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 over-fitting. 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 over-fitting 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

\[ d(\xi) = \sqrt{2 \int f_\xi(x) \log \left( \frac{f_\xi(x)}{f_0(x)} \right) dx}, \]

where \( \xi \) is the original flexibility parameter indexing model \( f_\xi \) 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
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 $2 \text{KLD}(N(0, 1) \parallel N(\mu, 1)) = \mu^2$, 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 play 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 do.

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 play 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 do.

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 Siltaonen, 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) $\xi$ 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 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), parameters in the Matérn covariance function (Harjanto et al., 2016), 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 $\mu$ and variance $\mu + \alpha^{-1} \mu^2$. When the base model is Poisson with mean $\mu$, the distance depends on $\mu$. While the distance can be computed numerically, the $\mu$-dependence makes it difficult to calibrate the prior. An alternative is to recall that $y \sim \text{NegBinom}(\mu, \phi)$ if $y \mid \epsilon \sim \text{Po}(\epsilon \mu)$, where $\epsilon \sim \text{Gamma}(\phi^{-1}, \phi^{-1})$. Using $\epsilon \equiv 1$ as the base model, the corresponding PC prior is

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

where $\psi$ is the digamma function and $\psi'$ 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


