Comment on: “Bayesian Computation Using Design of Experiments-based Interpolation Technique” by V. Roshan Joseph

John T. Ormerod and Matt P. Wand

April 20, 2012

The author is to be commended on the development of this new piece of methodology, which they name *DoIt*. We believe that the method, or later versions of the method, has the potential to be an important element in the kit-bag of non-MCMC based methods for approximate Bayesian inference. Throughout the article a number of criticisms have been leveled toward variational approximations (of which variational Bayes (VB) is a special case). As much of our recent research has been in this area we will focus our comments in defense of this methodology.

As a basis for comparison between methods we adapt the criteria listed in Ruppert, Wand & Carroll (2003, Section 3.16), upon which scatterplot smoothers may be judged, to criteria for general methodology.

1. *Convenience.* Is it available on the analyst’s favorite computer package?

2. *Implementability.* If not immediately available, how easy is it to implement in the analyst’s favorite programming language?

3. *Flexibility.* Is the method able to handle a wide range of models?

4. *Simplicity and Tractability.* Is it easy to understand how the technique processes the data to obtain answers? Is it easy to analyze the mathematical properties of the technique?

5. *Accuracy vs Efficiency* Does the method solve the problem to sufficient accuracy? How fast is the method?

We will argue that while DoIt performs well under several of these criteria VB compares favorably under others.

Under the criteria of convenience VB is most prominently implemented in the `Infer.NET` computing framework (Minka *et al.* 2010). The `Infer.NET` framework can be used in any of Microsoft’s .NET languages which includes C#, C++, Visual Basic, and Iron Python and implements the Expectation Propagation and Gibb’s sampling algorithms in addition to VB. The use of `Infer.NET` for some simple statistical models is illustrated in Wang & Wand (2011). While we freely admit that it is early days in this regard we look forward to an implementation of DoIt in a commonly used statistical environment such as `R`.

The `Infer.NET` framework is still in its infancy and does not implement every type of variational approximation and so, for a particular model, the analyst may have to implement both methods in their favorite programming language. Under this criteria VB can also have an advantage over DoIt. The paper gives a highly algebraic description of DoIt. Below we have attempted to summarize the DoIt algorithm to facilitate comparisons. We believe the core of the DoIt algorithm uses the following steps:

1. Choose a design: $\mathbf{D} = \{\boldsymbol{\nu}_1, \dots, \boldsymbol{\nu}_m\}$ and calculate $\mathbf{h} = [h_1, \dots, h_m]^T$ where $h_i = p(\mathbf{y}, \boldsymbol{\nu}_i)$.
2. Solve the minimization problem: $\widehat{\boldsymbol{\sigma}}^2 = \text{argmin}_{\boldsymbol{\sigma}^2 \geq \mathbf{0}} \, (\mathbf{e}^T \text{diag}(\mathbf{G}(\boldsymbol{\Sigma}))\mathbf{e})/m$ where $\mathbf{e} = (e_1, \dots, e_m)$, $e_i = (\mathbf{G}(\boldsymbol{\Sigma})^{-1})_i \mathbf{h}/(\mathbf{G}(\boldsymbol{\Sigma})^{-1})_{ii}$, $[\mathbf{G}(\boldsymbol{\Sigma})]_{ij} = \phi_{\boldsymbol{\Sigma}}(\boldsymbol{\nu}_i - \boldsymbol{\nu}_j)$ and $\boldsymbol{\Sigma} = \text{diag}(\boldsymbol{\sigma}^2)$.
3. Set $\widehat{\boldsymbol{\Sigma}} = \text{diag}(\widehat{\boldsymbol{\sigma}}^2)$, $[\mathbf{G}]_{ij} = \phi_{\boldsymbol{\Sigma}}(\boldsymbol{\nu}_i - \boldsymbol{\nu}_j)$, and solve the quadratic program
$$\widehat{\mathbf{c}} = \text{argmin}_{\mathbf{c} \geq 0} (\mathbf{h} - \mathbf{G}\mathbf{c})^T \mathbf{G}^{-1} (\mathbf{h} - \mathbf{G}\mathbf{c}).$$
4. Solve the minimization problem:
$$\widehat{\boldsymbol{\lambda}} = \text{argmin}_{\boldsymbol{\lambda} \geq \mathbf{0}} \frac{\widehat{\mathbf{b}}^T \text{diag}(\mathbf{G}(\boldsymbol{\Lambda}))\widehat{\mathbf{b}}}{m}$$
where $\mathbf{G}(\cdot)$ is as is defined in Step 2 above, $\widehat{\mathbf{b}} = \mathbf{G}(\boldsymbol{\Lambda})^{-1}(\mathbf{z} - a\mathbf{1})$, $\boldsymbol{\Lambda} = \text{diag}(\boldsymbol{\lambda})\widehat{\boldsymbol{\Sigma}}\text{diag}(\boldsymbol{\lambda})$, $z_i = h_i/\widehat{\mathbf{c}}^T \mathbf{g}(\boldsymbol{\nu}_i; \widehat{\boldsymbol{\Sigma}})$, $\mathbf{z} = [z_1, \dots, z_m]$, $\mathbf{g}(\boldsymbol{\nu}_i; \widehat{\boldsymbol{\Sigma}}) = [\phi_{\widehat{\boldsymbol{\Sigma}}}(\boldsymbol{\nu}_i - \boldsymbol{\nu}_1), \dots, \phi_{\widehat{\boldsymbol{\Sigma}}}(\boldsymbol{\nu}_i - \boldsymbol{\nu}_m)]$ and $a = \widehat{\mathbf{c}}^T \mathbf{G}(\widehat{\boldsymbol{\Sigma}} + \boldsymbol{\Lambda})\mathbf{G}(\boldsymbol{\Lambda})^{-1}\mathbf{z}/\widehat{\mathbf{c}}^T \mathbf{G}(\widehat{\boldsymbol{\Sigma}} + \boldsymbol{\Lambda})\mathbf{G}(\boldsymbol{\Lambda})^{-1}\mathbf{1}$.

The DoIt algorithm may need to follow Steps 1.–4. many times in order to determine a good design set $\mathbf{D}$ which is chosen differently depending on whether the posterior mode is known. If the posterior mode is known then $\mathbf{D}$ is chosen to follow a Latin hypercube design based on the Laplace approximation of the posterior density (note the citation title for Morris & Mitchel, 1995 is incorrect). If the posterior mode is unknown, or if the Laplace approximation is judged to be inaccurate, then $\mathbf{D}$ is built sequentially by minimizing $\boldsymbol{\nu}_{m+1} = \text{argmax}_{\boldsymbol{\theta}} \, (\widehat{\mathbf{c}}^T \mathbf{g}(\boldsymbol{\theta}; \widehat{\boldsymbol{\Sigma}}))^2 \{1 - \mathbf{g}(\boldsymbol{\theta}; \widehat{\boldsymbol{\Sigma}})^T \mathbf{G}(\widehat{\boldsymbol{\Lambda}})\mathbf{g}(\boldsymbol{\theta}; \widehat{\boldsymbol{\Sigma}})\}$ where $\widehat{\boldsymbol{\Lambda}} = \text{diag}(\widehat{\boldsymbol{\lambda}})\widehat{\boldsymbol{\Sigma}}\text{diag}(\widehat{\boldsymbol{\lambda}})$ and starting points for these maximization problems are obtained by choosing a point in the neighborhood of the $\boldsymbol{\nu}_i$ with the largest approximate leave-one-out error (specific details for this step are vague). The DoIt algorithm stops adding points to $\mathbf{D}$ when an approximate cross-validation criterion based criterion is judged to be sufficiently accurate. The minimization problems are solved using the Nelder-Mead algorithm which does not require derivative information, chosen we assume, to ease implementability.

The algorithm appears to contain many subproblems. Each of these subproblems may require some tuning for DoIt to obtain reasonable results. Termination criteria may need to be adjusted, multiple starting points may be required to ensure Steps 2 and 4 do not obtain poor results and the size of the neighborhood used for sequential updates of the design may need adjusting.

Suppose that we compare this for the longitudinal data analysis example considered in Section 4.1 of the paper. In comparison the VB algorithm, described in Ormerod & Wand (2010), can be programmed in around 10-15 lines of R code and requires virtually no tuning due to convergence properties of VB. In comparison the above algorithm, which lacks some detail, requires half a page or more to describe and may require at least minimal tuning. Clearly VB has an advantage in this case.

The DoIt algorithm has been custom designed for models involving continuous random variables with continuous joint distributions (implied by Theorem 1). Provided that the problem falls into this category DoIt appears quite flexible. In particular results for the nonlinear regression in

Section 3.1 are quite impressive and we do not know of a variational approximation for obtaining suitably accurate approximations for problems of this type. Furthermore, the the only other non-MCMC method, that we are aware of, suitable for this type of problem is the iterLap method of Bornkamp (2011).

However, VB is applicable in situations for models with both discrete and continuous random variables and it is not limited to joint distributions which are continuous. For example, the VB method has been successfully applied to Gaussian mixture models (McGrory & Titterington, 2007) and hidden Markov models (McGrory & Titterington, 2009) and has an advantage over DoIt in this setting. Furthermore, when the prior is discontinuous, for example when if the horseshoe prior of Carvalho, Polson & Scott (2010) is employed, then VB can be applied (Neville, Ormerod & Wand 2012). In such a setting it is unclear whether DoIt does not need a prohibitively large number of design points to obtain a sufficiently accurate approximation.

The counter claim against VB is that VB is only applicable to conjugate-family type models. While we admit that VB cannot be applied to every model much of our recent research has been to substantially widen recently the applicability of VB to some non-conjugate-family models (Ormerod & Wand, 2012; Wand *et al.*, 2011). In short, for the criteria of flexibility, VB can handle some models DoIt cannot and vice-versa.

Both methods are simple and fairly easy to understand how answers are obtained. We admit that few theoretical development for variational approximations have been made and those that exist are context specific (Humphreys & Titterington, 2000; Wang & Titterington, 2006; Hall, Ormerod & Wand, 2011; Hall *et al.*, 2011; Ren *et al.*, 2011; Ormerod & Wand, 2012). In terms of tractability Gaussian interpolation is a reasonably well understood technique (for example, Fasshauer, 2007). As noted in the paper most results for bounding such interpolation methods rely on the fill-distance of the design points. We do not know of results for obtaining good designs in high dimensional spaces. Thus we concur a direct application of DoIt, without using some type of dimension reduction, would be unsuitable for high dimensional problems. In comparison VB has been successfully applied in genetic association studies where the problems can involve hundreds of thousands, if not millions, of dimensions (Logsdon, Hoffman & Mezey, 2010; Carbonetto & Stephens, 2011).

We believe that criteria accuracy and efficiency should be considered together as one is often traded against the other. Furthermore, these should be considered in the context of the problem the method is trying to solve. Consider again the longitudinal data analysis example considered in Section 4.1. The paper describes the VB approximation as poor. We would describe the posterior approximations for the coefficients β_i as quite accurate. while for the variance components σ_ε^2 and σ_u^2 the posterior means are estimated quite well while the posterior variances are slightly underestimated. Furthermore, these approximations, using a näive implementation in R (which does not take advantage of the random effects structure), takes around 0.01 seconds to compute. If, in the context of the analysis, the analyst was only interested in the posterior approximations of the coefficients, then VB would be the ideal choice for this problem. It is hard to compare DoIt with this in mind as the paper does not report how long DoIt takes to solve this problem, but we anticipate that VB would compare favorably.

Our second objection to their comparison of variational approximations with DoIt is that all variational approximations are lumped together. For example, it Section 2.5 of the paper, DoIt is compared to the tangent variational approximation (TVA) of Jaakkola & Jordan (2000). We regard, for this problem, TVA to be inferior to the Gaussian variational approximation (GVA, Ormerod & Wand, 2012). Consider the example presented in Wand (2009, Section 6) in Figure

1 where TVA and GVA are applied. Clearly GVA, like DoIt, appears adequately accurate for this problem whereas TVA does not. Similarly, again considering the longitudinal data analysis example considered in Section 4.1, the paper compares the VB method described in Ormerod & Wand (2010) when other variational approximations are superior in terms of accuracy. Consider, in Figure 2 the grid-based variational approximation of Ormerod (2011). This approximation like the structured mean field variational approximation described in Wand *et al.* (2011) offer a general method for improving variational approximations, albeit at the expense of speed. Using grid-based variational approximations adequate approximations for the marginal posterior densities of the variance components can be obtained. In this regard the paper appears to be making a straw-man argument against variational approximations.

Clearly, we believe that while DoIt is a worthy addition to non-MCMC analysis and that the results presented in the paper are impressive, that variational approximations still offer a competitive alternative for many problems.

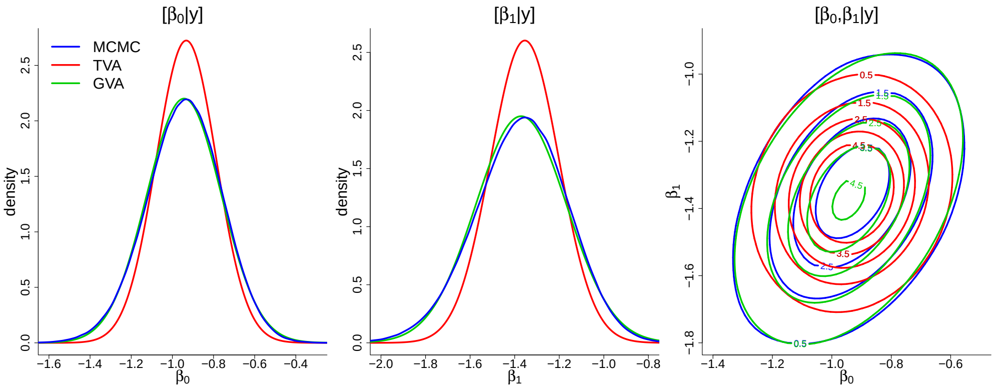

Figure 1: *A comparison of tangent based variationa approximations (TVA), Gaussian variational approximations (GVA) and MCMC for the bronchopulmonary dysplasia example in Wand (2009).*

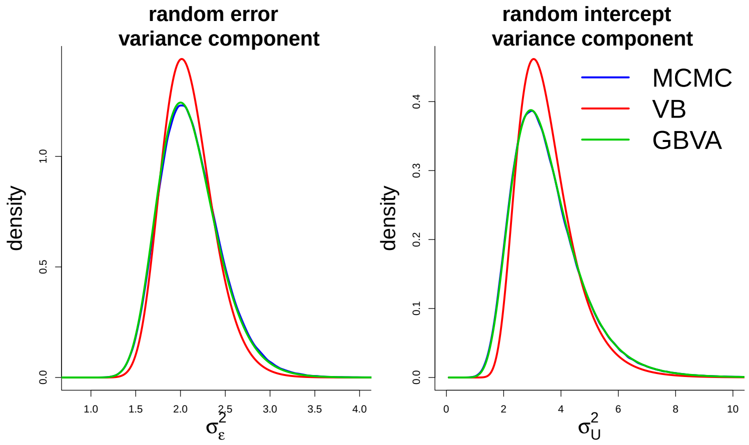

Figure 2: *Posterior density estimates for the inverse variance components using VB and the grid-based variational approximation described in Ormerod (2011).*

References

Bornkamp, B. (2011). Approximating Probability Densities by Iterated Laplace Approximations, *Journal of Computational and Graphical Statistics*, **20**, 656–669.

Carvalho, C.M., Polson, N.G. & Scott, J.G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, **97**, 465–480.

Carbonetto, P. & Stephens, M. (2011). Scalable variational inference for Bayesian variable selection in regression, and its accuracy in genetic association studies. *Bayesian Analysis*, **6**, 1–42.

Fasshauer, G.E. (2007). *Meshfree Approximation Methods with Matlab*, Volume 6 of Interdisciplinary Mathematical Sciences. Singapore: World Scientific.

Hall, P., Ormerod, J.T. and Wand, M.P. (2011). Theory of Gaussian variational approximation for a Poisson mixed model. *Statistica Sinica*, **21**, 369–389.

Hall, P., Pham, T., Wand, M.P. and Wang, S.S.J. (2011) Asymptotic Normality and Valid Inference for Gaussian Variational Approximation. The Annals of Statistics, **39**, 2502–2532.

Humphreys, K. & Titterington, D. M. (2000). Approximate Bayesian inference for simple mixtures. In Proceedings of Computational Statistics 2000 (Edited by J. G. Bethlehem and P. G. M. van der Heijden), 331–336. Physica, Heidelberg.

Jaakkola, T.S. & Jordan, M.I. (2000). Bayesian Parameter Estimation via Variational Methods. *Statistics and Computing*, **10**, 25–37.

Logsdon, B.A., Hoffman, G.E. & Mezey, J.G. (2010). A variational Bayes algorithm for fast and accurate multiple locus genome-wide association analysis. *BMC Bioinformatics*, **11**, 1–13.

McGrory, C.A. & Titterington, D.M. (2007). Variational approximations in Bayesian model selection for finite mixture distributions. *Computational Statistics & Data Analysis*, **51**, 5352–5367.

McGrory, C.A. & Titterington, D.M. (2009). Variational Bayesian Analysis for Hidden Markov Models. *Australian & New Zealand Journal of Statistics*, **51**, 227–244.

Minka, T., Winn, J., Guiver, J., and Knowles, D. (2010), "Infer.Net 2.4," Microsoft Research Cambridge, Cambridge, U.K., available at `http://research.microsoft.com/infernet`.

Ormerod, J.T. & Wand, M.P. (2010). Explaining variational approximations. *The American Statistician*, **64**, 140–153.

Ormerod, J.T. & Wand, M.P. (2012). Gaussian variational approximate inference for generalized linear mixed models. Journal of Computational & Graphical Statistics, **21**, 2–17.

Ren, Q., Banerjee, S., Finley, A.O. & Hodges, J.S. (2011). Variational Bayesian methods for spatial data analysis. *Computational Statistics & Data Analysis*, **55**, 3197–3217.

Ruppert, D., Wand, M.P. & Carroll, R.J. (2003). *Semiparametric Regression*. New York: Cambridge University Press.

Wand, M.P. (2009). Semiparametric Regression and Graphical Models. *Australian and New Zealand Journal of Statistics*, **51**, 9–41.

Wand, M.P., Ormerod, J.T. Padoan, S.A. and Frühwirth, R. (2011). Mean field variational Bayes for elaborate distributions. *Bayesian Analysis*, **6**, Number 4, 847–900.

Wang, B. & Titterington, D. M. (2006). Convergence properties of a general algorithm for calculating variational Bayesian estimates for a normal mixture model. *Bayesian Analysis*, **1**, 625–650.