Atmos. Chem. Phys. Discuss.,
https://doi.org/10.5194/acp-2018-784-RC2, 2018

Atmospheric Chemistry and Physics
Open Access
Discussions

Interactive comment on "Receptor modelling of both particle composition and size distribution from a background site in London, UK – the two step approach" *by* David C. S. Beddows and Roy M. Harrison

P. Paatero (Referee)

pentti.paatero86@gmail.com

Received and published: 5 November 2018

RECEPTOR MODELLING OF BOTH PARTICLE COMPOSITION AND SIZE DISTRIBUTION FROM A BACKGROUND SITE IN LONDON, UK – THE TWO STEP APPROACH

by David C.S. Beddows and Roy M. Harrison

submitted to ACP

This manuscript deals with PMF analyses of "combined" data matrices such as [X Z] where X contains elemental composition profiles of aerosol samples and Z contains aerosol number size distributions measured simultaneously with composition profiles. This is an important problem that occurs often in modern aerosol research. There are specific problems in this task; these problems have not been studied in depth in literature so far.

This manuscript studies one specific combined data matrix and reports a PMF model for this matrix. Thus the ms might deserve publication despite of certain serious problems. These problems are in part related to misunderstandings found in earlier papers that discuss this same topic. For this reason, the present review contains a lengthy general discussion of the task of modeling combined matrices. The specific questions regarding this ms are based on this general discussion.

The ms might also be suitable for publication in the sister Journal AMT, Atmospheric Measurement Techniques. My personal view is slightly in favour of AMT. However, both ACP and AMT seem possible, and this review considers publication in either Journal.

The structure of this review is as follows:

===

The hidden factor, aka Nucleation factor

Miscellaneous

==

Recommendations

There are very many problems of different kinds in this manuscript. For this reason, I hesitantly recommend that this ms should NOT be published by ACP or AMT. However, if it is desired to publish this ms because of the importance of the problem, then a thorough rewriting of the text and mathematical details must be undertaken. I recommend that the following enhancements be performed:

There has apparently been lack of communication between the person(s) who did the actual computations and those who wrote the paper. For this reason, the mathematical description is erratic, chaotic and impossible to understand or replicate. In order to create an accurate description, the person(s) who did the computations should be included in the group of authors. Without such help, it may not be possible to achieve a satisfactory mathematical description of what was done.

The entire mathematical discussion about problems attributed to PMF analysis of joint matrices containing different units is erroneous, based on a widespread misunderstanding. This discussion must be rewritten according to suggestions given below. It might be good to include in the author group somebody familiar with the quantitative mathematical structure of the PMF model. In particular, it seems that lines 79,80 are not based on quantitative understanding of the model. These lines, and other similar sentences, must be removed.

Much of Conclusions must be rewritten so that the claims against using variables with different dimensions/units are replaced by opposite sentences stating e.g. that a joint analysis of matrices of variables with different dimensions/units is not harmed by these differences but unfortunately the opposite was believed to be true when the work was

carried out.

The mathematical description of what was done must be totally rewritten so that systematic matrix-form notation is used. Equations must be corrected and written in correct notation, using correct terminology and correct numbering. Details of PMF modeling must be reported, such as dimensions of matrices, used parameters such as uncertainties of data values, robust/nonrobust, obtained Q values, numbers of observed outliers, unique or multiple minima, and so on.

Rotational questions are an ever-present problem in factor analytic modeling, independently of what programs are used. It is alarming that the word "rotation" does not occur in this manuscript. Pay attention to rotational questions.

There are certain weaknesses in the plan of this work, such as assuming that the rotational status of the original PMF model of X was correct or best possible (see below). These weaknesses cannot be corrected in an enhanced ms but they should be briefly discussed. This is important because otherwise, colleagues following the example of this work will feel the need to replicate everything that was done here, being unaware that some details may not have been optimal.

Enhance figure captions so that readers do not need to guess what is shown. Have the enhanced ms proofread by colleagues. Check also the references. This ms illustrates, once again, how difficult it is to find ones own mistakes and typos.

Notation used in this review.

The notation "[X Z]" indicates here attached or joined matrices, i.e. placing X and Z side by side so that they form one larger matrix.

The notations G(X) and F(X) will indicate factor matrices (G and F) obtained from an individual PMF model of X only, and similarly G(Z) and F(Z) for Z only.

The left and right parts of F, when modeling [X Z], are denoted by F[Xz] and F[xZ].

Q(X) and Q(Z) indicate Q values from separate analyzes of X and Z.

Similarly, Q[Xz] and Q[xZ] denote Q sums computed over elements of X and over elements of Z in the joint analysis of [X Z]. Hence, Q[X Z] = Q[Xz] + Q[xZ].

Total weight of X means the sum of squares of X_ij/s_ij over X, where s_ij is the uncertainty assumed for X_ij. If both X and Z are equally important, and if X and Z are of different sizes, all s_ij reported for the larger matrix should be increased so that total weights of X and Z become approximately equal. This implies a deviation from the general principle of determining weights from std-dev of values.

Background

Before examining this manuscript in detail, it is necessary to discuss the model that it tries to solve and the problems that make this task difficult. It is known that PMF of combined matrices often leads to disappointing results, such that some factors only (or mainly) fit X while other factors only/mainly fit Z. Such result is worthless in cases where X and Z are caused by the same emission sources whose emission profiles should be determined for X and Z.

It is important to realize what advantages may be expected from the joint analysis of X and Z. Three Cases are possible: PMF models computed separately for X and for Z may be valid and rotationally unique for (A) both X and Z, (B) one of them (for X, say), or (C) neither one of them.

Case A: If individually computed factors G(X) and G(Z) are practically identical, then a straight-forward joint model is successful for this case. Then G_[X Z] = G(X) = G(Z). If G(X) and G(Z) are significantly different, however, then the joint model will fail, producing too large residual values and hence too large Q. Such result might be caused e.g. by "common-mode errors" (see below) in X and/or in Z.

Case B: Now a joint model should be specified so that total weight (see Notations, above) of better-analyzed matrix X is significantly higher than total weight of Z. Then X

will "drive the model", and G_[X Z] will be approximately equal to G(X). If a reasonable Q[xZ] is obtained, then it indicates that X and Z are compatible, i.e. a joint PMF model is meaningful. Larger Z residuals and larger Q[xZ] would be obtained e.g. if X and Z do not have common sources or if there are common-mode errors. Then the joint PMF model is not meaningful for the chosen number of factors.

Case C: individual PMF models of both X and Z contain rotational ambiguity and/or other problems such as unidentifiable factors or missing factors. In this case, the approach of Case B cannot be used because the obtained ambiguous rotation, based mostly on X, may not be the best rotation for fitting Z. Ideally, equal total weights should be applied on X and Z, hoping that the best rotation for fitting both will be obtained when rotational information from Z is combined with information from X. Experience shows that quite often, such modeling fails. Few, if any, studies have been made about the reasons of such failures. It must be stressed that these failures must not be ascribed to "different units used in X and Z" (see below). As a first remedy, one might inspect the residuals in order to see if common mode errors are visible. Such errors might be corrected by hand, or by using an enhanced PMF model that automatically corrects for common mode errors. One might also inspect individual variables in order to see if only few variables are causing incompatibility of X and Z. Such variables might be downweighted in order to obtain a better overall model. Of course, one must also consider the possibility that in addition to their joint sources, X and Z may also have one or several unique sources. An enhanced PMF model may be developed for analysing such joint matrices containing common and non-common sources.

Summary of Case C: too little is known about reasons why this case fails. Well-documented case studies are needed. Singular value decompositions of G matrices computed for X, Z, and [X Z] may be useful for demonstrating the root of the problem. Reliable remedies may only be suggested when more is known about the reasons for failures in joint PMF modeling.

Common mode errors

Certain problems in measurements will cause so-called "common mode" errors. E.g. an error in air volume control in an aerosol sampler, when measuring sample i, causes that all aerosol concentrations on row i of X will change by the same fractional amount. Such common mode deviation does not contribute to residuals in customary PMF analysis of such aerosol data. Instead, common mode disturbance of sample i will change all elements of row i of matrix G. In a combined matrix, the other part Z is often measured using another instrument. Then Z may have its own common mode errors, different than those of X. In a joint analysis of X and Z, two independent sets of common mode errors will cause increased residuals when factors are common to X and Z. It appears highly probable that such common mode errors are an important reason for those PMF results where individual factors tend to fit either X or Z but not both.

Joint matrices containing different units

This ms claims that quantitative PMF modeling of a joint matrix [X Z] is not possible if variables in X and Z are measured in different units, such as mass concentration (expressed in mass/airvolume) and particle number concentration (expressed in particles/airvolume). These claims are based on a widespread misunderstanding, as explained in this section.

Customary aerosol PMF models are often scaled so that the sum of all elements in each row of matrix F equals unity. Then factor element F_pj indicates the fraction of species j in profile of source p. With joint matrices containing different units, summation over a row of F is not meaningful. The following workflow should be used instead in order to preserve the quantitative nature of the model:

In PMF (or after PMF), scale factors so that the average of each column of G is scaled ("normalized") to unity. Then elements of F have the following quantitative meaning: F_pj indicates the average contribution of source p to observations in column j, both for species j in matrix X and for species j in Z. The average total amount of all aerosol species in source p is obtained by summing values F_pj over all species j in F[Xz], i.e.

in the part of F corresponding to aerosol matrix X. In this way, the customary interpretation of F_pj as fractions of total may be obtained "off-line" after PMF computations by dividing the F_pj values by their sums taken over F[Xz].

The ms also suggests that presence of other variables (Z) in PMF model somehow makes the model non-quantitative or unreliable:

ms lines 79-80: there can be no confidence as to whether the sources are apportioned by units of number concentration (1/cm3) or any of the other units used in the auxiliary data.

Units may be entirely ignored in PMF modeling if all variables are represented in same units. If different units are present in different columns of matrix X, then the following practice is followed: elements of factor matrix G are pure numbers. Elements in column j of factor matrix F carry the same dimension and unit as column j of data matrix X. In the present case, all elements of left part F[Xz] of factor matrix F will be in mass/airvolume (same as X) while all elements of the right part F[xZ] are in units of number concentration (1/cm3) (same as Z). There is no confusion regarding dimensions or units.

Disturbance of quantitative modeling of X by "other variables" in Z may only be present if Z variables make the fit of X extremely poor, so that Q[Xz] increases to unacceptable levels in comparison to the original Q(X). This can be seen from Eq. (1) which defines PMF model: all values in column j of X are fitted using F factor elements from column j of F only. The "other columns" in F, corresponding to "other variables" in Z, do not enter in the fit of any X variables.

If Q[Xz] remains normal, model of X remains quantitative even when Z is introduced in modeling. However, if introduction of Z requires that number of factors must be increased, then the two models are different. Then rotational uniqueness and interpretatability of the joint model of [X Z] may well be better or worse in comparison to the original model of X only.

On the other hand, G(X) and G[X Z] may appear significantly different even when all Q values are normal. In this sense, including Z may interfere with the fit of X although the new fit of X remains as quantitative (or better) than the original fit of X. Such effect depends on rotational ambiguity of the original PMF fit of X: when Z is introduced, it may "rotate" a rotationally ambiguous model of X so that Z obtains a better fit while Q[Xz] does not increase from Q(X) or increases a little. Such rotation may only occurr if the original model of X is rotationally ambiguous, "non-quantitative". If such ambiguity is not understood by the scientist, it might appear that introduction of other variables "harms" the original model. In contrast, however, modifying the original model of X by a rotation is what is desired when using the joint model: both X and Z should be fitted as well as possible. This effect does not harm the quantitative nature of the model, as long as Q value of X does not grow too much.

Summary of this section: if Q computed over X elements increases significantly when modeling [X Z] instead of X, this indicates that X and Z are not compatible (when assuming this number of factors). Then analysis of [X Z] should be rejected. In all other cases, the joint model of X is equally good or better than the original model of X. If original model is rotationally ambiguous, then factors usually change: G[X Z] is different from G(X) and similarly F[Xz] is different from F(X). These new factors fit X as well as the original factors, thus they are as quantitative as the original factors. The rotation of these new factors takes into account information from matrix Z. In some cases, the new factors are rotationally unique, without any ambiguity. More often, the ambiguity of new factors is less than the original ambiguity.

Discussion of the manuscript

This manuscript suffers badly from almost complete avoidance of equations and mathematical symbols and mathematical notation in general. Also, there are serious problems in the few equations that are present. A more compact and easier to read presentation is obtained if mathematical notation is used as the primary means of communication. It is possible that part of my criticism in this review is simply based on

misunderstanding unclear and/or ambiguous verbal explanations of mathematical concepts.

The ideal of scientific work is repeatability. This ms does not provide facts that might enable repeatability, even in principle. E.g., I could not find dimensions of data matrices or obtained Q values. How were NSD data preprocessed before PMF computations? Using averages or medians? How were outliers handled? How many factors were used in each case? And so on.

The basic assumption of factor analytic modeling is that for each source, chemical profile and size distribution stay constant throughout the measurement campaign. On the other hand, it is well known that whenever nucleation happens, aerosol size distributions do vary. Also, largest particles tend to settle down more during longer transit times. In this work, constancy of size distributions was silently assumed. It might be good to discuss this fundamental question in future versions of this work.

Two-stage PMF model vs. customary PMF model

In the present ms, the goal was to determine the size distributions corresponding to the previously determined aerosol composition sources. It was assumed (on what grounds?) that the rotation of the original PMF result was correct, so that the originally obtained G matrix was deemed suitable for the PMF model of NSD matrix Z. In other words, it was desired that X "drive" the modeling of [X Z]. Essentially, this method corresponded to Case B, discussed above. Apparently, the authors were unaware of the one-stage method suggested for Case B. In hindsight, the best approach might have been to follow both Case B and Case C, especially if there was no positive information confirming that the original PMF model of X was rotationally unique and correct. An enhanced version of the ms should briefly discuss the one-stage possibilities of doing this work according to Case B and/or Case C. The one-stage method, with suitably weighted X and Z, would be easier to explain and much easier to understand. However, it is not reasonable to expect that the work be redone using the one-stage approach.

I understand step 2 so that the computed G factors from step 2 were forced to be practically identical to G factors from step 1. Is this right? If this is right, then step 2 appears to be equivalent to non-negative weighted regression (non-negative weighted linear least squares fit) of matrix Z by columns of matrix G. This should be mentioned. There are easy-to-use computer programs for computing such LS fits. Although PMF may also be used for this fit, using simpler tools would make the process more transparent, so avoiding unnecessary complications. Equations for defining the hidden factor should be given. The verbal definition is hard to understand and I did not manage to understand it.

The hidden factor, aka Nucleation factor

It is a good idea to assume that due to its higher time resolution, the NSD matrix Z may contain factors that are not visible in matrix X of chemical profiles. Unfortunately, the method for defining the hidden factor(s) in Z is questionable. First of all, why did you assume that there is only one hidden factor?

It seems that in stage 2, 6 factors were used. This is not defined (why not) but this is how I understand the ms. Why did you not use in 2nd stage PMF a 7th (and maybe an 8th) factor that may only fit the NSD part of the data matrix? This simple arrangement would determine hidden factor(s) avoiding the bias that non-negativity constraints may introduce in your method (see below). This alternative must be mentioned in a future version of the paper.

The second Equation (3) is incorrectly formulated. Symbol j is used as a summation index on the right side. Then it cannot appear on the left side. There is a symbol "x". It is not defined, what does it mean? The text says: "The Cran R package Non-Linear Minimization (nlm) (R Core Team, 2018) was used to minimise equation 3." You must not say "minimize equation". You must specify the expression that is minimized, and also specify the free variable(s) that are varied in order to minimize. I cannot understand the expression to minimize nor the free variables. For this reason, I cannot

comment more on determining the hidden factor. Maybe it is properly determined, maybe not. This part of the work is certainly not reproducible by others.

Bias: It seems that the second Equation (3) is not applied to all data because of non-negativity constraints (however, there seems to be an error in the constraints, it is impossible to guess what was really intended). When some data are excluded, this creates a bias. It is impossible to know from the outside if this bias was negligible or if it distorted the results. The bias question must be documented.

Miscellaneous

Lines 415-417 in Conclusion: "This generates confidence that the NSD and PM10 factors ascribed to one source are in fact attributable to that same source." This is a very important statement, good!

There are two equations numbered (3). This caused a LOT of trouble when trying to understand the discussion of the "hidden profile" a.k.a. "nucleation profile". The first Equation 3 does not appear correctly on my computer. Possibly, it uses a symbol font that is not present on my computer so that one symbol is not visible. There is also another problem in this equation: symbol "a" is used as summation index, and symbol "a" appears also on left side. A summation index cannot be present on left side. Please check your equations before submitting new versions of the ms. Make sure that the .pdf file contains all non-standard fonts that are used e.g. in equations.

The presentation should be helpful for the reader. The symbols used in text and in equations should be defined. Example: in first Eq. (3), there is symbol j. What does it mean? Is it the index of size bin? Why not help the reader and say so? In second Eq. (3), there is again a symbol j. What is it now? Please update the ms so that symbols are used in a systematic way, in order to help the reader. The following method is recommended in order to avoid confusion with symbols:

For your own use, create a table where each symbol, however trivial, is entered. When needing more symbols, check first with the table if the symbol is already reserved for another use. When you are ready, include short definitions from the table into the ms, either in a table of notation or to the location of first use of each symbol. Use customary matrix element notation whenever possible. In this way, you could avoid using scalar "a" first as an index and then vector a_j as a vector of unknowns.

Description of the linear regression model (section 2.4) is strange. I have never seen that the coefficients are called "gradients". Also, correlations should not be mentioned when discussing linear least squares. It would be best to simply show the equation. I recommend that explanation of regression be omitted, except that the equation, using matrix element notation, should be shown.

Figure 6 is unclear. What is illustrated by the bivariate plots? Figure caption only tells that they are bivariate plots, plotted using the Openair program. Instead of naming the plotting program, it would be more important to define what is plotted vs. what, and what are the dimensions in individual diagrams. After working with the ms for a long time, I tend to guess that the "bivariate plots" might represent NSD concentrations in polar plots of wind direction and wind speed. Why did you not say this? Saving one sentence from the ms may cost hours for your new readers.

Interactive comment on Atmos. Chem. Phys. Discuss., https://doi.org/10.5194/acp-2018-784, 2018.