Submitted to the Statistical Science

Discussion of Big Bayes stories, and BayesBag

Peter Bühlmann

ETH Zürich

1. INTRODUCTORY REMARKS

I congratulate all the authors for their insightful papers with wide-ranging contributions. The articles demonstrate the power and elegance of the Bayesian inference paradigm. In particular, it allows to incorporate prior knowledge as well as hierarchical model building in a convincing way. Regarding the latter, the contribution by Raftery, Alkema and German is a very fascinating piece as it addresses a set of problems of great public interest and presents predictions for the world populations and other interesting quantities with uncertainty regions. Their approach is based on a hierarchical model, taking various characteristics into account (e.g. fertility projections). It would have been very difficult to come up with a "better" solution which would be as clear in terms of interpretation (in contrast to a "black-box machine") and which would provide (model-based) uncertainties for the predictions into the future.

2. UNCERTAINTY, STABILITY AND BAGGING THE POSTERIOR

Many of the papers quantify in one or another form various notions of uncertainties. In the Bayesian framework, this is usually based on the posterior distribution. An old "debate" is how much the results are sensitive to the choice of the prior, and I believe that some reasonable sensitivity analysis can lead to much insight. The sensitivity with respect to "perturbed data" though is not easily captured by the Bayesian framework. In the context of prediction, Leo Breiman (Breiman, 1996a,b) has pointed to issues of stability with respect to perturbations of the data, Bousquet and Elisseeff (2002) provide some mathematical connections to prediction performance while Meinshausen and Bühlmann (2010) present some theory and methodology for controlling the frequentist error of expected false positives.

As an example, the (frequentist) Lasso (Tibshirani, 1996) is very unstable for estimating the unknown parameters in a linear model, in particular if the correlation among the covariates is high (for two highly correlated variables where at least one of them has a substantially large regression coefficient, the Lasso selects either one or the other in an unstable fashion). Thus, the MAP for a Gaussian linear model with a Double-Exponential prior for the regression coefficients is unstable. The posterior distribution is probably more stable but presumably, it is still "rather" sensitive with respect to perturbation of the data: if the data would look a bit different, the posterior might be "rather" different. The situa-

Seminar for Statistics, ETH Zürich, CH-8092 Zürich, Switzerland (e-mail: buhlmann@stat.math.ethz.ch).

tion becomes more exposed to stability problems when using spike and slab priors (Mitchell and Beauchamp, 1988), due to increased sparsity.

We can stabilize the posterior distribution by using a bootstrap and aggregation scheme, in the spirit of bagging (Breiman, 1996b). In a nutshell, denote by $\mathcal{D}^*$ a bootstrap- or sub-sample of the data $\mathcal{D}$. The posterior of the random parameters θ given the data $\mathcal{D}$ has c.d.f. $F(\cdot|\mathcal{D})$, and we can stabilize this using

$$F_{\text{BayesBag}}(\cdot|\mathcal{D}) = \mathbb{E}^*[F(\cdot|\mathcal{D}^*)],$$

where $\mathbb{E}^*$ is with respect to the bootstrap- or sub-sampling scheme. We call it the *BayesBag* estimator. It can be approximated by averaging over B posterior computations for bootstrap- or sub-samples, which might be a rather demanding task (although say $B = 10$ would already stabilize to a certain extent). Note that when conditioning on the data, the posterior $F(\cdot|\mathcal{D})$ is a fixed c.d.f. but when taking the view point that the data could change, it is useful to consider randomized perturbed versions $F(\cdot|\mathcal{D}^*)$ which are to be aggregated.

The following simple and rather stable example shows that such a bagging scheme outputs a larger uncertainty which is perhaps more appropriate.

Location model with conjugate Gaussian prior. Consider the model

$$\begin{aligned} &\theta \sim \mathcal{N}(0, \tau^2), \\ \text{conditional on } \theta : \quad &X_1, \ldots, X_n \text{ i.i.d. } \sim \mathcal{N}(\theta, \sigma^2). \end{aligned}$$

It is well known that the posterior distribution equals

$$\theta|\overline{X}_n \sim \mathcal{N}(\frac{\sum_{i=1}^n X_i}{n + \sigma^2/\tau^2}, (\frac{1}{\tau^2} + \frac{n}{\sigma^2})^{-1}).$$

Denote by $F(\cdot; \overline{X}_n)$ the c.d.f. of the posterior distribution, i.e.,

$$F(u; \overline{X}_n) = \Phi(u, \text{mean} = \frac{n\overline{X}_n}{n + \sigma^2/\tau^2}, \text{var} = (\frac{1}{\tau^2} + \frac{n}{\sigma^2})^{-1}), \tag{2.1}$$

where $\Phi(u, \text{mean} = m, \text{var} = s^2) = \Phi((u - m)/s)$ and $\Phi(\cdot)$ denotes the c.d.f. of $\mathcal{N}(0, 1)$. We can either use the nonparametric bootstrap, with resampling the data with replacement, or a parametric bootstrap (assuming here that σ^2 is known):

$$X_1^*, \ldots, X_n^* \text{ i.i.d. } \mathcal{N}(\hat{\theta}, \sigma^2),\ \hat{\theta} = \overline{X}_n. \tag{2.2}$$

With the parametric bootstrap in (2.2), we can easily calculate the *BayesBag* estimator:

$$\begin{aligned} &\mathbb{E}^*[F(u; \overline{X}_n^*)] \\ &= \int \Phi(\frac{u - r}{\sqrt{(\frac{1}{\tau^2} + \frac{n}{\sigma^2})^{-1}}})\varphi(r, \text{mean} = \frac{n\overline{X}_n}{n + \sigma^2/\tau^2}, \text{var} = \frac{n\sigma^2}{(n + \sigma^2/\tau^2)^2})dr, \end{aligned} \tag{2.3}$$

where $\varphi(r, \text{mean} = m, \text{var} = s^2) = s^{-1}\varphi((r - m)/s)$ and $\varphi(\cdot)$ denotes the p.d.f. of $\mathcal{N}(0, 1)$. We consider the posterior credible region by computing the 2.5% and 97.5% quantiles of $F(\cdot; \overline{X}_n)$ and we compare these quantiles with the corresponding ones from the BayesBag $\mathbb{E}^*[F(\cdot; \overline{X}_n^*)]$ above in (2.3). We only consider here

sample size	(2.5%,97.5%) posterior	(2.5%,97.5%) *BayesBag*
$n = 1$	(-0.69, 2.81)	(-1.30, 3.41)
$n = 10$	(0.10, 1.32)	(-0.16, 1.56)

TABLE 1

2.5% and 97.5% quantiles of the posterior $F(\cdot|\overline{X}_n)$ in (2.1) and of the BayesBag (bagged posterior) in (2.3). The data was generated once using a single realized value of $\theta = 1.31$.

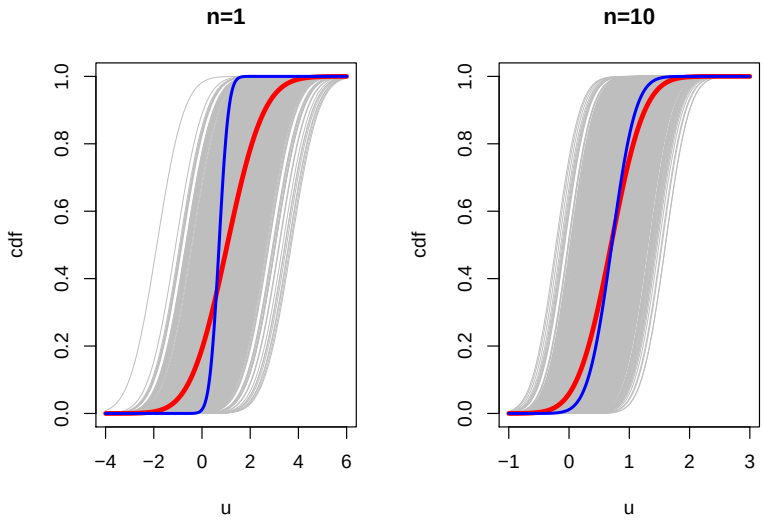

FIG 1. *1000 bootstrapped cumulative distribution functions $F(u|\overline{X}_n^*)$ of $\theta|\overline{X}_n^*$. The BayesBag (i.e., mean) $\mathbb{E}^*[F(u|\overline{X}_n^*)]$ in (2.3) (thick red line) and the cumulative distribution function $F(u|\overline{X}_n)$ of the classical posterior of $\theta|\overline{X}_n$ in (2.1) (blue line). Left panel for $n = 1$ and right panel for $n = 10$, and note the different scales for the x-axis. The data is as in Table 1.*

the case with $\sigma^2 = 1$ and $\tau^2 = 4$, and the results are given in Table 1. Of course, we can also look at the variability of the posterior via the bootstrapped c.d.f.'s $F(\cdot|\overline{X}_n^*)$, instead of considering the bootstrap mean (*BayesBag*) only. Figure 1 illustrates that variability can be rather high, but the situation obviously improves as sample size increases.

It is worth pointing out that in general, one could use a parametric bootstrap when using $\hat{\theta}$ as the MAP of the posterior distribution, and such a scheme could be used in models with complex hierarchical and dependence structures.

The frequentist approach usually does not address stability issues either and in addition, assigning p-values and confidence intervals in complex scenarios is a non-trivial problem. Recent progress has been achieved for high-dimensional sparse models (Minnier et al., 2011; Bühlmann, 2013; Zhang and Zhang, 2011; van de Geer et al., 2013; Bogdan et al., 2013, cf.); regarding the issue of constructing "stable p-values", an approach based on sub-sampling and appropriate aggregation of p-values is described in Meinshausen et al. (2009). Yet, much more work in frequentist inference would be needed to cope with e.g. high-dimensional hierarchical models in non-i.i.d. settings such as space-time processes or clustered data, or as another example, the population dynamic model in the beautiful paper by Kuikka, Vanhatalo, Pulkkinen, Mäntyniemi and Corander in this issue.

3. NETWORKS AND GRAPHICAL MODELS

The paper by Johnson, Abal, Ahern an Hamilton presents an interesting application by using Bayesian inference for a Bayesian networks (as is well known, the term "Bayesian network" does not require Bayesian inference at all – and it is somewhat confusing). The arrows in the directed acyclic graph often indicate causal relations (Spirtes et al., 2000; Pearl, 2000) and as such, the model allows for causal conclusions. Great care is needed, of course, when the DAG is mis-specified or estimated from observational data since causal conclusions are depending in a very "sensitive way" on the underlying DAG. A lot of work exists regarding identifiability of the DAG from observational data (Spirtes et al., 2000; Pearl, 2000; Shpitser and Pearl, 2008; Hoyer et al., 2009; Peters and Bühlmann, 2013, cf.), and obviously, there are ill-posed situations such as with a bivariate Gaussian distribution where one cannot identify the causal direction between two variables. In the Bayesian framework, the problem of identifiability does not exist in a "direct sense": but I believe it must come in through another channel, presumably by a high sensitivity with respect to prior specifications. Due to severe identifiability problems, causal inference based on observational data is ill-posed or depends on non-testable assumptions. However, one can nevertheless (under some assumptions) derive lower bounds on absolute values of causal effects (Maathuis et al., 2009). As lower bounds, they are conservative and a Bayesian average bound would be interesting.

In view of non-testable assumptions, causal models should be validated with randomized experiments. Often though, this cannot be done due to limited resources or ethical reasons. The field of molecular biology with simple organisms is an interesting application where causal model validation is feasible thanks to gene knock-out or other manipulation methods. We pursued this in the past, for estimated causal structures and models based on frequentist approaches, for the organisms yeast (Maathuis et al., 2010) and arabidopsis thaliana (Stekhoven et al., 2012). These two papers indicate that it is indeed possible to predict to a certain extent lower bounds of causal strength and relations based on observational (and very high-dimensional) data. Such a conclusion can only be made post-hoc, after validation – and validation has nothing to do whether a Bayesian or any other inference machine has been used.

Acknowledgments. I thank Nicolai Meinshausen for interesting comments and suggesting the name *BayesBag*.

REFERENCES

Bogdan, M., van den Berg, E., Su, W., and Candès, E. (2013). Statistical estimation and testing via the sorted L1 norm. arXiv:1310.1969.

Bousquet, O. and Elisseeff, A. (2002). Stability and generalization. *Journal of Machine Learning Research*, 2:499–526.

Breiman, L. (1996a). Bagging predictors. *Machine Learning*, 24:123–140.

Breiman, L. (1996b). Heuristics of instability and stabilization in model selection. *Annals of Statistics*, 24:2350–2383.

Bühlmann, P. (2013). Statistical significance in high-dimensional linear models. *Bernoulli*, 19:1212–1242.

Hoyer, P., Janzing, D., Mooij, J., Peters, J., and Schölkopf, B. (2009). Nonlinear causal discovery with additive noise models. In *Advances in Neural Information Processing Systems 21, 21st Annual Conference on Neural Information Processing Systems (NIPS)*, pages 689–696.

Maathuis, M., Colombo, D., Kalisch, M., and Bühlmann, P. (2010). Predicting causal effects in large-scale systems from observational data. *Nature Methods*, 7:247–248.

Maathuis, M., Kalisch, M., and Bühlmann, P. (2009). Estimating high-dimensional intervention effects from observational data. *Annals of Statistics*, 37:3133–3164.

Meinshausen, N. and Bühlmann, P. (2010). Stability Selection (with discussion). *Journal of the Royal Statistical Society Series B*, 72:417–473.

Meinshausen, N., Meier, L., and Bühlmann, P. (2009). P-values for high-dimensional regression. *Journal of the American Statistical Association*, 104:1671–1681.

Minnier, J., Tian, L., and Cai, T. (2011). A perturbation method for inference on regularized regression estimates. *J. of the American Statistical Association*, 106:1371–1382.

Mitchell, T. and Beauchamp, J. (1988). Bayesan variable selection in linear regression. *Journal of the American Statistical Asssociation*, 83:1023–1032.

Pearl, J. (2000). *Causality: models, reasoning and inference*. Cambridge Univ. Press.

Peters, J. and Bühlmann, P. (2013). Identifiability of Gaussian structural equation models with equal error variances. *To appear in Biometrika; arXiv:1205.2536*.

Shpitser, I. and Pearl, J. (2008). Complete identification methods for the causal hierarchy. *Journal of Machine Learning Research*, 9:1941–1979.

Spirtes, P., Glymour, C., and Scheines, R. (2000). *Causation, Prediction, and Search*. MIT Press, second edition.

Stekhoven, D., Moraes, I., Sveinbjörnsson, G., Hennig, L., Maathuis, M., and Bühlmann, P. (2012). Causal stability ranking. *Bioinformatics*, 28:2819–2823.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society Series B*, 58:267–288.

van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2013). On asymptotically optimal confidence regions and tests for high-dimensional models. arXiv:1303.0518.

Zhang, C.-H. and Zhang, S. (2011). Confidence intervals for low-dimensional parameters with high-dimensional data. *To appear in the Journal of the Royal Statistical Society, Series B; arXiv:1110.2563*.