Short comment to Lu et al.: Global methane budget and trend, 2010–2017: complementarity of inverse analyses using in situ (GLOBALVIEWplus CH_4 ObsPack) and satellite (GOSAT) observations

Luke Western[1]

[1]Atmospheric Chemistry Research Group, University of Bristol, Bristol, UK

October 2020

1 Reason for comment

I write this short comment to discuss a minor part of the discussion paper by Lu et al. (2020), in which a statistical condition is erroneously interpreted. This condition erroneously appears in many other sources of literature, and the belief in this condition has seemed to be confounded as a result.

The condition is that in equations (6) and (7) of Lu et al. (2020), which states that

$$J_A(\hat{\mathbf{x}}) = (\hat{\mathbf{x}} - \mathbf{x}_A)^T \mathbf{S}_A^{-1} (\hat{\mathbf{x}} - \mathbf{x}_A) \approx n, \tag{1}$$

and

$$J_O(\hat{\mathbf{x}}) = (\mathbf{y} - \mathbf{K}\hat{\mathbf{x}})^T \mathbf{S}_O^{-1} (\mathbf{y} - \mathbf{K}\hat{\mathbf{x}}) \approx m, \tag{2}$$

using the variables in Lu et al. (2020). These conditions state that the sum of log-likelihood and log-prior terms in the 'cost function' should be approximately equal to the number of observations, m for $\mathbf{y}$, or inferred parameters, n for $\mathbf{x}$, respectively at the maximum a posteriori value of $\mathbf{x}$.

The paper elaborates on this condition, for example for the component concerning the prior distribution, saying that "$J_A(\hat{\mathbf{x}}) >> n$ implies overfit to the observations because the posterior state vector estimates are far outside the estimated errors on the prior estimates." In addition, there is the statement "In our case the prior error covariance matrix is not strictly diagonal because of covariance for the wetland terms (Bloom et al., 2017), so one may expect $J_A(\hat{\mathbf{x}})$ to be somewhat deviated from n."

I hope in the following that I will demonstrate that these statements have no foundations in Bayesian probability theory, and likely have become pervasive due to an earlier misinterpretation of the mathematics, and subsequent adoption of this. I will explain what the mathematics show with respect to the error distribution of such distributions.

2 Properties of the Multivariate Normal

My assumption is that the confusion has stemmed from a misinterpretation of the condition outlined in, for example, Tarantola (2005), Ch6, which discusses the application of the Chi-squared distribution. It is worth noting that there are many other texts with a more mathematical description of this concept (e.g. Mardia et al., 1979). We can apply the properties of the Chi-squared distribution in Tarantola (2005), using less ambiguous notation, to the problem as framed in Lu et al. (2020), for example to the likelihood, where

$$J_O(\mathbf{x}) = (\mathbf{y} - \mathbf{K}\mathbf{x})^T \mathbf{S}_O^{-1} (\mathbf{y} - \mathbf{K}\mathbf{x}), \tag{3}$$

where the random variable $J_O(\mathbf{x})$ is distributed for all possible values according to the χ^2 distribution with

$$\nu = \dim(\mathbf{y}) = m \tag{4}$$

degrees of freedom. Note here that it doesn't say that $J_O(\mathbf{x}) = \nu$, nor $J_O(\mathbf{x}) \approx \nu$, but $J_O(\mathbf{x}) \sim \chi^2_\nu$, i.e. it is distributed with this distribution.

This is where I assume much of the confusion has come from. The earliest erroneous statement that I can find is in Michalak et al. (2005), but there may be others before this. Note also that the presence or lack of off-diagonal elements in the covariance matrix makes no difference to the statement in equation 3 and its subsequent distribution.

So what does $J_O(\mathbf{x}) \sim \chi^2_\nu$ mean practically? In a frequentist (i.e. non-Bayesian) setting, the 'cost' corresponds to a particular probability contour, following the quantile function of the Chi-squared distribution. Figure 1 shows a toy frequentist 'inversion' of two parameters, shown by the values on the x and y axis. The true values are 1 and 2. These were informed using 5 observations. The coloured background shows the 'cost' over the parameter space and the contours show the corresponding probability content according to the Chi-squared distribution. This has a practical application, for example to define the uncertainty in an estimated value (see e.g. Western et al., 2020).

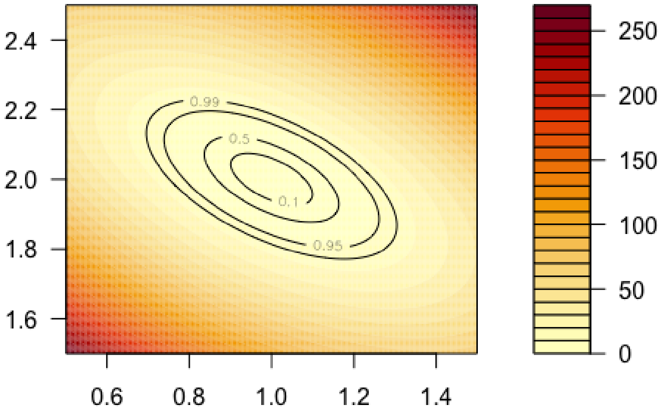

Figure 1: A toy frequentist 'inversion' of two parameters, shown by the values on the x and y axis. The true values are 1 and 2. These was informed using 5 observations. The coloured background show the 'cost' at the parameter values and the contours show the corresponding probability content according to the Chi-squared distribution.

This idea can also be readily applied, for example to equation 2. In a frequentist setting, if $J_O(\hat{\mathbf{x}}) = m$, this also has a corresponding probability following the quantile function of the Chi-squared distribution. That is, all values of $\mathbf{x}$ which are less or equal to some value of $J_O(\mathbf{x})$ can be translated into a confidence region with a defined probability. For example, if $m = 100$, for all values of $\mathbf{x}$ where $J_O(\mathbf{x}) \leq 124.3421$, then all these values fall within the 95% confidence region of the maximum likelihood estimate, or we can say with 95% certainty that the 'true' value falls within this parameter space. Figure 2 shows this probability for $J_O(\hat{\mathbf{x}}) = m$ for $1 \leq m \leq 100$. Or, in other words, Figure 2 shows the probability contour of a confidence region at $J_O(\hat{\mathbf{x}}) = m$, for a problem with m degrees of freedom. This probability asymptotes for large m, but what is key is that this probability content at m (or n) changes depending on the degrees of freedom.

Therefore, unless m and n are equal, or at least both very large, equations 1 and 2 are not making a comparison to the same probability. If $m = 1000$, then this probability contour for all values of $J_O(\hat{\mathbf{x}}) \leq m$ is around 51%, whereas if $n = 10$, all values where $J_A(\hat{\mathbf{x}}) \leq m$ is around 56%. I do not see a reason why (even assuming $m = n$) it is supposed that each term should be evaluated with a probability content ∼0.5 at $\hat{\mathbf{x}}$. If $J_A(\hat{\mathbf{x}}) < n$, and $J_O(\hat{\mathbf{x}}) < m$, why would this suggest an overfit?

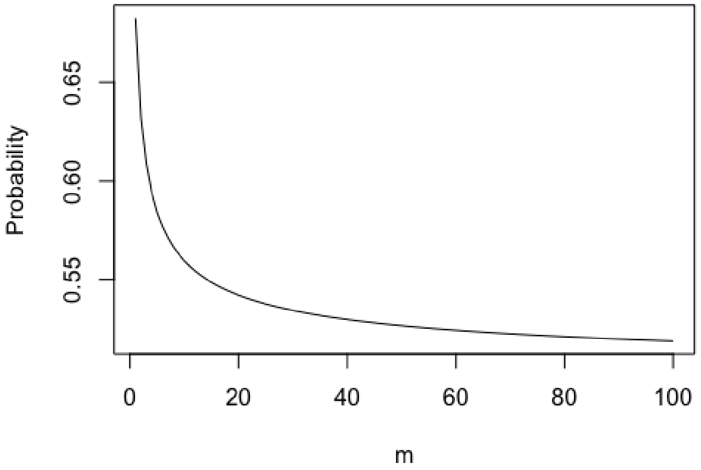

Figure 2: The probability contour at $J_O(\hat{\mathbf{x}}) = m$ according to the quantile function of the Chi-squared distribution with m degrees of freedom.

3 Does any of this matter?

The reason I am talking about uncertainty regions is that this seems to be implicit in the concept applied. My interpretation of, for example, equation 2 (Lu et al., 2020, equation 7), is that if $J_O(\hat{\mathbf{x}}) \leq m$ in equation 2, one would assume that the inversion is over confident in its estimated value, and hence the uncertainty is smaller than it should be. In Figure 1, this would translate as the contours on the plot being much smaller than they should be – the results are showing too much confidence in the inversion's estimates. This makes intuitive sense (even if equations 1 and 2 do not make sense statistically). However, a problem with the discussion in Section 2 is that the connection to uncertainty regions is valid for frequentist statistics, but not for Bayesian statistics, which is the stated approach to inference taken. Instead, measures of uncertainty in Bayesian inference rely on integration over the parameter space, which results in a fixed interval in which the 'truth' resides, as opposed to the uncertainty about a fixed most probable value in frequentist statistics. This means that the derived uncertainty is not the uncertainty in $\hat{\mathbf{x}}$ (a frequentist idea), but rather a fixed uncertainty region for $\mathbf{x}$ in which some metric $\hat{\mathbf{x}}$ resides. An example of a suitable Bayesian uncertainty region is the Highest Posterior Density (see e.g. Box and Tiao, 1992, Ch.2), defined as the narrowest region, R, in the total posterior parameter space that holds probability content $(1 - \alpha)$, or

1. $\mathrm{p}\{\mathbf{x} \in R\} = (1 - \alpha)$
2. for $\mathbf{x}_1 \in R$ and $\mathbf{x}_2 \notin R$, $\mathrm{p}(\mathbf{x}_1 \mid \mathbf{y}) \leq \mathrm{p}(\mathbf{x}_2 \mid \mathbf{y})$.

Although, in the case presented in the paper due to the Gaussian likelihood and prior, and resultant Gaussian posterior, such an integration is simple and readily available.

In my opinion, an improvement on trying to *post hoc* adjust probability distributions can take one of two paths. The first is to explicitly include uncertainty in parameters within the inversion itself, following either an empirical Bayes or hierarchical approach (e.g. Michalak et al., 2005; Ganesan et al., 2014), and thus formally considering the probabilities. The second is to invest some time in creating a better prior probability distribution that is representative of your actual prior belief of the possible parameter space. See e.g. Rougier (2007), Sect 2, for a more thorough discussion on this topic.

The second suggestion raises an interesting question – is the inversion approach taken in this work, and many others, actually probabilistic or is it a regularisation but explained using concepts from Bayesian probability? This has previously been raised in the context of remote sensing by e.g. Cressie (2018). The adjustment of 'probability distributions' to better fit models using the concept of a 'cost function' in my opinion falls closer to a regularisation problem. That is, if your posterior probability indicates that the mean inferred parameters have a low probability according to your prior probability, then this does not mean that the posterior/prior is wrong, and you may miss low-frequency events by removing this. If this happens consistently, then of course some reevaluation of the model or prior knowledge should take place. Using *post hoc* adjustment instead gives the impression that the prior probability (its functional form and parameter values), the uncertainties in the likelihood and the use of the extra variable γ, are instead weightings given to guide an optimisation procedure. The use of a the regularisation factor γ (as used in Lu et al. (2020)) in inverse modelling comes from regularisation rather than anything probabilistic (Tikhonov, 1963), which is somewhat 'un-Bayesian' in its current application unless included within the probabilistic hierarchy. Regularisation is fine – the machine learning community in particular has had a lot of success in working with optimisation through regularisation – but it then means that concepts such as uncertainty in the posterior estimate is not probabilistic and as such is difficult to interpret. As a result, the approach taken in the work should probably not be described as Bayesian, or probabilistic.

4 Final remarks

I do not want this to seem like an attack on the paper – it is not. In fact, I think the paper is very good and hence a suitable platform to raise this issue (rather than some work which generally has more pressing issues). I commend the work that has been done and hope for its eventual publication.

I have also been purposefully slightly provocative in my arguments, in order to facilitate discussion, which I hope others in the community will contribute to – including the nominated reviewers. I am willing to be proved wrong in my arguments, and indeed welcome a proof of the statement that has thus far, to my knowledge, not been sufficiently presented, even as a heuristic.

References

Bloom, A. A., Bowman, K. W., Lee, M., Turner, A. J., Schroeder, R., Worden, J. R., Weidner, R., McDonald, K. C., and Jacob, D. J. (2017). A global wetland methane emissions and uncertainty dataset for atmospheric chemical transport models (WetCHARTs version 1.0). *Geoscientific Model Development*, 10(6):2141–2156.

Box, G. E. P. and Tiao, G. C. (1992). *Bayesian inference in statistical analysis.* A Wiley-Interscience publication. Wiley, New York, wiley classics library ed edition. OCLC: 25247039.

Cressie, N. (2018). Mission CO $_2$ ntrol: A Statistical Scientist's Role in Remote Sensing of Atmospheric Carbon Dioxide. *Journal of the American Statistical Association*, 113(521):152–168.

Ganesan, A. L., Rigby, M., Zammit-Mangion, A., Manning, A. J., Prinn, R. G., Fraser, P. J., Harth, C. M., Kim, K.-R., Krummel, P. B., Li, S., Mühle, J., O'Doherty, S. J., Park, S., Salameh, P. K., Steele, L. P., and Weiss, R. F. (2014). Characterization of uncertainties in atmospheric trace gas inversions using hierarchical Bayesian methods. *Atmospheric Chemistry and Physics*, 14(8):3855–3864.

Lu, X., Jacob, D. J., Zhang, Y., Maasakkers, J. D., Sulprizio, M. P., Shen, L., Qu, Z., Scarpelli, T. R., Nesser, H., Yantosca, R. M., Sheng, J., Andrews, A., Parker, R. J., Boech, H., Bloom, A. A., and Ma, S. (2020). Global methane budget and trend, 2010–2017: complementarity of inverse analyses using in situ (GLOBALVIEWplus CH&lt;sub&gt;4&lt;/sub&gt; ObsPack) and satellite (GOSAT) observations. preprint, Gases/Atmospheric Modelling/Troposphere/Physics (physical properties and processes).

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979). *Multivariate analysis.* Probability and mathematical statistics. Academic Press, London ; New York.

Michalak, A. M., Hirsch, A., Bruhwiler, L., Gurney, K. R., Peters, W., and Tans, P. P. (2005). Maximum likelihood estimation of covariance parameters for Bayesian atmospheric trace gas surface flux inversions. *Journal of Geophysical Research: Atmospheres*, 110(D24).

Rougier, J. (2007). Probabilistic Inference for Future Climate Using an Ensemble of Climate Model Evaluations. *Climatic Change*, 81(3-4):247–264.

Tarantola, A. (2005). *Inverse problem theory and methods for model parameter estimation.* Society for Industrial and Applied Mathematics, Philadelphia, PA. OCLC: ocm56672375.

Tikhonov, A. N. (1963). Solution of incorrectly formulated problems and the regularization method. *Soviet Math. Dokl.*, 4:1035–1038.

Western, L. M., Rougier, J. C., Watson, I. M., and Francis, P. N. (2020). Evaluating nonlinear maximum likelihood optimal estimation uncertainty in cloud and aerosol remote sensing. *Atmospheric Science Letters*, 21(8).