Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: http://www.tandfonline.com/loi/uasa20

Comment

Dustin Tran & David M. Blei

To cite this article: Dustin Tran & David M. Blei (2017) Comment, Journal of the American Statistical Association, 112:517, 156-158, DOI: 10.1080/01621459.2016.1270044

To link to this article: http://dx.doi.org/10.1080/01621459.2016.1270044

Stan Development Team (2016), "Stan: A C++ Library for Probability and Sampling, (version 2.9.0)," available at *http://mc-stan.org*. [138,149,151,153]
Tan, L. S. L., and Nott, D. J. (2013), "Variational Inference for Generalized Linear Mixed Models Using Partially Noncentered Parametrizations," *Statistical Science*, 28, 168–188. [151]
Uhler, C., Lenkoski, A., and Richards, D. (2014), "Exact Formulas for the Normalizing Constants of Wishart Distributions for Graphical Models," unpublished manuscript (*arXiv:1406.490*). [147]
Verbyla, A. P., Cullis, B. R., Kenward, M. G., and Welham, S. J. (1999), "The Analysis of Designed Experiments and Longitudinal Data by Using Smoothing Splines" (with discussion), *Applied Statistics*, 48, 269–312. [149]
Wainwright, M. J., and Jordan, M. I. (2008), "Graphical Models, Exponential Families and Variational Inference," *Foundations and Trends in Machine Learning*, 1, 1–305. [141]
Wand, M. P. (2009), "Semiparametric Regression and Graphical Models," *Australian and New Zealand Journal of Statistics*, 51, 9–41. [137,140]
Wand, M. P. (2014), "Fully Simplified multivariate normal updates in Non-Conjugate Variational Message Passing," *Journal of Machine Learning Research*, 15, 1351–1369. [151]
Wand, M. P., and Ormerod, J. T. (2008), "On Semiparametric Regression with O'sullivan Penalized Splines," *Australian and New Zealand Journal of Statistics*, 50, 179–198. [140]
Wand, M. P., and Ormerod, J. T. (2011), "Penalized Wavelets: Embedding Wavelets into Semiparametric Regression," *Electronic Journal of Statistics*, 5, 1654–1717. [137,140]
Wang, S. S. J., and Wand, M. P. (2011), "Using Infer.NET for Statistical Analyses," *The American Statistician*, 65, 115–126. [149]
Winn, J., and Bishop, C. M. (2005), "Variational Message Passing," *Journal of Machine Learning Research*, 6, 661–694. [137,138,140]
Wood, S. N. (2006), *Generalized Additive Models: An Introduction with R*, Boca Raton, FL: CRC Press. [149]
Wood, S. N., Scheipl, F., and Faraway, F. F. (2013), "Straightforward Intermediate Rank Tensor Product Smoothing in Mixed Models," *Statistics and Computing*, 23, 341–3601. [149]

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2017, VOL. 112, NO. 517, Theory and Methods
http://dx.doi.org/10.1080/01621459.2016.1270044

Comment

Dustin Tran and David M. Blei

Department of Computer Science and Statistics, Columbia University, New York, NY

We commend Wand (2017) for an excellent description of message passing (MP) and for developing it to infer large semiparametric regression models. We agree with the author in fully embracing the modular nature of message passing, where one can define "fragments" that enables us to compose localized algorithms. We believe this perspective can aid in the development of new algorithms for automated inference.

Automated inference. The promise of automated algorithms is that modeling and inference can be separated. A user can construct large, complicated models in accordance with the assumptions he or she is willing to make about their data. Then the user can use generic inference algorithms as a computational backend in a "probabilistic programming language," that is, a language for specifying generative probability models.

With probabilistic programming, the user no longer has to write their own algorithms, which may require tedious model-specific derivations and implementations. In the same spirit, the user no longer has to bottleneck their modeling choices to fit the requirements of an existing model-specific algorithm. Automated inference enables probabilistic programming systems, such as Stan (Carpenter et al. 2016), through methods like automatic differentiation variational inference (ADVI; Kucukelbir et al. 2016) and no U-turn sampler (NUTS; Hoffman and Gelman 2014).

Though they aim to apply to a large class of models, automated inference algorithms typically need to incorporate modeling structure to remain practical. For example, Stan assumes that one can at least take gradients of a model's joint density. (Contrast this with other languages that assume one can only sample from the model.) However, more structure is often necessary: ADVI and NUTS are not fast enough by themselves to infer very large models, such as hierarchical models with many groups.

We believe MP and Wand's work could offer fruitful avenues for expanding the frontiers of automated inference. From our perspective, a core principle underlying MP is to leverage structure when it is available—in particular, statistical properties in the model—which provides useful computational properties. In MP, two examples are conditional independence and conditional conjugacy.

From conditional independence to distributed computation. As Wand (2017) indicated, a crucial advantage of message passing is that it modularizes inference; the computation can be performed separately over conditionally independent posterior factors. By definition, conditional independence separates a posterior factor from the rest of the model, which enables MP to define a series of iterative updates. These updates can be run asynchronously and in a distributed environment.

We are motivated by hierarchical models, which substantially benefit from this property. Formally, let y_{nk} be the nth data point in group k, with a total of N_k data points in group k and K many groups. We model the data using local latent variables α_k associated with a group k, and using global latent variables ϕ, which are shared across groups. The model is depicted in Figure 1.

CONTACT David Blei *david.blei@columbia.edu* Department of Computer Science and Statistics, Columbia University, New York, NY 10027, USA.

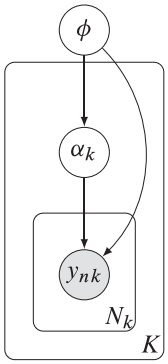

Figure 1. A hierarchical model, with latent variables α_k defined locally per group and latent variables ϕ defined globally to be shared across groups.

The posterior distribution of local variables α_k and global variables ϕ is

$$p(\alpha, \phi|y) \propto p(\phi|y) \prod_{k=1}^{K} \left[p(\alpha_k \mid \phi) \prod_{n=1}^{N_K} p(y_{nk} \mid \alpha_k, \phi) \right].$$

The benefit of distributed updates over the independent factors is immediate. For example, suppose the data consist of 1000 data points per group (with 5000 groups); we model it with 2 latent variables per group and 20 global latent variables. Passing messages, or inferential updates, in parallel provides an attractive approach to handling all 10,020 latent dimensions. (In contrast, consider a sequential algorithm that requires taking 10,019 steps for all other variables before repeating an update of the first.)

While this approach to leveraging conditional independence is straightforward from the message passing perspective, it is not necessarily immediate from other perspectives. For example, the statistics literature has only recently come to similar ideas, motivated by scaling up Markov chain Monte Carlo using divide and conquer strategies (Huang and Gelman 2005; Wang and Dunson 2013). These first analyze data locally over a partition of the joint density, and second aggregate the local inferences. In our work in Gelman et al. (2014), we arrive at the continuation of this idea. Like message passing, the process is iterated, so that local information propagates to global information and global information propagates to local information. In doing so, we obtain a scalable approach to Monte Carlo inference, both from a top-down view, which deals with fitting statistical models to large datasets and from a bottom-up view, which deals with combining information across local sources of data and models.

From conditional conjugacy to exact iterative updates. Another important element of message passing algorithms is conditional conjugacy, which lets us easily calculate the exact distribution for a posterior factor conditional on other latent variables. This enables analytically tractable messages (see Eqs. 7 and 8 of Wand 2017).

Consider the same hierarchical model discussed above, and set

$$p(y_k, \alpha_k \mid \phi) = h(y_k, \alpha_k) \exp\left\{\phi^\top t(y_k, \alpha_k) - a(\phi)\right\},$$
$$p(\phi) = h(\phi) \exp\left\{\eta^{(0)\top} t(\phi) - a(\eta_0)\right\}.$$

The local factor $p(y_k, \alpha_k \mid \phi)$ has sufficient statistics $t(y_k, \alpha_k)$ and natural parameters given by the global latent variable ϕ. The global factor $p(\phi)$ has sufficient statistics $t(\phi) = (\phi, -a(\phi))$, and with fixed hyperparameters $\eta^{(0)}$, which has two components: $\eta^{(0)} = (\eta_1^{(0)}, \eta_2^{(0)})$.

This exponential family structure implies that, conditionally, the posterior factors are also in the same exponential families as the prior factors (Diaconis and Ylvisaker 1979),

$$p(\phi \mid y, \alpha) = h(\phi) \exp\left\{\eta(y, \alpha)^\top t(\phi) - a(y, \alpha)\right\},$$
$$p(\alpha_k \mid y_k, \phi) = h(\alpha_k) \exp\left\{\eta(y_k, \phi)^\top t(\alpha_k) - a(y_k, \phi)\right\}.$$

The global factor's natural parameter is $\eta(y, \alpha) = (\eta_1^{(0)} + \sum_{k=1}^{K} t(y_k, \alpha_k), \eta_2^{(0)} + \sum_{k=1}^{K} N_k)$.

With this statistical property at play—namely, that conjugacy gives rise to tractable conditional posterior factors—we can derive algorithms at a conditional level with exact iterative updates. This is assumed for most of the message passing of semiparametric models in Wand (2017). Importantly, this is not necessarily a limitation of the algorithm. It is a testament to leveraging model structure: without access to tractable conditional posteriors, additional approximations must be made. Wand (2017) provided an elegant way to separate out these nonconjugate pieces from the conjugate pieces.

In statistics, the most well-known example that leverages conditionally conjugate factors is the Gibbs sampling algorithm. From our own work, we apply the idea to access fast natural gradients in variational inference, which accounts for the information geometry of the parameter space (Hoffman et al. 2013). In other work, we demonstrate a collection of methods for gradient-based marginal optimization (Tran, Gelman, and Vehtari 2016). Assuming forms of conjugacy in the model class arrives at the classic idea of iteratively reweighted least squares as well as the EM algorithm. Such structure in the model provides efficient algorithms—both statistically and computationally—for their automated inference.

Open challenges and future directions. Message passing is a classic algorithm in the computer science literature, which is ripe with interesting ideas for statistical inference. In particular, MP enables new advancements in the realm of automated inference, where one can take advantage of statistical structure in the model. Wand (2017) made great steps following this direction.

With that said, important open challenges still exist to realize this fusion.

First is about the design and implementation of probabilistic programming languages. To implement Wand's (2017) message passing, the language must provide ways of identifying local structure in a probabilistic program. While that is enough to let practitioners use MP, a much larger challenge is to then automate the process of detecting local structure.

Second is about the design and implementation of inference engines. The inference must be extensible, so that users cannot only employ the algorithm in Wand (2017) but easily build on top of it. Further, its infrastructure must be able to encompass a variety of algorithms, so that users can incorporate MP as one of many tools in their toolbox.

Third, we think there are innovations to be made on taking the stance of modularity to a further extreme. In principle, one can compose not only localized message passing

updates but compose localized inference algorithms of any choice—whether it be exact inference, Monte Carlo, or variational methods. This modularity will enable new experimentation with inference hybrids and can bridge the gap among inference methods.

Finally, while we discuss MP in the context of automation, we point out that fully automatic algorithms are not possible. Associated with all inference are statistical and computational tradeoffs (Jordan 2013). Thus, we need algorithms along the frontier, where a user can explicitly define a computational budget and employ an algorithm achieving the best statistical properties within that budget; or conversely, define desired statistical properties and employ the fastest algorithm to achieve them. We think ideas in MP will also help in developing some of these algorithms.

References

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2016), "Stan: A Probabilistic Programming Language," *Journal of Statistical Software*, 76, 1–32. [156]

Diaconis, P., and Ylvisaker, D. (1979), "Conjugate Priors for Exponential Families," *The Annals of Statistics*, 7, 269–281. [157]

Gelman, A., Vehtari, A., Jylänki, P., Robert, C., Chopin, N., and Cunningham, J. P. (2014), "Expectation Propagation as a Way of Life," *arXiv preprint arXiv:1412.4869*. [157]

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. (2013), "Stochastic Variational Inference," *Journal of Machine Learning Research*, 14, 1303–1347. [157]

Hoffman, M. D., and Gelman, A. (2014), "The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo," *Journal of Machine Learning Research*, 15, 1593–1623. [156]

Huang, Z., and Gelman, A. (2005), "Sampling for Bayesian Computation With Large Datasets," Technical Report. [157]

Jordan, M. I. (2013), "On Statistics, Computation and Scalability," *Bernoulli*, 19, 1378–1390. [158]

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., and Blei, D. M. (2016), "Automatic Differentiation Variational Inference," *Journal of Machine Learning Research,* to appear. [156]

Tran, D., Gelman, A., and Vehtari, A. (2016), "Gradient-Based Marginal Optimization," Technical Report. [157]

Wand, M. P. (2017), "Fast Approximate Inference for Arbitrarily Large Semiparametric Regression Models via Message Passing," *Journal of the American Statistical Association*, this issue. [156,157]

Wang, X., and Dunson, D. B. (2013), "Parallelizing MCMC via Weierstrass Sampler," *arXiv preprint arXiv:1312.4605*. [157]

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2017, VOL. 112, NO. 517, Theory and Methods
http://dx.doi.org/10.1080/01621459.2016.1270045

Comment

Wanzhu Tu

Department of Biostatistics, Indiana University School of Medicine, Indianapolis, IN

1. Introduction

Matt Wand's article describes a computational paradigm for semiparametric regression through variational message passing (VMP), an approximate inference technique originated from Bayesian networks but found increasing application in mainstream statistical modeling. By expressing semiparametric models as factor graphs, Wand showed that factorized modeling components could be used as bases for fast estimation and inference, through mean field variational Bayes and VMP algorithms. I applaud Wand's effort in identifying and formulating common factor graph fragments for frequently used parametric and semiparametric models; these fragments are designed to function as the basic building blocks for more complex models and algorithms. Wand's work, in my opinion, represents an initial but important step toward the development of general-purpose model fitting algorithms and user-friendly software for larger semiparametric models.

Technical contributions aside, the article's dissemination of the computational details, illustrated by frequently used models together with real-data applications, makes it an excellent introduction to variational Bayes and related message passing algorithms in a more familiar modeling setting. The educational values of the piece should not be underestimated, as adoption of a new methodology is usually contingent on the pedagogical quality of its initial introduction. Wand deserves compliments for his clear articulation of the problem and explicit provision of the solution.

In this short discussion, I would like to recap some of the basic tenets of VMP, comment on its use in semiparametric regression, and give my reasons on why we should be excited about Wand's article, all from the viewpoint of someone who uses semiparametric methods in real scientific investigations. Of course, it would not be so interesting if I am in complete agreement with Wand—I will comment on a few things that I thought deserved greater attention.

2. VMP in a Nutshell

VMP is an approximate inference algorithm that initially arises in the context of graphical models. An earlier algorithm based on a similar idea appeared under the name of belief propagation (Pearl 1986; Spiegelhalter 1986; Lauritzen and Spiegelhalter 1988). The VMP technique as we currently understand is framed by Winn (2003) and Winn and Bishop (2005), upon which this review is based.

CONTACT Wanzhu Tu *wtu1@iu.edu* Department of Biostatistics, Indiana University School of Medicine, Indianapolis, IN 46202.