Referee Comments for Manuscript AMT-2021-51 "A Software Package to Simplify Tikhonov Regularization with Examples for Matrix-Based Inversion of SMPS and HTDMA Data" Markus D Petters

Disclaimer: Other than just some broad principles, this reviewer is not familiar with regularization techniques or the Julia syntax and is therefore ill-equipped to properly review the technical nature of that aspect of this work. Attention is generally focused on other aspects of this paper. Also, the lack of full comprehensive documentation of all the notation used in the equations presented here has frequently hampered a thorough understanding of these equations. However, it is still possible to discern the general meaning of most equations. Equation (10) is a good example of this. The definition of the map() function and the interpretation of the right arrow (→) are not given in the text here. At least the map() function is defined in the Petters (2018) reference. It appears the arrow notation is part of notation for a series or sequence.

This manuscript addresses the important issue of automating the processing of tandem DMA data. The idea of inverting data with regularization is sound. However, there are problems with the forward model of calculating system response from a known input distribution. If these issues can be properly addressed, the resulting software package should prove of great utility.

Major Comments

There appears to be a problem with proper accounting of diffusional losses and broadening of the transfer function, Ω, for DMA 2. Eq. (10) in the form of $\mathbf{A}$ characterizes the transfer through DMA 1 while the equation for $\mathbf{O}$ (line 230) characterizes transfer through DMA 2. As noted in Petters (2018), these two expressions are analogous except for the inclusion of T_c in the former and the limitation of the summation to $k=1$ in the latter. In the DMA, a particle is sized according to its apparent mobility diameter whereas diffusional losses as well as broadening of the transfer function are dependent on the true mobility diameter via particle diffusivity. Given one of these diameters, the particle charge is required to calculate the other and ultimately $T_{size}^{\Lambda,\delta}$. Thus, it is important to sum over all charge states individually to calculate the diffusing transfer through a DMA. As this is not done for the second DMA, the given expression cannot be properly accounting for transfer of multiply charged particles.

The interpretation of Eq. (11) and its components would be greatly facilitated by an explicit indication of the independent parameters of distribution for the input size distribution n^{cn}. Also, the precise form of n^{cn} (*e.g.* dN/dD_p, $dN/d\ln D_p$, or $dN/d\log D_p$) is important. The most obvious set of independent particle parameters would be (true) mobility diameter, D_1, and charge, k. However, it appears that n^{cn} is distributed according to apparent mobility diameter,

D_k, and k in order to have the balance of the equation work out. The apparent mobility diameter is then pre-multiplied by the effective, or apparent, growth factor, $gf_k(z^s, gf_0)$, and then by the ratio of true to apparent mobility diameters, D_1/D_k. However, this ratio is being evaluated at the DMA 1 centroid mobility, z^s, but applied to the Z grid after growth. Since this ratio is a function of size, this does not work out. Also, this means that the input distribution to the $\mathbf{O}$ operator characterizing DMA 2 transfer is in terms of true mobility diameter, in contrast to the $\mathbb{n}^{cn}$ input to DMA 1 and $\mathbf{A}$. All of this switching back and forth between true and apparent mobility diameter seems overly complicated.

Minor Comments and Corrections

line 158: Insert a space between "as" and "x".

line 186: The description of a DMA here is a bit too brief, saying nothing about the flow. Try "Charged particles in <u>a flow between the electrodes</u> are deflected to an exit slit …"

lines 188-189: "The functions … and tandem DMAs ~~is~~ <u>are</u> well understood …"

line 200: "$T\cdot\mathbb{n}$" should be "$T.\cdot\mathbb{n}$" according to Petters (2018). Presumably T is a vector, but this differs from the notation conventions given in lines 87-88.

Eq. (10): Here $T_{size}^{\Lambda,\delta}(k, z^s)$ alone characterizes transfer through the DMA. Evidently the balance of this expression puts this into the required form for later matrix manipulation. Some additional explanation of how this matrix is created from $T_{size}^{\Lambda,\delta}(k, z^s)$ would be useful here. And though perhaps only parentheses may be used in programming, the readability of this equation would be greatly improved by alternating "( )" with "[ ]" and "{ }".

lines 213, 230: Though z^s is defined in lines 223-224, what is z_k^s in the indicated lines? Since z^s is used in Eq. (10), it might be more conveniently defined in line 211 (rather than line 224) along with Z as "… Z is a vector of centroid mobilities, z^s, scanned by the DMA …"

lines 216-217: "The size distribution after passage through the DMA is given by $\mathbb{r} = \mathbf{A}\mathbb{n} + \varepsilon$, where $\mathbb{r}$ is the response function … ." The size distribution exiting the DMA and the response of the detector are not the same thing. The former is usually given as $dN/d\log D_p$ while the latter, as in the case of a CPC, is given by N_{CPC}, a simple number concentration. Also, there is the matter of the detector efficiency as well as the transport efficiency between the DMA and the detector, unless the latter has been subsumed into the DMA transport efficiency. As $\mathbf{A}$ is to later serve as the operator corresponding to transfer through DMA 1 in a tandem DMA setup, $\mathbf{A}\mathbb{n}$ must represent a size distribution, not a response function.

Eq. (11) and following: The double character notation for growth factor as "*gf*" is atypical as far as normal mathematical notation is concerned. It is too easily interpreted as g times f, rather than as a single parameter. And in this draft of the manuscript there is actually extra space between the two letters, increasing the likelihood of the wrong interpretation. However, it is seen that this space is eliminated in Petters (2018) so presumably it can and will be eliminated in the final typeset form. If not for this preexisting work and a strong preference to remain consistent with that, it would be better to change this to a single character form such as simply "g". Also, the reason for the choice of the cn superscript on n for the input distribution is quite obscure. Does that stand for something?

line 230: Given the length and complexity of the expression for the operator $\mathbf{O}$, it would be better placed on a line by itself and numbered.

Discretization: As noted (lines 461-462), the forward model for the TDMA represents a triple integral. The parameters of integration may be denoted as D_i and D_o, the mobility diameters before and after growth, and gf_0, the size-independent growth factor. Though the discretization of these parameters is automated in the software, some discussion of the constraints on this discretization should be included here. For instance, is there a restriction between the number of particle diameter bins and the number of measurement bins? Eq. (15) and the statement (line 248) "The size of $\mathbf{A}_2$ is n^2, …" would imply that the number of gf bins must be equal to the number of measurement bins. Is this a necessary condition and, if so, why?

Figs. 1-4: Frequency vs. Growth Factor: Growth factor gf and its frequency distribution P_{gf} are naturally continuous functions, though the former is (artificially) discretized for the purposes of inversion. Just as the size distribution, n, is explicitly written as dN/dD_p or $dN/d\ln D_\mathrm{p}$ with total integral N, the growth factor frequency distribution is also a derivative, dF/dgf or $dF/d\ln gf$, with total integral $F=1$. However, in the indicated plots, the frequency is plotted as for a parameter with truly discrete values such that the sum of the heights, rather than the areas, of the bars is equal to 1. That is, the height of each bar is given by $(dF/d\ln gf)\cdot\Delta\ln gf$ where $\Delta\ln gf$ is the width of the bar. If the growth factor is discretized such that $\Delta\ln gf$ is constant, then what is plotted is simply a uniformly scaled version of the more traditional $dF/d\ln gf$ plot, though this would normally be versus $\ln gf$. As plotted, the area under these curves is not equal to 1.

Number Concentration vs. Apparent Growth Factor: In these plots, the Apparent Growth Factor is evidently given by

$$gf_\mathrm{app} = D_1\left(z_2^s\right)\Big/D_1\left(z_1^s\right) .$$

The "Concentration" parameter is apparently the first-order inverted number distribution function given by

$$dN_{app}/d\ln D_{p2} = (dN_{app}/d\ln Z_{p2})(d\ln Z_p/d\ln D_{p2}) = (N_{CPC}/\beta_2)(d\ln Z_p/d\ln D_{p2})$$

where $\beta_2 = Q_{aerosol}/Q_{sheath}$ for DMA 2. This is also seen to be a scaled version of the apparent growth factor frequency distribution as

$$dN_{app}/d\ln D_{p2} = N_{t,2}\cdot(dF_{app}/d\ln gf_{app})$$

where $N_{t,2}$ is the total concentration exiting DMA 2. If this is to be compared to the Frequency vs. Growth Factor plot, this would need to be multiplied by $\Delta\ln gf_{app} = \Delta\ln D_1(z_2^s)$. For the two plots to be directly comparable, $\Delta\ln D_1(z_2^s)$ would have to be a constant.

line 315-316: "… the residual is high ~~is~~ if the true input is a broad growth factor frequency distribution …"

lines 332-333: "Errors from scans with low non-zero concentration at the edge of the size distribution propagate back into the inversion at other dry sizes."

line 345: "… a cylindrical DMA column (TSI 3080)." Model "3080" does not specify the actual DMA column. Assuming it is the TSI long DMA, this should be specified as either "TSI 3080L" for the whole system or "TSI 3081" for just the column.

lines 386-387: "… with the timestamp closest to the ~~a~~ scan …" Eliminate "a".

line 419: "… a marine inflow event on March 27−28 2015." Use a date format consistent with the other dates, *i.e*. 27-28 March 2015. However, this date is beyond the limits of the plot in Fig. 6.

line 422: "… 9 February 2015, …". Shouldn't this be 11 February 2015?

Lines 505-513: "The inverted dataset … closure(Mahish et al., 2018)." This is a very long run-on sentence. It needs to be broken up into several sentences.

"Best fit" vs "good fit": Though regularization produces what might be considered a best fit solution to the inversion problem, this does not necessarily imply it is a good fit. It would be best to calculate a fit parameter such as the chi square of the normalized residuals over the degrees of freedom. For a good fit, this should be near 1. That is, the residuals are on the order of what is predicted by Poisson statistics. Values an order of magnitude or more greater than that would suggest some sort of problem either with the dataset or the model.