Atmos. Meas. Tech. Discuss., 6, C3673–C3680, 2013
www.atmos-meas-tech-discuss.net/6/C3673/2013/

Atmospheric Measurement Techniques Discussions

Interactive comment on "Methods for estimating uncertainty in factor analytic solutions" *by* P. Paatero et al.

P. Paatero et al.

norris.gary@epa.gov

Received and published: 13 December 2013

>This article describes three different methods, i.e. bootstrap analysis (BS), displacement analysis (DISP) and a combination of the two (BS-DISP), for estimating the confidence interval for PMF solutions. In the manuscript the confidence interval is well-described as the combination of the random errors and the rotational uncertainty occurring in bilinear multivariate factor analytic models. The article is very nicely written, clearly structured and totally within the scope of the AMT journal. An important point that is overlooked is that normally PMF users would estimate the amount of rotational uncertainty using the global fpeak tool. It would have been more complete to have discussed this approach as a fourth case in order to show the strengths/limitations of

the fpeak tool.

Response: The reviewer writes: An important point that is overlooked is that normally PMF users would estimate the amount of rotational uncertainty using the global fpeak tool

Yes, we are aware of this practice with Fpeak. However, varying Fpeak explores only a portion of the rotationally available space. Varying Fpeak through both positive and negative values traces a one-dimensional path through the entire rotationally accessible domain. If the accessible domain is essentially one-dimensional, then Fpeak explores the entire domain. In most cases, though, the rotational domain is more complex and exists in many dimensions. and for these cases, the path traced by Fpeak does not reflect the variability of the entire domain. Fpeak will demonstrate that rotational variability is at least this large (this was the original intention in developing the Fpeak tool) but it does not give an upper limit for the extent of this space in all possible directions. This conclusion was expressed in the rotational paper of Paatero et al (2002): "Thus, there is also no guarantee that the true solution is contained in the domain covered by the rotations driven by different values of Fpeak." It is partially because Fpeak explores only a portion of the rotational domain that these new techniques for uncertainty estimation have been developed.

Because of the importance of this question, we have added a short paragraph in the end of section 2.2, as follows:

In literature, atmospheric scientists have used the Fpeak rotational tool of program PMF2 to understand rotational uncertainty of the solution. This practice provides only a lower limit for rotational uncertainty. Specifically, varying the Fpeak parameter traces a one-dimensional path through the rotationally accessible domain. In most cases, though, the rotationally accessible domain is many-dimensional; for these cases, Fpeak will demonstrate only a lower limit for rotational uncertainty (Paatero et al (2002)). Rotational error analysis requires an upper limit, and this is not achievable

by the Fpeak of program PMF-2 nor by the simpler one-parameter Fpeak of program ME-2. DISP, BS, and BS-DISP provide such upper limits.

>Below are listed some suggestions/questions to the manuscript: p. 7597, l.17-20 The idea that the rotational ambiguity scales with the amount of zero entries in G and F is, to my understanding, only valid within the discussion of rotations understood as linear combinations of the entire factor time series or factor profile in G and/or F. What about rotations that involve only a change in one or few entries in G and/or F? For these rotations, whether the neighboring variable in G and/or F shows a zero or not should be irrelevant.

Response: A true "Rotation" is a special transformation of G and F such that it leaves the fitted matrix (G*F) unchanged (see Paatero et al 2002). This necessarily involves linear combinations of the entire factor time series or factor profiles in G and/or F. If only a few factor elements do change, then it is not a "rotation", then G*F necessarily changes. When we consider changes ("perturbations") in one or few entries in G and/or F, then it is irrelevant whether the neighboring variable in G and/or F shows a zero or not, as the reviewer correctly remarks.

>p. 7598, l.28 Why have you chosen dQ 20? Is there a specific reason for this value? If not, then "for example" before dQ 20 could be inserted.

Response: "for example" was inserted before "20 units".

> I can't really follow why this limit should be independent. Shouldn't dQ be always scaled by Qexp to account for the size and the number of factors?

Response: No. It may be counter-intuitive but it is a fact that the appropriate dQ does not depend on the number of degrees of freedom, provided that there are no modeling errors, so that then all data errors are uncorrelated and all specified uncertainties for data values are correct. In the paragraph under consideration here, we specifically say "If there were absolutely no modeling errors...". Thus in this paragraph, it is correct to

discuss a dQ without reference to Qexp.

>Since the PMF users are more used to the discussion based on Q/Qexp, a comparison with this entity could be of interest.

Response: Quite right. In the presence of modeling errors, the appropriate dQ depends on the number of data values and hence on Qexp. In the continuation, we discuss this question and come to the conclusion that dQ should not be assumed to be in a fixed relationship to Qexp. Rather, the relationship will depend on types of data, as stated at end of following pargraph.

>p. 7599, l.22-27 This paragraph describes the uncertainty estimation based on the perturbed original data. On page 7600 line 13 this approach is repeated. I suggest merging this to one paragraph.

Response: The first occasion, p. 7599, l.22-27, discusses the general concept of utilizing perturbations of original data in order to derive error estimates, without narrowing the discussion on any specific method of perturbation. The second occasion, page 7600 line 8, discusses one particular method of perturbation, viz. noise insertion. Later, another perturbation is discussed, viz. bootstrap analysis. In order to clarify this, the second occasion is enhanced, to read like this: "Noise insertion is a computation-intensive variation of the classical error propagation method. In this method, a number (br) of perturbed versions of the original data set are generated in the following way: Each perturbed version is of same dimensions as the original data set."

>p. 7601, l.7 I was asking myself why only the entries of the matrix F have been displaced. Why haven't the entries of the matrix G also been displaced?

Response: We understand the need for G error estimates and they are on our priority list. A new subsection 3.9 has been inserted in order to discuss the problems associated with evaluating G uncertainties. Please see also our responses to reviewers 1 and 2, and the general response, at top of this discussion.

>p. 7601, l.11 and l.17 It is somehow contradictory, since in line 11 DISP is supposed to capture first the data error (random errors) and afterwards the rotational ambiguity in line 17. I would agree with the statement in line 17.

Response: We do not see here any problem. DISP captures both kinds of uncertainties because all factor elements (except the displaced one) are free variables when determining the dQ that is due to the displacement. Thus the deformation of factors comprises both rotations and changes of individual variables. In line 11, we discuss how well DISP may capture uncertainty due to data errors. Inline 17, we remark that DISP captures rotational uncertainty well, essentially independent of assumed data errors.

>p. 7603, l.1 Since the Frobenius norm is explicitly mentioned I would quickly define it. Not all PMF users are aware of this norm.

Response: Definition of Frobenius norm was inserted here.

>p. 7603, l.9-1 I can't really follow this sentence. Does the word "problem" at the end refer to the G-T approach?

Response: No. It refers equally to both methods. We try to say that although the two methods produce different results, neither of them is wrong because they solve different mathematical problems. The sentence has been rewritten in order to clarify the meaning.

>p. 7606, l.14 The fact that only a subset of variables in F and or G is displaced is the biggest limitation for the full estimation of the rotational uncertainty. This is most probably due to the high computational effort involved with such immense model runs. For the reader, it would be interesting to know what is feasible using the aforementioned approaches with the actual technology. In other words, how many tests are necessary or how much time is required for a series of test runs that you carried out? Could one make some suggestions on the number of tests in relation to the size of the data matrix

and/or the number of factors? This is especially important, since the data acquisition techniques are generating more and more input data for e.g. the PMF algorithm. Could such an approach still make sense for data set with e.g. 50'000 to 100'000 points in time? It would also be interesting to address this question to the approach dealing with the perturbed original data for the estimation of the random errors for the model solution that was briefly mentioned in the introduction.

Response: These are important questions that could not be examined in this work. See the general response, at top of this discussion. Some information may be stated here, however. We have inserted a new subsection 3.8 about computational loads of DISP and BS-DISP. The amount of computer time used in the tests reported in this work cannot be estimated because no diary was kept of computer time.

It is not known how many displaced F factor elements are needed in order to examine the entire rotationally accessible domain. At most, this domain may be of dimension P(P-1). Because zeros are present in factors, the dimension may in practice be e.g. of the order of P. Thus the statement by the reviewer "The fact that only a subset of variables in F and or G is displaced is the biggest limitation for the full estimation of the rotational uncertainty" is not necessarily valid for many typical data sets. Thus, in certain cases it might be possible to explore all rotations by DISPing only about P well-chosen F factor elements. – All this is speculative and cannot be included in the paper. It is hoped that colleagues will continue this research. The necessary tools will be available as soon as EPA PMF v5 is released.

>p. 7609, l.20 Could a valid alternative be the internal definition, i.e. within the script file of a slightly higher dQmax, such as 5 or 10%. This could allow finding the maximal increase of e.g. fkj for the given dQmax and as such would avoid the assumptions of interpolation and linearization

Response: No. Such large dQmax would produce, for all or most of fkj, the lower limit =0 and the upper limit approximately = sum of F values in column j of F, = f1j+f2j+...fPj.

The purpose of interpolation and linearization (done automatically by the script) is to find the distance where Q has increased by dQmax. As demonstrated by the simulations, such distance makes sense if there are no modeling errors. If interpolations are forbidden, then the workload increases steeply because then many more tentative displacements are needed in order to determine maximum displacement with reasonable accuracy, such as 10% accuracy.

>p. 7610, l.7 Does DISP alone really comprise both the rotational uncertainty and the random er- rors? I understood that this would be only in combination with BS, i.e. BS-DISP.

Response: Please see our response regarding p. 7601, l.11 and l.17, above. Yes, DISP comprises both. The problem with DISP is caused by not-well-known data uncertainties. If all data uncertainties are badly off, e.g. by a factor of 2, then in the worst case, (with no rotational ambiguity) DISP estimated uncertainties are off by the same factor of 2. The use of BS-DISP helps in this respect, then the assumed data uncertainties play a lesser role.

>p. 7610, l.18 What does “well” describe? The full rotational ambiguity is not captured here, right? If this is the case, then I would state this The _full_ rotational ambiguity _is_ captured for all passive elements if a “sufficient” number of F elements are displaced actively.

Response: The question is only about exploring the entire rotationally accessible domain, cf. discussion of dimension of the rotationally accessible domain, above. If the entire domain is explored, then the full rotational ambiguity is determined for all passive elements. How many are “sufficient”? This depends on factors, how many zeros they contain. No recipe may be given here. Users might wish to perform experiments with different numbers of active elements. In order to clarify this situation, the wording has been enhanced to read “If a sufficient number of F elements are displaced actively, then passive interval estimates reflect rotational ambiguity well for the remaining

passive elements.”

Interactive comment on Atmos. Meas. Tech. Discuss., 6, 7593, 2013.