TEST (2015) 24:29–30
DOI 10.1007/s11749-014-0412-4

DISCUSSION

Comments on: Comparing and selecting spatial predictors using local criteria

Alan E. Gelfand

Published online: 29 November 2014

This is a timely and well-targeted paper in the sense that, currently, there is a substantial escalation in the size of spatial datasets being collected and that the primary objective for many of these datasets is spatial prediction rather than explanation. Moreover, the flexibility of the authors' approach, accommodating an arbitrary number of arbitrarily selected predictors, treats a usually challenging "apples and oranges" problem. By the same token, for me, this suggests that the applicability for this paper resides primarily in bigger data settings.

That is, with datasets of order at most 10^3, as a statistician, full stochastic modeling can be implemented and I would see no reason to settle for less. The inferential limitations of the authors' approach would leave me less than satisfied. I need to be able to assess uncertainty; and I need to know something about the variance associated with my predictions. For me, this would not be a version of predictive mean square error (PMSE). In the absence of a stochastic specification, one cannot really calibrate "how much better" one predictor is compared with another using PMSE. (This connects to points below.) Furthermore, I would not find PMSE (or other "observed" vs. "expected" criteria) to be a satisfying way to compare models. (Here, I am thinking stochastically, in terms of predictive distributions that need not be unimodal). I would want to see something regarding empirical coverage of predictive intervals, or, even better, perhaps some version of a proper scoring rule comparison like a continuous rank probability score.

I must confess to discomfort with the arbitrariness of the choice of D^{val}. The authors do discuss this issue but, unlike in customary cross-validation, it seems that

This comment refers to the invited paper available at doi:10.1007/s11749-014-0415-1.

A. E. Gelfand (✉)
Department of Statistical Science, Duke University, Durham, USA
e-mail: alan@stat.duke.edu

model comparison will be more sensitive to the choice here. In fact, it seems we need the validation data to calculate the locally selected predictor (LSP) across all of D. Usually, we only employ the fitting data to obtain the predictor over D; the authors' definition of the LSP seems to revise the interpretation of validation data as well as the sample size for creating the estimator. Why should the value of the predictor and thus the performance of the predictor depend on this choice. Furthermore, it would seem evident that overfitting becomes a potential concern. However, because the estimator requires the validation data to be calculated, this issue is obscured. Do the authors think overfitting is something to worry about? Also, customarily, the model performance criterion is calculated using the hold out data. Typically, we do not create a third set for "testing", a set which adds further arbitrariness.

While the authors' approach does not limit the number of or nature of the predictors, this freedom raises some unclear issues. How should one choose the number of predictors to work with? (This would seem to connect to the issue of overfitting.) For instance, it is easy to imagine arbitrarily many predictors of a particular type, e.g., with usual kriging, by changing the covariance function. So, in developing the set of predictors to apply the LSP strategy to, we can work with many from a single class or a single from each of many classes (the authors' illustration) or all sorts of combinations in between. How would this balancing affect the resulting LSP? In different words, with LSP, in principle, we can identify which type of predictor worked best overall (was chosen by the most test locations), which type worked best in which part of D (assessments of interest to the authors). However, with different mixes of predictors, this would seem to muddy such determination.

Finally, I am curious what motivates the mean structure in the simulation specification in expression (18). Would not a more explicit form be easier to interpret with regard to performance? Also, I missed the reason why there are no proposed test locations in the simulation example while they are introduced with the real data.

In summary, my comments are not intended to be too critical. Rather, this paper is an example of a personal current struggle with the future of the field of Statistics. Perhaps I am an old stick in the mud, but, for me, Statistics is about inference within a probabilistic framework, not just developing an algorithm to obtain a good predictor. With a collection of models, for prediction, I would prefer to do some version of Bayesian Model Averaging. I am sure the authors will say that we have to pay a price to handle sufficiently large datasets and, though I remain unsatisfied, perhaps their approach is worth the price.