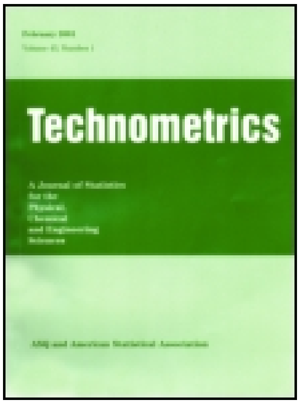

Technometrics
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/utch20

Comment

John T. Ormerod [a] & M. P. Wand [b]

[a] School of Mathematics and Statistics, University of Sydney, Sydney, 2006, Australia

[b] School of Mathematical and Sciences, University of Technology, Sydney, Broadway, 2007, Australia

Version of record first published: 16 Aug 2012

To cite this article: John T. Ormerod & M. P. Wand (2012): Comment, Technometrics, 54:3, 233-236

To link to this article: http://dx.doi.org/10.1080/00401706.2012.697244

Comment

John T. Ormerod

School of Mathematics and Statistics
University of Sydney
Sydney 2006, Australia
(john.ormerod@sydney.edu.au)

M. P. Wand

School of Mathematical Sciences
University of Technology
Sydney, Broadway 2007, Australia
(Matt.Wand@uts.edu.au)

The author is to be commended on the development of this new piece of methodology, which he names *DoIt*. We believe that the method has the potential to be an important element in the kit-bag of 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, and Carroll (2003, sec. 3.16), upon which scatterplot smoothers may be judged, to criteria for general statistical methodology:

- *Convenience*: Is it available in a computer package?
- *Implementability*: If not immediately available, how easy is it to implement in the analyst's favorite programming language?
- *Flexibility*: Is the method able to handle a wide range of models?
- *Tractability*: Is it easy to analyze the mathematical properties of the technique?
- *Accuracy*: Does the method solve the problem to sufficient accuracy?
- *Speed*: Are answers obtained sufficiently quickly for the analyst's application?
- *Extendibility*: Is the method easily extended to more complicated settings?

Concerning the convenience criterion, we note that VB is part of the Infer.NET computing framework (Minka et al. 2010). The Infer.NET framework can be used in any of the .NET languages, which includes C#, C++, and Visual Basic, 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 and Wand (2011). Although DoIt is a new idea, we look forward to its implementation in a commonly used statistical environment such as R.

The Infer.NET framework is still in its infancy and does not support all models for which VB algorithms can be derived. In such cases, the analyst has to implement VB in his/her favorite programming language.

Under this implementability criteria, VB can also have an advantage over DoIt. The article describes DoIt over several pages. But the algorithm can summarized in the following set of steps, with some notational changes that we believe improve digestibility. Joseph uses $\mathrm{diag}(\boldsymbol{v})$ to denote the diagonal matrix with diagonal entries corresponding to the vector $\boldsymbol{v}$ and $\mathrm{diag}(\boldsymbol{M})$ to denote the diagonal matrix formed when the off-diagonal entries of the square matrix $\boldsymbol{M}$ are set to zero. Following Magnus and Neudecker (1988), we use $\mathrm{dg}(\boldsymbol{M})$ for the latter to avoid having different meanings of "diag." We also use $\boldsymbol{v} > \mathbf{0}$ to denote all entries of a vector $\boldsymbol{v}$ being positive:

1. Choose a design $D = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_m\}$ within the parameter space (discussed below) and set
$$\boldsymbol{h} = \begin{bmatrix} p\,(\boldsymbol{y}, \boldsymbol{\theta}_1) \\ \vdots \\ p\,(\boldsymbol{y}, \boldsymbol{\theta}_m) \end{bmatrix}.$$
2. Define the $m \times m$ matrix $\boldsymbol{G}(\boldsymbol{\sigma})$ to have (i, j)th entry
$$|\mathrm{diag}(\boldsymbol{\sigma})|^{-1} \exp\left[-\frac{1}{2}(\boldsymbol{\theta}_i - \boldsymbol{\theta}_j)^T \{\mathrm{diag}(\boldsymbol{\sigma})\}^{-2} (\boldsymbol{\theta}_i - \boldsymbol{\theta}_j)\right].$$
Solve
$$\widehat{\boldsymbol{\sigma}} = \underset{\boldsymbol{\sigma} > \mathbf{0}}{\mathrm{argmin}} \left[\boldsymbol{h}^T \boldsymbol{G}(\boldsymbol{\sigma})^{-1} \{\mathrm{dg}(\boldsymbol{G}(\boldsymbol{\sigma}))\}^{-1} \boldsymbol{G}(\boldsymbol{\sigma})^{-1} \boldsymbol{h}\right].$$
3. Solve
$$\widehat{\boldsymbol{c}} = \underset{\boldsymbol{c} \geq \mathbf{0}}{\mathrm{argmin}} \left\{\frac{1}{2} \boldsymbol{c}^T \boldsymbol{G}(\widehat{\boldsymbol{\sigma}}) \boldsymbol{c} - \boldsymbol{h}^T \boldsymbol{c}\right\}.$$
4. Define
$$\boldsymbol{z} \equiv \{\mathrm{diag}(\boldsymbol{G}(\widehat{\boldsymbol{\sigma}}) \widehat{\boldsymbol{c}})\}^{-1} \boldsymbol{h} \quad \text{and}$$
$$a(\boldsymbol{\lambda}) \equiv \frac{\widehat{\boldsymbol{c}}^T \boldsymbol{G}\big(\sqrt{\widehat{\boldsymbol{\sigma}}^2 + \boldsymbol{\lambda}^2}\big) \boldsymbol{G}(\boldsymbol{\lambda})^{-1} \boldsymbol{z}}{\widehat{\boldsymbol{c}}^T \boldsymbol{G}\big(\sqrt{\widehat{\boldsymbol{\sigma}}^2 + \boldsymbol{\lambda}^2}\big) \boldsymbol{G}(\boldsymbol{\lambda})^{-1} \mathbf{1}}$$
for $m \times 1$ vectors $\boldsymbol{\lambda}$. Here, $\sqrt{\widehat{\boldsymbol{\sigma}}^2 + \boldsymbol{\lambda}^2}$ is the $m \times 1$ vector defined by taking element-wise squares and square roots, and $\mathbf{1}$ is an $m \times 1$ vector of 1's. Solve
$$\widehat{\boldsymbol{\lambda}} = \underset{\boldsymbol{\lambda} > \mathbf{0}}{\mathrm{argmin}} [\{\boldsymbol{z} - a(\boldsymbol{\lambda})\mathbf{1}\}^T \boldsymbol{G}(\boldsymbol{\lambda} \odot \widehat{\boldsymbol{\sigma}})^{-1} \{\mathrm{dg}(\boldsymbol{G}(\boldsymbol{\lambda} \odot \widehat{\boldsymbol{\sigma}}))\}^{-1}$$
$$\times \boldsymbol{G}(\boldsymbol{\lambda} \odot \widehat{\boldsymbol{\sigma}})^{-1} \{\boldsymbol{z} - a(\boldsymbol{\lambda})\mathbf{1}\}],$$

© 2012 American Statistical Association and
the American Society for Quality
TECHNOMETRICS, AUGUST 2012, VOL. 54, NO. 3
DOI: 10.1080/00401706.2012.697244

where $\widehat{\boldsymbol{\lambda}} \odot \boldsymbol{\sigma}$ denotes the element-wise product of $\widehat{\boldsymbol{\lambda}}$ and $\boldsymbol{\sigma}$.

5. The approximation to the posterior density function $p(\boldsymbol{\theta}|\mathbf{y})$ involves simple calculations involving D, $\widehat{\boldsymbol{\sigma}}$, $\widehat{\mathbf{c}}$, and $\widehat{\boldsymbol{\lambda}}$, given by (13) and (14) in the article.

The DoIt algorithm may need to follow steps 1–4 many times to determine a good design set D, which is chosen differently depending on whether the posterior mode is known. If the posterior mode is known, then D is chosen to follow a Latin hypercube design based on the Laplace approximation of the posterior density. If the posterior mode is unknown, or if the Laplace approximation is judged to be inaccurate, then D is built sequentially by solving an additional suite of multidimensional optimization problems. The starting points for these maximization problems are obtained by choosing a point in the neighborhood of the $\boldsymbol{\theta}_i$ with the largest approximate leave-one-out error (specific details for this step are vague on how this neighborhood is chosen). The DoIt algorithm stops adding points to 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. The algorithm contains 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. Consider the longitudinal data analysis example considered in section 4.1 of the article. The VB algorithm for this analysis, corresponding to algorithm 3 of Ormerod and Wand (2010), requires 10–15 lines of simple R code to implement and no tuning. In comparison, DoIt requires several multidimensional constrained optimizations and, possibly, some tuning.

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 only other non-MCMC (Markov chain Monte Carlo) method that we are aware of, suitable for this type of problem, is the iterLap method of Bornkamp (2011a). However, VB is applicable in situations for models with both discrete and continuous random variables, and it is not limited to joint distributions that are continuous. For example, the VB method has been successfully applied to Gaussian mixture models (McGrory and Titterington 2007) and hidden Markov models (McGrory and Titterington 2009), and has an advantage over DoIt in this setting. Furthermore, when the prior is discontinuous, for example, if the horseshoe prior of Carvalho, Polson, and Scott (2010) is employed, then VB can be applied (Neville, Ormerod, and Wand 2012). In such a setting, it is unclear whether DoIt needs a prohibitively large number of design points to obtain a sufficiently accurate approximation. 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 developments for variational approximations have been made and those that exist are context-specific (Hall, Humphreys and Titterington 2002; Wang and Titterington 2006; Hall, Ormerod and Wand 2011; Hall et al. 2011; Ormerod and Wand 2012). In terms of tractability, Gaussian interpolation is a reasonably well-understood technique (e.g., Fasshauer 2007). As noted in the article, most results for bounding errors for 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 that 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 parameters numbering in

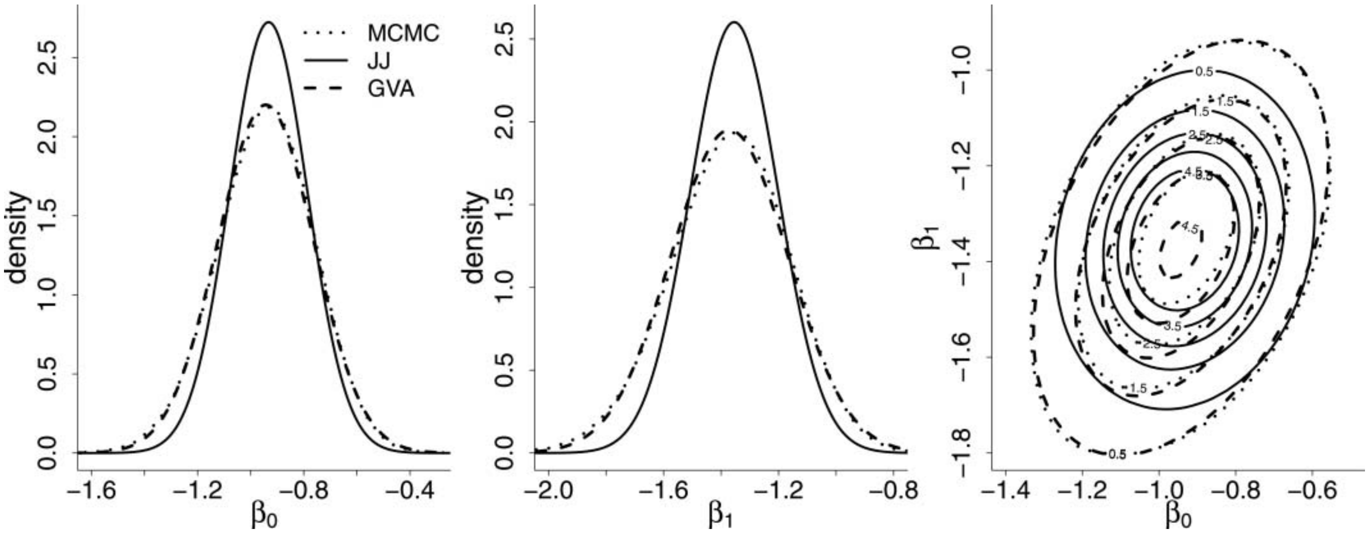

Figure 1. A comparison of tangent-based variational approximations (JJ), Gaussian variational approximations (GVA), and MCMC for the bronchopulmonary dysplasia example in Wand (2009).

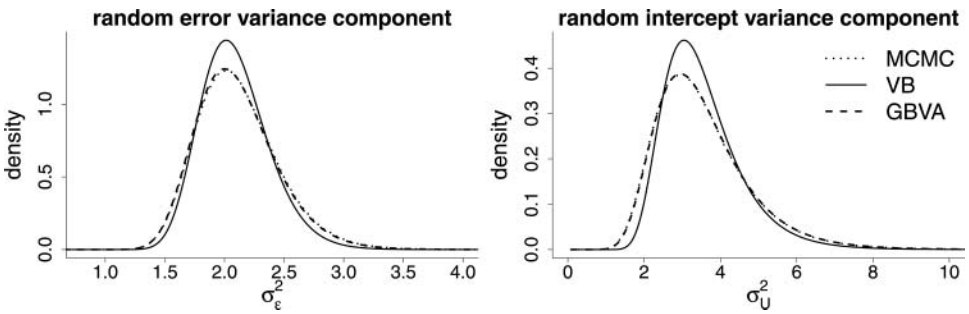

Figure 2. Posterior density estimates for the inverse variance components using VB and the grid-based variational approximation described in Ormerod (2011).

hundreds of thousands (Logsdon, Hoffman, and Mezey 2010; Carbonetto and Stephens 2011).

Criteria accuracy and speed could be considered together as one is often traded against the other. Furthermore, these should be considered in the context of the application at hand. Consider again the longitudinal data analysis example considered in section 4.1. Joseph describes the VB approximations for the variance components as "poor." We would call them "reasonable." Furthermore, these approximations, using a näive implementation in R (which does not take advantage of the random-effects structure), takes around 0.01 sec to compute on the first author's laptop. 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 article does not report how long DoIt takes to solve this problem, but we anticipate that VB would compare favorably.

Our second objection to the comparison with variational approximations with DoIt is that all variational approximations are lumped together. For example, in section 2.5 of the article, DoIt is compared with the tangent-based variational approximation of Jaakkola and Jordan (2000), which we denote by JJ. For this problem, JJ can be markedly inferior to Gaussian variational approximation (GVA) (Ormerod and Wand 2012), as we now demonstrate. Consider the example presented in Wand (2009, sec. 6) in Figure 1 where JJ and GVA are applied. Clearly, GVA, like DoIt, appears adequately accurate for this problem, whereas JJ does not. Similarly, again considering the longitudinal data analysis example considered in section 4.1, the article compares the VB method described in Ormerod and 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), offers 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 article appears to be making a straw-man argument against variational approximations.

An attraction of VB is that relative ease with which it can be extended to handle complications such as missing data. This follows from the locality property of VB, which, as with MCMC, means that algorithmic components are localized on the directed acyclic graph of the Bayesian model (e.g., Wand et al. 2011, sec. 3). In Faes, Ormerod, and Wand (2011), we demonstrated the extendibility of VB to handling missingness in regression models. Since missingness leads to an increase in the size of the Bayesian model (an increase in the number of *hidden nodes* in graph theoretical language), we would expect DoIt to run into difficulties for such models.

In summary, we believe that, while DoIt is a worthy addition to non-MCMC analysis and that the results presented in the article are impressive, variational approximations still offer a competitive alternative for many problems, depending on the analyst's weighting of the aforementioned criteria.

ADDITIONAL REFERENCES

Bornkamp, B. (2011a), "Approximating Probability Densities by Iterated Laplace Approximations," *Journal of Computational and Graphical Statistics*, 20, 656–669. [234]

Carbonetto, P., and Stephens, M. (2011), "Scalable Variational Inference for Bayesian Variable Selection in Regression, and Its Accuracy in Genetic Association Studies," *Bayesian Analysis*, 6, 1–42. [234]

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010), "The Horseshoe Estimator for Sparse Signals," *Biometrika*, 97, 465–480. [234]

Faes, C., Ormerod, J. T., and Wand, M. P. (2011), "Variational Bayesian Inference for Parametric and Nonparametric Regression With Missing Data," *Journal of the American Statistical Association*, 106, 959–971. [235]

Fasshauer, G. E. (2007), *Meshfree Approximation Methods With Matlab*, (Vol. 6: Interdisciplinary Mathematical Sciences). Singapore: World Scientific. [234]

Hall, P., Humphreys, K., and Titterington, D. M. (2002), "On the Adequacy of Variational Lower Bound Functions for Likelihood-Based Inference in Markovian Models With Missing Values," *Journal of the Royal Statistical Society,* Series B, 64, 549–564. [234]

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. [234]

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. [234]

Jaakkola, T. S., and Jordan, M. I. (2000), "Bayesian Parameter Estimation via Variational Methods," *Statistics and Computing*, 10, 25–37. [235]

Logsdon, B. A., Hoffman, G. E., and Mezey, J. G. (2010), "A Variational Bayes Algorithm for Fast and Accurate Multiple Locus Genome-Wide Association Analysis," *BMC Bioinformatics*, 11, 1–13. [234]

Magnus, J. R., and Neudecker, H. (1988), *Matrix Differential Calculus With Applications in Statistics and Econometrics*, Chichester: Wiley. [233]

McGrory, C. A., and Titterington, D. M. (2007), "Variational Approximations in Bayesian Model Selection for Finite Mixture Distributions," *Computational Statistics and Data Analysis*, 51, 5352–5367. [234]

——— (2009), "Variational Bayesian Analysis for Hidden Markov Models," *Australian and New Zealand Journal of Statistics*, 51, 227–244. [234]

Minka, T., Winn, J., Guiver, J., and Knowles, D. (2010), "Infer.Net 2.4," Available at *http://research.microsoft.com/infernet*. [233]

Neville, S. E., Ormerod, J. T., and Wand, M. P. (2012), "Mean Field Variational Bayes for Continuous Sparse Signal Shrinkage: Pitfalls and Remedies," unpublished manuscript. Available at *http://www.uow.edu.au/∼mwand/papers.html* [234,235]

Ormerod, J. T. (2011), "Grid Based Variational Approximations'," *Computational Statistics and Data Analysis*, 55, 45–56. [234,235]

Ormerod, J. T., and Wand, M. P. (2010), "Explaining Variational Approximations," *The American Statistician*, 64, 140–153. [234,235]

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003), *Semiparametric Regression*, Cambridge: Cambridge University Press. [233]

Wand, M. P. (2009), "Semiparametric Regression and Graphical Models," *Australian and New Zealand Journal of Statistics*, 51, 9–41. [233]

Wand, M. P., Ormerod, J. T., Padoan, S. A., and Frühwirth, R. (2011), "Mean Field Variational Bayes for Elaborate Distributions," *Bayesian Analysis*, 6 (4), 847–900. [235]

Wang, B., and 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. [234]

Wang, S. S. J., and Wand, M. P. (2011), "Using Infer.NET for Statistical Analyses," *The American Statistician*, 65, 115–126. [233]

Comment: DoIt—Some Thoughts on How to Do It

David M. Steinberg

Department of Statistics and Operations
Research, The Raymond and Beverly
Sackler Faculty of Exact Sciences
Tel Aviv University
Tel Aviv 69978, Israel
(*dms@post.tau.ac.il*)

Bradley Jones

SAS Institute, SAS Campus Drive
Cary, NC 27513
(*Bradley.Jones@jmp.com*)

Computational methods to explore posterior distributions, in particular Markov chain Monte Carlo (MCMC), have played a dominant role in Bayesian statistics over the last 30 years. These methods have enabled statisticians and researchers to tackle problems that defy closed-form solution, greatly expanding the scope of Bayesian analysis.

Joseph's ingenious DoIt algorithm uses ideas developed over the last 20–25 years on statistical modeling of deterministic functions to develop a direct approximation to complex posterior distributions, without the need for the large sequential samples required by MCMC. The method can be applied to a wide variety of problems and offers the promise of accurate results with substantially reduced computing. The approximation is a weighted sum of Gaussians, which leads to the significant advantages that it is simple to normalize and it is easier to compute marginal densities. We think that the method has great potential and applaud Dr. Joseph for this important new idea. Our comments focus on some issues where we think further work might lead to additional improvements in the method.

1. THE DOIT POSTERIOR DENSITY APPROXIMATION

The DoIt approximation is a linear combination of basis functions of the form $g(\boldsymbol{\theta}; \boldsymbol{\nu}_i, \boldsymbol{\Sigma}) = \exp\{-0.5(\boldsymbol{\theta} - \boldsymbol{\nu}_i)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta} - \boldsymbol{\nu}_i)\}$, where $\boldsymbol{\nu}_i$ is an evaluation point and $\boldsymbol{\Sigma}$ plays the role of a covariance matrix in a multivariate Gaussian density. The matrix $\boldsymbol{\Sigma}$ is clearly important in determining the quality of the DoIt approximation. But there are three important issues to consider: (1) What happens if the variances are too large? (2) What happens if the variances are too small? (3) What happens if the orientation is chosen poorly?

To see what can happen when the variances are too large, we consider the nonelliptical posterior density from Haario et al. that is studied in section 3.2. There is a single posterior mode at (0,3). The second derivatives of $\log[h(\boldsymbol{\theta})]$ at the mode lead to a diagonal covariance matrix whose entries are 100 and 1, respectively. Even for evaluation points very close to the mode, the associated Gaussian basis functions assign nonnegligible density to $\boldsymbol{\theta}$ values that have negligible posterior density (e.g., $\boldsymbol{\theta} = (15,3)$, with ν at the mode). With evaluation sites in the same region, the kriging predictor will "correct" for this error, but to do so, it must assign some basis functions positive coefficients and others negative coefficients. So, the problem of potentially negative density values is compounded. In this case, we think that it would be beneficial to "shrink" the variances (relative to the second derivatives) for the purpose of fitting the DoIt approximation. The fraction of negative coefficients in the initial DoIt fit could be a useful diagnostic here—a large fraction of negative coefficients may suggest that the variances are too large.

Large variances can also cause computational problems. The initial kriging estimator involves solving the linear system $\mathbf{Gc} = \mathbf{h}$, in which $\mathbf{G}$ is a correlation matrix. With large variances, $\mathbf{G}$ may be an ill-conditioned matrix for which the solution is numerically unstable. For the Haario et al. example with our cross-validation estimates of the variances, 20 of the 100 singular values of $\mathbf{G}$ were effectively 0.

On the other hand, the problem of having variances that are too small is that the Gaussians centered near the mode will fail to

© 2012 American Statistical Association and
the American Society for Quality
TECHNOMETRICS, AUGUST 2012, VOL. 54, NO. 3
DOI: 10.1080/00401706.2012.697247