Response to Associate Editor

We appreciate your comments and suggestions on ways to improve the article. Thank you.

1. In addition to addressing the two reviewer's comments, I think that a well-designed simulation study would greatly improve the quality of this work. Right now, it is difficult to have insights about the performance (accuracy of the estimation of the accuracies, and computation time) of the proposed method, compared to existing alternative methods. In particular, the simulation study should aim at justifying the author's claims about scalability. Hence, some scenarios with $p >> n$ should be included. The authors could use any setting and methods they see fit but they should at least include a simple linear regression setting with different combinations of (p, n). In that case, the authors should make some links with the burgeoning "post-selection inference" literature.

   Thanks for the suggestion. We have added a regression simulation study with different combinations of (p, n) in both sparse and non-sparse settings. Your comment also led us to add extra individual weights into the penalty. This helps in the large p case. As for CPU time, WBB is comparable (approximately 1.5 times) to Bayesian LASSO in the linear case, but of course, for the nonlinear deep learning example, WBB provides significant advantages over full MCMC.

Response to Referee 1

We appreciate your comments and suggestions on ways to improve the article. Thank you for your detailed comments.

Major Concerns

1. Page 8 line 18 (Eqn (4)). Does λ remain the same for all sets of $\mathbf{w}$, or it is tuned for each set of $\mathbf{w}$? I think the latter would be more appropriate. If the latter is used, and λ is selected through CV, I assume the $\mathbf{w}$ associated with the training and testing cases will stay when estimating parameters in the training set and calculating the testing loss? But you suggest the former, which is practically convenient, how would be select that one λ? Please add some discussion regarding the above questions.

   Thanks for your comment. We have a discussion on the selection of λ right after Algorithm 1, Section 2.5. λ is fixed for all sets of $\boldsymbol{w}$ and can be chosen by marginal maximum likelihood or CV, based on the unweighted posterior function.

2. In Fig 1, do the authors have any explanation why the WBB mean is more sparse? If the loss function is divided by $\mathbf{w}_1$, then for each set of ($\mathbf{w}_0$, $\mathbf{w}_1$), it is just the regular lasso with a single tuning parameter $\lambda \mathbf{w}_0/\mathbf{w}_1$, where $\mathbf{w}_0/\mathbf{w}_1$ is the ratio of two independent exponential

distribution, whose mean would be much larger than 1, is this the reason why WBB results in more sparsity? This would also explain why the fixed prior WBB ($\mathbf{w}_0 = 1$) provides more sparse solutions than the weighted prior WBB in Figure 2.

Thanks for this good point. A second referee pointed out the ratio of w's can lead to extra sparsity. We have added individual weight on each coordinate in the prior term, given that for the LASSO prior $\phi(\theta)$ is separable. This helps in the case where p is large – and also your point on sparsity.

3. In Application 3 (deep learning on MNIST), I would suggest the authors add some narratives that the purpose of this case study is just for demonstration of WBB being applied in NNs, and the NN structure (2 hidden layers, feedforward) used is not the state of the art NN, and as a result, the accuracy rate is way below the state of the art accuracy rate on MNIST data (which is around 99%).

   Thanks for the suggestion. We've added this point in the end of MNIST example as well as two related references.

Minor Concerns

1. Page 7 line 39: I would suggest providing the specific URL on the optimization view (https://cran.r-project.org/web/views/Optimization.html) instead of the generic CRAN URL.

   Thanks for the suggestion. The URL has been changed.

2. Why not directly state that $f(\theta^*_{\mathbf{w}}|y)$approximates $p(\theta|y)$ (given in Eqn (6) and (9)?

   Thanks for your comment. You are right. The right hand sides of Eqn (6) and (9) are the same, which builds the approximate equivalence between the target $p(\theta|y)$ and our WBB variation $p(\theta^*_{\boldsymbol{w}}|y)$.

3. Page 8 line 45: I was somewhat confused by introducing sets B when defining Eqn (5).

   Thanks for your comment. $\mathcal{B}$ can be any measurable set in the parameter space that we are interested in. The conditional distribution of $\theta^*_{\mathbf{w}}$ given data gives the probability of set $\mathcal{B}$ which approximately equal to its posterior probability.

4. In the 2nd application, why $\alpha = 1/2$ in bridge used, why not use $\alpha = 1$ to obtain Bayesian lasso, which seems to be a more appropriate benchmark for WBB that operates on the lasso regression.

   Thanks for the good suggestion. We have switched to Bayesian LASSO in this example as well as in the simulation study.

5. In Fig 2, do the authors have any explanation why the WBB means are more sparse than bridge? $\alpha = 1/2$ in bridge, would it provide more sparse estimator than lasso (where WBB) is used if the same tuning parameter is used?

   Thanks for the suggestion. We have removed Bayesian bridge and used Bayesian LASSO instead. In WBB, we now allowed a separate weight for each β_j.

6. Page 20 line 26. Remove $K = 10$.

   Thanks for the suggestion. This is done.

7. Page 21, line 31, how did you decide on $\lambda = 10^{-4}$?

   Thanks for the suggestion. As you mentioned, this section (MNIST Example) is for illustration purpose, showing how WBB can be easily implemented in deep learning. Our main goal is not the optimization of classification accuracy. $\lambda = 10^{-4}$ is chosen manually by authors.

Response to Referee 2

We appreciate your comments and suggestions on ways to improve the article. Thank you.

1. (Page 3, Line 50) "Thus, uncertainty assessments are provided at little extra computational cost over the original training computations." – WBB needs to repeatedly carry out optimization for different draws of the random weights, right? Why would that be considered "little extra computational cost" over just a single run of optimization?

   Thanks for your comment. Your statement is right. We've added a sentence after Algorithm 1, noting that WBB can be carried out in a parallel way.

2. (Page 3, Line 53) "...it is straightforward to add a regularization path across hyperparameters..." & (Page 22, Line 41) "...we obtain a full regularization path as a form of prior sensitivity analysis..." – The statement suggests that the full regularization path can be obtained as a bi-product of the required computation for WBB (without extra computational costs), say, as in the least angle regression algorithm used in LASSO. However, I do not find any discussion of how one can actually achieve this. The random weight on the penalty term does induce a variation in the level of regularization, but it is misleading to call this 'full regularization path.'

   Thanks for your comment. You're right. By "straightforward" we just mean that a full regularization path can be obtained by repeating WBB on different λ. This indeed requires extra computational costs. We don't claim the regularization path as a bi-product of WBB. We've revised this sentence to make it clear.

3. Neither the theoretical nor empirical analysis supports WBB's ability to approximate complex posterior distributions in high dimensions. The provided theoretical analysis is based on the first-order Taylor expansion – which is essentially the Laplace approximation of the posterior – which of course is valid in the large sample limit by the Bayesian central limit theorem. But there is no need to impose regularization in this asymptotic regime, and thus the analysis provides little support for the intended use cases of WBB. The authors also note that the random exponential weights are partly motivated by the Bayesian bootstrap of Rubin (1981). This motivation again provides little support for the theoretical validity of WBB; in fact, Rubin (1981) provides the Bayesian interpretation of bootstrap as a criticism against the (frequentist) bootstrap.

   Empirically, the authors compare the WBB uncertainty estimate to that of the true posterior (computed by MCMC) in the Bayesian bridge regression applied to the diabetes data set ($n = 442 \& p = 10$). This example hardly demonstrates how well the WBB would approximate the true posterior distribution in more realistic applications of the Bayesian bridge (or other Bayesian sparse regression models) in which $p \geq n$ (or at least p is much larger). Since p is very small and n is significantly larger than p in the diabetes data set, we expect that the posterior is well approximated by the Laplace approximation – the theoretical basis of WBB. In fact, the posterior marginal distributions from MCMC shown in Figure 2 appear quite Gaussian (and the apparent slight non-Gaussianity may simply be due to Monte Carlo errors).

   The Bayesian bridge example actually indicates that WBB fails to accurately approximate the posterior even in this favorable setting; note the discrepancy between the true posteriors and WBB estimates for "age", "ldl", and "tch." The example also indicates the sensitivity of WBB to the random weight on the penalty term. The fact that WBB tends to yield a sparser solution (as pointed out by the authors themselves) can be explained by occasional large weights on the penalty term.

   Thanks for your very good point. It led us to reconsider how to deal with the case $p > n$. We now add a separate weight on each penalty term to mitigate the problem you identified, although the problem is not completely eliminated. We have added a simulation study to illustrate your point – we used Bayesian LASSO.

4. Despite the issue raised, I think there are potential contributions in the manuscript the authors just need to make the case for why WBB is "better" or "more promising" than alternative methods for quantifying uncertainty. For example, given that most of the existing approximate Bayesian procedures rather poorly estimate the true posterior, perhaps the accuracy of WBB should be measured against other approximate Bayesian procedures such as variational inference. (Such comparisons are not performed in the manuscript.)

   Thanks for your suggestion. We have added a simulation study which compares WBB with Bayesian LASSO. This provides good evidence of applicability of WBB as an alternative method for sampling posterior.