References

Davis, T. A. (2006), *Direct Methods for Sparse Linear Systems*, Philadelphia, PA: SIAM. [164]

Lee, C. Y. Y., and Wand, M. P. (2016), "Streamlined Mean Field Variational Bayes for Longitudinal and Multilevel Data Analysis," *Biometrical Journal*, 58, 868–895. [164]

R Core Team (2014), *R: A Language and Environment for Statistical Computing*, Vienna, Austria: R Foundation for Statistical Computing. [165]

Rigby, R., and Stasinopoulos, D. M. (2005), "Generalized Additive Models for Location, Scale and Shape," *Journal of the Royal Statistical Society*, Series C, 54, 507–554. [165]

Rue, H., Martino, S., and Chopin, N. (2009), "Approximate Bayesian Inference for Latent Gaussian Models by Using Integrated Nested Laplace Approximations," *Journal of the Royal Statistical Society*, Series B, 71, 319–392. [164]

Wood, S. N., and Fasiolo, M. (2017), "A Generalized Fellner-Schall Method for Smoothing Parameter Estimation With Application to Tweedie Location, Scale and Shape Models," *Biometrics,* to appear. *arXiv preprint arXiv:1606.04802.* [164,165]

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2017, VOL. 112, NO. 517, Theory and Methods
http://dx.doi.org/10.1080/01621459.2016.1270051

Rejoinder

M. P. Wand

School of Mathematical and Physical Sciences, University of Technology Sydney, Sydney, Australia, and Australian Research Council Centre of Excellence for Mathematical and Statistical Frontiers, Queensland University of Technology (QUT), Brisbane, Australia

1. Mean Field Variational Bayes Iterative Scheme Variants

For a given model and dataset, the overarching goal is to minimize the Kullback–Leibler divergence of the joint posterior density function from the q-density function subject to a particular product restriction. This is mean field variational Bayes (MFVB). Apart from the traditional iterative scheme, listed as Algorithm 1 of Ormerod and Wand (2010), there were at least three other alternative MFVB iterative schemes in existence before this discussion took place—given in Winn and Bishop (2005), Minka (2005) and Minka and Winn (2008) under the label of variational message passing (VMP). It could be argued that there is no need for another term (and another acronym) and all approaches are just iterative scheme variants of MFVB. After all, the related approximation approach known as expectation propagation only has one name regardless of the iterative scheme used to obtain its version of optimal q-densities.

The comment of Reiss and Goldsmith offers two more MFVB iterative scheme variants: one without stochastic node to factor messages (Loop B) and one with factor to stochastic node messages replaced by factor-specific subcomponents at the end of their sec. 2. These variants provide interesting enlightenment regarding VMP-type approaches to MFVB. Nevertheless, their comment makes no mention of the arbitrarily large model viewpoint and compartmentalization of MFVB iterations via factor graph fragments, which is the main point of Wand (2017). Their claim that one could "jettison the message passing metaphor altogether" is made in the context of MFVB/VMP for a fixed factor graph. But the more relevant issue is whether this MFVB iterative scheme variant allows for compartmentalization of the algebra and computing in the same way that the Minka (2005) approach does. Put another way, can sec. 4.1 and 5.1– 5.3 of Wand (2017) be reexpressed in terms of factor-specific subcomponents with compartmentalization preserved?

2. Exponential Family Density Functions and Conjugacy

There is no doubt that exponential family density functions and conjugacy play an important role in practical MFVB/VMP. Discussion and additional insight on this aspect appears in the comments of Reiss and Goldsmith, Tran and Blei, and Tu. Each of the fragments presented in sec. 4 of Wand (2017) involve exponential family density functions and conjugate sub-graphs.

Nevertheless, exponential density functions and conjugacy are not intrinsic to MFVB/VMP. Equations (7)– (9) of Wand (2017) apply to messages of *any* form and can be used to compartmentalize MFVB iterative schemes for more elaborate model components such as Negative Binomial and t-distribution likelihoods and non-Gaussian penalization of random effects. Current work with Matt McLean is concerned with fragments for elaborate distribution likelihoods. It will allow for the replacement of updates presented in sec. 4.1.5 of Wand (2017) for Gaussian likelihood semiparametric regression models with similar models having other likelihoods. As with the fragments in Wand (2017), the forthcoming McLean and Wand fragment updates only need to be implemented once within a VMP software project.

3. Distributed Computing

The comments of Reiss and Goldsmith and Tran and Blei include discussion on distributed computing. I was fascinated to read about Pearl (1982), which demonstrates that ideas such as

CONTACT M. P. Wand matt.wand@uts.edu.au School of Mathematical and Physical Sciences, University of Technology Sydney, P. O. Box 123, Broadway 2007, Australia.

asynchronous updating for inference in large Bayesian networks go back at least 35 years.

As mentioned in sec. 1 of Wand (2017), Luts (2015) developed an approach to semiparametric regression on distributed datasets using MFVB approximate inference. Expectation propagation approaches to semiparametric regression and related models have also received significant attention recently, with distributed computing being one of the driving forces. Interesting yet-to-be-published work on this front includes Gelman et al. (2014), described in the Tran and Blei comment, and Kim and Wand (2017).

4. Accuracy/Speed Trade-Offs

Wand (2017) is on the interface between statistics and computer science and inevitably gets caught up in differing philosophies of these areas. One point of contention involves trade-offs between accuracy and speed. Biostatistician Tu says "it seems reasonable to demand a more detailed understanding of VMP's capacity for valid inference" and "inference is what science demands. A proven ability to produce valid inference is always expected of a new method," while computer scientists Tran and Blei say "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" in accordance with the recent statistics journal article, Jordan (2013), that advocates the incorporation of runtime into statistical risk measures. My work on fast approximate inference is driven by the belief that as the sizes of datasets and models continue to grow, alternatives to accurate but computationally intensive approaches are useful. Examples include applications for which interpretation matters more than valid inference, applications where speed is paramount, the problem of sifting through a large set of candidate models and experimental design. For the last of these, Ryan et al. (2016) is a recent review article with some mention of MFVB.

More than 30 years ago statisticians Breiman et al. (1984) developed classification and regression trees. It has since become a mainstay in business analytics and many other applications due to its speed, ability to handle messy data and interpretability—despite it lacking a proven ability to produce valid inference. Do all of the semiparametric regression methods that Tu uses in his impressive biomedical research have a proven ability to produce valid inference? In 30 years from now, will biomedical researchers be able to afford such a restriction to make the best use of the very large amounts of data at their disposal? Despite its imperfections, variational inference is a principled and versatile paradigm that can be of great use for confronting upcoming deluges of data.

There is scope for improvements in the accuracy of VMP. Section 5.1 of Knowles and Minka (2011) describes two alternatives to the Jaakkola-Jordan device for the logistic likelihood fragment that can lead to more accurate inference, as demonstrated by their Figure 1. One of them is simply the Knowles–Minka–Wand updates given in sec. 5.3 of Wand (2017) but with the Poisson likelihood replaced by its logistic counterpart. Current work with Tui Nolan is concerned with addressing numerical issues that arise with these more accurate logistic likelihood fragments and describing them within the Wand (2017) framework.

5. Model Sparsity

I agree with Wood that model sparsity and sparse matrix methods are very important for semiparametric regression as datasets and models continue to grow in size. As mentioned in Wand (2017), sparse matrix-type refinement (referred to there as matrix algebraic streamlining) of the fragment updates relevant to grouped data is still required for efficient handling of large longitudinal and multilevel datasets. This is a tedious endeavor and lacking in statistical glamor, which can make publication in good statistics journals more challenging. I am grateful to Simon Wood and co-authors for their continual recognition of the importance of numerical analysis in practical semiparametric regression and the solutions provided in their several articles and software such as the mgcv package (Wood 2016) in Wood and Fasiolo (2016) is the latest in an impressive sequence of contributions of this type.

6. Software

Tu, Tran and Blei and Wood each mention software. As discussed in Wand (2017) there is already one software product, Infer.NET, that uses VMP for fast approximate inference. A newer one is BayesPy (Luttinen 2016) for the Python computing environment. The niche that Wand (2017) endeavors to carve out stems from the fact that general purpose software products are limited in terms of how well they handle specific classes of models. Wand (2017) explained the ideas of VMP in statistical contexts and introduces factor graph fragments for arbitrarily large models. It puts this still very young methodology into the hands of statistical programmers who want to develop their own suite of programs for carrying out fast approximate inference for models relevant to their particular data analytic problems. In the case of my main area of research, semiparametric regression, there is enormous potential for software packages that allow large models to be fit quickly to massive datasets. In Lee and Wand (2016a, 2016b), we show how MFVB for large longitudinal and multilevel semiparametric regression analysis can be optimized for speed by streamlining the matrix algebra, and this work can be transferred to the VMP framework. As mentioned by Tran and Blei, parallelization of the required computations is also a possibility.

Acknowledgment

The author thanks each of the discussants for their thoughtful and thought-provoking comments.

ORCID

M. P. Wand http://orcid.org/0000-0003-2555-896X

Additional References

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984), *Classification and Regression Trees*, Belmont, CA: Wadsworth Publishing. [167]

Kim, A. S. I., and Wand, M. P. (2017). "On Expectation Propagation for Generalized, Linear and Mixed Models," *Australian and New Zealand Journal of Statistics*, in press. [167]

Luttinen, J. (2016), "BayesPy: Variational Bayesian Inference in Python," *Journal of Machine Learning Research*, 17, 1–6. [167]

Ryan, E. G., Drovandi, C. C., McGree, J. M., and Pettitt, A. N. (2016), "A Review of Modern Computational Algorithms for Bayesian Optimal Design," *International Statistical Review*, 84, 128–154. [167]

Wood, S. N. (2016), "mgcv: Mixed GAM Computation Vehicle With GCV/AIC/REML Smoothness Estimation," R package version 1.8. Available at *http://cran.r-project.org* [167]