The Annals of Statistics
2015, Vol. 43, No. 4, 1437–1443
DOI: 10.1214/15-AOS1270B
Main article DOI: 10.1214/14-AOS1270

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

BY ISMAËL CASTILLO

Universités Paris VI & VII

First I would like to congratulate the three authors for a very nice paper. During a visit to Eindhoven in 2010, Botond Szabó and Harry van Zanten mentioned the first steps of this work to me, which concerned the understanding of certain empirical Bayes procedures in the white noise L^2-setting. Since then, together with Aad van der Vaart, they have broadened their original goals and have produced an impressive and very interesting series of papers on the subject. The present paper is indeed one aspect of a larger body of work, and we will mention a few connections with these related papers below.

The authors start from the signal in white noise model, that after projection in L^2 onto an appropriate basis, typically related to the SVD of the operator K of the inverse problem, is translated into a sequence formulation. They choose a prior distribution that makes coordinates independent:

$$\Pi_\alpha \sim \bigotimes_{i\geq 1} N\bigl(0, i^{-1-2\alpha}\bigr). \tag{1}$$

If the true parameter belongs to a regularity space defined from a decay of coefficients in the previous basis, the authors prove that certain credible sets constructed from the posterior distribution coupled with a (marginal-likelihood) empirical Bayes (EB) procedure for α achieve excellent performance: they are honest confidence sets with adaptive, optimal asymptotic diameter if one restricts to certain classes of "self-similar"-type true parameters. These are the first results of this type in Bayesian nonparametrics.

We organize this discussion around two main themes:

1. Priors for Bayesian credible sets.
2. Bayesian credible regions and simulations.

1. Priors for Bayesian credible sets. Several aspects of the prior scheme (1) are investigated by the authors in [10], [9] together with Bartek Knapik and in [14]. In [10], a fixed regularity parameter α is considered; in [9], adaptative contraction rates are derived. In [14], the prior (1) is used for fixed α and the use of a different empirical Bayes scheme is advocated.

Received January 2015.

RELATED PRIORS. Staying with priors defined on the SVD of K, some other adaptation schemes have been considered recently. One is (see [13])

$$\Pi_\tau \sim \bigotimes_{i\geq 1} N\big(0, \tau^2 i^{-1-2\alpha}\big), \qquad \tau > 0, \tag{2}$$

and adaptation is made by empirical Bayes or full Bayes on τ.

Another prior is obtained by setting, for a sequence $\{\lambda_i\}_{i\geq 1}$ of positive nondecreasing real numbers,

$$\Pi_t \sim \bigotimes_{i\geq 1} N\big(0, e^{-\lambda_i t}\big), \qquad t > 0. \tag{3}$$

In the case where K is the identity and, for example, $\lambda_i = i^2$, this falls into the framework considered in [2], where a full Bayes method is considered by putting a well-chosen hyperprior on t.

A natural question is whether the same construction as in the paper with a slightly blown up L^2-ball and regularity estimated by empirical or full Bayes would work the same for the priors (2) or (3), with self-similarity constraints expressed in a similar way. One can conjecture that the answer is yes and that one may study the empirical Bayes procedure from the explicit form of the marginal likelihood.

RELATED PRIORS AND SHARP RATES. Rates of convergence for Bayes procedures are sometimes shown to be optimal up to a slowly varying factor in n, for instance, logarithmic. In some cases it is not so clear whether such a logarithmic term should be present in the rate or not. The present work points to interesting questions with this respect, with connections to the related prior schemes (2)–(3).

For prior (2), it is shown in [13] that the minimax rate $n^{-\beta/(1+2\beta)}$ in L^2 over hyperrectangles is achieved by the marginal-likelihood-empirical Bayes procedure. This comes, however, to a cost: one should assume that the true regularity β of the signal satisfies $\beta < 1/2 + \alpha$, for α the regularity parameter in (2), otherwise the (uniform) rate can be shown to be suboptimal.

For prior (3), we obtained in [2] the rate $(\log n/n)^{\beta/(1+2\beta)}$ in L^2 over a class containing hyperrectangles and for which the minimax rate is $n^{-\beta/(1+2\beta)}$, so *without* the log-term, thus showing the unavoidable loss of a logarithmic factor when using prior (3).

In [9], the authors obtain an upper-bound rate for prior (1) in L^2 that contains a logarithmic factor. However, Proposition 3.8 of the present paper shows that the radius of the credible set is proportional to $n^{-\beta/(1+2\beta)}$, while Theorem 3.6 implies coverage of the credible set for polished tail parameters. Combining these results, one deduces that the posterior mean $\hat{\theta}_{n,\hat{\alpha}_n}$ verifies $\|\hat{\theta}_{n,\hat{\alpha}_n} - \theta_0\|_2 = O_P(n^{-\beta/(1+2\beta)})$. This presumably implies that the posterior itself converges at the minimax rate, without extra log-terms, if the true θ_0 has polished tails. One may conjecture that this is also true without the polished tail assumption. If so, it would be interesting to better understand what makes that priors (1)–(3) behave differently.

DIFFERENT PRIORS AND CONDITIONS. The prior scheme (1) is, by definition, somewhat tied to the SVD of K. As this type of basis may not be well-localized, this may cause some difficulties if the goal is a result in terms of a different loss function than L^2.

Also, smoothness classes for f_0 are defined in terms of this basis and thus connected to K. This may not always correspond to natural assumptions of the practical problem at hand; see, for instance, [6]. The same can probably be said about the polished tail or self-similarity conditions. As they stand, they refer to coefficients in the basis associated to K, which may not always be canonical.

For these reasons, it would probably be interesting for future works to consider different types of priors. It is unclear whether in general a direct analysis of the explicit expression of the likelihood (and marginal likelihood for the EB approach) will be possible. It would certainly be desirable, if possible, to develop some general understanding of empirical Bayes methods. On the other hand, it would also be interesting to develop indirect (or qualitative) techniques, similar to those of the meta-theorem of [7] for these problems. Although this may not be easy for inverse problems, some recent work for these include [11] and [8]. Other recent results on functionals using arguments allowing implicit expressions can be found in [5] and [1].

DIFFERENT APPROACHES TO NONPARAMETRIC CREDIBLE SETS. As the authors mention at the end of their introduction, for parametric models the Bernstein–von Mises (BvM) theorem is a canonical tool to justify that Bayesian credible sets are frequentist confidence sets. In [3] and [4], R. Nickl and myself proposed a possible approach for the nonparametric BvM and showed that it could be applied to the construction of fixed-regularity nonparametric confidence sets. I am not sure I understand the authors' sentence "no method that avoids dealing with the bias–variance trade-off will properly quantify the uncertainty... current practice." In [3] and [4], no adaptation claims were made, and the confidence sets there are for fixed regularity, although the proposed methodology to build such sets does not per se exclude adaptive priors. Recently, a first application of this programme with adaptive priors in white noise was carried out in [12], leading to L^2 and L^∞ adaptive confidence sets computable in practice, under appropriate self-similarity conditions. The "bias–variance" trade-off mentioned by the authors I guess typically appears when estimating the "regularity" of the signal, for instance, by an empirical Bayes technique.

BIAS–VARIANCE TRADE-OFF AND CHOICE OF THE PRIOR. There are several interesting questions mentioned by the authors beyond the L^2-results of the paper. One is obtaining Bayesian confidence sets for other norms, related to the problem of estimating certain functionals, such as the value of the function at a point; see the discussion on these in [9] for the prior (1). Another question is building different types of adaptive L^2-confidence sets, where the regularity is assumed to belong to an interval $[\alpha, 2\alpha]$, as considered in [14], again with the scheme (1).

In both cases the authors seem to conclude that marginal likelihood empirical Bayes or full Bayes methods have some trouble, related to the choice of the regularity parameter: for instance, the marginal-likelihood EB method does not seem to perform the correct bias–variance trade-off in the two problems. The proposed solution is then to choose the tuning parameter $\hat{\alpha}_n$ independently, by a possibly non-Bayes method. We agree, but one may note that all these results are for the given prior scheme (1). Is it not conceivable that, for a given problem (e.g., adaptive estimation of a functional), there exists a prior for which the two steps are performed optimally? Perhaps this is too much to ask in general, but, after all, this is the remarkable result that the authors show in the present paper: at least for the present problem, the Bayes method performs well in (1) rate-adaptation and (2) providing an (EB-)estimate $\hat{\alpha}_n$ so that the confidence set has the desired coverage.

2. Bayesian credible sets and simulations. The authors present interesting simulations and a representation of the credible sets in the case of the Volterra operator.

WHAT IS EXACTLY A PLOT OF A CREDIBLE SET? The credible ball considered in the paper is, with $L = 1$,

$$\hat{C}_n = \{\theta \in \ell^2, \|\theta - \hat{\theta}_{n,\hat{\alpha}_n}\|_2 \leq r_{n,\gamma}(\hat{\alpha}_n)\}. \tag{4}$$

In their Figure 1, the authors plot random draws from the posterior distribution. The idea is that all (but possibly a few) of these draws belong to the credible ball. From this definition, we can make two comments:

1. Curves that are not typical posterior draws belong to $\hat{C}_n$.
2. There is typically much more "information" in the posterior (coming from the prior) than the fact of belonging to such an ℓ^2-ball.

To illustrate the fact that $\hat{C}_n$ is in some sense larger than the "support" of the posterior distribution, we have generated random draws within $\hat{C}_n$ using a distribution different from the posterior. First, consider the sequence, given the data,

$$\mu = (\mu_k)_{k\geq 1} \sim \left(\hat{\theta}_{n,\hat{\alpha}_n} + a\frac{\xi_k}{(k\log^2 k)^{1/2}}\right)_{k\geq 1},$$

for $a > 0$ some small constant and ξ_k i.i.d. $N(0, 1)$ variables. Consider the law

$$\mathcal{L}(\mu|\mu \in \hat{C}_n), \tag{5}$$

the distribution of μ conditioned to belong to the set $\hat{C}_n$. Curves whose coefficients are sampled from this law are represented in the left column of Figure 1, where we took $a = r_{n,\gamma}(\hat{\alpha}_n)$, while the right column corresponds to posterior draws. One notices that the typical curves on the left are more "wiggly" than those from the

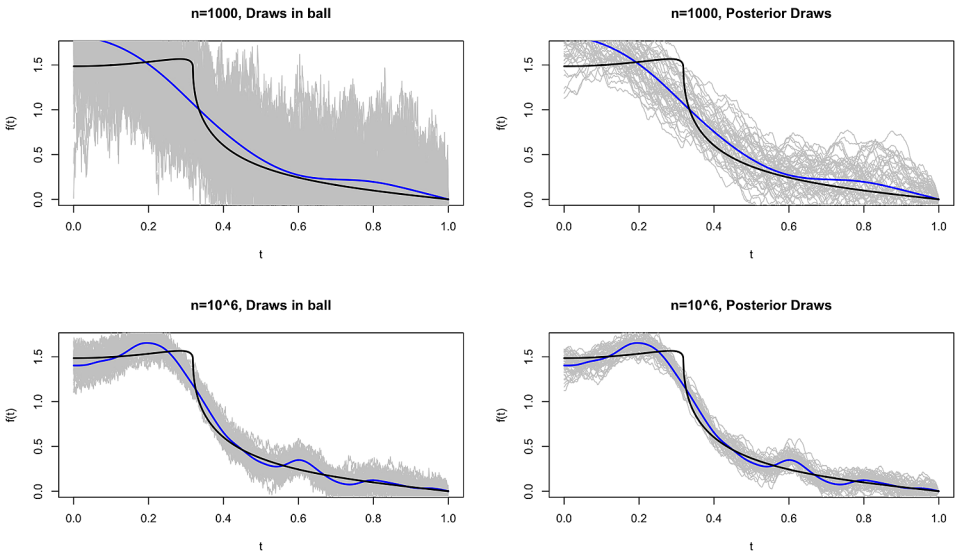

FIG. 1. *In gray, on each plot, $N = 50$ sampled curves from the posterior distribution (right column) and the law induced by (5) (left column), for $n = 10^3$ (top) and $n = 10^6$ (bottom). Posterior mean and true function are in blue and black, respectively.*

posterior distribution and also tend to spread more, depending on how much curves N are simulated, here $N = 50$.

On the other hand, the posterior distribution itself admits a series of features that are not necessarily present in a typical element of the L^2-ball. For instance, if f is a draw from the posterior on the signal function, and is $\hat{\alpha}_n$ concentrates, which is the case for self-similar-type truths, the supremum norm $\|f - \hat{f}_{n,\hat{\alpha}_n}\|_\infty$ is a stochastically bounded quantity that only depends on the data via $\hat{\alpha}_n$, as can be seen from equation (6) below. So with high probability the posterior draws stay within a tube centered at the posterior mean. If $\alpha > 1/2$, one could presumably also prove at least some supremum-norm consistency of the posterior around f_0, following, for example, [5].

Given that the mathematical definition of the credible set is (4), it seems natural to ask whether one should report draws from posterior or from (5). Or rather, would it be possible to define a credible set directly from the posterior draws themselves, instead of reporting a full L^2-ball, while still retaining the desired coverage properties?

IMPROVING ON THE ESTIMATE OF THE RADIUS. The authors simulate $N = 2000$ draws from the empirical Bayes posterior and retain the $1 - \gamma = 95\%$ closest to the posterior mean. This means that an implicit "built-in" estimator of the radius of the credible set is used. More precisely, if $R_1, \ldots, R_N$ denote the observed L^2-radii of N draws under the posterior $\Pi_{\hat{\alpha}_n}[\cdot|X]$, only the curves with radius, respectively, $R_{(1)} \leq \cdots \leq R_{(\lfloor 0.95 \cdot N \rfloor)}$ are retained. In other words, $R_{(\lfloor 0.95 \cdot N \rfloor)}$ is used as an estimator of $r_{n,\gamma}(\hat{\alpha}_n)$.

This methodology is simple and certainly reasonable for relatively large N, the precision of the "built-in" quantile estimator being of order $N^{-1/2}$. In case one likes to be precise about the $(1-\gamma)$-coverage or, in cases where the posterior can only be approximated, if one wants to detect possible outliers, one may suggest an improvement based on a separate estimation of $r_{n,\gamma}(\hat{\alpha}_n)$. First, one may note that, in general, the posterior distribution of the radius could be more easily accessible (or sampling from it could require less computing time) than the full posterior. In the considered white noise model example, computing a precise approximation of $r_{n,\gamma}(\hat{\alpha}_n)$ is simple, as the posterior distribution re-centered at the posterior mean has distribution, if τ_n is the map $\theta \to \theta - \hat{\theta}_{n,\hat{\alpha}_n}$,

$$\Pi_{\hat{\alpha}_n}[\cdot|X] \circ \tau_n^{-1} \stackrel{\mathcal{L}}{=} \bigotimes_{i\geq 1} N\left(0, \frac{1}{i^{1+2\hat{\alpha}_n} + n\kappa_i^2}\right). \tag{6}$$

It is then straightforward to simulate the random variable $\|\zeta\|_2$, where ζ is a draw from the distribution in the last display, and then estimate $r_{n,\gamma}(\hat{\alpha}_n)$ based, for example, on a quantile as before, but this time using a much larger sample size (not necessarily $N = 2000$ as before). This can be made before running the program simulating the posterior draws of the function f. For instance, in the Volterra example with $n = 1000$, one obtains the estimate $\bar{r}_{n,\gamma}(1) := 0.42 \approx r_{n,\gamma}(1)$ using a sample of size 10^5 [we set $\alpha = 1$ for simplicity, but an approximation of $r_{n,\gamma}(\hat{\alpha}_n)$ is obtained similarly, as soon as $\hat{\alpha}_n$ has been computed].

We have run a few iterations of the algorithm proposed by the authors, with the previous slight modification and setting $\alpha = 1$ for simplicity. As the estimate of the radius is improved, the rule for discarding draws is more precise. For the results in Table 1, we have taken the precise estimate $\bar{r}_{n,\gamma}(1)$ as "true."

As shown in Table 1, a few curves per experiment typically were either incorrectly included or excluded. Quantitatively, the number of such curves is not very high, but, on the other hand, these are the curves the farthest away from the posterior mean, so visually this has (sometimes) some impact on the pictures. This

TABLE 1

Experiment using the original algorithm compared to a program with separate precise estimation $\bar{r}_{n,\gamma}(1)$ (taken as "true") of $r_{n,\gamma}(1)$. After 10 repetitions, N_{fp} is the mean number of "incorrectly" retained curves (false positive) by original algorithm and N_{fn} of "incorrectly" discarded curves (false negative). In parenthesis percentage of occurrence

n	1000		10^6		10^8	
N	500	2000	500	2000	500	2000
N_{fp}	6 (40%)	5 (70%)	4 (50%)	6 (40%)	6 (50%)	4 (20%)
N_{fn}	3 (50%)	14 (20%)	6 (50%)	12 (50%)	3 (50%)	8 (80%)

observation can be applied as well for pictures of credible bands, as recently considered, for example, in [12].

Congratulations again to the authors for their inspiring series of works. Developing tools to build Bayesian credible sets for other models and priors is a very interesting topic, and we expect to see more on the subject soon.

REFERENCES

[1] CASTILLO, I. (2014). On Bayesian supremum norm contraction rates. *Ann. Statist.* **42** 2058–2091. MR3262477

[2] CASTILLO, I., KERKYACHARIAN, G. and PICARD, D. (2014). Thomas Bayes' walk on manifolds. *Probab. Theory Related Fields* **158** 665–710. MR3176362

[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] CASTILLO, I. and ROUSSEAU, J. (2013). A Bernstein–von Mises theorem for smooth functionals in semiparametric models. Available at arXiv:1305.4482.

[6] DONOHO, D. L. (1995). Nonlinear solution of linear inverse problems by wavelet–vaguelette decomposition. *Appl. Comput. Harmon. Anal.* **2** 101–126. MR1325535

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

[8] KNAPIK, B. and SALOMOND, J.-B. (2014). A general approach to posterior contraction in nonparametric inverse problems. Available at arXiv:1407.0335.

[9] KNAPIK, B. T., SZABÓ, B. T., VAN DER VAART, A. W. and VAN ZANTEN, J. H. (2012). Bayes procedures for adaptive inference in inverse problems for the white noise model. Available at arXiv:1209.3628.

[10] 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. MR2906881

[11] RAY, K. (2013). Bayesian inverse problems with non-conjugate priors. *Electron. J. Stat.* **7** 2516–2549. MR3117105

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

[13] SZABÓ, B. T., VAN DER VAART, A. W. and VAN ZANTEN, J. H. (2013). Empirical Bayes scaling of Gaussian priors in the white noise model. *Electron. J. Stat.* **7** 991–1018. MR3044507

[14] SZABÓ, B. T., VAN DER VAART, A. W. and VAN ZANTEN, J. H. (2015). Honest Bayesian confidence sets for the L^2 norm. *J. Statist. Plann. Inference*. To appear. Available at arXiv:1311.7474.

CNRS—LABORATOIRE PROBABILITÉS
ET MODÈLES ALÉATOIRES
UNIVERSITÉS PARIS VI & VII
BÂTIMENT SOPHIE GERMAIN
75205 PARIS CEDEX 13
FRANCE
E-MAIL: ismael.castillo@math.cnrs.fr