10

Consistency of Bayes Estimates for Nonparametric Regression: A Review

P. Diaconis[1]
D. A. Freedman[2]

ABSTRACT This paper reviews some recent studies of frequentist properties of Bayes estimates. In nonparametric regression, natural priors can lead to inconsistent estimators; although in some problems, such priors do give consistent estimates.

10.1 Introduction

Consider a sequence of iid pairs (Y_1, ξ_1), (Y_2, ξ_2),... with $E(Y_i \mid \xi_i) = f(\xi_i)$. Here, f is an unknown function to be estimated from the data. A Bayesian approach postulates that f lies in some class of functions Θ and puts a prior distribution π on Θ. This generates a posterior distribution $\tilde{\pi}_n$ on Θ: the conditional law of the regression function f given the data $(Y_1, \xi_1), (Y_2, \xi_2), \ldots, (Y_n, \xi_n)$. The prior π is said to be consistent at f if $\tilde{\pi}_n$ converges to point mass at f almost surely as $n \to \infty$.

When Θ is finite-dimensional, π will be consistent at any f in the support of π; some additional regularity conditions are needed. If Θ is infinite-dimensional, the situation is quite different, and inconsistency is the rule rather than the exception. Section 2 reviews examples where commonly-used "hierarchical priors" lead to inconsistency in infinite-dimensional binary regression problems (the response variable Y_i takes only the values 0 and 1).

Section 3 discusses the root problem behind these inconsistencies, which is deceptively simple: if you choose p at random and toss a p-coin once, that is just like tossing a $\bar{p}$-coin, where $\bar{p}$ is the average of p. In other words, a mixture of Bernoulli variables is again Bernoulli. When the relevant class of measures is not closed under mixing, in a sense to be made precise below, we believe that hierarchical priors will be consistent under standard regularity

[1]Harvard University
[2]University of California at Berkeley

D. Pollard et al. (eds.), *Festschrift for Lucien Le Cam*

conditions. Section 4 contains one such theorem, for normal regression. Our mixing condition is satisfied, because a mixture of $N(\mu, 1)$ variables cannot be $N(\mu, 1)$.

In the balance of this section, we discuss some history. Lucien Le Cam is a major contributor to the study of frequentist properties of Bayes procedures. Le Cam's first boss in France, Etienne Halphen, was a staunch Bayesian who sought to convert the young Lucien. This was not to be, but it did stimulate a lifelong interest in Bayes procedures.

Formal work began with the thesis: Le Cam (1953) proved a version of what has come to be known as the Bernstein-von Mises theorem. Le Cam's theorems were almost sure results, with respect to the true underlying measure that had generated the data, and he proved convergence in total variation norm. Previous authors had demonstrated only convergence of distribution functions, in probability. Furthermore, Le Cam seems to have been the first to condition on all the data, not just a summary statistic (like the sample mean).

Le Cam (1958) explained how localizing the prior affects convergence of the posterior. Breiman, Le Cam & Schwartz (1964) gave versions of Doob's theorem showing consistency, starting from the joint distribution of parameters and data. Le Cam (1982) gave bounds—rather than asymptotic theory—for Bayes risk. Also see Le Cam & Yang (1990).

A more complete exposition of these results can be found in Lucien's book (Le Cam 1986). Convergence properties of Bayes estimates are closely related to the behavior of maximum likelihood estimates. Le Cam (1990) gives a beautiful overview of counter-examples in the latter area.

Frequentist properties of Bayes rules have been studied since Laplace (1774), who showed that in smooth, finite-dimensional problems, the posterior concentrates in a neighborhood of the maximum likelihood estimates. Modern versions of the result can be found in Bernstein (1934), von Mises (1964), Johnson (1967, 1970), Le Cam (1982), or Ghosh, Sinha & Joshi (1982). These results hold for almost all data sequences. In very simple settings, we obtained bounds that hold for all sequences (Diaconis & Freedman 1990).

Freedman (1963) was an early paper on nonparametric Bayes procedures, with a counter-example: there is a prior supported on all of the parameter space, whose posterior converges almost surely to the wrong answer. This paper introduced the Dirichlet and tail free priors, and showed them to be consistent. For reviews, see Ferguson (1974) or Diaconis & Freedman (1986).

Bayesian regression, with hierarchical priors, was developed in finite-dimensional settings by Lindley & Smith (1972). In non-parametric regression, there is an early paper by Kimeldorf & Wahba (1970), who use Gaussian processes to define priors; for a review, see Wahba (1990). Cox (1993) studies frequentist coverage properties of posterior confidence sets. Also see Kohn & Ansley (1987). Diaconis (1988) traces the history back to

Poincaré and gives many further references.

The simplest possible regression problem has a constant regression function. That is the location problem: $Y_i = \mu + \epsilon_i$, where μ is an unknown constant and the errors ϵ_i are iid. Diaconis & Freedman (1986) studied nonparametric priors on μ and the law of the errors; also see Doss (1984, 1985*a*, 1985*b*). Some natural priors lead to inconsistent estimates, while other priors give consistent results.

10.2 Binary regression

This section summarizes results from Diaconis & Freedman (1993*a*, 1993*b*). There is a binary response variable Y, which is related to a covariate ξ:

$$P\{Y = 1|\xi\} = f(\xi) \tag{1}$$

The problem is to estimate f from the data.

Following de Finetti (1959, 1972), we think of ξ as a sequence of 0s and 1s. Sequence space is given the usual product topology, and the parameter space Θ is the set of measurable functions f from sequence space to $[0, 1]$. The L_2 topology is installed on Θ, relative to coin-tossing measure λ in sequence space. A basic neighborhood of $f \in \Theta$ is

$$N(f, \epsilon) = \{g : \int (g - f)^2 d\lambda < \epsilon\} \tag{2}$$

We will consider a prior π on Θ, with posterior $\tilde{\pi}_n$. Then π is consistent at f provided $\tilde{\pi}_n\{N(f, \epsilon)\} \to 1$ almost surely, for all positive ϵ.

The next step is to define the hierarchical priors on Θ. Begin with a prior π_k supported on the class of functions f that depend only on the first k coordinates, or bits, in ξ. Under π_0, the function f does not depend on ξ at all. Under π_1, f depends only on ξ_1. And so forth. Then treat k as an unknown "hyper-parameter", putting prior weight w_k on k. We refer to k as the "theory index"; theory k says that $f(x)$ depends only the first k bits of x; and w_k are the "theory weights." Our prior is of the form

$$\pi = \sum_{k=0}^{\infty} w_k \pi_k / \sum_{k=0}^{\infty} w_k \tag{3}$$

where

$$w_k > 0 \text{ for all } k \text{ and } \sum_{k=0}^{\infty} w_k < \infty \tag{4}$$

To complete the description of the prior, π_k must be specified. According to π_k, only the first k bits in ξ matter, so f depends only on $\xi_1, \ldots, \xi_k$. Thus, π_k is determined by specifying the joint distribution of the 2^k possible values for f. Here, we take these to be independent and uniformly

distributed over $[0,1]$. More general priors are considered in Diaconis & Freedman (1993*a*, 1993*b*).

We turn now to the data. For technical reasons, it is simplest to consider "balanced" data, as in Diaconis & Freedman (1993*a*); more conventional sampling plans are discussed in Diaconis & Freedman (1993*b*). At stage n, there are 2^n subjects. Each has a covariate sequence; the first n bits of these covariate sequences cover all possible patterns of length n; each pattern appears once and only once. The remaining bits from $n+1$ onward are generated by coin tossing. Given the covariates, response variables are generated from (1); the response of subject i depends only on the covariates for that subject. The preliminaries are now finished, and we can state a theorem.

Theorem 1 *With nonparametric binary regression, balanced data, and a hierarchical uniform prior:*

(a) *π is consistent at f unless $f \equiv 1/2$;*

(b) *Suppose $f \equiv 1/2$. Then π is consistent at f provided that for some $\delta > 0$, for all sufficiently large n,*

$$\sum_{k=n}^{\infty} w_k < 2^{-n(\frac{1}{2}+\delta)}$$

On the other hand, π is inconsistent at f provided that for some $\delta > 0$, for infinitely many n,

$$\sum_{k=n}^{\infty} w_k > 2^{-n(\frac{1}{2}-\delta)}$$

The surprising point is the inconsistency result in part (b). Suppose the data are generated by tossing a fair coin: $f \equiv 1/2$. Theory 0 is true: f does not depend on ξ at all. You don't know that, and allow theories of finite but arbitrary complexity in your prior, according to (3) and (4). In the face of all these other theories, the posterior loses faith in theory 0. The curse of dimensionality strikes again.

Regression is a natural problem, hierarchical priors are often used, and the one defined by (3) and (4) charges every weak star neighborhood of the parameter space Θ. Still, inconsistency may result. In high-dimensional problems, little can be taken for granted. "Rational use of additional information" is not a slogan to be adopted without reflection.

10.3 Why inconsistency?

What is the root cause of the inconsistency? Suppose $f \equiv 1/2$, so the data result from coin tossing, and the covariates do not matter. Thus, theory 0

is the truth. The statistician does not know this, however, and high-order theories may be deceptively attractive because they have many parameters.

To make this a little more precise, consider a design of order n, so there are 2^n subjects. According to theory n, the response of each subject is determined by the toss of a coin, where the probability is uniform on $[0, 1]$. Now one toss of a uniform coin is like one toss of a fair coin—you get heads with probability $1/2$ and tails with probability $1/2$. Thus, theory n competes with theory 0. Indeed, the predictive probability of the data under theory n is

$$\pi_n\{\text{data}\} = 1/2^{2^n}.$$

Let S be the sum of the response variables—the total number of heads. Under theory 0, the predictive probability of the data is

$$\pi_0\{\text{data}\} = \left[(2^n + 1)\binom{2^n}{S}\right]^{-1} \approx \frac{\sqrt{\pi/2}}{2^{n/2}}\pi_n\{\text{data}\}$$

because $S \approx 2^n/2$. Thus,

$$\pi_n\{\text{data}\} = \text{const.}\, 2^{n/2}\pi_0\{\text{data}\} \tag{5}$$

The prior π is a mixture $\sum_{k=0}^{\infty} w_k\pi_k / \sum_{k=0}^{\infty} w_k$. The posterior is a similar mixture, the posterior weight on theory k being w_k times the predictive probability of the data under π_k. If $f \equiv 1/2$, then, it is the theory weights w_k that decide consistency. If w_k declines rapidly, for example, $w_k = 1/2^k$, the weight on theory n compensates for the factor $2^{n/2}$ in (5); and the prior is consistent at $f \equiv 1/2$. On the other hand, if w_k declines slowly, for example, $w_k = 1/(k+1)^2$, the factor $2^{n/2}$ dominates, and inconsistency is the result.

The heart of the problem seems to be that a mixture of Bernoulli variables is again Bernoulli. For example, suppose the response variable takes three values, 0, 1 and 2; and, given the covariates ξ, the response is distributed as the number of heads when an $f(\xi)$-coin is tossed twice. A mixture of bin $(2, p)$ variables cannot be bin $(2, p)$; the heuristic suggests that Bayes estimates will be consistent.

To discuss this kind of theorem in any degree of generality, we would need to impose smoothness conditions like those which underly the usual asymptotics of maximum likelihood estimates, including the Bernstein-von Mises theorem; and integrability conditions of the kind which underly the usual theory of entropy bounds. The second set of conditions would enable us to localize the problem, and the first set would enable us to make local estimates. Rather than pursue such technical issues here, we discuss one simple form of the theorem, for normal response variables.

10.4 Normal regression

Suppose the response variable is normal with mean μ and variance 1, where $\mu = f(\xi)$. The covariates are a sequence of 0's and 1's. We require the subjects to satisfy the balance condition as before. We assume:

Assumption 1 *Given the covariates, the response variables are independent across subjects, and normally distributed, with common variance 1 and* $E\{Y \mid \xi\} = f(\xi)$.

The function f is assumed to be measurable; f may be unbounded, but we require f to be square integrable (relative to coin-tossing measure). We define hierarchical priors, consistency, etc., as before. However, the π_k are assumed for convenience to be "normal" in the following sense: theory k says that $f(x)$ depends only on the first k bits of x; under π_k, the 2^k possible values of f are independent $N(0,1)$ variables. Thus, π is a conventional hierarchical normal prior.

This completes the setup. The main theorems can now be stated.

Theorem 2 *Suppose the design is balanced and normal in the sense of Assumption 1. Suppose the prior* π *is hierarchical, and the* π_k *are normal. Then* π *is consistent at all* f.

Let C_k be the class of L_2 functions f which depend only on the first k bits of the argument x. Recall that $\tilde{\pi}_n$ is the posterior given the data at stage n.

Theorem 3 *Suppose the design is balanced, and normal in the sense of Assumption 1. Suppose the prior* π *is hierarchical, the* π_k *are normal, and* $f \in C_k$ *for some* k. *Then* $\tilde{\pi}_n\{C_k\} \to 1$ *a.e. as* $n \to \infty$.

Theorem 2 demonstrates consistency, while Theorem 3 says that the Bayesian gets the order of a finite model right. This is a bit surprising, because many model selection algorithms over-estimate the order of a finite model. For proofs, see Diaconis & Freedman (1994).

Cox (1993) has results for a similar problem, and it may be worth a moment to indicate the differences. That paper uses a different prior, based on Gaussian processes; the covariates are deterministic and equally spaced, rather than completely at random; and results depend on the behavior of $f(x)$ at rational x, rather than a.e. properties of f.

Acknowledgments: Research of Diaconis partially supported by NSF Grant DMS 86-00235. Research of Freedman partially supported by NSF Grant DMS 92-08677.

10.5 References

Bernstein, S. (1934), *Theory of Probability*, GTTI, Moscow. (Russian).

Breiman, L., Le Cam, L. & Schwartz, L. (1964), 'Consistent estimates and zero-one sets', *Annals of Mathematical Statistics* **35**, 157–161.

Cox, D. (1993), 'An analysis of Bayesian inference for nonparametric regression', *Annals of Statistics* **21**, 903–923.

de Finetti, B. (1959), *La probabilità, la statistica, nei rapporti con l'induzione, secondo diversi punti di vista*, Centro Internazionale Matematica Estivo Cremonese, Rome. English translation in de Finetti (1972).

de Finetti, B. (1972), *Probability, Induction, and Statistics*, Wiley, New York.

Diaconis, P. (1988), Bayesian numerical analysis, *in* S. S. Gupta & J. O. Berger, eds, 'Statistical Decision Theory and Related Topics IV', Vol. 1, pp. 163–177.

Diaconis, P. & Freedman, D. (1986), 'On the consistency of Bayes estimates (with discussion)', *Annals of Statistics* **14**, 1–67.

Diaconis, P. & Freedman, D. (1990), 'On the uniform consistency of Bayes estimates for multinomial probabilities', *Annals of Statistics* **18**, 1317–1327.

Diaconis, P. & Freedman, D. (1993*a*), 'Nonparametric binary regression: a Bayesian approach', *Annals of Statistics* **21**, 2108–2137.

Diaconis, P. & Freedman, D. (1993*b*), Nonparametric binary regression with random covariates, Technical Report 291, Department of Statistics, University of California, Berkeley. (To appear in *Probability and Mathematical Statistics.*).

Diaconis, P. & Freedman, D. (1994), Consistency of Bayes estimates for nonparametric regression: normal theory, Technical Report 414, Department of Statistics, University of California, Berkeley.

Doss, H. (1984), 'Bayesian estimation in the symmetric location problem', *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **68**, 127–147.

Doss, H. (1985*a*), 'Bayesian nonparametric estimation of the median; part I: Computation of the estimates', *Annals of Statistics* **13**, 1432–1444.

Doss, H. (1985*b*), 'Bayesian nonparametric estimation of the median; part II: Asymptotic properties of the estimates', *Annals of Statistics* **13**, 1445–1464.

Ferguson, T. (1974), 'Prior distributions on spaces of probability measures', *Annals of Statistics* **2**, 615–629.

Freedman, D. (1963), 'On the asymptotic behavior of Bayes estimates in the discrete case', *Annals of Mathematical Statistics* **34**, 1386–1403.

Ghosh, J. K., Sinha, B. K. & Joshi, S. N. (1982), Expansions for posterior probability and integrated Bayes risk, *in* S. S. Gupta & J. O. Berger, eds, 'Statistical Decision Theory and Related Topics III', Vol. 1, Academic Press, New York, pp. 403–456.

Johnson, R. (1967), 'An asymptotic expansion for posterior distributions', *Annals of Mathematical Statistics* **38**, 1899–1906.

Johnson, R. (1970), 'Asymptotic expansions associated with posterior distributions', *Annals of Mathematical Statistics* **41**, 851–864.

Kimeldorf, G. & Wahba, G. (1970), 'A correspondence between Bayesian estimation on stochastic processes and smoothing by splines', *Annals of Mathematical Statistics* **41**, 495–502.

Kohn, R. & Ansley, C. (1987), 'A new algorithm for spline smoothing and interpolation based on smoothing a stochastic process', *SIAM Journal on Scientific and Statistical Computing* **8**, 33–48.

Laplace, P. S. (1774), 'Memoire sur la probabilité des causes par les évènements', Memoires de mathématique et de physique presentés a l'académie royale des sciences, par divers savants, et lûs dans ses assemblées. Reprinted in Laplace's Oeuvres Complètes 8 27-65. English translation by S. Stigler (1986) *Statistical Science* **1** 359-378.

Le Cam, L. (1953), 'On some asymptotic properties of maximum likelihood estimates and related Bayes estimates', *University of California Publications in Statistics* **1**, 277–330.

Le Cam, L. (1958), 'Les propriétés asymptotiques des solutions de Bayes', *Publications de l'Institut de Statistique de l'Université de Paris* **7**, 17–35.

Le Cam, L. (1982), On the risk of Bayes estimates, *in* S. S. Gupta & J. O. Berger, eds, 'Statistical Decision Theory and Related Topics III', Vol. 2, Academic Press, New York, pp. 121–138.

Le Cam, L. (1986), *Asymptotic Methods in Statistical Decision Theory*, Springer-Verlag, New York.

Le Cam, L. (1990), 'Maximum likelihood: an introduction', *International Statistical Review* **58**, 153–172.

Le Cam, L. & Yang, G. L. (1990), *Asymptotics in Statistics: Some Basic Concepts*, Springer-Verlag.

Lindley, D. & Smith, A. (1972), 'Bayes estimates for the linear model', *Journal of the Royal Statistical Society* **67**, 1–19.

von Mises, R. (1964), *Mathematical Theory of Probability and Statistics*, Academic Press, New York. H. Geiringer, ed.

Wahba, G. (1990), *Spline Models for Observational Data*, Society for Industrial and Applied Mathematics, Philadelphia.