Atmos. Chem. Phys. Discuss., 9, C3329–C3336, 2009
www.atmos-chem-phys-discuss.net/9/C3329/2009/

Atmospheric Chemistry and Physics Discussions

Interactive comment on "*Est modus in rebus*: analytical properties of multi-model ensembles" *by* S. Potempski and S. Galmarini

A. Riccio

angelo.riccio@uniparthenope.it

Received and published: 27 July 2009

Comments on your paper:

'Est modus in rebus: ...'

A. Riccio

First remark

You state:
. . . while previously we did not take any particular assumption on pdf (lines 5-6, page 14278)

I wish to point out that your results (equation (10), page 14273) can also be derived as a particular application of a more general variational principle. Several variants of this principle are known as *maximum entropy principle*, *minimum cross-entropy*, and other names. The application of this principle clearly point out the connection existing between the minimization of a given function and the 'implied' pdf. In the following, I will explain my point of view.

Suppose that the i-th model value, x_i, is a random variable and is as distributed as an (unknown) probability distribution function, p_i. If we know the first, μ, and second central moment, σ_i^2, of our random variable, what we can say about p_i? How information about the mean and variance of our process can be coded into a probabilistic description?

The answer has been clearly outlined by Jaynes [1]. First, introduce the entropy:

$$S = -\sum p_i \ln p_i \tag{1}$$

Next, maximize the Lagrangian

$$L(p, \lambda_1, \lambda_2) = -\sum p_i \ln p_i + \lambda_1 \left(1 - \sum p_i\right) + \lambda_2 \left(\sigma_i^2 - \sum (x_i - \mu)^2 p_i\right) \tag{2}$$

λ_1 and λ_2 are the Lagrangian multipliers.

Note that the knowledge of mean and variance is exactly what you postulate: *In this context we do not need to specify any particular form of the pdf neither for the models nor for the observations. What we need to know however, are biases and variances* (lines 26-28, page 14269).

To find the extremum of L, set $\partial L/\partial p_i = 0$. It is easy to shown that we obtain

$$p_i = \mathrm{e}^{-(1+\lambda_1)} \mathrm{e}^{-\lambda_2 (x_i - \mu)^2} \tag{3}$$

i.e.

$$p(x_i|\mu, \sigma_i^2) \propto \exp\left[-\lambda_2 (x_i - \mu)^2\right] \tag{4}$$

The normalization constant and λ_2 are easily evaluated if the limits of integration are $\pm\infty$; it results in the standard Gaussian pdf

$$p(x_i|\mu, \sigma_i^2) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left[-\frac{1}{2}\frac{(x_i - \mu)^2}{\sigma_i^2}\right] \tag{5}$$

This demonstrates that there is a close link between the normal distribution and quantities characterized solely by their mean and variance.

In the case of m independent model, then equation (5) becomes

$$p(x_1, \ldots, x_m|\mu, \sigma_1^2, \ldots, \sigma_m^2) \propto \exp\left[-\frac{1}{2}\sum_i \frac{(x_i - \mu)^2}{\sigma_i^2}\right] \tag{6}$$

If we introduce the sufficient statistics, $\bar{x} = \sum_i x_i/\sigma_i^2$ and $s^2 = 1/\sum_i 1/\sigma_i^2$, it is not difficult to 'invert' equation (6) and show that

$$p(\mu|\bar{x}, s^2) \propto \exp\left[-\frac{1}{2}\frac{\left(\mu - \bar{x}s^2\right)^2}{s^2}\right] \tag{7}$$

that is, the expected average of model values can be calculated as the average of the m $\{x_1, x_2, \ldots, x_m\}$ values, each weighted by a coefficient given by

$$\alpha_i = \frac{1/\sigma_i^2}{\sum_k 1/\sigma_k^2} \tag{8}$$

which is the same result derived in your equation (10).

In summary, the knowledge of mean and variance (+ the maximum entropy principle) implies a Gaussian distribution, and they are the sufficient statistics for this kind of function.

About the justification of this principled approach, Jaynes states: ... *the maximally non-committal probability distribution is the one with the maximum entropy; hence, of all possible probability distributions we should choose the one that maximizes S* [1].

Again: *By applying MAXENT we mean assigning a distribution* $p_1 \cdots p_m$ *on some 'hypothesis space' by the criterion that it shall maximize the information entropy,* $S = -\sum p_i \ln p_i$*, subject to constraints that express properties we wish the distribution to have, but are not sufficient to determine it* [2].

I'm not surprised at all that you (correctly) recover the same result from the minimization of your Lagrangian function. In a certain sense, you are implying a Gaussian distribution, otherwise it would have introduced additional information.

In a more general context, the maximization of entropy in a variational approach can be justified in a variety of different ways: arguments ranging from information theory [3] to logical consistency [4], and several others besides, all lead to the conclusion that the '$-\sum p_i \ln p_i$' criterion is highly desirable.

You can also have a look at [5] for a recent application of maximum entropy principle in a meteorological context.

Second remark

In my opinion pages 14274 (starting from line 17) to page 14276 could be more clearly written. I suggest:

If the optimal weights are not chosen, what can be said about the statistical properties of the ensemble mean? What conditions guarantee that the variance of the ensemble mean is lower than the lowest model's variance?

It can be demonstrated that models should have similar variances, precisely

$$\frac{\sigma_m^2}{\sigma_1^2} \leq m + 1 \tag{9}$$

where we assumed that we were able to enumerate models in the ascending order of their variances, $\sigma_1^2 \leq \sigma_2^2 \leq \ldots \leq \sigma_m^2$. Condition (9) defines the relation between the maximal and minimal variance of model ensemble dataset. If this condition holds, then the variance of the ensemble mean is lower than the minimum variance

$$V(\bar{x}_m) \leq \sigma_1^2$$

$\langle$ *your considerations follow* $\rangle$.

Equation (9) shows that it would be preferable to aggregate models whose individual variances are not very different (i.e. their relative ratio is close to 1). If it could be guessed that a model has a variance very large and different from the others, than it should be preferable to exclude it from the ensemble mean.

Of course, when models' variances are known, there is no need to exclude any model, since the optimization constraints, given by equations $\langle$ *insert the correct equation number* $\rangle$, assure that a small weight is assigned to a model with a large variance, and the optimal ensemble average variance is always lower than the lowest model variance. Moreover, if a model were excluded, the optimal variance calculated by $m-1$ models is greater than the optimal variance calculated by m models.

. . . *your text follows (from line 14, page 14276)*.

If you accept this suggestion, then similar following sections should be changed accordingly.

Third remark

I strongly agree with your comment: *These concepts or definitions are most of the time ignored, or given for granted thus leaving every reader with his own interpretation and sometime misunderstanding* (lines 3-5, page 14270).

I think that a source of confusion (which is usually not underlined in scientific works) is the difference between ***models variances*** and ***data variances***. Model variances are usually obtained by Monte Carlo techniques or through sensitivity analysis, while data variances are estimated by comparing the expected model output with reference

values, e.g. observations.

Your nice paper concentrates several considerations about the effect of ***models variances*** on the ensemble average, but the reader usually confuses these variances with data variances.

Why these clarification may be critical? Models variances play an important role in defining the number of '***potentially accessible states***', i.e. the number of different states a model is able to predict.

If we consider a model as a nonlinear application, mapping the set of input states onto the set of output states

$$\mathcal{M} : x \longmapsto y$$

where x and y are both random variables, and assume that $y \sim \mathcal{P}(\mu, \sigma^2)$[1], then the integral of function $\mathcal{P}$ over the number of accessible states

$$\int \mathcal{P}(y|\mu, \sigma^2) dy \tag{10}$$

is known with different names, e.g. *evidence* in the statistical literature [6], *partition function* in statistical mechanics [1], etc..

The value of this integral is of great interest if one would select (or average) models results, since it is related to the weight to be assigned in an ensemble averaging procedure.

It is not surprising that the minimization of your Lagrangian assigns a weight proportional to $1/\sigma_i^2$, because an unbiased model (i.e. a model correctly reproducing the corresponding observation in the mean), but unnecessarily powerful (i.e. with a large variance, so that a large number of different states can be predicted) is *penalized* with respect to an unbiased model whose predictions are concentrated around the mean.

[1] this notation means that y is a generic random variable with mean μ and variance σ^2, but nothing is said about the form of the function $\mathcal{P}$.

This is nothing else that a manifestation of the famous Ockham's razor: '*entia non sunt multiplicanda praeter necessitatem*' (entities should not be multiplied beyond necessity). A very delightful discussion about the Ockham's razor can be found in [7].

I'd like if authors can comment on this.

References

Jaynes, E.T., (1957): Information Theory and Statistical Mechanics, The Physical Review, vol. 106(4), pp. 620-630

Jaynes, E.T., (1988): The Relation Of Bayesian And Maximum Entropy Methods. In Maximum-Entropy and Bayesian methods in Science and Engineering. Vol. 1, G.J. Erikson and C.R. Smith Eds., pp. 25-29

Shannon, C.E., (1948): A mathematical theory of communication. Bell. Sys. Tech. J., 27, pp. 379-423 & 623-656

Shore, J.E. and Johnson R.W., (1980): Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. IEEE Trans. Inform. Theory, IT 26, pp. 26-37

Bocquet, M., (2005): Reconstruction of an atmospheric tracer source using the principle of maximum entropy. I: Theory. Q.J.R. Meteorol. Soc., 131, pp. 2191-2208

Hoeting, J.A., D.M. Madigan, A.E. Raftery and C.T. Volinsky, (1999): Bayesian model averaging: A tutorial (with discussion), Stat. Sci., 14, pp. 382-401

McKay, D.J.C., (1991): Bayesian interpolation, Neural Comp., 4, pp. 415-447