REJOINDER TO DISCUSSIONS OF "FREQUENTIST COVERAGE OF ADAPTIVE NONPARAMETRIC BAYESIAN CREDIBLE SETS"

BY BOTOND SZABÓ, A. W. VAN DER VAART AND J. H. VAN ZANTEN

TU Eindhoven, Leiden University and University of Amsterdam

We thank the discussants for their supportive comments and interesting observations. Many questions are still open and not all methodological or philosophical questions may have an answer. Our reply addresses only a subset of questions and is organized by topic. A final section reviews recent work.

1. Hierarchical Bayes credible sets. Our paper considers empirical Bayes tuning of the posterior distribution, whereas many Bayesians might prefer to use a hierarchical Bayes approach. Ghosal and Rousseau ask whether, or conjecture that, the hierarchical Bayes procedure behaves similarly as the likelihood based empirical Bayes procedure. Indeed, we can show exactly the same coverage of hierarchical Bayes credible sets for polished tail truths. A counterexample showing that hierarchical Bayes credible sets also do *not* cover without some restriction was already given in [14], while the size of such sets follows from [7]. Thus, within the context of our paper there is no difference between the two schemes.

In the hierarchical Bayes approach we endow the regularity hyperparameter α with a hyperprior distribution λ, and then apply an ordinary Bayes method with the overall prior, for some upper bound A (possibly dependent on n),

$$\Pi(\cdot) = \int_0^A \Pi_\alpha(\cdot) \lambda(\alpha) \, d\alpha.$$

For $\Pi(\cdot | X^{(n)})$ the posterior distribution relative to this prior, a hierarchical Bayes credible ball centered at the posterior mean $\hat{\theta}_n$ is defined by its radius $\hat{r}_{n,\gamma}$:

$$(1.1) \qquad \Pi\big(\theta : \|\theta - \hat{\theta}_n\|_2 \leq \hat{r}_{n,\gamma} | X^{(n)}\big) = 1 - \gamma.$$

We blow this up a bit, and for $L > 0$ consider

$$(1.2) \qquad \hat{C}_n(L) = \big\{\theta \in \ell^2 : \|\theta - \hat{\theta}_n\|_2 \leq L\hat{r}_{n,\gamma}\big\}.$$

Under a mild regularity condition on λ, similar to that in [7], these sets cover polish tail truths.

THEOREM 1.1. *Suppose that there exist $c_1, c_2 \geq 0$, c_3 and $c_4, c_5 > 0$, with $c_3 > 1$ if $c_2 = 0$, such that $c_4^{-1}\alpha^{-c_3}\exp(-c_2\alpha) \leq \lambda(\alpha) \leq c_4\alpha^{-c_3}\exp(-c_2\alpha)$, for*

all $\alpha > c_1$ and $\lambda(\alpha) \geq c_5$ for all $0 < \alpha \leq c_1$. Then for any positive A, L_0, N_0 there exists a constant L such that

$$(1.3) \qquad \inf_{\theta_0 \in \Theta_{pt}(L_0)} P_{\theta_0}(\theta_0 \in \hat{C}_n(L)) \to 1.$$

Furthermore, for $A = A_n \leq \sqrt{\log n}/(4\sqrt{\log \rho \vee e})$ this is true with a slowly varying sequence $[L := L_n \lesssim (3\rho^{3(1+2p)})^{A_n}$ works].

PROOF.   The probability of interest $P_{\theta_0}(\|\hat{\theta}_n - \theta_0\|_2 \leq L\hat{r}_{n,\gamma})$ is bounded below by

$$P_{\theta_0}(\|\theta_0 - \mathrm{E}_{\theta_0}\hat{\theta}_{n,\underline{\alpha}_n}\|_2 + \|\hat{\theta}_{n,\underline{\alpha}_n} - \mathrm{E}_{\theta_0}\hat{\theta}_{n,\underline{\alpha}_n}\|_2 + \|\hat{\theta}_n - \hat{\theta}_{n,\underline{\alpha}_n}\|_2 \leq L\hat{r}_{n,\gamma}).$$

Therefore, the theorem follows from Theorem 5.1 of [16], if

$$\|\theta_0 - \mathrm{E}_{\theta_0}\hat{\theta}_{n,\underline{\alpha}_n}\|_2 \lesssim n^{-\underline{\alpha}_n/(1+2\underline{\alpha}_n+2p)},$$

$$P_{\theta_0}(\|\hat{\theta}_{n,\underline{\alpha}_n} - \mathrm{E}_{\theta_0}\hat{\theta}_{n,\underline{\alpha}_n}\|_2 \leq C_1 n^{-\underline{\alpha}_n/(1+2\underline{\alpha}_n+2p)}) \to 1,$$

$$(1.4) \qquad P_{\theta_0}(\|\hat{\theta}_n - \hat{\theta}_{n,\underline{\alpha}_n}\|_2 \leq C_2 n^{-\underline{\alpha}_n/(1+2\underline{\alpha}_n+2p)}) \to 1,$$

$$(1.5) \qquad P_{\theta_0}(\hat{r}_{n,\gamma} \geq C_3 n^{-\overline{\alpha}_n/(1+2\overline{\alpha}_n+2p)}) \to 1.$$

The first two assertions follow immediately from (5.8) and (5.9) of [16].

For the proof of (1.5), we first note that, for any given $C_3 > 0$,

$$\Pi(\theta : \|\theta - \hat{\theta}_n\|_2 < C_3 n^{-\overline{\alpha}_n/(1+2\overline{\alpha}_n+2p)}|X^{(n)})$$

$$= \int_{\underline{\alpha}_n}^{\overline{\alpha}_n} \Pi_\alpha(\theta : \|\theta - \hat{\theta}_n\|_2 < C_3 n^{-\overline{\alpha}_n/(1+2\overline{\alpha}_n+2p)}|X^{(n)})\lambda(\alpha|X^{(n)})\,d\alpha$$

$$+ o_{P_{\theta_0}}(1).$$

The right side becomes bigger if we replace $\hat{\theta}_n$ by $\hat{\theta}_{n,\alpha}$, as the latter is the center of the Gaussian distribution $\Pi_\alpha(\cdot|X^{(n)})$, and again bigger if we replace $\overline{\alpha}_n$ in the rate inside the probability by α. From the proof of (5.7) of [16], it follows that there exists a constant C_3, such that for every α,

$$\Pi_\alpha(\theta : \|\theta - \hat{\theta}_{n,\alpha}\|_2 \leq C_3 n^{-\alpha/(1+2\alpha+2p)}|X^{(n)}) \leq 1 - 2\gamma.$$

Then the integral in the preceding display is asymptotically smaller than $1 - 2\gamma$, whence (1.5) follows by the definition of $\hat{r}_{n,\gamma}$.

To prove (1.4), we proceed similarly to the proof of (4.4) of [14]. By Jensen's inequality,

$$\|\hat{\theta}_n - \hat{\theta}_{n,\underline{\alpha}_n}\|_2^2 \leq \int \|\hat{\theta}_{n,\alpha} - \hat{\theta}_{n,\underline{\alpha}_n}\|_2^2 \lambda(\alpha|X^{(n)})\,d\alpha$$

$$(1.6) \qquad \leq \sup_{\alpha \in [\underline{\alpha}_n, \overline{\alpha}_n]} \|\hat{\theta}_{n,\underline{\alpha}_n} - \hat{\theta}_{n,\alpha}\|_2^2$$

$$+ \sup_{\alpha \notin [\underline{\alpha}_n, \overline{\alpha}_n]} \|\hat{\theta}_{n,\underline{\alpha}_n} - \hat{\theta}_{n,\alpha}\|_2^2 \int_{\alpha \notin [\underline{\alpha}_n, \overline{\alpha}_n]} \lambda(\alpha|X^{(n)})\,d\alpha.$$

We separately bound the two terms on the right side. First, as $\underline{\alpha}_n \in [\underline{\alpha}_n, \overline{\alpha}_n]$, by several applications of the triangle inequality,

$$\sup_{\alpha \in [\underline{\alpha}_n, \overline{\alpha}_n]} \|\hat{\theta}_{n,\alpha} - \hat{\theta}_{n,\underline{\alpha}_n}\|_2$$

$$\leq 2 \sup_{\alpha \in [\underline{\alpha}_n, \overline{\alpha}_n]} \|E_{\theta_0} \hat{\theta}_{n,\alpha} - \theta_0\|_2 + 2 \sup_{\alpha \in [\underline{\alpha}_n, \overline{\alpha}_n]} \|\hat{\theta}_{n,\alpha} - E_{\theta_0} \hat{\theta}_{n,\alpha}\|_2.$$

As a consequence of (5.8) and (5.9) of [16], this is bounded above by a multiple of $n^{-\underline{\alpha}_n/(1+2\underline{\alpha}_n+2p)}$, with P_{θ_0}-probability tending to one. For the second term, we first note that similar to the preceding display, with P_{θ_0}-probability tending to one,

$$\sup_{\alpha} \|\hat{\theta}_{n,\alpha} - \hat{\theta}_{n,\underline{\alpha}_n}\|_2 \leq 2 \sup_{\alpha} \|E_{\theta_0} \hat{\theta}_{n,\alpha} - \theta_0\|_2 + 2 \sup_{\alpha} \|\hat{\theta}_{n,\alpha} - E_{\theta_0} \hat{\theta}_{n,\alpha}\|_2.$$

As a consequence of (5.10) and (5.11) of [16], this is uniformly bounded by a constant times $\|\theta_0\|_2^2 + 1 \lesssim 1$, with P_{θ_0}-probability tending to one. Furthermore, in view of Section 7 of [7],

$$E_{\theta_0} \int_{\alpha \notin [\underline{\alpha}_n, \overline{\alpha}_n]} \lambda\big(\alpha | X^{(n)}\big) \, d\alpha \leq 2 e^{-(C_4 n^{1/(1+2\overline{\alpha}_n+2p)})/(1+2\overline{\alpha}_n+2p)} (\log n)^{C_5} e^{C_6 e^{\sqrt{\log n}}/3}$$

$$\leq 2 e^{-(C_4 n^{1/(1+2\overline{\alpha}_n+2p)})/(2(1+2\overline{\alpha}_n+2p))}$$

$$\lesssim n^{-(2\overline{\alpha}_n)/(1+2\overline{\alpha}_n+2p)}.$$

Therefore, by Markov's inequality, the second term on the right-hand side of (1.6) is bounded above by a multiple of $n^{-(2\overline{\alpha}_n)/(1+2\overline{\alpha}_n+2p)}$, which is smaller than the same rate at $\underline{\alpha}_n$, with P_{θ_0}-probability tending to one. $\square$

For the adaptive size we note that similar to the proof of assertion (4.5) of [14], it can be shown that there exists a positive constant C_7 such that

$$P_{\theta_0}\big(\hat{r}_{n,\gamma} \leq C_7 n^{-\underline{\alpha}_n/(1+2\underline{\alpha}_n+2p)}\big) \to 1.$$

Then following [16], we get the rate adaptive size for Sobolev balls, hyperrectangles, analytic balls,... etc.

2. Shape of the credible sets: Bands versus balls. All discussants pointed out that L_2-confidence sets are harder to visualize than confidence bands, that is, L_∞-balls. We fully agree. See our remarks on plotting below.

We chose to consider L_2-balls because they fit naturally in our inverse problem setup and can be studied theoretically with reasonable ease. At the same time, we believe that they provide an accurate (or at least not misleading) rendering of the general phenomena surrounding adaptive credible sets. We fully agree that it is of interest to work out similar results for other norms and other situations.

One of the theoretical difficulties to handle credible bands is to describe the L_∞-norm in terms of quantities that are controllable under the prior and posterior. Several authors (starting with [4]) have recently obtained contraction rates in this norm, and their work may well be extendible to adaptive credible sets.

Ghosal proves a rate of contraction for the uniform norm, for parameters such that $\sum_i i^\alpha |\theta_i| < \infty$ and a prior that depends on α. He next argues heuristically that the resulting credible sets with an *adaptive* choice of α will cover relative to the uniform norm. This is possible, but the particular empirical Bayes $\hat\alpha$ from our paper may for many true parameters not estimate Ghosal's α, but a different value.

We have encountered similar phenomena when deriving contraction rates and credible intervals for (not necessarily continuous) linear functionals of the parameter; see [13]. Since point evaluations are linear functionals, such credible intervals can be glued together into L_∞-credible bands, where due to the Gaussianity one would expect at most a logarithmic factor to be necessary to pass from pointwise to simultaneous intervals. A difficulty is that Sobolev regularity is not the most useful concept when estimating a function at a point; one would like to employ a Hölder norm. As a worst case one loses a $1/2$ when passing from Sobolev to Hölder, and this loss was seen to be real for the minimax contraction rate in [7, 8]. The likelihood-based empirical Bayes method seems to "estimate" the Sobolev regularity of the truth. In [13] we have shown that coverage can be retained by subtracting $1/2$ from the estimate, thus under-smoothing the empirical Bayes posterior distribution. In forthcoming work with Sniekers, we note that the ordinary empirical Bayes procedure may still give good coverage for many true parameters, the loss of $1/2$ being really a worst case comparison of the two norms and coverage being connected to more subtle properties of the true parameter.

3. Simulation and plotting. We included some pictures in the paper, and we feel that they nicely illustrate the limitations and strengths of adaptive Bayesian credible sets. The pictures consist of individual plots of all curves in the 95% out of 2000 curves simulated from the posterior distribution that are closest to the posterior mean. Within the resolution of the pictures these curves form a ragged grey band and it is tempting to view this as a confidence band.

We may have mislead the reader to think that the pictures show the L_2-credible set that we study theoretically in the paper. However, as already noted, L_2-balls are difficult to plot. To relate our plots to these balls, it seems one would have to "visually compute" the L_2-distance of the plotted curves to the center of the band (the posterior mean), take the maximum distance, and compare this to the L_2-distance of a tentative function to this center, in order to see whether this function is in the ball. This is hard to do. The pictures are not formal credible bands either. Still, they manage to give an impression of where the posterior distribution puts its mass.

Low and Ma describe this difficulty very accurately. In particular, our choice of making exactly 2000 draws was rather arbitrary and, indeed, at other places we

have also produced pictures showing just 20 draws (without rejecting any). All these pictures seem to *illustrate* the effect of the bias–variance trade-off, and its possible failure, on credible sets reasonably well.

Low and Ma also suggest a method for constructing L_∞-confidence bands from the L_2-credible balls and apply it to the adaptive posterior distribution. Bayesians will be delighted to see that the empirical Bayes method performs satisfactorily in their simulation study. The new concept of coverage introduced by these authors, together with Cai, is interesting.

Castillo also addresses the discrepancy between our analytic definition of a credible ball and our small simulation study. He points out that the radius can be simulated more precisely. He also suggests that simulating curves from distributions that are rougher than the posterior might be useful to fill out the gap between the support of the posterior and the ball. This is an interesting suggestion, but we would be reluctant to simulate from other distributions than the posterior distribution. We imagine that this could be queried in many ways, for example, to produce bands, intervals for specific functionals or perhaps even of qualitative aspects of parameters, but we would support the Bayesian view that the posterior distribution gives a full report of the analysis.

Nickl and Castillo [3] have introduced an approach toward credible sets based on a nonparametric Bernstein–von Mises theorem. Nickl writes to be "unsure to which extent ℓ_2-credible balls are applied in current practice as claimed in the introduction of (our paper)," and next suggests that "Practitioners may prefer (. . .) to compute credible balls in $\mathbb{H}$-spaces." Castillo wonders about our opinion that "no method that avoids dealing with the bias–variance trade-off will properly quantify the uncertainty." We do not believe we have claimed that ℓ_2-*balls* are routine in practice; if we did, then we retract that claim here. We do claim that posterior distributions are routinely used for uncertainty quantification, often by simulating from it. Then a main finding of 50 years of nonparametric statistics, theory and practice, is that the bias–variance trade-off drives everything, setting it apart from classical, parametric statistics, which deals mostly with variance, as bias is negligible, particularly in the large-sample limit. The ℓ_2-setting of our paper incorporates the bias–variance trade-off, and hence we believe that our theoretical results are relevant. It appears that Nickl and Castillo's "Bernstein–von Mises theorem in $\mathbb{H}$-space" removes bias, essentially by parameterizing the function as a collection of smooth functionals that can be estimated as the parameters in classical parametric models, with neglect of bias. Their work is very intriguing and pretty. However, as it explains away bias, we found it difficult to believe that it solves the nonparametric problem. It is still more intriguing that pictures by Ray in [9], which are based on the $\mathbb{H}$-spaces of [3], look similar to ours. Possibly that is because these pictures do not show their suggested set, just as our pictures are deficient in this sense. This deserves further investigation.

4. Other priors. The discussants pose the question whether our results extend to other priors than the $N(0, i^{-1-2\alpha})$-priors in our paper. We believe the answer is affirmative: it appears that the polished tail condition is not linked to the form of the priors.

One reason to believe this are preliminary results, of ourselves and in a forthcoming thesis of Sniekers at Leiden University, about priors of the form

$$\Pi_\tau = \prod_{i=1}^{\infty} N(0, \tau^2 i^{-1-2\alpha}),$$

where α is fixed, but τ is adapted to the data, by either an empirical or hierarchical Bayes method. For empirical Bayes we plug the marginal maximum likelihood estimator $\hat{\tau}_n$ of τ into the posterior distributions for fixed τ, and construct adaptive credible sets of the form

(4.1) $$\hat{C}_n(L) = \{\theta \in \ell^2 : \|\theta - \hat{\theta}_n(\hat{\tau}_n)\|_2 \le L r_{n,\gamma}(\hat{\tau}_n)\},$$

where $\hat{\theta}_n(\tau)$ is the posterior mean and $r_{n,\gamma}(\tau)$ satisfies

(4.2) $$\Pi(\theta : \|\theta - \hat{\theta}_n(\tau)\|_2 \le r_{n,\gamma}(\tau)|X^{(n)}) = 1 - \gamma.$$

THEOREM 4.1.   *For any* A, L_0, N_0 *there exists a constant* L *such that*

$$\inf_{\theta_0 \in \Theta_{pt}(L_0)} P_{\theta_0}(\theta_0 \in \hat{C}_n(L)) \to 1.$$

Hierarchical Bayes credible sets will similarly cover. However, these sets have the disadvantage that they may be unnecessarily big. In our paper [15] we proved that the corresponding posterior distributions contract at the minimax rate over Sobolev balls of regularity $\beta < \alpha + 1/2$, but only at the suboptimal rate $n^{-(1+2\alpha)/(4+4\alpha)}$ if $\beta > \alpha + 1/2$. The latter suboptimal rate is partially due to the variance of the posterior distribution, and hence, in the case that $\beta > \alpha + 1/2$, the radius of the credible balls will be suboptimal as well.

5. Choice of basis. Rousseau and Castillo point out that the polished tail condition is dependent on the chosen basis, whereas one might hope or expect the set of "good behaving" true parameters not to depend on the basis.

In inverse problems the eigenbasis of the operator K^*K plays, implicitly or explicitly, an important role to describe the problem [1, 6, 8, 10] and, hence, it is natural to assume the polished tail condition with respect to this basis. Other bases were explored in recent work [5], but a good link between the operator and the prior seems always needed.

In "direct problems" one can consider any basis. This then determines both the prior and the polished tail condition. The prior, or rather collection of priors, will be chosen to model a scale of models that is thought to capture the true parameter. In

our situation these were Sobolev spaces, which are naturally described in a basis. That the polished tail condition will adopt the same basis seems not unnatural. After all, "good-behaving" is not an absolute property of a parameter, but is relative to a method, which is the one induced by the prior in this case.

There is a good scope for extensions to other models and priors. In our case the coefficients could be modeled differently than independent and Gaussian, although both seem natural. We imagine that similar results as in our paper can easily be written down for double-indexed bases, as wavelets, thus moving closer to the earliest works on self-similarity. More challenging will be priors such as Dirichlet mixtures, which are known to adapt to the bandwidth in the (normal) kernel. What can be said about their coverage?

6. Further references. The paper [9] derives an adaptive and nonparametric version of the Bernstein–von Mises theorem, using techniques developed in [3] and [7], under a self-similarity restriction, and next applies this result to construct adaptive credible sets. The same paper also considers spike and slab priors and L_∞-credible bands. The author of [2] investigates credible sets from an oracle perspective. He considers truncated (finite dimensional) Gaussian priors and shows that the empirical Bayes approach chooses the optimal truncation level under a (slightly) extended version of the polished tail condition. This family of priors is relatively wide and contains a member that attains the minimax posterior contraction rate for every regularity class S^β. The authors of [12] have followed up their work with investigating adaptive pointwise credible sets using rescaled (integrated) Brownian motion as a prior in the nonparametric regression model. Random smoothing spline priors with Gaussian weights on the spline coefficients are shown in [11] to give honest credible sets in the nonparametric regression problem under the self-similarity condition.

REFERENCES

[1] AGAPIOU, S., LARSSON, S. and STUART, A. M. (2013). Posterior contraction rates for the Bayesian approach to linear ill-posed inverse problems. *Stochastic Process*. *Appl*. **123** 3828–3860. MR3084161

[2] BELITSER, E. (2014). On coverage and oracle radial rate of DDM-credible sets under excessive bias restriction. Available at arXiv:1407.5232.

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

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

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

[6] KNAPIK, B. T., VAN DER VAART, A. W. and VAN ZANTEN, J. H. (2013). Bayesian recovery of the initial condition for the heat equation. *Comm*. *Statist*. *Theory Methods*. **42** 1294–1313. MR3031282

[7] KNAPIK, B. T., SZABÓ, B. T., VAN DER VAART, A. W. and VAN ZANTEN, J. H. (2015). Bayes procedures for adaptive inference in inverse problems for the white noise model. *Probab. Theory Related Fields*. DOI:10.1007/s00440-015-0619-7.

[8] 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

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

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

[11] SERRA, P. and KRIVOBOKOVA, T. (2014). Adaptive empirical Bayesian smoothing splines. Available at arXiv:1411.6860.

[12] SNIEKERS, S. and VAN DER VAART, A. (2014). Credible sets in the fixed design model with Brownian motion prior. *J. Statist. Plann. Inference*. DOI:10.1016/j.jspi.2014.07.008.

[13] SZABÓ, B. (2014). On Bayesian based adaptive confidence sets for linear functionals. Available at arXiv:1412.0459.

[14] SZABÓ, B., VAN DER VAART, A. and VAN ZANTEN, H. (2014). Honest Bayesian confidence sets for the L2-norm. *J. Statist. Plann. Inference*. DOI:10.1016/j.jspi.2014.06.005.

[15] 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

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

B. SZABÓ
DEPARTMENT OF MATHEMATICS
EINDHOVEN UNIVERSITY OF TECHNOLOGY
P.O. BOX 513
5600 MB EINDHOVEN
THE NETHERLANDS
E-MAIL: b.szabo@tue.nl

A. W. VAN DER VAART
MATHEMATICAL INSTITUTE
LEIDEN UNIVERSITY
P.O. BOX 9512
2300 RA LEIDEN
THE NETHERLANDS
E-MAIL: a.w.van.der.vaart@umail.leidenuniv.nl

J. H. VAN ZANTEN
KORTEWEG-DE VRIES INSTITUTE FOR MATHEMATICS
UNIVERSITY OF AMSTERDAM
P.O. BOX 94248
1090 GE AMSTERDAM
THE NETHERLANDS
E-MAIL: hvzanten@uva.nl