Statistical Science
2017, Vol. 32, No. 1, 29–32
DOI: 10.1214/16-STS591
Main article DOI: 10.1214/16-STS576

Prior Specification Is Engineering, Not Mathematics

James G. Scott

In their thought-provoking paper, Drs. Simpson et al. argue that "the current practice of prior specification is not in a good shape." I agree, and offer some reasons why this is so, rooted in the culture and practice of Bayesian statistics as it stands today.

The Bayesian collaborator on a scientific project is often put in the position of asking, with all appropriate tact, why a particular prior has been chosen and whether something else might actually be a bit wiser. The experience is, I imagine, like working at a tattoo parlor: it widens your perspective about what kinds of poor choices are even possible.

For the R package maintainer, this experience must be magnified ten- or a hundred-fold. I suspect that Drs. Simpson et al. never would have imagined some of the things that people do with priors, until they undertook the job of writing and supporting an R package that does Bayesian inference for a wide class of models. I appreciate very much the authors' effort here to share their wisdom from the front lines of prior specification, and to formulate some general principles arising from this hard-won practical experience. I will organize my discussion of their article, which is both thought-provoking and excellent, around two broad questions that surround the practical art of prior specification.

DOES THE AUTHORS' PROPOSAL ADDRESS THE PROBLEM?

There is tremendous value in the authors' discussion of criteria for good default priors. Here, they identify many common mistakes, which to my eye have a common theme: choices that make the prior rather more informative than you intended. In particular, my vocabulary has been enriched by the concept of "forced overfitting," in which a default prior has the unintended consequence of rewarding a needlessly complex model. This is most obvious in the case of a variance component for random effects in a hierarchical model, a prototypical kind of nuisance parameter.

However, while it does not diminish my appreciation of the paper, I am not convinced about the "PC prior" formalism itself. For multivariate parameters, in particular, I have not yet been convinced that this formalism, or any other, is adequate to the task of answering the questions of prior choice that I have confronted in my recent scientific collaborations. More generally, I am leery of transferring intuitions gleaned from the scalar case to the high-dimensional case.

I will give a simple example. In Tansey et al. (2017), we describe an application in which the goal is to estimate the background radiation intensity across a wide spatial area. The details are unimportant here, but the essence is this: we have an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ that describes spatial adjacencies among locations, and a parameter $\{\theta^{(s)} : s \in \mathcal{V}\}$ at each node in the graph, parametrizing the background radiation at that location. To estimate θ, we used a prior that penalizes first-differences across edges in the graph:

$$p(\theta) \propto \prod_{(r,s)\in\mathcal{E}} p\big(\theta^{(r)} - \theta^{(s)} \mid \tau\big),$$

where τ is a precision parameter. If $p(\cdot)$ is a Gaussian distribution, then this is a traditional Gaussian Markov random field (specifically, an intrinsic CAR prior). This fits in the class of random-effects models described in the authors' Section 3.3, and we could therefore have used equation (3.3) as a prior for τ in a Gaussian CAR model.

But instead, we chose a Laplace prior for these first differences. Why? For several reasons. First, the Laplace prior leads to a nonlinear spatial smoother that adapts to different degrees of smoothness in different regions of the graph, which our situation called for. The Gaussian CAR prior, on the other hand, leads to linear shrinkage, which has important consequences

James G. Scott is Associate Professor of Statistics, Department of Information, Risk, and Operations Management and Department of Statistics and Data Sciences, University of Texas at Austin, 2110 Speedway, Box B6500, Austin, Texas 78712, USA (e-mail: james.scott@mccombs.utexas.edu).

for estimation accuracy, both theoretically and practically; see, for example, Sadhanala, Wang and Tibshirani (2016) and Donoho and Johnstone (1998). Second, the Laplace prior was better at handling discontinuities in our spatial field, due to the presence of line-of-sight occlusions that scattered radiation even from nearby sources. Finally, even in situations where the Gaussian prior had better *average* performance across the graph, the Laplace prior nearly always had better *worst-case* performance, that is, better accuracy for the nodes that were the hardest to estimate. For our application, this mattered a lot, because the performance of the whole radiation-monitoring system was limited by its weakest spatial link.

The authors, of course, explicitly endorse this kind of reasoning, when they write: "As always, if the user knows of a better prior for their case, then they should use it." By relating this example, I do not mean to imply that the authors would disagree with the particulars of our model. I merely wish to draw out two points.

The first point is that formal prior-specification rules may encourage us to focus on the wrong question, because doing so allows a mental shortcut. For our problem, the PC formalism provides an immediately available answer to the question: what should the prior be for the precision in a Gaussian CAR model? We are therefore nudged in the direction of *asking* this question, even in a situation like ours, where the relevant question is: should the prior be Gaussian in the first place? (No, it should not, and not because of the criteria that motivate PC priors.) To be sure, the expert statistician will probably have the wisdom to ask the second question, and the design language to answer it. But most statistical analyses are done by nonexperts.

In my experience, these first-order questions about functional form are much more important than questions about scalar hyperparameters, to which the PC prior admittedly provides an appealing answer. I do appreciate that the multivariate extension of the PC construction can provide an answer to the question of functional form. But is it a good answer? The crux of the matter seems to be the condition of constant rate penalization; this seems just as reasonable *a priori* as many other reasonable conditions leading to very different formalisms, which I mean as a criticism. There is also the matter of the user's choice of a "sensible" scale parameter arising from Principle 4. In Sections 6 and 7 of the paper, I see many examples of how we *could* use the PC formalism for a multivariate parameter. I do not see any convincing evidence, however, that we *should*, or that users are any better at specifying this scale than they are at specifying the scale of the original parameter.

The second point is that good prior specification is more like engineering than like math. (I expect that the authors would agree, although I hope they will correct me if not.) To be sure, there are many principles that cut across different areas of engineering, whether of airplanes, circuit boards or software. For example:

- Clear integration with a given context. A well-engineered product fits into its environment with no confusion. (A bridge is for cars, and has a clearly posted height and weight limit; Twitter's API streams tweets, and has a documented interface for accessing it.)
- Testability. There are clear and repeatable tests that can be performed to continuously validate whether a product behaves as intended. (Airplanes are put in wind tunnels; software, through unit tests.)
- Maintainability. The product is designed to last, but also has room to grow or change as the future may require. (Televisions come with external ports; modern cars have software updates.)

But it does not follow that these general principles will be embodied in precisely the same way in all engineering contexts. Electrical engineers use different math and a different design language than aerospace engineers, because they are responding to different problems.

To my mind, statistical models for analyzing fMRI data are at least as different from those for financial time series as circuit boards are from airplanes. The same general design principles apply, but the hard part is how these principles interact with the specifics of a problem. Drs. Simpson et al. have done us a great service in collecting many general design principles for priors in one place, and I have no doubt that I will refer to this list many times over the ensuing years. But the PC formalism seems too tidy to match my own experience, which has left me skeptical that these principles can ever be codified in some universally applicable formal rule.

WHY DO WE HAVE THIS PROBLEM IN THE FIRST PLACE?

The authors end their paper with the observation that "the current practice of prior specification is not in a good shape." I agree, and offer a simple reason for this: Bayesians do not write software. Of course, there are notable exceptions, including the authors; I speak generally.

To quantify this, I visited the "Advance publication" page of the *Bayesian Analysis* journal website in early September 2016. I clicked through to all 36 articles that were available ahead of their formal publication. I searched for the words "code," "software," "library" and "package" in the full text of every article, in an attempt to discern whether it was accompanied by publicly available software. This informal survey captured only a moment in time of Bayesian research, but a broad cross-section of topics, from variable selection to approximate Bayesian computation, and from phylogenetics to climate modeling. Here is what I found:

- 3 papers with public software, including documentation and a web page.
- 2 papers with undocumented R scripts and no data, or simulated data only.
- 31 papers with no code provided in any form.

Of the three software packages I did find, one was an R package that was available only through a link on an author's website, and not through any of the usual venues that are a proxy for quality control, like CRAN, BioConductor or even GitHub. A second was in Java and the third in Julia, neither of which are in widespread use by the statistics community. Thus the final tally stands: 36 papers, 0 R packages available through the usual channels and 0 Python modules.

I grant that my convenience sample may contain the odd theory paper, for which software could be irrelevant. Nonetheless, this is a lamentable record for a field that purports to invent useful methodology for data analysis. It seems that most Bayesians would rather be star architects than engineers, producing pretty drawings of buildings and leaving others to figure out how to actually build them so that the roof does not leak.[1] Let me be the first to say that I am guilty, too: I have written two papers that appeared in *Bayesian Analysis*, neither of which had software. One had theorems, at least; the other, in retrospect, has had no lasting value, which is at least partially due to the lack of software.

What does this have to do with priors? Everything. In my experience, if you write software for some statistical approach you have proposed, you are forced to think carefully about how someone will use it, in the same way that an aerospace engineer must build a plane with the idea that someone will fly it. But if you do not write software for public use, you can get by with sloppy, narrow thinking—about priors, and about many other things besides, like how efficient and readable your code is, how robust your overall approach is, whether your experiments can be reproduced and whether there is even a nail out there for the fancy hammer you have made.

Some methodological fields, like natural language processing, operate under the mantra: no software, no paper. We do not. Most editors and referees do not expect that papers have software, or reward those that do. Most departments do not hire or promote people based on the software they have written. Therefore, because writing good software is hard, people tend to not do it. These cultural tendencies excuse Bayesians from engaging seriously with prior specification as an act of engineering, rather than an act of mathematics. They also prevent us from playing a meaningful role in the vast majority of all scientific data analyses, which involve complicated "pipelines" of different statistical models joined together, and which require software that is robust, efficient, and stable (as well as priors that will not mess anything up).

Nothing but our collective inertia prevents this culture from changing overnight. Imagine if we took some basic and obvious steps as a field—as editors, as referees, as members of hiring committees, as teachers, and as scholars—to make this happen. My conjecture is that mastery of the practical art of prior specification—in other words, the wisdom and experience to engineer priors that are fit for purpose, scientifically speaking—would become much more widespread, and this whole conversation would seem quaint.

In closing, I embrace the tradition of what the authors call "risk-averse" priors, anchored in scientific consensus. (After all, conservatism is an engineering design principle, too.) I simply believe that we should work to nudge whatever consensus may exist in some field toward a better, more informed place. Thus my solution to this problem sounds a lot simpler than the authors', but is actually harder. Talk to scientists. Publish in their journals. Write software, so that others can reproduce your work and use your method in their scientific data-analysis pipelines. In short, do what the authors of this paper do. The priors will take care of themselves—not by magic, and not because of a formal rule, but because of what you will learn.

[1] See: "M.I.T. Sues Frank Gehry, Citing Flaws in Center He Designed," by Robin Pogrebin and Katie Zezima. *The New York Times*, November 7, 2007, page A20.

REFERENCES

DONOHO, D. L. and JOHNSTONE, I. M. (1998). Minimax estimation via wavelet shrinkage. *Ann. Statist.* **26** 879–921. MR1635414

SADHANALA, V., WANG, Y.-X. and TIBSHIRANI, R. J. (2016). Total variation classes beyond 1d: Minimax rates, and the limitations of linear smoothers. Available at arXiv:1605.08400.

TANSEY, W., REINHART, A., ATHEY, A. and SCOTT, J. G. (2017). Multiscale spatial density estimation: An application to large-scale radiological survey and anomaly detection. *J. Amer. Statist. Assoc.* To appear.