Submitted to the Annals of Statistics

DISCUSSION: "COVERAGE OF BAYESIAN CREDIBLE SETS"

By Subhashis Ghosal

North Carolina State University

First I like to congratulate the authors Botond Szabó, Aad van der Vaart and Harry van Zanten for a fine piece of work on the extremely important topic of frequentist coverage of adaptive nonparametric credible sets. Credible sets are used by Bayesians to quantify uncertainty of estimation, which is typically viewed as more informative than point estimation. Such sets are often easily constructed, for instance by sampling from the posterior, while confidence sets in the frequentist setting may need evaluating limiting distributions, or resampling, which needs additional justification. Bayesian uncertainty quantification in parametric problems from the frequentist view is justified through the Bernstein-von Mises theorem. In the recent years, such results have been also obtained for the parametric part in certain semiparametric models, guaranteeing coverage of Bayesian credible sets for it. However, as mentioned by the authors, inadequate coverage of nonparametric credible sets has been observed [Cox (1993), Freedman (1999)] in the white noise model, arguably the simplest nonparametric model. A clearer picture emerged after the work of Knapik, van der Vaart and van Zanten (2011) that undersmoothing priors can resolve the issue of coverage; see also Leahu (2011) and Castillo and Nickl (2013).

In the present paper, the authors address the issue of coverage of credible sets in a white noise model under the inverse problem setting, when the underlying smoothness (i.e. regularity) of the true parameter is not known, so a procedure must adapt to the smoothness. The authors follow an empirical Bayes approach where a key regularity parameter in the prior is estimated from its marginal likelihood function. As the authors mentioned, undersmoothing leads to inferior point estimation and is also difficult to implement when the smoothness of the parameter is not known. We shall see that the issue of coverage can also be addressed by two other alternative approaches.

Before entering a discussion on the contents of the paper, let us take another look at the coverage problem for Bayesian credible sets in an abstract setting. Suppose that we have a family of experiments based on observations $Y^{(n)}$ and indexed by a parameter $\theta \in \Theta$, some appropriate metric space. Let

ϵ_n be the minimax convergence rate for estimating θ. Let $\gamma_n \in [0,1]$ be a sequence which can be fixed or may tend to 0. For some $m_n \to \infty$, typically a slowly varying sequence, the goal is to find a subset $\mathcal{C}(Y^{(n)}) \subseteq \Theta$ such that uniformly on $\theta_0 \in \mathcal{B}$

(i) $\Pi(\theta \in \mathcal{C}(Y^{(n)})|Y^{(n)}) \geq 1-\gamma_n$,
(ii) $P_{\theta_0}^{(n)}(\theta_0 \in \mathcal{C}(Y^{(n)})) \to 1$,
(iii) $\operatorname{diam}(\mathcal{C}(Y^{(n)})) = O_{P_{\theta_0}^{(n)}}(m_n\epsilon_n)$,

where $\mathcal{B}$ varies over a class of compact balls in Θ.

In the formulation, credibility may increase with the sample size. We find it natural that when the information content is increasing, a researcher should quantify uncertainty with more and more confidence, instead of staying at a fixed level, just like one seeks for more precise point estimators or tests. If $\gamma_n \to 0$ it can be seen that the problems of mismatch of credibility and coverage pointed out in Cox (1993) and Freedman (1999) goes away. Thus although the uncertainty quantification of a Bayesian and a frequentist may not match at finite levels, they do match at the infinitesimal level. For finer matching, one may also like to impose some requirement on how fast $P_{\theta_0}^{(n)}(\theta_0 \in \mathcal{C}(Y^{(n)}))$ should approach 1, but we shall forgo the issue in this discussion. Another approach is to obtain a $(1-\gamma_n)$-credible ball around the posterior mean typically with fixed γ_n and inflate the region by a factor m_n, to be called the inflation factor, to ensure adequate frequentist coverage. The size of the original credible region is typically of the order the minimax convergence rate ϵ_n so that the third condition will be met. The factor m_n can be considered as a reasonable price for the increased level of coverage. Typically, the resulting extra cost m_n is low, for instance, in an asymptotic normality setting, while adopting $(1-\gamma_n)$-credible sets with $\gamma_n \to 0$, the additional cost is $m_n = o(\sqrt{\log(1/\gamma_n)})$. In the setting we shall discuss, the inflation factor may be taken as a sufficiently large constant. The supremum over compact sets in the formulation imposes honesty of the coverage.

As mentioned by the authors, fully adaptive honest nonparametric confidence regions are not possible by any means, so in the adaptive context Θ will be replaced by an appropriate subset of the parameter space, such as the set of self-similar sequences or polished tailed sequences in the context of the paper. The concept of polished tail is pretty elegant as it blends nicely in the adaptive setting without any direct reference to the smoothness of the parameter.

The main result proved in the paper, namely, honest coverage of adaptive posterior credible regions for $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots)$ in the model $Y_i = \kappa_i\theta_i + n^{-1/2}\varepsilon_i$, where $\boldsymbol{\theta} \in \ell_2$ and $\varepsilon_i \overset{\text{iid}}{\sim} N(0,1)$, for all polished tail sequences is certainly exciting. In terms of the equivalent (and perhaps more directly relevant) white noise inverse problem model $dY(t) = Kf(t)dt + n^{-1/2}dW(t)$, this translates into honest coverage of credible regions for f through Parseval's identity, where the distance on f is measured in terms of the L_2-distance. However, L_2-regions for functions do not look like bands, and may be a little harder to visualize. This aspect may be relevant for covering a true function that has a bump like the one given by equation (4.1) in the paper under discussion, since L_2-closeness does not even imply pointwise closeness, let alone uniform closeness. Regions on function spaces given by L_∞-neighborhoods are easier to visualize and interpret. Moreover such uniform closeness has other implications. For instance, if derivatives of the functions in a region are uniformly ϵ_n-close to the derivative of the true function and the derivative of true function has a well separated mode, then mode of a function in that region is $O(\epsilon_n)$-close to the true mode. The observation can be used to induce honest confidence regions for the mode from those for the derivative function under the L_∞-distance.

Study of coverage of L_∞-regions with chosen credibility needs studying posterior contraction rates under the L_∞-norm, which is easier if conjugacy is present, like in the white noise model or nonparametric regression using a random series with normal coefficients. Below we shall argue that in the white noise model credible regions for L_∞-norm can also be characterized and computed relatively easily, and their coverage can be shown to be adequate. Interestingly we can use fixed level of credibility (any value higher than 1/2 works) and the inflation factor can be taken to be a constant. We shall follow techniques similar to those used in Yoo and Ghosal (2014), who considered the problem of multivariate nonparametric regression using a random series of tensor product B-splines in the known smoothness setting. In a sense the present treatment of the simpler white noise model will be easier, but there are certain differences as well, particularly since the number of basis elements used in constructing the prior is infinite in the present case, unlike the case treated by Yoo and Ghosal (2014). For the sake of simplicity of the discussion we restrict to the direct problem, i.e. $\kappa_i \equiv 1$ and consider a Fourier basis $\phi_1(x) = 1$, $\phi_{2i}(x) = \sqrt{2}\cos(2\pi i x)$, $\phi_{2i+1}(x) = \sqrt{2}\sin(2\pi i x)$, $i = 1, 2, \ldots$. Let the true function be denoted by f_0 and the true sequence by $\boldsymbol{\theta}_0 = (\theta_{01}, \theta_{02}, \ldots)$. Since we intend to study L_∞-contraction rate and coverage of L_∞-regions, we need to assume that the true function f_0 belongs to a Hölder class, or stated in terms of coefficients $\sum_{i=1}^{\infty} i^\alpha |\theta_{0i}| < \infty$, which is stronger than the analogous Sobolev condition $\sum_{i=1}^{\infty} i^{2\alpha}\theta_{0i}^2 < \infty$. The logarithmic factor we obtain in the rate is not optimal — it is off by the factor $(\log n)^{1/2(2\alpha+1)}$. Using a more refined analysis or perhaps using a different basis like B-splines or wavelets, optimal logarithmic factor may be obtained as in Yoo and Ghosal (2014) or Giné and Nickl (2012). Also because we use a Fourier basis, we assume that the true function is periodic, but this does not dampen the essential spirit of the argument.

Consider the white noise model $dY(t) = f(t)dt + n^{-1/2}dW(t)$ and its equivalent normal sequence model $Y_i = \theta_i + n^{-1/2}\varepsilon_i$, where $Y_i = \int \phi_i(x)dY(x)$, $\theta_i = \int \phi_i(x)f(x)dx$ and $\varepsilon_i = \int \phi_i(x)dW(x)$, $i = 1, 2, \ldots$. Let the prior Π be defined by $\theta_i \stackrel{\text{ind}}{\sim} N(0, i^{-2\alpha+1})$. Let $\hat{f} = \mathrm{E}(f|D_n)$, where D_n stands for the data. Note that $\hat{f}(x) = \sum_{i=1}^{\infty} \hat{\theta}_i \phi_i(x)$, where $\hat{\theta}_i = \mathrm{E}(\theta_i|Y_i) = nY_i/(i^{2\alpha+1}+n)$, and $\mathrm{var}(\theta_i|Y_i) = (i^{2\alpha+1}+n)^{-1}$. Let $B(\alpha, R) = \{f : \sum_{i=1}^{\infty} i^{\alpha}|\theta_i| \le R\}$. Below we shall write "$\lesssim$" for inequality up to a constant and "$\asymp$" for equality in order.

THEOREM 1. *For any* $M_n \to \infty$, $\mathrm{E}_{f_0}\Pi_\alpha(f : \|f - f_0\|_\infty > M_n\epsilon_n|D_n) \to 0$ *uniformly for all* $f_0 \in B(\alpha, R)$, *where* $\epsilon_n = n^{-\alpha/(2\alpha+1)}\sqrt{\log n}$ *and for a sufficiently large constant* $M > 0$, $P_{f_0}\{\|f_0 - \hat{f}\|_\infty \le Mh_n\} \to 1$ *where* h_n *is determined by* $\Pi_\alpha(f : \|f - \hat{f}\|_\infty \le h_n|D_n) = 1 - \gamma$, $\gamma \ge 1/2$ *is a predetermined constant. Moreover* $h_n \lesssim \epsilon_n$.

The theorem implies that the $(1-\gamma)$-credible region for L_∞-distance around the posterior mean for any $\gamma \le 1/2$ inflated by a sufficiently large factor M has asymptotic coverage 1 and its size h_n is not larger than the posterior contraction rate, which is nearly optimal. It is interesting to note that h_n is actually deterministic since the posterior distribution of $f - \hat{f}$ is free of the observations. Analytical computation of h_n may be difficult, but can be easily determined by simulations.

PROOF OF THE THEOREM. We have $f(x) = \sum_{i=1}^{\infty} \theta_i\phi_i(x)$. Thus given D_n, $Z = f - \hat{f}$ is a mean-zero Gaussian process with covariance kernel

$$\mathrm{cov}\left(\sum_{i=1}^{\infty} \theta_i\phi_i(s), \sum_{i=1}^{\infty} \theta_i\phi_i(t)\big|D_n\right) = \sum_{i=1}^{\infty}(i^{2\alpha+1}+n)^{-1}\phi_i(s)\phi_i(t)$$

and

$$\begin{aligned}
\mathrm{E}(|Z(s) - Z(t)|^2|D_n) &= \sum_{i=1}^{\infty} \mathrm{var}(\theta_i|D_n)|\phi_i(s) - \phi_i(t)|^2 \\
&\lesssim \sum_{i=1}^{\infty}(i^{2\alpha+1}+n)^{-1}i^2|s-t|^2 \\
&\lesssim n^{2(\alpha-1)/(2\alpha+1)}|s-t|^2
\end{aligned}$$

by standard estimates and the fact $|\phi_i(s) - \phi_i(t)| \le 2\sqrt{2}\pi i|s-t|$, a consequence of the mean value theorem and the boundedness of trigonometric functions. Using a uniform grid with mesh-width $\delta_n \asymp n^{-p}$ for $p > 0$ sufficiently large and a chaining argument for Gaussian processes with values of Z at the chosen grid points, Lemma 2.2.2 and Corollary 2.2.8 of van der Vaart and Wellner (1996) gives the estimate $\mathrm{E}\|Z\|_\infty \le \sqrt{\mathrm{E}\|Z\|_\infty^2} \lesssim n^{-\alpha/(2\alpha+1)}\sqrt{\log n}$.

Let $V(x) = \hat{f}(x) - \mathrm{E}_{f_0}\hat{f}(x) = \sum_{i=1}^\infty \sqrt{n}\varepsilon_i\phi_i(x)/(i^{2\alpha+1}+n)$. Then V is a mean-zero Gaussian process with covariance kernel $\sum_{i=1}^\infty n(i^{2\alpha+1}+n)^{-2}\phi_i(s)\phi_i(t)$ and

$$\mathrm{E}|V(s)-V(t)|^2 = \sum_{i=1}^\infty n(i^{2\alpha+1}+n)^{-2}|\phi_i(s)-\phi_i(t)|^2.$$

Arguing as before, it follows that $\mathrm{E}_{f_0}\|V\|_\infty \lesssim n^{-\alpha/(2\alpha+1)}\sqrt{\log n}$.

Now using the uniform boundedness of the basis functions and $\sum_{i=1}^\infty i^\alpha|\theta_{0i}| \le R$, uniformly for $f_0 \in B(\alpha, R)$, we have for any k, $\sum_{i>k}|\theta_{0i}| \le Rk^{-\alpha}$. Therefore

$$\begin{aligned}\|\mathrm{E}_{f_0}\hat{f} - f_0\|_\infty &= \left\|\sum_{i=1}^\infty \left(\frac{n}{i^{2\alpha+1}+n} - 1\right)\theta_{0i}\phi_i\right\|_\infty \\ &\le \sqrt{2}R\left(\frac{k^{\alpha+1}}{n} + k^{-\alpha}\right) \lesssim n^{-\alpha/(2\alpha+1)}\end{aligned}$$

by choosing $k = k_\alpha \asymp n^{1/(2\alpha+1)}$.

Combining the three pieces, it follows using Chebyshev's inequality that the posterior contraction rate under the L_∞-distance is ϵ_n.

Now we find a lower bound for the size of the credible region. By definition h_n, the $(1-\gamma)$-quantile of the distribution of $\|Z\|_\infty$ for the mean-zero Gaussian process Z with covariance kernel $\sum_{i=1}^\infty (i^{2\alpha+1}+n)^{-1}\phi_i(s)\phi_i(t)$, is at least as large as the median of the distribution of $\|Z\|_\infty$. Now $\sigma_Z^2 = \sup \mathrm{E}|Z(t)|^2$ is easily seen to be $O(n^{-2\alpha/(2\alpha+1)})$. Since $\mathrm{E}\|Z\|_\infty^2 \ge \sigma_Z^2$, standard facts about Gaussian processes imply that $\mathrm{E}\|Z\|_\infty$ and the median of $\|Z\|_\infty$ are of the same order (cf. Ledoux and Talagrand (1991), pages 52 and 54). Hence to find a lower bound for h_n, it suffices to lower bound $\mathrm{E}\|Z\|_\infty$. We shall show that the order of the lower bound is $n^{-\alpha/(2\alpha+1)}\sqrt{\log n}$.

To this end, we observe that $\|Z\|_\infty \ge \max\{Z(j/k_\alpha) : j = 1, \ldots, k_\alpha\}$, and

$$\mathrm{E}|Z(j/k_\alpha) - Z(l/k_\alpha)|^2 = \sum_{i=1}^\infty (i^{2\alpha+1}+n)^{-1}|\phi_i(j/k_\alpha) - \phi_i(l/k_\alpha)|^2.$$

With a sufficiently small fixed $\epsilon > 0$, there exists a $\delta > 0$ such that $|\sin s - \sin t| > \epsilon$ if $|s-t| > \delta$ and $|s+t-\pi| > \delta$, and a similar assertion holds for the cosine function. Therefore it is observed that for $j, l = 1, \dots, k_\alpha$, $j \neq l$, $\phi_i(j/k_\alpha)$ and $\phi_i(l/k_\alpha)$ differ by at least a fixed positive number for a positive fraction of $i \in \{2, \dots, k_\alpha\}$. From this we obtain that there exists $c > 0$ such that

$$\mathrm{E}|Z(j/k_\alpha) - Z(l/k_\alpha)|^2 \geq cn^{-2\alpha/(2\alpha+1)}.$$

Let $U_j = \sqrt{2}c^{-1/2}n^{\alpha/(2\alpha+1)}Z(j/k_\alpha)$, $j = 1, \dots, k_\alpha$, so that $\mathrm{E}(U_j - U_l)^2 \geq \mathrm{E}(V_j - V_l)^2$, where $V_1, \dots, V_{k_\alpha} \overset{\text{iid}}{\sim} N(0,1)$. Hence by Slepian's inequality (cf. Corollary 3.14 of Ledoux and Talagrand (1991)) and equation (3.14) of Ledoux and Talagrand (1991), we obtain

$$\mathrm{E}(\max_j U_j) \geq \mathrm{E}(\max_j V_j) \gtrsim \sqrt{\log k_\alpha} \asymp \sqrt{\log n},$$

which upon rescaling gives $\mathrm{E}\|Z\|_\infty \geq \mathrm{E}(\max Z(j/k_\alpha)) \gtrsim n^{-\alpha/(2\alpha+1)}\sqrt{\log n}$.

Now turning to coverage, the lack of coverage of the credible set inflated by a sufficiently large constant M is given by

$$\begin{aligned} P_{f_0}\{\|f_0 - \hat{f}\|_\infty > Mh_n\} &\leq \mathrm{P}\{\|V\| > M'\epsilon_n - \|\mathrm{E}\hat{f} - f_0\|_\infty\} \\ &\leq 2e^{-C\log n} \to 0 \end{aligned}$$

by virtue of Borell's inequality (cf. second assertion of Proposition A.2.1 of van der Vaart and Wellner (1996)) since $\sup_t \mathrm{var}(Z(t)) \lesssim n^{-2\alpha/(2\alpha+1)}$ and $\epsilon_n \asymp n^{-\alpha/(2\alpha+1)}\sqrt{\log n}$, uniformly for $f_0 \in B(\alpha, R)$, where M' and C are positive constants.

Finally we estimate the size of the inflated credible region. For that we need to find an upper bound for the $(1-\gamma)$-quantile of the distribution of $\|Z\|_\infty$ given D_n. By Borell's inequality (cf. third assertion of Proposition A.2.1 of van der Vaart and Wellner (1996)), it is clear that the $(1-\gamma)$-quantile is bounded by $\sqrt{8\mathrm{E}\|Z\|_\infty^2 \log(2/\gamma)}$, which is of the order $n^{-\alpha/(2\alpha+1)}\sqrt{\log n}$. □

We also wish to study the coverage problem for L_∞-credible regions when the regularity α is not known. Consider the empirical Bayes device of the paper under discussion, and assume that the true sequence has polished tail. The heuristic arguments given below seem to indicate that the credible region constructed by plugging-in the empirical Bayes estimate of α should have adequate coverage.

Because we deal with various values of α simultaneously, let us include α in the notations Π_α for the prior, Z_α and V_α for the Gaussian processes introduced in the proof, and $\epsilon_{n,\alpha} = n^{-\alpha/(2\alpha+1)}\sqrt{\log n}$ for the sup-norm posterior contraction rate. We observe that $\epsilon_{n,\alpha}$ is decreasing in α. By Theorem 5.1 of the paper under discussion, it follows that the empirical Bayes estimate $\hat{\alpha}$ of α lies, with high probability, between two deterministic bounds $\underline{\alpha}$ and $\overline{\alpha}$, and that $\epsilon_{n,\underline{\alpha}} \asymp \epsilon_{n,\overline{\alpha}}$.

In the proof of the result on coverage of credible region, one needs to lower bound the radius of the credible ball around the estimate and show that its order is at least as large as the convergence rate of the point estimator given by the center of the credible region. When $\hat{\alpha}$ is plugged-in, the radius of the credible region is of the order of the expected value of the supremum of the Gaussian process $Z_{\hat{\alpha}}$. The randomness of this process comes from posterior variation conditioned on the sample, and hence $\hat{\alpha}$ can be considered as a constant. Therefore, as argued in the proof of the theorem, radius of the credible region is of the order $\epsilon_{n,\hat{\alpha}} \asymp \epsilon_{n,\underline{\alpha}} \asymp \epsilon_{n,\overline{\alpha}}$.

The sampling error of the Bayes estimator $\hat{f}_\alpha$ using Π_α has two parts — variability around its expectation Z_α and its bias. Now for any $t, s \in [0,1]$, $\mathrm{E}|Z_\alpha(t) - Z_\alpha(s)|^2$ is decreasing in α, so by Slepian's inequality

$$\sup\{\mathrm{E}\|Z_\alpha\|_\infty : \underline{\alpha} \le \alpha \le \overline{\alpha}\} = \mathrm{E}\|Z_{\underline{\alpha}}\|_\infty \asymp \epsilon_{n,\underline{\alpha}} \asymp \epsilon_{n,\overline{\alpha}},$$

and fixed quantiles of $\|Z_\alpha\|_\infty$ also have the same order as the expectation of $\|Z_\alpha\|_\infty$ by Borell's inequality. On the other hand, the bias of $\hat{f}_\alpha$ increases with α, and hence its maximum is attained at $\overline{\alpha}$ for $\underline{\alpha} \le \alpha \le \overline{\alpha}$. Note that if $\overline{\alpha}$ underestimates the true α, then the order of the bias is $\epsilon_{n,\overline{\alpha}} \asymp \epsilon_{n,\underline{\alpha}}$, and so for every α lying in the range $[\underline{\alpha}, \overline{\alpha}]$, the posterior contraction rate would be the same. Lemma 3.11 seems to indicate that this may be the case. This will ensure adequate coverage of the empirical Bayes credible set.

Another issue that might of interest for future investigation is the handling of unknown variance. In the non-adaptive setting, both empirical and hierarchical Bayes approach can fruitfully address the issue of unknown variance as demonstrated by Yoo and Ghosal (2014) for nonparametric regression. In the adaptive setting, this is somewhat unclear, as the empirical Bayes estimate of smoothness and variance will depend on each other.

It is also natural to ask if the hierarchical Bayes credible sets can also have adequate coverage in the adaptive setting. This may not have an affirmative answer, as indicated by Rivoirard and Rousseau (2012).

Finally, for other curve estimation problems like density estimation or nonparametric regression, what should be a proper analog of conditions like self-similarity of polished tail, and how may that help in establishing coverage? The nonparametric regression problem may be more tractable than the density estimation, since for the former a basis expansion approach reduces the function of interest to a sequence of real-valued parameters which are typically given normal priors as well and conjugacy holds in the model. Usually it is more convenient to use a truncated series expansion, but then the sequence of parameters form a triangular array. It seems that the main challenge will be to identify a proper analog of a condition on the tail of the sequence in such a setting.

REFERENCES

[1] Castillo, I. and Nickl, R. (2013). Nonparametric Bernstein-von Mises theorems in Gaussian white noise. *Ann. Statist.* **41** 1999–2028.

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

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

[4] Giné, E. and Nickl, R. (2011). On the Bernstein-von Mises theorem with infinite dimensional parameters. *Ann. Statist.* **39** 2795–3443.

[5] Knapik, B. T., van der Vaart, A. W. and van Zanten, J. H. (2011). Bayesian inverse problems with Gaussian priors. *Ann. Statist.* **39** 2626–2657.

[6] Leahu, H. (2011). On the Bernstein-von Mises phenomenon in the Gaussian white noise model. *Electron. J. Statist.* **5** 373–404.

[7] Ledoux, M. and Talagrand, M. (1991). *Probability in Banach Spaces.* Springer-Verlag, Berlin Heidelberg,

[8] Rivoirard, V. and Rousseau, J. (2012). Bernstein-von Mises theorem for linear functional of the density. *Ann. Statist.* **40** 1489–1523.

[9] van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Process With Applications to Statistics.* Springer-Verlag, New York.

[10] Yoo, W. W. and Ghosal, S. (2014). Supremum norm posterior contraction and credible sets for nonparametric multivariate regression. arXiv:1411.6716.

Department of Statistics
North Carolina State University
4276 SAS Hall, 2311 Stinson Drive
Raleigh, NC 27695-8203, USA
E-mail: sghosal@ncsu.edu