Statistical Science
2017, Vol. 32, No. 1, 41–43
DOI: 10.1214/16-STS607
Main article DOI: 10.1214/16-STS576

Toward Automated Prior Choice

David B. Dunson

1. GENERAL THOUGHTS ON PRIOR CHOICE

There is a pressing need for more work providing general guidelines for prior choice in realistically complex Bayesian models for real world applications. I find that the rich literature on "objective Bayes" (O'Bayes) lacks useful suggestions, with too much focus on "flat" and noninformative priors, and on approaches designed to mimic "old school" (i.e., prior to the modern era of penalization) frequentist inferences. In practice, I find that it is almost always a bad idea to choose a noninformative or very high variance/diffuse prior in complex modeling settings. Such priors tend to only work well in very simple settings; for example, when the data contain ample information and the model under consideration is regular and contains a modest number of parameters.

In practice, priors that tend to have good performance in realistically complex models almost always favor some degree of shrinkage toward some notion of a low-dimensional structure. If the prior is overly vague and the data are potentially not very informative about certain model parameters, then instabilities can result computationally and Bayesian inferences can have relatively poor behavior (e.g., in a mean square prediction or estimation error sense). Although shrinkage is most famously important in high-dimensional low sample size data settings, it can lead to gains much more broadly. There is an increasingly vast literature proposing shrinkage priors that are targeted toward specific settings and do not require subjective elicitation of hyperparameters using domain knowledge. Although most of the focus (by far) has been on Gaussian linear regression and closely related modeling contexts, there is an increasing literature on more elaborate settings ranging from factor modeling of high-dimensional covariance matrices (e.g., Bhattacharya and Dunson, 2011) to analysis of many way contingency tables and high-dimensional categorical data Zhou et al., 2015.

David B. Dunson is Arts and Sciences Professor, Department of Statistical Science, Duke University, Durham, North Carolina 27708, USA (e-mail: dunson@duke.edu).

Much of my own research agenda focuses on designing new and better classes of priors for complex data and models, with a particular emphasis on high dimensional and object data settings. In our work, we often attempt to design priors that will lead to appealing frequentist properties, such as efficient rates of concentration of the posterior distribution in asymptotic regimes in which the dimension of the data increases with the sample size (refer, e.g., to Bhattacharya et al., 2015 and Zhou et al., 2015). In addition, a common theme is designing the prior in such a manner that a very small number of tuning parameters control the degree of shrinkage toward some simple structure (zero coefficient values, low rank factorization, etc.). However, often it can be complicated to choose such priors and validate their properties. Hence, it is appealing to have new prescriptive approaches that can help one to target design of new priors. Current thinking in the "pragmatic" Bayes community is that priors should be chosen to be (a) weakly informative in the sense of placing high probability on a wide range of plausible values while avoiding an overly-vague specification; (b) concentrated near some lower-dimensional structure (e.g., zero parameter values) while having heavy tails to be robust to deviations from this structure; (c) have a simple form favoring interpretation and computation. Of course, in practice it is often not clear how exactly to choose a prior having properties (a)–(c); although there are many widely used families that satisfy (a)–(c) in certain common classes of problems, it is typically not clear how to choose hyperpriors and best select from among the members of a family of priors. In addition, it is difficult to develop appropriate priors in classes of problems that have not been as widely studied; for example, outside of locally Gaussian and/or linear models.

2. PENALIZED COMPLEXITY PRIORS AND THE SIMPSON ET AL. APPROACH

The Simpson et al. article provides an important and thought-provoking contribution to the rich literature on penalized complexity (PC) priors. Many (most?) of the existing shrinkage priors in the Bayesian literature can also be said to penalize complexity in shrinking toward a simple baseline model structure. However, the

main contribution of Simpson et al. is to provide a formalization of how to explicitly penalize *complexity*, using Kullback–Leibler (KL) divergence from the base model, in a simple and potentially broadly useful manner. In addition, in the process of considering the idea of PC priors, they introduce some intriguing and thought-provoking ideas, which should hopefully help to stimulate research in the important area of automated prior choice in general problems. In the following, I include some brief comments on their work.

Definition of Overfitting

This was one aspect of the article I found particularly interesting and surprising. In complex Bayesian models, over-fitting is a common concern and shrinkage toward a simple base model is a well-known and widely used solution. However, I was surprised to read their definition that a prior overfits if "its density in a sensible parameterization is zero at the base model." This is not a definition that I would have thought to apply—yes we want the prior to assign sufficient mass around the base model in some sense, but it is unusual to focus just on the density at zero in considering whether or not the prior overfits. A prior could overfit (i.e., lead to posteriors that tend to overfit) even if the density is nonzero at a simple base model if the density does not drop off rapidly enough for larger distances. On the other hand, a prior could badly under-fit (i.e., over-shrink interesting structure in the data) even if the density is nonzero at the baseline model if the density drops off too rapidly away from zero before reaching the "true" complexity level. The authors bypass this issue and make their simple definition more reasonable by focusing on constant rate penalization (their principle 3), leading to an exponential tail. This exponential form will presumably lead to good behavior only in certain settings, but is appealing as a simple choice.

Measure of Complexity

The Kullback–Leibler (KL) divergence for a more complex model relative to a simple base model provides a natural notion of complexity, which leads to some appealing properties for the proposed class of PC priors. First, the KL is a measure of the amount of additional information in the more complex model over the base model, and hence is a good choice philosophically. In addition, the KL divergence can be easily calculated analytically for multivariate Gaussian likelihoods, providing analytic tractability which the authors take advantage of in the latent Gaussian examples presented in the manuscript (one wonders how useful the proposed PC priors are in non-Gaussian models). I was quite excited and interested to see how some of the examples presented work out, leading to slightly unusual but well-motivated priors for a random effects precision, the degrees of freedom of a t-distribution and the shrinkage parameter in the normal means problem. These examples are all essentially toy cases, but still the results are intriguing. The interpretation of the PC priors as having an approximately tilted Jeffrey's form near the base model was also quite interesting.

One wonders whether there is any relationship between the proposed class of PC priors based on KL, and recent work on robust alternatives to the usual Bayesian paradigm that also utilize KL in their specifications. For example, Miller and Dunson (2015) proposed a robust alternative to the posterior distribution based on conditioning on the event that the observed data are close to data generated from the presumed model. They used KL to define closeness, and placed an exponential prior on the neighborhood size. One could view this prior on neighborhood size as an alternative type of penalized complexity prior.

Generality

The authors' goals are extremely ambitious in attempting to define principles and a concrete class of priors that are useful in routine implementations for very broad classes of models limiting the need for complex prior elicitation in each new application. Although this article contains a lot of very interesting and thought provoking ideas, the authors certainly fall far short of their goal (as they acknowledge). The main issues to me in terms of generalizability relate to (a) feasibility of automatically calculating and doing computation with the PC prior for arbitrary models within a probabilistic programming language, such as BUGS, JAGS, STAN, etc.; (b) the fundamental limitations of defining the priors locally ignoring the global structure of the model entirely; (c) the lack of ability to characterize dependence within different hyperparameters; (d) whether or not the proposed class of PC priors actually has generally good properties in some sense. It seems difficult to attack (a)–(d) simultaneously but I nonetheless feel that the proposed article takes more of a sizable step toward automatic prior specification in complex models than most previous relevant articles.

3. NONPARAMETRIC BAYES MODELS

Much of my research agenda focuses on designing and applying nonparametric (NP) Bayes methods in

complex applications. As a general rule of thumb, I typically attempt to choose the nonparametric prior in such a way that it is centered on a simple baseline parametric model, with one or perhaps two tuning (hyperparameters) controlling collapsing back onto the base model. Such a structure allows the posterior to concentrate near the baseline model when that model provides an adequate approximation. One example is the problem of modeling an unknown conditional density $f(y|x)$ of a response variable y given predictors x. In this setting, I typically choose the base model as a (possibly sparse) Gaussian linear regression, with mixtures used to allow flexible deviations (Dunson and Park, 2008; Chung and Dunson, 2009). Certain hyperparameter values will favor large weight on one mixture component, inducing collapsing on a normal linear regression as appropriate.

This type of approach to NP Bayes modeling tends to protect against over-fitting in broad settings and to have a close philosophical relationship to the notion of PC priors introduced by Simpson et al. Potentially the principles defined in their article may provide a useful manner in which one can choose good hyperpriors on key tuning parameters in NP Bayes models, though a major practical question is computability of such priors in the NP Bayes setting as the K–L divergence is typically not available in a simple form or easy to approximate accurately analytically for complex models.

4. CLOSING COMMENTS

I congratulate the authors on an interesting and thought provoking contribution, and hope that this work inspires more research in the very important topic of how to automatically choose priors in realistically complex Bayesian models.

REFERENCES

BHATTACHARYA, A. and DUNSON, D. B. (2011). Sparse Bayesian infinite factor models. *Biometrika* **98** 291–306. MR2806429

CHUNG, Y. and DUNSON, D. B. (2009). Nonparametric Bayes conditional distribution modeling with variable selection. *J. Amer. Statist. Assoc.* **104** 1646–1660. MR2750582

DUNSON, D. B. and PARK, J.-H. (2008). Kernel stick-breaking processes. *Biometrika* **95** 307–323. MR2521586

MILLER, J. and DUNSON, D. B. (2015). Robust Bayesian inference via coarsening. Available at arXiv:1506.06101.

ZHOU, J., BHATTACHARYA, A., HERRING, A. H. and DUNSON, D. B. (2015). Bayesian factorizations of big sparse tensors. *J. Amer. Statist. Assoc.* **110** 1562–1576. MR3449055