Atmospheric
Chemistry
and Physics

Open Access

Discussions

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

Anonymous Referee #2

Received and published: 26 April 2017

Summary of review

The article describes Bayesian model averaging for combining inferences from multiple flux inversion experiments. The approach seems reasonable, although I have some reservations on some of the steps taken, which I detail below. I hope the author finds the comments helpful.

General Comments

- The Introduction (Section 1) is well-presented. I particularly agree with the comment "the posterior PDFs for unknowns are implicitly dependent on the choice of model," a fact commonly overlooked.

- 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?

- 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.

- Are you sure that if we replace $H_i B H_i^T + R$ 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.

- 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.

- Related to the previous comment, the theory shows that the posterior flux is a weighted Gaussian mixture with weights $\{W_i\}$. What is the theoretical justification of using the marginal log-likelihood to weight when combining (as I think is implied when using the expression 'JIC-weighted')? The $\{W_i\}$ being degenerate is unfortunate but not a valid reason. The same comment holds for the cross-validation metrics.

- 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.

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?

- Section 2 and Appendix A need work to make the notation consistent. Just some things I picked up:

  - H_i 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 $K(\boldsymbol{H}_i)$ as prior is confusing since K is a normalising constant in Equation (6).
  - The matrix $\boldsymbol{B}$ in Section 2 has become $\boldsymbol{P}$ in Appendix A.
  - In Appendix A, the PDF $G(\cdot)$ 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.

- 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.

- P4 L7: The statement 'Eq.6 is also the same expression as the maximum likelihood estimate in Michalak et al. (2005)' is inaccurate. Estimate of what? Eq.6 is the marginal likelihood of the data under the ith component.

- P5 L32: What is the difference between JIC/N and JIC if N is constant? Does scaling affect any of the results and conclusions?

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

- 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. Using the author's notation, first the exponents in (A10) should be 1/2 so that

  $p(\boldsymbol{H}_i) \propto |\boldsymbol{A}\boldsymbol{P}^{-1}\boldsymbol{R}^{-1}|^{1/2} \exp\left(-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{H}_i\boldsymbol{x}^b)^T \cdot (\boldsymbol{R}+\boldsymbol{H}_i\boldsymbol{P}\boldsymbol{H}_i^T)^{-1} \cdot (\boldsymbol{y}-\boldsymbol{H}_i\boldsymbol{x}^b)\right)$

  $\propto |\boldsymbol{A}|^{1/2} \exp\left(-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{H}_i\boldsymbol{x}^b)^T \cdot (\boldsymbol{R}+\boldsymbol{H}_i\boldsymbol{P}\boldsymbol{H}_i^T)^{-1} \cdot (\boldsymbol{y}-\boldsymbol{H}_i\boldsymbol{x}^b)\right)$. Now let $D = |\boldsymbol{A}|^{1/2} = |\boldsymbol{A}^{-1}|^{-1/2}$. From here simply substitute $\boldsymbol{A}^{-1} = \boldsymbol{P}^{-1} + \boldsymbol{H}_i^T\boldsymbol{R}^{-1}\boldsymbol{H}_i$ into D to obtain

  $D = |\boldsymbol{A}^{-1}|^{-1/2}$
  $= |\boldsymbol{P}^{-1} + \boldsymbol{H}_i^T\boldsymbol{R}^{-1}\boldsymbol{H}_i|^{-1/2}$
  $= |\boldsymbol{P}^{-1}|^{-1/2}|\boldsymbol{I} + \boldsymbol{P}\boldsymbol{H}_i^T\boldsymbol{R}^{-1}\boldsymbol{H}_i|^{-1/2}$
  $= |\boldsymbol{P}^{-1}|^{-1/2}|\boldsymbol{I} + \boldsymbol{R}^{-1}\boldsymbol{H}_i\boldsymbol{P}\boldsymbol{H}_i^T|^{-1/2}$ (Sylvester's Determinant Theorem)
  $= |\boldsymbol{P}^{-1} + \boldsymbol{P}^{-1}\boldsymbol{R}^{-1}\boldsymbol{H}_i\boldsymbol{P}\boldsymbol{H}_i^T|^{-1/2}$
  $= |\boldsymbol{P}^{-1}|^{-1/2}|\boldsymbol{R}^{-1}|^{-1/2}|\boldsymbol{R} + \boldsymbol{H}_i\boldsymbol{P}\boldsymbol{H}_i^T|^{-1/2}$
  $\propto |\boldsymbol{R} + \boldsymbol{H}_i\boldsymbol{P}\boldsymbol{H}_i^T|^{-1/2}$ as required. 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.