ELSEVIER

Statistics & Probability Letters 58 (2002) 23–30

STATISTICS & PROBABILITY LETTERS

www.elsevier.com/locate/stapro

Bayesian nonparametric point estimation under a conjugate prior

Xuefeng Li, Linda H. Zhao*

Statistics Department, University of Pennsylvania, The Wharton School, 3620 Locust Walk, 19104-6302 Philadelphia, PA, USA

Received October 2001; received in revised form January 2002

Abstract

Estimation of a nonparametric regression function at a point is considered. The function is assumed to lie in a Sobolev space, S_q, of order q. The asymptotic squared-error performance of Bayes estimators corresponding to Gaussian priors is investigated as the sample size, n, increases. It is shown that for any such fixed prior on S_q the Bayes procedures do not attain the optimal minimax rate over balls in S_q. This result complements that in Zhao (Ann. Statist. 28 (2000) 532) for estimating the entire regression function, but the proof is rather different. © 2002 Elsevier Science B.V. All rights reserved.

Keywords: Nonparametric regression; Point estimation; Bayesian procedure; Gaussian prior; Optimal rate

1. Introduction

Within the past two decades nonparametric regression has become an important, widely used statistical methodology. More recently there has been increasing interest in the possibility of effectively using a Bayesian approach for such situations. This paper involves one step in that direction.

We investigate an aspect of the performance of the Bayes estimator for a natural conjugate prior. (These priors correspond to infinite dimensional Gaussian distribution.) Of interest is the asymptotic performance of the estimator of the regression function, f, at a given point, x_0.

As is customary, we assume that regression function lies in a standard function space—in this case a Sobolev space of specified smoothness. Consistent with this we derive the prior to be supported on this Sobolev space.

* Corresponding author. Fax: +1-215-898-1280.

E-mail address: lzhao@wharton.upenn.edu (L.H. Zhao).

PII: S0167-7152(02)00097-4

We show that for any such prior the Bayes estimators for sample size n do not attain the optimal minimax rate of squared error risk.

Zhao (2000) demonstrates an analogous deficiency of Bayes procedures for the problems of estimating the entire regression function. For that problem she also constructs a nonconjugate prior distribution whose Bayes estimators do attain the optimal minimax rate. It follows, however, from Cai et al. (2001) that at almost every x_0, these estimators will not attain the optimal minimax rate for estimating $f(x_0)$. The question is thus an open one of whether there exists a prior on the Sobolev space whose Bayes procedures attain this rate for estimating $f(x_0)$. We suspect such priors exist but have not (yet) succeeded in constructing them.

For further background on problems of this nature and additional references we refer the reader to Zhao (2000). See also Pinsker (1980). We close the introduction by noting two additional features of the results in the present manuscript.

First, we prove here an additional result that shows there do exist Gaussian priors supported outside the given Sobolev-space whose estimators of $f(x_0)$ do attain the optimal minimax rate. This result is analogous to one in Zhao (2000) for estimating all of f; however, the appropriate priors are different in the two problems. (This is consistent with the fact that the minimax rates are also different.) Second, although the main theorem here has analogies to the main result in Zhao (2000), the proof is rather different.

2. Preliminaries

In a standard nonparametric regression problem one observes (x_i, z_i), $i = 1, \ldots, n$ where

$$z_i = f(x_i) + \varepsilon_i, \quad i = 1, \ldots, n. \tag{1}$$

Here, we take $x_i = i/(n+1)$ to be equally spaced on $[0,1]$ and we take $\varepsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$. For simplicity assume $\sigma^2 = 1$. Our goal is to estimate $f(x_0)$, the value of f at a given point, $x_0 \in [0,1]$. The loss function is squared-error loss:

$$L(\hat{f}(x_0), f(x_0)) = (\hat{f}(x_0) - f(x_0))^2. \tag{2}$$

Donoho et al. (1990), Brown and Low (1996) and Brown and Zhao (2001) show the following equivalence results. Suppose f is expressed by an orthonormal basis $\{\varphi_i(x)\}$ on $\mathscr{L}_2 = \{f\colon \int_0^1 f^2(x)\,\mathrm{d}x < \infty\}$, i.e.,

$$f(x) = \sum \theta_i \varphi_i(x).$$

Then we can construct $\{y_i\}$ as a (randomized) function of $\{z_i\}$ such that

$$y_i = \theta_i + \frac{1}{\sqrt{n}} \varepsilon_i, \quad i = 1, \ldots, \varepsilon_i \overset{\text{i.i.d.}}{\sim} N(0, 1). \tag{3}$$

Estimating a functional such as $f(x_0)$ is asymptotically equivalent to estimating the matching functional of $\theta = \{\theta_i\}$. (Brown and Zhao (2001) give an equivalence construction that is also valid if the x_i are themselves observations of i.i.d. random variables on $[0,1]$.)

To be explicit we take φ_i to be the usual Fourier basis on [0,1]. Thus, for $x \in [0,1]$

$$\varphi_0(x) = 1,$$
$$\varphi_{2k-1}(x) = 2^{-1/2}\cos(2\pi kx),$$
$$\varphi_{2k}(x) = 2^{-1/2}\sin(2\pi kx), \quad k = 1,2,\ldots .$$

Then, $\theta_i = \int_0^1 f(x)\varphi_i(x)\,dx$, $i = 0,\ldots$. If $\hat{\theta} = \{\hat{\theta}_i\}$ is an estimator of $\theta = \{\theta_i\}$ then

$$\hat{f}(x_0) = \sum a_i\hat{\theta}_i, \quad a_i = \varphi_i(x_0)$$

is the matching estimator of $f(x_0)$. This can conveniently be rewritten as $\hat{f}(x_0) = a'\hat{\theta}$ where $a = \{a_i\}$, $\hat{\theta} = \{\hat{\theta}_i\}$. If T denotes an estimator of $f(x_0)$ then the risk function is of course

$$R(T, f(x_0)) = E(T - f(x_0))^2.$$

When f is assumed to be in a Sobolev ball $S_q(B) = \{\{\theta_i\}: \sum i^{2q}\theta_i^2 \leqslant B\}$ when $q > \frac{1}{2}$ the optimal minimax rate in the present problem is known to be $n^{-(2q-1)/2q}$, i.e.,

$$0 < \inf_T \sup_{\theta\in S_q(B)} n^{(2q-1)/2q} R(T, f(x_0)) < \infty. \tag{4}$$

This rate was established by Wahba (1975); see also Donoho and Low (1992). We assume throughout that $q > \frac{1}{2}$.

3. Main results

How does a Bayesian estimator perform for this nonparametric point estimation problem? We are especially interested in the question of whether the Bayes solution resulting from a Gaussian prior possesses the optimal property defined in (4).

Zhao (2000) dealt with Bayesian estimation of the entire function. We establish similar results for estimating $f(x_0)$ under square error loss. The general results have points of similarity, but there are also some differences and the methods of proof are rather different.

A product Gaussian prior

$$\pi(\theta) = \prod N(0, \tau_i^2) \tag{5}$$

has support on $S_q \triangleq \{\{\theta_i\}: \sum i^{2q}\theta_i^2 < \infty\}$ if and only if

$$\sum_i i^{2q}\tau_i^2 < \infty.$$

It is straightforward to compute the posterior mean of the prior to be

$$\hat{\theta}_i = \frac{\tau_i^2}{\tau_i^2 + \frac{1}{n}}\, y_i. \tag{6}$$

For details of both assertions see Zhao (2000). The Bayes estimator of $f(x_0)$ is then easily calculated to be

$$\hat{T} = \sum a_i \hat{\theta}_i = a'\hat{\theta}. \tag{7}$$

We first show that there exist independent normal priors whose Bayes procedure attains the optimal minimax rate, but these priors are not supported on S_q.

Theorem 3.1. *Let $\tau_i^2 = i^{-2p}$ in* (5). *When $p > \max(q/2, \frac{1}{2})$ the Bayes estimator $\hat{T}$ in* (7) *has the minimax rate $n^{-m(p,q)}$, where $m(p,q) = \min(1 - 1/2p, (2q-1)/2p)$. To be more precise,*

$$0 < \lim_{n\to\infty} \sup_{\theta\in S_q(B)} n^{m(p,q)} R(a'\hat{\theta}, f(x_0)) < \infty.$$

In particular, the Bayes estimator attains the optimal minimax rate if and only if $p = q$.

Proof. Take $B = 1$ with no loss of generality. The risk function can be written as

$$R(a'\hat{\theta}, a'\theta) = \mathrm{Var}(a'\hat{\theta}) + \mathrm{Bias}^2(a'\hat{\theta}, a'\theta). \tag{8}$$

Note that if $b_i > 0$, then

$$\sup_{\sum b_i\theta_i^2 \leqslant 1} \left(\sum w_i\theta_i\right)^2 = \sum \frac{w_i^2}{b_i}. \tag{9}$$

Then, for the squared bias in (8)

$$\begin{aligned}\sup_{\sum i^{2q}\theta_i^2 \leqslant 1} \mathrm{Bias}^2(a'\hat{\theta}, a'\theta) &= \sup_{\sum i^{2q}\theta_i^2 \leqslant 1} \left[\sum a_i \left(\frac{\tau_i^2}{\tau_i^2 + 1/n}\theta_i - \theta_i\right)\right]^2 \\ &= \sup_{\sum i^{2q}\theta_i^2 \leqslant 1} \left[\sum a_i \frac{\theta_i}{1 + n\tau_i^2}\right]^2 \\ &= \sum a_i^2 \frac{(1 + ni^{-2p})^{-2}}{i^{2q}} \\ &\sim n^{(1-2q)/(2p)},\end{aligned} \tag{10}$$

where the last assertion comes from

$$\sum \frac{(1 + ni^{-2p})^{-2}}{i^{2q}} \sim n^{(1-2q)/(2p)} \tag{11}$$

and from

$$a_{2k-1}^2 + a_{2k}^2 = 1 \quad \forall k \geqslant 1. \tag{12}$$

For the variance term

$$\mathrm{Var}(a'\hat{\theta}) = \frac{1}{n}\sum a_i^2 \left(\frac{i^{-2p}}{i^{-2p} + \frac{1}{n}}\right) \sim n^{-(1-1/2p)},$$

where we have again used (12).

Hence, the minimax rate for $\mathbf{a}'\hat{\theta}$ is $n^{-\min(1-1/(2p),(2q-1)/2p)}$. And when $p=q$, $\min(1-1/(2p),(2q-1)/2p)$ achieves its maximum value $(2q-1)/2q$, for which the corresponding rate is just the optimal rate. □

Remark. Among priors with $\tau_i^2 = i^{-2p}$ the Bayes estimator is optimal only when $p=q$. But in that case both the prior and the posterior distribution have measure 0 on the space S_q of the interest.

The next theorem builds from Theorem 3.1 and gives a more general result about Bayesian approaches.

Theorem 3.2. *There does not exist a Gaussian prior on S_q such that the corresponding sequence of Bayes procedures attains the optimal minimax rate. That is, if Σ is the covariance matrix of a Gaussian measure on S_q, then the Bayes estimator $\hat{T}$ of $f(x_0)=a'\theta$ must have*

$$\lim_{n\to\infty} \sup_{\theta\in S_q(B)} n^{(2q-1)/2q} R(\hat{T}, a'\theta) = \infty. \tag{13}$$

Before we prove the above theorem let us derive some basic facts as lemmas.

Lemma 3.1. *Let D, W be positive definite $m\times m$ matrices and b an m dimensional vector. Then*

$$\sup_{\xi' D\xi\leqslant 1} (b'W\xi)^2 = b'WD^{-1}Wb$$

Proof. This standard result is the matrix generalization of (9). We omit the proof. □

Lemma 3.2. Let $\mathscr{P}_m = \{P$: *P is an $m\times m$ positive definite matrix with maximum eigenvalue* $<1\}$ *and $\vec{u}$ be a unit vector. Then*

$$\inf_{P\in\mathscr{P}_m} \{\vec{u}'P^2\vec{u} + \mathrm{Tr}(P^{-1}-I)\} > 0.889. \tag{14}$$

Proof. Write $P = O\Lambda O'$, with O some orthonormal matrix and $\Lambda = \mathrm{Diag}(\lambda_i)$. Here $\{\lambda_i\}$ are the eigenvalues of P. Let $\vec{v} = O'\vec{u}$. Then $\vec{v}$ is also a unit vector, i.e. $\sum v_i^2 = 1$. Then,

$$\begin{aligned}
&\vec{u}'P^2\vec{u} + \mathrm{Tr}(P^{-1}-I)\\
&\quad = \vec{v}'\Lambda^2\vec{v} + \sum \frac{1}{\lambda_i} - m\\
&\quad = \sum \lambda_i^2 v_i^2 + \sum \frac{1}{\lambda_i} - m.
\end{aligned}$$

Hence

$$\inf_{P\in\mathscr{P}_m} \vec{u}'P^2\vec{u} + \mathrm{Tr}(P^{-1}-I) \geqslant \inf_{0<\lambda_i\leqslant 1\ \sum v_i^2=1} \left\{\sum \lambda_i^2 v_i^2 + \sum \frac{1}{\lambda_i} - m\right\}. \tag{15}$$

If $\frac{1}{2} < v^2 \leqslant 1$ the function $\lambda^2 v^2 + 1/\lambda$ attains its minimum over $0 \leqslant \lambda \leqslant 1$ at $\lambda = 1/(2v^2)^{1/3}$. If $0 \leqslant v^2 \leqslant \frac{1}{2}$ then this function attains its minimum on this region at $\lambda = 1$. At most one v_j^2 can satisfy $v_j^2 > \frac{1}{2}$ since $\sum v_j^2 = 1$. Suppose there is one such v_j. Then

$$\inf_{0<\lambda_i \leqslant 1 \; \sum v_i^2=1} \left\{ \sum \lambda_i^2 v_i^2 + \sum \frac{1}{\lambda_i} - m \right\}$$

$$\geqslant \inf_{1/2 \leqslant v_j^2 \leqslant 1} (2^{-2/3} + 2^{1/3}) v_j^{2/3} - v_j^2$$

$$\geqslant \inf_{1/2 \leqslant v_j^2 \leqslant 1} (2^{-2/3} + 2^{1/3}) v_j^{2/3} - v_j^2$$

$$= 0.88988.$$

On the other hand, if all $v_j^2 \leqslant \frac{1}{2}$ then

$$\inf_{0<\lambda_i \leqslant 1 \; \sum v_i^2=1} \left\{ \sum \lambda_i^2 v_i^2 + \sum \frac{1}{\lambda_i} - m \right\}$$

$$\geqslant \sum v_j^2 + m - m = 1. \qquad \square$$

Proof. (1) It suffices to consider prior having mean 0. Since $f(x_0) = a'\theta$ the Bayes estimator will be $\hat{T} = a'\hat{\theta}$ with

$$\hat{\theta} = a' \sum \left(\frac{I}{n} + \Sigma \right)^{-1} Y.$$

Now,

$$\begin{aligned} R(\hat{T}, a'\theta) &\geqslant \mathrm{Bias}^2(\hat{T}, a'\theta) \\ &= \left[a' \left(\sum \left(\frac{I}{n} + \Sigma \right)^{-1} - I \right) \theta \right]^2 \\ &= (a'(I + n\Sigma)^{-1}\theta)^2 = (a'V\theta)^2 \end{aligned} \tag{16}$$

with V defined as

$$V = (I + n\Sigma)^{-1}. \tag{17}$$

Notice that all eigenvalues of V are between 0 and 1 and

$$\Sigma = n^{-1}(V^{-1} - I). \tag{18}$$

Given an integer m let D denote the $(m \times m)$ diagonal matrix with diagonal entries $(m+i)^{2q}$, $i = 1, \ldots, m$. Let W denote the $(m \times m)$ matrix composed of the $(m+1)$th, $\ldots 2m$th rows and columns of V and let b consist of the corresponding coordinates of a. Note that D, W, b all depend on m, but the dependence is suppressed in the notation. From (16) and (17),

$$\sup_{\sum i^{2q}\theta_i^2 \leqslant 1} \mathrm{Bias}^2(\hat{T}, a'\theta) \geqslant \sup_{\sum_{i=m+1}^{2m} i^{2q}\theta_i^2 \leqslant 1 : \theta_i = 0 \text{ if } i \notin [m+1, 2m]} \mathrm{Bias}^2(\hat{T}, a'\theta)$$

$$= \sup_{\xi' D \xi \leqslant 1} (b' W \xi)^2, \tag{19}$$

where ξ corresponds to the vector in the previous expression having coordinates $(\theta_{m+1}, \ldots, \theta_{2m})$.

Hence, by Lemma 3.1

$$R(\hat{T}, a'\theta) \geqslant b' W D^{-1} W b. \tag{20}$$

(2) Recall that $a_{2k-1}^2 + a_{2k}^2 = 1$, $k = 1, \ldots$. Hence in (20) $\|b\|^2 \sim m/2$. More precisely, for all $m > 100$

$$\|b\|^2 \geqslant 0.49m. \tag{21}$$

Suppose that the assertion of the theorem is false. Then by (20) there is a $c < \infty$ such that

$$n^{(2q-1)/(2q)} b' W D^{-1} W b \leqslant c. \tag{22}$$

Note that $D^{-1} - (2m)^{-2q} I$ is positive semi-definite. Hence (21) and (22) imply

$$n^{(2q-1)/(2q)} (0.49m) \frac{1}{(2m)^{2q}} \vec{u}' W^2 \vec{u} \leqslant c, \tag{23}$$

where $\vec{u} = b / \|b\|$ is a unit vector.

Now take

$$m = \left[n^{1/2q} \left(\frac{2^{2q}}{0.4 \times 0.49} c \right)^{1/(1-2q)} \right]. \tag{24}$$

Then from (23)

$$\vec{u}' W^2 \vec{u} \leqslant 0.4. \tag{25}$$

(3) Now, for m as in (24)

$$\sum_{i=m+1}^{2m} i^{2q} \Sigma_{ii} = \frac{1}{n} \sum_{i=m+1}^{2m} i^{2q} ((V^{-1})_{ii} - 1)$$

$$\geqslant \frac{1}{n} m^{2q} \sum_{i=m+1}^{2m} ((V^{-1})_{ii} - 1)$$

$$\geqslant c_1 \operatorname{Tr}(W^{-1} - I_m) \quad \text{for } (V^{-1})_{\text{diag}} > (V_{\text{diag}})^{-1} \tag{26}$$

with $c_1 > 0$, independent of m. Apply Lemma 3.2 and (25) to get

$$\operatorname{Tr}(W^{-1} - I_m) \geqslant 0.48.$$

Hence

$$\sum_{i=m+1}^{2m} i^{2q}\Sigma_{ii} \geqslant 0.48c_1. \tag{27}$$

As $n \to \infty$ there is an infinite sequence of arbitrarily large corresponding values of m given by (24). Thus (27) yields

$$\sum_{i=1}^{\infty} i^{2q}\Sigma_{ii} = \infty.$$

This establishes that the Gaussian prior is not supported on S_q. □

Remark. Note that the preceding proof establishes a slightly stronger fact than claimed in (13). Namely, for any Gaussian prior on S_q the squared bias does not converge at the optimal rate.

Acknowledgements

Research supported in part by NSF Grant DMS-9971848.

References

Brown, L.D., Low, M., 1996. Asymptotic equivalence of nonparametric regression and white noise. Ann. Statist. 24, 2384–2398.

Brown, L.D., Zhao, L., 2001. Direct asymptotic equivalence of nonparametric regression and the infinite dimensional location problem. Technical Report (available from www-stat.wharton.upenn.edu/~lzhao).

Cai, T.T., Low, M.G., Zhao, L., 2001. Tradeoffs between global and local risks in nonparametric function estimation. Technical Report (available from www-stat.wharton.upenn.edu/~lzhao).

Donoho, D.L., Liu, R., MacGibbon, B., 1990. Minimax risk over hyperrectangles and implications. Ann. Statist. 18, 1416–1437.

Donoho, D.L., Low, M.G., 1992. Renormalization exponents and optimal pointwise rates of convergence. Ann. Statist. 20, 944–970.

Pinsker, M.S., 1980. Optimal filtering of square integrable signals in Gaussian white noise. Probl. Reredachi Informatsii 16, 52–68 ((in Russian) (Probl. Imform. Transmission, 120–133)).

Wahba, G., 1975. Smoothing noisy data by spline functions. Numer. Math. 24, 383–393.

Zhao, L., 2000. Bayesian aspects of some nonparametric problems. Ann. Statist. 28, 532–552.