Bayesian Analysis (2010) **5**, Number 3, pp. 451–456

Comment on Article by Manolopoulou et al.

Fabio Rigat[*]

Abstract. We approach a discussion of this paper by asking: to what extent the specific models and computational techniques used succeed in bridging data and research questions? This discussion is mostly seen as an opportunity to pose useful questions rather than providing definitive answers, pointing to relevant literature when necessary.

Introduction

Experimental science has traditionally advanced through relatively well-posed questions, which inspired the construction of measurement technologies and the development of theories explaining consistently observed patterns (Kuhn (1962)). A modeler's job here is naturally integrated in and constrained by the process of scientific research from the design of an experiment to the formulation of stochastic models characterising measurement errors and uncontrolled variation. The rapid collection of multivariate data, starting in the 1990s, increases the complexity of this scenario. This radical change is one of the most relevant markers of the quantitative revolution in the natural sciences, in particular in genetics (Golub et al. (1999), Sebastiani et al. (2003), Jung et al. (2008), Stephens and Balding (2009)). As the experimentalist is inundated by noisy measurements, the theorist can't keep up developing deterministic reference frameworks encompassing large networks of relations involving different data dimensions. This paper illustrates the role of the Bayesian statistician as the necessary third pillar of the learning process in the data-rich scenario of contemporary science.

1 How many parameters?

Since Bayesian nonparametrics came to age, with the marriage to suitable numerical approximation techniques such as Markov chain Monte Carlo and with the development of an appropriate vocabulary of processes, we have been fascinated by its endless flexibility. Following this well-established tradition, the authors propose as their workhorse a multivariate Dirichlet process mixture of Normal-Wishart priors in conjunction with a Gaussian likelihood. A Gamma hyper-prior for the Dirichlet precision parameter is also introduced. Given that in this paper the number of mixture components K is fixed and typically large, an inspection of the posterior (11) and of the predictive (3) shows that under this model the posterior inferences for the Gaussian moments are largely driven by the centering prior $G_0(\cdot)$ and by the likelihood function. This could be easily verified by fitting model $(x_i \mid z_i = k, \phi_i) \sim N(x_i \mid \phi_i)$, $(\phi_i \mid G_0) \sim G_0$ to the flow cytometry data. The fundamental issue here is that, when K is large and fixed, the

[*]Department of Statistics and Centre for Analytical Science, University of Warwick, Coventry, U.K., mailto:f.rigat@warwick.ac.uk

DOI:10.1214/10-BA517A

gains of Dirichlet mixing the Gaussian mixture model are not obvious. Since Dirichlet mixing increases the complexity of the posterior distributions, especially those of the covariance matrices, was it really worth here? Also, should our ability to construct complex models be somewhat tempered by the need to explain in simple terms what Bayesians do to the experimentalists providing us with motivating data?

2 Weighty sub-samples

In a first instance, the authors propose a two-step procedure to better focus their posterior inferences on the mixture component of interest in a computationally affordable manner. The sub-sample X^R is chosen here uniformly among all possible sets of size $n^R < N$ whereas the targeted sub-sample X^T of size $n^T << n^R < N$ and moments (μ_{k^*}, Σ_k^*) is chosen using weights proportional to the Gaussian probability density function with the same mean and with a tuned covariance matrix. A key advantage of the latter is that the tuning vector v allows weighting differently the components of the covariance matrix, reflecting a focus on selected data dimensions. As noted by the authors, both this two-stage procedure and its multi-stage sequential Monte Carlo formulation can perform poorly when n^T is small. In this case, X^R might contain little or no information about (μ_{k^*}, Σ_k^*) unless n^R approaches N, which defeats the purpose of the selection sampling approach. This is not an academic question inasmuch as the proportion of cell sub-types k^* along with their associated moments might reveal therapeutically relevant differences among candidate immunization strategies. From a Bayesian perspective, it is natural to seek an improvement of this key aspect of the methodology by asking: what does the experimentalist know about the rare cell sub-type prior to observing this particular data set? To what extent can this existing prior knowledge be reflected in improved weighting of observations allocated respectively to X^R and X^T? From this angle, a striking feature of the multi-stage selection sampling algorithm proposed here is that the starting value of the weights w_i is effectively $1/(n^T + n^R)$ with no further reference to, for instance, the hyperparameters of the prior G.

3 Beyond Metropolis-Hastings

Bayesian statistics is overcoming a certain fascination for Markov chain Monte Carlo, mostly by recognising the limitations of traditional sampling recipes and by exploring alternative strategies (Jordan et al. (1998), Marjoram et al. (2003), Haario et al. (2006), Kou et al. (2006), Rue et al. (2008), Del Moral et al. (2006)). In this paper, computations are carried out by coupling an approximation of the joint posterior density with a numerical maximisation step using sequential Monte Carlo and a stopping rule. Here we focus on the former two, leaving a discussion of stopping rules to the next section.

Approximation (b) in equation (21) states that the separation of the sample in random and targeted subsets is sensible when the moments of the latter are sufficiently different from those of the former. In this case, the posterior distribution of the indicators z^R is essentially unaffected by data belonging to X^T. Under this assumption, the authors are faced with the daunting task of summing over all possible configurations of the vector z^R so as to approximate the left-hand side of (21). A first solution is proposed, based on L parallel two-stage samplers. The main unaddressed issues are how many of these samplers one should ideally run and, even more importantly, what relationships to induce among the chains generated by the L samplers. Intuitively, L should scale with the number of configurations of z^R one cares about, which itself is a function of (n^R, K) and thus of N. Also, when block-sampling of the z^R produces highly correlated draws, inducing negative correlations among the L chains along the lines of Frigessi et al. (2000) or using multiple-try strategies (Liu et al. (2000)) exploits the multiplicity of the multiple-chains sampler to improve mixing.

The authors proceed to elaborating their two-stage sampling strategy by generalising to a multi-stage SMC algorithm. The main difference with respect to the two-stage procedure here is the incremental updating of the targeted sub-sample by sampling new observations without replacement using their weights w_i. As such, the stability of this maximisation step hinges on that of the weights w_i. When estimates of the Gaussian mixture moments are updated, as new components are identified, the weights w_i might change dramatically for all observations. In other terms, this incremental procedure can be unstable to the extent that an observation which has been sampled in the targeted set early on might be recognised as not belonging to X^T at a later stage. Thankfully, the authors provide their code along with the paper, so that numerical stability can be promptly checked by the interested readers using their favorite datasets.

4 Stopping rules

Two stopping rules are proposed to terminate the incremental sampling of observations in X^T. The first prescribes stopping when no observations are found with a Bayes factor larger than a given threshold. The second rule states that sampling stops when less than $N_{threshold}$ unsampled observations have weights larger than a fixed value $c_{threshold}$. It is not clear whether the first stopping rule is computable and the authors do not make use of it in either of the examples included in the paper. Implementation of the second stopping rule requires setting a value of $c_{threshold}$. Here an empirical rule is proposed and then used for the motivating flow cytometry example, setting lower threshold values as the data dimensionality p increases. Unfortunately, the motivations underpinning this particular choice are not stated and the simulated data example is not helpful in this sense. In the latter, the authors first state that, using their empirical stopping rule, the number of targeted observations is 200. Then a targeted sub-sample with $n^T = 900$ is used, which together with their choice $n^R = 700$ violates one of the desiderata of the selection sampling approach, that is $n^T << n^R$. Figure 1 is not conclusive either, as the shrinkage of the posterior credible interval of μ_{11} from its left-most values to its middle values, when both X^R and X^T are used, is not compared with the distribution of intervals one would obtain using, for instance, a second random sub-sample of size 900 in addition to X^R. What needs being clarified here is what $c_{threshold}$ stands for. A possible interpretation is that if this is a critical threshold above which an observation should fall within the targeted sub-sample, then $N_{threshold}$ should be zero. In this case, $c_{threshold}$

should be set to a sufficiently high value to allow for efficient computations. Otherwise, before stopping the algorithm one should decide how to allocate the remaining samples with weights larger than $c_{threshold}$.

5 Feeding back to science

The flow cytometry example demonstrates the payoffs of the authors'modeling and computational efforts to address in a timely fashion an important problem in immunology. The implementation of their methods illustrates the strengths of the Bayesian approach, through which the ability of specific viral antigens to elicit a selected response of the immune system in conjunction with the secretion of effector cytokines is characterised. In this respect, this paper illustrates the extent to which Bayesian inference is a fundamental pillar ensuring that complex, high-dimensional data measuring the behavior of biological systems under realistic experimental conditions is coherently summarised and turned into information. The conclusions of the paper illustrate one of the key challenges Bayesian statisticians face today: feeding back to science their findings in a way that promotes scientific understanding. This is clearly a challenge, requiring communication and social skills along with the traditional mathematical and modeling abilities. It is also a challenge for the Editorial boards, faced with the need to ensure an appropriate balance between methodological developments, their motivations and applications. Why should Bayesian statisticians be concerned by this challenge? Because the impact of the methods we develop, and thus our reputation, critically hinge on our response.

References

Del Moral, P., Doucet, A., and Jasra, A. (2006). “Sequential Monte Carlo samplers.” *Journal of the Royal Statistical Society B*, 68: 411–436. 452

Frigessi, A., Gasemyr, J., and Rue, H. (2000). “Antithetic Coupling of Two Gibbs Sampler Chains.” *The Annals of Statistics*, 28: 1128–1149. 453

Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, j., Coller, H., Loh, M., Downing, J., Caligiuri, M., Bloomfield, C., and Lander, E. (1999). “Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.” *Science*, 286: 531–537. 451

Haario, H., Laine, M., Mira, A., and Saksman, E. (2006). “DRAM: efficient adaptive MCMC.” *Statistics and Computing*, 16: 339–354. 452

Jordan, M., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1998). “An introduction to variational methods in graphical models.” *In M. I. Jordan (Ed.), Learning in graphical models. Norwell, MA: Kluwer*. 452

Jung, K., An, G., and Ronald, P. (2008). “Towards a better bowl of rice: assigning function to tens of thousands of rice genes." *Nature Reviews Genetics*, 9: 91–101. 451

Kou, S., Zhou, Q., and Wong, W. (2006). "Equi-energy Sampler with applications in statistical inference and statistical mechanics." *The Annals of Statistics*, 34: 1581–1619. 452

Kuhn, T. (1962). *The structure of scientific revolutions*. University of Chicago press. 451

Liu, J., Liang, F., and Wong, W. (2000). "The multiple-try method and local optimization in Metropolis sampling." *Journal of the American Statistical Association*, 95: 121–134. 453

Marjoram, P., Molitor, J., Plagnol, V., and Tavare, S. (2003). "Markov chain Monte Carlo without likelihoods." *Proceedings of the National Academy of Sciences*, 100: 15324–15328. 452

Rue, H., Martino, S., and Chopin, N. (2008). "Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations." *Journal of the Royal Statistical Society B*, 71: 1–35. 452

Sebastiani, P., Gussoni, E., Kohane, I., and Ramoni, M. (2003). "Statistical challenges in functional Genomics." *Statistical Science*, 18: 33–70. 451

Stephens, M. and Balding, D. (2009). "Bayesian statistical methods for genetic association studies." *Nature Reviews Genetics*, 10: 681–690. 451