Atmos. Chem. Phys. Discuss.,
https://doi.org/10.5194/acp-2017-10-AC1, 2017

Atmospheric Chemistry and Physics Discussions

Interactive comment on "Data Assimilation using an Ensemble of Models: A hierarchical approach" *by* Peter Rayner

P. Rayner

prayner@unimelb.edu.au

Received and published: 4 July 2017

article times natbib

Response to Referees' Comments

Peter Rayner

July 4, 2017

I thank Amy Braverman and an anonymous referee for their comments. Both have highlighted a series of problems of presentation which I have addressed in the revision. The reviews have also prompted me to read more widely in the statistical literature and realise that the paper is, as I suspected, an application of existing theory. I now point to this theory, and focus more on developing the example and some of the possibilities and problems that arise in biogeochemical applications. I have made a series of general and specific responses to reviewers' comments. I detail the general responses first and address specific concerns from each referee below. I have placed referee comments in Typewriter font and my responses in Roman.

General Comments

1. I have removed the appendix and replaced it with a refernce to (*MacKay*, 2003, Ch. 28). thanks to the anonymous reviewer for pointing this out.
2. I have expanded the description of the TRANSCOM inversion as requested by Amy Braverman.

3. I have made the conditioning on the data explicit in the various PDFs.

4. I have made a careful pass through the manuscript to regularise the typography.

Amy Braverman

General comments: In general I like this paper a lot. However, I find it extremely difficult to follow because of some type-o's and much notation with which I am not familiar. The notation seems inconsistent in distinguishing between fixed quantities and random ones, and indicating where conditioning has taken place. It is easy at this point to get side-tracked into a discussion of the interpretation of probability. I think that Dr. Braverman's real concern is the discussion of the TRANSCOM case where the target variables and data should be more clear. See general response 2. I agree that it should be clear when conditioning has taken place, see general response 3.

Page 3, lines 6 and 7: Please define the random variables x and Hi . In what sense is P (x|Hi) "the conventional data assimilation problem"? I hav added some explanatory text to clarify this.

2. Page 3, line 9: To what "linear model" are you referring? A linear transport model represented by Hi ? What do you mean by "over enough of the relevant pdf's"? Or, do you mean "over enough of the support of the random variable Hi ? I should have said a linear observation operator. I have corrected this and expanded the text.

3. Page 3, line 12 and 13: Is Hi the same as Hi ? H1 , . . . , HN are defined here as Jacobian matrices corresponding to N different transport models "...with un- knowns defined by

the multivariate Gaussian G(xb , B)...". Which unknowns? I am following the notation of *Rayner et al.* (2016) so that $\mathbf{H}_i$ is the linearised form of H_i. This distinction is unnecessary here so I have changed to the linearised form throughout. $\mathbf{x}$ are the continuous variables described in the text now added at the head of the section.

4. Page 3, line 15: "For each Hi our problem is the simple linear Gaussian inver- sion..." What does this mean? What is it you are trying to solve for or infer? Is it the flux that gave rise to the observed concentrations? The problem is more general than fluxes and concentrations although that is a common example and one I use later. Again, I hope the explanatory text at the head of the section explains the meaning of the symbols.

5. Page 3, line 16: "Most importantly for us P (xa |Hi) is Gaussian." Please define xa . Should it be xa ? These should be bold throughout following *Rayner et al.* (2016). I have corrected this. 4$\mathbf{x}^a$ is the analysis or posterior.

6. Page 3, lines 16 and 17: P (x, Hi) appears to be a joint distribution of two quanti ties: the vector-valued x and the matrix-valued Hi . It's unclear from the notation whether Hi is a random matrix or a fixed matrix. (On line 21, Hi is treated as random.) My guess is that it is fixed since the right side of the equal sign ap- pears to show the pdf of just one variable; presumably x. Is μi a vector or a scalar? Please define μi , Ui , and Wi . I have added these definitions. I have also switched from using $\mathbf{H}_i$ as the variable in the PDF to i since it is the index into the set of observation operators which is the target variable.

The expression P (x, Hi) = Wi G(μi , Ui) does not define

a proper pdf unless Wi = 1 since the area under the pdf must equal one. A more precise definition of a mixture would be in terms of ran- P 1 with probability wi , PK dom variables: X = K i=1 Ai Xi , Ai = , i=1 wi = 0 otherwise 1, and Xi âĹij G(μi , Ui). I don't agree with this. The expression represents the probability that i is the correct model and $\mathbf{x}$ the value of the continuous target variable. The normalisation requirement is defined by the integral over continuous target variables and sum over models. That is expressed by the extra constraint in the next equation.

7. Page 3, line 23: Either x should be bold, or not. Do not mix within the same equation. Also, the notation G(μi , Ui)(x) seems is very confusing (to me, at least). Do you mean that x is an argument to the function G? Why not write G(x|μi , Ui)? See general point above for typography. The other reviewer also noted the number of arguments in the definition of the Gaussian. I have regularised this, choosing to make the argument of the function, as well as its parameters, explicit. Note that this is different from the normal shorthand.

8. Page 3, line 26: In this equation Hi is treated as a non-random quantity. Above in line 21 it was random. Have you conditioned on it? If so, this distribution should be written as a conditional distribution. If not, then Wi is a random variable, not a fixed weight. W_i represents the probability that the ith model is the true model. I have added an explanation of this at the beginning of the section, hopefully clarifying several of Dr. Braverman's concerns at once.

9. Page 3, line 26: I can't check this equation because I can't follow the derivation in Appendix A. See below. It appears from reviewer 2's comments that the derivation is a standard result which I now quote.

10. Page 3, line 27: I think there is an extra "v" at the

end of this line. removed.

11. Page 4, line 2: I assume that xb has a prior distribution somewhere because it is being treated as both random and fixed in various places. What is the prior distribution? $\mathbf{x}^b$ is the mean of the prior distribution for the target variable $\mathbf{x}$. The confusion here raises a general question on which I seek editorial guidance. I have relied heavily on the notation and explanations in *Rayner et al.* (2016) thus making the current paper less self-contained. Should I move away from that and define the notation locally?

12. Page 4, line 16: Type-o. I am not seeing this.

13. Page 4, Footnote is missing. Removed.

Appendix A, through page 12
1. Page 12, line 18: K was defined in the main text as a normalizing constant. What is K(Hi) here? Do you mean P (Hi)?
2. Page 12, line 21: I am confused by this equation. G is a function that has an argument and parameters. What are the parameters and what are the arguments in this expression? The definitions from Section 2, lines 13 and 14 should be restated here and clarified as indicated earlier.
3. Page 12, line 23: Please define σ. Why is x in bold while dx is not?
4. Page 12, line 25: I find the use of Hi as both a Jacobian and an indicator of model identity to be very confusing. Why not let Hi be the Jacobian of model i, and introduce a model indicator, say Δ, an integer-value random variable taking values in 1, 2, . . . , M , where M is the number of models?

See general comment 1.

Reviewer 2

It is important to note that all the theory from Section 2 is conditional on the same data y being used in all of the inversions. Is this the case in the Gurney study? What can be done when y differs across models? Yes, the TRANSCOM inversion studies kept the PDFs for prior and data constant and changed only the observation operator. I could imagine weighting model 1 by its match to data $\mathbf{Y}_1$ and model 2 to data $\mathbf{Y}_2$ but this would get quite tricky if the dimensions were different and might be difficult to interpret in any case.

I disagree with the introduction of JIC. Isn't this just the marginal log-likelihood of the ith component? I think it should be described as such. Also it is well-known that the marginal log-likelihood penalises for complex models and the marginal likelihood is commonly employed in model selection. I agree and thank the reviewer for sending me back to the literature. I have expanded the introductory section and reduced the theoretical development accordingly and now use the conventional terminology. I am more concerned with model weighting than model selection here but agree that the marginal likelihood rather than its logarithm is the more useful quantity. The logarithm is still a more convenient tool for presentation however and I continue to use it for figures.

Are you sure that if we replace H i BH Ti + R with I we retrieve the BIC? The BIC plugs-in maximum likelihood estimates of x into the log-likelihood, while the marginal likelihood (which you label JIC) integrates out x, which is different. I

think they are equivalent. The same thing happens with the conventional χ^2 for an inversion. this uses the MAP for the target variables but, in the linear Gaussian case, when one substitutes this value, one obtains an expression involving the innovations and uncertainty projected into observation space.

P3 L30: The comment 'We don't believe that the relative quality of two model depends on the amount of data used to compare then...' is slightly out of place. There are many texts that show that asmyptotically these criteria hold (under some assumptions). If there is reason to believe that the assumptions are not valid (and the criteria are hence not valid) for the flux inversion problem discussed, then some other theoretical foundation for the use of a different criterion is needed. My point is pedagogical rather than theoretical. Modellers tend to use comparison with data as a measure of relative quality. This is quite reasonable but the relative quality of models doesn't diverge as we use more data to compare them. I have rewritten this point.

Related to the previous comment, the theory shows that the posterior flux is a weighted Gaussian mixture with weights Wi . What is the theoretical justifica- tion of using the marginal log-likelihood to weight when combining (as I think is implied when using the expression 'JIC-weighted')? The Wi being degenerate is unfortunate but not a valid reason. The same comment holds for the cross- validation metrics. This was an error of writing rather than calculation. I used the marginal likelihood throughout for weights but was incorrect in several places in calling this the JIC.

P8 L25: I cannot find a reasonable justification as to why one should add R and R sample . There are justifiable alternatives - for example in spatial analysis one would fit a space-time

model to the residuals with nugget and use this for the new R. This will be guaranteed to be non-singluar and positive definite. I think that adding these two matrices can lead to unforeseen consequences in other settings. I have now expanded this section and discuss in more detail the task of describing the model contribution to uncertainty. The reviewer's example is an interesting example. This kind of fitting of spatial models is certainly desirable if we have enough data to do it. In biogeochemical applications this is rare and we need something else. I agree that the addition of the sample covariance of residuals is ad hoc and I now describe it as an example, along with another of the sample covariance of a priori simulations.

Other Comments

P1 L24: I do not agree with the comment 'When structural uncertainty enters the problem only ensemble methods are available'. Although less direct, cannot one introduce an additional unstructured residual term in the model and see if that dominates? I agree, the comment was too strong and I have moderated it.

Section 2 and Appendix A need work to make the notation consistent. Just some things I picked up: - Hi should be bold everywhere - The lack of conditioning on the data y in all equations makes it hard to distinguish between prior and posterior distributions. Also the use of K(H i) as prior is confusing since K is a normalising constant in Equation (6). - The matrix B in Section 2 has become P in Appendix A. - In Appendix A, the PDF G(Âů) takes three arguments, while in Section 2 it only takes 2. - Appendix A needs to be cleaned up, there are expressions like 'mathbf' appearing, μ is not bold, the subscript i could be apparent on the LHS but not on the RHS, etc. See also my comment on clarifying the maths in

Section A further below.
P12: I found that the maths from Equation (A10) onwards becomes a bit obscure. The result is correct, however I think more details are needed. (some algebra deleted) However, as the author stated I also think this proof should be readily available in some textbook as it is just the marginal log-likelihood of a Gaussian density. I group all these comments together since I have dealt with them all by essentially quoting the standard result on marginal likelihood and leaving out most of the algebra.

P3 L27: Instead of "assumption that we must choose one" I would instead say "assumption that the sum of the probabilities of the models given the data equals one." Strictly speaking, unless you are using a Bayesian estimator you do not need to choose any specific one model. I was trying to make a different point that "none of the above" is not an option. I have reworded this.

P4 L7: The statement 'Eq.6 is also the same expression as the maximum likeli- hood estimate in Michalak et al. (2005)' is inaccurate. Estimate of what? Eq.6 is the marginal likelihood of the data under the ith component. But the expressions *are* the same. I now think this is because they represent the same thing, the marginal likelihood for a hyperparameter. I mention this in passing now but want to keep the paper focused on the practical use of the machinery.

P5 L32: What is the difference between JIC/N and JIC if N is constant? Does scaling affect any of the results and conclusions? Not here since we only use one dataset. See the above response on relative model quality.

P7 L9: Why 'model seven'? From Figure 1 it looks like Model 8

is the best model? Apologies, this was a numbering from 0 vs numbering from 1 problem, I have corrected the text.

References

MacKay, D. J. C., *Information Theory, Inference, and Learning Algorithms*, Cambridge University Press, available from http://www.inference.phy.cam.ac.uk/mackay/itila/, 2003.

Rayner, P., A. M. Michalak, and F. Chevallier, Fundamentals of data assimilation, *Geoscientific Model Development Discussions*, *2016*, 1–21, doi:10.5194/gmd-2016-148, 2016.