Statistical Science
2017, Vol. 32, No. 1, 44–46
DOI: 10.1214/17-STS576REJ
Main article DOI: 10.1214/16-STS576

You Just Keep on Pushing My Love over the Borderline: A Rejoinder

Daniel Simpson, Håvard Rue, Andrea Riebler, Thiago G. Martins and Sigrunn H. Sørbye

The entire reason that we wrote this paper was to provide a concrete object around which to focus a broader discussion about prior choice and we are extremely grateful to the editorial team at *Statistical Science* for this opportunity. David Dunson (DD), Jim Hodges (JH), Christian Robert, Judith Rousseau (RR) and James Scott (JS) have taken this discussion in diverse and challenging directions and over the next few pages, we will try to respond to the main points they have raised.

1. "IF I COULD LOVE, I WOULD LOVE YOU ALL."—KIKI DURANE

The point of departure for our paper is that most modern statistical models are built to be flexible enough to model diverse data generating mechanisms. Good statistical practice requires us to limit this flexibility, which is typically controlled by a small number of parameters, to the amount "needed" to model the data at hand. The Bayesian framework provides a natural method for doing this although, as DD points out, this trend for penalising model complexity casts a broad shadow over all of modern statistics and data science.

The PC prior framework argues for setting priors on these flexibility parameters that are specifically built to penalise a certain type of complexity and avoid overfitting. The discussants raised various points about this core idea. First, DD pointed out that while over-fitting a model is a bad thing, under-fitting is not better: we do not want Occam's razor to slit our throat. We saw this behaviour when using a half-Normal prior on the distance, while the exponential prior does not lead to obvious attenuation of the estimates. This is confirmed experimentally by Klein and Kneib (2016).

Both DD and RR note our focus on a specific parameterisation and DD (as well as a large number of reviewers) note that our informal definition of overfitting is parameterisation dependent. We did this on purpose: most people who use complex statistical models do not understand prior mass conditions in terms of Kullback–Leibler balls and the theoretical results in the paper do not justify this level of mathematical sophistication. Our choice to sacrifice generality (and annoy reviewers) in the search for a clear exposition has lead us to a revelation: we can replace questions about prior choice with questions about parameterisation. This leads us to re-phrase DD's implied question: *How should we parameterise a flexibility parameter so that we can use an exponential prior?*

The parameterisation we chose was

$$d(\xi) = \sqrt{2 \int f_\xi(\boldsymbol{x}) \log\left(\frac{f_\xi(\boldsymbol{x})}{f_0(\boldsymbol{x})}\right) d\boldsymbol{x}},$$

where ξ is the original flexibility parameter indexing model f_ξ and f_0 is the base model. JH correctly tweaks our nose over our inability to communicate this distance in a meaningful way (a heinous sin for people who abandoned measure theory in a quest for clarity). While we personally find our interpretation—$d(\xi)$ is the amount of information you lose by abandoning the flexible component in favour of the base model—appealing, it is a bit dry and abstract. JH suggests communicating the distance by considering how much a coin would be weighted to achieve that distance from a fair coin. While we agree that some sort of physical analogy would be appealing (see Roos et al., 2015, for work in this direction), we think that there is still some distance to go. For instance, the fairly

Daniel Simpson is a Reader in Statistics, Department of Mathematical Sciences, University of Bath, Bath, BA2 7AY, United Kingdom (e-mail: d.simpson@bath.ac.uk). Håvard Rue is a Professor of Statistics, CEMSE Division, King Abdullah University of Science and Technology, Thuwal 23955-6900, Saudi Arabia (e-mail: hrue@math.ntnu.no). Andrea Riebler is an Associate Professor, Department of Mathematical Sciences, NTNU, Norway (e-mail: andrea.riebler@math.ntnu.no). Thiago G. Martins was a PhD student, Department of Mathematical Sciences, NTNU, Norway (e-mail: thigm85@gmail.com). Sigrunn H. Sørbye is an Associate Professor, Department of Mathematics and Statistics, UiT The Arctic University of Norway, Norway (e-mail: sigrunn.sorbye@uit.no).

stringent condition that $\sigma < 1$ for a Gaussian random effect would be mapped to a weighted coin with the alarming property that the probability of getting a head was less than 96%. A different option is to note that $2\,\mathrm{KLD}(N(0,1) \parallel N(\mu,1)) = \mu^2$, and interpret the distance in terms of a changing mean of a Gaussian. This still fails to communicate the asymmetry of the distance measure.

To conclude this mini-tour of parameterisations, we can address RR's question of why chose one particular direction for the Kullback–Leibler divergence. They partially answer the question themselves: much like Variational Bayes, it just does not work the other way. Perhaps a more satisfying justification would be to recall the early method of building shrinkage estimators through "testimation" (Brewster and Zidek, 1974). This proceeded by first performing a hypothesis test to see if the data was drawn from the base model and the flexible model was only used if that null hypothesis was rejected. Our distance measure is very much in this spirit: we are asking what the penalty would be if we just used the base model instead of the more flexible machinery.

2. AN ARROW'S FLIGHT

You do not have to be Arrow to realise that we set ourselves an impossible task. It is not mathematically impossible to build a systematic method of prior specification from a small set of principles as long as you are also allowed to define what a good prior is: the theory underneath reference priors demonstrates this. So maybe all we need to do is find a sufficiently compelling concept of what a "good" prior is. Our desiderata come from a different direction. They provide tools to ask "is this existing prior good?" As RR point out, this does not lead to a useful mathematical construction for prior distributions. It is easy to say that a child should not play with a chainsaw; it is considerably harder to set the exact boundaries of what they should play with!

Does our set of principles lead to a universally good prior distributions? We have no idea. We do know that in all of the examples we have tried, PC priors work well. But, as both JH and RR point out, good experience does not a programme make. We also know that this prior encodes known pieces of prior information in a direct manner, which makes them communicable. And for the class of Structured Additive Regression models, Klein and Kneib (2016) demonstrated empirically that PC priors perform better than commonly used priors when the data generating mechanism is close to the base model or when there is little information about the parameter in question in the likelihood. It also performs no worse than commonly used priors when the model is far from the base model.

We have hit two problems with theoretically validating the PC prior process. We have found, as DD recognises, that statistical theory focusses almost exclusively on either classical models or a small class of modern nonparametric models. There are almost no mathematical tools developed for the types of hierarchical model we are interested in. The second problem is that we have been unable to come up with a mathematical way of validating models that matches practical use. While far from uninformative, asymptotic analysis is only of limited use to us. Le Cam's seventh (mildly tongue-in-cheek) principle is "If you need to use asymptotic arguments, do not forget to let your number of observations tend to infinity" (Le Cam, 1990). In all of the examples we consider, and the ones that we have in mind when building this system, the data does not push us into the asymptotic regime for all model components simultaneously and the priors are, therefore, important. In this situation, priors are (and should be) influential, but it is incredibly important to specify exactly what influence we want them to have. That has been our aim with this project: precise elucidation of the four principles and the model structure that encodes extra information that has been inserted into the model. With PC priors we did not shoot for optimality, instead we aimed for robustness and utility.

There is also the problem that priors act in concert with the other parts of the model. JS rightly suggests, no matter how good a prior is it usually will not overcome deficiencies in the modelling higher up the hierarchy. For example, while the PC prior for the scaling parameter in the Laplace prior is the same as the PC prior for the corresponding parameter in the multivariate Gaussian, nothing will fix the fact that putting a Laplace prior on the differences will, asymptotically, not preserve sharp changes (Lassas and Siltanen, 2004).

3. SOME HORSES ARE DESIGNED BY COMMITTEE

Sparsity crept into the paper like a thief in the night (less poetically, an early reviewer wanted us to com-

ment on sparsity). JH picked up on our ambivalence to the assigned topic. The only comment that we really could make (that if we measure complexity by the number of nonzero components, we should penalise it) is fairly asinine and uninspired. Sparsity is an important topic with a rich literature that our paper does not add much to.

When a reviewer gave us lemons, we tried to make lemon drizzle cake. Our aim was to argue that the structural assumptions that (1) ξ is a flexibility parameter and (2) the priors on the flexibility parameters can be set independently are restrictive. While later in the paper we outlined a method for relaxing the independence assumption, the key point remains: having a system for specifying prior distributions is not an invitation to ignore assumptions.

4. GIVE THE PEOPLE WHAT THEY WANT

Since writing this paper, PC priors have been derived an applied in a whole variety of situations. PC priors now exist for models of tail dependence (Kereszturi, Tawn and Jonathan, 2016), the Hurst parameter for fractional Gaussian noise (Sørbye and Rue, 2016a), the degrees of freedom for P-splines (Ventrucci and Rue, 2016), parameters in the Matérn covariance function (Fuglstad et al., 2015), the correlation parameter in bivariate meta-analysis models (Guo, Rue and Riebler, 2015), the autoregressive parameters in an AR(p) process (Sørbye and Rue, 2016b) and the variance in the mean-variance parameterisation of the Beta distribution (Harjanto et al., 2016).

To finish this response, we will answer JH's request for a PC prior for the over-dispersion parameter in a negative binomial distribution parameterised by its mean μ and variance $\mu+\alpha^{-1}\mu^2$. When the base model is Poisson with mean μ, the distance depends on μ. While the distance can be computed numerically, the μ-dependence makes it difficult to calibrate the prior. An alternative is to recall that $y \sim \text{NegBinom}(\mu, \phi)$ if $y \mid \varepsilon \sim \text{Po}(\varepsilon\mu)$, where $\varepsilon \sim \text{Gamma}(\phi^{-1}, \phi^{-1})$. Using $\varepsilon \equiv 1$ as the base model, the corresponding PC prior is

$$\pi(\phi) = \frac{\lambda}{\phi^2} \frac{\left|\psi'(\phi^{-1}) - \phi\right|}{\sqrt{2\log(\phi^{-1}) - 2\psi(\phi^{-1})}} \cdot \exp\left[-\lambda\sqrt{2\log(\phi^{-1}) - 2\psi(\phi^{-1})}\right],$$

where ψ is the digamma function and ψ' is its derivative. In the context of regression modelling of over-dispersed data, this construction can be justified as a prior on the variation of the expected number of counts rather than directly on the over-dispersion parameter.

REFERENCES

BREWSTER, J. F. and ZIDEK, J. V. (1974). Improving on equivariant estimators. *Ann. Statist.* **2** 21–38. MR0381098

FUGLSTAD, G.-A., SIMPSON, D., LINDGREN, F. and RUE, H. (2015). Interpretable priors for hyperparameters for Gaussian random fields. ArXiv Preprint ArXiv:1503.00256.

GUO, J., RUE, H. and RIEBLER, A. (2015). Bayesian bivariate meta-analysis of diagnostic test studies with interpretable priors. ArXiv Preprint ArXiv:1512.06217.

HARJANTO, D., PAPAMARKOU, T., OATES, C. J., RAYON-ESTRADA, V., PAPAVASILIOU, F. N. and PAPAVASILIOU, A. (2016). RNA editing generates cellular subsets with diverse sequence within populations. *Nat. Commun.* **7** 12145.

KERESZTURI, M., TAWN, J. and JONATHAN, P. (2016). Assessing extremal dependence of North Sea storm severity. *Ocean Eng.* **118** 242–259.

KLEIN, N. and KNEIB, T. (2016). Scale-dependent priors for variance parameters in structured additive distributional regression. *Bayesian Anal.* **11** 1071–1106. MR3545474

LASSAS, M. and SILTANEN, S. (2004). Can one use total variation prior for edge-preserving Bayesian inversion? *Inverse Probl.* **20** 1537–1563. MR2109134

LE CAM, L. (1990). Maximum likelihood: An introduction. *Int. Stat. Rev.* 153–171.

ROOS, M., MARTINS, T. G., HELD, L., RUE, H. et al. (2015). Sensitivity analysis for Bayesian hierarchical models. *Bayesian Anal.* **10** 321–349.

SØRBYE, S. H. and RUE, H. (2016a). Fractional Gaussian noise: Prior specification and model comparison. ArXiv Preprint ArXiv:1611.06399.

SØRBYE, S. H. and RUE, H. (2016b). Penalised complexity priors for stationary autoregressive processes. ArXiv Preprint ArXiv:1608.08941.

VENTRUCCI, M. and RUE, H. (2016). Penalized complexity priors for degrees of freedom in Bayesian P-splines. *Stat. Model.* **16** 429–453. MR3589051