arXiv: math.PRxxxx.xxxxx

The Bayesian Analysis of Complex, High-Dimensional Models: Can it be CODA?

Y. Ritov[†], **P. J. Bickel**, **A. C. Gamst**[*], **B. J. K. Kleijn**,

Department of Statistics, The Hebrew University, 91905 Jerusalem, Israel; e-mail: yaacov.ritov@gmail.com*; url:* http://pluto.mscc.huji.ac.il/∼yaacov
Department of Statistics, University of California, Berkeley, CA 94720-3860, USA; e-mail: bickel@stat.berkeley.edu*; url:* http://www.stat.berkeley.edu/∼bickel
Biostatistics and Bioinformatics, University of California, San Diego, CA 92093-0717, USA; e-mail: acgamst@math.ucsd.edu*; url:* http://biostat.ucsd.edu/acgamst.htm
Korteweg-de Vries Institute for Mathematics, P.O.Box 94248, 1090 GE Amsterdam, The Netherlands; e-mail: B.Kleijn@uva.nl*; url:* http://staff.science.uva.nl/∼bkleijn/

Abstract: We consider the Bayesian analysis of a few complex, high-dimensional models and show that intuitive priors, which are not tailored to the fine details of the model and the estimated parameters, produce estimators which perform poorly in situations in which good, simple frequentist estimators exist. The models we consider are: stratified sampling, the partial linear model, linear and quadratic functionals of white noise, and estimation with stopping times. We present a strong version of Doob's consistency theorem which demonstrates that the existence of a uniformly $\sqrt{n}$-consistent estimator ensures that the Bayes posterior is $\sqrt{n}$-consistent for values of the parameter in subsets of prior probability 1. We also demonstrate that it is, at least, in principle, possible to construct Bayes priors giving both global and local minimax rates, using a suitable combination of loss functions. We argue that there is no contradiction in these apparently conflicting findings.

Keywords and phrases: Foundations, CODA, Bayesian inference, White noise models, Partial linear model, Stopping time, Functional estimation, Semiparametrics.

1. Introduction

We show, through a number of illustrative examples of general phenomena, some of the difficulties faced by application of the Bayesian paradigm in the analysis of data from complex, high-dimensional models. We do not argue against the use of Bayesian methods. However, we judge the success of these methods from the frequentist/robustness point of view, in the tradition of Bernstein, von Mises, and Le Cam; and more recently Cox (1993). Some references are: Bayarri and Berger (2004), Diaconis and Freedman (1993), Diaconis and Freedman (1998), Freedman (1963), Freedman (1999), Le Cam and Yang (1990), and Lehmann and Casella (1998).

[*]Research supported by grants from the NSF and DOE.
[†]Research supported by an ISF grant.

The extent to which the subjective aspect of data analysis is central to the modern Bayesian point of view is debatable. See the dialog between Goldstein (2006) and Berger (2006a) and the discussion of these two papers. However, central to any Bayesian approach is the posterior distribution and the choice of prior. Even those who try to reconcile Bayesian and frequentist approaches, cf. Bayarri and Berger (2004), in the case of conflict, tend to give greater weight on considerations based on the posterior distribution, than on those based on frequentist assessments; cf. Berger (2006b).

An older and by now less commonly held point of view, is that rational inquiry requires the choice of a Bayes prior and exclusive use of the resulting posterior in inference, cf. Savage (1961) and Lindley (1953). A modern weaker version claims: "Bayes theorem provides a powerful, flexible tool for examining the actual or potential ranges of uncertainty which arise when one or more individuals seek to interpret a given set of data in light of their own assumptions and 'uncertainties about their uncertainties'," Smith (1986). This point of view, which is the philosophical foundation of the Bayesian paradigm, has consequences. Among them are the strong likelihood principle, which says that all of the information in the data is contained in the likelihood function, and the stopping time principle, which says that stopping rules are irrelevant to inference. We argue that a commitment to these principles can easily lead to absurdities which are striking in high dimensions. We see this as an argument against ideologues.

We discuss our examples with these two types of Bayesian analysts in mind:

I. The Bayesian who views his prior entirely as reflecting his beliefs and the posterior as measuring the changes in these beliefs due to the data. Note that this implies strict adherence to the likelihood principle, a uniform plug-in principle, and the stopping time principle. Loss functions are not specifically considered in selecting the prior.

II. The pragmatic Bayesian who views the prior as a way of generating decision theoretic procedures, but is content with priors which depend on the data, insisting only that analysis starts with a prior and ends with a posterior.

For convenience we refer to these Bayesians as type I and type II.

The main difference we perceive between the type II Bayesian and a frequentist is that, when faced with a specific problem, the type II Bayesian selects a unique prior, uses Bayes rule to produce the posterior, and is then committed to using that posterior for all further inferences. In particular, the type II Bayesian is free to consider a particular loss function in selecting his prior and, to the extent that this is equivalent to using a data-dependent prior, change the likelihood; see Wasserman (2000). That the loss function and prior are strongly connected has been discussed by Rubin; see Bock (2004).

We show that, in high-dimensional (non or semi-parametric) situations Bayesian procedures based on priors chosen by one set of criteria, for instance, reference priors, selected so that the posterior for a possibly infinite dimensional parameter β converges at the minimax rate, can fail badly on other sets of criteria,

in particular, in yielding asymptotically minimax, semi-parametrically efficient, or even $\sqrt{n}$-consistent estimates for specific one-dimensional parameters, θ. We show by example that priors leading to efficient estimates of one-dimensional parameters can be constructed but that the construction can be subtle, and typically does not readily also give optimal global minimax rates for infinite dimensional features of the model. It is true, as we argue in the section 7, that by general considerations, Bayes priors giving minimax rates of convergence for the posterior distributions for both single or 'small' sets of parameters and optimal rates in global metrics can be constructed, in principle. Although it was shown in Bickel and Ritov (2003) that this can be done consistently with the "plug-in principle", the procedures optimal for the composite loss are not natural or optimal, in general, for either component. There is no general algorithm for constructing such priors and we illustrate the failure of classical type II Bayesian extensions (see below) such as the introduction of hyperparameters. Of course, Bayesian procedures are optimal on their own terms and we prove an extension of a theorem of Doob at the end of this paper which makes this point. As usual, the exceptional sets of measure zero in this theorem can be quite large in non-parametric settings.

For smooth, low-dimensional parametric models, the Bernstein-von Mises theorem ensures that for priors with continuous positive densities, all Bayesian procedures agree with each other and with efficient frequentist methods, asymptotically, to order $n^{-1/2}$; see, for example, Le Cam and Yang (1990). At the other extreme, even with independent and identically distributed data, little can be said about the extreme nonparametric model $\mathcal{P}$, in which nothing at all is assumed about the common distribution of the observations, P. The natural quantities to estimate, in this situation, are bounded linear functionals of the form $\theta = \int g(x)\,\mathrm{d}P(x)$, with g bounded and continuous. There are unbiased, efficient estimates of these functionals and Dirichlet process priors, concentrating on small but dense subsets of $\mathcal{P}$ yielding estimates equivalent to order $n^{-1/2}$ to the unbiased ones; see Ferguson (1973), for instance.

The interesting phenomena occur in models between these two extremes. To be able to even specify natural unbounded linear functionals such as the density p at a point, we need to put smoothness restrictions on P and, to make rate of convergence statements, global metrics such as L_2 must be used. Both Bayesians and frequentists must specify not only the structural features of the model but smoothness constraints. Some of our examples will show the effect of various smoothness assumptions on Bayesian inference.

For ease of exposition, in each of our examples, we consider only independent and identically distributed (i.i.d) data and our focus is on asymptotics and estimation. Although our calculations are given almost exclusively for specific Bayesian decision theoretic procedures under L_2-type loss, we believe (but do not argue in detail) that the difficulties we highlight carry over to other inference procedures, such as the construction of confidence regions. Here is one implication of such a result. Suppose that we can construct a Bayes credible region C for an infinite dimensional parameter β which has good frequentist and Bayesian properties, e.g. asymptotic minimax behavior for the specified model,

as well as $P(\beta \in C \mid X)$ and $P(\beta \in C(X) \mid \beta) > 1 - \alpha$. Then we automatically have a credible region $q(C)$ for any $q(\beta)$. Our examples will show, however, that this region can be absurdly large. So, while a Bayesian might argue that parameter estimation is less important than the construction of credible regions, our examples carry over to this problem as well.

Our examples will be discussed heuristically rather than exhaustively, but we will make it clear when a formal proof is needed. There is a body of theory in the area, cf. Ghosal et al. (2000), Kleijn and van der Vaart (2006), and Bickel and Kleijn (2012), among others, giving specific conditions under which some finite dimensional intuition persists in higher dimensions. However, in this paper we emphasize how easily these conditions are violated and the dramatic consequences of such violations. Our examples can be thought of as points of the parameter space to which the prior we use assigns zero mass. Since all points of the parameter space are similarly assigned zero mass, we have to leave it to the readers to judge whether these points are, in any sense, exceptional.

In Section 2, we review an example introduced in Robins and Ritov (1997). The problem is that of estimating a real parameter in the presence of an infinite dimensional "nuisance" parameter. The parameter of interest admits a very simple frequentist estimator which is $\sqrt{n}$-consistent without any assumptions on the nuisance parameters at all, as long as the sampling scheme is reasonable. In this problem, the type I Bayesian is unable to estimate the parameter of interest at the $\sqrt{n}$-rate at all, without making severe smoothness assumptions on the infinite dimensional nuisance parameter. In fact, we show that if the nuisance parameters are too rough, a type I Bayesian is unable to find any prior giving even a consistent estimate of the parameter of interest. On the other hand, we do construct priors, tailored to the parameter we are trying to estimate, which essentially reproduce the frequentist estimate. Such priors may be satisfactory to a type II Bayesian, but surely not to Bayesians of type I. The difficulty here is that a commitment to the strong likelihood principle forces the Bayesian analyst to ignore information about a parameter which factors out of the likelihood and he is forced to find some way of connecting that parameter to the parameter of interest, either through reparameterization, which only works if the nuisance parameter is smooth enough, or by tailoring the prior to the parameter of interest.

In Section 3, we turn to the classical partial linear regression model. We recall results of Wang et al. (2011) which give simple necessary and sufficient conditions on the nonparametric part of the model for the parametric part to be estimated efficiently. We use this example to show that a natural class of Bayes priors, which yield minimax estimates of the nonparametric part of the model under the conditions given in Wang et al. (2011), lead to Bayesian estimators of the parametric part which are inefficient. In this case, there is auxiliary information in the form of a conditional expectation which factors out of the likelihood but is strongly associated with the amount of information in the data about the parameter of interest. The frequentist can estimate this effect directly, but the type I Bayesian is forced to ignore this information and, depending on smoothness assumptions, may not be able to produce a consistent estimate of

the parameter of interest at all. The fact that, for a sieve-based frequentist approach, two different bandwidths are needed for local and global estimation of parameters in this problem has been known for some time; see Chen and Shiau (1994).

In Section 4, we consider the Gaussian white noise model of Ibragimov and Hasminskii (1984), Donoho and Johnstone (1994), and Donoho and Johnstone (1995). Here we show that from a frequentist point of view we can easily construct uniformly $\sqrt{n}$-consistent estimates of all bounded linear functionals. However, both the type I and type II Bayesian, who are restricted to the use of one and only one prior, must fail to estimate some bounded linear functionals at the $\sqrt{n}$-rate. This is because both are committed to the plug-in principle and, as we argue, any plug-in estimator will fail to be uniformly consistent. On the positive side, we show that it is easy to construct tailor-made Bayesian procedures for any of the specific functionals we consider in this section. Again, reparameterization, which in this case is a change of basis, is important. The resulting Bayesian procedures are capable of simultaneously estimating both the infinite dimensional features of the model at the minimax rate and the finite dimensional parameters of interest efficiently, but linear functionals which might be of interest in subsequent inferences and can not be estimated consistently remain. We give a graphic example, in this section, to demonstrate our claims.

A second example, examined in Section 5, concerns the estimation of the norm of a high-dimensional vector of means, $\boldsymbol{\beta}$. Again, for a suitably large set of $\boldsymbol{\beta}$, we can show that the priors normally used for minimax estimation of the vector of means in the L_2 norm do not lead to Bayesian estimators of the norm of $\boldsymbol{\beta}$ which are $\sqrt{n}$-consistent. Yet there are simple frequentist estimates of this parameter which are efficient. We then give a constructive argument showing how a type I Bayesian can bypass the difficulties presented by this model at the cost of selecting a non-intuitive prior and various inconsistencies. A type II Bayesian can use a data-dependent prior which allows for simultaneous estimation of $\boldsymbol{\beta}$ at the minimax rate and this specific parameter of interest efficiently. These examples show that, in many cases, the choice of prior is subtle, even in the type II context, and the effort involved in constructing such a prior seems unnecessary, given that good, general-purpose frequentist estimators are easy to construct for the same parameters.

In Section 6, we give a striking example in which, for Gaussian data with a high-dimensional parameter space, we can, given any prior, construct a stopping time such that the Bayesian, who must ignore the nature of the stopping times, estimates the vector of means with substantial bias. This is a common feature of all our examples. In high dimensions, even for large sample sizes, the bias induced by the Bayes prior overwhelms the data.

In Section 7 we extend Doob's theorem, showing that if a suitably uniform $\sqrt{n}$-consistent estimate of a parameter exists then necessarily the Bayesian estimator of the parameter is $\sqrt{n}$-consistent on a set of parameter values which has prior probability one. We also give another elementary result showing that it is in principle possible to construct Bayes priors giving both global and local minimax rates, using a suitable combination of loss functions. We summarize

our findings in Section 8.

In the appendix we give proofs of many of the assertions we have made in the previous sections. Throughout this paper, θ is a finite-dimensional parameter of interest, β is an infinite-dimensional nuisance parameter, and g is an infinite-dimensional parameter which is important for estimating θ efficiently, but is missing from the joint likelihood for (θ, β); g might describe the sampling scheme, the loss function, or the specific functional $\theta(\beta) = \theta(\beta, g)$ of interest. We use π for priors and g and β are given as $\mathbf{g}$ and $\boldsymbol{\beta}$ when it is easier to think of them as infinite-dimensional vectors than functions.

2. Stratified Random Sampling

Robins and Ritov (1997) consider an infinite-dimensional model of continuously stratified random sampling in which one has i.i.d. observations $W_i = (X_i, R_i, Z_i)$, $i = 1, \ldots, n$; the X_i are uniformly distributed in $[0,1]^d$; and $Z_i = R_i Y_i$. The variables R_i and Y_i are conditionally independent given X_i and take values in the set $\{0, 1\}$. The function $g(X) = \mathsf{E}(R|X)$ is known, with $g > 0$ almost everywhere, and $\beta(X) = \mathsf{E}(Y|X)$ is unknown. The parameter of interest is $\theta = \mathsf{E}(Y)$.

It is relatively easy to construct a reasonable estimator for θ in this problem. Indeed, the classical Horvitz-Thompson (HT) estimator, cf. Cochran (1977),

$$\widehat{\theta} = n^{-1} \sum_{i=1}^{n} Z_i / g(X_i),$$

solves the problem nicely. Because,

$$\begin{aligned} \mathsf{E}\{RY/g(X)\} &= \mathsf{E}\left\{\mathsf{E}(R|X)\mathsf{E}(Y|X)/g(X)\right\} \\ &= \mathsf{E}\mathsf{E}(Y|X) = \theta, \end{aligned}$$

the estimator is consistent without any further assumptions. If we assume that g is bounded from below, the estimator is $\sqrt{n}$-consistent and asymptotically normal.

2.1. Type I Bayesian Analysis

As g is known and we have assumed that the X_i are uniformly distributed, the only parameter which remains is β, where $\beta(X) = \mathsf{E}(Y|X)$. Let π be a prior density for β with respect to some measure μ. The joint density of β and the observations $W_1, \ldots, W_n$ is given by

$$\begin{aligned} p(\beta, \mathbf{W}) = \pi(\beta) & \prod_{i\,:\,R_i=1} \beta(X_i)^{Y_i} \left(1 - \beta(X_i)\right)^{1-Y_i} \\ & \times \prod_{i=1}^{n} g(X_i)^{R_i} \left(1 - g(X_i)\right)^{1-R_i}, \end{aligned}$$

as $Z_i = Y_i$ when $R_i = 1$. But this means that the posterior for β has a density $\pi(\beta|\mathbf{W})$ with,

$$\pi(\beta|\mathbf{W}) \propto \pi(\beta) \prod_{i\,:\,R_i=1} \beta(X_i)^{Y_i} \left(1-\beta(X_i)\right)^{1-Y_i}. \tag{1}$$

Of course, this is a function of only those observations for which $R_i = 1$, i.e. for which the Y_i are directly observed. The observations for which $R_i = 0$ are deemed uninformative.

If β is assumed to range over a smooth parametric model, and the known g is bounded away from 0, one can check that the Bernstein-von Mises theorem applies, and that the Bayesian estimator of θ is efficient, $\sqrt{n}$-consistent and necessarily better than the HT estimator. Heuristically, this continues to hold for minimax estimation of θ and β over "small" nonparametric models for β; that is, sets of very smooth β; see Bickel and Kleijn (2012).

In the nonparametric case, if we assume that the prior for β does not depend on g, then, because the likelihood function does not depend on g, the type I Bayesian will use the same procedure whether g is known or unknown, see (1). That is, the type I Bayesian will behave as if g were unknown. This is problematic because, as Robins and Ritov (1997) argued and we now show, unless β or g are sufficiently smooth, the type I Bayesian can not produce a consistent estimator of θ. To the best of our knowledge, the fact that there is no consistent estimator of θ when g is unknown, unless β or g are sufficiently smooth, has not been emphasized before.

Note that our assumption that the prior for β does not depend on g is quite plausible. Consider, for example, an in-depth survey of students, concerning their scholastic interests. The design of the experiment is based on all the information the university has about the students. However, the statistician is interested only in whether a student is firstborn or not. At first, he gets only the list of sampled students with their covariates. At this stage, he specifies his prior for β. If he is now given g, there is no reason for him to change what he believes about β, and no reason for him to include information about g in his prior.

The fact that, if g is unknown, θ cannot be estimated unless either g or β is smooth enough, is true even in the one-dimensional case. Our analysis is similar to that in Robins et al. (2009). Suppose the X_i are uniformly distributed on the unit interval, and g is given by,

$$g(x) = \frac{1}{2} + \frac{1}{4} \sum_{i=0}^{m-1} s_i\, \psi\left(mx - i\right),$$

where $m = m_n$ is such that $m_n/n \to \infty$; the sequence $s_1, \ldots, s_m \in \{-1, 1\}$ is assumed to be exchangeable with $\sum s_i = 0$, and $\psi(x) = \mathbf{1}\left(0 \le x < \frac{1}{2}\right) - \mathbf{1}\left(\frac{1}{2} \le x < 1\right)$. Furthermore, assume that $\beta(x) \equiv 5/8$ or $\beta(x) \equiv g(x)$. With probability converging to 1, there will be no interval of length $1/m$ with more than one X_i. However, given that there is one $X_i \in (j/m, (j+1)/m)$, then the

distribution of (R_i, Z_i) is the same whether $\beta(x) \equiv 5/8$ or $\beta(x) \equiv g(x)$, and hence θ is not identifiable; it can be either 5/8 or 1/2. This completes the proof.

Note that, in principle, both $\mathsf{E}(YR|X) = \beta(X)g(X)$ and $E(R|X) = g(X)$ are, in general, estimable, but not uniformly to adequate precision on "rough" sets of (g, β). One can also reparameterize in terms of $\xi(X) = \mathsf{E}(YR|X)$ and θ. This forces g into the likelihood, but one still needs to assume $\xi(X)$ is very smooth. In the above argument, the roughness of the model goes up with the sample size, and this is what prevents consistent estimation.

2.2. Bayesian Procedures with Good Frequentist Behavior

In this section we study plausible priors for Type II Bayesian inference. These priors are related to those in Wasserman (2004), Harmeling and Toussaint (2007), and Li (2010). We need to build knowledge of g into the prior, as we argued in Section 2.1. We do so first by following the suggestion in Harmeling and Toussaint (2007) for Gaussian models.

Following Wasserman (2004), we consider now a somewhat simplified version of the continuously stratified random sampling model, in which the X_i are uniformly distributed on $1, \ldots, N$, with $N = N_n \gg n$, such that with probability converging to 1, there are no ties. In this case, the unknown parameter β is just the N-vector, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_N)$. Our goal is to estimate $\theta = N^{-1} \sum_{i=1}^{N} \beta_i$.

To construct the prior, we proceed as follows. Assume that the components β_i are independent, with β_i distributed according to a Beta distribution with parameters $p_\tau(i)$ and $1 - p_\tau(i)$, and

$$p_\tau(i) = \frac{e^{\tau/g_i}}{1 + e^{\tau/g_i}},$$

with τ an unknown hyperparameter. Let $\theta^* = N^{-1} \sum_{i=1}^{N} p_\tau(i)$. Note that under the prior $\theta = N^{-1} \sum_{i=1}^{N} \beta_i = \theta^* + O_P(N^{-1/2})$, by the CLT. We now aim to estimate θ^*. In the language of Lindley and Smith (1972), we shift interest from a random effect to a fixed effect. This is level 2 analysis in the language of Everson and Morris (2000). The difference between θ and θ^* is apparent in a full population analysis, e.g., Berry et al. (1999) and Li (1999), where the real interest is in θ^*.

In this simplified model, marginally, $X_1, \ldots, X_n$ are i.i.d. uniform on $1, \ldots, N$, Y_i and R_i are independent given X_i, with $Y_i|X_i \sim \text{Binomial}\,(1, p_\tau(X_i))$, and $R_i|X_i \sim \text{Binomial}\,(1, g(X_i))$. The log-likelihood function for τ is given by,

$$\ell(\tau) = \sum_{R_i=1} \left[Y_i \log p_\tau(X_i) + (1 - Y_i) \log\left(1 - p_\tau(X_i)\right) \right].$$

This is maximized at $\hat{\tau}$ satisfying,

$$\begin{aligned}
0 &= n^{-1} \sum_{R_i=1} \left(Y_i \frac{\dot{p}_{\hat{\tau}}(X_i)}{p_{\hat{\tau}}(X_i)} - (1-Y_i) \frac{\dot{p}_{\hat{\tau}}(X_i)}{1-p_{\hat{\tau}}(X_i)} \right) \\
&= n^{-1} \sum_{R_i=1} \frac{\dot{p}_{\hat{\tau}}(X_i)}{p_{\hat{\tau}}(X_i)\left(1-p_{\hat{\tau}}(X_i)\right)} \left(Y_i - p_{\hat{\tau}}(X_i)\right) \\
&= n^{-1} \sum_{R_i=1} \left(Y_i - p_{\hat{\tau}}(X_i)\right) / g(X_i) \\
&= \hat{\theta}_{HT} - \frac{1}{n} \sum_{i=1}^{n} \frac{R_i}{g(X_i)} p_{\hat{\tau}}(X_i),
\end{aligned}$$

where $\dot{p}_\tau$ is the derivative of p_τ with respect to τ. A standard Bernstein-von Mises argument shows that $\hat{\tau}$ is within $o_P(n^{-1/2})$ of the Bayesian estimator of τ, thus $\hat{\theta}^*_B$, the Bayesian estimator of θ^*, satisfies:

$$\begin{aligned}
\hat{\theta}^*_B &= \frac{1}{N} \sum_{i=1}^{N} p_{\hat{\tau}}(i) + o_P\left(n^{-1/2}\right) \\
&= \frac{1}{n} \sum_{i=1}^{n} \frac{R_i}{g(X_i)} p_{\hat{\tau}}(X_i) + O_P\left(n^{-1/2}\right) \\
&= \hat{\theta}_{HT} + O_P\left(n^{-1/2}\right).
\end{aligned}$$

(where O_P and o_P are evaluated under the population model).

The estimator presented in Li (2010) is somewhat similar; however, his estimator is inconsistent, in general, and consistent only if $\mathsf{E}(Y|R=1) = \mathsf{E}Y$ (as, in fact, his simulations demonstrate).

With this structure, it is unclear how to define sets of β on which uniform convergence holds. This construction merely yields an estimator equivalent to the nonparametric HT estimator.

This prior produces a good estimator of θ^* but, for other functionals, e.g. $\mathsf{E}\left(Y|g(X) > a\right)$ or $\mathsf{E}\left(\boldsymbol{\beta}'\boldsymbol{\beta}\right)$, the prior leads to estimators which aren't even consistent. So, if we are stuck with the resulting posterior, as a type II Bayesian would be, we have solved the specific problem with which we were faced at the cost of failing to solve other problems which may come to interest us.

3. The Partial Linear Model

In this section we consider the partial linear model, also known as the partial spline model, which was originally discussed in Engle et al. (1986); see also Schick (1986). In this case, we have observations $W_i = (X_i, U_i, Y_i)$ such that,

$$Y_i = \theta X_i + \beta(U_i) + \varepsilon_i. \tag{2}$$

where the (X_i, U_i) form an i.i.d. sample from a joint density $p(x, u)$, relative to Lebesgue measure on the unit square, $[0, 1]^2$; β is an element of some class of functions $\mathcal{B}$; and the ε_i are i.i.d. standard-normal. The parameter of interest is θ and β is a (possibly very non-smooth) nuisance parameter. Let $g(U) = \mathsf{E}(X \mid U)$. For simplicity, assume that U is known to be uniformly distributed on the unit interval.

3.1. The Frequentist Analysis

Up to a constant, the log-likelihood function equals,

$$\ell(\theta, h, p) = -\left(y - \theta x - \beta(u)\right)^2/2 - \log p(x, u).$$

It is straightforward to argue that the score function for θ, the derivative of the log-likelihood in the least favorable direction for estimating θ, cf. Schick (1986) and Bickel et al. (1998), is given by,

$$\tilde{\ell}_\theta(\theta, h) = (x - g(u))\left(y - \theta x - \beta(u)\right) = (x - g(u))\,\varepsilon,$$

and that the semiparametric information bound for θ is,

$$I = \mathsf{E}\left[\mathrm{var}(X|U)\right].$$

We assume that $I > 0$ (which implies, in particular, that X is not a function of U). Regarding estimation of θ, intuition based on (2) says that for small neighborhoods of u, the conditional expectation of Y given X is linear with intercept $\beta(u)$, and slope θ which does not depend on the neighborhood. The efficient estimator should average the estimated slopes over all such neighborhoods.

Indeed, under some regularity conditions, an efficient estimator can be constructed along the following lines. Find initial estimators $\tilde{g}$ and $\tilde{\beta}$ of g and β, respectively, and estimate θ by computing,

$$\hat{\theta} = \frac{\sum\left(X_i - \tilde{g}(U_i)\right)\left(Y_i - \tilde{\beta}(U_i)\right)}{\sum\left(X_i - \tilde{g}(U_i)\right)^2}.$$

The idea here is that θ is the regression coefficient associated with regressing Y on X, conditioning on the observed values of U. In order for this estimator to be $\sqrt{n}$-consistent (or minimax), we need to assume that the functions g and β are smooth enough that we can estimate them at reasonable rates.

We could, for example, assume that the functions β and g satisfy Hölder conditions of order α and and δ, respectively. That is, there is a constant $0 \leq C < \infty$ such that $|\beta(u) - \beta(v)| \leq C|u - v|^\alpha$ and $|g(u) - g(v)| \leq C|u - v|^\delta$ for all u, v in the support of U. We also need to assume that $\mathrm{var}(X|U)$ has a version which is continuous in u. In this case, it is proved in Wang et al. (2011) that a necessary and sufficient condition for the existence of a $\sqrt{n}$-consistent and semiparametrically efficient estimator of θ is that $\alpha + \delta > 1/2$.

3.2. The Type I Bayesian Analysis

We assume that the type I Bayesian places independent priors on $p(u,x)$, β and θ, $\pi = \pi_p \times \pi_\beta \times \pi_\theta$. For example, the prior on the joint density may be a function of the environment, the prior on the nonparametric regression function might be a function of an underlying physical process, and the third component of the prior might reflect our understanding of the measurement engineering. We have already argued that such assumptions are plausible. The log-posterior-density is then given by,

$$-\sum_{i=1}^{n}\left(Y_i-\theta X_i-\beta(U_i)\right)^2/2+\log\pi_\theta(\theta)+\log\pi_\beta(\beta) \\ +\sum_{i=1}^{n}\log p(U_i,X_i)+\log\pi_p(p)+A,$$

where A depends on the data only. Note that the posterior for (θ,β) does not depend on p. The type I Bayesian would use the same estimator regardless of what is known about the smoothness of g.

Suppose now that, essentially, it is only known that β is Hölder of order α, while the range of U is divided up into intervals such that, on each of them, g is either Hölder of order δ_0 or of order δ_1, with,

$$\alpha+\delta_0<1/2<\alpha+\delta_1.$$

A $\sqrt{n}$-consistent estimator of θ can only make use of data from the intervals on which g is Hölder of order of δ_1. The rest should be discarded. *Suppose these intervals are disclosed to the statistician.* If the number of observations in the "good" intervals is of the same order as n, then the estimator is still $\sqrt{n}$-consistent. For a frequentist, there is no difficulty in ignoring the nuisance intervals – θ is assumed to be the same everywhere. However, the type I Bayesian cannot ignore these intervals. In fact, his *posterior* distribution can not contain any information on which intervals are good and which are bad.

More formally, let us consider a discrete version of the partial linear model. Let the observations be $Z_i=(X_{i1},X_{i2},Y_{i1},Y_{i2})$, with $Z_1,\ldots,Z_n$ independent. Suppose,

$$\begin{aligned} X_{i1} &\sim N(g_i,1),\\ X_{i2} &\sim N(g_i+\eta_i,1),\\ Y_{i1} &= \theta X_{i1}+\beta_i+\varepsilon_{i1},\\ Y_{i2} &= \theta X_{i2}+\beta_i+\mu_i+\varepsilon_{i2},\\ \varepsilon_{i1},\varepsilon_{i2} &\overset{\text{iid}}{\sim} N(0,1), \end{aligned}$$

where $X_{i1},X_{i2},\varepsilon_{i1},\varepsilon_{i2}$ are all independent, while g_i,η_i,β_i, and μ_i are unknown parameters. We assume that under the prior $(g_1,\eta_1),\ldots,(g_n,\eta_n)$ are i.i.d. independent of θ and the $(\beta_1,\mu_1),\ldots,(\beta_n,\mu_n)$ are i.i.d. This model is connected to

the continuous version, by considering isolated pairs of observations in the model with values differing by $O(1/n)$. The Hölder conditions become $\eta_i = O_P(n^{-\delta_i})$, and $\mu_i = O_P(n^{-\alpha})$, where $\delta_i \in \{\delta_0, \delta_1\}$, as above.

From a frequentist point of view, the $(X_{i1}, X_{i2}, Y_{i1}, Y_{i2})$ have a joint normal distribution and we would then consider the statistic,

$$\begin{bmatrix} X_{i2} - X_{i1} \\ Y_{i2} - Y_{i1} \end{bmatrix} \sim N\left(\begin{bmatrix} \eta_i \\ \theta\eta_i + \mu_i \end{bmatrix}, \begin{bmatrix} 2 & 2\theta \\ 2\theta & 2\theta^2 + 2 \end{bmatrix} \right).$$

Now consider the estimator,

$$\begin{aligned} \hat{\theta} &= \frac{\sum_{\delta_i=\delta_1}(X_{i2} - X_{i1})(Y_{i2} - Y_{i1})}{\sum_{\delta_i=\delta_1}(X_{i2} - X_{i1})^2} \\ &= \theta + \frac{\sum_{\delta_i=\delta_1}(X_{i2} - X_{i1})(\varepsilon_{i2} - \varepsilon_{i1})}{\sum_{\delta_i=\delta_1}(X_{i2} - X_{i1})^2} + \frac{\sum_{\delta_i=\delta_1}(X_{i2} - X_{i1})\mu_i}{\sum_{\delta_i=\delta_1}(X_{i2} - X_{i1})^2} \\ &= \theta + O_P\left(n^{-1/2}\right) + R, \end{aligned}$$

where,

$$R = \frac{\sum_{\delta_i=\delta_1} \eta_i\mu_i}{\sum_{\delta_i=\delta_1}(X_{i2} - X_{i1})^2} = o_P\left(n^{-1/2}\right),$$

since $\alpha + \delta_1 > 1/2$.

Note that if the sum were over all pairs, and if the number of pairs with $\delta_i = \delta_0$ is of order n, then the estimator would not be $\sqrt{n}$-consistent, since now $\sqrt{n}R$ may diverge, almost surely. In general, this model involves $2n + 1$ parameters and the parameter of interest can not be estimated consistently unless the nuisance parameters can be ignored, at least, asymptotically. However, these parameters can only be ignored if we consider the smooth pairs – that is, those pairs for which $\alpha + \delta_i > 1/2$, making the connection between variability, here, and smoothness, in the first part of this section. Of course, the information on which pairs to use in constructing the estimator is unavailable to the type I Bayesian.

The type I Bayesian does not find any logical contradiction in this failure. The parameter combinations on which the Bayesian estimator fails to be $\sqrt{n}$-consistent have negligible probability, *a priori*. He assumes that, *a priori*, β and g are independent, and short intervals are essentially independent since β and g are very rough. Under these assumptions, the intervals on which g is Hölder of order δ_0 contribute, on average, 0 to the estimator. There are no data in these intervals that contradict this *a priori* assessment. Hence assumptions, made for convenience in selecting the prior, dominate the inference. The trouble is that, as discussed in the appendix, even if we assume *a priori* that β and g are independent, their cross-correlation may be non-zero with high probability, in spite of the fact that this random cross-correlation has mean 0.

4. The White Noise Model and Bayesian Plug-In Property

We now consider the white noise model in which we observe the process,

$$dX(t) = \beta(t)\,dt + n^{-1/2}dW(t), \quad t \in (0,1),$$

where β is an unknown L_2-function and $W(t)$ is standard Brownian motion. This model is asymptotically equivalent to models in density estimation and nonparametric regression; see Nussbaum (1996) and Brown and Low (1996). It is also clear that this model is equivalent to the model in which we observe,

$$X_i = \beta_i + n^{-1/2}\varepsilon_i, \quad \varepsilon_i \overset{\text{iid}}{\sim} N(0,1), \quad i = 1,2,\ldots, \tag{3}$$

where X_i, β_i, and ε_i are the i-th coefficients in an orthonormal (e.g., Fourier) series expansion of $X(t)$, $\beta(t)$, and $W(t)$, respectively. Note that all the sequence $X_1, X_2, \ldots$ is observed, and n serves only as a scaling parameter. We are interested in estimating $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots)$ as an object in ℓ_2 with the loss function $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2$ and linear functionals $\theta = g(\boldsymbol{\beta}) = \sum_{i=1}^{\infty} g_i\beta_i$ with $(g_1, g_2, \ldots) \in \ell_2$, also under squared error loss. From a standard frequentist point of view, estimation in this problem is straightforward. Simple estimators achieving the optimal rate of convergence are given in the following proposition:

Proposition 4.1 *Assume that $\boldsymbol{\beta} \in \mathcal{B}_\alpha = \{\boldsymbol{\beta} : |\beta_i| \le i^{-\alpha}\}$ and $\alpha > 1/2$. The estimator $\widehat{\theta} = \sum g_i X_i$ is $\sqrt{n}$-consistent for any $g \in \ell_2$ and the estimator,*

$$\widehat{\beta}_i = \begin{cases} X_i & i^\alpha \le n^{1/2}, \\ 0 & i^\alpha > n^{1/2}, \end{cases}$$

achieves the minimax rate of convergence, $n^{-(2\alpha-1)/2\alpha}$.

The proof is given in the appendix.

4.1. The failure of Type I Bayesian analysis

A critical feature of Bayesian procedures for estimating linear functionals is that they necessarily have the plug-in property (PIP). For example, for squared error loss, since,

$$\mathsf{E}g(\widehat{\boldsymbol{\beta}}) = \sum_{i=1}^{n} g_i \mathsf{E}\widehat{\beta}_i,$$

we have $\widehat{g(\boldsymbol{\beta})} = g(\widehat{\boldsymbol{\beta}})$, for any Bayesian estimators of $g(\boldsymbol{\beta})$ and $\boldsymbol{\beta}$ based on the same prior.

We say that $\widehat{\boldsymbol{\beta}}$ is a *uniformly efficient* plug-in estimator for a set Θ of functionals and model $\mathcal{P}$ if,

$$\left\{ r_n^{-2}\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2 + n \sup_{\theta \in \Theta} \left(\theta(\widehat{\boldsymbol{\beta}}) - \theta\right)^2 \right\} = O_P(1),$$

and $\widehat{\theta} = \theta(\widehat{\boldsymbol{\beta}})$ is semiparametrically efficient for θ, where r_n is the minimax rate for estimation of $\boldsymbol{\beta}$.

Bickel and Ritov (2003) argued that there is no uniformly efficient plug-in estimator in the white noise model when Θ is large enough; for example, the set of all bounded linear functionals. Every plug-in estimator fails to achieve either the optimal nonparametric rate for estimating $\boldsymbol{\beta}$ or $\sqrt{n}$-consistency as a plug-in-estimator (PIE) of at least one bounded linear functional $g(\boldsymbol{\beta})$. The argument given in Bickel and Ritov (2003), that no estimator with the PIP can be uniformly efficient in the white noise model, can be refined, slightly, as follows.

We need the following lemma, the proof of which is given in appendix B.

Lemma 4.2 *Suppose $X \sim N(\beta, \sigma^2)$, $|\beta| \le a \le \sigma$. Let $\widehat{\beta} = \widehat{\beta}(X)$ be the posterior mean when the prior is π, assuming π is supported on $[-a, a]$, and let b_β be its bias under β. Then $|b_\beta| + |b_{-\beta}| > 2(1 - (a/\sigma)^2)|\beta|$. In particular, if π is symmetric about 0, then $|b_\beta| > (1 - (a/\sigma)^2)|\beta|$.*

This lemma shows that any Bayesian estimator is necessarily biased and puts a lower bound on this bias. We use this lemma to argue that any Bayesian estimator will fail to yield $\sqrt{n}$-consistent estimators for at least one linear functional.

Theorem 4.3 *For any Bayesian estimator $\widehat{\boldsymbol{\beta}}$ with respect to prior π supported on $\mathcal{B}_\alpha$, with $\alpha > 1/2$, there is a pair $(g, \boldsymbol{\beta}) \in \ell_2 \times \mathcal{B}_\alpha$ such that $n\left[g(\widehat{\boldsymbol{\beta}}) - g(\boldsymbol{\beta})\right]^2 \xrightarrow{p} \infty$. In fact, $\liminf_{n\to\infty} n^{(2\alpha-1)/4\alpha}\left[\mathsf{E}_{\boldsymbol{\beta}} g(\widehat{\boldsymbol{\beta}}) - g(\boldsymbol{\beta})\right] > 0$.*

Proof. It follows from Lemma 4.2 that for any $i > 2n^{1/2\alpha}$ there are β_i such that if $b_i = \mathsf{E}\widehat{\beta}_i - \beta_i$ then $|b_i| > 3i^{-\alpha}/4$. Define,

$$g_i = \begin{cases} 0, & i \le 2n^{1/2\alpha}, \\ Cn^{(2\alpha-1)/4\alpha} i^{-\alpha}, & i > 2n^{1/2\alpha}, b_i > \quad i^{-\alpha}/2, \\ -Cn^{(2\alpha-1)/4\alpha} i^{-\alpha}, & i > 2n^{1/2\alpha}, b_i < -i^{-\alpha}/2, \end{cases}$$

where C is such that $\sum_{i=1}^{\infty} g_i^2 = 1$. (Note that C is bounded away from 0 and ∞.) We have,

$$\begin{aligned} \mathsf{E}\left[\sum_{i=1}^{\infty} g_i\left(\widehat{\beta}_i - \beta_i\right)\right] &\ge 3Cn^{(2\alpha-1)/4\alpha} \sum_{i>2n^{1/2\alpha}} i^{-2\alpha}/4 \\ &\ge 3Cn^{-(2\alpha-1)/4\alpha}/4. \end{aligned}$$

□

Thus, any Bayesian estimator will fail to achieve optimal rates on some pairs $(\mathbf{g}, \boldsymbol{\beta})$. These pairs are not unusual. Actually they are pretty 'typical' members of $\ell_2 \times \mathcal{B}_\alpha$. In fact, for any Bayesian estimator $\widehat{\boldsymbol{\beta}}$ and for almost all $\boldsymbol{\beta}$ with respect to the distribution with independent uniform coordinates on $\mathcal{B}_\alpha$, there is a $\mathbf{g}$ such that $g(\widehat{\boldsymbol{\beta}})$ is inconsistent and asymptotically biased, as in the theorem. Formally,

let μ be a probability measure such that the β_i are independent and uniformly distributed on $[-i^{-\alpha}, i^{-\alpha}]$. Then, for any sequence of Bayesian estimators, $\{\widehat{\boldsymbol{\beta}}_n\}$,

$$\liminf_{n\to\infty} \mu\left\{\boldsymbol{\beta} : \sup_{\mathbf{g}\in\ell_2} n^{(2\alpha-1)/4\alpha}\left[\mathsf{E}_{\boldsymbol{\beta}} g\left(\widehat{\boldsymbol{\beta}}_n\right) - g(\boldsymbol{\beta})\right] > M\right\} = 1,$$

for some $M > 0$. This statement follows from the proof of the theorem, noting that $\mu\{|b_i| > i^{-\alpha}/2\} > 1/2$.

What makes the pairs that yield inconsistent estimators special, is only that the sequences $\beta_1, \beta_2, \ldots$ and $g_1, g_2, \ldots$ are non-ergodic. Each of them have a non-trivial auto-correlation function, and the two auto-correlation functions are similar (see Appendix A). The prior suggests that such pairs are unlikely, and therefore, that the biases of the estimators of each component cancel each other out. If the prior distribution represents a real physical phenomenon, this exact cancelation is reasonable to assume, by the law of large numbers, and the statistician should not worry about it. If, on the other hand, the prior is a way to express ignorance or subjective belief, then the analyst should worry about these small biases. This is particularly true if the only reason for assuming that these small biases are not going to accumulate is mathematical convenience. Indeed, in high-dimensional spaces, auto-correlation functions may be complex, with unknown neighborhood structures which are completely hidden from the analyst.

We consider a Bayesian model to be honestly nonparametric on $\mathcal{B}_\alpha$, if the distribution of β_i, given X_{-i}, is symmetric around 0, and $P(\beta_i > \epsilon i^{-\alpha} \,|\, X_{-i}) > \epsilon$, for some $\epsilon > 0$, where $X_{-i} = X_1, \ldots, X_{i-1}, X_{i+1}, \ldots$. That is, at least in some sense, all the components of β_i are free parameters. In this case, we have:

Theorem 4.4 *Let the prior π be honestly non-parametric on $\mathcal{B}_\alpha$ and $1/2 < \alpha < 3/4$. Suppose $\mathbf{g} = (g_1, g_2, \ldots) \in \mathcal{B}_\alpha$, and $\limsup \sqrt{n}\,|\sum_{i=\nu n^{1/2\alpha}}^{\infty} g_i\beta_i| = \infty$ for some $\nu > 1$. Then the Bayesian estimator of $g(\boldsymbol{\beta}) = \sum_{i=1}^{\infty} g_i\beta_i$ is not $\sqrt{n}$-consistent.*

Note that if the last condition is not satisfied, then an estimator that simply ignores the tails ($i > n^{1/2\alpha}$) could be $\sqrt{n}$-consistent. However, for $\mathbf{g}, \boldsymbol{\beta} \in \mathcal{B}_\alpha$, in general, all the first $n^{1/(4\alpha-2)}$ terms must be used, a number which is much greater than $n^{1/2\alpha}$ for α in the range considered.

Proof. Again, we consider the bias as in the second part of Lemma 4.2. Under our assumptions, we have,

$$\begin{aligned}
\sqrt{n}\left|\sum_{i>\nu n^{1/2\alpha}} g_i\left(\mathsf{E}\widehat{\beta}_i-\beta_i\right)\right| &= \sqrt{n}\left|\sum_{i>\nu n^{1/2\alpha}}(1-d_i)g_i\beta_i\right|, \quad 0\le d_i\le ni^{-2\alpha}\\
&\ge \sqrt{n}\left|\sum_{i>\nu n^{1/2\alpha}} g_i\beta_i\right| - \sqrt{n}\sum_{i>\nu n^{1/2\alpha}} n\left|g_i\beta_i\right| i^{-2\alpha}\\
&\ge \sqrt{n}\left|\sum_{i>\nu n^{1/2\alpha}} g_i\beta_i\right| - \sqrt{n}\sum_{i>\nu n^{1/2\alpha}} ni^{-4\alpha}\\
&= \sqrt{n}\left|\sum_{i>\nu n^{1/2\alpha}} g_i\beta_i\right| - o(1).
\end{aligned}$$

□

Note that the assumptions of the theorem are natural if the prior corresponds to the situation in which the β_i tend to 0 slowly, so that we need essentially all the available observations to estimate $g(\boldsymbol{\beta})$ at the $\sqrt{n}$-rate. As in the last two examples, if either β_i or g_i converges to 0 quickly enough – that is, β or g are smooth enough – then the difficulty disappears, as the tails do not contribute much to the functional $g(\boldsymbol{\beta})$ and they can be ignored. However, when the prior is supported on $\mathcal{B}_\alpha$, then the estimator $\widehat{\beta}_i = X_i$ is unavailable to the Bayesian (whatever the prior!) and $g(\boldsymbol{\beta})$ can not be estimated at the minimax rate with $\mathbf{g}\in\mathcal{B}_\alpha$, much less ℓ_2.

4.2. Type II Analysis

It is easy to construct priors which give the global and local minimax rates separately. For the nonparametric part $\boldsymbol{\beta}$, one can select a prior for which the β_i are independent and the estimator of β_i based on $X_i\sim N(\beta_i, n^{-1})$ with β_i restricted to the interval $[-i^{-\alpha}, i^{-\alpha}]$ is minimax; see Bickel (1981). For the parametric part, one can use an improper prior under which the β_i are independent and uniformly distributed on the real line. This prior works, but it completely ignores the constraints on the coordinates of $\boldsymbol{\beta}$. If one permits priors which are not supported on the parameter space, then this prior is perfect, in the sense that any linear functional can be estimated at the minimax rate.

If we are permitted to work with a prior which is not supported by the parameter space, then we can construct a prior which yields good estimators for both $\boldsymbol{\beta}$ and any *particular* linear functional. Indeed, suppose that $g_i\neq 0$, infinitely often, and change bases so that $\tilde{X}=B'X$, where B is an orthonormal basis for ℓ_2 with first column equal to $\mathbf{g}/\|\mathbf{g}\|$. Note that $\tilde{X}_1=\sum_{j=1}^{\infty} g_jX_j/\|\mathbf{g}\|$ and the $\tilde{X}_i$ are independent, with $\tilde{X}_i\sim N\left(\tilde{\beta}_i, n^{-1}\right)$, $i=0,1,\ldots$, where $\tilde{\beta}_1$ is the

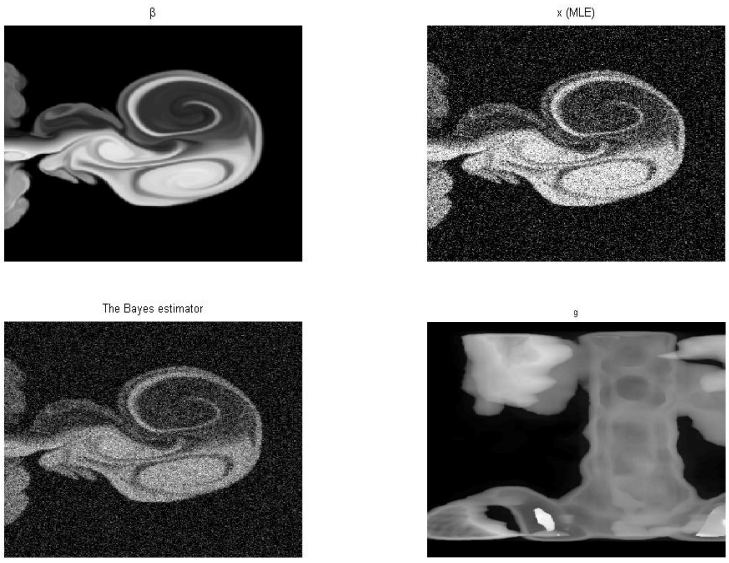

FIG 1. *Estimating Linear Functionals: (a) the vector $\boldsymbol{\beta}$; (b) the observations X; (c) the Bayesian estimator; (d) the functional g.*

parameter of interest, and $\|\tilde{\boldsymbol{\beta}}\|_2 = \|\boldsymbol{\beta}\|_2$. Thus, a Bayesian who places a flat prior on $\theta = \tilde{\beta}_1$ and a standard nonparametric prior on the other coordinates of $\tilde{\boldsymbol{\beta}}$, such that $\tilde{\beta}_i$ is estimated by $\tilde{X}_i$, properly thresholded, will be able to estimate θ efficiently and $(\tilde{\beta}_2, \tilde{\beta}_3, \ldots)$ at the minimax rate, simultaneously, cf. Zhao (2000). Of course, this prior was tailor-made for the specific functional $\theta = g(\boldsymbol{\beta})$ and would yield estimators of other linear functionals which are not $\sqrt{n}$-consistent, should the posterior be put to such a task.

4.3. An Example

To demonstrate that the effects described above have real, practical consequences, consider the following example. Take $\boldsymbol{\beta} = \text{vec}(M_0)$ and $\mathbf{g} = \text{vec}(M_1)$, where M_0 and M_1 are the two images shown in Figure 1 (a) and (d), respectively. That is, each image is represented by the matrix of the gray scale levels of the pixels, and $\text{vex}(M)$ is the vector obtained by piling the columns of M together to obtain a single vector. These images were sampled at random from the images which come bundled in the standard distribution of `Matlab`. The images have been modified slightly, so they both have the same 367×300 geometry, but nothing else has been done to them. To each element of $\boldsymbol{\beta}$ we added an independent $N(0, 169)$ random variable. This gives us X, shown in Figure 1 (b). Let π be that prior which takes the β_i i.i.d. $N(\mu, \tau^2)$, where $\mu = \sum w_i \beta_i / \sum w_i$, with w_i independent and identically uniformly-distributed on $(0, 1)$ and $\tau^2 = 315.786$,

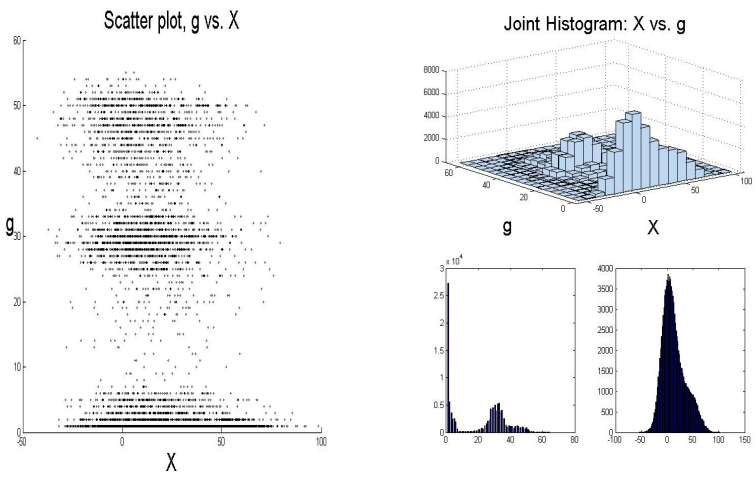

FIG 2. *A scatter plot and histograms of the data X and functional g. (a) A scatter plot of* 5% *of all pairs, chosen at random. (b) Joint and marginal histograms.*

the true empirical variance of the β_i. The resulting nonparametric Bayesian estimator is shown in Figure 1 (c). The mean squared error (MSE) of this Bayesian estimator is approximately 65% smaller than that of the MLE. Now consider the functional defined by $\mathbf{g}$, shown in Figure 1 (d). Applying g to X yields an estimator with root mean squared error (RMSE) of 1.04, but plugging-in the much cleaner Bayesian estimator of Figure 1 (c) gives an estimator with a RMSE of 19.01, almost twenty times worse than the frequentist estimator. Of course, the biggest difference between these two estimators is bias: 0.01 for the frequentist versus 19.00 for the Bayesian. These RMSE calculations were based on 1000 Monte Carlo simulations.

There is no reason to suspect that these images are correlated – they were sampled at random from an admittedly small collection of images – and they are certainly unrelated, one image shows the results of an astrophysical fluid jet simulation and the other is an image of the lumbar spine, but neither is permutation invariant nor ergodic, and this implies that the two images may be strongly positively or negatively correlated, just by chance; see Figure 2 and Appendix A.

5. Estimating the Norm of a High-Dimensional Vector

We continue with our analysis of the white noise model, but we consider a different, non-linear Euclidean parameter of interest: $\theta = \sum_{i=1}^{\infty} \beta_i^2$.

A natural estimator of β_i is given in Proposition 4.1, and one may consider a plug-in estimator of the parameter, given by $\tilde{\theta} = \sum \tilde{\beta}_i^2 = \sum_{i<n^{1/2\alpha}} X_i^2$. This estimator achieves the minimax rate for estimating $\boldsymbol{\beta}$ and $\tilde{\theta}$ is an efficient estimator of the Euclidean parameter, so long as $\alpha > 1$. But $\tilde{\beta}_i^2$ has bias $1/n$

as an estimator of β_i^2. Summing i from 1 to n, we see that the total bias is $n^{-1+1/2\alpha}$, which is much larger than $n^{-1/2}$ if $\alpha < 1$. The traditional solution to this problem is to simply unbias the estimator, cf. Bickel and Ritov (1988).

Proposition 5.1 *Suppose* $3/4 < \alpha < 1$*, then an efficient estimator of* θ *is given by,*

$$\widehat{\theta} = \sum_{i \leq m} \left(X_i^2 - n^{-1} \right), \tag{4}$$

for $n^{1/(4\alpha-2)} < m \leq n$.

Proof. Clearly the bias of the estimator is bounded by,

$$\sum_{i>m} i^{-2\alpha} < m^{-(2\alpha-1)} = o_P(n^{-1/2}),$$

and its variance is bounded by,

$$n^{-1} \sum_{i \leq m} (4\beta_i^2 + 2/n) = 4\theta n^{-1} + o_P(n^{-1}),$$

demonstrating $\sqrt{n}$-consistency. The estimator is efficient since $\hat{\theta}$ is asymptotically normal, and $1/4\theta$ is the semiparametric information for the estimation of θ. □

This is a standard frequentist approach: there is a problem and the solution is justified because it works – it produces an asymptotically efficient estimator of the parameter of interest – not because it fits a particular paradigm. The difficulty with the naive, plug-in estimator $\sum_{i \leq m} \hat{\beta}_i^2 = \sum_{i \leq m} X_i^2$ is that it is biased, but this is a problem that is easy to correct. Of course, this simple fix is not available to the Bayesian, as we show next.

5.1. The Bayesian Analysis: An Even Simpler Model

We start with a highly simplified version of the white noise model. To avoid confusion, we change notation slightly and consider,

$$Y_1, \ldots, Y_k \text{ independent with } Y_i \sim N(\mu_i, \sigma^2), \tag{5}$$

$$\theta = \theta(\mu_1, \ldots, \mu_k; g_1, \ldots, g_k) = \sum_{i=1}^{k} g_i \mu_i^2, \tag{6}$$

where the g_i are known constants. Here, we consider the asymptotic performance of estimators of θ with $\sigma^2 = \sigma_k^2 \to 0$ as $k \to \infty$. Let,

$$\widehat{\theta} = \sum_{i=1}^{k} g_i \left(Y_i^2 - \sigma^2 \right).$$

Clearly,

$$\mathsf{E}\widehat{\theta} = \theta, \qquad \mathrm{var}\widehat{\theta} = 4\sigma^2 \sum_{i=1}^{k} g_i^2 \mu_i^2 + 2\sigma^4 \sum_{i=1}^{k} g_i^2.$$

Suppose that the μ_i are *a priori* i.i.d. $N(0, \tau^2)$, with $\tau^2 = \tau_k^2$ known, and consider the situation in which $g_1 \sim \cdots \sim g_k$. If $k^{-1/2}\sigma_k^2 \ll \tau_k^2 \ll \sigma_k^2$, then the signal-to-noise ratio τ^2/σ^2 is strictly less than 1 and no estimator of μ_i performs much better than simply setting $\widehat{\mu}_i = 0$. On the other hand, $\widehat{\theta}$ remains a good estimator of θ, with coefficient of variation, $O\left(\sqrt{k}\sigma^2/k\tau^2\right)$, converging to 0. We call this paradoxical regime the *non-localizable* range, as we can estimate global parameters, like θ, but not the local parameters, $\mu_1, \ldots, \mu_k$.

A posteriori, the $\mu_i \sim N\left(\tau^2 Y_i/(\sigma^2+\tau^2), \tau^2\sigma^2/(\sigma^2+\tau^2)\right)$ and the Bayesian estimator of θ is given by,

$$\sum_{i=1}^{k} g_i \mathsf{E}\left(\mu_i^2 \,|\, Y_i\right) = \frac{\sigma^4 + 2\tau^2\sigma^2}{(\sigma^2+\tau^2)^2} \sum_{i=1}^{k} g_i^2 \tau^2 + \frac{\tau^4}{(\sigma^2+\tau^2)^2} \sum_{i=1}^{k} g_i \left(Y_i - \sigma^2\right).$$

This expression has the structure of a Bayesian estimator in exponential families: a weighted average of the prior mean and the unbiased estimator. If the signal-to-noise ratio is small, $\tau^2 \ll \sigma^2$, almost all the weight is put on the prior. This is correct, since the variance of θ, under the prior, is much smaller than the variance of the unbiased estimator. So, if we really believe the prior, the data can be ignored at little cost. However, in frequentist terms, the estimator is severely biased and, for a type II Bayesian, non-robust.

The Achilles heel of the Bayesian approach is the plug-in property. That is, $\mathsf{E}\left(\sum_{i=1}^m \mu_i^2 | \mathrm{data}\right) = \sum_{i=1}^m \mathsf{E}\left(\mu_i^2 | \mathrm{data}\right)$. However, when the signal-to-noise ratio is infinitesimally small, any Bayesian estimator must employ shrinkage. Note that, in particular, the unbiased estimator $Y_i^2 - \sigma^2$ of μ_i^2 can not be Bayesian, because it is likely to be negative and is an order of magnitude larger than μ_i^2.

A 'natural' fix to the non-robustness of the i.i.d. prior, is to introduce a hyperparameter. Let τ^2 be an unknown parameter, with some smooth prior. Marginally, under the prior, $Y_1, \ldots, Y_k$ are i.i.d. $N(0, \sigma^2 + \tau^2)$. By standard calculations, it is easy to see that the MLE of τ^2 is $\widehat{\tau}^2 = k^{-1}\sum_{i=1}^k \left(Y_i^2 - \sigma^2\right)$. By the Bernstein-von Mises theorem, the Bayesian estimator of τ^2 must be within $o_P(k^{-1/2})$ of $\widehat{\tau}^2$. If $g_1 = \cdots = g_k$ and we plug $\tau = \widehat{\tau}$ into the formula for the Bayesian estimator, we get a weighted average of two estimators of θ, both of which are equal to $\widehat{\theta}$. But, in general, $\widehat{\tau}$ is strictly different from $\widehat{\theta}$ and this estimator is inconsistent. Of course, the Bayesian estimator is not obtained by plugging-in the estimated value of τ, but the difference would be small here, and the Bayesian estimator would perform poorly.

Although the prior would be arbitrary, we can, of course, select the prior so that the marginal variance is directly relevant to estimating θ. One way to do this is to assume that τ^2 has some smooth prior and, given τ^2, the μ_i are i.i.d. $N(0, (\tau^2/g_i) - \sigma^2)$. Then, $Y_i \sim N(0, \tau^2/g_i)$, marginally, and the marginal

log-likelihood function is,

$$-k\log(\tau^2)/2 - \sum_{i=1}^{k} g_i Y_i^2/2\tau^2.$$

In this case, $\widehat{\tau}^2 = k^{-1}\sum_{i=1}^k g_i Y_i^2$ and the posterior mean of $\sum_{i=1}^k g_i\mu_i^2$ is approximately $\sum_{i=1}^k g_i\left(\widehat{\tau}^2/g_i - \sigma^2\right) = \sum_{i=1}^k g_i\left(Y_i^2 - \sigma^2\right)$, as desired.

This form of the prior variance for the μ_i is not accidental. Suppose, more generally, that $\mu_i \sim N\big(0, \tau_i^2(\rho)\big)$, *a priori*, for some hyperparameter ρ. Then the score equation for $\widehat{\rho}$ is $\sum_{i=1}^k w_i(\widehat{\rho})Y_i^2 = \sum_{i=1}^k w_i(\widehat{\rho})\left(\tau_i(\widehat{\rho}) + \sigma^2\right)$, where $w_i(\rho) = \tau_i(\rho)\dot{\tau}_i(\rho)/\left(\tau_i(\widehat{\rho}) + \sigma^2\right)^2$. If we want the weight w_i to be proportional to g_i, then we get a simple differential equation, the general solution of which is given by $\left(\tau_i(\rho) + \sigma^2\right)^{-1} = g_i\rho + d_i$. Hence, the general form of the prior variance is,

$$\tau_i^2(\rho) = \left(g_i\rho + d_i\right)^{-1} - \sigma^2.$$

The prior suggested above simply takes $d_i = 0$, for all i. If the type II Bayesian really believes that all the μ_i should have some known prior variance τ_0^2, he can take $d_i = \left(\tau_0^2 + \sigma^2\right)^{-1} - g_i$, obtaining the expression,

$$\tau_i^2(\rho) = \frac{\tau_0^2 + (\rho-1)(\tau^2+\sigma^2)\sigma^2 g_i}{1 + (\rho-1)(\tau^2+\sigma^2)\sigma^2 g_i}.$$

If the variance of the μ_i really is τ_0^2, then the posterior for the hyperparameter ρ will concentrate on 1 and the τ_i^2 will concentrate on τ_0^2. If, on the other hand, τ^2 is unknown, the resulting estimator will still perform well, although the expression for τ_i^2 is quite arbitrary.

The discussion above holds when we are interested in estimating the hyperparameter $\sum_{i=1}^k g_i\tau_i^2(\rho)$. This is a legitimate change in the rules and the resulting estimator can be used to estimate θ in the non-localizable regime, because the main contribution to the estimator is the contribution of the prior, conditioning on $\tau_i^2(\rho)$. However, when $\tau_i^2(\rho) \approx \sigma^2$, there may be a clear difference between the Bayesian estimators of $\sum_{i=1}^k \tau_i^2(\rho)$ and $\sum_{i=1}^k \mu_i^2$, respectively.

We conjecture that a construction based on stratification might be used to avoid the problems discussed above: the use of an unnatural prior and the difference between estimating the hyperparameter and estimating the norm. In this case, we would stratify based on the values of the g_i and estimate $\sum \mu_i^2$ separately in each stratum. The price paid by such an estimator is a large number of hyperparameters and a prior suited to a very specific task.

The discussion above shows that $\widehat{\theta}$ can at least be approximated by a Bayesian estimator, but the corresponding prior has to have a specific form and would have to have been chosen for convenience rather than prior belief. This presents no difficulty for the type II Bayesian, who is free to select his prior to achieve a particular goal. However, problems with the prior remain. The prior is tailor-made for a specific problem: while $\beta_1, \ldots, \beta_k$ i.i.d. $N(0, \tau^2)$ is a very good prior

for estimating $\sum_{i=1}^{k} \mu_i^2$, when the parameter of interest is not permutation invariant, the estimator is likely to perform poorly in frequentist terms. Also, the prior is appropriate for regular models but not sparse ones. Consider again the non-localizable regime in which $\sqrt{k}\sigma^2 \ll \theta \ll k\sigma^2$, but suppose that most of the μ_i are very close to zero, with only a few taking values larger than σ^2 in absolute value. A Bayesian estimator based on the prior suggested above will shrink all the Y_i toward 0, strongly biasing the estimates of the μ_i, whereas a standard (soft or hard) thresholding estimator will have much better performance. A completely different prior is need to deal with sparsity. See Greenshtein et al. (2008) and van der Pas et al. (2013) for an empirical Bayes solution to the sparsity problem.

5.2. A Bayesian Analysis of the White Noise Model

Returning to original model, $X_i \sim N(\beta_i, 1/n)$, $|\beta_i| < i^{-\alpha}$, with $\theta = \sum_{i=1}^{n} \beta_i^2$, we can use a prior for which the β_i are i.i.d. $N(0, \tau^2)$, for $i = 1, \ldots, m$, and 0, otherwise, where $m = n^{1/(4\alpha-2)+\nu}$, for some $\nu > 0$. This gives us a Bayesian estimator of θ which is asymptotically equivalent to the unbiased estimator, $\widehat{\theta} = \sum_{i=1}^{n} \left(X_i^2 - n^{-1}\right)$, and asymptotically efficient. However, the corresponding estimator for $\boldsymbol{\beta}$ is not even consistent and, when we try to estimate β_i, even for i relatively small, we see that the Bayesian estimator shrinks X_i toward 0 by a factor of $1 - \rho$ where ρ is asymptotically larger than $\theta m/n = \theta n^{-(4\alpha-3)/(4\alpha-2)-\nu} \gg n^{-1/2}$. So our estimate of β_i fails to be $\sqrt{n}$-consistent.

A more reasonable approach, in this situation, is to partition the set $X_1, \ldots, X_n$ into blocks, $\{X_{k_{j-1}}, \ldots, X_{k_j}\}$, $j = 1, \ldots, J$, and use a mean-zero Gaussian prior with unknown variance in each of the blocks. One possible assignment is $k_0 = 1$, $k_1 = o(\sqrt{n})$, and $k_j = 2k_{j-1}$, $j > 1$. Thus, $O(\log n)$ blocks are needed. The analysis presented above shows that this prior would yield a good estimator of θ without, hopefully, sacrificing our ability to estimate the β_i at the $\sqrt{n}$-rate. Of course, this prior is not supported on the parameter space $\mathcal{B}_\alpha$: it forces uniform shrinkage of the observations in each block (and bypasses the plug-in property by estimating block-wise hyperparameters). But there is nothing 'natural' about these blocks and nothing in the problem statement suggests this grouping.

As before, this "objective" prior was constructed with a specific parameter in mind and is unlikely to be effective for other parameters; it can not represent prior beliefs. The prior will also fail when sparsity makes the block structure inappropriate. The unbiased, frequentist estimator has no such difficulty. The Bayesian is obliged to conform to the plug-in principle and, because of this, at some stage, must get stuck with the wrong prior for some parameter which wasn't considered interesting initially.

Consider a general prior π. Let π_i be the prior for β_i given $X_{-i} = (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots)$. For $i > n^{1/2\alpha+\nu}$ with $\nu > 0$ arbitrarily small and $m =$

$n^{1/(4\alpha-2)+\nu}$, as in Proposition 5.1,

$$\mathsf{E}_\pi(\beta_i^2 \mid X_1,\dots,X_m) = \frac{\int_{-i^{-\alpha}}^{i^{-\alpha}} t^2\varphi\left(n(X_i-t)\right)d\pi_i(t)}{\int_{-i^{-\alpha}}^{i^{-\alpha}} \varphi\left(n(X_i-t)\right)d\pi_i(t)} \in (a^{-1}\mathsf{E}_{\pi_i}\beta_i^2, a\mathsf{E}_{\pi_i}\beta_i^2), \quad (7)$$

where for $\mathcal{I} = \{i : n^{1/2+\nu} < i \le n^{1/(4\alpha-2)+\nu}\}$,

$$\max_{i\in\mathcal{I}} \log a \le \max_{\substack{i\in\mathcal{I}\\ |t_i|<i^{-\alpha}}} n\left|(X_i-t_1)^2-(X_i-t_2)^2\right| \xrightarrow{p} 0,$$

since $\max_{i\in\mathcal{I}} n^{1/2-\nu}|X_i| \xrightarrow{p} 0$. But this means that the estimate of β_i^2 depends only weakly on X_i itself. It is mainly a function of X_{-i} and the prior. Moreover, if the estimate of θ is to be close to the unbiased one, then this must be achieved through the influence of X_i on the estimates of β_j, for $j \neq i$. This is the case in the construction above where, formally, we are estimating a hyperparameter of the prior, rather than θ, itself. The result is a non-robust estimator which works for the particular functional of interest but not others. In fact, we have the following theorem.

Theorem 5.2 *Let $\boldsymbol{\beta}_{-i} = (\beta_1,\dots,\beta_{i-1},\beta_{i+1},\dots)$. Let π be the prior on $\boldsymbol{\beta}$. Suppose that there is an $\eta > 0$ such that a.s. under the prior π: $P_\pi(\lceil 4i^{2\alpha}\beta_i^2\rceil = \kappa|\boldsymbol{\beta}_{-i}) > \eta$, $i = 1,2,\dots$, and $\kappa = 1,\dots,4$. There exists a set $S = S_n$ with $\pi(S_n) \to 1$, such that for all $\boldsymbol{\beta} \in S$ there is a sequence $g_1, g_2, \dots$, for which the Bayesian estimator of $\sum g_i\beta_i^2$ with respect to π is not $\sqrt{n}$-consistent.*

The proof is given in the appendix. The conditions in the theorem are needed to ensure that support of the the prior does not degenerate to a finite-dimensional parametric model.

6. Data-Dependent Sample Sizes and Stopping Times

The stopping rule principle (SRP) says that, in a sequential experiment, with final data $\mathbf{x}^N(\tau)$, inferences should not depend on the stopping time τ; see Berger and Wolpert (1988). In so much as Bayesian techniques follow the strong likelihood principle (SLP), they must also follow the SRP.

To see that high dimensional data represents a challenge for the SRP, consider another version of the white noise model. Let $n^{-2\alpha} < \beta_i < 3n^{-2\alpha}$, $i = 1,\dots,k = \lfloor n^{2\alpha}\rfloor$, and $1/6 < \alpha < 1/4$. Suppose that, for each i, $X_i(\cdot)$ is a Brownian motion with drift β_i, and that X_i is observed until some random time T_i. Take $\bar{X}_i(t) = X_i(t)/t$ and note that this is the sufficient statistic for β_i given $\{X_i(s) : s < t\}$. Of course, $\bar{X}_i$ is also the MLE. Finally, let π_i be the prior for β_i given $X_{-i} = (X_1,\dots,X_{i-1},X_{i+1},\dots)$. Let $f_i(\cdot)$ be the density of the distribution of $X_i(T_i)$ given X_{-i}; $f_i = \pi_i * N(0,1/T_i)$. We assume that the prior π_i is non-parametric in the sense that π_i is bounded away from 0 on the allowed support, so that X_{-i} does not give us too much information about β_i.

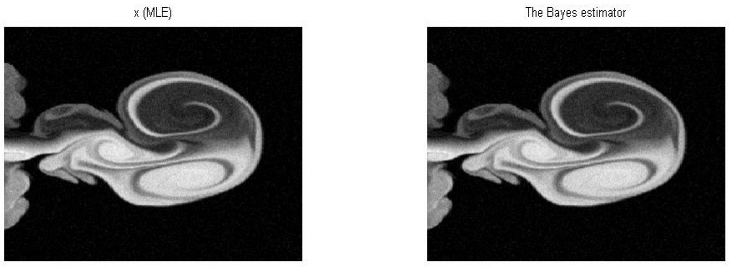

FIG 3. *(a) the MLE; (b) the Bayesian estimator.*

It is well known that the posterior mean of β_i given the data satisfies,

$$\mathsf{E}(\beta_i \,|\, \text{data}) = \bar{X}_i(T_i) + \frac{1}{T_i}\frac{f_i'\left(X_i(T_i)\right)}{f_i\left(X_i(T_i)\right)}.$$

If $T_i = O(n)$, then $f_i \approx \pi_i$ and $\bar{X}_i(T_i) \approx \beta_i$. Suppose T_i is correlated with $g_i(\beta_i)$, where $g_i = f_i'/f_i$, then the MLE of $\sum_{i=1}^k \beta_i$, given by $\sum_{i=1}^k X_i(T_i)$ is unbiased and has a random error on the order of $n^{\alpha}n^{-1/2}$, while the Bayesian estimator has a bias which is $\sim n^{2\alpha}n^{2\alpha}/n$, with $n^{2\alpha}$ terms each contributing $O(n^{2\alpha})$ to the bias, from g_i, and a term of $O(1/n)$ from $1/T_i$. With $1/6 < \alpha < 1/4$, the Bayes bias dominates the random error!

6.1. An Example

We consider again the vector $\boldsymbol{\beta}$ represented in Figure 1 (a), but this time the vectorized version of the spine image shown in Figure 1 (d) is used to specify the random number of observations associated with each element of $\boldsymbol{\beta}$. Adding noise to Figure 1 (a), we get the observed data and MLE, shown in Figure 3 (a). This SNR is higher here than before (+2.72db) and, as a result, the Bayesian estimator shown in Figure 3 (b) is much smoother.

Here, we used a prior with independent Gaussian components, each with a mean equal to the mean of $\boldsymbol{\beta}$ and variance equal to the variance of the β_i. We have two processes on the unit square: one represents $\boldsymbol{\beta}$ and the other corresponds to random stopping times, with the number of observations proportional to the gray-scale value of the corresponding pixel in the image of the spine. As we have already seen, these images are correlated, although there is no reason, *a priori*, to expect they would be, having been chosen at random from a collection of unrelated images. This correlation causes trouble: In 500 Monte Carlo simulations, the RMSE of the Bayesian estimator of the sum of the β_i is 0.05, whereas the RMSE of the MLE is 0.009. The difference is due almost entirely to bias. If we replace the stopping times with a fixed time, the Bayesian estimator performs better, achieving a RMSE of 0.0071 versus the RMSE of the MLE = 0.0072. This example shows clearly that the Bayesian estimator can be

badly biased when the stopping times and the unknown parameters happen to be correlated.

7. Bayesian Procedures are Efficient under Bayesian Assumptions

Freedman (1965) proves that in some very weak sense consistency of Bayesian procedures is 'rare'. We, however, start with a version of Doob's consistency result and show that the existence of a uniformly $\sqrt{n}$-consistent estimator ensures that the posterior distribution is $\sqrt{n}$-consistent with prior probability 1.

To simplify notation we consider the Markov chain $\beta_0 \to X_n \to \beta_n$, where $\beta_0, \beta_n \in \mathcal{B}$, $\beta_0 \sim \pi$, $X_n \sim P_{\beta_0}$, and given X_n, β_0 and β_n are i.i.d. That is, given X_n, β_n is distributed according to the posterior distribution π_{X_n}. Let P be the joint distribution of the chain. With some abuse of notation, P_{β_0} is also the conditional distribution of the chain given that it starts at β_0. Let d_n be a semi-metric on the parameter space, normalized to the sample size. Typically, in the nonparametric situation considered in this paper, $d_n(\beta, \beta') = \sqrt{n}|\theta(\beta) - \theta(\beta')|$ for some real-valued functional θ of the parameter.

We consider an estimator $\tilde{\beta}_n$ to be d_n *consistent uniformly on* $\mathcal{B}$, if for every $\varepsilon > 0$ there is an $M < \infty$ such that for all $\beta \in \mathcal{B}$ and n large enough, $P_\beta \left\{ d_n(\tilde{\beta}_n, \beta) \geq M \right\} \leq \varepsilon$. The posterior is d_n *consistent uniformly on* $\mathcal{B}$ if for every $\varepsilon > 0$ there is an $M < \infty$ such that for all $\beta_0 \in \mathcal{B}$ and n large enough, $P_{\beta_0} \{ d_n(\beta_n, \beta_0) \geq M \} \leq \varepsilon$.

Thus we consider the inference to be d_n uniformly consistent if the frequentist Markov chain, $\beta_0 \to X_n \to \tilde{\beta}_n$, or the Bayesian one, $\beta_0 \to X_n \to \beta_n$ lands in an $O_p(1)$ d_n-ball.

Theorem 7.1 *Suppose there is an estimator which is d_n consistent uniformly on $\mathcal{B}$. Then there is a $\mathcal{B}' \subseteq \mathcal{B}$ such that $\pi(\mathcal{B}') = 1$ and the posterior is d_n consistent uniformly on $\mathcal{B}'$.*

The proof is given in appendix B.

Thus, the existence of a uniformly good frequentist estimator ensures that the there is a set with prior probability one such that the Bayes posterior is uniformly consistent at the right rate on that set. The difficulty with this statement is that, in high dimensional spaces, there is no natural extension of Lebesgue measure and null sets of very natural-looking priors are sometimes much larger than one would expect. For a simple example, consider a prior with hyperparameters of the type we considered for the white noise models: τ has standard exponential distribution, and $\beta_i, \ldots, \beta_k$ are, given τ, i.i.d. $N(0, \tau^2)$. Consider the set $S = \{\boldsymbol{\beta} : k^{-1} \sum_{i=1}^k (\beta_i - \bar{\beta}_k)^4 < 2.5(k^{-1} \sum_{i=1}^k (\beta_i - \bar{\beta}_k)^2)^2\}$. The probability of S is 0.82 if $k = 5$. It drops to approximately 0.27 when $k = 50$. It is 0.0025 for $k = 500$, and negligible when $k = 5000$. (These numbers are based on 100,000 Matlab simulations.) The set S is not so unusual or unexpected that it can be really ignored *a priori* and, unlike most sets, S is simple to comprehend. If inferences depend on whether or not the fourth moment of the parameter is exactly

three times the square of the second, as implied by the normality assumption, which was made for convenience, these inferences would not be robust.

Theorem 7.1 does not contradict our findings. In the stratified sampling and partial linear model examples of Sections 2 and 3, the difference between the Bayesian estimator and the frequentist one, is that the former ignores the information that restricts the model to a subset of the parameter space which has prior probability 0. In the white noise models of Sections 4 and 5, the requirement that the prior be "honestly non-parametric" limits $\beta_1, \beta_2, \ldots$ to regular sequences obeying a law of large numbers and, as a result, the set of non-ergodic sequences is given prior probability 0. And, in these examples, there are two phenomena which make this theorem irrelevant. First, Bayesian estimators must obey the plug-in principle, restricting estimators to those of the form $\theta(\tilde{\beta})$ for $\tilde{\beta} \in \mathcal{B}$, while the frequentist estimator can not be written in this form. Second, each prior fails for a different functional, but, if the functional and the parameter are chosen together, as we have argued might well happen, this theorem has no consequences.

The second result of this section gives an easy abstract construction which shows that, under some conditions, a type II Bayesian is able to choose a prior with good frequentist properties. Our setup is as follows. In the n-th problem we observe $X^{(n)} \sim P \in \mathcal{P}^{(n)} \ll \nu$, with density $p = dP/d\nu$. Estimators take values in the set $\mathcal{A}$, and a loss function $\ell_n : \mathcal{P}^{(n)} \times \mathcal{A} \to \mathbb{R}^+$ is used to assess the "cost" associated with a particular estimate. We assume that ℓ_n is bounded by $L_n < \infty$ for all n and that,

A1 The loss function is Lipschitz: for all $a \in \mathcal{A}$ and $P, P' \in \mathcal{P}^{(n)}$: $|\ell_n(P, a) - \ell_n(P', a)| \leq c_n \|p - p_n\|$, where $\|\cdot\|$ is the variational norm.

A2 Given $\varepsilon > 0$ there exists a finite set $\mathcal{P}_K^{(n)} \subset \mathcal{P}^{(n)}$ with cardinality $\kappa_{n,\varepsilon}$, such that $\sup_{P \in \mathcal{P}^{(n)}} \inf_{P' \in \mathcal{P}_K^{(n)}} \|P - P'\| \leq \varepsilon$.

A3 Let $R_n(P, \delta) = \mathsf{E}_P \ell_n(P, \delta(X))$, where $\delta : \mathcal{X}^{(n)} \to \mathcal{A}$, or more generally, δ is a randomized procedure (or Markov kernel from $\mathcal{X}^{(n)}$ to $\mathcal{A}$). Let $R_n(\delta) = \sup_{P \in \mathcal{P}^{(n)}} R_n(P, \delta)$. There exist δ^* such that $R_n(\delta^*) = \inf_\delta R_n(\delta) \equiv r_n \leq r < \infty$ for all n.

Let μ_n be a probability measure on $\mathcal{P}_K^{(n)}$. The corresponding posterior distribution is $\mu_n(P_j|X^{(n)}) = \mu_n(P_j)p_j(X^{(n)}) / \sum_{k=1}^{\kappa} \mu_n(P_k)p_k(X^{(n)})$. Let δ_{μ_n} be the Bayesian procedure with respect to μ_n.

Theorem 7.2 *If conditions A1–A3 hold, then for all $\varepsilon' > 0$, there exist $\mu_{n,\varepsilon'}$ on $\mathcal{P}^{(n)}$, such that $R_n(\delta_{\mu_{n,\varepsilon'}}) \leq r_n + \varepsilon'$.*

The proof is given in appendix B and can be used to argue that, under the conditions above, it is always possible (for a type II Bayesian) to select a prior such that the corresponding Bayesian procedure estimates both the global and local parameters at their minimax rates:

Corollary 7.3 *Consider an estimation problem in which $\mathcal{P}^{(n)}$ satisfies the conditions of Theorem 7.2; $\ell_{1n}(P, a)$, $\ell_{2n}(P, a)$ are two loss functions, each satisfy-*

ing condition A1, with Lipschitz constants c_{1n} and c_{2n}, respectively, and,

$$\inf_{\delta} \max_{P \in \mathcal{P}^{(n)}} \mathsf{E}_P \ell_{kn}(P, \delta) = O(b_{kn}^{-1}), \quad k = 1, 2,$$

For some b_{1n}, b_{2n}. Then, given $\varepsilon > 0$, there exist μ_n on $\mathcal{P}^{(n)}$ such that, simultaneously,

$$\max_{P \in \mathcal{P}^{(n)}} \mathsf{E}_P \ell_{kn}(P, \delta_{\mu_n}) = O(b_{kn}^{-1}) \quad k = 1, 2.$$

The corollary follows by applying the theorem to the combined loss function $\ell_n(P, (a_1, a_2)) = b_{1n}\ell_{1n}(P, a_1) + b_{2n}\ell_{2n}(P, a_2)$.

The conditions essentially hold in our examples (technically, in the stratified sampling and partial linear model examples, before applying the theorem, one should restrict the parameter space to a compact set). However, note that the prior may depend on information that may not be known *a priori*, such as the loss function, and on parameters that "should not" be part of the loss, such as the weight function in the stratified sampling example, the (smoothness of the) conditional expectation of U given X in the partial linear model, and the linear functional in the white noise model.

Note, however, that the theorem as proven does not say that there exists a prior such that the two Bayesian estimators for each of the two loss functions achieve the corresponding minimax rates. Indeed, a single estimator is produced which balances the two objectives.

8. Summary

In this paper we presented a few toy examples in which a nonparametric prior fails to produce estimators of simple functionals that are $\sqrt{n}$-consistent, in spite of the fact that efficient frequentist procedures exist (and are often easy to construct). In these examples, minimal smoothness was assumed, but we do not believe that this is necessary in order for the Bayesian paradigm to have difficulty with high-dimensional models. With minimal smoothness, it is easy to prove that bias accumulates and global functionals cannot be estimated at minimax rates (while with smoother objects, this would be more difficult to demonstrate).

Bayesian procedures are always unbiased with the respect to the prior on which they are based. Bayesian estimators tend to replace parameters buried in noise by their *a priori* means. This would be a reasonable strategy if the prior represented a physical reality, but is not workable if the prior represents a subjective belief or is selected for computational convenience. In the latter case, to the extent that the beliefs or assumptions fail to match the physical reality, the Bayesian paradigm will run into difficulty.

Several difficulties with the Bayesian approach were demonstrated by our examples, including:

1. The possibility of *de facto* cross-correlation between two independent processes, as discussed in Appendix A, is ignored by the Bayesian estimator.

The effect of such *spurious correlations* can be seen in the stratified sampling example of Section 2, the partial linear model of Section 3, and the discussion on estimating linear functionals in the white noise model of Section 4. Because the spurious correlations observed have mean value 0, the Bayesian estimators are unbiased, on average, but this average is only with respect to the prior. In any other sense, the Bayesian estimators are biased.

2. For linear functionals with squared error loss, the Bayesian paradigm requires the analyst to follow the *plug-in principle*, estimating functionals θ of high-dimensional parameters β by $\tilde{\theta} = \theta(\tilde{\beta})$. The fact that universal plug-in estimators do not exist shows that strict adherence to the Bayesian paradigm is too rigid. This was shown in Section 4.
3. Having selected a prior, the Bayesian may assume that some functionals of the unknown parameter are known – for example, weighted means of many unknown parameters. But, as a matter of fact, these unverified assumptions, hidden in the selected prior, force the resulting estimator to be *non-robust*. See, for example, the discussion of the partial linear model in Section 3.
4. On the other hand, replacing components of signal buried deeply in noise by their prior means may cause an *accumulation of bias*, destroying estimators of functionals which can be estimated without bias and with bounded asymptotic variance. This is clear from the discussion in Section 5.
5. Finally, the Bayesian paradigm forces the analyst to follow the strict likelihood principle, cf. Berger and Wolpert (1988), and this may force him to *ignore auxiliary information* which could be used to produce asymptotically unbiased, efficient estimators. This was the core of the argument in the stratified sampling example of Section 2 in which the type I Bayesian can not make use of information on the sampling probabilities, at all, and can not produce a $\sqrt{n}$-consistent estimator of the population mean, in general, as a result. The same is true in the partial linear model example of Section 3, in which the Bayesian analyst can not make use of information on smoothness, and in the stopping times example of Section 6.

Real-life examples are more complex and less tractable than the toy problems we have played with in this paper and, as a result, it would be more difficult to determine the real-life effect of assumptions hidden in the prior on the frequentist behavior of Bayesian estimators in such situations. It is very difficult to build a prior for a very complicated model. Typically, one would assume a lot of independence. However, with many independent or nearly-independent components, the law of large numbers and central limit theorem will take effect, concentrating what was supposed to have been a vague prior in a small corner of the parameter space. The resulting estimator will be efficient for parameters in this small set, but not in general. It is safe to say that Bayes is not curse of dimensionality appropriate (or CODA, see Robins and Ritov (1997)).

Appendix A: Independent but Correlated Series

Much of the analysis in this paper is based on presenting counterexamples on which a given estimation procedure fails. This is satisfactory from a minimax frequentist point of view: one example is enough to argue that the result depends on the unknown parameter and is not uniformly valid, or asymptotically minimax. However, this may not convince a Bayesian, who might claim that the counter example is *a priori* unreasonable. A typical example of the argument was presented in the stratified sampling example of Section 2. This argument can be characterized by constructing two *a priori* independent processes (β and g), which happen to be "similar". For the Bayesian this is a very unlikely event. After all, he assumes that they are independent; for example, one of them depends on biology and the other on budget constraints. In this section, we argue that such correlations can actually be quite likely. Harmeling and Toussaint (2007) write: "Let us now get to the core of Robins and Ritov (1997). The authors consider uniform unbiasedness of an estimator. This means that the estimator has to be unbiased for every possible choice of θ and ξ. In the experiment we performed above, though, we chose θ and ξ independently and thus it was very unlikely that we ended up with an accidentally correlated θ and ξ, e.g., where θ tends to be large whenever also ξ is (or inversely)." (We should remark that they consider also a scenario in which the process are correlated.) We claim that this criticism ignores the fact that two processes can be independent and yet, with high probability, have an empirical cross-correlation which is far from 0. This would be the case, for example, if the processes are non-ergodic and have similar autocorrelation functions.

Suppose $U_1, \ldots, U_n$ and $V_1, \ldots, V_n$ are two independent simple random walks. Then of course U_n and V_n are uncorrelated. But we may consider the correlation between these two series $R = n^{-1} \sum_{i=1}^{n} (U_i - \bar{U}_n)(V_i - \bar{V}_n)$, where $\bar{U}_n$ and $\bar{V}_n$ are the empirical means of the two series, respectively. R is a random variable and clearly it has mean 0. However, it is far from being close to 0, even if n is large. In fact, asymptotically, R is almost uniformly distributed on most of the interval $(-1, 1)$, cf. McShane and Wyner (2011). The reason for this somewhat surprising fact is that random walks and Brownian motions are less wild than they are sometimes thought to be. In fact given U_n, the best predictor of $U_{\lfloor n/2 \rfloor}$ is $U_n/2$, where $\lfloor a \rfloor$ is the largest integer less than a, and the sequence tends to be, very roughly speaking, monotone. But if both $U_1, \ldots, U_n$ and $V_1, \ldots, V_n$ are "somewhat" monotone, then they will be cross-correlated; maybe positively correlated, maybe negatively, but rarely uncorrelated. Consider now two general, independent mean 0 random, non-mixing sequences $U_1, \ldots, U_n$ and $V_1, \ldots, V_n$. Suppose that the two sequences have the autocorrelation functions $A(i, j) = \mathrm{cov}(U_i, U_j)$ and $B(i, j) = \mathrm{cov}(V_i, V_j)$, where we assume $\mathrm{var}(U_i) = \mathrm{var}(V_i) = 1$ (although, in the standard engineering usage, autocorrelation refers to what some would like to call autocovariance). We do not assume that the series are stationary, and we do not know their autocorrelation functions. The picture we have in mind is that each (U_i, V_i) is a characteristic of points in a large graph, and neighboring nodes are highly correlated, but we do not know the

neighborhood structure of the graph. Define,

$$R = \langle U, V \rangle_0 \equiv n^{-1} \sum_{i=1}^{n} U_i V_i - n^{-2} \sum_{i=1}^{n} U_i \sum_{i=1}^{n} V_i,$$

where $\langle \cdot, \cdot \rangle_0$ is the empirical cross-covariance between two sequences. Then $\mathsf{E}R = 0$, while direct calculations give,

$$\mathrm{var}(R) = n^{-1} \sum_{i=1}^{n} \langle A(i,\cdot), B(i,\cdot) \rangle_0 - \left\langle n^{-1} \sum_{j=1}^{n} A(\cdot, j), n^{-1} \sum_{j=1}^{n} B(\cdot, j) \right\rangle_0.$$

To get some sense of the size of $\mathrm{var}(R)$, suppose that $n^{-1} \sum_{j=1}^{n} A(i,j) \equiv n^{-1} \sum_{j=1}^{n} B(i,j) \equiv c$. Then we get,

$$\mathrm{var}(R) = n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} \left(A(i,j) - c\right) \left(B(i,j) - c\right).$$

Clearly, if the two series are mixing and $\sum_j A(i,j) = \sum_j B(i,j) = O(1)$, then $\mathrm{var}(R) = O(n^{-1})$. However, if they are not mixing, and have similar autocorrelation functions, then most realizations of these two series will have non-zero cross-correlation.

Appendix B: Proofs

Proof. [Proof of Proposition 4.1] Clearly,

$$\begin{aligned} \mathsf{E} \sum_{i=1}^{\infty} \left(\widehat{\beta}_i - \beta_i\right)^2 &= \lfloor n^{1/2\alpha} \rfloor / n + \sum_{i > n^{1/2\alpha}} \beta_i^2 \\ &\leq n^{-(2\alpha-1)/2\alpha} + \sum_{i > n^{1/2\alpha}} i^{-2\alpha} \\ &\leq 2\alpha n^{-(2\alpha-1)/2\alpha} / (2\alpha - 1). \end{aligned}$$

That this is the minimax rate is established by considering the prior Π which makes $\beta_1, \beta_2, \ldots$ independent, with $\Pi\left(\beta_i = \pm i^{-\alpha}\right) = 1/2$. □

Proof. [Proof of Lemma 4.2] First note that because of the monotone likelihood ratio property, $\widehat{\theta}(x)$ is a monotone increasing function of x. We have,

$$\begin{aligned} 1 + \dot{b}_\theta &= \partial \mathsf{E}_\theta \mathsf{E}_\pi \left(\Theta \,|\, X\right) / \partial \theta \\ &= \frac{\partial}{\partial \theta} \mathsf{E}_\theta \frac{\int t e^{-(X-t)^2/2\sigma^2} d\pi(t)}{\int e^{-(X-t)^2/2\sigma^2} d\pi(t)}, \end{aligned}$$

where E_θ is the expectation assuming the true value of the parameter is θ, (Θ, X) is a pair of random variables such that $\Theta \sim \pi$, and $X|\Theta \sim N(\Theta, \sigma^2)$, and E_π is the expectation under this joint distribution. Note that E_π is a formal expression, since we assume that $X \sim N(\theta, \sigma^2)$. Let $Z \sim N(0, \sigma^2)$ then,

$$\begin{aligned}
1 + \dot{b}_\theta &= \frac{\partial}{\partial\theta}\mathsf{E}_\theta \frac{\int t e^{-(Z+\theta-t)^2/2\sigma^2} d\pi(t)}{\int e^{-(Z+\theta-t)^2/2\sigma^2} d\pi(t)} \\
&= \frac{1}{\sigma^2}\mathsf{E}\Big\{ \frac{\int t(t-Z-\theta) e^{-(Z+\theta-t)^2/2\sigma^2} d\pi(t)}{\int e^{-(Z+\theta-t)^2/2\sigma^2} d\pi(t)} \\
&\qquad - \frac{\int t e^{-(Z+\theta-t)^2/2\sigma^2} d\pi(t)}{\int e^{-(Z+\theta-t)^2/2\sigma^2} d\pi(t)} \frac{\int (t-Z-\theta) e^{-(Z+\theta-t)^2/2\sigma^2} d\pi(t)}{\int e^{-(Z+\theta-t)^2/2\sigma^2} d\pi(t)} \Big\} \\
&= \frac{1}{\sigma^2}\mathsf{E}_\theta\{\mathrm{var}(\Theta \mid X)\}.
\end{aligned}$$

Hence $0 \le 1 + \dot{b}_\theta \le (a/\sigma)^2$, or $\dot{b}_\theta \in [-1, -(1-(a/\sigma)^2)]$. The lemma then follows from the mean value theorem. □

Proof. [Proof of Theorem 5.2] Let $\boldsymbol{\beta} \sim \pi$. For any i, let F_i be the distribution of $b_i = \mathsf{E}\left(\beta_i^2 | X_{-i}\right)$. Note that b_i and β_i are independent given $\boldsymbol{\beta}_{-i}$. By assumption, conditionally on $\boldsymbol{\beta}_{-i}$, $P_{\pi_i \times F_i}(|\beta_i^2 - b_i| > i^{2\alpha}/4) > \eta$. But then it follows from (7) that for n large enough, $P_\pi(|\beta_i^2 - \widehat{\beta}_i^2| > i^{2\alpha}/4) > \eta/2$. Let $c_i'(\boldsymbol{\beta}) = \mathbf{1}\left\{E_{\boldsymbol{\beta}}(\widehat{\beta}_i^2 - \beta_i^2) < -i^{2\alpha}/4\right\}$, $c_i''(\boldsymbol{\beta}) = \mathbf{1}\left\{E_{\boldsymbol{\beta}}(\widehat{\beta}_i^2 - \beta_i^2) > i^{2\alpha}/4\right\}$, and

$$c_i(\boldsymbol{\beta}) = \begin{cases} c_i'(\boldsymbol{\beta}), & \sum_{i=n^{1/2\alpha+\nu}}^m c_i'(\boldsymbol{\beta}) > \eta/3(m - n^{1/2\alpha+\nu}) \\ c_i''(\boldsymbol{\beta}), & \text{otherwise.} \end{cases}$$

Now, $\sum c_i(\boldsymbol{\beta})\widehat{\beta}_i^2$ picks exactly those β_i^2 which are estimated with bias, positive bias if $c_i = c_i''$ and negative if $c_i = c_i'$. □

Proof. [Proof of Theorem 7.1] The proof is based on the two lemmas which follow. Suppose the posterior is not d_n consistent on $\mathcal{B}'$ with $\pi(\mathcal{B}') > 0$. Then, by Lemma B.1, (8) must hold for $\beta_0 \in \mathcal{B}'$. By Lemma B.2, (10) must hold. But (10) contradicts $\pi(\mathcal{B}) = 1$, since then, for all M, we have $\pi\{\beta : P_\beta(d_n(\tilde{\beta}_n, \beta) \ge M)\} > 0$. □

Recall that β_0 is the true parameter. It has a prior probability π. β_n is a random variable which, given the data X_n, has the posterior distribution π_{X_n}. The first lemma says that if there is a d_n consistent estimator, but $d_n(\beta_n, \beta_0)$ is not tight, then neither is $d_n(\beta_n, \tilde{\beta}_n)$:

Lemma B.1 *Suppose that,*

1. *There is a statistic $\tilde{\beta}_n$ such that $\limsup_n P_{\beta_0}\left(d_n(\tilde{\beta}_n, \beta_0) \ge M\right) \to 0$ as $M \to \infty$.*

2. For all $M < \infty$, $\limsup_n P_{\beta_0}\left(\pi_{X_n}(d_n(\beta_n, \beta_0) \geq 2M) \geq 2\varepsilon\right) \geq 2d$.

Then there is an M which may depend on β_0 such that

$$\limsup_{n\to\infty} P_{\beta_0}\left(\pi_{X_n}(d_n(\beta_n, \tilde{\beta}_n) \geq M) \geq \varepsilon\right) \geq d. \tag{8}$$

Proof.

$$\begin{aligned}
&P_{\beta_0}\left(\pi_{X_n}(d_n(\beta_n, \tilde{\beta}_n) \geq M) \geq \varepsilon\right) \\
&\quad \geq P_{\beta_0}\left(\{\pi_{X_n}(d_n(\beta_n, \beta_0) \geq 2M) \geq 2\varepsilon\} \cap \{d_n(\tilde{\beta}_n, \beta_0) \leq M\}\right) \\
&\quad \geq P_{\beta_0}\left(\pi_{X_n}(d_n(\beta_n, \beta_0) \geq 2M) \geq 2\varepsilon\right) - P_{\beta_0}\left(d_n(\tilde{\beta}_n, \beta_0) \geq M\right)
\end{aligned}$$

By assumption the limsup of the first term on the right-hand side is bounded by $2d$, while we can choose M large enough such that the second term on the right-hand side is bounded by d for all n large enough. The lemma follows. □

The reverse is given in the following lemma:

Lemma B.2 *Suppose there is a statistic $\tilde{\beta}_n$ and $M, \varepsilon, d > 0$ such that*

$$P_{\beta_0}\left(\pi_{X_n}(d_n(\tilde{\beta}_n, \beta_n) \geq M) \geq \varepsilon\right) \geq d \tag{9}$$

for all $\beta_0 \in \mathcal{B}'$ and $\pi(\mathcal{B}') \geq \gamma > 0$. Then for all $M < \infty$:

$$P(d_n(\tilde{\beta}_n, \beta_0) \geq M) \geq \varepsilon d\gamma, \tag{10}$$

Proof. If U, V, W are three random variables, then $E(E(E(U|V)|W) = E(U)$. Computing the expected value of (9), we obtain (10). □

Proof. [Proof of Theorem 7.2] Let $P, P' \in \mathcal{P}^{(n)}$. Then

$$\begin{aligned}
|R_n(P, \delta) - R_n(P', \delta)| &= \left| \int \ell\left(p, \delta(x)\right)\, p(x)\, d\nu(x) - \int \ell\left(p', \delta(x)\right)\, p'(x)\, d\nu(x) \right| \\
&\leq \int |\ell(p, \delta(x)) - \ell(p', \delta(x))|\, p(x)\, d\nu(x) \\
&\qquad + \int \ell(p', \delta(x))\, |p(x) - p'(x)|\, d\nu(x).
\end{aligned}$$

The first term on the right-hand side is bounded by $c_n \|P - P'\|$, and the second by $L_n \|P - P'\|$, so that,

$$|R_n(P, \delta) - R_n(P', \delta)| \leq (c_n + L_n)\|P - P'\|, \tag{11}$$

for all δ, P, and P'.

By the complete class theorem, for any $\varepsilon' > 0$ there is a μ_n supported on $\mathcal{P}_K^{(n)}$ such that,

$$\max_{P\in\mathcal{P}_K^{(n)}} R_n(P,\delta_{\mu_n}) \le \inf_\delta \max_{P\in\mathcal{P}_K^{(n)}} R_n(P,\delta) + \varepsilon'. \tag{12}$$

By (11), we also have,

$$\left|\max_{P\in\mathcal{P}_K^{(n)}} R_n(P,\delta) - \max_{P\in\mathcal{P}^{(n)}} R_n(P,\delta)\right| \le (c_n + L_n)\varepsilon. \tag{13}$$

Combining (12) and (13), applied to δ and δ_{μ_n},

$$\max_{P\in\mathcal{P}^{(n)}} R_n(P,\delta_{\mu_m}) \le \inf_\delta \max_{P\in\mathcal{P}^{(n)}} R_n(P,\delta) + (c_n + L_n)\varepsilon + \varepsilon'.$$

Since $\varepsilon, \varepsilon' > 0$ are arbitrary, the assertion follows. □

References

Bayarri, M. and Berger, J. (2004). The interplay between Bayesian and frequentist analysis. *Statist. Sc.*, **19**, 58–80.

Berger, J. (2006a). The case for objective Bayesian analysis. *Bayesian Analysis*, **1**, 385–402.

Berger, J. (2006b). Rejoinder. *Bayesian Analysis*, **1**, 457–464.

Berger, J. O. and Wolpert, R. L. (1988). *The Likelihood Principle: A Review, Generalizations, and Statistical Implications* (2nd ed.)., volume 6 of *Lecture Notes—Monograph Series*. IMS, Hayward, California.

Berry, S. M., Reese, C. S., and Larkey, P. D. (1999). Bridging different eras in sports. *J. Amer. Statist. Assoc.*, **84**, 661–676.

Bickel, P. and Ritov, Y. (1988). Estimating integrated squared density derivatives. *Sankhya*, **A-50**, 381–393.

Bickel, P. J. (1981). Minimax estimation of the mean of a normal distribution when the parameter space is restricted. *Ann. Statist.*, **9**, 1301–1309.

Bickel, P. J., Klaassen, C. A. J., Ritov, Y., and Wellner, J. A. (1998). *Efficient and adaptive estimation in semiparametric models*. Springer-Verlag, New York.

Bickel, P. J. and Kleijn, B. J. K. (2012). The semiparametric Bernstein-von Mises theorem. *Ann. Statist.*, **40**, 206–237.

Bickel, P. J. and Ritov, Y. (2003). Nonparametric estimators which can be "plugged-in". *Ann. Statist.*, *31*(4), 1033–1053.

Bock, M. E. (2004). Conversations with Herman Rubin. In *A Festschrift for Herman Rubin*, (pp. 408–417). IMS Press.

Brown, L. D. and Low, M. G. (1996). Asymptotic equivalence of nonparametric regression and white noise. *Ann. Statist.*, *24*(6), 2384–2398.

Chen, H. and Shiau, J.-J. H. (1994). Data-driven efficient estimators for a partially linear model. *Ann. Statist.*, *22*(1), 211–237.

Cochran, W. G. (1977). *Sampling Techniques* (3rd ed.). Wiley, New York.

Cox, D. (1993). An analysis of Bayesian inference for nonparametric regression. *Ann. Statist.*, **21**, 903–924.

Diaconis, P. and Freedman, D. (1993). Nonparametric binary regression: a Bayesian approach. *Ann. Statist.*, *21*(4), 2108–2137.

Diaconis, P. and Freedman, D. (1998). Consistency of Bayes estimates for nonparameteric regression: Normal theory. *Bernoulli*, **4**, 411–444.

Donoho, D. L. and Johnstone, I. M. (1994). Minimax risk over l_p-balls for l_q-error. *Probab. Theory Related Fields*, **99**, 277–303.

Donoho, D. L. and Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Assoc.*, **90**, 1200–1224.

Engle, R. F., Granger, C. W. J., Rice, J., and Weiss, A. (1986). Nonparametric estimates of the relation between weather and electricity sales. *J. Amer. Statist. Assoc.*, **81**, 310–320.

Everson, P. J. and Morris, C. N. (2000). Inference for multivariate normal hierarchical models. *J. Roy. Statist. Soc.*, **B 62**, 399–412.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.*, *1*(2), 209–230.

Freedman, D. (1963). On the asymptotic behavior of Bayes estimates in the discrete case I. *Ann. Math. Statist.*, **34**, 1386–1403.

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

Freedman, D. A. (1965). On the asymptotic behavior of Bayes estimates in the discrete case II. *Ann. Math. Statist.*, **36**, 454–456.

Ghosal, S., Ghosh, J., and van der Vaart, A. (2000). Convergence rates of posterior distributions. *Ann. Statist.*, **28**, 500–531.

Goldstein, M. (2006). Subjective Bayesian analysis: Principles and practice. *Bayesian Analysis*, **1**, 403–420.

Greenshtein, E., Park, J., and Ritov, Y. (2008). Estimating the mean of high valued observations in high dimensions. *J. Statist. Th. Practice*, **2**, 407–418.

Harmeling, S. and Toussaint, M. (2007). Bayesian estimators for Robins-Ritov's problem. Technical report, University of Edinburgh, School of Informatics Research Report EDI-INF-RR-1189.

Ibragimov, I. A. and Hasminskii, R. Z. (1984). On nonparametric estimation of a linear functional in Gaussian white noise. *Theory of Probability and its Applications*, *29*(1), 19–32.

Kleijn, B. and van der Vaart, A. (2006). Misspecification in infinite-dimensional Bayesian statistics. *Ann. Statist.*, **34**, 837–877.

Le Cam, L. and Yang, G. (1990). *Asymptotics in Statistics: Some Basic Concepts*. Springer, New York.

Lehmann, E. and Casella, G. (1998). *Theory of Point Estimation*. Springer, New York.

Li, K. (1999). Testing symmetry and proportionality in ppp. *J. Bus. Econom. Statist.*, **17**, 409–418.

Li, L. (2010). Are Bayesian inferences weak for Wasserman's example? *Communications in Statistics—Simulation and Computation*, **39**, 657–667.

Lindley, D. V. (1953). Statistical inference. *JRSS-B*, **15**, 30–76.

Lindley, D. V. and Smith, A. F. M. (1972). Bayes estimates for the linear model. *J. Roy. Statist. Soc.*, **B34**, 1–41.

McShane, B. B. and Wyner, A. J. (2011). A statistical analysis of multiple temperature proxies: Are reconstructions of surface temperatures over the last 1000 years reliable? *Ann. Appl. Stat.*, **5**, 5–44.

Nussbaum, M. (1996). Asymptotic equivalence of density estimation and gaussian white noise. *Ann. Statist.*, *24*(6), 2399–2430.

Robins, J., Tchetgen, E. T., Li, L., and van der Vaart, A. (2009). Semiparametric minimax rates. *Electron. J. Statist.*, **3**, 1305–1321.

Robins, J. M. and Ritov, Y. (1997). Toward a curse of dimensionality appropriate (coda) asymptotic theory for semiparametric models. *Statistics in Medicine*, **17**, 285–319.

Savage, L. J. (1961). The foundations of statistical inference reconsidered. In Neyman, J. (Ed.), *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, volume I, (pp. 575–586). Cambridge University Press, London.

Schick, A. (1986). On efficient estimation in regression models. *Ann. Statist.*, **14**, 1486–1521.

Smith, A. F. M. (1986). Some Bayesian thoughts on modelling and model choice. *J. Roy. Statist. Soc., Series D (The Statistician)*, *35*(2), 97–101.

van der Pas, S. L., Kleijn, B. J. K., and van der Vaart, A. W. (2013). The horseshoe estimator: Posterior concentration around nearly black vectors. *(submitted to Electron. J. Statist.)*.

Wang, L., Brown, L. D., and Cai, T. T. (2011). A difference based approach to the semiparametric partial linear model. *Electron. J. Statist.*, **5**, 619–641.

Wasserman, L. (2000). Asymptotic inference for mixture models using data-dependent priors. *J. Roy. Statist. Soc.*, *62*(1), 159–180.

Wasserman, L. (2004). *All of Statistics: A Concise Course in Statistical Inference*. Springer, New York.

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