The Annals of Statistics
2015, Vol. 43, No. 4, 1429–1436
DOI: 10.1214/14-AOS1270A
Main article DOI: 10.1214/14-AOS1270

DISCUSSION OF "FREQUENTIST COVERAGE OF ADAPTIVE NONPARAMETRIC BAYESIAN CREDIBLE SETS"

BY RICHARD NICKL

University of Cambridge

1. Introduction. I would like to congratulate Botond Szabó, Aad van der Vaart and Harry van Zanten [12] for a fundamental and thought provoking article on a highly important topic. One of the key contributions of statistics to modern science may arguably be *the theory of uncertainty quantification*. Assessing the accuracy of an estimate by a confidence statement goes beyond the mere search for an efficient statistical algorithm. In particular, within the contemporary search for *adaptive* procedures, research of the last decade has revealed that the construction of adaptive confidence statements is fundamentally *harder*—in an information theoretic sense—than the construction of adaptive algorithms. Confidence statements are at the same time crucial for the main application of modern data analysis, which is to accept or reject hypotheses.

Szabó, van der Vaart and van Zanten tackle the important topic as to whether increasingly popular Bayesian methodology can actually provide objective uncertainty quantification methods in nonparametric models or not. The nonparametric setting is a key test-case for the general paradigm of high-dimensional modeling that has emerged recently in statistics.

My discussion of the paper surrounds the two focal points of why "Bayesian uncertainty quantification" is a mathematically and conceptually nontrivial problem: the first has nothing to do with adaptation and addresses some of the probabilistic subtleties intrinsic to the Bayesian approach to provide "credible sets." The second point is common to all frequentist procedures and is about the fact that adaptive uncertainty quantification is in general only possible under "signal-strength" conditions on the underlying parameter.

2. Freedman's paradox and the nonparametric Bernstein–von Mises theorem. I first want to discuss the fact that the frequentist coverage probabilities obtained by Szabó, van der Vaart and van Zanten for their credible sets are not *exact*, that is to say, not of the precise asymptotic level $1-\alpha$, and the related question of why obtaining exact posterior asymptotics in the nonparametric situation is a subtle matter.

Consider observations $Y \sim P_\theta$ with parameter space $\theta \in \Theta$, a prior Π on Θ and resulting posterior distribution $\Pi(\cdot|Y)$ of $\theta|Y$. The classical finite-dimensional

Received September 2014.

($\Theta \subseteq \mathbb{R}^p$) *Bernstein–von Mises theorem* asserts—under fairly mild assumptions on Π and on the parameterization $\theta \mapsto P_\theta$—that we have approximately

$$\mathcal{L}\big(\sqrt{n}(\theta - \bar{\theta})|Y\big) \approx N\big(0, I(\theta_0)^{-1}\big) \qquad \text{when } Y \sim P_{\theta_0}, \theta_0 \in \Theta.$$

Here $\bar{\theta} = \bar{\theta}(Y)$ is any efficient estimator of θ such as the maximum likelihood estimator (MLE) or the posterior mean $E(\theta|Y)$, $I(\theta_0)$ is the Fisher information, and the approximation holds in the small noise or large sample limit, in total variation distance. As a consequence, computing posterior probabilities is approximately equivalent to computing "optimal frequentist" probabilities under $N(\bar{\theta}, I(\theta_0)^{-1}/n)$, and the natural level $1 - \alpha$ Bayesian credible set

$$C_n = \{\theta : \|\theta - E(\theta|Y)\| \le r_{\alpha,n}\} \qquad \text{with } r_{\alpha,n} \text{ s.t. } \Pi(C_n|Y) = 1 - \alpha \tag{1}$$

asymptotically coincides with the classical one based on the MLE. In particular, we have frequentist coverage

$$P_{\theta_0}(\theta_0 \in C_n) \to 1 - \alpha \qquad \text{as } n \to \infty. \tag{2}$$

For the frequentist the main idea behind this phenomenon is similar to the bootstrap: if $Y \sim P_{\theta_0}$, the (known) posterior distribution of $\theta|Y - E(\theta|Y)$ serves as a proxy for the (unknown) distribution of $E(\theta|Y) - \theta_0$.

In his influential 1999 Wald lecture, Freedman [5] has shown that in the case where Θ is infinite-dimensional, the above phenomenon need not occur. Freedman considered precisely the setting of Szabó, van der Vaart and van Zanten: in the standard nonparametric sequence space model

$$Y_k = \theta_k + \frac{1}{\sqrt{n}} g_k, \qquad k \in \mathbb{N}; \qquad g_k \overset{\text{i.i.d.}}{\sim} N(0,1), \qquad \theta \in \ell_2, \tag{3}$$

one considers natural conjugate Gaussian priors

$$\Pi = \bigotimes_{k \in \mathbb{N}} N(0, k^{-1-2\gamma}), \qquad \gamma > 0, \tag{4}$$

for a γ-regular signal θ. Freedman then showed that even when the true signal θ_0 is β-regular with $\beta > \gamma$—so in a favorably "well-specified" situation where β is known—the natural posterior credible set paralleling (1),

$$\begin{aligned} C_n = \{\theta \in \ell_2 : \|\theta - E(\theta|Y)\|_{\ell_2} \le r_{\alpha,n}\} \\ \text{with } r_{\alpha,n} \text{ s.t. } \Pi(C_n|Y) = 1 - \alpha, \end{aligned} \tag{5}$$

does in fact *not* satisfy (2) as $n \to \infty$, rather the frequentist and Bayesian variances of $\|\theta - E(\theta|Y)\|_{\ell_2}^2$ scale differently and the Bernstein–von Mises theorem does not hold in this infinite-dimensional setting. See Freedman's original paper [5] and also Leahu [8] for a recent account.

We note that this "paradox" has nothing to do with "adaptation" (since β is known above), but is a mathematical artefact of the Bayesian formalism to construct credible sets in the infinite-dimensional setting. It prevents Bayesian $1 - \alpha$

credible balls in ℓ_2 from being asymptotically *exact* frequentist $1-\alpha$ confidence balls. The pathology occurs because one insists on "exact" asymptotic level $1-\alpha$, and the results by Szabó, van der Vaart and van Zanten show that if one "blows up the Bayesian radius $r_{\alpha,n}$" by a factor of L as in equation (3.2) of their paper [12], then as $L \to \infty$ this allows to obtain "conservative" frequentist confidence sets in the sense that, as $n \to \infty$,

$$P_{\theta_0}(\theta_0 \in C_n) \to 1 \geq 1-\alpha.$$

While such a construction is theoretically satisfactory from a frequentist point of view, this approach has a practical drawback: in applications one does not know how to choose L and prefers to have a simple, fully Bayesian, rule that discards 5% of all posterior draws and uses the remaining 95% to graphically describe a credible region.

In the recent papers [3, 4] by Castillo and myself, a new approach to nonparametric Bernstein–von Mises theorems has been put forward. In essence, the idea is to modify the geometry of the credible set in (5) and to replace ℓ_2-balls by other shapes. These shapes correspond to norms in sequence space that induce weaker topologies than ℓ_2 and for which a "weak functional Bernstein–von Mises theorem" can be proved. For instance, if we consider ellipsoids in a sequence space of the form

$$\text{(6)}\quad \mathcal{E}(M) = \left\{(\theta_k) : \sum_k \frac{\theta_k^2}{w_k} \leq M^2\right\}, \qquad \frac{w_k}{k(\log k)^\delta} \uparrow \infty, \delta > 1, 0 < M < \infty$$

or, in case the sequence space model corresponds to a double-indexed basis $\{e_{lk} : l \in \mathbb{N} \cup \{0\}, k = 0, \dots, 2^l - 1\}$, multi-scale sup-norm balls of the form

$$\text{(7)}\qquad \mathcal{E}(M) = \left\{(\theta_{lk}) : \sup_{l \geq 0} \frac{\max_k |\theta_{lk}|}{w_l} \leq M\right\}, \qquad \frac{w_l}{\sqrt{l}} \uparrow \infty, 0 < M < \infty,$$

then, under mild assumptions on the prior, [3, 4] prove that, as $n \to \infty$,

$$\text{(8)}\qquad \mathcal{L}\big(\sqrt{n}(\theta - \bar{\theta})|Y\big) \to \mathcal{N} \qquad \text{weakly in } \mathbb{H},$$

in P_{θ_0}-probability. Here $\mathbb{H}$ are the sequence spaces that have norm balls $\{\theta : \|\theta\|_{\mathbb{H}} \leq 1\} = \mathcal{E}(1)$, $\mathcal{N}$ is the Gaussian measure on $\mathbb{H}$ corresponding to a pure white noise $\bigotimes_{k \in \mathbb{N}} N(0,1)$, and $\bar{\theta} = \bar{\theta}(Y)$ is equal to the maximum likelihood estimator $Y = (Y_k : k \in \mathbb{N})$ or to the posterior mean $E(\theta|Y)$.

As a consequence of weak convergence toward $\mathcal{N}$, one can show that

$$\text{(9)}\qquad \sup_M \big|\Pr\big(\sqrt{n}(\theta - E(\theta|Y)) \in \mathcal{E}(M)|Y\big) - \mathcal{N}\big(\mathcal{E}(M)\big)\big| \overset{P_{\theta_0}}{\to} 0$$

as $n \to \infty$, and from this [3, 4] deduce that credible sets

$$C_n = \{\theta : \|\theta - E(\theta|Y)\|_{\mathbb{H}} \leq r_{\alpha,n}\} \qquad \text{where } r_{\alpha,n} \text{ is such that } \Pi(C_n|Y) = 1-\alpha$$

have correct asymptotic frequentist coverage: as $n \to \infty$,

$$P_{\theta_0}(\theta_0 \in C_n) \to 1 - \alpha.$$

Two main questions arise from this construction, one theoretical, one practical. The theoretical one asks whether such confidence sets can reconstruct nonparametric signals in a minimax optimal way. In [3, 4] it is shown that this can be the case by using high-frequency information in the posterior appropriately. This can be extended to the adaptive setting (see Ray [11]), where nonparametric Bernstein–von Mises theorems in $\mathbb{H}$ are proved for the empirical Bayes procedure of Szabó, van der Vaart and van Zanten.

The second question is as follows: is such a construction practical, and do such credible sets look substantially different from the (perhaps) more intuitive ℓ_2-type credible sets? From a computational point of view the sets C_n are quite tractable: for instance, in the multi-scale case the computation of C_n consists of finding constants $r_{\alpha,n}$ such that

$$\frac{|\theta_{lk} - E(\theta_{lk}|Y)|}{w_l} \leq r_{\alpha,n} \qquad \forall k, l$$

happens for $(1-\alpha) \times 100\%$ of the posterior draws. The theory in [4] implies

$$\sqrt{n} \cdot r_{\alpha,n} \overset{P_{\theta_0}}{\to} \text{const} \neq 0,$$

and so a multi-scale posterior credible ball has a natural interpretation as a simultaneous credible set for a large class of semi-parametric coordinate projection functionals obtained from thresholding each projection at a level slightly larger than $1/\sqrt{n}$ (recalling that w_l is slightly larger than $\sqrt{l}$).

More concretely, simulations by Ray [11] show that the differences to the standard ℓ_2-approach are marginal in several practical examples; see Figure 1 below.

It is striking to observe that, although the norms $\|\cdot\|_{\ell_2}$ and $\|\cdot\|_{\mathbb{H}}$, as well as the rules these norms induce to accept or reject posterior draws in the construction of a credible set, are quite different, the visualized credible sets of both approaches look very similar.

It is also worthwhile noting that in both cases the credible ball actually "covers the true function" despite the graph suggesting that pointwise coverage fails. The reason is that ℓ_2-confidence balls are insensitive to lack of coverage on intervals of small Lebesgue measure. A frequentist theory for simultaneous credible "bands" is thus also of interest—some first results in this direction are given in [4, 11].

I am unsure to which extent ℓ_2-credible sets are "applied in current practice" as claimed in the Introduction of [12], particularly if one has to choose "blow-up" constants L. Practitioners may prefer to avoid such choices, and instead compute posterior credible balls in $\mathbb{H}$-spaces. At any rate, it remains a mathematical fact that the nonparametric Bernstein–von Mises theorem *does* hold in the spaces $\mathbb{H}$, whereas it does *not* hold in ℓ_2.

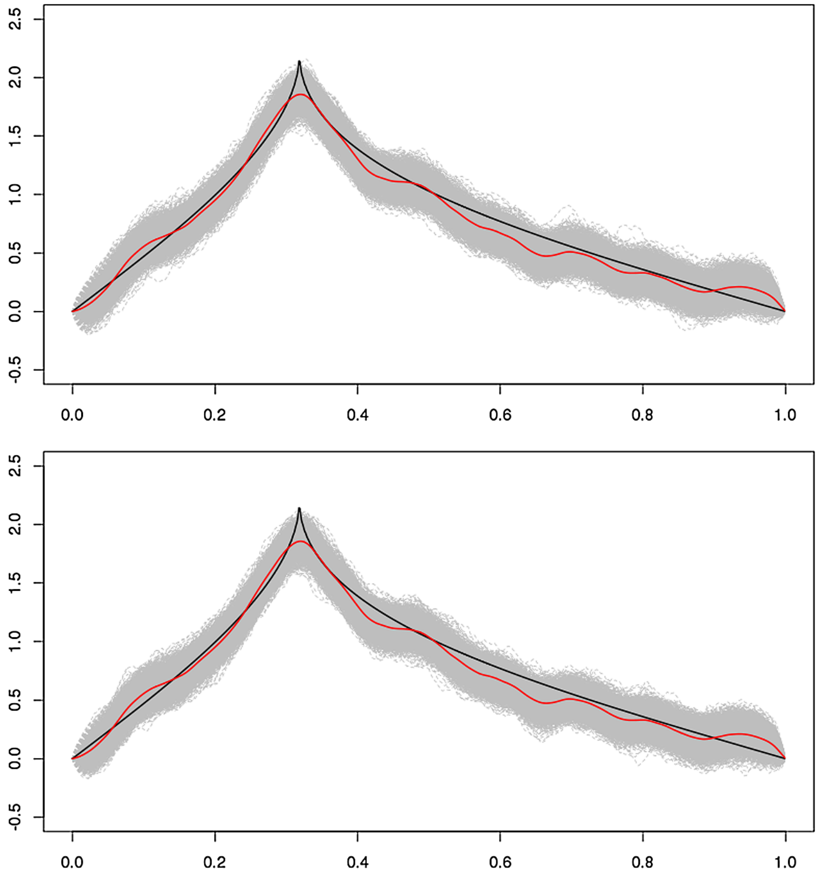

FIG. 1. *The top display shows a credible set generated from ellipsoids* (6) *and the bottom from an ℓ_2 ball. Both credible sets are based on observations in Gaussian white noise* ($n = 1000$), *and with Gaussian priors*, *with* 100,000 *posterior draws plotted as grey clouds. The red curve depicts the posterior mean and the black curve the true function. See* [11] *for details and more extensive simulation results*.

3. "Honest" nonparametric models and "self-similar" functions. Perhaps more important than the question of how to obtain *exact* coverage statements for Bayesian credible sets (discussed in the previous section) is the question of existence of *adaptive* confidence sets—Bayesian or not. It is one of the more surprising insights of the theory of nonparametric and high-dimensional inference that estimators that adapt to unknown regularity properties (such as smoothness or sparsity) exist, whereas associated confidence sets in general *do not*. Roughly speaking, the reason behind this is that an "honest" (=uniform in the parameter θ) adaptive confidence set implicitly solves the testing problem of whether a signal belongs to a given regularity class or not, and that such tests simply do not exist over the entire parameter spaces considered in nonparametric *estimation*. Rather, some kind of *signal-strength condition* needs to be enforced on the elements of the parameter space to construct confidence sets for adaptive estimators. See Hoffmann and Nickl [7] and Nickl and van de Geer [10] for two basic instances of this fact (in nonparametrics and sparse regression, resp.).

Among such signal-strength conditions, the "self-similarity" assumptions introduced in Giné and Nickl [6] have proved compatible with commonly used adaptive frequentist procedures (such as Lepski's method). In the L^∞-setting of confidence bands they are also shown to be necessary (see [1]) if one wants to adapt to a continuum of smoothness parameters, as is usually the case in nonparametric statistics. The starting point of Szabó, van der Vaart and van Zanten is to transpose the L^∞-type self-similarity condition from [6] into their ℓ_2-risk setting:

$$\sum_{k=N}^{\rho N} \theta_k^2 \geq \varepsilon \|\theta\|_{S^\beta}^2 N^{-2\beta} \qquad \forall N \geq N_0 \text{ with "tolerance" factor } \varepsilon > 0, \tag{10}$$

whenever θ belongs to a Sobolev space S^β with norm

$$\|\theta\|_{S^\beta}^2 = \sum_k \theta_k^2 k^{2\beta}, \qquad S^\beta(B) = \{\theta : \|\theta\|_{S^\beta} \leq B\}.$$

Here $\rho > 2$, $N_0 \in \mathbb{N}$ are fixed constants; see equation (3.4) in [12]. Note that finiteness of the Sobolev norm implies

$$\sum_{k \geq N} \theta_k^2 \leq \|\theta\|_{S^\beta}^2 N^{-2\beta} \qquad \forall N \in \mathbb{N}, \tag{11}$$

and the idea behind (10) is hence that over repeated blocks $\{N, \dots, \rho N\}$ the signal θ indicates that it is actually exactly β-regular. A nice observation of Szabó, van der Vaart and van Zanten is that this condition can in fact be substituted by the slightly more general "polished tail" condition

$$\sum_{k=N}^{\rho N} \theta_k^2 \geq L_0^{-1} \sum_{k \geq N} \theta_k^2 \qquad \forall N \geq N_0 \text{ for some } L_0 > 0, \tag{12}$$

which effectively means that the blocks in (10) have, for every N large enough and up to a small constant L_0^{-1}, the same signal strength as the full tail series $\sum_{k \geq N} \theta_k^2$. This condition is conceptually somewhat cleaner than (10), as it does not require the identification of the unknown regularity parameter β, although it implicitly does so in the sense that (12) implies that (10) and (11) cannot hold for multiple values of β.

> The key issue I want to discuss here is in which sense exactly conditions like (10) or (12) are necessary for adaptive inference procedures to exist in the setting of the paper [12] under discussion.

Since Szabó, van der Vaart and van Zanten are considering ℓ_2-risk, the situation is qualitatively different from the L^∞-setting for which the lower bounds in [1] apply. First of all, as also noted by the authors, when adaptation is sought after for β contained in fixed smoothness windows $[\beta_0, 2\beta_0]$, a direct construction of an adaptive confidence set is possible without any restrictions on the parameter space. However, the constraint $\beta \in [\beta_0, 2\beta_0]$ is not satisfactory in the typical situations of nonparametric inference. Once relaxed, information-theoretic arguments imply that restrictions on the parameter space S^β become necessary (e.g., Theorems 1 or 4 in [2]). Employing conditions of the kind (10) or (12) to enforce such restrictions, one notices that these assumptions can be weakened quantitatively by increasing the windows $[N, \rho N]$ over which the lower bound of the signal is allowed to accrue. The question arises whether the window sizes $[N, \rho N]$ with $\rho > 2$ are *minimal* conditions for the existence of adaptive confidence sets or whether larger windows are admissible, pertaining to larger parameter spaces for which inference is possible. *For self-similar classes it is shown in* [9] *that condition* (10) *is not optimal*, *and that in turn* (12) *can also not be*.

Let us describe the results from [9] to understand in what sense weaker conditions are possible: let $N_0 \in \mathbb{N}$, $0 < b < B < \infty$. For $\varepsilon \in (0, 1]$ and $c_\beta = 16 \times 2^{2\beta+1}$, define the set

$$S_\varepsilon^\beta = \left\{\theta \in S^\beta(B) : \sum_{k=N^{(1-\varepsilon)}}^{N} \theta_k^2 \geq c_\beta \|\theta\|_{S^\beta}^2 N^{-2\beta} \ \forall N \geq N_0\right\}. \tag{13}$$

Again, as in (10), sufficiently large signal blocks have to appear repeatedly. But now these blocks are allowed to have increased *window-width* since

$$N^\varepsilon \gg \rho \qquad \text{as } N \to \infty,$$

and allow for an *asymptotically shrinking tolerance factor*

$$\varepsilon = N^{-2\varepsilon\beta} c_\beta \to 0 \qquad \text{as } N \to \infty$$

in the lower bound. *In particular*, (13) *only approximately identifies the smoothness of* θ *in the sense that it can be satisfied*, *unlike* (10) *or* (12), *for multiple values of* β *simultaneously*.

As shown in [9], signal strength conditions enforced through (13) allow for the construction of honest adaptive confidence ℓ_2-balls for signals

$$\theta \in \bigcup_{\beta_{\min} \leq \beta \leq \beta_{\max}} S_{\varepsilon(\beta)}^\beta, \qquad 0 < \beta_{\min} < \beta_{\max} < \infty,$$

under (effectively) the following conditions on ε:

$$\varepsilon(\beta) < \tfrac{1}{2} \qquad \forall \beta \in [\beta_{\min}, \beta_{\max}] \text{ is necessary,}$$

whereas

$$\varepsilon(\beta) < \frac{\beta}{2\beta + 1/2} \qquad \forall \beta \in [\beta_{\min}, \beta_{\max}] \text{ is sufficient.}$$

Note that $\beta_{\min} < \beta_{\max}$ are arbitrary, and hence the lower bound cannot be improved in general, since in the limit $\beta \to \infty$ we have $\beta/(2\beta + 1/2) \to 1/2$.

We conclude that requiring lower bounds in windows of size $[N, \rho N]$ as in (10), (12) is too strong a requirement for adaptive ℓ_2-confidence sets, and the results in the paper by Szabó, van der Vaart and van Zanten are suboptimal from an information-theoretic perspective. It would be interesting to know whether this suboptimality is an artefact of the proofs or of the particular Bayesian inference procedure used, although it may be difficult to answer this question.

Acknowledgment. I would like to thank Kolyan Ray for allowing me to reproduce Figure 1 from his paper [11].

REFERENCES

[1] BULL, A. D. (2012). Honest adaptive confidence bands and self-similar functions. *Electron. J. Stat.* **6** 1490–1516. MR2988456

[2] BULL, A. D. and NICKL, R. (2013). Adaptive confidence sets in L^2. *Probab. Theory Related Fields* **156** 889–919. MR3078289

[3] CASTILLO, I. and NICKL, R. (2013). Nonparametric Bernstein–von Mises theorems in Gaussian white noise. *Ann. Statist.* **41** 1999–2028. MR3127856

[4] CASTILLO, I. and NICKL, R. (2014). On the Bernstein–von Mises phenomenon for nonparametric Bayes procedures. *Ann. Statist.* **42** 1941–1969. MR3262473

[5] FREEDMAN, D. (1999). On the Bernstein–von Mises theorem with infinite-dimensional parameters. *Ann. Statist.* **27** 1119–1140. MR1740119

[6] GINÉ, E. and NICKL, R. (2010). Confidence bands in density estimation. *Ann. Statist.* **38** 1122–1170. MR2604707

[7] HOFFMANN, M. and NICKL, R. (2011). On adaptive inference and confidence bands. *Ann. Statist.* **39** 2383–2409. MR2906872

[8] LEAHU, H. (2011). On the Bernstein–von Mises phenomenon in the Gaussian white noise model. *Electron. J. Stat.* **5** 373–404. MR2802048

[9] NICKL, R. and SZABÓ, B. (2014). A sharp adaptive confidence ball for self-similar functions. Preprint. Available at arXiv:1406.3994.

[10] NICKL, R. and VAN DE GEER, S. (2013). Confidence sets in sparse regression. *Ann. Statist.* **41** 2852–2876. MR3161450

[11] RAY, K. (2014). Bernstein–von Mises theorems for adaptive Bayesian nonparametric procedures. Preprint. Available at arXiv:1407.3397.

[12] SZABÓ, B., VAN DER VAART, A. W. and VAN ZANTEN, H. (2015). Frequentist coverage of adaptive nonparametric Bayesian credible sets. *Ann. Statist.* **43** 1391–1428.

STATISTICAL LABORATORY
DEPARTMENT OF PURE MATHEMATICS
AND MATHEMATICAL STATISTICS
UNIVERSITY OF CAMBRIDGE
CB3 0WB CAMBRIDGE
UNITED KINGDOM
E-MAIL: r.nickl@statslab.cam.ac.uk