



EGUsphere, author comment AC1
<https://doi.org/10.5194/egusphere-2022-213-AC1>, 2022
© Author(s) 2022. This work is distributed under
the Creative Commons Attribution 4.0 License.

Reply on RC1

Rona L. Thompson and Ignacio Pisso

Author comment on "A flexible algorithm for network design based on information theory"
by Rona L. Thompson and Ignacio Pisso, EGU sphere,
<https://doi.org/10.5194/egusphere-2022-213-AC1>, 2022

Comment

This paper presents a metric for the evaluation of observing networks. Previous studies have used some scalar function of the posterior covariance. Most commonly this is also the uncertainty in some linear functional of the posterior estimate (such as the total flux over a given region) but metrics like the trace of the posterior covariance are also used. The metric proposed by this paper is the information content defined as the difference in entropy between prior and posterior. Under the common linear Gaussian assumption this turns out to be closely related to the log of the generalised variance or determinant of the posterior covariance. The authors develop the necessary mathematics for their metric then present two examples, adding isotope measurements for methane or radiocarbon measurements for estimating fossil fuel CO₂ emissions. The mathematics is presented well and the examples are clear and pertinent. Furthermore the paper clearly lies within scope for the journal.

My concern with the paper is its lack of evaluation of the metric itself. There is a comparison with the independence metric but not with the covariance metrics. It is not clear to me that the information theory metric does the same job as the covariance metric or a better or worse job. There are two related problems:

First the generalised variance is one metric and probably not a very flexible one. It must include all potential sources. Normally this is not what we want. We have some target quantity like national emissions for which we are designing the network. Normally our target quantity will be a subset of pixels (e.g. pixels in one country). I'm not sure we can easily calculate the determinant of a submatrix of the Hessian.

Response

We thank the reviewer for his well-thought through comments, especially about the metric we use. Concerning the flexibility of our metric (i.e., the information content, calculated as the log of the determinant of the ratio of prior to posterior uncertainty), yes it can be used to target specific flux types or regions, such as national emissions. In this case, the Hessian matrix is calculated specifically for the region/flux of interest. In the manuscript, we also give examples of this (see e.g. lines 204 to 221). Here we describe how in our example the metric can be set to target only emissions in EU27+3 countries (although the domain is larger) or set to target only e.g. anthropogenic emissions. Furthermore, we

describe how the contribution to changes in the mixing ratios from fluxes that are not the target should be accounted for (see Eq. 21). As far as we are aware, our manuscript is the first to formally describe how the influence of non-targeted fluxes on the observations can be accounted for.

Comment

Next there is the choice of uncertainty quantity to minimise. Rayner (1996) pointed out that the preferred network depended on details of this quantity, such as the total ocean flux vs the average uncertainty for each ocean basin. The determinant is the volume of the hyperellipse described by the posterior covariance. It might be a good general choice but is likely to obscure these differences.

Response

We chose to the information content metric because it describes the change in probability space from before an observation is made (prior probability) compared to after an observation is made (posterior probability) and thus is more closely linked to the observations themselves than to the choice of metric (e.g. minimum of the mean posterior uncertainty of the target fluxes). However, we agree with the reviewer that it would be interesting to include a comparison to a more standard metric, e.g. minimum of the mean posterior uncertainty.

We now include this comparison for the CH₄ example in the Discussion section. Using the posterior uncertainty metric, the selected network is quite different to that using the information content (only one site is common between them). Therefore, to test the performance of each network, we ran two pseudo-data inversions each with the same prior fluxes. The performance was assessed using the ratio of the distance of the posterior from the true fluxes versus the distance of the prior from the true fluxes. We found that the inversion using the network based on information content performed slightly better (i.e., the posterior fluxes were closer to the true fluxes) than that based on the posterior uncertainty.

Comment

Finally I think the computational advantages of the new metric need a bit more justification. The authors claim that the covariance metric requires the inversion of a large matrix. Depending on the uncertainty metric we choose to minimise this might not be true. In general our target quantity is a linear functional of the posterior sources. Examples include the sum over some subregion and average over time. From what I learned to call Riesz's Representation Theorem (though there seem to be several of these) for any linear functional f on R^n there is a vector v such that:

$$f(x) = vx \text{ for all } x \text{ in } R^n$$

Thus for any target quantity t we can find some vector v_t such that

$$t^a = v_t x^a$$

The superscript "a" refers to the analysis or posterior. An example v_t contains "1" for every pixel in a region and "0" otherwise. This will sum over the region of interest. By the Jacobian law of probabilities the uncertainty in t is given by

$$v_t^T A v_t$$

where A is the posterior or analysis covariance. A is the inverse of the Hessian, G , so we

need to calculate:

$$v_t^T G^{-1} v_t$$

I believe this calculation can be efficiently accomplished by the Cholesky decomposition of G if we write (Cholesky decomposition):

$$G = LL^T$$

then I believe:

$$G^{-1} = L^{-1T} L^{-1}$$

Substituting this we see:

$$t = y^T y$$

where

$$y = L^{-1} v_t$$

Thus I think the target uncertainty can be performed with a Cholesky decomposition, a matrix-vector product and a dot-product. This may even be less costly than the determinant via the Cholesky decomposition. I may just as easily be wrong here but think the comparison of the cost and generality of the new metric compared to the existing uncertainty metric does need more consideration than it gets here.

Response

We have included an analysis of the algorithmic complexity and efficiency of using the information content as the metric at the end of the Discussion section and compare it with the posterior uncertainty metric.

Obtaining the posterior covariance is one of the specific cases that cannot be done without inverting matrices. Even if these are relatively well behaved the cost of the operation (and the theoretical complexity) is of the order of $O(n^3)$. The matrix multiplication $M(n)$ of two $n \times n$ matrices is also of asymptotic complexity $O(n^3)$. It is true that faster algorithms exist in theory starting with the celebrated Strassen (1969) that is $O(n^{\log_2(7)})$; but even for matrix larger than 1000×1000 the time reduction can be less than 10% with respect to the naive algorithm due to hardware optimization and acceleration. Besides the practical considerations related to a particular software/hardware implementation it has been established that the algorithmic complexity and hence the computational cost of the calculation of the determinant is the same as that of matrix multiplication (Strassen, 1969; Aho et al., 1974). Therefore, in comparing the two approaches we can abstract the particularities of the algorithm used for matrix multiplication and of its hardware implementation and simplify the analysis by comparing the number of matrix multiplications ($O(n^3)$ or $M(n)$), matrix inversions ($O(n^3)$ or $M(n)$), LU ($2O(n^3)/3$) or Cholesky ($O(n^3)/3$) decompositions and determinant calculations ($O(n^3)$ or $M(n)$).

In the case of synthesis inversion, obtaining the posterior covariance of the maximum a posteriori estimate requires two inversions, two matrix multiplications, one addition and one additional inversion (e.g. Tarantola (2005)). Subsequently for obtaining the covariance metric a Cholesky decomposition, a matrix-vector product and a dot-product will be applied as suggested above.

The cost of the covariance metric algorithm can therefore be analysed as follows:

1. The calculation of the Hessian $H^T R^{-1} H + B^{-1}$ requires
 1.1 one inversion of $B \sim O(n^3) / 3$
 1.2 one inversion of $R \sim O(k^3) / 3$
 1.3 product of three matrices requires: $O(n k^2 + n^2 k) \sim O(n^2)$ if $k \ll n$
 This yields G in: $O(n^3) / 3 + O(k^3) / 3 + O(n k^2 + n^2 k)$

2. The Cholesky decomposition of G to obtain L requires $\sim O(n^3) / 3$
 The subsequent matrix-vector product and a dot-product are of strictly lower order than 3

Therefore, the general algorithm suggested above is of complexity:
 $2O(n^3) / 3 + O(k^3) / 3$.
 The terms in $O(n^2)$ and lower order can be neglected.

The metric of information content requires the calculation of the Cholesky decomposition of the matrix B and the matrix $H^T R^{-1} H + B^{-1}$ and the trace of their logarithms (both linear).

The calculation of B^{-1} from B requires: $O(n^3) / 3$ operations
 The calculation of R^{-1} from R requires: $O(k^3) / 3$ operations
 The calculation of: $H^T R^{-1} H$ requires: $O(n k^2 + n^2 k) \sim O(n^2)$ if $k \ll n$.

Then remains the calculation of the determinant of the Hessian matrix:
 $|H^T R^{-1} H + B^{-1}|$

Which given that it is symmetric and positive definite takes $O(n^3) / 3$ operations. The subsequent logarithm and the trace operations and are linear i.e. $O(n)$.

Therefore, the total complexity yields:
 $O(n^3) / 3 + O(k^3) / 3 + O(n^2) + O(n^3) / 3 + O(n) \sim 2O(n^3) / 3 + O(k^3) / 3$

As the computational complexities of both procedures are comparable, we can state that the information content procedure is not worse than the posterior uncertainty procedure. We would like to underline that in the submitted manuscript we pointed out that the calculation done in this way is more computationally efficient than applying naively the definition of information content to both the posterior and the prior and then calculating separately the determinant for each of those, and to the best knowledge of the authors this is a novel remark of the present work.

We could remark that information content procedure is intended to be applied to a reduced observation space. In contrast, the general procedure proposed above is intended to be applied to the full space of observations first and then from this generic point the space is reduced with the linear form v_t . Any target quantity as a linear function of the posterior sources can be calculated, but the generality of the approach brings the cost of calculating the inverse of the full matrix B . In case of a reduced B the information content metric could be less costly than a generic uncertainty metric applicable to any linear combination of sources. Aggregation of observations would conduce to equivalent reductions in the dimension of R for both algorithms and therefore are not analysed here.

Strassen, V.: "Gaussian Elimination is not Optimal". Numer. Math. 13 (4): 354–356, 1969. doi:10.1007/BF02165411. S2CID 121656251.

Aho, A.V., Hopcroft, J.E. and Ullman, J.D.: "The Design and Analysis of Computer Algorithms". Addison-Wesley, 1974.

Comment

I only have two specific comments on the paper:

1) L45: When citing early literature it is probably fair to cite the paper that gave rise to the field, Hardt et al. 1994

Response

We presume the reviewer is referring to this paper:

Hardt, M. and Scherbaum, F.: The design of optimum networks for aftershock recordings, *Geophysical Journal International*, 117, (3), 1994.

<https://doi.org/10.1111/j.1365-246X.1994.tb02464.x>

In this case, we now refer to this paper in the introduction (L44).

Comment

2) L64: Summing over the submatrix does indeed account for the covariance of uncertainty but that isn't it's most important property. This is that it calculates the uncertainty on the summed regional flux rather than the individual pixels.

Response

We thank the reviewer for pointing this out. We now also mention at L63 that summing over a submatrix of the posterior uncertainty provides the regional uncertainty.