J. Aerosol Sci., Vol. 22, Suppl. 1, pp. S273–S276, 1991.
Printed in Great Britain.

0021-8502/91 $3.00 + 0.00
Pergamon Press plc

Matrix Factorization Methods for Analysing Diffusion Battery Data

Pentti Paatero, Unto Tapper, Pasi Aalto and Markku Kulmala
University of Helsinki, Department of Physics
Siltavuorenpenger 20D, SF-00170 Helsinki, Finland

Abstract

A special form of Factor Analysis is presented. The method is shown for inverting sequences of diffusion battery data. The term "Positive Matrix Factorization", or PMF, is introduced in order to distinquish the present method from traditional Factor Analysis. Factor Analysis has a well-established meaning in many scientific areas.

Keywords

Factor Analysis, Inversion, Aerosol Dynamics, Diffusion Battery.

Introduction

In the present paper we introduce a factor analytical method for analysing long sequences of diffusion battery (DB) measurements. Let $\boldsymbol{A}$ denote the measured experimental DB data matrix of dimension $n \times m$ (in this case $m = 11$ and $n = 191$). The essence of factor analysis (FA) is to find two suitable smaller matrices $\boldsymbol{G}$ and $\boldsymbol{F}$ so that the product $\boldsymbol{G} \cdot \boldsymbol{F}$ approximates the original data matrix $\boldsymbol{A}$ with reasonable accuracy (see eg. Malinowski and Howery, 1980).

In our approach there are two major differences with the traditional FA: 1) FA may produce both positive and negative entries in the matrices $\boldsymbol{G}$ and $\boldsymbol{F}$, whereas PMF only produces zero or positive entries. 2) In FA it is not possible to utilize error estimates $\boldsymbol{S}_{ij}$ of numbers $\boldsymbol{A}_{ij}$. In PMF, error estimates $\boldsymbol{S}_{ij}$ are used optimally: any inaccurate number $\boldsymbol{A}_{ij}$ does not distort the result if corresponding error estimate $\boldsymbol{S}_{ij}$ is large. There are a few recent papers where better handling of non-negativity constraints has been studied (see eg. Shen and Israël, 1989).

Inversion of diffusion battery measurements has been difficult because a single measurement contains so little information. In PMF it is possible to include inversion step into the computation (see later three matrix model). Then, information of many individual measurements is combined into a more stable result.

The Positive Matrix Factorization (PMF) model

The basic assumption for alysis of DB data with PMF is: The aerosol consists of a sum of r basic size distributions of unknown shape. The shapes are assumed constant with respect to time but the amount of each base distribution varies with time. Mathematically this can be formulated, assuming that the number of base distributions is r,

$$\boldsymbol{A}_{ij} = \boldsymbol{G}_{i1} \cdot \boldsymbol{F}_{1j} + \boldsymbol{G}_{i2} \cdot \boldsymbol{F}_{2j} + \ldots \boldsymbol{G}_{ir} \cdot \boldsymbol{F}_{rj} + \boldsymbol{E}_{ij}, \tag{1}$$

$$\begin{aligned} \boldsymbol{G}_{ik} &\geq 0, \quad & i &= 1, \ldots, n \\ \boldsymbol{F}_{kj} &\geq 0, \quad & j &= 1, \ldots, m \\ & & k &= 1, \ldots, r \end{aligned}$$

where each entry in the matrices $\boldsymbol{F}$ and $\boldsymbol{G}$ is only allowed to take positive values and the matrix $\boldsymbol{E}$ is the difference between the factor model and the measured matrix $\boldsymbol{A}$.

The interpretation for the expression above is that there are n sets of measured values, each consisting of m values assembled into a matrix $\boldsymbol{A}$. In our diffusion battery example n is typically quite large, up to $n \simeq 200$. The number m is 11, corresponding to 11 sampling ports of our diffusion battery. Each row of $\boldsymbol{A}$ is then approximated by r fixed distributions $\boldsymbol{F}_{1j}, \ldots, \boldsymbol{F}_{rj}$ with coefficients $\boldsymbol{G}_{i1}, \ldots, \boldsymbol{G}_{ir}$ multiplying the distributions for making up each row of $\boldsymbol{A}$. In practice, the factors $\boldsymbol{G}_{\bullet k}$ show the time behaviour of each base distribution $\boldsymbol{F}_{k\bullet}$ and the $\boldsymbol{F}$ factors $\boldsymbol{F}_{k\bullet}$ should be of such shapes that could be a result of a diffusion battery measurement: they are the response of DB to each base distribution.

In matrix notation the model can be written as

$$\boldsymbol{A} = \boldsymbol{G} \cdot \boldsymbol{F} + \boldsymbol{E}. \tag{2}$$

The existence of any useful model with a meaningful r is taken as an empirical question: if a model with meaningful r and sufficiently small error $\boldsymbol{E}$ is found, then it can be utilized. Sometimes, finding or not finding a suitable r may be a result in itself. The model is considered good enough if the χ^2 value is small enough. This criterium is approximately

$$Q(G, F) = \sum_i \sum_j (\boldsymbol{E}_{ij}/\boldsymbol{S}_{ij})^2 < m \cdot n, \quad \boldsymbol{E} = \boldsymbol{E}(G, F), \tag{3}$$

where the numbers $\boldsymbol{S}_{ij}$ are the error estimates (standard deviations) for the measured values $\boldsymbol{A}_{ij}$.

The simple factor model (1) does not use the *a priori* information that the DB result must be a positive sum of exponentials. A three matrix factor model (inversion factorization) was constructed for diffusion battery: in matrix notation the model is

$$\boldsymbol{A} = \boldsymbol{G} \cdot \boldsymbol{F} \cdot \boldsymbol{W} + \boldsymbol{E}. \tag{4}$$

The matrix $\boldsymbol{W}$ is now a discretized kernel, it contains exponentials which are the base vectors for representing $\boldsymbol{A}$. In component writing,

$$\boldsymbol{A}_{ij} = \sum_{k=1}^{r} \sum_{h=1}^{w} \boldsymbol{G}_{ik} \cdot \boldsymbol{F}_{kh} \cdot \boldsymbol{W}_{hj} + \boldsymbol{E}_{ij}, \tag{5}$$

$$\begin{aligned} &\boldsymbol{G}_{ik} \geq 0, && i = 1, \ldots, n, \\ &\boldsymbol{F}_{kh} \geq 0, && h = 1, \ldots, w, \\ & && k = 1, \ldots, r, \\ &\boldsymbol{W}_{hj} = e^{-d_h \cdot N_j} && j = 1, \ldots, m, \end{aligned}$$

where the numbers N_j are the numbers of screens in a DB up to and including port j. The coefficients d_h correspond to the slopes of the DB results for different particle sizes, $d_h = 0.01, \ldots, 2$. The dimension of w was set to 12 ($\boldsymbol{W}$ is 12×11), which gave good results.

Remarks on algorithm

Minimization of the sum in (3) under equations (5) is a non-linear least squares fit problem. The numerical task is difficult due to the non-negativity constraints, (near-) singularity of matrix $\boldsymbol{W}$, and the non-linearity caused by products of unknowns $\boldsymbol{G}$ and $\boldsymbol{F}$.

Full details of the algorithm will be published later. Typically the algorithm requires 50 to 70 steps for convergence. Solution of the shown example case took approximately 10 minutes on a 386 computer, when algorithm was in matlab language.

A practical example for three matrix factorization

The example data ($n = 191$, $m = 11$) was measured by letting the diffusion battery (TSI, model 3040) + CNC (TSI, model 3022) run appr. 20 hours in a laboratory room (June, 25–26, 1991),

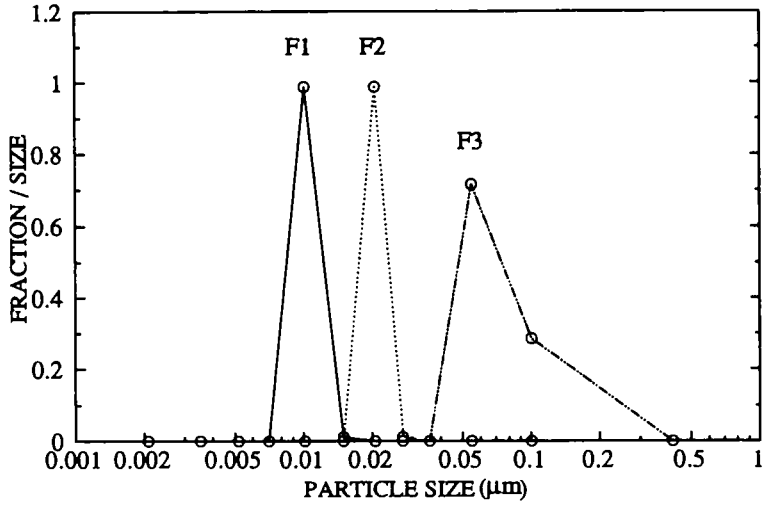

Figure 1. The $\boldsymbol{F}$ factor matrix produced by three matrix inversion PMF of DB data. The size factors are indicated as $\boldsymbol{F}_1$, $\boldsymbol{F}_2$ and $\boldsymbol{F}_3$.

making one measurement set in 6.2 minutes. The background information concerning the on/off times of air circulation, weather conditions during the measurement, etc., is also available. The measurement took place after a rainy period, therefore the particle concentration during the measurement was rather low.

The analysis with three factors is presented. The error estimates of 5% were used for all measured values except for port no.1, 6% error estimates were utilized. During the computation the program was instructed to search a solution having many zeros in $\boldsymbol{F}$. The resulting factorization of three matrix (inversion) PMF is shown in Figs. 1 and 2. The three factors are plotted as $\boldsymbol{F}_1$ to $\boldsymbol{F}_3$ and as $\boldsymbol{G}_1$ to $\boldsymbol{G}_3$. $\boldsymbol{F}_1$ is the size distribution of factor component no.1, and $\boldsymbol{G}_1$ is the time behavior of the factor $\boldsymbol{F}_1$. Thus the first component, written in matrix form, is $\boldsymbol{G}_1 \cdot \boldsymbol{F}_1$. The same applies for other two components.

The adopted result is quite typical for our PMF method when applied to diffusion battery measurements: the aerosol distribution is represented as a sum of three base distributions of "small", "medium" and "large" particles. The factors $\boldsymbol{F}_1$ and $\boldsymbol{F}_2$ are quite trivial or almost monodisperse. One must not make to deep conclusions on such a result. Perhaps it only should be taken as data compression or as extraction of data from noise. The most interesting feature in the result is the almost steady time behavior of the "large particle" factor $\boldsymbol{F}_3$ which consists of accumulation mode particles.

Conclusions

Traditional FA is unsymmetrical: it concentrates either on $\boldsymbol{G}$ or on $\boldsymbol{F}$. Furthermore, rotations must be utilized on matrices $\boldsymbol{G}$ or $\boldsymbol{F}$ in order to make these factors more useful. However, even then the physical interpretation for the resulting factors may be difficult or even impossible. In PMF the resulting factor matrices $\boldsymbol{G}$ and $\boldsymbol{F}$ are considered simultaneously (PMF is symmetrical, conceptually and algorithmically) giving added power through constraints imposed on both $\boldsymbol{G}$ and $\boldsymbol{F}$: there is a clear physical interpretation for resulting factor matrices.

The traditional FA is accused of subjectivity: rotations can be performed in different ways, resulting in different factor structures. In PMF the non-negativity constraints are sometimes enough to produce unique factorization. Sometimes, however, the non-negativity contraints are not enough. Then there is freedom in the factorization: there are many solutions producing the

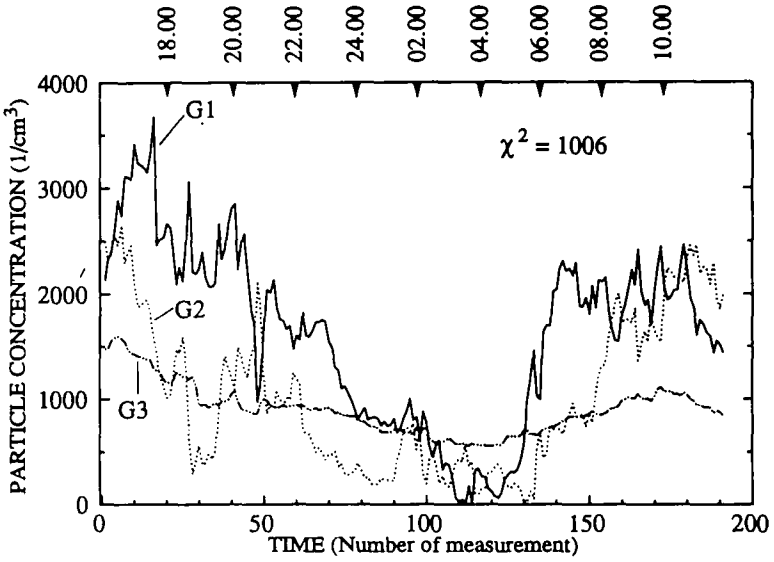

Figure 2. The G factor matrix produced by three matrix PMF of DB data. G_1 is the time behaviour of size distribution F_1, and so on. The actual time scale is indicated in the upper part of the figure.

The traditional FA is accused of subjectivity: rotations can be performed in different ways, resulting in different factor structures. In PMF the non-negativity constraints are sometimes enough to produce unique factorization. Sometimes, however, the non-negativity contraints are not enough. Then there is freedom in the factorization: there are many solutions producing the same χ^2 value. In PMF, the user can specify different criteria for the solutions: the program may favor a "smooth" solution, or a solution with many zeros in G or in F (in our example above, solution with many zeros in F was favoured). In this sense there is also subjectivity in PMF.

The standard practice of experiments has been to measure a stationary situation, with as little variation as possible. With PMF, best results are obtained if there is a fair amount of variation in the experiment. Therefore, PMF is a possible tool for analysing and interpreting aerosol dynamics in field measurements or even in laboratory conditions. Analysing dynamical behaviour of aerosol distribution enables us to make comparisons between computer simulations and in situ measurements.

References

Malinowski, E. R. and D. G. Howery (1980). *Factor Analysis in Chemistry.* Wiley, New York.

Shen, J. and G. W. Israël (1989). A receptor model using a specific non-negative transformation technique for ambient aerosol. *Atmospheric Environment*, 23, pp. 2289–2298.