The Annals of Statistics
2015, Vol. 43, No. 4, 1444–1447
DOI: 10.1214/15-AOS1270C
Main article DOI: 10.1214/14-AOS1270

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

BY JUDITH ROUSSEAU

Université Paris-Dauphine and CREST-ENSAE

1. Introduction. I would like first to thank the editors for giving me the opportunity to discuss the paper *Frequentist coverage of adaptive Bayesian credible sets* by B. Szabo, A. van der Vaart and Harry van Zanten. This is a very significant contribution to the literature on nonparametric credible sets. Before discussing some specific phenomena enlightened by the paper on frequentist coverage of adaptive credible sets, I would like to emphasize why I believe this to be a crucial problem, in particular, in large or infinite-dimensional models. Over the last ten years or so, there has been a growing literature on frequentist properties of the posterior distribution in large or infinite-dimensional models. The results obtained concern mainly posterior concentration rates, initiated by the seminal paper of Ghosal, Ghosh and van der Vaart [2]. These results have shown that Bayesian nonparametric approaches often lead to estimates having very good frequentist properties such as adative and minimax (or near minimax with possibly a $\log n$ penalty term) convergence rates under standard loss functions. However, one of the interesting aspects of Bayesian methods is that they provide much more than point estimates via the posterior distribution, in particular, various sorts of measures of uncertainty can be derived. An important question is then, how can we understand these measures of uncertainty? This is all the more important when the models are complex or large, as it becomes impossible to (1) elicit fully a subjective prior and (2) to assess perfectly the influence of the prior. Hence, looking at the frequentist properties of measures of uncertainty is a way to answer—at least partially—these questions.

In [3] it is observed that when the posterior distribution has concentration rate ε_n, under some loss function $\ell(\cdot,\cdot)$, and if there exists an estimate, say, the posterior mean, whose convergence rate is also ε_n under the loss $\ell(\cdot,\cdot)$, then if $r_n(\gamma)$ is defined as the posterior $1-\gamma$ quantile of $\ell(\theta,\hat{\theta})$, we have

$$1-\gamma = \int P_\theta\big(\ell(\theta,\hat{\theta}) \le r_n(\gamma)\big)\, d\pi(\theta), \qquad E_\theta\big(|r_n(\gamma)|\big) = 0(\varepsilon_n), \tag{1.1}$$

where P_θ and E_θ denote, respectively, the probability and expectation under the sampling distribution associated to the parameter θ. This means that on average the credible region $\{\theta; \ell(\theta,\hat{\theta}) \le r_n(\gamma)\}$ has the correct frequentist coverage and

Received January 2015.
Key words and phrases. Empirical Bayes, Bayesian credible sets, frequentist coverage.

if ε_n is the minimax rate, it has also optimal size, asymptotically. The question is then, can we describe precisely the set of parameters θ such that

$$P_\theta\big(\ell(\theta,\hat{\theta}) \le r_n(\gamma)\big) \ge 1-\gamma \tag{1.2}$$

or is at least large enough? In their paper, the authors answer a similar question by allowing $r_n(\gamma)$ to be inflated by a constant L, possibly large, in the special case of a Gaussian white noise model with an adaptive empirical Bayes Gaussian prior.

2. On polished tail parameters. A key notion in this paper is that of polished tail parameters $\theta \in \ell_2$,

$$\sum_{j=N}^{\infty} \theta_j^2 \le L_0 \sum_{j=N}^{\rho N} \theta_j^2 \qquad \forall N > N_0,$$

where N_0, L_0 and ρ are fixed. This definition is not intrinsic to the function $f = \sum_{j=1}^{\infty} \theta_j \phi_j$ which is to be estimated since f can lead to polished tail parameter θ when represented in some orthonormal basis $(\phi_j)_j$ and nonpolished tail parameter η in some other orthonormal basis. Although this might seem disturbing at first, I think this is inherent to the Bayesian approach. Let π be a prior on some functional set containing a collection of Hölder balls or Sobolev balls which leads to adaptive minimax posterior concentration rates over this collection. Then, since there exists no adaptive concentration rate in L_2 and probably in many other metrics on f, the set of *good parameters* (for which credible regions have good frequentist coverage) is necessarily a subset of Θ. The question is which subset? Bull and Nickl [1] (among others) construct procedures such that the subset of *badly behaved* parameters (those that either do not lead to good coverage or do not lead to optimal size) is rendered as small as possible. But these procedures require to artificially withdraw the badly behaved points from the confidence regions, which is not entirely satisfying. In their case, however, the definition of the well-behaved parameters does not depend on the representation of the function f in some particular basis. In the paper of Szabo et al. the advantage is that the credible region is constructed using standard methods and it has good frequentist properties for a subset of parameters. Hence, a natural question arises: Is it possible to construct a prior whose set of well-behaved parameters is similar to that of Bull and Nickl [1], without having to modify the credible region by taking out the badly behaved parameters?

3. On the generalization of the results. More importantly, I think that even though the results are presented in a very specific context, they pave the way for controlling frequentist coverage of posterior credible regions. Indeed, consider an (inflated) credible region in the form

$$C_\alpha^\pi = \big\{\ell(\theta,\hat{\theta}) \le L r_n(\gamma)\big\},$$

where $r_n(\gamma)$ is the posterior $1-\gamma$ quantile of $\ell(\theta,\hat{\theta})$ and $\hat{\theta}$ is some minimax (adaptive) estimator with rate $\varepsilon_n(\theta)$ (the dependence on θ is here to emphasize the adaptation property). Under the condition of posterior concentration rate $\varepsilon_n(\theta)$ at θ,

$$E_\theta\big[r_n(\gamma)\big] \lesssim \varepsilon_n;$$

see [3] for details on this result. So the only thing that remains to be verified is that

$$\liminf_n \inf_{\Theta^o} P_\theta\big[\theta \in C_\alpha^\pi\big] \geq 1-\gamma \tag{3.1}$$

for some well-identified subset Θ^o of Θ. If the posterior distribution satisfies

$$\Pi\big(\theta;\, \ell(\theta,\hat{\theta}) > \delta\varepsilon_n(\theta)|X\big) \geq \alpha + o_{P_\theta}(1) \tag{3.2}$$

uniformly on Θ^o, then $P_\theta(r_n(\gamma) > \delta\varepsilon_n(\theta)) = 1 + o(1)$ uniformly on Θ^o and

$$P_\theta\big(\theta \in C_\alpha^\pi\big) \geq P_\theta\big(\ell(\theta,\hat{\theta}) \leq L\delta\varepsilon_n(\theta)\big) + o(1),$$

and we can choose $L = 1/\delta$ to ensure (3.1). So the main difficulty is to verify (3.2). When $\hat{\theta}$ is the posterior mean and $\ell(\cdot,\cdot)$ is the L_2 loss, as in the present paper, this boils down to bound from below the trace of the posterior variance. This typically requires that the posterior distribution asymptotically lives in a space that has an effective dimension large enough so that the bias of $\hat{\theta}$ is of the same order as its variance. The polished tail condition ensures that when the prior is based on a sequence parameter $\theta \in \ell_2$.

Obviously, this is a rough description of the underlying mechanisms and this does not take into account some more subtle aspects of the paper. For instance, the maximum marginal likelihood empirical Bayes approach is used to simplify the computations since the empirical Bayes distribution remains Gaussian, while leading to an adaptive posterior distribution. From the above comments it seems that hierarchical posterior distribution could be treated using similar ideas, though less directly. However, how influential are some of the specific aspects of this empirical Bayes procedure? If instead of maximizing the marginal likelihood in α, the posterior had been computed by considering a family of priors in the form

$$\theta_i \overset{\text{ind}}{\sim} \mathcal{N}\big(0, \tau i^{-1-2\alpha}\big)$$

and either consider a hierarchical procedure with a prior on τ or a empirical Bayes procedure maximizing the marginal likelihood in τ, then adaptive posterior concentration rates would be achieved on the range $\beta \in (0, \alpha + 1/2)$ if β represents the Sobolev smoothness of the true parameter. Similar results should be obtained in this case on the range $\beta \in (0, \alpha + 1/2)$. Now, if instead the prior model had the form f, conditional on τ is a Gaussian process with kernel

$$K_\tau(x, y) = e^{-\tau^2(x-y)^2}$$

and τ follows a Gamma random variable as in [4]. Then the posterior has an adaptive concentration rate over collections of Hölder balls with smoothness β, with $\beta \in (0, +\infty)$. How does it impact the behavior of credible regions?

4. How honest should a confidence region be? As I said in Section 1, the questions answered by the authors in this paper are important questions, as they help to understand some subtle effects of the prior in large dimensional models. However, as the nonexistence of adaptive confidence regions over a wide collection of Sobolev or Hölder classes of functions show, the *full* minimax paradigm (i.e., having a uniform lower bound on the confidence and an adaptive minimax upper bound on the size of the confidence region) has its limits. One might wonder what is the most important? Weakening the requirement on the confidence or on the size of the credible regions or considering smaller classes of functions? Somehow the adaptive Bayesian approach naturally adapts on the size while losing slightly on the confidence properties of the credible regions, as shown by (1.1). Confidence regions constructed in the frequentist literature are typically honest, however, their sizes are not uniformly optimal. Using this starting point as a construction of honest with optimal size confidence regions over smaller functional classes requires withdrawing from these regions badly behaved functions. This leads to a somewhat artificial construction. It seems thus better to be slightly dishonest and start with confidence regions that have optimal size and to understand over which subclasses of functions they are honest confidence regions. Obviously, this construction need not be necessarily Bayesian; however, I believe that the Bayesian methodology naturally leads to such a construction.

REFERENCES

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

[2] GHOSAL, S., GHOSH, J. K. and VAN DER VAART, A. W. (2000). Convergence rates of posterior distributions. *Ann. Statist.* **28** 500–531. MR1790007

[3] HOFFMANN, M., ROUSSEAU, J. and HIEBER, J. S. (2014). On adaptive posterior concentration rate. Technical report.

[4] VAN DER VAART, A. W. and VAN ZANTEN, J. H. (2009). Adaptive Bayesian estimation using a Gaussian random field with inverse gamma bandwidth. *Ann. Statist.* **37** 2655–2675. MR2541442

UNIVERSITÉ PARIS-DAUPHINE AND CREST-ENSAE
PLACE DU MARÉCHAL DE LATTRE DE TASSIGNY
75016 PARIS
FRANCE
E-MAIL: rousseau@ceremade.dauphine.fr