Discussion of Uncertainty quantification for the horseshoe

Ismaël Castillo

The presently discussed paper by Stéphanie van der Pas, Botond Szabó and Aad van der Vaart is a third of a series of very interesting works on convergence properties of posterior distributions associated to the horseshoe prior in the sparse normal means model. The horseshoe prior distribution as considered in the paper is a specific scale mixture of normal distributions. Given τ, it is the distribution of θ_1 obtained from

$$\theta_1 \,|\, \lambda, \tau \sim \mathcal{N}(0, \lambda^2\tau^2), \qquad \lambda \sim C^+(0,1). \tag{1}$$

Convergence rates are obtained in van der Pas et al. (2014) and adaptive counterparts are derived in van der Pas et al. (2017). In the present paper the authors make an important step further and study uncertainty quantification: they demonstrate that under certain conditions credible sets derived from the horseshoe posterior distribution, either local marginal credible intervals or global ℓ^2 credible balls, can be used as confidence sets, asymptotically in the number of observations. This is, after Belitser and Nurushev (2015), one of the first works on the subject using Bayesian methods in sparse settings.

I really enjoyed reading this paper and the previous ones. Below I discuss two main points and then close my discussion with a couple of more specific questions. The first comment draws some analogies with spike and slab priors with sparsity parameter calibrated by empirical Bayes (EB) and asks for possibly more general horseshoe-type distributions. In a second comment, we discuss model selection properties and credible sets for the horseshoe.

Some of the comments below are inspired by current work in progress with Romain Mismer Castillo and Mismer (2017) and Botond Szabó Castillo and Szabó (2017), in which we consider related questions for spike and slab prior distributions

$$\theta_1 \sim (1-\alpha)\delta_0 + \alpha G, \tag{2}$$

for some absolutely continuous distribution G and α calibrated by an empirical Bayes approach: following the steps of Johnstone and Silverman (2004), who studied risks of a class of point estimators derived from the EB approach, we consider the convergence of the full EB-posterior and related credible sets properties.

1. More flexible horseshoe prior distributions?

[*] Université Pierre et Marie Curie – Paris 6; Laboratoire Probabilités et Modèles Aléatoires (LPMA); UMR 7599; 5, place Jussieu; 75005 Paris, France
ismael.castillo@upmc.fr

DOI: 0000

Following Carvalho et al. (2010), if $\pi(\theta_1)$ denotes the marginal density of θ_1 in (1),

$$\frac{1}{\tau}\log\left(1+\frac{4\tau^2}{\theta_1^2}\right) \lesssim \pi(\theta_1) \lesssim \frac{1}{\tau}\log\left(1+\frac{2\tau^2}{\theta_1^2}\right). \tag{3}$$

This implies that the horseshoe prior, given τ, has a pole at zero and Cauchy tails. The pole at zero guarantees shrinkage of small signals while heavy tails avoid over-shrinkage of large signals.

There seems to be a striking correspondance between the tuning parameter τ of the horseshoe and the success probability α in the spike and slab prior, especially when G is taken to be a distribution with Cauchy tails. For instance, when using a marginal maximum likelihood empirical Bayes (MMLE) method to estimate α for such a spike and slab prior with Cauchy tails, one can show Castillo and Mismer (2017) (thereby slightly improving, in the case one restricts to $\ell_0[p_n]$ classes, upon the estimate from Lemma 10 and Eq. (101) in Johnstone and Silverman (2004)) the estimate $\hat{\alpha}$ is such that, as $n \to \infty$,

$$\sup_{\theta_0\in\ell_0[p_n]} P_{\theta_0}\left[\hat{\alpha} > (p_n/n)\sqrt{\log(n/p_n)}\right] = o(1),$$

where p_n is the sparsity parameter. This is the same as the upper boundary for τ obtained by the authors who established in van der Pas et al. (2017) that the MMLE $\hat{\tau}_n$ verifies

$$\sup_{\theta_0\in\ell_0[p_n]} P_{\theta_0}\left[\hat{\tau}_n > \tau(p_n)\right] = o(1),$$

which is part of Condition 1 of the present paper. This suggests that tails of the horseshoe and tails of the slab distribution play a similar role, also at level of precise conditions arising in the proofs.

This naturally leads to the question of whether it is possible to allow for other tail distributions for the marginal distribution of θ_1 for horseshoe-type priors. Another reason why we mention this is that it appears from Castillo and Mismer (2017)-Castillo and Szabó (2017) that in the spike and slab case, tails of G are particularly critical in obtaining optimal adaptive rates and confidence sets when using an empirical Bayes method. While Cauchy tails are fine in the spike and slab case when the squared ℓ^2 loss $\|\theta-\theta'\|^2 = \sum_i(\theta_i-\theta_i')^2$ is considered, they presumably lead to suboptimal rates if the loss is measured in terms of d_q-distances $d_q(\theta,\theta) := \sum_i|\theta_i-\theta_i'|^q$ (as in Castillo and van der Vaart (2012)) when $q<1$ (we note here that we are talking about results for the full EB posterior distribution, not aspects of it such as the median or mode as in Johnstone and Silverman (2004), for which this phenomenon does not arise).

Perhaps heavier tails, such as $\theta_1^{-1-\delta}$ with $\delta<1$ could be obtained by considering one of the other mixture priors mentioned in the paper such as the normal-exponential-gamma or the more general global-local scale mixture of normals, although we could not find any explicit results on tails of the marginal distribution in the mentioned references.

2. Model selection: 'sparsifying' the horseshoe?

By construction, a draw from the posterior distribution associated to the horseshoe prior does not set any component exactly to zero. In a sense, again by construction, the horseshoe prior is not exactly 'made for' $\ell_0[p_n]$ classes. Still, as the authors nicely prove, it leads to very good results for estimation and confidence sets for the squared ℓ^2 loss and $\ell_0[p_n]$ classes.

When one looks at a different type of results, such as model selection, or results for loss functions that are more sensible to missing the exact zeros, such as d_q–losses, something must be done, and the authors propose an additional *selection rule* to set some of the coefficients to zero.

The selection rule consists in looking at marginal credible intervals for individual coefficients θ_i and to select the given index i if the credible interval does not contain zero. This rule is very intuitive, but is there a qualitative justification of this specific choice? For instance, can something be said about its corresponding 'threshold' in the sense of the smallest signal strength that gives detection?

Part of the interesting message from Sections 2 (credible intervals) and 3 (model selection) from the paper is that, after the selection rule is applied, the resulting procedure does qualitatively something similar to what priors with a built-in selection procedure, such as spike and slab, would do: most true zero parameters are set to zero, large enough signals are always detected, while 'intermediate' signals are often set to zero.

One can wonder whether it is possible to recover some results obtained for priors with built-in selection with the horseshoe combined with the selection rule, for instance in the following two directions

a) Number of non-zero coefficients. From (i) of Theorem 3.1, it follows that the total number of selected coefficients is no larger than $p_n + (n - p_n)\gamma_n$ (I believe γ_n should be read $(n - p_n)\gamma_n$ in point (i) of the statement). The condition on γ_n implies that $n\gamma_n$ is of larger order than p_n. Could one prove that the bound is close to p_n, or rather here, say, a constant times $p_n\sqrt{\log(n/p_n)}$?

b) d_q–losses. In principle, one could also expect that, once some of the smallest coefficients of the horseshoe estimator are set to zero, the resulting 'after selection'-estimate would perform well also in terms of d_q–distances, at least for some qs in $(0, 2)$. This question arises for estimation as in van der Pas et al. (2017) but also for credible sets as in Section 4 of the present paper.

Specific questions

(i) Adaptive minimax rate with precise logarithmic term.

In the companion paper van der Pas et al. (2017), the authors obtain a nearly optimal minimax rate $Cp_n \log n$ for the horseshoe posterior, which may miss the minimax rate of the order $p_n \log(n/p_n)$ for signals that are nearly dense (e.g. $p_n = n/\log n$ or

$p_n = n/e^{\sqrt{\log n}}$). It would be interesting to see whether the precise logarithmic term can be obtained.

(ii) Simulations.

In principle, when looking at classes of sparse vectors that do not specifically contain zeros, such as strong or weak ℓ^p classes ($0 < p < 2$), the horseshoe estimator should perform even better, in the sense that it is not 'penalised' by the fact of not setting some coefficients to zero. Did the authors do some simulations in this type of setting?

Also, how does one choose in practice the blow-up factor L of the credible intervals or credible balls? Is there a recommended rule to chose it in simulations?

References

Belitser, E. and Nurushev, N. (2015). "Needles and straw in a haystack: empirical Bayes confidence for possibly sparse sequences." Preprint. 1

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). "The horseshoe estimator for sparse signals." *Biometrika*, 97(2): 465–480. 2

Castillo, I. and Mismer, R. (2017). "Empirical Bayes analysis of Spike and Slab posterior distributions." In preparation. 1, 2

Castillo, I. and Szabó, B. T. (2017). "Spike and Slab empirical Bayes sparse credible sets." In preparation. 1, 2

Castillo, I. and van der Vaart, A. W. (2012). "Needles and straw in a haystack: posterior concentration for possibly sparse sequences." *Ann. Statist.*, 40(4): 2069–2101. 2

Johnstone, I. M. and Silverman, B. W. (2004). "Needles and straw in a haystacks: empirical Bayes estimates of possibly sparse sequences." *Ann. Statist.*, 32(4): 1594–1649. 1, 2

van der Pas, S., Kleijn, B., and van der Vaart, A. (2014). "The horseshoe estimator: posterior concentration around nearly black vectors." *Electron. J. Stat.*, 8(2): 2585–2618. 1

van der Pas, S., Szabó, B., and van der Vaart, A. (2017). "Adaptive posterior contraction rates for the horseshoe." *Electron. J. Stat.*, 11(2): 3196–3225. 1, 2, 3