Statistical Science
2017, Vol. 32, No. 1, 33–35
DOI: 10.1214/16-STS601
Main article DOI: 10.1214/16-STS576

Swinging for the Fence in a League Where Everyone Bunts

James S. Hodges

I enjoyed this paper *very* much though sometimes it felt too rich, like eating an entire sheet of fudge. Many aspects of it deserve comment; I will discuss this paper as an example of the value of system building and then note what seems to be a missed opportunity.

1. THE VALUE OF SYSTEM-BUILDING

I congratulate the authors especially for building a system instead of devising yet another salami slice. Statistics does not have enough system-building and I do not mean theory-for-theory's-sake systems, like decision theory came to be, but rather systems built for practical purposes. The only recent examples that come to mind are computing systems like R, WinBUGS, JAGS and the authors' own INLA. To use a baseball analogy, it is refreshing to see the authors swing for the fence in an academic incentive system that almost forces people to bunt.[1]

An important virtue of system-building is that it bears fruit beyond the immediate products, which in this case are prior distributions. To build a system, you assemble tentative principles based on examples and what seems like good sense, then refine the system by applying it to more examples. After a while, the system merits enough confidence that when something odd happens in an example, it is permissible to question whether the oddity arose from an error in the customary way of thinking rather than from a flaw in the system. At this point, the system has begun to add value for problems besides those that motivated it. The danger, of course, is having too much confidence in your system, becoming an ideologue, and thus a menace. As I will argue below, however, the authors seem to have too little, not too much, confidence.

One example of using the system to challenge customary thinking is Section 5, which reconsiders the Besag–York–Mollié model. Based on Desideratum D2, the authors argue that the spatial and heterogenity components "cannot be seen independently" so that their priors "should . . . not [be] independent as . . . usually assumed." They implement this by re-parameterizing from the usual two parameters, one controlling each component, to a parameter controlling total precision in the prior and a parameter allocating total precision between the spatial and heterogeneity components. The authors are not the first to suggest such a parameterization for spatial models (e.g., Leroux, Lei and Breslow, 2000) or more generally (e.g., He et al., 2007) but I do think they are the first to show how this provides a convincing rationale for a prior.

In Section 8, the authors tantalize us by suggesting they could do something similar with the negative binomial distribution's over-dispersion parameter, replacing the now-standard parameterization (which I find uninterpretable) with the mean and variance-to-mean ratio. I encourage them to pursue this.

It does seem that the authors' confidence failed them for the sparsity priors example (Section 4.5), and I think they have done themselves an injustice. They begin by noting that the spike-and-slab prior has computing problems, then switch to "a more pleasant computational option [that] builds a prior on the scaling parameter of the individual model components," and treats them as independent. After some development, they say "does . . . the PC prior [for the independent-components formulation] make a good variable selection prior? . . . the answer is no. The problem with the basic PC prior . . . is that the base model has been incorrectly specified. The base model that a p-dimensional vector is sparse is not the same as the base model that each of the p components is independently zero, and hence the prior encodes the wrong information. A more correct application of [the authors'] principles . . . lead[s] to a PC prior that first selects the number of

James S. Hodges is Professor, Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, Minnesota 55455, USA (e-mail: hodge003@umn.edu).

[1]For those unfamiliar with baseball, to swing for the fence (i.e., try to hit a home run) is to try to accomplish much with one stroke, while a bunt is the minimum unit of aspiration. Although bunts have their place, a match consisting entirely of bunts would be, among other things, stupefyingly dull even by baseball's leisurely standard.

nonzero components and then puts i.i.d. PC priors on each of the selected components." Bravo! The authors' system has illuminated the problem. But later the authors conclude "The failure [!!] of PC priors to provide useful variable selection priors [arises from] the tails specified by the principle of constant rate penalisation. ... [T]his is the only situation we have encountered in which the exponential tails of PC priors are problematic." I would conclude that the authors' principles have not failed but rather that they have chosen to use a model they describe as wrong and then tried to force it to produce sensible answers. Have they done this because the "right" model (for which their system works) is too hard to compute? If so, is it not better to apply their system to the right model and then deal with the computing problem, instead of mutilating their system for the sake of a flawed model that is easy to compute? I encourage the authors to have more confidence and to follow their system's implications.

Although I do not think this example indicates an important flaw in the authors' system, it may be that someone will prove an analog to Arrow's Impossibility Theorem, which showed that it is impossible to construct a public-choice scheme satisfying a short list of reasonable-sounding criteria. The analogy is that the authors have proposed a short list of reasonable-sounding criteria for priors and we may end up concluding that no such scheme can exist. Arrow's theorem was, however, not a terminus but rather the beginning of a hugely fruitful area of inquiry. Thus even if the authors' system ultimately fails on its own ambitious terms, it could still be the beginning of a rich vein of work.

2. THE DISTANCE MEASURE d

I do have one problem with this paper, though I am optimistic that it is not an inherent flaw but merely a missed opportunity.

Principle 2 founds the authors' approach on a distance measure d based on the Kullback–Leibler divergence (KLD) of model f from model g, $d(f \parallel g) = \sqrt{2\,\mathrm{KLD}(f \parallel g)}$. In an earlier version of this paper, the authors said d is "a physically interpretable 'distance' scale." They have dropped that wording but d is still the key to their system, although the authors are vague about the extent to which d is essential to Principle 4, user-defined scaling, by burying d in $Q(\xi)$, "an interpretable transformation of the flexibility parameter." The crux is "interpretable": the authors have deliberately *not* interpreted d—and I know this because as a referee I urged them to interpret it—and by doing so, they may be missing an opportunity.

Instead of interpreting d, the authors interpret model parameters of which d happens to be a fairly simple function. One problem with this is that we cannot reasonably hope d will always be a fairly simple function of model parameters. A more immediate problem (perhaps just an apparent problem—more on this below) is that the model parameters the authors interpret are sometimes on arbitrary scales, for example, the standard deviation of a random effect implementing a penalized spline or an ICAR model. In such cases, intuition about the parameters is arguably impossible and certainly difficult. This was my main motivation for proposing priors on the degrees of freedom in a fit (which the authors cited in discussing Desideratum 3) though as the authors note, using degrees of freedom this way has the disadvantage of being far less general than KLD. However, by interpreting model parameters instead of d, the authors sacrifice part of the gain from using KLD and require users to do something new for every model.

(I said, "perhaps ... an apparent problem" because the authors try to solve this problem using scaling tricks that, frankly, I do not understand. Even if this problem truly is apparent and not real, working with model parameters instead of d still has the disadvantages mentioned above.)

To use d directly, the authors might consider investing in a body of interpretation and intuition for d. Some pertinent literature exists; I know only a bit of it, Section 2 of McCulloch (1989), which interprets KLD using simple models. For example, $\mathrm{KLD} = k$ corresponds to the divergence between a Bernoulli with probability 0.5 and another Bernoulli with probability $0.5(1 + (1 - e^{-2k})^{0.5})$. One might hope that even people with fairly modest statistical training and experience could have intuition about d based on models like this. I am nowhere near an expert in this literature and perhaps knowledgeable people have given up on interpreting KLD, and hence d. If so, this seems to be a weakness of the authors' system but if not, this line of inquiry could be worth pursuing.

Once again, I congratulate the authors on a virtuoso performance, which I expect will become a milestone in the literature on prior distributions.

REFERENCES

LEROUX, B. G., LEI, X. and BRESLOW, N. (2000). Estimation of disease rates in small areas: A new mixed model for spatial de-

pendence. In *Statistical Models in Epidemiology, the Environment, and Clinical Trials* (*Minneapolis, MN*, 1997). *IMA Vol. Math. Appl.* **116** 179–191. Springer, New York. MR1731684

MCCULLOCH, R. E. (1989). Local model influence. *J. Amer. Statist. Assoc.* **84** 473–478.