2000

Bayesian Aspects of Some Nonparametric Problems

Linda H. Zhao
University of Pennsylvania

Bayesian Aspects of Some Nonparametric Problems

Abstract
We study the Bayesian approach to nonparametric function estimation problems such as nonparametric regression and signal estimation. We consider the asymptotic properties of Bayes procedures for conjugate (= Gaussian) priors.

We show that so long as the prior puts nonzero measure on the very large parameter set of interest then the Bayes estimators are not satisfactory. More specifically, we show that these estimators do not achieve the correct minimax rate over norm bounded sets in the parameter space. Thus all Bayes estimators for proper Gaussian priors have zero asymptotic efficiency in this minimax sense.

We then present a class of priors whose Bayes procedures attain the optimal minimax rate of convergence. These priors may be viewed as compound, or hierarchical, mixtures of suitable Gaussian distributions.

Keywords
white noise, nonparametric regression, Bayes, minimax, conjugate priors

Disciplines
Statistics and Probability

The Annals of Statistics
2000, Vol. 28, No. 2, 532–552

BAYESIAN ASPECTS OF SOME NONPARAMETRIC PROBLEMS[1]

BY LINDA H. ZHAO

University of Pennsylvania

We study the Bayesian approach to nonparametric function estimation problems such as nonparametric regression and signal estimation. We consider the asymptotic properties of Bayes procedures for conjugate (=Gaussian) priors.

We show that so long as the prior puts nonzero measure on the very large parameter set of interest then the Bayes estimators are not satisfactory. More specifically, we show that these estimators do not achieve the correct minimax rate over norm bounded sets in the parameter space. Thus all Bayes estimators for proper Gaussian priors have zero asymptotic efficiency in this minimax sense.

We then present a class of priors whose Bayes procedures attain the optimal minimax rate of convergence. These priors may be viewed as compound, or hierarchical, mixtures of suitable Gaussian distributions.

1. Introduction. This paper investigates asymptotic properties of Bayes estimators in nonparametric regression and in the closely related nonparametric white noise with drift signal estimation problem. The primary problem considered is estimation of an unknown regression function under mean integrated squared error loss.

For most of our results we assume this regression function lies in a Sobolev space of functions of prescribed degree of smoothness. This is a common assumption for such problems, and minimax rates over bounded balls in this space are well known.

We first discuss the nature of conjugate priors having linear Bayes estimators. We next construct such a prior whose Bayes estimates attain the optimal minimax rate. See Theorem 5.1. However this prior is not supported on the assumed Sobolev parameter space. Instead, it is supported on a larger space, also of Sobolev type. Furthermore, the assumed parameter space has both prior and posterior probability zero under this prior. Consequently, this conjugate prior has the undesirable property that any Bayesian confidence regions for the regression function consist almost surely of functions outside the assumed parameter space.

We then investigate the question of whether there exists a Gaussian prior supported on the assumed parameter space whose Bayes estimates attain the optimal minimax rate. In a series of results that end with Theorem 5.2 we show that the answer to this question is negative.

---

[1]Supported in part by NSF Grant DMS-99-71848.

AMS 1991 *subject classifications*. Primary 62G07; secondary 62A15, 62G20.

Key words and phrases. White noise, nonparametric regression, Bayes, minimax, conjugate priors.

The preceding negative results lead to the question of whether there exists any prior distribution supported on the parameter space whose Bayes estimates attain the optimal minimax rate. In the latter part of this paper we show there do exist such priors. The priors we construct for this purpose are mixtures of Gaussian priors.

Ferguson (1973) was a relatively early attempt to provide a Bayesian framework for nonparametric problems. His approach is very useful for a certain subclass of nonparametric problems, but for various technical reasons (which are now well understood) it cannot be adapted to the types of curve estimation problems in which we are interested.

Wahba (1978, 1990) established that certain nonparametric spline estimators could be viewed as Bayesian estimators. The prior distributions appearing in her results depend on the sample size, n, and are not supported on the set of potential curves in the estimation problem.

Van der Linde (1993, 1995) contains additional discussion of smoothing splines as Bayes estimates. She points out that even though Wahba's prior and posteriors have measure zero on the function space of interest, the marginal distribution of the function evaluated at the design points can be used for the purpose of estimation at those points.

Cox (1993), Diaconis and Freedman (1986, 1998) and Freedman (1999) discuss aspects of the asymptotic behavior of Bayes estimates for a special subfamily of conjugate priors. In particular they find that Bayes estimators resulting from these priors exhibit an undesirable asymptotic property related to the usual Bernstein-von Mises criterion. These papers do not investigate minimax risk in the way we do, nor do they establish that the undesirable property they establish holds for all conjugate priors, as we do in our Theorem 5.2.

In familiar parametric problems conjugate priors have several well-recognized advantages as well as some disadvantages. See, for example, Berger [(1985), Chapter 4]. At least, in those problems such priors behave satisfactorily as the sample size n tends to infinity. Thus, the corresponding Bayes procedures are asymptotically consistent as well as asymptotic efficient and locally asymptotically minimax. See, for example, LeCam [(1986) especially Section 16.6].

In sections 2 and 3 we discuss the definition and construction of conjugate priors for our problems and the calculation of their Bayes estimates. The key reference here is Mandelbaum (1983) which builds on general results in Kuo (1975).

In Section 4 we study the consistency of Bayes procedures for conjugate priors. We find that a large subclass in fact yields consistent estimators. However the more diffuse of these priors actually lead to asymptotically inconsistent estimators.

Another feature already mentioned for parametric problems is the asymptotic efficiency of Bayes estimators. (Indeed, in that setting virtually all reasonable priors yield asymptotically efficient estimators.) One version of this efficiency in parametric problems is that the Bayes estimators are locally asymptotically minimax. A similar property can be studied for nonparametric problems. Specifically one can study formulations in which a nontrivial asymptotic minimax result is valid, and then ask whether Bayes estimators for conjugate priors have this asymptotic minimax property.

In Section 5 we study problems of this sort. We show that no Gaussian prior supported on the space of possible distributions yields asymptotically minimax estimators. Indeed, the negative result is somewhat more striking - not only are the Bayes estimators not locally asymptotically minimax (over norm-bounded sets in the parameter space), they do not even achieve the correct minimax rate. Thus they have zero asymptotic efficiency in this minimax sense.

In Section 6 we then construct a family of priors whose Bayes procedures do attain the minimax rate. These priors are carefully chosen mixtures of particular Gaussian priors. Theorem 6.1 establishes that these priors achieve the minimax rate so long as the mixing distribution has tails that do not decay faster than exponentially.

2. White noise model. Consider the following prototypical problem. Given n, we observe $\{Y(t)\}$ of the form

$$\mathbf{Y(t)} = \int_{-1}^{t} f(s)\,ds + \frac{\sigma}{\sqrt{n}} B_t\,, \tag{2.1}$$

where f belongs to some subset of $\mathscr{L}^2_{[-1,1]} = \{f : \int_{-1}^{1} f^2(t)dt < \infty\}$ and B_t is a Brownian motion, while $\sigma^2 > 0$ is known. We may assume $\sigma^2 = 1$. One wants to estimate some function of f, in terms of $\{Y(t)\}$. For example the entire functional f, or the function value at a point, say $f(0)$.

Equation (2.1) is often written as

$$dY(t) = f(t)dt + \frac{1}{\sqrt{n}} dB(t) \tag{2.2}$$

and is called the white noise model.

Much research has been conducted in this area. In particular, Brown and Low (1996) constructively showed under mild conditions that for any nonparametric regression problem there is a white noise problem which is asymptotically equivalent to it and conversely. Nussbaum (1996) and Klemelä and Nussbaum (1998) showed later (again under mild conditions) that a density estimation problem is also asymptotically equivalent to a corresponding white noise problem. See also Donoho and Liu (1991) and Donoho and Johnstone [(1998), Section 8]. These remarkable results make the white noise model more useful in addition to its own theoretical and applied importance.

In particular, the nature of the equivalence established in Brown and Low (1996) implies that all asymptotic results established here for the white noise model also hold for the standard nonparametric regression model where one observes

$$Y_i = f(i/n) + \varepsilon_i, \quad \varepsilon_i \overset{i.i.d}{\sim} N(0,1), \quad i = 1, \dots, n.$$

See especially Theorem 4.1 and Remark 4.3 in Brown and Low (1996).

The white noise model is equivalent to the normal mean problem. To be more precise, if we express $f(t)$ in terms of its Fourier series, then to estimate $f(t)$ is equivalent to estimating its Fourier coefficients. This connection was beautifully exploited in Donoho, Liu and MacGibbon (1993) to study the efficiency of linear minimax rules. By the properties of Brownian motion, the Fourier coefficients turn out to be normal means, as follows.

Let $\{\phi_i(t)\}_{i=1}^{\infty}$ be an orthonormal basis of $\mathscr{L}^2_{[-1,1]}$. Then $f(t)$ can be uniquely expressed as

$$f(t) = \sum \theta_i \, \phi_i(t),$$

where $\theta = \{\theta_i\} \in l_2 = \{\theta : \sum \theta_i^2 < \infty\}$.

In the white noise problem, we observe $Y(t)$. Let

$$y_i = \int_{-1}^{1} \phi_i(t) dY(t).$$

Then

$$y_i = \theta_i + \frac{1}{\sqrt{n}} \varepsilon_i$$

where

$$\varepsilon_i \overset{iid}{\sim} N(0, 1).$$

Furthermore, let $\tilde{f}(t)$ denote an estimator of $f(t)$ and let $\tilde{\theta}$ denote the corresponding estimator of θ. Thus

$$\tilde{\theta}_i = \int_{-1}^{1} \phi_i(t) \tilde{f}(t) dt.$$

Then

$$\begin{aligned} \|f(t) - \tilde{f}(t)\|^2 &= \int (f(t) - \tilde{f}(t))^2 dt \\ &= \|\theta - \tilde{\theta}\|^2 = \sum (\theta_i - \tilde{\theta}_i)^2. \end{aligned}$$

This shows that estimating the unknown function $f(t)$ in the white noise model based on observation of $\{Y(t)\}$ is the same as estimating the infinite dimensional normal mean vector $\{\theta_i\}$ based on observation of $\{y_i\}$, and $\mathscr{L}^2$ loss in the first problem is equivalent to ℓ_2 loss in the second.

3. Conjugate priors. In this section we consider observations of the random variable $\mathbf{Y} = (y_1, y_2, \dots)'$ with $y_i \overset{ind}{\sim} N(\theta_i, \sigma^2/n)$. $\sigma^2 > 0$ is known, and so we may as well assume $\sigma^2 = 1$. The parameter $\theta = (\theta_1, \theta_2, \dots)' \in \Theta$ is to be estimated under loss

$$L(\theta, \mathbf{a}) = \sum_i (\theta_i - a_i)^2, \tag{3.1}$$

which is equivalent to estimating f under the loss $\|\hat{f} - f\|^2$ in $\mathscr{L}^2_{[-1,1]}$.

We will establish a class of conjugate priors and find the Bayes solutions to the above problem. For completeness of the presentation we start with a general discussion of Gaussian measures on a Hilbert space H.

DEFINITION 3.1. A Gaussian measure μ on H is a Borel measure on H such that for each $x \in H$, the measurable function $< x, \cdot >$ is normally distributed, that is, there exist real numbers m_x and ${\sigma_x}^2$ such that $< x, \cdot > \sim N(m_x, {\sigma_x}^2)$.

We combine Lemma 2.1 and Theorem 2.3 in Kuo (1975) to get the following version of a theorem of Prohorov.

THEOREM 3.1 (PROHOROV). *An operator S corresponds to the covariance operator of a Gaussian measure on H if and only if*:

(i) *S is self-adjoint.*
(ii) *S is positive definite.*
(iii) $\sum_i < Se_i, e_i > < \infty$, *where $\{e_i\}$ is an orthonormal basis of H.*

Let ℓ_2 denote the Hilbert space

$$\ell_2 = \{\theta : \ \sum {\theta_i}^2 < \infty\}$$

with inner product $(\theta, \phi) = \theta' \phi$.

To any Gaussian measure on a Hilbert space $H \subset \{\theta : \theta = (\theta_1, \ldots)'\}$ there corresponds a covariance matrix $\Sigma = (\sigma_{ij})$ defined by

$$\sigma_{ij} = E(\theta_i - E(\theta_i))(\theta_j - E(\theta_j)).$$

DEFINITION 3.2. We say Σ is trace-class if

$$Tr\Sigma = \sum \sigma_{ii} < \infty.$$

Chapter 1.3 of Mandelbaum (1983) [or see Mandelbaum (1984)] yields

THEOREM 3.2. *Let π be a Gaussian measure on ℓ_2 with mean $m \in \ell_2$. Assume it has a trace-class covariance matrix Σ.*

(a) *Then the conditional distribution of θ given Y is a Gaussian measure on ℓ_2 with*

$$\hat{T} = E(\theta|Y) = \Sigma(\Sigma + I/n)^{-1} Y + (I - \Sigma(\Sigma + I/n)^{-1})m. \tag{3.2}$$

(b) *With probability* 1, $E(\theta|Y) \in \ell_2$. *This measure has a trace-class covariance matrix $\Sigma|Y = \Sigma(n\Sigma + I)^{-1}$.*

We have so far been using the term "conjugate prior" in either of two well accepted senses. One definition of the term is particularly suited to exponential families and refers to priors whose Bayes procedures are linear. See, for example, Diaconis and Ylvisaker (1979). Another definition of the term refers to any class of priors whose posterior distributions also lie within the same class. See, for example, Berger (1985). The following corollary verifies that both meanings of the term apply to the Gaussian priors in the present problem.

COROLLARY 3.1. *Let $\mathscr{C}$ denote the class of Gaussian measures on ℓ_2. Then $\mathscr{C}$ is a class of conjugate priors. The corresponding Bayes estimator is as in* (3.2).

4. Consistency of Bayes estimators. In finite dimensional parameter estimation, Bayesian analysis virtually always yields consistent estimators on the support of the prior. But Diaconis and Freedman (1986) pointed out the existence of inconsistent Bayes estimators in infinite dimensional estimation problems. See also Freedman (1999) and references therein. It then becomes necessary and useful too to consider the consistency of Bayes estimators in nonparametric, that is, infinite parametric, problems.

We will use the same notation as in the last section, that is, let

$$\begin{aligned} y_i &\overset{ind}{\sim} N\left(\theta_i, \frac{1}{n}\right), \quad i = 1, 2, \ldots, \\ \theta &= \{\theta_i\} \in \ell_2. \end{aligned}$$

DEFINITION 4.1. *An estimator T of θ is consistent if*

$$\|T - \theta\|^2 \xrightarrow{P} 0 \quad \forall \theta \in \ell_2.$$

We first give a result for a Gaussian prior whose covariance matrix, Σ, is trace-class. The corresponding Bayes estimator is shown to be consistent in the following theorem.

THEOREM 4.1. *Let $y_i \overset{ind}{\sim} N(\theta_i, \frac{1}{n})$, $\theta = \{\theta_i\} \in \ell_2$. Suppose the prior is Gaussian with mean in ℓ_2 and a trace-class covariance matrix Σ. Then the Bayes estimator is consistent.*

PROOF. It suffices to consider the case where the prior mean is zero and $\Sigma = \mathrm{d}iag(\Sigma_{ii})$.

Based on formula (3.2), the Bayes estimator T has

$$T_i = \frac{{\tau_i}^2 y_i}{{\tau_i}^2 + \frac{1}{n}} \tag{4.1}$$

where ${\tau_i}^2 = \Sigma_{ii}$. Note,

$$\begin{aligned} E_\theta \|T - \theta\|^2 &= E_\theta \sum (T_i - \theta_i)^2 \\ &= \sum var(T_i) + \sum (E(T_i) - \theta_i)^2. \qquad (4.2) \\ &= \sum_i \left(\frac{{\tau_i}^2}{{\tau_i}^2 + 1/n}\right)^2 \frac{1}{n} + \sum \left(\frac{1/n}{{\tau_i}^2 + 1/n}\right)^2 {\theta_i}^2 \to 0 \end{aligned}$$

since the dominated convergence theorem applies to both sums above. □

If Σ is not trace-class then the consistency conclusion of Theorem 4.1 can easily fail to hold. For this purpose, if M is positive definite, let H_M denote the Hilbert space

$$H_M = \{\theta : \quad \theta' M \theta < \infty\}$$

with inner product $< \theta, \phi >= \theta' M \phi$. Note that if the supremum of the eigenvalues of M is finite then $\theta \in \ell_2$ implies $\theta \in H_M$.

DEFINITION 4.2. *We say Σ is Hilbert–Schmidt if*

$$\mathrm{Tr}(\Sigma^2) < \infty.$$

Note that if Σ is Hilbert-Schmidt and μ is a Gaussian measure on H_Σ with covariance operator $S = \Sigma^2$ then its covariance matrix is Σ. [Condition (iii) of Theorem 3.1 is satisfied because Σ is Hilbert–Schmidt.] This observation combined with Chapter 1.3 of Mandelbaum's (1983) thesis [or see Mandelbaum (1984)] shows that if π is a Gaussian measure on H_Σ with mean $m \in \ell_2$ and Hilbert-Schmidt covariance matrix Σ, then the conditional distribution of θ given Y is a Gaussian measure on H_Σ with

$$E(\theta|Y) = \Sigma(\Sigma + I/n)^{-1} Y + (I - \Sigma(\Sigma + I/n)^{-1})\, m.$$

With probability one $E(\theta|Y) \in \ell_2$. This measure has Hilbert–Schmidt covariance matrix $\Sigma|Y = \Sigma(n\Sigma + I)^{-1}$.

THEOREM 4.2. *Suppose Σ is diagonal with $\Sigma_{ii} = {\tau_i}^2 = i^{-2p}$. Then for $1/4 < p \le 1/2$ the Bayes estimator is inconsistent.*

PROOF. For $1/4 < p \le 1/2$, Σ is Hilbert-Schmidt but not trace class. By (4.2) the variance term satisfies

$$\begin{aligned} \mathrm{var} &= \frac{1}{n} \sum_i \frac{i^{-4p}}{(1/n + i^{-2p})^2} \\ &\sim \frac{1}{n} \int_0^\infty \frac{x^{-4p}}{(1/n + x^{-2p})^2} dx \qquad (4.3) \\ &= \frac{1}{n^{1-\frac{1}{2p}}} \int_0^\infty \frac{1}{(1 + t^{2p})^2} dt\,. \end{aligned}$$

If $p = 1/2$ the right side converges to a finite non-zero number, and if $1/4 < p < 1/2$ it diverges to infinity.

By the Lindeberg–Lévy central limit theorem, $\sum (T_i - \theta_i)^2$ is asymptotically normal with mean $E(\sum (T_i - \theta_i)^2)$ as in (4.3) above. Its variance is

$$\begin{aligned} \sum \mathrm{Var}(T_i - \theta_i)^2 &= 2 \sum \frac{1}{n^2} \left(\frac{{\tau_i}^2}{{\tau_i}^2 + 1/n} \right)^4 \\ &\quad + 4 \sum \frac{1}{n^2} \left(\frac{1}{{\tau_i}^2 + 1/n} \right)^2 {\theta_i}^2 \frac{1}{n} \left(\frac{{\tau_i}^2}{{\tau_i}^2 + 1/n} \right)^2 . \end{aligned}$$

It is then easy to check that for $p > 1/4$

$$\lim_{n\to\infty} \sum \mathrm{Var}(T_i - \theta_i)^2 = 0 \quad \text{for } \theta \in \ell^2.$$

Hence $\|T - \theta\|^2$ has a mean bounded away from 0 and a variance which converges to 0. It follows that for every $\varepsilon > 0$ sufficiently small

$$\lim_{n\to\infty} Pr(\|T - \theta\|^2 < \varepsilon) = 0.$$

Hence the Bayes estimator is inconsistent. □

5. Some Bayesian results in function estimation. In this section, the parameter family will be restricted to Sobolev-type subspaces of ℓ_2. To be precise we will for convenience assume that

$$E_q = \left\{\{\theta_i\} : \sum i^{2q} \theta_i{}^2 < \infty\right\}, \quad q > \tfrac{1}{2}. \tag{5.1}$$

Minimax results hold over bounded balls in this space, which we denote as

$$E_q(B) = \left\{\{\theta_i\} : \sum i^{2q} \theta_i{}^2 \le B\right\}.$$

Under the usual Fourier basis, this essentially corresponds, when q is an integer, to the class of all periodic $f \in L_2$ with absolutely continuous $(q-1)$th derivative and whose qth derivative has uniformly bounded L_2 norm. Note that many other Sobolev subspaces can be written similarly as $\{\{\theta_i\} : \sum a_i \theta_i^2 < \infty\}$. Our results apply immediately to such spaces for which $a_i \sim ci^{2q}$. This generalization allows one for example to remove the above assumption of periodicity.

Estimating the entire function under integrated square loss is considered, that is, if $\hat\theta = \{\hat\theta_i\}$ is an estimator of θ, then the risk function is

$$R(\hat\theta, \theta) = E_\theta \sum_i (\hat\theta_i - \theta_i)^2. \tag{5.2}$$

The optimal rate in the minimax sense is well known to be $n^{-2q/(2q+1)}$, that is,

$$0 < \lim_{n\to\infty} \inf_{\hat\theta} \sup_{\theta \in E_q(B)} n^{2q/(2q+1)} R(\hat\theta, \theta) < \infty\,. \tag{5.3}$$

Details can be found in Ibragimov and Hasminskii (1977), Pinsker (1980), Donoho, Liu and MacGibbon (1990) or Zhao (1993). Our attempt is to try to find a Bayesian procedure attaining this rate. Surprisingly, it will be later seen that no conjugate prior supported on E_q can produce (5.3).We will proceed in steps to establish this result, looking at conjugate priors with progressively more general covariance functions.

We start with independent normal priors having a special power-variance structure. The first main result is that there exist priors of this sort whose Bayes procedure attains the optimal minimax rate, but these priors are not supported on E_q.

THEOREM 5.1. *Let*

$$\pi(\theta) = \prod_i \frac{1}{\sqrt{2\pi i^{-2p}}} \exp\left\{-\frac{{\theta_i}^2}{2i^{-2p}}\right\}. \tag{5.4}$$

When $p > \max(\frac{q}{2}, \frac{1}{2})$, *then the Bayes estimator*

$$\hat{T}_n = \left\{\hat{T}_{ni} = \frac{i^{-2p}}{i^{-2p} + 1/n} y_i\right\} \tag{5.5}$$

has the minimax rate $n^{-\min(1-1/(2p),\ q/p)}$, *that is,*

$$0 < \lim_n \sup_{\theta \in E_q(B)} n^{\min(1-1/(2p),\ q/p)} R(\hat{T}_n, \theta) < \infty. \tag{5.6}$$

In particular, if we take $p = q + \frac{1}{2}$, *then the Bayes estimator attains the optimal rate.*

PROOF. For simplicity, we take $B = 1$. By (4.2),

$$\sup_{\theta \in E_q(1)} R(\hat{T}_n, \theta) = \sup_{\theta \in E_q(1)} \left\{ \sum_i \left(\frac{1/n}{i^{-2p} + 1/n} \right)^2 {\theta_i}^2 + \sum_i \left(\frac{i^{-2p}}{i^{-2p} + 1/n} \right)^2 \frac{1}{n} \right\}. \tag{5.7}$$

Now

$$\sum_i \left(\frac{i^{-2p}}{i^{-2p} + 1/n} \right)^2 \frac{1}{n} \asymp n^{-(1-1/2p)} \tag{5.8}$$

and

$$\begin{aligned} &\sup_{\sum i^{2q} {\theta_i}^2 \le 1} \sum_i \left(\frac{1/n}{i^{-2p} + 1/n} \right)^2 {\theta_i}^2 \\ &\quad = \sup_{\sum i^{2q} {\theta_i}^2 \le 1} \sum \frac{\left(\frac{1/n}{i^{-2p}+1/n} \right)^2}{i^{2q}} i^{2q} {\theta_i}^2 \\ &\quad = \max_i \left\{ \frac{1}{(1 + n i^{-2p})^2 i^{2q}} \right\}. \end{aligned} \tag{5.9}$$

Finding the critical point of $1/(1 + n x^{-2p})^2 x^{2q}$, as long as

$$2p - q \ge 0\ , \tag{5.10}$$

shows that (5.9) will attain its maximum value at

$$i = \left[\frac{(2p - q)n}{q} \right]^{\frac{1}{2p}}, \tag{5.11}$$

where $[x]$ denote the largest integer less than or equal to x. Consequently, the maximum value in (5.9) is

$$\max_i \left\{ \frac{1}{(1 + n i^{-2p})^2 i^{2q}} \right\} = c n^{-q/p} \tag{5.12}$$

for some $c > 0$.

(5.8) and (5.12) lead to (5.6).

Finally, when $p = q + 1/2$, $\min(1 - 1/2p,\ q/2p) = 2q/(2q + 1)$. □

The above theorem and the fact in (5.3) yield a very plausible result: there is a Bayes estimator corresponding to independent normal priors which attains the optimal minimax rate. Unfortunately, such a prior, that is, (5.4) with $p = 2q + 1/2$, has measure 0 on the parameter space E_q. To see this note that as in Section 3, a positive definite, self-adjoint operator Σ corresponds to the covariance matrix of a Gaussian measure μ with mean zero on E_q, iff

$$\sum i^{2q}\Sigma_{ii} < \infty. \tag{5.13}$$

But when $\Sigma_{ii} = i^{-(2q+1)}$ as in Theorem 5.1, $\sum i^{2q}\Sigma_{ii} = \infty$.

This prior has a further undesirable property. From Theorem 3.2 note that the posterior distribution is normal with independent coordinates having variances

$$\Sigma_{ii} = \frac{i^{-(2q+1)}}{1 + ni^{-(2q+1)}}.$$

These posterior coordinate variances also do not satisfy (5.13) and hence the posterior also is not supported in E_q.

In summary, we have also proved the following negative result:

PROPOSITION 5.1. *There is no independent normal prior as defined in* (5.4) *with support on* E_q *such that the corresponding Bayes estimator attains the optimal rate as in* (5.3).

This leads us to work on a more general result.

PROPOSITION 5.2. *Let*

$$\pi(\theta) = \prod_i N(0, \tau_i^2) \tag{5.14}$$

such that

$$\sum i^{2q}\tau_i^2 < \infty\ . \tag{5.15}$$

Then the Bayes estimator

$$\hat{T} = \left\{ \frac{\tau_i^2 y_i}{\tau_i^2 + 1/n} \right\} \tag{5.16}$$

cannot attain the optimal rate $n^{-2q/(2q+1)}$, *that is,*

$$\limsup_{n\to\infty} n^{\frac{2q}{2q+1}} \sup_{\theta\in E_q(B)} R(\hat{T}, \theta) = \infty. \tag{5.17}$$

PROOF. Without loss of generality, we will assume $B = 1$ in $E_q(B)$. Since

$$\sup_{\theta\in E_q(1)} R(\hat{T}, \theta) = \sup_{\theta\in E_q(1)} \sum \left(\frac{1/n}{\tau_i^2 + 1/n} \right)^2 \theta_i^2 + \sum \left(\frac{\tau_i^2}{\tau_i^2 + 1/n} \right)^2 \frac{1}{n}.$$

It suffices to show that there is no $c > 0$ such that

$$\sup_{\theta \in E_q(1)} \sum \left(\frac{1/n}{\tau_i^2 + 1/n} \right)^2 \theta_i^2 \le c n^{-\frac{2q}{2q+1}} \tag{5.18}$$

is true for all sufficiently large n. Now

$$\begin{aligned} &\sup_{\theta \in E_q(1)} \sum \left(\frac{1/n}{\tau_i^2 + 1/n} \right)^2 \theta_i^2 \\ &\quad = \sup_{\theta \in E_q(1)} \sum \left(\frac{1/n}{\tau_i^2 + 1/n} \right)^2 i^{-2q} i^{2q} \theta_i^2 \\ &\quad = \sup_i \left(\frac{1/n}{\tau_i^2 + 1/n} \right)^2 i^{-2q} \\ &\quad = \frac{1}{\min_i (n\tau_i^2 + 1)^2 i^{2q}}. \end{aligned}$$

Suppose (5.18) were right. Then

$$\min_i (n\tau_i^2 + 1)^2 i^{2q} \ge \frac{1}{c} n^{\frac{2q}{2q+1}}.$$

Hence, for all large n, say $n \ge n_0$,

$$(n\tau_i^2 + 1)^2 i^{2q} \ge \frac{1}{c} n^{\frac{2q}{2q+1}} \tag{5.19}$$

for every i. This is equivalent to

$$\tau_i^2 \ge \frac{\frac{1}{\sqrt{c}} n^{\frac{q}{2q+1}} i^{-q} - 1}{n} . \tag{5.20}$$

Choose

$$n = \left[\left(\sqrt{c} \frac{2q+1}{q+1} \right)^{\frac{2q+1}{q}} i^{2q+1} \right].$$

Substitute this value in (5.20) to find

$$\tau_i^2 \ge (c_q + o(1)) i^{-(2q+1)}$$

as $i \to \infty$, where

$$c_q = \frac{q}{q+1} \left(\sqrt{c}\, \frac{2q+1}{q+1} \right)^{-\frac{2q+1}{q}} > 0.$$

Hence

$$\sum_{i \ge i_0} i^{2q} \tau_i^2 \ge \sum_i c_q \frac{1}{i} = \infty .$$

This contradicts the assumption (5.15) and leads the conclusion to that (5.18) cannot hold. □

The results so far are upsetting. Even worse, we will show next that there is no Gaussian prior supported on E_q whose corresponding Bayes estimator attains the optimal rate. Let us start by noting the following facts about matrices. See, for example, Brown [(1986), formula 1.17 (2)].

LEMMA 5.1. *Let the symmetric matrix P be positive definite and let P_D denote the diagonal matrix of P. Then*

$$(P^{-1})_D \geq (P_D)^{-1}. \tag{5.21}$$

If P is positive definite, then

$$(P^{-2})_D \geq (P_D)^{-2}\ . \tag{5.22}$$

We are now ready to state the penultimate theorem of this section.

THEOREM 5.2. *There is no Gaussian measure supported on E_q whose corresponding Bayes estimator attains the optimal rate. More precisely, if Σ is the covariance matrix of a Gaussian measure on E_q, then the Bayes estimator, $\hat{T}$, satisfies*

$$\limsup_{n\to\infty} n^{\frac{2q}{2q+1}} \sup_{\theta\in E_q(B)} R(\hat{T}, \theta) = \infty\ . \tag{5.23}$$

PROOF. It suffices to prove the result when the prior has mean zero and $B = 1$.

(i) From formula (3.2), when the prior has mean zero

$$\hat{T} = n(nI + \Sigma^{-1})^{-1} Y.$$

Since

$$n(nI + \Sigma^{-1})^{-1} - I = -(n\Sigma + I)^{-1},$$

it follows that

$$\mathrm{Bias}^2(\theta, \hat{T}) = \|(n\Sigma + I)^{-1}\theta\|^2. \tag{5.24}$$

Suppose (5.23) fails, that is,

$$n^{\frac{2q}{2q+1}} \sup_{\theta\in E_q(1)} R(\hat{T}, \theta) \leq c \tag{5.25}$$

for some constant $c > 0$ and all sufficiently large n. Then

$$n^{\frac{2q}{2q+1}} \sup_{\theta\in E_q(1)} \mathrm{Bias}^2(\theta, \hat{T}) \leq c,$$

that is,

$$n^{\frac{2q}{2q+1}} \sup_{\theta\in E_q(1)} \|(n\Sigma + I)^{-1}\theta\|^2 \leq c. \tag{5.26}$$

If we can show that

$$\sup_{\theta\in E_q(1)} \|(n\Sigma_D + I)^{-1}\theta\|^2 \le \sup_{\theta\in E_q(1)} \|(n\Sigma + I)^{-1}\theta\|^2, \tag{5.27}$$

then we will have

$$n^{\frac{2q}{2q+1}} \sup_{\theta\in E_q} \|(n\Sigma_D + I)^{-1}\theta\|^2 \le c\ . \tag{5.28}$$

Notice that $\|(n\Sigma_D + I)^{-1}\theta\|^2$ is the Bias^2 term of the Bayes estimator corresponding to the independent normal prior with variance Σ_{ii}. The variance term in the risk of the independent normal prior is of course the same as that corresponding to the original prior. Hence (5.28) will contradict Proposition 5.2.

(ii) It remains to show (5.27).

Let

$$E_0 = \left\{\theta : \theta_i = \frac{1}{i^q},\quad \theta_j = 0 \text{ if } j \ne i,\quad i = 1, 2, \ldots\right\} \subset E_q(1).$$

Let us denote

$$[(n\Sigma + I)^{-2}]_D = (\lambda'_{ii})$$

and

$$(n\Sigma_D + I)^{-2} = (\lambda_{ii})\ .$$

Then

$$\begin{aligned}
&\sup_{\theta\in E_q(1)} \|(n\Sigma + I)^{-1}\theta\|^2 \\
&\ge \sup_{\theta\in E_0} \|(n\Sigma + I)^{-1}\theta\|^2 \\
&= \sup_{\theta\in E_0} \|((n\Sigma + I)^{-1})_D\theta\|^2 \\
&= \max_i \frac{\lambda'_{ii}}{i^{2q}}.
\end{aligned} \tag{5.29}$$

On the other hand

$$\begin{aligned}
&\sup_{\theta\in E_q(1)} \|(n\Sigma_D + I)^{-1}\theta\|^2 \\
&= \sup_{\sum i^{2q}\theta_i^2\le 1} \sum \lambda_{ii}\theta_i^2 \\
&= \max_i \frac{\lambda_{ii}}{i^{2q}}\ .
\end{aligned} \tag{5.30}$$

By Lemma 5.1, we have

$$[(n\Sigma + I)^{-2}]_D \ge (n\Sigma_D + I)^{-2}, \tag{5.31}$$

that is,

$$\lambda'_{ii} \ge \lambda_{ii}\ .$$

(5.29) – (5.31) prove (5.27). This completes the proof of Theorem 5.2. □

REMARK (Generalization). The preceding results can be immediately generalized to hold for ellipsoids of the form $\{\{\theta_i\} : \sum a_i^2 \theta_i{}^2 \leq 1\}$ when $a_i \sim ci^q$. Because of this, our results also cover the white noise problem with $\{f\}$ consisting of $f \in L_2$ with absolutely continuous $(q-1)th$ derivative and qth derivative bounded in L_2. Other Sobolev type spaces are also included.

6. Compound priors. Because of the negative results in the previous section it is necessary to look for a prior supported on E_q whose Bayes estimator attains the optimal rate. We will demonstrate the existence and construct such a prior. The prior constructed has a special compound (or hierarchical) conjugate structure.

Take $\tau_i^2 = i^{-(2q+1)}$. Let G_k be the prior distribution function on E_q for which the coordinates, θ_i, are independent with $\theta_i \sim N(0, \tau_i^2)$ for $i \leq k$ and $\theta_i \sim N(0,0)$ for $i > k$. Each G_k is a conjugate, Gaussian prior.

Let $\{a_k\}$ be any sequence such that $\sum a_k = 1$ and for some $A < \infty, \quad A_1 > 0$

$$a_k \geq A_1 e^{-Ak} \quad k = 1, \ldots. \tag{6.1}$$

The prior of interest is the compound prior given by

$$G = \sum_{k=1}^{\infty} a_k G_k. \tag{6.2}$$

This prior can be considered as a hierarchical prior, with the value of k as the hierarchical index having the prior probabilities $P(K = k) = a_k$.

This collection of priors was introduced and discussed in Zhao (1996) to establish the optimal mean square error rate results given below. Shen and Wasserman (1998) have also used a subset of this collection of priors and refer to them as "sieve priors." They consider the posterior distribution under G, and show for each $\theta \in E_q$ that it converges in probability to a point mass at θ at the desired rate.

Let $\hat{\delta}_k$ denote the Bayes estimator for the prior G_k. The coordinates of $\hat{\delta}_k$ are given by

$$(\hat{\delta}_k)_i = \begin{cases} b_i y_i, & \text{if } i = 1, \ldots, k, \\ 0, & \text{if } i = k+1, \ldots, \end{cases} \tag{6.3}$$

where $b_i^{(n)} = \tau_i^2/(n^{-1} + \tau_i^2)$.

Since G is a hierarchical prior its Bayes estimator, $\hat{\theta}$, can be represented as a weighted average of the estimators $\hat{\delta}_k$ where the weights are from the posterior distribution of G_k given Y. Thus

$$\hat{\theta}_i^{(n)}(Y) = \sum_k w_k^{(n)}(Y) \hat{\delta}_{ki}(Y) \tag{6.4}$$

where

$$w_k^{(n)}(Y) = \frac{c_k^{(n)}(Y)}{\sum_j c_j^{(n)}(Y)},$$

with

$$c_j^{(n)}(Y) = a_j \int p_\psi^{(n)}(Y) G_j(d\psi) \tag{6.5}$$

where

$$p_\psi^{(n)}(Y) = \exp\left[n \sum(\psi_i y_i - \psi_i^2/2)\right].$$

In view of their special form (6.3) and (6.4) can be rewritten as

$$\hat{\theta}_i^{(n)}(Y) = W_i^{(n)}(Y) b_i^{(n)} y_i \tag{6.6}$$

with

$$W_i^{(n)}(Y) = \frac{\sum_{j \geq i} c_j^{(n)}(Y)}{\sum_{j=1}^{\infty} c_j^{(n)}(Y)}. \tag{6.7}$$

(For notational convenience, the dependence on n and/or Y of b_i, p_ψ, c_i and W_i, etc., will often be suppressed in later expressions.)

As previously noted, the minimax rate for this problem is $n^{-2q/(2q+1)}$. Our main result is that the Bayes procedure $\hat{\theta}$ for this prior achieves this rate.

THEOREM 6.1. *Let G be the compound prior described above. In particular, assume that the tail condition* (6.1) *is satisfied. Then its Bayes procedure*, $\hat{\theta}$, *achieves the minimax rate, that is, for any $B < \infty$,*

$$\lim_{n\to\infty} \sup_{\theta \in E_q(B)} n^{\frac{2q}{2q+1}} E(||\hat{\theta}^{(n)} - \theta||^2) < \infty. \tag{6.8}$$

PROOF. Routine computation yields

$$c_j(Y) = a_j \exp\left[\frac{1}{2} \sum_{i \leq j} \left(\frac{n\tau_i^2}{\frac{1}{n} + \tau_i^2} y_i^2 - \ln(1 + n\tau_i^2)\right)\right]. \tag{6.9}$$

Let $l_n = n^{1/(2q+1)}$. Choose $C > \max\{8A, 16(q+1)\}$. Fix $B < \infty$. Define $\lambda = \lambda^{(n)}(\theta)$ by

$$\lambda = \sup\left\{I : \sum_{i \geq I} \theta_i^2 \geq (C+B) n^{\frac{-2q}{2q+1}}\right\}. \tag{6.10}$$

Note that

$$\begin{aligned}\sum_{l_n}^{\infty} \theta_i^2 &\leq \frac{1}{l_n^{2q}} \sum_{i=1}^{\infty} i^{2q} \theta_i^2 \\ &\leq \frac{B}{l_n^{2q}} \\ &= Bn^{\frac{-2q}{2q+1}}.\end{aligned}$$

Hence for any $i \le \lambda^{(n)}(\theta)$,

$$V_i = \sum_{j=i}^{l_n} \theta_j^2 \ge Cn^{\frac{-2q}{2q+1}}. \tag{6.11}$$

[In particular, $\lambda^{(n)}(\theta) \le l_n$ for all $\theta \in E_q(B)$, and $V_\lambda \ge Cn^{-2q/(2q+1)}$.]

For any $i \le \lambda$,

$$\begin{aligned} W_i &= \sum_{j=i}^{\infty} \frac{c_j}{\sum_{j=1}^{\infty} c_j} \\ &\ge 1 - \sum_{j=1}^{i-1} \frac{c_j}{c_{l_n}} \end{aligned} \tag{6.12}$$

Note that the index j on the right of (6.12) satisfies $j \le \lambda - 1$. Apply Lemma A1 (Appendix). There is a constant γ_5 such that

$$P\left\{W_i \ge 1 - a_{l_n}^{-1} \exp\left(-\frac{nV_i}{8}\right)\right\} \ge 1 - \frac{\gamma_5}{nV_i}.$$

Because $\hat\theta_i = W_i b_i y_i$ as in (6.6), we apply Lemma A2 (Appendix) by taking $\alpha = \gamma_5/nV_i$ and $\beta_i = a_{l_n}^{-1}\exp(-nV_i/4)$. This yields that for some $A_2 < \infty$ and all $\theta \in E_q(B)$ and $i \le \lambda^{(n)}(\theta)$,

$$\begin{aligned} E_\theta(\hat\theta_i - \theta_i)^2 &\le 2\frac{b_i^2}{n} + 3(1-b_i)^2\theta_i^2 + 2a_{l_n}^{-2}\exp\left(-\frac{nV_i}{4}\right)\theta_i^2 + \frac{\gamma_5}{nV_i}\theta_i^2 \\ &\le 2\frac{b_i^2}{n} + 3(1-b_i)^2\theta_i^2 + A_2 n^{-\frac{2q}{2q+1}}\theta_i^2 + \frac{\gamma_5}{n} \end{aligned} \tag{6.13}$$

In verifying (6.13) we have used

$$\begin{aligned} 2a_{l_n}^{-2}\exp\left(-\frac{nV_i}{4}\right) &\le 2A_1^{-2}\exp\left(2Al_n - \frac{nV_\lambda}{4}\right) \quad (\text{since } \ i \le \lambda) \\ &\le 2A_1^{-2}\exp\left(2\frac{A}{C}nV_\lambda - \frac{n}{4}V_\lambda\right) \\ &\le 2A_1^{-2}\exp\left(-c_1 n^{-\frac{1}{2q+1}}\right) \le A_2 n^{-\frac{2q}{2q+1}} \end{aligned} \tag{6.14}$$

for some constants c_1 and A_2, since $nV_\lambda \ge Cn^{1/(2q+1)}$ and $\exp\{-c_1(n^{1/(2q+1)})\} = O(n^{-2q/(2q+1)})$. We have also used $V_i \ge \theta_i^2$.

Thus

$$\begin{aligned} \sum_{i=1}^{\lambda^{(n)}(\theta)} E_\theta(\hat\theta_i^2 - \theta_i)^2 \le& \sum_{i=1}^{\lambda^{(n)}} \left[2\frac{b_i^2}{n} + 3(1-b_i)^2\theta_i^2\right] + A_2 B n^{-2q/(2q+1)} \\ &+ \gamma_5 n^{-2q/(2q+1)} \end{aligned} \tag{6.15}$$

since $\sum \theta_i^2 \le B$ and $\lambda \le n^{1/(2q+1)}$.

The remaining part of the risk is easier to bound. Note first that

$$\sum_{i\geq\lambda^{(n)}(\theta)+1} \theta_i^2 < (C+B)n^{-2q/(2q+1)}. \tag{6.16}$$

Now apply Lemma A2 with $\beta = 1$ and $\alpha = 0$ to write for all $i \geq \lambda^{(n)}$

$$\begin{aligned} E(\hat{\theta}_i - \theta_i)^2 &= E(W_i b_i y_i - \theta_i)^2 \\ &\leq 2\frac{b_i^2}{n} + 3(1-b_i)^2\theta_i^2 + 2\theta_i^2 \\ &\leq 2\frac{b_i^2}{n} + 5\theta_i^2 \end{aligned} \tag{6.17}$$

Hence there is an $A_3 < \infty$ such that for all $\theta \in E_q(B)$ we have

$$\begin{aligned} \sum E(\hat{\theta}_i - \theta_i)^2 &\leq \sum\{2b_i^2/n + 3(1-b_i)\theta_i^2\} \\ &\quad + (A_2B + \gamma_5 + 5C + 5B)n^{-2q/(2q+1)} \\ &\leq A_3 n^{-2q/(2q+1)}. \end{aligned} \tag{6.18}$$

Here we use the fact from Theorem 5.1 that

$$E_\theta(||\{b_i y_i\} - \theta||^2) = \sum(b_i^2/n + (1-b_i)^2\theta_i^2) = O(n^{-2q/(2q+1)})$$

uniformly for $\theta \in E_q(B)$. $\square$

APPENDIX

The following two lemmas are needed in the proof of Theorem 6.1.

LEMMA A1. *For any $\theta \in E_q(B)$, if $i \leq \lambda$, defined as in* (6.10)*, then for some* $0 < \gamma_5 < \infty$

$$P\left\{W_i \leq 1 - a_{l_n}^{-1}\exp(-\frac{nV_i}{8})\right\} \leq \frac{\gamma_5}{nV_i}.$$

PROOF. By (6.9), for $j \leq \lambda - 1$ we have

$$\begin{aligned} \frac{c_j}{c_{l_n}} &= \frac{a_j}{a_{l_n}}\exp\left(-\frac{1}{2}\sum_{h=j+1}^{l_n}\left(\frac{n\tau_h^2}{\frac{1}{n}+\tau_h^2}y_h^2 - \ln(1+n\tau_h^2)\right)\right) \\ &\leq \frac{a_j}{a_{l_n}}\exp\left\{-\frac{1}{2}\sum_{h=j+1}^{l_n}\left(\frac{n\tau_h^2}{\frac{1}{n}+\tau_h^2}y_h^2\right) + \frac{1}{2}\sum_{h=1}^{l_n}\ln(1+n\tau_h^2)\right\} \end{aligned} \tag{A.1}$$

Now,

$$\sum_{h=1}^{l_n}\ln(1+n\tau_h^2) = n^{\frac{1}{2q+1}}\sum_{h=1}^{l_n} n^{-\frac{1}{2q+1}}\ln\left(1+x_i^{-(2q+1)}\right) \quad \left(\text{where } x_i = \frac{i}{n^{1/(2q+1)}}\right)$$

$$\leq n^{\frac{1}{2q+1}} \int_0^1 \ln(1+t^{-(2q+1)})dt \tag{A.2}$$
$$< 2(q+1)n^{\frac{1}{2q+1}}$$

Next, for given $\theta \in E_q(B)$ we write

$$(Y_i^{(n)})^2 = \theta_i^2 + \frac{1}{n} + R_i^{(n)} \tag{A.3}$$

where the

$$R_i^{(n)} = (Y_i^{(n)} - \theta_i)^2 + 2\theta_i(Y_i^{(n)} - \theta_i) - \frac{1}{n}$$

are independent random variables with

$$E\left(R_i^{(n)}\right) = 0, \quad \mathrm{Var}\left(R_i^{(n)}\right) = \frac{2}{n^2} + \frac{4\theta_i^2}{n}, \quad E\left((R_i^{(n)})^4\right) \leq \gamma_1\left(\frac{1}{n^4} + \frac{\theta_i^2}{n^3} + \frac{\theta_i^4}{n^2}\right),$$

where γ_1 is a constant independent of θ_i. It follows that

$$\begin{aligned} E\left(\left(\sum_{h=j+1}^{l_n} R_h\right)^4\right) &= \sum_{h=j+1}^{l_n} E(R_i)^4 + 6 \sum\sum_{j+1\leq i<h\leq l_n} E(R_i^2)E(R_h^2) \\ &\leq \gamma_3\left(\frac{l_n^2}{n^4} + \frac{l_n V_{j+1}}{n^3} + \frac{V_{j+1}^2}{n^2}\right). \end{aligned} \tag{A.4}$$

Hence the fourth-moment version of the Chebyshev–Markov inequality yields for $j \leq \lambda - 1$

$$\begin{aligned} P\left(\left|\sum_{h=j+1}^{l_n} R_h\right| > V_{j+1}/4\right) &\leq \frac{4^4\gamma_3}{V_{j+1}^4}\left(\frac{l_n^2}{n^4} + \frac{l_n V_{j+1}}{n^3} + \frac{V_{j+1}^2}{n^2}\right) \\ &\stackrel{\text{def}}{=} \tilde{\alpha}_n(V_{j+1}). \end{aligned} \tag{A.5}$$

Then for $i \leq \lambda$ we have

$$\begin{aligned} P\left(\bigcup_{j=1}^{i-1}\left\{\left|\sum_{h=j+1}^{l_n} R_h\right| > V_{j+1}/4\right\}\right) &\leq \sum_{j=1}^{i-1} \tilde{\alpha}_n(V_{j+1}) \\ &\leq \gamma_4\left[\frac{l_n^3}{n^4 V_i^4} + \frac{l_n^2}{n^3 V_i^3} + \frac{l_n}{n^2 V_i^2}\right] \\ &\leq \frac{\gamma_5}{n V_i} \end{aligned} \tag{A.6}$$

since $V_i \leq V_{j+1}$ for $j+1 \leq i$ and $l_n/nV_i \leq 1$ for $i \leq \lambda$.

Note that $\tau_h^2/(1/n+\tau_h^2) \geq 1/2$ for $h \leq l_n$ and that $C \geq 16(q+1)$, $nV_{j+1} \geq nV_\lambda \geq Cn^{1/(2q+1)}$, and hence

$$\begin{aligned}\sum_{h=j+1}^{l_n} \frac{n\tau_h^2}{\frac{1}{n}+\tau_h^2} y_h^2 &\geq \frac{n}{2} \sum_{h=j+1}^{l_n} y_h^2 \\ &\geq \frac{n}{2} \sum_{h=j+1}^{l_n} (\theta_h^2 + R_h) \\ &\geq \frac{n}{2}\left(\frac{3V_{j+1}}{4}\right)\end{aligned} \tag{A.7}$$

with probability $\geq 1-\gamma_5/nV_i$. From (A.1)–(A.8) it then follows that for all $\theta \in E_q(B)$ and $i \leq \lambda^{(n)}(\theta)$ it is true with probability $\geq 1-\gamma_5/nV_i$ that

$$W_i \geq 1 - a_{l_n}^{-1} \sum_{j=1}^{i-1} a_j \exp\{-nV_{j+1}/8\} \geq 1 - a_{l_n}^{-1} \exp\left(-\frac{nV_i}{8}\right). \qquad \square$$

LEMMA A2. *Let W, Y be jointly distributed real random variables. Suppose $Y \sim N(\theta, 1/n)$ and $0 \leq W \leq 1$ with*

$$Pr(W \geq 1-\beta) \geq 1-\alpha. \tag{A.8}$$

Let $0 \leq b \leq 1$ be a constant. Then

$$E(WbY-\theta)^2 \leq 2\frac{b^2}{n} + 3(1-b)^2\theta^2 + 2\beta^2\theta^2 + \alpha\theta^2. \tag{A.9}$$

PROOF. Let $S = \{(W,Y) : W \geq 1-\beta\}$. Then for $(W,Y) \in S$

$$(WbY-\theta)^2 \leq (bY-\theta)^2 + ((1-\beta)bY-\theta)^2$$

and for $(W,Y) \notin S$

$$(WbY-\theta)^2 \leq (bY-\theta)^2 + \theta^2.$$

Hence,

$$\begin{aligned}&E(WbY-\theta)^2 \\ &\quad\leq E\{(bY-\theta)^2\chi_S + ((1-\beta)bY-\theta)^2\chi_S + (bY-\theta)^2(1-\chi_S) + \theta^2(1-\chi_S)\} \\ &\quad\leq E\{(bY-\theta)^2 + ((1-\beta)bY-\theta)^2\} + \alpha\theta^2 \\ &\quad= b^2/n + (1-b)^2\theta^2 + (1-\beta)^2 b^2/n + (1-b(1-\beta))^2\theta^2 + \alpha\theta^2 \\ &\quad\leq 2b^2/n + 3(1-b)^2\theta^2 + 2\beta^2\theta^2 + \alpha\theta^2\end{aligned}$$

since

$$(1-b)^2 + (1-b(1-\beta))^2 \leq (1-b)^2 + 2(1-b)^2 + 2b^2\beta^2. \qquad \square$$

Acknowledgments. The author would like to thank all the referees and especially the Associate Editor for many very useful suggestions.

REFERENCES

BERGER, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*, 2nd ed. Springer, New York.

BROWN, L. D. (1986). *Fundamentals of Statistical Exponential Families with Applications in Statistical Decision Theory.* IMS, Hayward, CA.

BROWN, L. D. and LOW, M. G. (1996). Asymptotic equivalence of nonparametric regression and white noise. *Ann. Statist.* **24** 2384–2398.

COX, D. (1993). An analysis of Bayesian inference for nonparametric regression. *Ann. Statist.* **21** 903–923.

DIACONIS, P. and FREEDMAN, D. (1986). On the consistency of Bayes estimates (with discussion). *Ann. Statist.* **14** 1–67.

DIACONIS, P. and FREEDMAN, D. (1998). On the Bernstein-von Mises theorem with infinite dimensional parameters. Technical Report 492, Dept. Statistics, Univ. California, Berkeley

DIACONIS, P. and YLVISAKER, D. (1979). Conjugate priors for exponential families. *Ann. Statist.* **7** 269–281.

DONOHO, D. L. and JOHNSTONE, I. M. (1998). Minimax estimation via wavelet shrinkage. *Ann. Statist.* **26** 879–921.

DONOHO, D. L. and LIU, R. (1991). Geometrizing rates of convergence III. *Ann. Statist.* **19** 668–701.

DONOHO, D. L., LIU, R. and MACGIBBON, B. (1990). Minimax risk over hyperrectangles and implications. *Ann. Statist.* **18** 1416–1437.

DONOHO, D. L. and LOW, M. G. (1992). Renormalization exponents and optimal pointwise rates of convergence. *Ann. Statist.* **20** 944–970.

FERGUSON, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1** 209–230.

FREEDMAN, D. (1999). On the Bernstein-von Mises theorem with infinite dimensional parameters. *Ann. Statist.* **27** 1119–1140.

IBRAGIMOV, I. A. and HASMINSKII, R. Z. (1977). Estimation of infinite-dimensional parameter in white Gaussian noise. *Dokl. Akad. Nauk SSSR* **236** 1053–1056.

KLEMELÄ, J. and NUSSBAUM, M. (1998). Constructive asymptotic equivalence of density estimation and Gaussian white noise. Unpublished manuscript.

KUO, H. (1975). *Gaussian Measures in Banach Spaces. Lecture Notes in Math.* **403**. Springer, New York.

LECAM, L. (1986). *Asymptotic Methods in Statistical Decision Theory*. Springer, New York.

MANDELBAUM, A. (1983). Linear estimators of the mean of a Gaussian distribution on a Hilbert space. Ph. D. dissertation, Cornell Univ.

MANDELBAUM, A. (1984). All admissible linear estimators of the mean of a Gaussian distribution on a Hilbert space. *Ann. Statist.* **12** 1448–1466.

NUSSBAUM, M. (1996). Asymptotic equivalence of density estimation and white noise. *Ann. Statist.* **24** 2399–2430.

PINSKER, M. S. (1980). Optimal filtering of square integrable signals in Gaussian white noise. *Problemy Peredachi Informatsii* **16** 52–68 (in Russian). (Translation in *Problems Imform. Transmission* **16** 120–133.)

SHEN, X. and WASSERMAN, L. (1998). Rates of convergence of posterior distributions. Unpublished manuscript.

VAN DER LINDE, A. (1993). A note on smoothing splines as Bayesian estimates. *Statist. Decisions* **11** 61–67.

VAN DER LINDE, A. (1995). Splines from a Bayesian point of view. *Test* **4** 63–81.

WAHBA, G. (1978). Improper priors, spline smoothing and the problem of guarding against model errors in regression. *J. Roy. Statist. Soc. Ser. B* **40** 364–372.

WAHBA, G. (1990). *Spline Models for Observational Data.* SIAM, Philadelphia.

ZHAO, L. H. (1993). Frequentist and Bayesian Aspects of some nonparametric estimation problems. Ph.D. dissertation, Cornell Univ.

ZHAO, L. H. (1996). Bayesian aspects of some nonparametric estimation problems (abstract). *IMS Bulletin* **25** 20.

DEPARTMENT OF STATISTICS
3000 STEINBERG-DIETRICH HALL
UNIVERSITY OF PENNSYLVANIA
PHILADELPHIA, PENNSYLVANIA 19104-6302
E-MAIL: lzhao@wharton.upenn.edu