Published in final edited form as:
J Am Stat Assoc. 2017 ; 112(517): 161–164. doi:10.1080/01621459.2016.1270049.

Comment

Philip T. Reiss[a,b] and **Jeff Goldsmith**[c]

[a]Department of Statistics, University of Haifa, Haifa, Israel

[b]The Department of Child and Adolescent Psychiatry, New York University School of Medicine, New York, NY

[c]Department of Biostatistics, Columbia University Mailman School of Public Health, New York, NY

1. Introduction

This very stimulating article reminded us of the following remark in a review article by Wand and co-authors (Ruppert, Wand, and Carroll 2009):

> Interplay with Computer Science is one of the most exciting recent developments in semiparametric regression. We anticipate this to be an increasingly fruitful area of research.

Bringing message passing to bear on semiparametric regression, as Wand has done here, is very much in the spirit of such interplay. The notion of message passing is ubiquitous in some areas of computer science, such as distributed computing and object-oriented programming. More specifically, within the field of artificial intelligence, the influential problem-solving model of Hewitt (1977) is based upon message passing, while Pearl (1982, 1988) proposed the passing of messages among neighboring nodes as a way to update beliefs efficiently in large Bayesian networks.

Against this backdrop it is unsurprising that, whereas variational message passing (VMP) is formulated quite differently in the three papers that Wand cites (Winn and Bishop 2005;Minka 2005; Minka and Winn 2008), in each case the authors find it natural to portray the algorithm as passing messages among the nodes of a network. But for readers like us with a mainstream statistics background, a message passing scheme such as that described by Wand comes across, at least at first, as uncomfortably mysterious (see Wainwright and Jordan 2008, p. 36).

We begin by posing three questions that arose for us as we read Wand's article. The first crystallizes our unease with the very notion of message passing, and addressing this key question will pave the way toward answering the other two.

1. As presented by Wand, following Minka (2005), VMP works by iteratively updating two types of messages: messages $m_{\theta_i \to f_j}(\theta_i)$ from variables (stochastic nodes) to factors, and messages $m_{f_j \to \theta_i}(\theta_i)$ from factors to variables. What, exactly, is the statistical meaning of these messages?

CONTACT Philip T. Reiss, reiss@stat.haifa.ac.il Department of Statistics, University of Haifa, Haifa Israel.

2. How is the VMP algorithm related to the traditional approach to mean field variational Bayes (MFVB)?
3. Wand’s message updates are given in (W7)–(W9) (here and below, to avoid confusion with our own equation numbers, we use (W*x*) to denote Wand’s equation (*x*)). How do these reduce to natural parameter updates, as presented from Section 3.2 onward?

In the following sections, we attempt to answer these questions and thereby, we hope, to shed some light on VMP.

2. A Closer Look at Messages

To address Question 1, we consider first the variable-to-factor messages and then the factor-to-variable messages.

Variable-to-factor messages

Recall that the form of the messages in Wand’s presentation of VMP flows from the factor graph representation. In an article popularizing this representation, Kschischang, Frey, and Loeliger (2001) developed a generic *sum–product* algorithm in which messages are passed back and forth between factors and variables, as in Wand’s presentation of VMP. Bishop (2006, p. 408) noted that one can eliminate the variable-to-factor messages in the sum–product algorithm, and reformulate it with only factor-to-variable messages. We find it helpful to reformulate VMP in a similar way.

Let us first recall Wand’s generic algorithm. In Section 3.2, following Minka (2005), he presents an iteration loop for VMP that could be stated as (1) choose a factor; (2) update messages from neighboring stochastic nodes to that factor; (3) update messages from that factor to neighboring stochastic nodes. While the schedule for updating factors may be flexible in some applications (Winn and Bishop 2005, sec. 3.5), for our purposes we can assume the factors are updated serially in a fixed order. Thus, a single iteration of the VMP algorithm might be written as a loop over *j*, with each step comprising two subloops:

Loop A. For $j = 1, \ldots, N$:

1. For each $i' \in S_j$, perform the update

$$m_{\boldsymbol{\theta}_{i'} \to f_j}(\boldsymbol{\theta}_{i'}) \overset{\propto}{\leftarrow} \prod_{j' \neq j : i' \in S_{j'}} m_{f_{j'} \to \boldsymbol{\theta}_{i'}}(\boldsymbol{\theta}_{i'}). \qquad (1)$$

   This is just (W7), but with $S_{j'}$ (defined in (W5)) replacing the equivalent “neighbors (j').”

2. For each $i \in S_j$:
   - a. Define the density in (W9), which is proportional to

$$\prod_{i' \in S_j \setminus \{i\}} m_{f_j \to \boldsymbol{\theta}_{i'}}(\boldsymbol{\theta}_{i'}) m_{\boldsymbol{\theta}_{i'} \to f_j}(\boldsymbol{\theta}_{i'}). \qquad (2)$$

b. Update the factor-to-variable message $m_{f_j \to \boldsymbol{\theta}_i}(\boldsymbol{\theta}_i)$ by (W8), which we repeat for convenience, again using S_j in place of "neighbors(j)":

$$m_{f_j \to \boldsymbol{\theta}_i}(\boldsymbol{\theta}_i) \leftarrow\!\propto \exp\left[E_{f_j \to \boldsymbol{\theta}_i}\left\{\log f_j(\boldsymbol{\theta}_{S_j})\right\}\right], \qquad (3)$$

where the expectation is with respect to the density in (2).

The messages on the right-hand side of (1) emanate from factors other than f_j, and thus are not updated within the current step of the loop over j. Therefore, the density (2) is unchanged if we substitute the right-hand side of (1) for $m_{\boldsymbol{\theta}_{i'} \to f_j}(\boldsymbol{\theta}_{i'})$ in (2). Doing so renders the first subloop redundant, so that a single iteration of VMP can be rewritten in the following mathematically equivalent form.

Loop B. For $j = 1, \ldots, N$:

For each $i \in S_j$,

a. Define the density proportional to

$$\prod_{i' \in S_j \setminus \{i\}} \left[m_{f_j \to \boldsymbol{\theta}_{i'}}(\boldsymbol{\theta}_{i'}) \prod_{j' \neq j : i' \in S_{j'}} m_{f_{j'} \to \boldsymbol{\theta}_{i'}}(\boldsymbol{\theta}_{i'}) \right] = \prod_{i' \in S_j \setminus \{i\}} \prod_{j' : i' \in S_{j'}} m_{f_{j'} \to \boldsymbol{\theta}_{i'}}(\boldsymbol{\theta}_{i'}). \qquad (4)$$

b. Update the factor-to-variable message $m_{f_j \to \boldsymbol{\theta}_i}(\boldsymbol{\theta}_i)$ using (3), with the expectation taken with respect to the density in (4).

If the jth factor depends on more than two of the $\boldsymbol{\theta}_i$ (i.e., $|S_j| > 2$) then Loop A may save some computation by performing the multiplication (1) just once, whereas Loop B must do the same multiplication, in (4), $|S_j| - 1$ times. We suspect the savings would typically be small, since $|S_j| \leq 2$ for most j and the multiplication reduces to the summing natural parameters (see Section 4). At any rate, the conceptual simplicity that Loop B achieves by doing away with variable-to-factor messages will facilitate our development in Sections 3 and 4.

Factor-to-variable messages

As Wand notes, in MFVB we seek component densities $q_1^*(\boldsymbol{\theta}_1), \ldots, q_N^*(\boldsymbol{\theta}_N)$ (here and below, unlike Wand, we include subscripts for these densities) such that $q^*(\boldsymbol{\theta}) = \prod_{i=1}^{M} q_i^*(\boldsymbol{\theta}_i)$

minimizes the Kullback–Leibler divergence $\mathrm{KL}[q\|p(\cdot|\boldsymbol{D})]=\int q(\boldsymbol{\theta})\ \log\ \left[\frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\boldsymbol{D})}\right]\,d\boldsymbol{\theta}$ over all product densities $q(\boldsymbol{\theta})=\prod_{i=1}^{M} q_i(\boldsymbol{\theta}_i)$. By (W10), in the VMP implementation of MFVB, we have

$$q_i^*(\boldsymbol{\theta}_i) \propto \prod_{j:i\in S_j} m_{f_j\to\boldsymbol{\theta}_i}(\boldsymbol{\theta}_i) \tag{5}$$

upon convergence. Alternatively, one can view $q_1^*(\boldsymbol{\theta}_1),\ldots,q_M^*(\boldsymbol{\theta}_M)$ as quantities that are being updated throughout the iterative algorithm (Minka 2005 does this). We can then view the factor-to-variable messages as (proportional to) iteratively updated subcomponent densities, where component $q_i^*(\boldsymbol{\theta}_i)$ is divided into subcomponents for each factor f_j of which $\boldsymbol{\theta}_i$ is an argument.

To summarize: we understand the variable-to-factor messages as a bookkeeping device with no independent statistical meaning, such that VMP can be formulated without them; and we interpret the factor-to-variable messages as factor-specific subcomponents of each component density as in (5). One could, then, jettison the message passing metaphor altogether, and replace the notation $m_{f_j\to\boldsymbol{\theta}_i}(\boldsymbol{\theta}_i)$ by $q_i^{f_j}(\boldsymbol{\theta}_i)$ or $q_i^{j}(\boldsymbol{\theta}_i)$.

3. Relating VMP to Traditional MFVB

We can now answer Question 2 posed in the Introduction. As explained in, for example, Ormerod and Wand (2010) and Goldsmith, Wand, and Crainiceanu (2011), the traditional MFVB algorithm initializes $q_1^*(\boldsymbol{\theta}_1),\ldots,q_M^*(\boldsymbol{\theta}_M)$ and then updates these iteratively via coordinate descent steps

$$q_i^*(\boldsymbol{\theta}_i) \leftarrow\!\propto \exp\ \left[\int \log\{p(\boldsymbol{\theta},\boldsymbol{D})\}\prod_{i'\neq i}\{q_{i'}^*(\boldsymbol{\theta}_{i'})d\boldsymbol{\theta}_{i'}\}\right] \tag{6}$$

for $i = 1, \ldots, M$.

By (5), (4) reduces to $\prod_{i'\in S_j\setminus\{i\}} q_{i'}^*(\boldsymbol{\theta}_{i'})$, and thus the VMP update (3) can be rewritten as

$$m_{f_j\to\boldsymbol{\theta}_i}(\boldsymbol{\theta}_i) \leftarrow\!\propto \exp\ \left[\int \log\{f_j(\boldsymbol{\theta}_{S_j})\}\prod_{i'\in S_j\setminus\{i\}}\{q_{i'}^*(\boldsymbol{\theta}_{i'})d\boldsymbol{\theta}_{i'}\}\right]. \tag{7}$$

This is a factor-specific analogue of the usual MFVB update (6): the ith component $q_i^*(\boldsymbol{\theta}_i)$ is replaced by its jth-factor-specific subcomponent $m_{f_j\to\boldsymbol{\theta}_i}(\boldsymbol{\theta}_i)$, the joint density $p(\boldsymbol{\theta}, \boldsymbol{D})$ by its jth factor $f_j(\boldsymbol{\theta}_{S_j})$, and the product of $q_{i'}^*(\boldsymbol{\theta}_{i'})d\boldsymbol{\theta}_{i'}$ over all $i' \neq i$ by one restricted to those in S_j

(this last is not a real difference, since taking the product over all $i' \neq i$ in (7) would be equivalent).

The traditional MFVB algorithm cycles over all i (the variables) to update the component densities. VMP, on the other hand, cycles over j (the factors), and within each factor, cycles over $i \in S_j$ to update subcomponent densities.

4. Reduction to Natural Parameter Updates

We now turn to Question 3. Throughout his Sections 3 and 4, Wand exploits conjugacy (as defined for factor graphs in Section 3.2.2) to simplify a number of special cases of the VMP algorithm, and in particular, to reduce updates (W7)–(W9) for the *messages* to updates for the *natural parameters* of the messages. The following is our attempt to make explicit what Wand's treatment assumes implicitly.

In the exponential family case, it is natural to write the factor-to-variable messages in the form

$$m_{f_j} \to \boldsymbol{\theta}_i(\boldsymbol{\theta}_i) \propto \exp\left[T_i(\boldsymbol{\theta}_i)^T \boldsymbol{\eta}_{f_j \to \boldsymbol{\theta}_i}\right] \quad (8)$$

for a natural parameter $\boldsymbol{\eta}_{f_j \to \boldsymbol{\theta}_i}$. But since these messages are not really defined—only their updates are, by (3)—there is a hint of vagueness in the definition of $\boldsymbol{\eta}_{f_j \to \boldsymbol{\theta}_i}$. It seems to us that the key to avoiding such ambiguity is to start by defining the jth factor density as an exponential family density respect to $\boldsymbol{\theta}_i$, for each $i \in S_j$: namely

$$f_j(\boldsymbol{\theta}_{S_j}) \propto \exp\left[T_i(\boldsymbol{\theta}_i)^T \boldsymbol{\eta}_{f_j \backslash \boldsymbol{\theta}_i}\right], \quad (9)$$

where $\boldsymbol{\eta}_{f_j|\boldsymbol{\theta}_i}$ does not depend on $\boldsymbol{\theta}_i$. For instance, consider the factor $f_j(\sigma^2, a) = p(\sigma^2|a)$ in Wand's linear regression example. In Section S.2.1 of the supplement, the logarithm of this factor is written in the two forms:

$$\log\, p(\sigma^2|a) = T(\sigma^2)^T \boldsymbol{\eta}_{p(\sigma^2|a)\backslash\sigma^2} = T(a)^T \boldsymbol{\eta}_{p(\sigma^2|a)\backslash a}$$

where $T(\cdot)$ denotes the inverse chi-squared sufficient statistic vector,

$$\boldsymbol{\eta}_{p(\sigma^2|a)\backslash\sigma^2} = \begin{bmatrix} -3/2 \\ -1/(2a) \end{bmatrix}, \text{ and } \boldsymbol{\eta}_{p(\sigma^2|a)\backslash a} = \begin{bmatrix} -1/2 \\ -1/(2\sigma^2) \end{bmatrix}. \quad (10)$$

Inserting (9) into (the log of) (3) yields

$$\log\, m_{f_j \to \boldsymbol{\theta}_i}(\boldsymbol{\theta}_i) \leftarrow E_{f_j \to \boldsymbol{\theta}_i}[\log\, f_j(\boldsymbol{\theta}_{S_j})] + \text{constant} = T_i(\boldsymbol{\theta}_i)^T E_{f_j \to \boldsymbol{\theta}_i}(\boldsymbol{\eta}_{f_j \backslash \theta_i}) + \text{constant}.$$

This update equation serves as the justification both for writing factor-to-variable messages in the form (8) and for updating their natural parameters by

$$\boldsymbol{\eta}_{f_j \to \boldsymbol{\theta}_i} \leftarrow E_{f_j \to \boldsymbol{\theta}_i}(\boldsymbol{\eta}_{f_j \backslash \boldsymbol{\theta}_i}). \quad (11)$$

The density in Loop A is proportional to (2), which can be expressed as

$$\prod_{i' \in S_j \backslash \{i\}} \exp\left[T_{i'}(\boldsymbol{\theta}_{i'})^T(\boldsymbol{\eta}_{f_j \to \theta_{i'}} + \boldsymbol{\eta}_{\theta_{i'} \to f_j})\right] = \prod_{i' \in S_j \backslash \{i\}} \exp\left[T_{i'}(\boldsymbol{\theta}_{i'})^T \boldsymbol{\eta}_{f_j \leftrightarrow \theta_{i'}}\right], \quad (12)$$

with $\boldsymbol{\eta}_{f_j \leftrightarrow \theta_{i'}}$ generalizing the notation that first appears in (W22). Eliminating variable-to-factor messages as in Loop B above, (12) becomes

$$\prod_{i' \in S_j \backslash \{i\}} \exp\left[T_{i'}(\boldsymbol{\theta}_{i'})^T \sum_{j' : i' \in S_{j'}} \boldsymbol{\eta}_{f_{j'} \to \theta_{i'}}\right] \quad (13)$$

(cf. (4)). If, as suggested in Section 3 above, $q^*_{i'}(\boldsymbol{\theta}_{i'})$ is defined using (5) throughout the iterations rather than just at convergence, then (13) can be rewritten as

$$\prod_{i' \in S_j \backslash \{i\}} \exp\left[T_{i'}(\boldsymbol{\theta}_{i'})^T \boldsymbol{\eta}_{q^*_{i'}(\theta_{i'})}\right].$$

This obviates the need for the notation $\boldsymbol{\eta}_{f_j \leftrightarrow \theta_{i'}}$, which is equivalent to $\boldsymbol{\eta}_{q^*_{i'}(\theta_{i'})}$.

We can thus summarize the natural-parameter-updating version of VMP as follows. First, we must obtain the right side of (11), a function of $\{\boldsymbol{\eta}_{q^*_{i'}(\theta_{i'})} : i' \in S_j \backslash \{i\}\}$, for each i, j. Then, after initializing all the natural parameters $\boldsymbol{\eta}_{f_j \to \theta_i}$, each iteration of the algorithm is a modified, more concrete version of Loop B:

Loop C. For $j = 1, \ldots, N$:

For each $i \in S_j$,

a. Compute $\boldsymbol{\eta}_{q^*_{i'}(\theta_{i'})} = \sum_{j' : i' \in S_{j'}} \boldsymbol{\eta}_{f_{j'} \to \theta_{i'}}$ for each $i' \in S_j \backslash \{i\}$;

b. Plug these into the right side of (11) to update $\boldsymbol{\eta}_{f_j \to \theta_i}$.

5. Bayesian Linear Regression Revisited

The remarkably simple Loop C can be visualized using a less snazzy alternative to a factor graph: an $N \times M$ table of natural parameters for the factor-to-variable messages, with a row for each of the variables $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N$ and a column for each of the factors $f_1, \ldots, f_M$. Table 1 illustrates this for the linear regression example of Wand's Section 3 (here, like Wand, we dispense with the subscripts in q_i^*).

In this case, only the second and third columns require updates (see (W21)), so each iteration consists of updating these two columns in turn. Consider updating the third column, that is, the natural parameters for the messages from $p(\sigma^2|a)$ to σ^2 and a. By (11), these updates are expectations of (10) with respect to densities proportional to (13); formulas for these expectations are as given in (W24), using information in Wand's Table S.1. Given these update formulas, in Loop C we simply add the natural parameters in the third row of Table 1 and insert the result ($\boldsymbol{\eta}_{q^*(a)}$, or $\boldsymbol{\eta}_{p(\sigma^2|a)\leftrightarrow a}$ in Wand's notation) into the formula for $\boldsymbol{\eta}_{p(\sigma^2|a)\rightarrow\sigma^2}$ to update this natural parameter; and then, similarly, we add up the natural parameters in the second row of Table 1 and insert the result into the formula for updating $\boldsymbol{\eta}_{p(\sigma^2|a)\rightarrow a}$.

6. Summary and Conclusion

In this comment, we have sought to build on the pedagogical component of Wand's achievement: that is, rendering VMP intelligible to statisticians who, like us and unlike Wand, have *not* spent years painstakingly reformulating and extending the methodology. Our main proposal, Loop B, is a modified VMP algorithm that is equivalent to the original, mathematically if not computationally, but does away with the variable-to-factor messages. This modification offers a reduced notational load, a clearer connection to the traditional implementation of MFVB, and a streamlined account of the reduction from message updates to natural parameter updates.

Of course, while explaining VMP to statisticians is an important contribution in itself, Wand has gone a great deal further. He has masterfully laid the groundwork for VMP-based semiparametric regression, which should be a huge step forward for flexible modeling with large datasets. Good on him (as they say Down Under) for this major contribution. We look forward to further advances in this area by Wand and coworkers, and by others who will draw inspiration from this landmark article.

Acknowledgments

Funding

Reiss's research was supported by award R01 MH095836 from the National Institute of Mental Health, and Goldsmith's research was supported in part by awards R01HL123407 from the National Heart, Lung, and Blood Institute and R21EB018917 from the National Institute of Biomedical Imaging and Bioengineering.

References

Bishop, CM. Pattern Recognition and Machine Learning. New York: Springer; 2006.

Goldsmith J, Wand MP, Crainiceanu C. Functional Regression via Variational Bayes. Electronic Journal of Statistics. 2011; 5:572–602. [PubMed: 22163061]

Hewitt C. Viewing Control Structures as Patterns of Passing Messages. Artificial Intelligence. 1977; 8:323–364.

Kschischang FR, Frey BJ, Loeliger H-A. Factor Graphs and the Sum–Product Algorithm. IEEE Transactions on Information Theory. 2001; 47:498–519.

Minka T. Divergence Measures and Message Passing. Technical Report MSR-TR-2005-173, Microsoft Research. 2005

Minka T, Winn J. Gates: A Graphical Notation for Mixture Models. Technical Report MSR-TR-2008-185, Microsoft Research. 2008

Ormerod JT, Wand MP. Explaining Variational Approximations. The American Statistician. 2010; 64:140–153.

Pearl J. Reverend Bayes on Inference Engines: A Distributed Hierarchical Approach. AAAI-82 Proceedings. 1982:133–136.

Pearl, J. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. San Mateo, CA: Morgan Kaufmann; 1988.

Ruppert D, Wand MP, Carroll RJ. Semiparametric Regression During 2003–2007. Electronic Journal of Statistics. 2009; 3:1193–1256. [PubMed: 20305800]

Wainwright MJ, Jordan MI. Graphical Models, Exponential Families, and Variational Inference. Foundations and Trends in Machine Learning. 2008; 1:1–305.

Winn J, Bishop CM. Variational Message Passing. Journal of Machine Learning Research. 2005; 6:661–694.

Table 1

Factor-to-variable message natural parameters for the linear regression example of Wand’s Section 3.

	$p(\beta)$	$p(y\mid\beta,\sigma^2)$	$p(\sigma^2\mid a)$	$p(a)$	
β	$\eta_{p(\beta)\to\beta}$	$\eta_{p(y\mid\beta,\sigma^2)\to\beta}$			$\eta_{q^*(\beta)}$
σ^2		$\eta_{p(y\mid\beta,\sigma^2)\to\sigma^2}$	$\eta_{p(\sigma^2\mid a)\to\sigma^2}$		$\eta_{q^*(\sigma^2)}$
a			$\eta_{p(\sigma^2\mid a)\to a}$	$\eta_{p(a)\to a}$	$\eta_{q^*(a)}$