \phi(x)')Q_g^{-1} \begin{pmatrix} \xi(x^1) \\ \phi(x^1) \end{pmatrix} \quad (4.4)$$

where

$$Q_g = \begin{pmatrix} KWK + n\lambda_g K & KWT \\ T'WK & T'WT \end{pmatrix}.$$

5 An Example

In this section, we will demonstrate the proposed method through a simple example. In this example, we consider three different sample sizes: $n = 125, 250, 500$. For each sample size, one hundred datasets were generated using the following setup.

$$x_i = (i - 0.5)/n, \quad i = 1, \dots, n; \quad (5.1)$$

$$y_i \sim N(\mu(x_i), \exp(g(x_i))), \quad (5.2)$$

where

$$\begin{aligned} \mu(x) &= 2 \left[\exp(-30(x - 0.25)^2) + \sin(\pi x^2) \right] - 2, \\ g(x) &= \sin(2\pi x). \end{aligned} \quad (5.3)$$

For each dataset, the DPLE was calculated using the algorithm given in the last section. The comparative Kullback-Leibler (CKL) distance is used to measure the performance of the estimators. More precisely, for an estimator $\hat{\mu}$ of the mean function μ , the CKL distance from $\hat{\mu}$ to μ is defined as

$$CKL(\hat{\mu}, \mu) = \frac{1}{n} \sum_{i=1}^n (\mu(x_i) - \hat{\mu}(x_i))^2 \sigma(x_i)^{-2}.$$

Similarly, the CKL distance from an estimator \hat{g} to g is defined as

$$CKL(\hat{g}, g) = \frac{1}{n} \sum_{i=1}^n \left[e^{g(x_i) - \hat{g}(x_i)} + \hat{g}(x_i) - g(x_i) \right],$$

which differs from the KL distance between \hat{g} and g only by a constant not depending on \hat{g} . Figure 1 gives the true test functions together with their 5th, 50th and 95th best fits for different sample sizes.

From Figure 1, we observe that the DPLE performs very well in most examples for sample size as small as 125. When the sample size increases, the performance improves. For sample size 500, even the 95th best fits are quite close to the truth.

Next, we investigate the empirical coverage of the Bayesian confidence intervals derived in Section 4. To this end, we repeated the above simulation with a fixed sample size 200. One hundred datasets were generated. For each dataset, we computed the Bayesian confidence interval at each sample point $x_i, i = 1, \dots, n$. We define the empirical coverage for each sample point as the percentage of datasets, for which the Bayesian confidence intervals successfully cover the true test functions.

Figure 2 gives the estimators and their 95% Bayesian confidence interval for a typical dataset. We also plotted the true test functions together with the empirical coverage probabilities on each sample points in figure 3. We find that 95% coverage is not achieved for each sample point. Instead, the overall coverage is approximately 95%. This phenomenon is called across-the-functions coverage, which is well understood for the smoothing splines (see Wahba, 1990 and Gu, 2002).

Over all the sample points, define the average coverage proportion by

$$\begin{aligned} ACP_\mu(\alpha) &= \frac{1}{n} \#\{i : |\hat{\mu}(x_i) - \mu(x_i)| < z_{\alpha/2} \sqrt{\widehat{var}(\hat{\mu}(x_i)|y)}\} \\ ACP_g(\alpha) &= \frac{1}{n} \#\{i : |\hat{g}(x_i) - g(x_i)| < z_{\alpha/2} \sqrt{\widehat{var}(\hat{g}(x_i)|y)}\}, \end{aligned}$$

where \widehat{var} stands for the approximate Bayesian posterior variance obtained in Section 4 and $z_{\alpha/2}$ is the $(1 - \alpha/2)$ th quantile of the standard normal distribution.

In the current example, we have

$$ACP_\mu(0.05) = 93.1\%, \quad ACP_g(0.05) = 92.5\%.$$

These coverages are slightly lower than 95%. This is reasonable because the approximate Bayesian confidence interval derived in Section 4 assumes that the smoothing parameters are fixed. But in our example, we automatically tuned the smoothing parameters instead of fixing them, which increases the variability of the DPLE.

These observations confirm that the Bayesian confidence interval constructed in Section 4 is valid.

6 Some Remarks on Generalizations and Applications

Modeling of covariances is an important problem in numerical weather prediction (NWP). See, for example Derber and Bouttier (1999), Dee *et. al.* (1999), Hollingsworth and Lönnberg (1986), Gong *et. al.* (1998), Purser and Parrish (2003) and Chapnik *et. al.*(2003). Note that although the example in this paper is univariate, the method applies to spatial data. A first generalization of the method in this paper would suppose that the ε_i in (1.1) have mean 0 and variance 1, but have a correlation matrix $C(x, x')$ which is known up to a small number of estimable parameters θ , and the challenge is to estimate $\sigma(x)$ and θ simultaneously. Generalizing further to a problem important in NWP, copious observations are available on the difference between observations and forecast. Since observation and forecast errors are generally modelled as independent, the y_i could be modeled as coming from a spatial process with covariance $\sigma_B(x)\sigma_B(x')C_B(x, x') + \sigma_R(x)\sigma_R(x')C_R(x, x')$, where B refers to forecast and R to observations, and the C 's are correlation matrices known up to possibly a few parameters. μ is the difference between the observation and forecast biases, but under the commonly made assumption that observations are unbiased, μ would be the forecast bias. In practice C_B and C_R will have very different structure, so that it may be possible in some cases to estimate the σ 's by generalization of the techniques in this paper. Chapnik *et. al.* discuss the case where σ_B and σ_R are constants to be estimated. It is believed that this and other generalizations have potential for a variety of important applications.

References

- [1] Andersen, T.G. and Lund, J. (1997), Estimating continuous time stochastic volatility function models of the short interest rate. *J. Economet.* **77**, 343-377.
- [2] Carroll, R.J. (1982), Adapting for heteroscedasticity in linear models. *Ann. Statist.* **10**, 1224-1233.
- [3] Chapnik, B., Desroziers, G., Rabier, F. and Talagrand, O. (2003), Properties and first application of an error statistics tuning method in variational assimilation, to appear in *Monthly Weather Review*.

- [4] Dee, D., Gaspari, G., Redder, C., Rukhovets, L. and A. da Silva (1999), Maximum-likelihood estimation of forecast and observation error covariance parameters. Part II: Applications, *Monthly Weather Review*, **8**, 1835-1849.
- [5] Derber, J. and Bouttier, F. (1999), A reformulation of the background error covariance in the ECMWF global data assimilation system, *Tellus*, **51A**, 195-221.
- [6] Fan, J.Q. and Yao, Q.W. (1998). Efficient estimation of conditional variance functions in stochastic regression. *Biometrika*, **85**, 645-660.
- [7] Gallant, A. R. and Tauchen, G. (1997), Estimation of continuous time models for stock returns and interest rates. *Macroeconomic Dynamics*, **1**, 135-168.
- [8] Gong, J., Wahba, G., Johnson, D. and Tribbia, J. (1998), Adaptive tuning of numerical weather prediction models: simultaneous estimation of weighting, smoothing and physical parameters, *Monthly Weather Review*, **126**, 210-231.
- [9] Gu, C. (2002), *Smoothing spline ANOVA models*. Springer-Verlag, New York.
- [10] Hall, P. and Carroll, R.J. (1989), Variance function estimation in regression: the effect of the estimation of the mean. *J. Roy. Statist. Soc. B* **51**, 3-14.
- [11] Hollingsworth, A. and Lönnerberg, P. (1986), The statistical structure of short-range forecast errors as determined from radiosonde data. Part I: The wind field, *Tellus*, **38A**, 111-136.
- [12] Jobson, J. D. and Fuller, W. A. (1980), Least Squares Estimation When the Covariance Matrix and Parameter Vector are Functionally Related. *J. Amer. Statist. Assoc.*, **75**, 176-181.
- [13] Muller, H.G. and Stadtmuller, U. (1987), Estimation of heteroscedasticity in regression analysis. *Ann. Statist.* **15**, 610-625.
- [14] Purser, J. and Parrish, D. (2003), A Bayesian technique for estimating continuously varying statistical parameters of a variational assimilation, *Meteor. Atmos. Physics*, **82**, 209-226.
- [15] Ruppert, D., Wand, M.P., Holst, U. and Hossjer, O. (1997), Local polynomial variance-function estimation. *Technometrics* **39**, 262-273.

- [16] Stadtmuller, V. and Tsybakov, A.B. (1995), Nonparametric recursive variance estimation. *Statistics* **27**, 55-63.
- [17] Wahba, G. (1978), Improper priors, spline smoothing and the problem of guarding against model errors in regression. *J. Roy. Statist. Soc. B*, **40(3)**, 364-372.
- [18] Wahba, G. (1990), Spline models for observational data. *CBMS-NSF Regional Conference Series in Applied Mathematics*, 59. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA.
- [19] Xiang, D. and Wahba, G. (1996), A generalized approximate cross validation for smoothing splines with non-Gaussian data. *Statist. Sinica*, **6**, 675–692.
- [20] Yau, P. and Kohn, R. (2003), Estimation and variable selection in nonparametric heteroscedastic regression, *Statistics and Computing* **13**, 191 - 208.

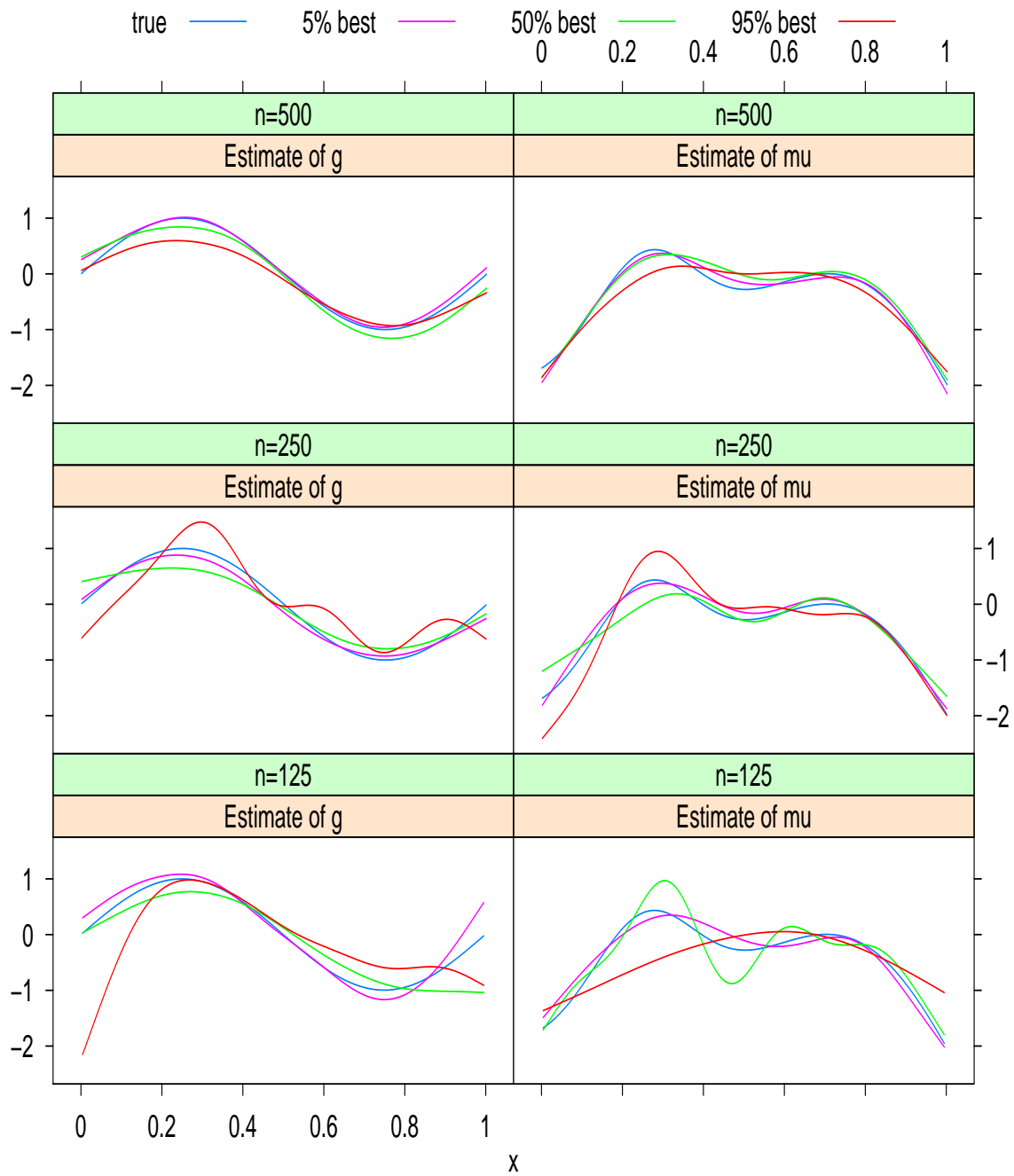


Figure 1: One hundred samples were simulated for each sample size. The DPLE were computed with automatically chosen smoothing parameters. True test functions together with their 5%, 50% and 95% best fits are given in the plot.

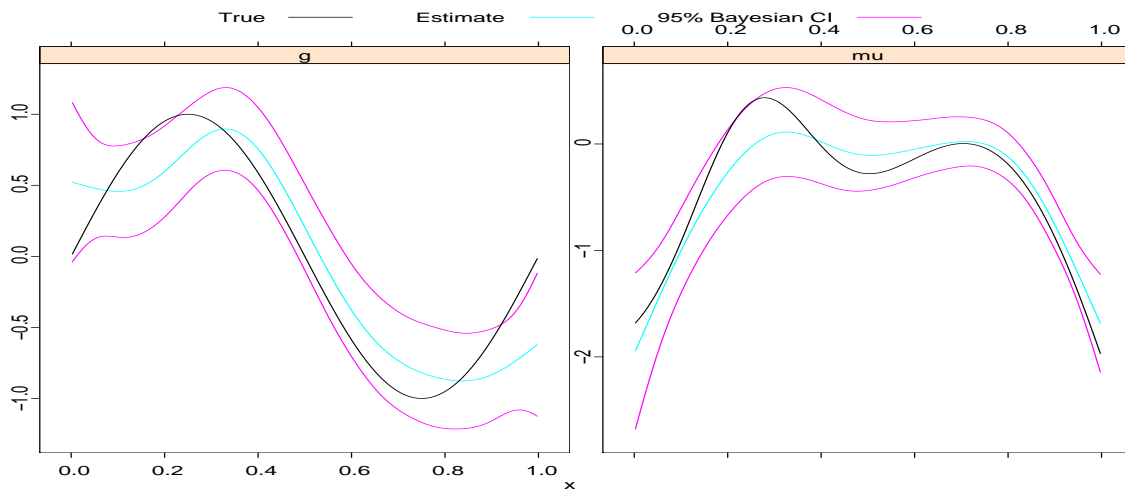


Figure 2: Bayesian Confidence Interval: For a typical dataset with 200 observations, the above plots give the true functions, the DPLE and the 95% Bayesian confidence intervals.

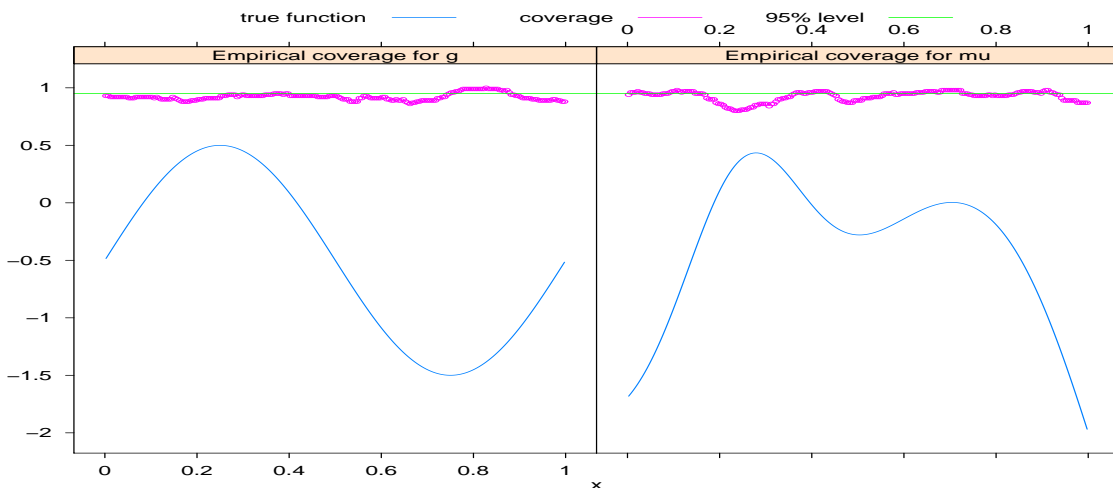


Figure 3: Coverage of Bayesian Confidence Intervals: The empirical coverage percentage for each sample point based on 100 simulated datasets are given in the above figure. Also plotted is the true test function. The variance function is shifted downward by 0.5.