OPTIMIZATION OF NUMERICAL INVERSION IN PHOTOPOLARIMETIC REMOTE SENSING

OLEG DUBOVIK[1,2]
[1]*Laboratory for Terrestrial Physics, Code 923, NASA Goddard Space Flight Center, Greenbelt, Maryland 20771 USA*
[2]*Also at Goddard Earth Science and Technology Center, University of Maryland Baltimore County, 1000 Hilltop Circle, Baltimore, Maryland 21250 USA*

Abstract. Remote sensing is one primary tool for studying the interactions of solar radiation with the atmosphere and surface and their influence on the Earth radiation balance. During the past three decades the radiation measured from satellite, aircraft and ground have been employed successfully for characterizing radiative properties of land, ocean, atmospheric gases, aerosols, clouds, etc. One of the challenges in implementing remote sensing is the development of a reliable inversion procedure required for deriving information about the atmospheric or surface component interaction with the measured radiation. The inversion is particularly crucial and demanding for interpreting high complexity measurements where many unknowns should be derived simultaneously. Therefore the deployment of remote-sensing sensors of the next generation with diverse observational capabilities inevitably would be coupled with significant investments into inverse-algorithm development. Numerous publications offer a wide diversity of inversion methodologies suggesting somewhat different inversion methods. Such uncertainty in methodological guidance leads to excessive dependence of inversion algorithms on the personalized input and preferences of the developer. This study is an attempt to outline unified principles addressing such important aspects of inversion optimization as accounting for errors in the data used, inverting multi-source data with different levels of accuracy, accounting for *a priori* and ancillary information, estimating retrieval errors, clarifying potential of employing different mathematical inverse operations (e.g. comparing iterative versus matrix inversion), accelerating iterative convergence, etc. The described concept uses the principles of statistical estimation and suggests a generalized multi-term least-square-type formulation

G. Videen et al. (eds.), Photopolarimetry in Remote Sensing, 65-106.
© 2004 Kluwer Academic Publishers, Printed in the Netherlands

that complementarily unites advantages of a variety of practical inversion approaches, such as *Phillips-Tikhonov-Twomey* constrained inversion, *Kalman filters*, *Gauss-Newton* and *Levenberg-Marquardt* iterations, etc. The proposed methodology has resulted from the multi-year efforts of developing inversion algorithms for retrieving comprehensive aerosol properties from ground-based remote sensing observations.

1. Introduction

For the last few decades remote sensing has provided the scientific community with the global distribution of climatically important information about radiative properties of the Earth atmosphere and surface. Future expectations are increasingly high, because remote sensing still has significant potential in improving the volume and accuracy of retrieved information. The indirect nature of observations is an inherent feature of remote-sensing measurements. Indeed, the atmospheric radiation measured from space, ground, etc. is a result of complex interactions of incident solar light with atmospheric components and surface scattering and absorbing radiation. Retrieving optical and radiative properties of natural objects from radiation measurements demands two types of development efforts. First, a capability of modeling atmospheric characteristics is required. That capability is vital for building a so-called "*forward model*" retrieval algorithm that adequately simulates the measured atmospheric radiation coming from the atmospheric or surface objects with known properties. The second necessary component of retrieval is the so-called "*inversion*" procedure that utilizes an inverse transformation by recovering unknown *input* parameters of the *forward model* from known *output* of the *forward model*. Investing in a particular atmospheric remote-sensing approach is motivated by the achievements in atmospheric radiation modeling. Therefore, in remote sensing applications the *forward-model* development usually is feasible and the main challenge is finding the most accurate model satisfying time constraints of operational processing. On the other hand, establishing a strategy for developing the best inversion method is a more convoluted task, in that the evaluation of inversion accuracy is an ambiguous question, especially for the case of simultaneous retrieval of many unknowns; - for example, replacing a scalar light-scattering model with a vector one that accounts for polarization results in accuracy improvements in scattered light reproduction. In contrast, identifying a preference between inversion methods is always rather uncertain. A change of inversion scheme in practical multi-parametric retrieval usually is accompanied by rather complex consequences: retrieval accuracy may improve for one parameter but degrade

for another parameter and that situation may alter for different observation configurations and circumstances. Hence, identifying a preferable inversion method from comparative tests is not always fruitful and should rely on consideration of rather fundamental principles of inversion optimization. However, existence of a very broad diversity of inversion methodologies leaves the researcher freedom in implementing the actual retrieval. Indeed, there are numerous publications describing different inversion techniques and procedures. On the other hand, the comparisons of different inverse methodologies are rather sparse and often limited to a particular application. Consequently, anyone presently designing a practical retrieval algorithm has to review rather fundamental principles of inversion optimization and make a number of principal decisions and choices in inversion implementation that largely predetermine the successes and limitations of the resulting retrieval. Obviously such "personalization" of inversion implementation raises ambiguity and diversification of retrieval development. For that reason, this study is aimed at analyzing the main principles of inversion optimization with an attempt to outline generalized guidance for inversion development in remote sensing. The considerations and results presented are based on the multi-year efforts of developing inversion algorithms for retrieving comprehensive properties of atmospheric aerosol from light-scattering observations. The proposed concept pursues the idea of establishing a unified formulation combining complementary principles of different inversion approaches.

Detailed reviews of inversion methods can be found in various textbooks [1-6]. However, the details given and descriptions of well established inversion procedures do not provide the reader with sufficient explanations as to which method and why it should be chosen for a particular application. Such a situation is partly a result of the fact that most innovations were proposed under pressure of different specific practical needs and, therefore, derived in scopes of rather different approaches. The inversion strategy described here was proposed and refined in the previous studies [7-9]. The approach is focused on clarifying the connection between different inversion methods established in atmospheric optics and unifying the key ideas of these methods into a single inversion procedure. This strategy is expected to be helpful for building optimized and flexible inversion techniques inheriting benefits from of a variety of methods well established in different applications. For example, considerations of this chapter reveal important connections of retrieval algorithms designed with the inversion methods widely adopted in atmospheric remote sensing and other geophysical applications, such as the methods given by *Kalman* [10], *Phillips* [11], *Tikhonov* [12], *Twomey* [13,14], *Strand and Westwater* [15-16], *Chahine* [17], *Turchin and Nozik* [18], *Rodgers* [19], etc.

Following the elaborations by study [9], the following aspects of inversion optimization will be outlined in the order of their importance and validity, starting from most important and most proven: (i) Optimizing the algorithm to the presence of measurement errors; (ii) Optimizing inclusion of *a priori* and ancillary data; (iii) improving performance of key mathematical operations (linear inversion, non-iteration convergence, etc.); and (iv) adjusting conventional assumption of noise-distribution to account for parameter non-negativity and data redundancy. Each of these aspects is discussed in numerous theoretical and practical studies. However, as a rule, theoretical analyses of inversions overemphasize single aspects of retrieval optimization and therefore the resulting conclusions have limited applicability. This study pursues the idea of formulating an inversion procedure based on harmonized consideration of different aspects of algorithm performance rather than on opposing one principle against another. With this purpose, the present chapter outlines the importance of addressing all above-listed aspects by specifying the role of each optimization principle in the development of successful inversion.

2. Basic inversions of linear systems

Commonly, remote sensing methods are set up to derive $N_{\mathbf{a}}$ unknown parameters $\mathbf{a}_i$ from $N_{\mathbf{f}}$ discrete observations $\mathbf{f}_j$ and a corresponding retrieval algorithm should solve the following system of equations:

$$\mathbf{f}^* = \mathbf{f}(\mathbf{a}) + \Delta\mathbf{f}^*, \tag{1}$$

where $\mathbf{f}^*$ is a vector of the measurements f_j, $\Delta\mathbf{f}^*$ is a vector of measurement errors $\Delta f_j^* = f_j^* - f_j^{real}$, $\mathbf{a}$ is a vector of unknowns a_i, $\mathbf{f}(\mathbf{a})$ denotes a physical forward model that allows adequate simulations of observations f_j from predefined parameters a_i. In remote-sensing applications, $\mathbf{f}^*$ usually includes the atmospheric radiation measurements conducted from ground, satellite or aircraft using detectors with various spectral, angular and polarimetric capabilities. The vector of unknowns $\mathbf{a}$ may include various parameters describing the optical properties of atmospheric or surface compounds, such as concentrations of gases and their vertical distributions, parameters describing composition and size distribution of aerosol, land or ocean reflectance, etc. Correspondingly, $\mathbf{f}(\mathbf{a})$ is usually modeled by solving the radiative-transfer equation accounting for transformations of solar radiation interacting with the atmosphere and the surface. Such physical models $\mathbf{f}(\mathbf{a})$ do not have an analytical inverse transformation and the system of Eq. (1) should be solved numerically. For example, in the simplest physical model $\mathbf{f}(\mathbf{a})$ with characteristics $\mathbf{f}_j$ being linearly dependent on $\mathbf{a}_i$ (i.e., $f_i = \sum_{i=1\ldots N_a} K_{ji}\, a_i$), Eq. (1) is reduced to the system of linear equations:

$$\mathbf{f}^* = \mathbf{f} + \Delta\mathbf{f}^* = \mathbf{K}\,\mathbf{a}, \tag{2}$$

where $\mathbf{K}$ is the matrix of the coefficients K_{ji}. If the number of measurements is equal to the number of unknowns, the solution of Eq. (2) is straightforward:

$$\hat{\mathbf{a}} = \mathbf{K}^{-1}\,\mathbf{f}^{*} \qquad (N_{\mathbf{f}} = N_{\mathbf{a}}), \tag{3}$$

where $\mathbf{K}^{-1}$ denotes the inverse matrix operator. For the matrix $\mathbf{K}$ with the linearly independent and non-zero rows, Eq. (3) gives a unique solution that always provides the equality of the left and right sides in Eq. (2), i.e. $\mathbf{f}^{*} = \mathbf{K}\,\hat{\mathbf{a}}$.

Equation (2) also can be solved by other methods without direct implementation of matrix inverse transformation $\mathbf{K}^{-1}$, for example, by means of linear iterations:

$$\mathbf{a}^{p+1} = \mathbf{a}^{p} - \mathbf{H}_{p}(\mathbf{K}\,\mathbf{a}^{p} - \mathbf{f}^{*}). \tag{4}$$

There are a number of the methods that use linear iterations, for example, the known *Jacobi* and *Gauss-Seidel* techniques, steepest descent method, etc. differing by the definition of matrix $\mathbf{H}_{p}$. This matrix should provide convergence of the iterations to a solution $\mathbf{a}^{p+1} \rightarrow \hat{\mathbf{a}}$ attaining equality in Eq. (2):

$$\begin{aligned}\mathbf{K}\,\mathbf{a}^{p+1} - \mathbf{f}^{*} &= \mathbf{K}\,(\mathbf{a}^{p} - \mathbf{H}_{p}\,(\mathbf{K}\,\mathbf{a}^{p} - \mathbf{f}^{*})) - \mathbf{f}^{*} = (\mathbf{I} - \mathbf{K}\mathbf{H}_{p})\,(\mathbf{K}\,\mathbf{a}^{p-1} - \mathbf{f}^{*}) \\ &= (\mathbf{I} - \mathbf{K}\mathbf{H}_{p})\,(\mathbf{I} - \mathbf{K}\mathbf{H}_{p-1})\ldots(\mathbf{I} - \mathbf{K}\mathbf{H}_{0})\,(\mathbf{K}\,\mathbf{a}^{0} - \mathbf{f}^{*}) \Rightarrow \mathbf{0} \quad (\text{for } p \rightarrow \infty),\end{aligned} \tag{4a}$$

where $\mathbf{I}$ is the unity matrix. Thus, the iterations converge from any initial guess $\mathbf{a}^{0}$ to $\hat{\mathbf{a}}$ if the following sequential transformation leads to a zero matrix:

$$(\mathbf{I} - \mathbf{K}\mathbf{H}_{p})\,(\mathbf{I} - \mathbf{K}\mathbf{H}_{p-1})\ldots(\mathbf{I} - \mathbf{K}\mathbf{H}_{0}) \Rightarrow \mathbf{0} \qquad (\text{for } p \rightarrow \infty). \tag{4b}$$

Obviously, if $\mathbf{H}_{p} = \mathbf{K}^{-1}$, Eq. (4) converges at the first iteration and is fully identical to Eq. (3). An important advantage of iterative techniques is that iterations are stable even if Eq. (2) does not have a unique solution. Indeed, if the matrix $\mathbf{K}$ has linearly dependent rows, applying Eq. (2) is problematic, since matrix $\mathbf{K}^{-1}$ does not exist. Iterating Eq. (4) works even in such a case, with the difference that use of Eq. (4) leads to one of many possible solutions that depend on initial guess.

Equation (2) also can be solved by other methods (see [4]) technically different from Eqs. (3)- (4). All these methods are equivalent in the sense that they lead to the same solution $\hat{\mathbf{a}}$ providing the equality in Eq. (2). Therefore, depending on the developer's preference and the requirements of the particular application, any of these methods can be employed in the retrieval algorithm (see discussion in Section 4.8).

3. Solution optimization in presence of measurement errors

In many remote-sensing applications the number of measurements $N_{\mathbf{f}}$ exceeds the number of retrieved parameters $N_{\mathbf{a}}$. This is characteristic of new advanced sensors with multi-spectral, multi-angle [20-23, etc.], and polarimetric

capabilities [24-26, etc.]. In addition to the increased volume of physical information, this redundancy ($N_{\mathbf{f}} > N_{\mathbf{a}}$), allows for minimization of retrieval errors in the presence of random noise in the measurement.

The errors $\Delta\mathbf{f}^*$ in Eq. (2) may have the following two components:

$$\Delta\mathbf{f}^* = \Delta\mathbf{f}^*_{sys} + \Delta\mathbf{f}^*_{ran}, \tag{5}$$

where $\Delta\mathbf{f}^*_{sys}$ – *systematic errors*, which are repeatable in different measurement realizations and $\Delta\mathbf{f}^*_{ran}$ – *random errors*, which differ in the different measurements, i. e.:

$$<\Delta\mathbf{f}^*_{sys}> = \mathbf{b} \neq \mathbf{0} \quad \text{and} \quad <\Delta\mathbf{f}^*_{ran}> = \mathbf{0}, \tag{6}$$

where <…> denotes averaging over measurement realizations, **b** is the average systematic error or so-called *bias*. The correction of the measured data for the *bias* is straightforward provided that **b** is identified and evaluated. The correction of the measurements for random errors is not possible, because their values are unpredictable in each individual act of measurement. Nevertheless, the statistical properties of the random errors can be used to improve the statistical properties of the retrievals.

The statistical properties of random errors are characterized by $\mathbf{P}(\Delta\mathbf{f}^*)$ - *Probability Density Function* (PDF) that provides the probabilities of observing various realizations of the errors $\Delta\mathbf{f}^* = \mathbf{f}^* - \mathbf{f}^{real}$. The retrieved estimates should be close to the real values of unknowns, i.e. $\hat{\mathbf{a}} \approx \mathbf{a}^{real}$. Using an adequate forward model [$\mathbf{f}^{real} = \mathbf{f}(\mathbf{a}^{real})$] the errors $\Delta\mathbf{f}^*$ can be modeled as

$$\Delta\mathbf{f}^* = \mathbf{f}^* - \mathbf{f}(\mathbf{a}^{real}) \approx \Delta\hat{\mathbf{f}}^* = \mathbf{f}^* - \mathbf{f}(\hat{\mathbf{a}}). \tag{7}$$

The known properties of the PDF can be used to improve the solution $\hat{\mathbf{a}}$. Indeed, the modeled measurement errors $\Delta\hat{\mathbf{f}}^* = \mathbf{f}^* - \mathbf{f}(\hat{\mathbf{a}})$ for $\hat{\mathbf{a}} \approx \mathbf{a}^{real}$ should reproduce the known statistical properties of measurement errors as closely as possible. The agreement of modeled $\Delta\hat{\mathbf{f}}^*$ with known error distribution can be evaluated using the known PDF as a function of modeled errors $P(\Delta\hat{\mathbf{f}}^*)$: the higher $P(\Delta\hat{\mathbf{f}}^*)$ the closer the modeled $\Delta\hat{\mathbf{f}}^*$ to the known statistical properties. Thus, the best solution $\hat{\mathbf{a}}^{best}$ should result in modeled errors corresponding to the most probable error realization, i.e. to PDF maximum:

$$P(\Delta\hat{\mathbf{f}}^*) = P(\mathbf{f}(\hat{\mathbf{a}}) - \mathbf{f}^*) = P(\mathbf{f}(\hat{\mathbf{a}}) \,|\, \mathbf{f}^*) = max. \tag{8}$$

In essence, this principle is the well-known *Method of Maximum Likelihood* (MML). The PDF written as a function of measurements $P(\mathbf{f}(\hat{\mathbf{a}}) \,|\, \mathbf{f}^*)$ is called *Likelihood Function*. The MML is one of the strategic principles of statistical estimation that provides statistically the best solution in many senses [27]. For example, the asymptotical error distribution (for infinite number of $\Delta\mathbf{f}^*$ realizations) of MML estimates $\hat{\mathbf{a}}$ have the smallest possible variances of $\Delta\hat{a}_i$.

Most statistical properties of the MML solution remain optimum for a limited number of observations [27].

The implementation of MML in the actual retrieval requires an assumption on PDF of errors $\Delta\mathbf{f}^*$. The normal (or *Gaussian*) function is most appropriate for describing random noise resulting from numerous additive factors:

$$P\left(\mathbf{f}(\mathbf{a})\middle|\mathbf{f}^*\right) = \left((2\pi)^{\mathrm{m}} \det(\mathbf{C})\right)^{-1/2} \exp\left(-\frac{1}{2}\left(\mathbf{f}(\mathbf{a})-\mathbf{f}^*\right)^{\mathrm{T}} \mathbf{C}^{-1}\left(\mathbf{f}(\mathbf{a})-\mathbf{f}^*\right)\right), \qquad (9)$$

where T denotes matrix transposition, $\mathbf{C}$ is the covariance matrix of the vector $\mathbf{f}^*$; det($\mathbf{C}$) denotes the determinant of $\mathbf{C}$, and m is the dimension of the vectors $\mathbf{f}$ and $\mathbf{f}^*$. Detailed discussions on the reasoning for using a normal PDF as the best noise assumption can be found in many textbooks [e.g. 3,27].

The maximum of the PDF exponential term in Eq. (9) corresponds to the minimum of the quadratic form in the exponent. Therefore, the MML solution is a vector $\hat{\mathbf{a}}$ corresponding to the minimum of the following quadratic form:

$$\Psi(\mathbf{a}) = \frac{1}{2}(\mathbf{f}(\mathbf{a})-\mathbf{f}^*)^{\mathrm{T}}\mathbf{C}^{-1}(\mathbf{f}(\mathbf{a})-\mathbf{f}^*) = \min. \qquad (10)$$

Thus, with the assumption of normal noise $\Delta\mathbf{f}^*$, the MML principle requires the search for a minimum in the product of the squared terms of ($\mathbf{f}(\mathbf{a})$ - $\mathbf{f}^*$) in Eq. (10). This is the basis for the widely known *Least Square Method* (LSM). The minimum of the quadratic form $\Psi(\mathbf{a})$ corresponds to a point with a zero gradient $\nabla\Psi(\mathbf{a})$, i.e. to a point where all partial derivatives of $\Psi(\mathbf{a})$ are equal to zero:

$$\nabla\Psi(\mathbf{a}) = \frac{\partial\Psi(\mathbf{a})}{\partial a_i} = \mathbf{0}, \quad (i = 1,..,N_a). \qquad (11a)$$

The gradient of $\Psi(\mathbf{a})$ can be written as (detailed derivations can be found elsewhere [28-29, etc.]):

$$\nabla\Psi(\mathbf{a}) = \mathbf{K}_a^{\mathbf{T}}\mathbf{C}^{-1}\,\mathbf{f}(\mathbf{a}) - \mathbf{K}_a^{\mathbf{T}}\mathbf{C}^{-1}\,\mathbf{f}^* = \mathbf{K}_a^{\mathbf{T}}\mathbf{C}^{-1}\,(\mathbf{f}(\mathbf{a}) - \mathbf{f}^*). \qquad (11b)$$

$\mathbf{K}_a$ is a matrix of first partial derivatives in vicinity of $\mathbf{a}$, i.e. $\{\mathbf{K}_a\}_{ji} = \partial f_j / \partial a_i\big|_{\mathbf{a}}$. Correspondingly, for linear forward models, ($\mathbf{f}(\mathbf{a}) = \mathbf{K}\ \mathbf{a}$), Eq. (11a) is equivalent to the following system:

$$\mathbf{K}^{\mathbf{T}}\mathbf{C}^{-1}\,\mathbf{K}\,\mathbf{a} = \mathbf{K}^{\mathbf{T}}\mathbf{C}^{-1}\,\mathbf{f}^*. \qquad (11c)$$

Using matrix inversion, the LSM solution can be written as

$$\hat{\mathbf{a}} = (\mathbf{K}^{\mathbf{T}}\mathbf{C}^{-1}\,\mathbf{K})^{-1}\ \mathbf{K}^{\mathbf{T}}\mathbf{C}^{-1}\,\mathbf{f}^*. \qquad (12)$$

This formula is valid if Eq. (2) has a unique solution, i.e. if det($\mathbf{K}^{\mathbf{T}}\mathbf{C}^{-1}\,\mathbf{K}$) $\neq 0$. The errors $\Delta\hat{\mathbf{a}}$ of the estimate $\hat{\mathbf{a}}$ are normally distributed and have random and systematic components resulting from $\Delta\mathbf{f}^*_{\mathrm{sys}}$ and $\Delta\mathbf{f}^*_{\mathrm{ran}}$ in the measurements:

$$\Delta\hat{\mathbf{a}} = \Delta\hat{\mathbf{a}}_{ran} + \Delta\hat{\mathbf{a}}_{sys} = (\mathbf{K}^{\mathbf{T}}\mathbf{C}^{-1}\mathbf{K})^{-1}\mathbf{K}^{\mathbf{T}}\mathbf{C}^{-1}(\Delta\mathbf{f}^*_{ran}+\Delta\mathbf{f}^*_{sys}). \quad (13)$$

As follows from Eq. (6) the mean $<\Delta\hat{\mathbf{a}}>$ is the resultant of measurement *bias*:

$$\hat{\mathbf{a}}_{bias} = <\Delta\hat{\mathbf{a}}> = <\Delta\hat{\mathbf{a}}_{sys}> = (\mathbf{K}^{\mathbf{T}}\mathbf{C}^{-1}\mathbf{K})^{-1}\mathbf{K}^{\mathbf{T}}\mathbf{C}^{-1}\mathbf{b}. \quad (14)$$

The covariance matrices of the estimate errors $\Delta\hat{\mathbf{a}}$ also have random and systematic components:

$$\mathbf{C}_{\hat{\mathbf{a}}} = \mathbf{C}_{\Delta\hat{\mathbf{a}}(ran)} + (\hat{\mathbf{a}}_{bias})(\hat{\mathbf{a}}_{bias})^T = (\mathbf{K}^{\mathbf{T}}\mathbf{C}^{-1}\mathbf{K})^{-1} + (\hat{\mathbf{a}}_{bias})(\hat{\mathbf{a}}_{bias})^T. \quad (15)$$

This equation is derived as follows:

$$\mathbf{C}_{\hat{\mathbf{a}}} = <\Delta\hat{\mathbf{a}}(\Delta\hat{\mathbf{a}})^{\mathbf{T}}> = <(\Delta\hat{\mathbf{a}}_{ran} + \Delta\hat{\mathbf{a}}_{sys})(\Delta\hat{\mathbf{a}}_{ran} + \Delta\hat{\mathbf{a}}_{sys})^T> =$$

$$= <\Delta\hat{\mathbf{a}}_{ran}(\Delta\hat{\mathbf{a}}_{ran})^{\mathbf{T}}> + <\Delta\hat{\mathbf{a}}_{sys}(\Delta\hat{\mathbf{a}}_{sys})^T> = \mathbf{C}_{\Delta\hat{\mathbf{a}}(ran)} + (\hat{\mathbf{a}}_{bias})(\hat{\mathbf{a}}_{bias})^T,$$

where

$$\mathbf{C}_{\Delta\hat{\mathbf{a}}(ran)} = <\Delta\hat{\mathbf{a}}_{ran}(\Delta\hat{\mathbf{a}}_{ran})^T> = <(\mathbf{K}^{\mathbf{T}}\mathbf{C}^{-1}\mathbf{K})^{-1}\mathbf{K}^T\mathbf{C}^{-1}\Delta\mathbf{f}^*_{ran}((\mathbf{K}^T\mathbf{C}^{-1}\mathbf{K})^{-1}\mathbf{K}^T\mathbf{C}^{-1}\Delta\mathbf{f}^*_{ran})^T>$$

$$= (\mathbf{K}^{\mathbf{T}}\mathbf{C}^{-1}\mathbf{K})^{-1}\mathbf{K}^{\mathbf{T}}\mathbf{C}^{-1} <\Delta\mathbf{f}^*_{ran}(\Delta\mathbf{f}^*_{ran})> \mathbf{C}^{-1}\mathbf{K}(\mathbf{K}^{\mathbf{T}}\mathbf{C}^{-1}\mathbf{K})^{-1} = (\mathbf{K}^{\mathbf{T}}\mathbf{C}^{-1}\mathbf{K})^{-1}.$$

The optimality of LSM solution $\hat{\mathbf{a}}$ is given by the *Cramer-Rao* inequality [27]:

$$<(\Delta g)^2> = <\mathbf{g}^T\Delta\tilde{\mathbf{a}}(\mathbf{g}^T\Delta\tilde{\mathbf{a}})^T> = \mathbf{g}^T\Delta\tilde{\mathbf{a}}(\Delta\tilde{\mathbf{a}})^T\mathbf{g} = \mathbf{g}^T\mathbf{C}_{\tilde{\mathbf{a}}}\mathbf{g} \geq \mathbf{g}^T\mathbf{C}_{LSM}\mathbf{g}, \quad (16)$$

where $\tilde{\mathbf{a}}$ denotes any estimate of the vector $\mathbf{a}$ with covariance of random errors $\mathbf{C}_{\tilde{\mathbf{a}}}$, $\mathbf{C}_{LSM}$ is the covariance matrix of LSM estimates [Eq. (15)], g is a characteristic linear dependence on $\mathbf{a}$ (i.e. $g = \mathbf{g}^T\mathbf{a}$, $\mathbf{g}$ is a vector of coefficients). Thus, according to the *Cramer-Rao* inequality the LSM estimates $\hat{\mathbf{a}}$ have the smallest variances of random errors and, moreover, the estimate $\mathbf{g}^T\hat{\mathbf{a}}$ of any function g obtained using $\hat{\mathbf{a}}$ also has the smallest variance determined by Eq. (16), i.e. any product $\mathbf{g}^T\hat{\mathbf{a}}$ of the LSM estimates $\hat{\mathbf{a}}$ is also optimum. These accuracy limits are related to the definition of *Fisher* information [27].

The values of minimized quadratic form of $\Psi(\hat{\mathbf{a}})$ [Eq. (10)] follows a χ^2 distribution with *m-n* degrees of freedom, i.e. the mean minimum is [27,30]:

$$\langle 2\Psi(\mathbf{a})\rangle = \left\langle (\hat{\mathbf{f}} - \mathbf{f}^*)^T\mathbf{C}^{-1}(\hat{\mathbf{f}} - \mathbf{f}^*)\right\rangle = \Delta\mathbf{f}^{*T}\mathbf{C}^{-1}\Delta\mathbf{f}^* - \Delta\hat{a}^T\mathbf{C}_{\hat{a}}^{-1}\Delta\hat{a} = N_{\mathbf{f}} - N_a, \quad (17)$$

where $\hat{\mathbf{f}} = \mathbf{f}(\hat{\mathbf{a}})$, $N_{\mathbf{f}} = rank(\mathbf{C})$ and $N_{\mathbf{a}} = rank(\mathbf{C}_{\hat{\mathbf{a}}})$. For the case when both measurement and estimate vectors have only statistically independent elements then $N_{\mathbf{f}}$ is equal to the number of measurements and $N_{\mathbf{a}}$ is equal to the number of retrieved parameters. The statistical property given by Eq. (17) is used to validate the assumptions on the noise $\Delta\hat{\mathbf{f}}^*$ in measurement and accuracy of the forward model (see Sections 4,6 and 7).

4. *A priori* constrains

In spite of its optimization properties, the basic LSM given by Eq. (12) is not often used in remote sensing. The modeling of interactions of solar light with the atmosphere and the surface requires a complex theoretical formalism with a large number of internal parameters. The integrative character and confines in viewing geometries limit the sensitivity of remote measurements to unique variations of each internal parameter in the radiative model. Therefore, the remote sensing of natural objects, in general, is inherently underdetermined and belongs to a class of so-called *ill-posed* problems. In fact, the frequent appearance of ill-posed problems in remote sensing and applied optics stipulated the development of methodologies that constrain standard inversion algorithms in order to overcome solution instability.

In terms of considerations given in Section 3, ill-posed problems have a non-unique and/or an unstable solution. For a non-unique solution, the matrix $\mathbf{K}^T\mathbf{C}^{-1}\mathbf{K}$ on the left side of Eq. (11c) has linearly dependent rows (and columns, since it is symmetrical), i.e. $\det(\mathbf{K}^T\mathbf{C}^{-1}\mathbf{K})=0$ (degenerated matrix) and the inverse operator $(\mathbf{K}^T\mathbf{C}^{-1}\mathbf{K})^{-1}$ does not exist. For a quasi-degenerated matrix ($\det(\mathbf{K}^T\mathbf{C}^{-1}\mathbf{K})\rightarrow 0$) the inverse operator $(\mathbf{K}^T\mathbf{C}^{-1}\mathbf{K})^{-1}$ exists. However, in this case, the covariances of the retrieval errors [Eq. (15)] become large due to uncertainty of the inverse operator:

$$\{\mathbf{C}_{\hat{\mathbf{a}}}\}_{ii} \sim \{(\mathbf{K}^T\mathbf{C}^{-1}\mathbf{K})^{-1}\}_{ii} \rightarrow \infty \quad (\text{for } \det(\mathbf{K}^T\mathbf{C}^{-1}\mathbf{K}) \rightarrow 0). \tag{18}$$

4.1 Basic formulations for constrained inversion

Constraining inversions by *a priori* information is an essential tool for achieving a unique and stable solution of an ill-posed problem. Most remote-sensing inverse techniques are based on the following equations:

$$\hat{\mathbf{a}} = (\mathbf{K}^T\mathbf{K} + \gamma\,\mathbf{\Omega})^{-1}\,\mathbf{K}^T\mathbf{f}^*, \tag{19}$$

$$\hat{\mathbf{a}} = (\mathbf{K}^T\mathbf{K} + \gamma\,\mathbf{I})^{-1}\,(\mathbf{K}^T\mathbf{f}^* + \gamma\,\mathbf{a}^*). \tag{20}$$

These equations originated the papers by *Phillips* [11], *Tikhonov* [12] and *Twomey* [13]. Equation (19) constrains the solution $\hat{\mathbf{a}}$ by minimizing its k-th differences Δ^k :

$$\begin{aligned} \Delta^1 &= \hat{a}_{i+1} - \hat{a}_i\,, && (k=1),\\ \Delta^2 &= \hat{a}_{i+2} - 2\,\hat{a}_{i+1} + \hat{a}_i\,, && (k=2),\\ \Delta^3 &= \hat{a}_{i+3} - 3\,\hat{a}_{i+2} + 3\,\hat{a}_{i+1} - \hat{a}_i\,, && (k=3). \end{aligned} \tag{21a}$$

The minimization of differences in Eq. (19) usually is considered [e.g. 12,13,18] to be an implicit constraint on derivatives. The correspondent smoothness matrix $\mathbf{\Omega}$ in Eq. (19) can be written as

$$\mathbf{\Omega} = (\mathbf{S}_k)^T(\mathbf{S}_k)\ , \tag{21b}$$

where $\mathbf{S}_k$ is the matrix of the k-th differences (i.e. $\mathbf{\Delta}^k = \mathbf{S}_k\ \hat{\mathbf{a}}$). For example, $\mathbf{S}_2$ (k=2) is:

$$\mathbf{S}_2 = \begin{pmatrix} 1 & -2 & 1 & 0 & \ldots & & & & \\ 0 & 1 & -2 & 1 & 0 & \ldots & & & \\ 0 & 0 & 1 & -2 & 1 & 0 & \ldots & & \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & 0 & 1 & -2 & 1 \end{pmatrix}. \tag{21c}$$

Correspondingly, for minimization of the second differences (introduced by *Phillips* [11]), the required smoothing matrix (written by *Twomey* [13]) is

$$\Omega = \begin{pmatrix} 1 & -2 & 1 & 0 & 0 & \ldots & & & & \\ -2 & 5 & -4 & 1 & 0 & 0 & \ldots & & & \\ 1 & -4 & 6 & 4 & 1 & 0 & 0 & \ldots & & \\ 0 & 1 & -4 & 6 & 4 & 1 & 0 & 0 & \ldots \\ & & & & \ldots & & & & \\ & & & \ldots & 0 & 1 & -4 & 5 & -2 \\ & & & & \ldots & 0 & 1 & -2 & 1 \end{pmatrix}. \tag{21d}$$

Twomey [13] also suggested employing the third differences in Eq. (19). *Tikhonov's* formulations consider a generalized definition of the smoothing or "regularization" function that usually is formulated via limitations on magnitudes of unknowns and/or their first differences, i.e. the smoothness matrix $\mathbf{\Omega}$ defined as a sum of the unity matrix (matrix of "zero differences") and matrix of first differences or one of these matrices simply used alone [2,5,12,31]. Equation (20) formulated by *Twomey* [13] constrains the solution $\hat{\mathbf{a}}$ to its *a priori* estimates $\mathbf{a}^*$ (i.e. formally constrains "zero differences"). The *Lagrange* multiplier γ in Eqs. (20-21) is defined as a nonnegative parameter that controls the strength of *a priori* constraints relative to the contribution of the measurements. The value of γ usually is evaluated by numerical tests and sensitivity studies (Section 4.6).

Unlike LSM, Eqs. (20-21) were derived without direct consideration of noise statistics. Nevertheless, Eqs. (20-21) are based on the minimization of quadratic norms of deviations ($\mathbf{f}(\mathbf{a}) - \mathbf{f}^*$) which is formally equivalent to assuming normal noise with unit covariance matrix (i.e. $\mathbf{C} = \mathbf{I}$). Thus, Eqs. (20-21) minimize the quadratic forms that have the additional term

$$2\Psi'(\hat{\mathbf{a}}) = (\mathbf{f}(\hat{\mathbf{a}}) - \mathbf{f}^*)^T(\mathbf{f}(\hat{\mathbf{a}}) - \mathbf{f}^*)\ + \gamma\ \hat{\mathbf{a}}^T\Omega\ \hat{\mathbf{a}} = \min, \tag{22a}$$

$$2\Psi'(\hat{\mathbf{a}}) = (\mathbf{f}(\hat{\mathbf{a}}) - \mathbf{f}^*)^T(\mathbf{f}(\hat{\mathbf{a}}) - \mathbf{f}^*) + \gamma\ (\hat{\mathbf{a}} - \mathbf{a}^*)^T(\hat{\mathbf{a}} - \mathbf{a}^*) = \min. \tag{22b}$$

The inclusion of a second *a priori* term [compare to Eq. (10)] into the minimization process results in the fact that Eqs. (19)-(20) provide stable solutions even for ill-posed problems ($\det(\mathbf{K}^T\ \mathbf{K}) \rightarrow 0$). Formally it can be

explained by the fact that an addition of the diagonal (**I**) or quasi-diagonal (**Ω**) matrices to ($\mathbf{K}^T\ \mathbf{K}$) results in non-degenerated matrices: $\det(\mathbf{K}^T\ \mathbf{K} + \gamma\ \mathbf{I}) > 0$ and $\det(\mathbf{K}^T\mathbf{K} + \gamma\mathbf{\Omega}) > 0$.

In atmospheric remote-sensing applications, the statistical interpretation of constrained inversion often is associated with the studies by *Strand and Westwater* [15-16] and *Rodgers* [19] and the following formulations:

$$\hat{\mathbf{a}} = (\mathbf{K}^T\mathbf{C}^{-1}\ \mathbf{K} + \mathbf{C}_{a*}^{-1})^{-1}\ (\mathbf{K}^T\mathbf{C}^{-1}\ \mathbf{f}^* + \mathbf{C}_{a*}^{-1}\ \mathbf{a}^*) \quad \text{and} \tag{23}$$

$$\hat{\mathbf{a}} = \mathbf{a}^* - \mathbf{C}_{a*}\mathbf{K}^T\ (\mathbf{C} + \mathbf{K}\mathbf{C}_{a*}\mathbf{K}^T)^{-1}\ (\mathbf{K}\mathbf{a}^* - \mathbf{f}^*), \tag{24}$$

where $\mathbf{a}^*$ is a normally distributed vector of *a priori* estimates called a "virtual solution" [19]. Equation (23) has an obvious similarity to Eqs. (19-20). Indeed, Eq. (23) can be transformed to Eq. (19) if $\mathbf{C}_{a*}^{-1} = \mathbf{\Omega}$ and $\mathbf{a}^* = \mathbf{0}$ and to Eq. (20) if $\mathbf{C}_{a*} = (1/\gamma)\,\mathbf{I}$. It should be noted that there are other statistical formulas for constrained inversions, for example, a statistical equivalent of Eq. (19) is discussed in studies [18, 32].

Equation (24) is fully equivalent to Eq. (23). This type of constrained inversion is popular (see [19]) in applications of satellite remote sensing for retrieving vertical profiles of atmospheric properties (pressure, temperature, gaseous concentrations, etc.). Equation (24) is also widely used in engineering (e.g. see textbook [30]) and other applications [33], such as assimilation of geophysical parameters [34], where Eq. (24) is known as a "*Kalman filter*" named after the author [10] who originated the technique.

The main difference between Eq. (23) and Eq. (24) is that the matrix ($\mathbf{K}^T\mathbf{C}^{-1}\mathbf{K} + \mathbf{C}_{a*}^{-1}$) inverted in Eq. (23) has dimension $N_\mathbf{f}$ (number of measurements) while ($\mathbf{C} + \mathbf{K}\mathbf{C}_{a*}\mathbf{K}^T$) inverted in Eq. (24) has the dimension $N_\mathbf{a}$ (number of retrieved parameters). Therefore, Eq. (23-24) are fully equivalent for the situation when $N_\mathbf{f} = N_\mathbf{a}$ and Eq. (23) generally is preferable for inverting redundant measurements ($N_\mathbf{f} > N_\mathbf{a}$); whereas, Eq. (24) is preferable when the measurement set is underdetermined ($N_\mathbf{f} < N_\mathbf{a}$). Indeed, in Eq. (23) [similarly as for Eqs. (19-20)], the estimate $\hat{\mathbf{a}}$ mostly is determined by the measurement term $\mathbf{K}^T\ \mathbf{C}^{-1}\ \mathbf{f}^*$ and minor *a priori* terms only are expected to provide uniqueness and stability of the solution. In contrast, in Eq. (24) the solution $\hat{\mathbf{a}}$ is expressed in the form of an *a priori* estimate $\mathbf{a}^*$ corrected or "filtered" by measurements, which is the situation when the small number measurements $N_\mathbf{f}$ ($N_\mathbf{f} < N_\mathbf{a}$) cannot fully determine the set of unknowns **a,** but can improve *a priori* assumed values $\mathbf{a}^*$.

4.2 Statistically optimized inversion of multi-source data

The similarities of the formulas for constrained inversion with basic non-constrained LSM were mentioned already in the previous section. This section further explores the use of statistical principles for implementing constrained

inversion by formulating a statistically optimized inversion of multi-source data that follows the developments [7-9]. Such an approach allows generalizing various inversion formulas into a single formalism.

Formally, both measured and *a priori* data can be written as

$$\mathbf{f}_k^* = \mathbf{f}_k(\mathbf{a}) + \Delta\mathbf{f}_k^*, \quad (k=1, 2,\ldots,K), \tag{25}$$

where index k denotes different data sets ("sources"). This assumes that the data from the same source have similar error structure independent of errors in the data from other sources. For example, direct Sun and diffuse sky radiances have different magnitudes and are measured by sensors with a different sensitivity, i.e., errors should be independent (due to different sensors) and likely have different magnitudes. Similarly, *a priori* data are independent of the measurements, i.e. they have errors with a different level of accuracy uncorrelated with remote-sensing errors. Formally, the statistical independence of $\mathbf{f}_k^*$ means that the covariance matrix of joint data $\mathbf{f}^*$ has array structure:

$$\mathbf{f}^* = \begin{pmatrix} \mathbf{f}_1^* \\ \mathbf{f}_2^* \\ \ldots \\ \mathbf{f}_K^* \end{pmatrix} \quad \text{and} \quad \mathbf{C}_{\mathbf{f}^*} = \begin{pmatrix} \mathbf{C}_1 & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_2 & \mathbf{0} & \mathbf{0} \\ \ldots & \ldots & \ldots & \ldots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{C}_K \end{pmatrix}, \tag{26}$$

where $\mathbf{f}^*$ is a vector-column with $(\mathbf{f}^*)^{\mathrm{T}} = (\mathbf{f}_1^*, \mathbf{f}_2^*, \ldots, \mathbf{f}_K^*)^{\mathrm{T}}$ and $\mathbf{C}_k$ is the covariance matrix of the k-th data set $\mathbf{f}_k^*$. Thus, from the formal viewpoint, the only difference of Eq. (25) from Eq. (1) is that Eq. (25) explicitly outlines an expectation of an array structure for the covariance matrix $\mathbf{C}_{\mathbf{f}^*}$. Such explicit differentiation of the input data makes the retrieval more transparent because the statistical optimization of the retrieval is driven by a covariance matrix of random errors. It should, be noted that Eq. (25) does not assume any relations between forward models $\mathbf{f}_k(\mathbf{a})$, i.e. forward models $\mathbf{f}_k(\mathbf{a})$ can be the same or different.

Following Eq. (26), the PDF of joint data $\mathbf{f}^*$ can be obtained by the simple multiplication of the PDFs of data from all K sources:

$$\mathrm{P}\left(\mathbf{f}(\mathbf{a})\middle|\mathbf{f}^*\right) = \mathrm{P}\left(\mathbf{f}_1(\mathbf{a}),\ldots,\mathbf{f}_K(\mathbf{a})\middle|\mathbf{f}_1^*,\ldots,\mathbf{f}_K^*\right) = \prod_{k=1}^{K} \mathrm{P}\left(\mathbf{f}_k(\mathbf{a})\middle|\mathbf{f}_k^*\right). \tag{27}$$

Then, under the assumptions of a normal PDF, one can write

$$\mathrm{P}\left(\mathbf{f}(\mathbf{a})\middle|\mathbf{f}^*\right) = \prod_{k=1}^{K} \mathrm{P}\left(\mathbf{f}_k(\mathbf{a})\middle|\mathbf{f}_k^*\right) \sim \exp\left(-\frac{1}{2}\sum_{k=1}^{K}\left(\mathbf{f}_k(\mathbf{a}) - \mathbf{f}_k^*\right)^{\mathrm{T}}(\mathbf{C}_k)^{-1}\left(\mathbf{f}_k(\mathbf{a}) - \mathbf{f}_k^*\right)\right). \tag{28}$$

Thus, for multi-source data, the LSM condition of Eq. (10) can be written as

$$2\Psi(\mathbf{a}) = \sum_{k=1}^{K}\left(\mathbf{f}_k(\mathbf{a}) - \mathbf{f}_k^*\right)^{\mathrm{T}}(\mathbf{C}_k)^{-1}\left(\mathbf{f}_k(\mathbf{a}) - \mathbf{f}_k^*\right) = \min. \tag{29}$$

This condition does not prescribe the value of the minimum and, therefore, Eq. (29) can be formulated via weight matrices:

$$2\Psi(\mathbf{a}) = 2\sum_{k=1}^{K}\gamma_k\,\Psi_k(\mathbf{a}) = \sum_{k=1}^{K}\gamma_k\left(\mathbf{f}_k(\mathbf{a}) - \mathbf{f}_k^*\right)^{\mathrm{T}}(\mathbf{W}_k)^{-1}\left(\mathbf{f}_k(\mathbf{a}) - \mathbf{f}_k^*\right) = \min, \tag{30a}$$

where

$$\mathbf{W}_k = \frac{1}{\varepsilon_k^2}\mathbf{C}_k \qquad \text{and} \qquad \gamma_k = \frac{\varepsilon_1^2}{\varepsilon_k^2}. \tag{30b}$$

Here ε_k^2 is the first diagonal element of $\mathbf{C}_k$, i.e. $\varepsilon_k^2 = \{\mathbf{C}_k\}_{11}$. Although, Eqs. (29) and (30) are equivalent, sometimes Eq. (30) is more convenient because in Eq. (30) the parameters γ_k are weighting the contribution of each source relative to the contribution of first data source (obviously, $\gamma_1 = 1$). Similarly, using weight matrices instead of covariance matrices allows for the analysis of the relative contribution of different measurements within each *k*-th data set.

The Minimum of the multi-term quadratic form $\Psi(\mathbf{a})$ can be found by solving the system of multi-term normal equations, i.e. Eq. (11) can be transformed as

$$\sum_{k=1}^{K}\gamma_k\,(\mathbf{K}_k)^{\mathrm{T}}(\mathbf{W}_k)^{-1}(\mathbf{K}_k)\;\mathbf{a} = \sum_{k=1}^{K}\gamma_k\,(\mathbf{K}_k)^{\mathrm{T}}(\mathbf{W}_k)^{-1}\mathbf{f}_k^*. \tag{31}$$

Correspondingly, using matrix inversion the multi-term equivalent of Eq.(12) is

$$\hat{\mathbf{a}} = \left(\sum_{k=1}^{K}\gamma_k\,(\mathbf{K}_k)^{\mathrm{T}}(\mathbf{W}_k)^{-1}(\mathbf{K}_k)\right)^{-1}\left(\sum_{k=1}^{K}\gamma_k\,(\mathbf{K}_k)^{\mathrm{T}}(\mathbf{W}_k)^{-1}\mathbf{f}_k^*\right). \tag{32}$$

The generalization of the basic LSM by the multi-term Eqs. (31-32) is useful for utilizing several observational data sets in a single flexible retrieval. In addition, these equations can be a basis for unifying various techniques for constrained inversion techniques. For example, constraining the solution by *a priori* estimates can be considered as a joint inversion of two data sets. For such case, Eq. (25) is

$$\begin{cases}\mathbf{f}_1^* = \mathbf{f}_1^*(\mathbf{a}) + \Delta\mathbf{f}_1^* \\ \mathbf{f}_2^* = \mathbf{f}_2^*(\mathbf{a}) + \Delta\mathbf{f}_2^*\end{cases} \Rightarrow \begin{cases}\mathbf{f}^* = \mathbf{f}^*(\mathbf{a}) + \Delta\mathbf{f}^* \\ \mathbf{a}^* = \mathbf{a} + \Delta\mathbf{a}^*\end{cases}. \tag{33}$$

The matrices $\mathbf{K}_k$ and $\mathbf{W}_k$ required in Eq. (32) are the following:

$$\mathbf{K}_1 = \mathbf{K}, \text{ and } \mathbf{K}_2 = \mathbf{I},$$

$$\mathbf{W}_1 = \mathbf{W} = (1/\varepsilon_{\mathbf{f}*})^2\, \mathbf{C}_{\mathbf{f}*}, \text{ and } \mathbf{W}_2 = \mathbf{W}_{\mathbf{a}*} = (1/\varepsilon_{\mathbf{a}*})^2\, \mathbf{C}_{\mathbf{a}*}, \tag{34a}$$

and the two-term Eq. (32) is

$$\hat{\mathbf{a}} = (\mathbf{K}^{\mathbf{T}}\, \mathbf{W}^{-1}\, \mathbf{K} + \gamma\, \mathbf{W}_{\mathbf{a}*}^{-1})^{-1}\, (\mathbf{K}^{\mathbf{T}}\, \mathbf{W}^{-1}\, \mathbf{f}^* + \gamma\, \mathbf{W}_{\mathbf{a}*}^{-1}\, \mathbf{a}^*). \tag{34b}$$

Equation (34) is an obvious analog of Eq. (23). This is not surprising because both Eq. (23) and Eq. (24) are derived and explained in several previous studies (e.g. [30, 19]) using the approach similar to the above considerations [Eqs. (25-32)]. Also, Eq. (34b) can be trivially reduced to the *Twomey* formula [Eq. (20)] by assuming the same accuracy for all measurements f_j^* (i.e. $\mathbf{W} = \mathbf{I}$) and the same accuracy for all *a priori* estimates a_i^* (i.e. $\mathbf{W}_{\mathbf{a}*} = \mathbf{I}$). It is interesting that the use of weight matrices gives a clear statistical interpretation to the *Lagrange* multiplier as the ratio of variances:

$$\gamma = \varepsilon_{\mathbf{f}*}^2 / \varepsilon_{\mathbf{a}*}^2 . \tag{34c}$$

Such an interpretation of the *Lagrange* multiplier is especially useful for cases when both $\varepsilon_{\mathbf{f}*}$ and $\varepsilon_{\mathbf{a}*}$ have the same unites or they are unitless. For example, if $\varepsilon_{\mathbf{f}*}$ and $\varepsilon_{\mathbf{a}*}$ are variances of relative errors: $\Delta f_j / f_j$ and $\Delta a_i / a_i$. In such a situation, a small value of the *Lagrange* multiplier γ logically is expected since *a priori* knowledge is always less certain than actual measurements.

4.3 Statistical interpretation of smoothing constraints

Constraining the inversion to a smooth solution as given by *Phillips – Tikhonov – Twomey* Eq. (19) has been proven to be very efficient in numerous applications, e.g. [1, 13, 35-40]. In contrast to Eqs. (20, 23-24) where the solution $\hat{\mathbf{a}}$ is constrained to the actual values of *a priori* estimates $\mathbf{a}^*$**,** Eq. (19) constrains only differences (derivatives) between elements of retrieved vector $\hat{\mathbf{a}}$ and does not restrict their values. Therefore, smoothing constraints may be preferable in applications where *a priori* magnitudes of unknowns are uncertain. For example, a smooth behavior with no sharp oscillations can naturally be expected for atmospheric characteristic $y(x)$ such as the size distributions of aerosol concentrations. Correspondingly, filtering of the solutions with strong oscillations of $a_i = y(x_i)$ (i=1, …,N_a) appears to be a logical constraint, while finding appropriate *a priori* estimates $a_i^* = y^*(x_i)$ would be problematic because the magnitudes of $a_i = y(x_i)$ (i.e. aerosol loading) may vary by ~ 100 times. Retrieval of vertical profiles of atmospheric gases and aerosol concentrations can be another example where some smoothness in the behavior of unknowns $a_i = y(x_i)$ can be expected.

Studies [11-13] originated Eq. (19) did not imply any statistical meaning to the smoothness constrains. Later studies suggest some statistical interpretation to smoothing constraints. For example, studies [18, 32] considered the

smoothness matrix Ω as the inverse matrix to the covariance matrix of *a priori* solutions. *Rodgers* [19] related smoothing constraints with the non-diagonal structure of the covariance matrix $\mathbf{C}_{\mathbf{a}^*}$ of the *a priori* estimates. The present analysis (as follows from [9]) explicitly considers smoothness constraints as *a priori* estimates of the derivatives of the retrieved characteristic $y(x_i)$.

The values of m-th derivatives g_m of the function $y(x)$ characterize the degree of its non-linearity and, therefore, can be used as a measure of $y(x)$ smoothness. For example, smooth functions $y(x)$, such as a constant, straight line, parabola, etc. can be identified by m-th derivatives as follows:

$$\begin{aligned} &g_1(x) = dy(x)/dx = 0 \quad \Rightarrow \quad y_1(x) = \mathrm{C}\,; \\ &g_2(x) = d^2y(x)/dx^2 = 0 \;\Rightarrow\; y_2(x) = \mathrm{B}\,x + \mathrm{C}\,; \\ &g_3(x) = d^3y(x)/dx^3 = 0 \;\Rightarrow\; y_3(x) = \mathrm{A}\,x^2 + \mathrm{B}\,x + \mathrm{C} \end{aligned} \quad . \tag{35}$$

These derivatives g_m can be approximated by differences between values of the function $a_i = y(x_i)$ in N_a discrete points x_i as:

$$\begin{aligned} &\frac{dy(x_{i'})}{dx} \approx \frac{\Delta^1 y(x_i)}{\Delta_1 x_i} = \frac{y(x_i+\Delta x_i) - y(x_i)}{\Delta_1 x_i} = \frac{y(x_{i+1}) - y(x_i)}{\Delta_1 x_i}\,; \\ &\frac{d^2y(x_{i''})}{dx^2} \approx \frac{\Delta^2 y(x_i)}{\Delta_2(x_i)} = \frac{\Delta^1 y(x_{i+1})/\Delta_1(x_{i+1}) - \Delta^1 y(x_i)/\Delta_1(x_i)}{(\Delta_1 x_i + \Delta_1 x_{i+1})/2} = \ldots\,; \quad , \\ &\frac{d^3y(x_{i'''})}{dz^3} \approx \frac{\Delta^3 y(x_i)}{\Delta_3(x_i)} = \frac{\Delta^2 y(x_{i+1})/\Delta_2(x_{i+1}) - \Delta^2 y(x_i)/\Delta_2(x_i)}{(\Delta_2(x_i) + \Delta_2(x_{i+1}))/2} = \ldots\,; \end{aligned} \tag{36}$$

where

$$\Delta_1(x_i) = x_{i+1} - x_i; \quad \Delta_2(x_i) = (\Delta_1(x_i) + \Delta_1(x_{i+1}))/2; \quad \Delta_3(x_i) = (\Delta_2(x_i) + \Delta_2(x_{i+1}))/2;$$

$$x_{i'} = x_i + \Delta_1(x_i)/2; \; x_{i''} = x_i + (\Delta_1(x_i) + \Delta_2(x_i))/2; \; x_{i''} = x_i + (\Delta_1(x_i) + \Delta_2(x_i) + \Delta_3(x_i))/2.$$

In retrievals of the function $y(x_i)$ in N_a discrete points x_i, the expectations of limited derivatives of $y(x)$ can be employed explicitly as smoothness constraints. Namely, if the retrieved function is expected to be close to a constant, straight line, parabola, etc., one can use zero m-th derivatives, as follows from Eq. (35), as *a priori* estimates: $\mathbf{g}_\mathrm{m}^* = \mathbf{0}$. Using this knowledge as a second source of information about $a_i = y(x_i)$, the multi-source Eq. (25) can be written:

$$\begin{cases} \mathbf{f}_1^* = \mathbf{f}_1^*(\mathbf{a}) + \Delta\mathbf{f}_1^* \\ \mathbf{f}_2^* = \mathbf{f}_2^*(\mathbf{a}) + \Delta\mathbf{f}_2^* \end{cases} \Rightarrow \begin{cases} \mathbf{f}^* = \mathbf{f}^*(\mathbf{a}) + \Delta\mathbf{f}^* \\ \mathbf{g}_\mathrm{m}^* = \mathbf{g}_\mathrm{m}(\mathbf{a}) + \Delta\mathbf{g}_\mathrm{m}^* \end{cases} \Rightarrow \begin{cases} \mathbf{f}^* = \mathbf{f}^*(\mathbf{a}) + \Delta\mathbf{f}^* \\ \mathbf{0}^* = \mathbf{G}_\mathrm{m}\mathbf{a} + \Delta_\mathbf{g}^* \end{cases}, \tag{37}$$

where $\mathbf{g}_\mathrm{m}$ is a vector of m-th derivatives $(g_\mathrm{m})_i = g_\mathrm{m}(x_i)$ $(i=m+1,\ldots,N_\mathbf{a})$, $\mathbf{G}_\mathrm{m}$ is the matrix of the coefficients required for matrix form $\mathbf{g}_\mathrm{m} = \mathbf{G}_\mathrm{m}\,\mathbf{a}$ of Eq. (32). The

errors $\Delta_{\mathbf{g}}$* reflect the uncertainty in the knowledge of the deviations of $y(x)$ from the assumed constant, straight line, parabola, etc. Correspondingly, assuming that $\Delta_{\mathbf{g}}$* have a normal distribution with the covariance matrix $\mathbf{C}_{\mathbf{g}*}$, one can use multi-term LSM Eq. (32) with the following matrices $\mathbf{K}_k$ and $\mathbf{W}_k$:

$$\mathbf{K}_1 = \mathbf{K}, \text{ and } \mathbf{K}_2 = \mathbf{G}_{\mathrm{m}},$$
$$\mathbf{W}_1 = \mathbf{W} = (1/\varepsilon_{\mathbf{f}*})^2\, \mathbf{C}_{\mathbf{f}*}, \text{ and } \mathbf{W}_2 = \mathbf{W}_{\mathbf{g}*} = (1/\varepsilon_{\mathbf{g}*})^2\, \mathbf{C}_{\mathbf{g}*}, \tag{38a}$$

and the two-term Eq. (32), solving Eq. (37), has the form:

$$\hat{\mathbf{a}} = \left(\mathbf{K}^{\mathrm{T}}\mathbf{W}^{-1}\mathbf{K} + \gamma\, \mathbf{G}_{\mathrm{m}}^{\mathrm{T}}\mathbf{W}_{\mathbf{g}*}^{-1}\mathbf{G}_{\mathrm{m}}\right)^{-1}\mathbf{K}^{\mathrm{T}}\mathbf{W}^{-1}\mathbf{f}*, \tag{38b}$$

where the multiplier γ is defined as

$$\gamma = \varepsilon_{\mathbf{f}*}^2 \big/ \varepsilon_{\mathbf{g}*}^2\,, \tag{38c}$$

where $\varepsilon_{\mathbf{g}*}^2$ is the first diagonal element of $\mathbf{C}_{\mathbf{g}*}$, i.e. $\varepsilon_{\mathbf{g}*}^2 = \{\mathbf{C}_{\mathbf{g}*}\}_{11}$.

Thus, Eq. (38b) minimizes the quadratic form [Eq. (30a)] with two terms (k=1,2), where the second term $\Psi_2(\mathbf{a})$ represents *a priori* constraints on the m-th derivatives. The inclusion of $\Psi_2(\mathbf{a})$ in the minimization can be considered as applying limitations on the quadratic norm of m-th derivatives of $y(x)$ that are commonly used as a measure of smoothness (e. g. see [32]). Indeed, if one assumes the diagonal covariance matrix $\mathbf{C}_{\mathbf{g}*}$ with diagonal elements

$$\left\{\mathbf{C}_{\mathbf{g}*}\right\}_{ii} \sim 1/\Delta_{\mathrm{m}}(x_i) \;\Rightarrow\; \left\{\mathbf{W}_{\mathbf{g}*}\right\}_{ii} = \left\{\mathbf{C}_{\mathbf{g}*}\right\}_{ii} \Big/ \left\{\mathbf{C}_{\mathbf{g}*}\right\}_{11} = \left(\Delta_{\mathrm{m}}(x_1)\right) \Big/ \left(\Delta_{\mathrm{m}}(x_i)\right), \tag{38d}$$

then the quadratic term $\Psi_2(\mathbf{a})$ can be considered as an estimate of the norm of the m-th derivatives obtained using the values of $y(x)$ at N_{a} discrete points x_i:

$$b_{\mathrm{m}} = \int \left(\frac{d^{\mathrm{m}} y(x)}{d^{\mathrm{m}} x}\right)^2 dx \approx \sum_{i=m+1}^{N_{\mathrm{a}}} \left(\frac{\Delta^{\mathrm{m}} y(x_i)}{\Delta_m(x_i)}\right)^2 \Delta_m(x_i) = \mathbf{a}^{\mathrm{T}}\mathbf{G}_m^{\mathrm{T}}\mathbf{C}_{\mathbf{g}*}^{-1}\mathbf{G}_m\mathbf{a} \sim \Psi_2(\mathbf{a}). \tag{39}$$

By the means of this equation, one can relate the variance $\varepsilon_{\mathbf{g}*}^2$ and the multiplier γ to the expected value of the norm b_{m}. Indeed, the estimates of the derivatives ($\mathbf{g}_{\mathrm{m}}$*= $\mathbf{0}$* = $\mathbf{0}$ + $\Delta_{\mathbf{g}}$*) employed in Eq. (38) assume the following mean value of the norm b_{m}:

$$\langle b_{\mathrm{m}} \rangle \approx \sum_{i=1}^{N_{\mathrm{a}}} \left\langle \left((\mathbf{g}_m)_i\right)^2 \right\rangle \Delta_m(x_i) = \sum_{i=m+1}^{N_{\mathrm{a}}} \left\{\mathbf{C}_{\mathbf{g}*}\right\}_{ii} \Delta_m(x_i) = (N_a - m)\, \Delta_m(x_1)\, \varepsilon_{\mathbf{g}*}^2. \tag{40}$$

Then, the variance $\varepsilon_{\mathbf{g}*}^2$ and the multiplier γ can be determined via $\langle b_{\mathrm{m}} \rangle$ as

$$\varepsilon^2_{\mathbf{g}*} = \frac{1}{\langle b_{\mathrm{m}} \rangle} \left(N_a - m\right) \Delta_m(x_1) \;\Rightarrow\; \gamma = \frac{\varepsilon^2_{\mathbf{f}*}}{\langle b_{\mathrm{m}} \rangle} \left(N_a - m\right) \Delta_m(x_1). \tag{41}$$

Equations (38) explicitly use the discrete approximation of derivatives via ratios of the differences of the function $\Delta^{\mathrm{m}}y(x_i)$ and differences of the arguments $\Delta_{\mathrm{m}}(x_i)$, while Eq. (19) uses only differences of the function $\Delta^{\mathrm{m}}y(x_i)$. Obviously, Eqs. (38) and Eq. (19) are nearly analogous when the differences of arguments $\Delta_{\mathrm{m}}(x_i)$ can be trivially accounted, i.e. when $y(x)$ is retrieved in N_{a} equidistant points $x_{i+1} = x_i + \Delta x$ ($i=1,\ldots,N_{\mathbf{a}}$-1). For this situation, $\Delta_m(x_i) = \Delta x$ and the derivatives and differences of the function $y(x)$ differ by a constant only:

$$\frac{d^{\mathrm{m}}y(x)}{d^{\mathrm{m}}x} \approx \frac{\Delta^{\mathrm{m}}y(x_i)}{(\Delta x)^m} \;\Rightarrow\; \mathbf{G}_{\mathrm{m}} = (\Delta x)^{-\mathrm{m}}\, \mathbf{S}_{\mathrm{m}}. \tag{42}$$

Correspondingly, Eq. (37) can be written using the differences $\Delta_{\mathrm{m}}y(x_i)$:

$$\begin{cases} \mathbf{f}^* = \mathbf{f}^*(\mathbf{a}) + \Delta\mathbf{f}^* \\ (\Delta^{\mathrm{m}}\mathbf{a})^* = \mathbf{S}_{\mathrm{m}}\mathbf{a} + \Delta(\Delta^{\mathrm{m}}\mathbf{a})^* \end{cases} \Rightarrow \begin{cases} \mathbf{f}^* = \mathbf{f}^*(\mathbf{a}) + \Delta\mathbf{f}^* \\ \mathbf{0}^* = \mathbf{S}_{\mathrm{m}}\mathbf{a} + \Delta^*_{\mathrm{m}} \end{cases}, \tag{43}$$

where the vectors $\mathbf{0}^*$ and Δ_{m}^* contain estimates of differences and errors of these estimates, respectively. Also, for equidistant points $x_{i+1} = x_i + \Delta$, the covariance matrix $\mathbf{C}_{\Delta*}$ of the differences differs from $\mathbf{C}_{\mathbf{g}*}$ by a constant only:

$$\left(\Delta_{\mathrm{m}}(x_i) = \Delta x\right) \;\Rightarrow\; (\Delta x_i)^{-2m}\mathbf{C}_{\Delta*} = \mathbf{C}_{\mathbf{g}*} \;\text{ and }\; \mathbf{W}_{\Delta*} = \mathbf{W}_{\mathbf{g}*} = \mathbf{1}. \tag{44a}$$

Correspondingly, Eq. (40) relates $\mathbf{C}_{\Delta*}$ with the norm b_{m} as follows:

$$\langle b_{\mathrm{m}} \rangle \approx \sum_{i=m+1}^{N_{\mathrm{a}}} \left\{\mathbf{C}_{\mathbf{g}*}\right\}_{ii} \Delta_m(x_i) = \left(\frac{1}{\Delta x}\right)^{2m+1} \sum_{i=m+1}^{N_{\mathrm{a}}} \left\{\mathbf{C}_{\Delta*}\right\}_{ii} = \left(\frac{1}{\Delta x}\right)^{2m+1} \left(N_a - m\right)\varepsilon^2_{\Delta*}. \tag{44b}$$

Finally, Eqs. (38) can be reduced to the equivalent of Eq. (19) as:

$$\mathbf{K}_1 = \mathbf{K}, \text{ and } \mathbf{K}_2 = \mathbf{S}_{\mathrm{m}},$$

$$\mathbf{W}_1 = \mathbf{W} = (1/\varepsilon_{\mathbf{f}*})^2\, \mathbf{C}_{\mathbf{f}*}, \text{ and } \mathbf{W}_2 = \mathbf{W}_{\Delta*} = (1/\varepsilon_{\Delta*})^2\, \mathbf{C}_{\Delta*} = \mathbf{1}, \tag{45a}$$

$$\hat{\mathbf{a}} = \left(\mathbf{K}^{\mathrm{T}}\mathbf{W}^{-1}\mathbf{K} + \gamma\, \mathbf{S}_{\mathrm{m}}^{\mathrm{T}}\mathbf{S}_{\mathrm{m}}\right)^{-1} \mathbf{K}^{\mathrm{T}}\mathbf{W}^{-1}\mathbf{f}^*, \tag{45b}$$

$$\gamma = \frac{\varepsilon^2_{\mathbf{f}*}}{\varepsilon^2_{\mathbf{g}*}} = \frac{\varepsilon^2_{\mathbf{f}*}}{\langle b_{\mathrm{m}} \rangle} \left(N_a - m\right)(\Delta x)^{-2m+1}. \tag{45c}$$

where $\varepsilon^2_{\Delta*}$ and $\varepsilon^2_{\mathbf{g}*}$ are the first diagonal elements of the covariance matrices: $\varepsilon^2_{\Delta*} = \{\mathbf{C}_{\Delta*}\}_{11}$ and $\varepsilon^2_{\mathbf{g}*} = \{\mathbf{C}_{\mathbf{g}*}\}_{11}$.

Thus the multi-term LSM is a useful approach for deriving the *Phillips – Tikhonov – Twomey* constrained inversion [Eq. (19)]. Also, it is shown above that constraining the solution by adding a smoothness term in Eq. (19) can be considered as explicit use of knowledge about the *m*-th derivatives of retrieved functions $y(x)$. In other words, the inversion of measurements $\mathbf{f}^*$ is replaced by the joint inversion of the measurements $\mathbf{f}^*$ and "measured" derivatives $\mathbf{g}_m^*$. In the scope of such considerations, the *Lagrange* multiplier γ has a clear quantitative interpretation [Eqs. (41) and (45)]. In addition, Eqs. (38-41) can be used in situations where utilizing the original Eq. (19) is not transparent. For example, Eq. (38) generalizes the use of smoothness constraints on situations where the retrieved function $y(x)$ is defined at points with non-equidistant ordinates x_i. Also Eq. (38) allows differentiating the smoothing strength for different x_i by weight matrix $\mathbf{W}_{\mathbf{g}*}$ non-equal to unity matrix. Although, using $\mathbf{W}_{\mathbf{g}*}$ other than defined by Eq. (38d) requires modifications in definition of $\varepsilon^2_{\mathbf{g}*}$ and γ. Namely, in addition to $\langle b_m \rangle$ the information about smoothness differentiation should be available. Illustrations of applying Eq. (38) can be found in the manuscript [9], where the spectral dependence $n(\lambda_i)$ of the aerosol refractive index is retrieved in non-equidistant λ_i fixed by the measurement specifications.

4.4 Combining multiple *a priori* constraints in the inversion

The consideration of *a priori* constraints as an equal component in the multi-term inversion Eqs. (30-32) is a useful tool for applying multiple constraints in a retrieval algorithm. For example, a simultaneously constraining solution by both *a priori* estimates and smoothness assumptions can be considered as inversion of the data from three independent sources and Eq. (25) is

$$\begin{cases} \mathbf{f}_1^* = \mathbf{f}_1^*(\mathbf{a}) + \Delta\mathbf{f}_1^* \\ \mathbf{f}_2^* = \mathbf{f}_2^*(\mathbf{a}) + \Delta\mathbf{f}_2^* \\ \mathbf{f}_3^* = \mathbf{f}_3^*(\mathbf{a}) + \Delta\mathbf{f}_3^* \end{cases} \Rightarrow \begin{cases} \mathbf{f}^* = \mathbf{f}^*(\mathbf{a}) + \Delta\mathbf{f}^* \\ \mathbf{0}^* = \mathbf{S}_m + \Delta(\Delta^m\mathbf{a})^* \\ \mathbf{a}^* = \mathbf{a} + \Delta\mathbf{a}^* \end{cases}. \tag{46}$$

The matrices $\mathbf{K}_k$ and $\mathbf{W}_k$ required in Eq. (32) are the following:

$$\mathbf{K}_1 = \mathbf{K}; \quad \mathbf{K}_2 = \mathbf{S}_m; \quad \text{and} \quad \mathbf{K}_3 = \mathbf{I},$$

$$\mathbf{W}_1 = \mathbf{W} = (1/\varepsilon_{\mathbf{f}*})^2\,\mathbf{C}_{\mathbf{f}*};\ \mathbf{W}_2 = \mathbf{W}_{\Delta*} = \mathbf{1};\ \text{and}\ \mathbf{W}_3 = \mathbf{W}_{\mathbf{a}*} = (1/\varepsilon_{a*})^2\,\mathbf{C}_{\mathbf{a}*}, \tag{47a}$$

and three-term Eq. (32) is:

$$\hat{\mathbf{a}} = (\mathbf{K}^T\mathbf{W}^{-1}\mathbf{K} + \gamma_2\,\mathbf{S}_m^T\,\mathbf{S}_m + \gamma_3\,\mathbf{W}_{\mathbf{a}*}^{-1})^{-1}\,(\mathbf{K}^T\mathbf{W}^{-1}\,\mathbf{f}^* + \gamma_3\,\mathbf{W}_{\mathbf{a}*}^{-1}\,\mathbf{a}^*), \tag{47b}$$

$$\gamma_2 = \frac{\varepsilon^2_{\mathbf{f}*}}{\varepsilon^2_{\mathbf{g}*}} = \frac{\varepsilon^2_{\mathbf{f}*}}{\langle b_{\mathrm{m}} \rangle} \left(N_a - m\right) \left(\Delta x\right)^{-2m+1} \quad \text{and} \quad \gamma_3 = \frac{\varepsilon^2_{\mathbf{f}*}}{\varepsilon^2_{\mathbf{g}*}}. \tag{47c}$$

This equation minimizes three quadratic forms simultaneously:

$$2\Psi(\hat{\mathbf{a}}) = 2\sum_{k=1}^{3} \gamma_k \Psi_k(\hat{\mathbf{a}}) = \left(\mathbf{f}(\hat{\mathbf{a}}) - \mathbf{f}^*\right)^{\mathrm{T}} \mathbf{W}^{-1} \left(\mathbf{f}(\hat{\mathbf{a}}) - \mathbf{f}^*\right) + \gamma_2 \hat{\mathbf{a}}^{\mathrm{T}} \Omega_{\mathrm{m}} \hat{\mathbf{a}} + \gamma_3 \hat{\mathbf{a}}^{\mathrm{T}} \mathbf{W}_{\mathbf{a}*}^{-1} \hat{\mathbf{a}} = \min \cdot \tag{48}$$

Thus, applying multiple *a priori* constraints is straightforward using multi-term LSM formulations, while multiple *a priori* constraints usually are not considered in the scope of basic formulas [Eqs. (19-20) and (23-24)].

4.5 Error evaluation

Equations (13-15) estimating errors of LSM solutions can be generalized in the case of multi-term solutions by Eqs. (31-32) as

$$\mathbf{C}_{\hat{\mathbf{a}}} = \mathbf{C}_{\Delta\hat{\mathbf{a}}(\mathrm{ran})} + (\hat{\mathbf{a}}_{\mathrm{bias}})\,(\hat{\mathbf{a}}_{\mathrm{bias}})^{\mathrm{T}}, \tag{49a}$$

$$\hat{\mathbf{a}}_{bias} = \left(\sum_{k=1}^{K} \gamma_k \left(\mathbf{K}_k\right)^{\mathrm{T}} \left(\mathbf{W}_k\right)^{-1} \left(\mathbf{K}_k\right)\right)^{-1} \left(\sum_{k=1}^{K} \gamma_k \left(\mathbf{K}_k\right)^{\mathrm{T}} \left(\mathbf{W}_k\right)^{-1} \mathbf{b}_k^*\right), \tag{49b}$$

$$\mathbf{C}_{\Delta\hat{\mathbf{a}}_{(ran)}} = \left\langle \Delta\hat{\mathbf{a}}_{(ran)} \left(\Delta\hat{\mathbf{a}}_{(ran)}\right)^{\mathrm{T}} \right\rangle = \left(\sum_{k=1}^{K} \gamma_k \left(\mathbf{K}_k\right)^{\mathrm{T}} \left(\mathbf{W}_k\right)^{-1} \left(\mathbf{K}_k\right)\right)^{-1} \varepsilon_1^2, \tag{49c}$$

where $\mathbf{b}_k$ denotes the *bias* vector in the k-th data set $\mathbf{f}_k$.

For example, for the three term solution by Eqs. (47), the retrieval bias $\hat{\mathbf{a}}_{\mathrm{bias}}$ and the covariance matrix of random errors $\mathbf{C}_{\Delta\hat{\mathbf{a}}(\mathrm{ran})}$ can be written as

$$\hat{\mathbf{a}}_{\mathrm{bias}} = (\mathbf{K}^{\mathbf{T}} \mathbf{W}^{-1} \mathbf{K} + \gamma_2 \boldsymbol{\Omega}_{\mathrm{m}} + \gamma_3 \mathbf{W}_{\mathbf{a}*}^{-1})^{-1} (\mathbf{K}^{\mathbf{T}} \mathbf{W}^{-1} \mathbf{b}_{\mathbf{f}*} + \gamma_2 \boldsymbol{\Omega}_{\mathrm{m}} \mathbf{b}_{\Delta*} + \gamma_3 \mathbf{W}_{\mathbf{a}*}^{-1} \mathbf{b}_{\mathbf{a}*}) \tag{50a}$$

$$\mathbf{C}_{\Delta\hat{\mathbf{a}}(\mathrm{ran})} = (\mathbf{K}^{\mathbf{T}} \mathbf{W}^{-1} \mathbf{K} + \gamma_2 \boldsymbol{\Omega}_{\mathrm{m}} + \gamma_3 \mathbf{W}_{\mathbf{a}*}^{-1})^{-1} \varepsilon_{\mathbf{f}*}^2, \tag{50b}$$

$$\mathbf{b}_{\mathbf{f}} = \langle \mathbf{f}^* - \mathbf{f}(\mathbf{a}^{\mathbf{real}}) \rangle = \langle \Delta\mathbf{f}^* \rangle;\ \{\mathbf{b}_{\Delta*}\}_{\mathrm{ii}} = (\Delta x)^{2\mathrm{m}} \left(d^m y(x_i) / d^m x\right)_{real};\ \mathbf{b}_{\mathbf{a}*} = \mathbf{a}^{\mathrm{real}} - \mathbf{a}^*, \tag{50c}$$

where $\mathbf{b}_{\mathbf{f}}$ is a *bias* in the measurements $\mathbf{f}^*$ or in the forward model: $\mathbf{f}^{\mathbf{real}} - \mathbf{f}(\mathbf{a}^{\mathbf{real}})$; $\boldsymbol{\Omega}_{\mathrm{m}} = \mathbf{S}_{\mathrm{m}}^{\mathrm{T}} \mathbf{S}_{\mathrm{m}}$ and vector $\mathbf{b}_{\Delta}$ (with elements $\{\mathbf{b}_{\Delta*}\}_{\mathrm{ii}}$) denotes a *bias* introduced by assuming zeros $\mathbf{0}^*$ as estimates of m-th differences; $\mathbf{b}_{\mathbf{a}*}$ is a *bias* in *a priori* estimates $\mathbf{a}^*$.

Equations (50) are helpful for analyzing the effects of constraints on the solution. It follows from Eq. (50b) that by strengthening *a priori* constraints (by γ_2 and γ_3), one formally can suppress the random errors of the retrieval to any desirable level. However, Eq. (50a) shows that, if *a priori* biases are non-zero, increasing γ_2 and γ_3 leads to increasing systematic errors. Therefore, *a*

priori constraints are useful only in the case when the increase of the systematic component $(\hat{\mathbf{a}}_{\text{bias}})(\hat{\mathbf{a}}_{\text{bias}})^{\text{T}}$ does not exceed the decrease of the random component of the retrieval errors in Eq. (49). Unfortunately, in practice the *a priori* biases are uncertain and, therefore, the selection of optimum *a priori* constraints is a very challenging issue in inversion developments. Generally, *a priori* estimates $\mathbf{a}^*$ always have non-zero bias $\mathbf{b}_{\mathbf{a}*}= \mathbf{a}^{\text{real}}-\mathbf{a}^*$. In contrast, the smoothness constraints are likely unbiased (i.e. $\mathbf{b}_{\Delta*}\rightarrow \mathbf{0}$), because for smooth functions, *m*-derivatives are close to zero. This is why smoothness constraints are preferable for the retrieval of smooth functions.

It should be noted that the multi-term estimates [Eqs. (32), (34), (47)] retain the optimality of LSM estimates, i.e. they have smallest errors as determined by the *Cramer-Rao* inequality, Eq. (16). However, the *Cramer-Rao* inequality is valid only if all assumptions about noise in both the measurements and the *a priori* terms are correct. Therefore, the validation of the assumptions is important, while problematic in reality. A useful consistency check can be performed using the achieved value of the minimized quadratic form $\Psi(\hat{\mathbf{a}})$ [Eq. (30)]. For example, in case of zero biases, the minimum value of the three-term $\Psi(\hat{\mathbf{a}})$ [Eq (48)] has a χ^2 distribution with mean

$$\begin{aligned}\left\langle\left(2\Psi(\hat{\mathbf{a}})\right)_{min}\right\rangle &= \left\langle 2\sum_{\text{i}=1,\ldots,3.}\Psi_i(\hat{\mathbf{a}})\right\rangle = \left\langle\left(\hat{\mathbf{f}}-\mathbf{f}^*\right)^{\text{T}}\mathbf{W}^{-1}\left(\hat{\mathbf{f}}-\mathbf{f}^*\right)+\gamma_2\hat{\mathbf{a}}^{\text{T}}\Omega_{\text{m}}\hat{\mathbf{a}}+\gamma_3\hat{\mathbf{a}}^{\text{T}}\mathbf{W}_{\mathbf{a}*}^{-1}\hat{\mathbf{a}}\right\rangle = \\ &= \left(\sum_{\text{i}=1,\ldots,3}(N_{\mathbf{f}_i})-N_{\mathbf{a}}\right)\varepsilon_1^2 = \left(N_{\mathbf{f}*}+N_{\Delta*}+N_{\mathbf{a}*}-N_{\mathbf{a}}\right)\varepsilon_1^2 = \left(N_{\mathbf{f}*}+N_{\mathbf{a}}-m\right)\varepsilon_1^2\end{aligned}, \quad (51)$$

where $N_{\Delta*}=N_{\mathbf{a}}-m$ and $N_{\mathbf{a}*}=N_{\mathbf{a}}$. Using this equation one can estimate the value ε_1^2 from a minimum value of $\Psi(\hat{\mathbf{a}})$ often called a *residual*:

$$\hat{\varepsilon}_1^2 \approx \frac{\left(2\Psi(\hat{\mathbf{a}})\right)_{min}}{\sum_{\text{i}=1,\ldots,3}(N_{\mathbf{f}_i})-N_{\mathbf{a}}} = \frac{\left(2\Psi(\hat{\mathbf{a}})\right)_{min}}{N_{\mathbf{f}*}+N_{\mathbf{a}}-m}. \quad (52)$$

If all assumptions are correct, the estimation by Eq. (52) should be close to the assumed ε_1^2. A significant increase of the estimated $\hat{\varepsilon}_1^2$ over expected ε_1^2 can be considered as an indication of unaccounted biases and/or inadequate assumptions about random errors in measurements or *a priori* data sets. The consistency checks relying on estimates $\hat{\varepsilon}_1^2$ from the residual commonly is used in remote sensing and other applications. For example, the effects of unaccounted biases in both measurements and forward modeling on retrievals of aerosol properties from ground-based observations using the residuals in observation fits have been analyzed [8].

4.6 *Lagrange multiplier* selection

Sections 4.3-4.4 provide quantitative definitions of *Lagrange* multipliers. However, in reality the detailed information required for an explicit definition of *a priori* constraints may not be available. In such situations the following recipes and discussions may be useful.

For constraining a retrieval by *a priori* estimates $\mathbf{a}^*$ as in Eqs. (33-34), one can use information about typical magnitudes and variabilities of the parameters $\mathbf{a}$. For example, in atmospheric remote sensing applications, climatological data sets are often used as $\mathbf{a}^*$ [9]. If actual observations are not available one can imply *a priori* estimates $\mathbf{a}^*$ from known ranges of a_i variability:

$$a^* = \langle a \rangle = (a_{\max} - a_{\min})/2 \text{ and } \varepsilon_{a^*} = (a_{\max} - a_{\min})/4. \tag{53}$$

This equation assumes the interval $[a_{\max}; a_{\min}]$ as 95% confidence interval $[\langle a \rangle + 2\,\varepsilon\,;\ \langle a \rangle - 2\,\varepsilon]$. These estimates a^* are biased to the middle of the interval $[a_{\max}; a_{\min}]$ (or to the climatological values). However, this bias is usually suppressed by a small γ. Indeed, the standard deviations ε_a determined from Eq. (53) (or from climatology) are usually much larger than the measurement errors. Correspondingly, the *Lagrange* multipliers defined via the ratio of variances defined in Eq. (34c) are likely to have small values.

The strength of smoothness constraints in Eqs. (43-45) is linked with known values of the derivatives of the retrieved $y(x)$. If an explicit analysis of derivatives $\partial^m y/\partial x^m$ is not feasible, the strength of smoothing can be implied from known least smooth of all *a priori* known $y(x)$. For example, Eq. (45) can be replaced by the inequality [9]:

$$\gamma \le \frac{\varepsilon_{\mathbf{f}*}^2}{(b_m)_{\max}} \left(N_a - m\right) \left(\Delta x\right)^{-2m+1}, \tag{54}$$

where $(b_m)_{\max}$ is the norm of the m-th derivatives of "most unsmooth" function $y(x)$. Indeed, the constrained inversion Eq. (45) with γ given by Eq. (54) limits the retrievals to the functions $y(x)$ with the norm of the m-th derivatives being comparable or smaller than $(b_m)_{\max}$, i.e. the retrieval of $y(x)$ much less smooth than the "most unsmooth" $y(x)$ is not allowed.

Thus, even if actual *a priori* data are not available, the values of *Lagrange* multipliers can be determined using Eqs. (53-54) before implementing actual inversion, or in another words *prior* to performing the inversion. This is a difference and a possible advantage of the approach [7-9] described here with respect to a majority of techniques established for determining *Lagrange* multipliers. Conventionally (see discussions [19,36,41]), the *Lagrange* multiplier is chosen from analysis of the sensitivity of the minimized quadratic form $\Psi(\mathbf{a})$ [such as given by Eq. (22)] to the weighting balance between contributions of measurements and *a priori* terms. The main idea, employed with some technical differences in many developments [1, 2, 36, 42] is that γ should be both large enough ($\gamma > 0$) for enforcing (via constraints) a stable unique solution and small enough for allowing the algorithm to achieve a reasonably small value of minimized quadratic form $\Psi(\mathbf{a})$. Usually a reasonably small value means a value that can be explained by expected presence of normal noise (with no biases) in the measurements. As discussed above [Eqs. (17, 51)], for such noise the residual is χ^2 distributed with m-n

degrees of freedom and an expected value of $<2\Psi(\mathbf{a})> = (N_{mes}-N_{par})\varepsilon^2_{mes}$ for $\Psi(\mathbf{a})$ defined via weight matrices.

The same idea is utilized in the *L-curve method* [42] where a burden between measurements and the *a priori* terms is visualized by plotting the *a priori* norm, i.e. the *a priori* term in the total $\Psi(\mathbf{a})$, versus the measurements norm, i.e. the measurements term in the total $\Psi(\mathbf{a})$, with γ as a function parameter. This plot has an L-shaped corner showing a point with a specific value of γ of optimum balance between measurement minimization and *a priori* terms.

In spite of the clear rationale and wide use of this optimum balance criterion for determining the *Lagrange* multiplier, there are shortcomings in employing this principle. For example, the technical implementation of this principle is very challenging if more than one *a priori* constraint is needed in the same inversion algorithm; i.e., if the determination of more than one *Lagrange* multiplier is required. Also, the implementation of a conventional determination of the *Lagrange* parameter is rather unclear in inversions of non-linear equations $\mathbf{f}(\mathbf{a})$. Non-linear inversions (see Section 5) use the first derivatives that are functions of the solutions. Therefore the optimum balance between measurements and *a priori* constraints is also a function of the solution. Correspondingly, finding optimum constraints for non-linear inversions requires extensive effort considering the entire space of possible solutions. In contrast, determining optimum constraints, e.g. by Eqs. (53-54), *prior* to the inversion relies only on the knowledge of the variances of errors in measured and *a priori* data. Such definition of a *Lagrange* multiplier is independent of the forward model and can be employed equally in both linear and non-linear algorithms. Also, knowledge about error statistics usually is established independently for each data set. Therefore, once the *Lagrange* multiplier is determined for a single type of *a priori* constraint it can be used with no changes in the inversions employing other types of *a priori* constraints.

Another and more fundamental issue is that the principle of optimum balance is based on an understanding of the limited sensitivity of observations with respect to the retrieved characteristic, but not on considerations of the available *a priori* information. For example, if observations are not sensitive to sharp oscillations of $y(x)$ then a unique inversion of such measurements is possible only if the search for solutions is restricted to smooth functions $y(x)$. Applying smoothness constraints enforces such restrictions and therefore enforces a unique and stable solution. However, the fact of achieving uniqueness and stability of the solution does not guarantee the reality of the solution. In principle, any unsmooth $y(x)$ resulting in the same observations can be a real solution. In other words, the development of a successful retrieval scheme should include two kinds of efforts: (i) identifying type and strength of constraints required for assuring the uniqueness of the solution, (ii) clarifying the realism of *a priori* information assumed by applying identified constraints. In this regard the approach discussed here gives useful insight for

relating employed *a priori* constraints to actual properties of retrieved characteristics; for example, Eqs. (53-54) allow researcher to relate the values of *Lagrange* parameters with variability ranges of magnitudes and derivatives of retrieved $y(x)$. In actual applications (e.g. see [9]), the strength of constraints with *Lagrange* multipliers determined by Eqs. (53-54) acquired from knowledge available *prior* to the inversion may be not sufficient for providing satisfactory solutions. In such situations, the estimated γ can be corrected by sensitivity studies similar to conventional approaches. Obviously, using these corrected (increased) constraints, researchers would face the above raised issue of constraint realism, since the increased constraints would exceed the actual available *a priori* knowledge. In these regards, Eqs. (53-54) can be useful for a quantitative evaluation of possible biases caused by increased *a priori* constraints.

4.7 Limitations of linear constrains.

The difficulty of enforcing non-negative solutions is an essential limitation of linear inversion methods. Indeed, the constrained *Phillips-Tikhonov-Twomey* type linear inversions defined by Eqs. (19-20) do not have a mathematical structure that allows filtering negative solutions even if the retrieved characteristic is physically positively defined. For example, remote sensing is known to suffer from the appearance of unrealistic negative values for retrieved atmospheric aerosol or gas concentrations that are positive by nature. Known techniques of securing non-negative solutions by Eqs. (19-20) force positive retrievals through enhancement of *a priori* smoothness constraints. For example, several studies [36, 42] suggest repeating linear inversion by changing strength of *a priori* constraints until the final solution both satisfies the positivity constraints and provides an admissibly accurate fit of the measurements. In such manner, *King* [36] iteratively adjusted the value of *Lagrange* parameter in Eq. (19). Similarly, *Turchin et al.* [42] iteratively corrected *a priori* terms in the statistical equivalent of Eq. (19), where $\gamma\mathbf{\Omega} = \mathbf{C}_a^{-1}$ and $\mathbf{C}_a$ is considered an *a priori* "correlation" matrix. Such iterative adjustments of γ or *a priori* matrix $\mathbf{C}_a$ require more computations than basic constrained inversions by Eqs. (19-20). However, the major concern of implying non-negativity constraints relates to difficulties of conforming these techniques to general methodological basis of constrained linear inversions. Indeed, as was mentioned in Section 4.1, minimization of quadratic forms in the constrained inversions formally is equivalent to assuming errors normality. Correspondingly, Eqs. (19-20) are harmonized with statistical LSM optimization. Study [42] attempted to integrate the non-negativity constraints within a normal-noise framework by introducing a "cutting" normal curve; i.e., forcing zero probability for negative values and retaining a normal distribution for positive values. Such artificial cutting can hardly be accepted because it contradicts the proven symmetry of a Gaussian curve that is a fundamental property of normal noise distribution.

In contrast to Eqs. (19-20), certain types of non-linear iterations invert linear systems and naturally provide non-negative solutions. For example, in atmospheric optics, the relaxation techniques [14, 17] often are considered as alternatives to linear methods (e.g., see discussions [1, 19]).

The solution of the linear system $\mathbf{f}^* = \mathbf{K}\,\mathbf{a}$ by non-linear iterations

$$a_i^{p+1} = a_i^p \left(f_i^* \big/ f_i^p \right). \tag{55}$$

is developed by *Chahine* [17]. This method is limited to application where measured and retrieved characteristics are positively defined and the number of measurements and unknowns are equal (i.e. $\mathbf{K}$ is square). Also, for convergence, square matrix $\mathbf{K}$ must be diagonally dominant (i.e., $K_{jj} > K_{jj' \neq j}$). One can see that *Chahine's* formula is different from both LSM Eqs. (12) and constrained inversions Eqs. (19-20), (22-23). Namely, instead of addition and subtraction in the linear methods, Eq. (50) is non-linear and includes multiplication and division, thereby eliminating the negative and highly oscillatory solutions occasionally appearing in linear matrix inversions. However, the applicability of Eq. (55) to square and diagonal matrices $\mathbf{K}$ is a serious limitation of *Chahine's* iterations. Many inversion studies adopted *Chahine's* iterations to other situations. The most known generalization of Eq. (55) was proposed by *Twomey* [14] for inverting overdetermined systems $\mathbf{f}^* = \mathbf{K}\,\mathbf{a}$ (where $N_\mathbf{f} > N_\mathbf{a}$ and $\mathbf{K}$ is rectangular):

$$a_i^{p+1} = a_i^p \prod_{j=1}^{N_\mathbf{f}} \left(1 + \left(f_j^* \big/ f_j^p - 1 \right) \tilde{K}_{ji} \right). \tag{56}$$

$\tilde{K}_{ji}$ denotes the elements of matrix $\mathbf{K}$ that are scaled to be less than unity. Eq. (56) provides non-negative solutions while it has much broader applicability than original Eq. (55). Nevertheless, Eq. (56) has been derived without formalized analysis of the noise effects in the initial data. Such empirical character of *Chahine*-like iterations makes it difficult for a researcher to use Eqs. (55-56) as a basis for rigorous inversion optimization.

Thus, the non-negativity of solution is not an established constraint in the theoretical foundation of linear methods. On the other hand, the empirically formulated non-linear methods [Eqs. (55-56)] effectively secure positive and stable solutions. Such a "weakness" of the rigorous linear methods indicates a possible inadequacy in criteria employed for formulating the optimum solutions. In Section 6 we discuss possible revisions in assumptions employed for accounting for random noise in inversions. For example, it will be shown that by using log-normal noise assumptions the non-negativity constraints can be imposed into inversion in a fashion consistent with the presented approach inasmuch as one considers the solution as a noise optimization procedure.

4.8 Alternatives to matrix inversion methods

The main formulas for implementing LSM [Eq. (12)] and linear constraint inversions [Eqs. (19, 20, 23, 24, 32)] are written via a matrix inversion operator. This operator is uncertain for ill-posed problems where the matrices to be inverted [$\mathbf{K}$ in Eq(3), $\mathbf{K}^T\mathbf{K}$ in Eqs. (19-21), $\mathbf{K}^T\mathbf{C}^{-1}\mathbf{K}$ in Eqs. (12, 23), etc.] tend to have zero determinant. This is why a number of studies associate solving ill-posed problems with the use of advanced numerical procedures introducing an inverse operator to degenerated matrices. For example, an inverse operator may be formulated by excluding eigensolutions, linear combinations of unknowns, with zero eigenvalues (e.g. see [43]). Singular value decomposition (SVD) is a particularly popular approach for inverting degenerated matrices. SVD is an operation of linear algebra (see details in [4]), that allow one to decompose a square matrix $\mathbf{M}$ as $\mathbf{M} = \mathbf{V}\mathbf{I}_w\mathbf{A}$, where matrices $\mathbf{V}$ and $\mathbf{A}$ are orthogonal in the sense that $\mathbf{V}^T\mathbf{V} = \mathbf{I}$ and $\mathbf{A}^T\mathbf{A} = \mathbf{I}$. Matrix $\mathbf{I}_w$ is diagonal with the elements on the diagonal equal to w_i. Inversion of matrix $\mathbf{M}$ trivially follows from this decomposition as $\mathbf{M} = \mathbf{A}^T\mathbf{I}_{1/w}\mathbf{V}^T$. In the case of a singular matrix $\mathbf{M}$, the inverse matrix of $\mathbf{M}^{-1}$ is uncertain, because some values w_i are equal or close to zero. Correspondingly, by means of replacing $w_i = 0$ by a moderately small non-zero w_i, singular matrix $\mathbf{M}$ can be replaced by a reasonably similar, non-singular "truncated" matrix $\mathbf{M}'$ that can be trivially inverted. Therefore, some theoretical developments consider applying SVD as an alternative way of constraining linear-system solutions. For example, the theoretical review by *Hansen* [42] considers a truncated SVD method as an essential equivalent of *Pillips-Tikhonov-Twomey* constrained inversion by Eqs. (19-20). The main concern of using the SVD technique instead of direct *a priori* constraints comes from the fact that replacement of matrix $\mathbf{M}$ with truncated matrix $\mathbf{M}'$ is formal and has no relation to the physics of an application. Therefore using SVD should be accompanied by an analysis clarifying how the solution space was restricted by using truncated matrix $\mathbf{M}'$. Such analysis is challenging, since some linear combinations of unknown parameters excluded by truncation may not have clear physical meaning. Also, SVD analysis becomes even more uncertain in non-linear inversions where matrix $\mathbf{M}$ and it's truncated analog $\mathbf{M}'$ changes during iterations. On the other hand, using SVD instead of direct matrix inversion is undoubtedly a useful tool for improving implementation of constrained inversion. For example, study [9] used SVD to solve a multi-term normal system [Eq. (32)]. In this way the initially ill-posed problem is constrained by *a priori* terms in quadratic form given by Eq. (30), and SVD is applied only for solving Eq. (32) that improves the technical performance of inversion algorithm. Indeed, in case of large matrices $\mathbf{M}$, applying standard methods for matrix inversion can be problematic even for non-singular $\mathbf{M}$, while SVD always gives an inverse operator.

Another alternative to matrix inversion is using linear iterations written by Eq. (4). As discussed in Section 2, linear iteration always provides a solution,

even if a linear system [Eq. (2)] has singular matrix **K**. For example, steepest descent method (e.g. see [44-45]) always converges to a solution (more in Section 5). However, in case of singular matrices **K,** iterative methods provide only one solution from many possible. Repeating iterations using different initial guesses may provide information about the entire space of possible solutions. However, building a domain of solutions in such a way is not straightforward because it requires establishing a set of initializations providing complete coverage of solution space. Also, in general, iterative methods are more time consuming than matrix inversion. For example, steepest decent may require an enormous number of iterations for convergence (e.g. see [3]).

The semi-iterative method of conjugated gradients is another popular method of solving linear systems of Eq. (1) via $N_{\mathbf{a}}$ iterations [4]. However, applying conjugated-gradients algorithms to solving a quasi-degenerated linear system ($\det(\mathbf{K}) \rightarrow 0$) results in similar problems as conventional numerical procedures used for inverting matrix (e.g., *Gauss-Jordan* elimination [4]). Although, in the framework of the conjugated-gradients technique, some uncertain components of solution can be suppressed and, as discussed in [41], conjugated gradients may provide a solution close to that of truncated SVD.

There are many other non-matrix inversion methods that are not considered here. Some techniques are based on concepts that are very different standard methods of numerical inversion. For example, using neural networks [46] is a technique that is popular in many applications for observations analysis. The basic idea of neural networking is that prior to interpretation of observations, the researcher establishes unique relations between observations and unknowns via network "training". Network "training" is an analysis that is, in some sense, similar to identifying non-linear regressions between output and input of forward model. Another example is generic-inversions methods that rely entirely on forward simulations [47-48]. Such techniques implement inversion by straightforward computer search for all solutions that admissibly agree with the observations to be inverted. Study [40] proposed a technique that can be considered as a combination of generic inversion with conventional linear inversion. Specifically, the technique implements a large number of inversions using *Phillips-Tikhonov-Twomey* [Eq. (19)] with different values of *Lagrange* multipliers. The average result of such inversions is considered as a suggested solution.

Thus, there are many techniques, only some of which are mentioned here, that can provide an appropriate inverse transformation without implementing direct matrix inversion. Some of these techniques may provide a reasonable solution of ill-posed problems without using explicit *a priori* constraints. However, in comparing those methods against LSM-based constrained inversion [Eqs. (19, 20, 23, 24, 32)], one should realize that each method *a priori* limits a space of possible solutions. For instance, SVD inversion excludes some solutions via matrix truncation; iterative solutions depend on the initial guess; generic inversion considers only solutions included in a

search; neural networks are constrained by training process, etc. Therefore, applying all these techniques to an ill-posed problem may result in different solutions from different techniques. These differences reflect differences in employed *a priori* constraints. Hence, the methods adopting the most reliable *a priori* constraints provide the best solution. In this regards, applying *a priori* constraints in a manner consistent with statistical optimization, as shown in Section 4.2, give rigorous and a clear concept for using *a priori* constraints and combining various data in single inversion. Contrary, absence of direct optimization of statistical properties in some inversion approaches can be considered disadvantageous.

5. Optimization of non-linear inversion

Sections 2-4 discussed inversions procedures only for the case of linear forward model $\mathbf{f}(\mathbf{a})$ in Eq. (1). However in practice, and particularly in remote-sensing applications, the majority of physical dependencies $\mathbf{f}(\mathbf{a})$ are non-linear. The purpose of this Section is to discuss inversion of a non-linear Eq. (1) and to outline the differences and similarities between linear and non-linear cases.

5.1. Basic inversions of non-linear equation system

For a case of non-linear functions $f_j(\mathbf{a})$, Eq. (1) usually is solved numerically by iterations relying on linear approximations. Namely, for points $\hat{\mathbf{a}}$ in the close neighborhood of solution $\mathbf{a}'$, $\mathbf{f}(\mathbf{a})$ can be expanded in *Taylor* series:

$$\mathbf{f}(\mathbf{a}') = \mathbf{f}(\hat{\mathbf{a}}) + \mathbf{K}_{\hat{\mathbf{a}}}\ (\mathbf{a}' - \hat{\mathbf{a}}) + o(\mathbf{a}' - \hat{\mathbf{a}})^2 + \ldots\ , \tag{57}$$

where $\mathbf{K}_{\hat{\mathbf{a}}}$ is the *Jacobi* matrix of the first derivatives $\partial f_j / \partial a_i$ in the near vicinity of $\hat{\mathbf{a}}$; $o(\mathbf{a}' - \hat{\mathbf{a}})^2$ denotes the function that approaches zero as $(\mathbf{a}' - \hat{\mathbf{a}})^2$ when $(\mathbf{a}' - \hat{\mathbf{a}}) \to 0$. Hence, neglecting all terms of second or higher order in Eq. (57), $f_j(\mathbf{a})$ can be considered as linear functions. Such a linear approximation is insufficient to solve Eq. (1) by Eq. (3) directly through inverse transformation, but can be employed successfully for iterative correction of guessed solution:

$$\mathbf{a}^{p+1} = \mathbf{a}^p - \Delta\mathbf{a}^p, \tag{58a}$$

$$\mathbf{K}_p\, \Delta\mathbf{a}^p \approx \mathbf{f}(\mathbf{a}^p) - \mathbf{f}^*, \tag{58b}$$

where $\mathbf{a}^p$ is p-th approximation of solution, $\Delta\mathbf{a}^p$ is a correction of $\mathbf{a}^p$ that is given by the solution of Eq. (58b), where $\mathbf{K}_p$ is the *Jacobi* matrix of $\partial f_j / \partial a_i$ calculated in the vicinity of $\mathbf{a}^p$. In the situation where $\mathbf{K}_p$ is square ($N_{\mathbf{f}} = N_{\mathbf{a}}$), the successive iterations can be implemented by employing inverse matrices:

$$\mathbf{a}^{p+1} = \mathbf{a}^p - \mathbf{K}_p^{-1}\, (\mathbf{f}(\mathbf{a}^p) - \mathbf{f}^*). \tag{59}$$

This is the basic formula of *Newton* iterations to solve non-linear systems.

5.2 Optimization of non-linear solution in presence of random noise

In case of over determined non-linear $f_j(\mathbf{a})$ in Eq. (1), statistical optimization of a solution can be included in the iterations by following MML as described in Section 2. Namely, the solution of Eq. (1) should be performed as minimization of quadratic form $\Psi(\mathbf{a})$ given by Eq. (10) and the resulting non-linear Eq. (11a) can be solved by Newton iterations. Replacing $\mathbf{f}(\mathbf{a}^p)$ by gradient $\nabla\Psi(\mathbf{a}^p)$ and $\mathbf{f}^*$ by $\mathbf{0}$, Eqs. (58) can be written as:

$$\mathbf{a}^{p+1} = \mathbf{a}^p - \Delta\mathbf{a}^p, \tag{60a}$$

$$\mathbf{K}_{\nabla,p}\,\Delta\mathbf{a}^p \approx \nabla\Psi(\mathbf{a}^p), \tag{60b}$$

where $\mathbf{K}_\nabla$ is a matrix of partial derivatives with elements

$$\left\{\mathbf{K}_{\nabla,p}\right\}_{ji} = \left.\frac{\partial\left(\nabla\Psi\right)_j}{\partial a_i}\right|_{\mathbf{a}^p}. \tag{61a}$$

The matrix $\mathbf{K}_{\nabla,p}$ follows from Eq. (11) as:

$$\mathbf{K}_{\nabla,p} = \mathbf{K}_p^{\mathbf{T}}\mathbf{C}^{-1}\,\mathbf{K}_p + \mathbf{D}_p \tag{61b}$$

where the matrix $\mathbf{D}_p$ depends on second partial derivatives of $\mathbf{f}(\mathbf{a}^p)$ as follows:

$$\left\{\mathbf{D}_p\right\}_{ik} = \sum_{j=1,\ldots,N_\mathbf{f}} \frac{\partial^2 f_j(\mathbf{a}^p)}{\partial a_i \partial a_k}\left\{\mathbf{C}^{-1}(\mathbf{f}(\mathbf{a}^p) - \mathbf{f}^*)\right\}_j. \tag{61c}$$

Assuming that the elements of matrix $\mathbf{D}_p$ are small (e.g. if second derivatives are close to zero) one can write:

$$\mathbf{K}_{\nabla,p} = \mathbf{K}_p^{\mathbf{T}}\mathbf{C}^{-1}\,\mathbf{K}_p + \mathbf{D}_p \approx \mathbf{K}_p^{\mathbf{T}}\mathbf{C}^{-1}\,\mathbf{K}_p, \tag{61d}$$

Finally, using inverse matrices, MML solution can be written as:

$$\begin{aligned}\mathbf{a}^{p+1} &= \mathbf{a}^p - (\mathbf{K}_p^T\mathbf{C}^{-1}\mathbf{K}_p)^{-1}\,\nabla\Psi(\mathbf{a}) \\ &= \mathbf{a}^p - (\mathbf{K}_p^T\mathbf{C}^{-1}\mathbf{K}_p)^{-1}\,\mathbf{K}_p^T\mathbf{C}^{-1}(\mathbf{f}(\mathbf{a}^p)-\mathbf{f}^*).\end{aligned} \tag{62}$$

This equation is known as the *Gauss-Newton* method [49]. This equation can also be considered as *Quasi-Newton* method [3] due to using approximated matrix $\mathbf{K}_{\nabla,p.}$ Eq.(61d). For square matrices $\mathbf{K}$, Eq.(62) can be reduced to *Newton* iterations using matrix identity:$(\mathbf{K}_p^T\mathbf{C}^{-1}\mathbf{K}_p)^{-1} = \mathbf{K}_p^{-1}\mathbf{C}(\mathbf{K}_p^{\mathbf{T}})^{-1}$. This is why solving Eq. (2) with square $\mathbf{K}$, as well as, *Newton* method [Eq.(58)] also can be considered as a minimization of quadratic form [45].

It should be noted that quadratic form $\Psi(\mathbf{a})$ [Eq. (11a)] can be minimized by iterations different from Eq. (62). Many such methods also utilize gradient $\nabla\Psi(\mathbf{a})$ for the solution search. The steepest descent method deserves particular attention among all other techniques. This method correct solution guess $\mathbf{a}^p$ relying only on gradient $\nabla\Psi(\mathbf{a}^p)$ in point $\mathbf{a}^p$:

$$\mathbf{a}^{p+1} = \mathbf{a}^{p} - t_p \nabla\Psi(\mathbf{a})$$

$$= \mathbf{a}^{p} - t_p \mathbf{K}_p^{\mathbf{T}}\mathbf{C}^{-1}(\mathbf{f}(\mathbf{a}^{p})-\mathbf{f}^*), \tag{63}$$

where the coefficient $0 < t_p \leq 1$ is selected empirically to provide convergence. Since gradient $\nabla\Psi(\mathbf{a}^p)$ shows the direction of the strongest local change of $\Psi(\mathbf{a}^p)$, the steepest descent always converges [3, 44-45]. However, implementing this method may take a very long time [3].

5.3 *Levenberg- Marquardt* optimization of iteration convergence

Implementing non-linear inversion by *Newton*-like methods requires assurance of iteration convergence. Iteration by Eqs. (59, 62) may not converge or converge to a wrong solution. The convergence difficulties may be caused by inadequate choice of the initial guess and/or limitations of the linear approximation used for guess correction. Indeed, for strongly non-linear functions $f_j(\mathbf{a})$, the minimized form $\Psi(\mathbf{a})$ may have a complex structure with several minima. The analysis of this structure is desirable prior to inversion. However, when three or more unknowns are to be retrieved, such analysis is practically not feasible. Usually, researchers repeat retrieval with a set of initializations and select the best solution. The initializations and the criteria for selecting the best solution are commonly established based on the physical constraints of the application, experience, and intuition of the developers. Also, a convergence of non-linear solutions can be improved by modifying Eqs. (59, 62). The most established modification of *Gauss-Newton* iterations is widely known as the *Levenberg-Marquardt* method [4,49]:

$$\mathbf{a}^{p+1} = \mathbf{a}^{p} - t_p (\mathbf{K}_p^{\mathbf{T}}\mathbf{C}^{-1}\mathbf{K}_p + \gamma\, \mathbf{D})^{-1}\ \mathbf{K}_p^{\mathbf{T}}\mathbf{C}^{-1}(\mathbf{f}(\mathbf{a}^{p})-\mathbf{f}^*), \tag{64}$$

where matrix $\mathbf{D}$ and the coefficients $0 < t_p$ and $0 \leq \gamma$ are selected empirically to provide convergence. The matrix $\mathbf{D}$ is predominantly diagonal (unity matrix is often chosen as $\mathbf{D}$) and addition of the term $\gamma\ \mathbf{D}$ to $\mathbf{K}_p^{\mathbf{T}}\mathbf{C}^{-1}\mathbf{K}_p$ in Eq. (64) is analogous to using *a priori* constraints in linear inversions. Specifically, the matrix $\mathbf{K}_p^{\mathbf{T}}\mathbf{C}^{-1}\mathbf{K}_p$ can be singular on some of *p*-th iterations even if it is non-singular in the solution neighborhood. Adding the term $\gamma\ \mathbf{D}$ to $\mathbf{K}_p^{\mathbf{T}}\mathbf{C}^{-1}\mathbf{K}_p$ helps to pass the iteration process through areas of $\mathbf{K}_p^{\mathbf{T}}\mathbf{C}^{-1}\mathbf{K}_p$ singularities. As pointed out in [4], the *Levenberg-Marquardt* formula generalizes the steepest descent method. Namely, Eq. (64) can be reduced to (63) by defining matrix $\mathbf{D}$ in Eq. (64) as the unit matrix $\mathbf{I}$ and prescribing a large value to the parameter γ. Thus, Eq. (64) always converge with appropriate γ.

The multiplier $0 < t_p \leq 1$ in Eq. (64) is invoked mainly to decrease the length of $\Delta\mathbf{a}^p$, because the linear approximation may overestimate the correction $\Delta\mathbf{a}^p$. Usually, t_p is decreased by a factor (e.g. by 2) until a condition $\Psi(\mathbf{a}^{p+1}) < \Psi(\mathbf{a}^p)$ is satisfied. Underestimation of $\Delta\mathbf{a}^p$ does not lead to a convergence failure and may only slow down the arrival to a solution. The addition of the term $\gamma\ \mathbf{D}$ also reduces $\Delta\mathbf{a}^p$. Correspondingly, using both $\gamma\ \mathbf{D}$ ($\gamma > 0$) and $0 < t_p \leq 1$ in the same iteration may seem redundant because both

operations reduce $\Delta\mathbf{a}^p$. However, using the multiplier t_p is straightforward (compare to adding γ **D**) and sufficient in the application with moderately non-linear forward model with non-singular $\mathbf{K}_p^{\mathbf{T}}\mathbf{C}^{-1}\mathbf{K}_p$. On the other hand, if the matrix $\mathbf{K}_p^{\mathbf{T}}\mathbf{C}^{-1}\mathbf{K}_p$ is singular in some points $\mathbf{a}^p$, using $t_p \le 1$ does not help and inclusion of constraining term γ **D** is necessary. Thus, use of both t_p and γ **D** modifications in *Levenberg-Marquardt* [Eq. (64)] complement each other in practice.

5.4 Formulation of *Levenberg- Marquardt* iterations using statistical formalism

This Section is aimed to show that statistical considerations analogous to those in Section 4 can be useful for optimizing *Levenberg-Marquardt* iterations.

Gauss-Newton Eq. (62) trivially can be generalized for simultaneous inversion of multi-source data. Specifically, Eq. (25) with non-linear forward models $\mathbf{f}_k(\mathbf{a})$ can be solved by multi-term equivalent of non-linear LSM :

$$\hat{\mathbf{a}}^{p+1} = \hat{\mathbf{a}}^p - \left(\sum_{k=1}^{K} \gamma_k \left(\mathbf{K}_{k,p}\right)^{\mathrm{T}} \left(\mathbf{W}_k\right)^{-1} \left(\mathbf{K}_{k,p}\right)\right)^{-1} \left(\sum_{k=1}^{K} \gamma_k \left(\mathbf{K}_{k,p}\right)^{\mathrm{T}} \left(\mathbf{W}_k\right)^{-1} \left(\mathbf{f}_k\left(\hat{\mathbf{a}}^p\right) - \mathbf{f}_k^*\right)\right), \quad (65)$$

where $\mathbf{K}_{k,p}$ is *Jakobi* matrix of the first derivatives from $\mathbf{f}_k(\mathbf{a})$ in the vicinity of $\mathbf{a}^p$. As discussed above, employing linear approximations for non-linear functions $\mathbf{f}_k(\mathbf{a})$ in Eq. (65) may result in a convergence failure. Therefore, if some $\mathbf{f}_k(\mathbf{a})$ are linear, it seems logical to expect fewer problems with convergence. This idea can be elaborated by considering linear constraints applied to non-linear iterations. Namely, the non-linear equivalent of Eq. (47b) that solves Eq. (46) with non-linear $\mathbf{f}(\mathbf{a})$ can be written as:

$$\hat{\mathbf{a}}^p = \hat{\mathbf{a}}^{p+1} - \Delta\hat{\mathbf{a}}^p, \quad (66a)$$

$$\begin{aligned} &\left(\mathbf{K}_p^{\mathrm{T}}\mathbf{W}^{-1}\mathbf{K}_p + \gamma_2\mathbf{\Omega}_{\mathrm{m}} + \gamma_3\mathbf{W}_{\mathbf{a}*}^{-1}\right)\Delta\hat{\mathbf{a}}^p = \\ &\qquad = \mathbf{K}_p^{\mathrm{T}}\mathbf{W}^{-1}\left(\mathbf{f}\left(\hat{\mathbf{a}}^p\right) - \mathbf{f}^*\right) + \gamma_2\mathbf{\Omega}_{\mathrm{m}}\hat{\mathbf{a}}^p + \gamma_3\mathbf{W}_{\mathbf{a}*}^{-1}\left(\hat{\mathbf{a}}^p - \hat{\mathbf{a}}^*\right) \end{aligned}, \quad (66b)$$

where $\mathbf{\Omega}_{\mathrm{m}} = \mathbf{S}_{\mathrm{m}}^{\mathrm{T}}\mathbf{S}_{\mathrm{m}}$. Although, Eq. (66b) is constrained by *a priori* terms, solution $\hat{\mathbf{a}}^p$ may fail to converge because at initial iterations (p=1,2,…) the constrained non-linear Eq. (66) does not differ significantly from non-constrained Eq. (62), similar to basic *Gauss-Newton* iterations by Eq. (62). Indeed, if the initial guess is far from the solution, the values ($f_j(\mathbf{a}^p)$- f_j*) are large and the measurement term dominates over *a priori* terms because the values of the *Lagrange* multipliers γ_2 and γ_3 are typically small. *A priori* terms start to matter only when fitting differences ($\mathbf{f}(\mathbf{a}^p)$-$\mathbf{f}$*) reach the level of measurement accuracy ε_1. Therefore, some enhancement of *a priori* terms at initial iterations may improve performance of Eq. (66). This idea can be elaborated in the following considerations.

Each p-th iteration in Eq. (66) assumes the solution of the following overdetermined linear system:

$$\begin{cases} \mathbf{K}_{1,p}\ \Delta\mathbf{a}^p \approx \mathbf{f}_1(\mathbf{a}^p) - \mathbf{f}_1^* + \Delta\mathbf{f}_1^* + \Delta\mathbf{f}_{1,\mathrm{p}}^{\mathrm{lin}} \\ \mathbf{K}_{2,p}\ \Delta\mathbf{a}^p \approx \mathbf{f}_2(\mathbf{a}^p) - \mathbf{f}_2^* + \Delta\mathbf{f}_2^* + \Delta\mathbf{f}_{2,\mathrm{p}}^{\mathrm{lin}} \\ \mathbf{K}_{3,p}\ \Delta\mathbf{a}^p \approx \mathbf{f}_3(\mathbf{a}^p) - \mathbf{f}_3^* + \Delta\mathbf{f}_3^* + \Delta\mathbf{f}_{3,\mathrm{p}}^{\mathrm{lin}} \end{cases} \Rightarrow \begin{cases} \mathbf{K}_p\ \Delta\mathbf{a}^p \approx \mathbf{f}(\mathbf{a}^p) - \mathbf{f}^* + \Delta\mathbf{f}^* + \Delta\mathbf{f}_{\mathrm{p}}^{\mathrm{lin}} \\ \mathbf{S}_{\mathrm{m}}\ \Delta\mathbf{a}^p = \mathbf{S}_{\mathrm{m}}\mathbf{a}^p - \mathbf{0}^* + \Delta(\Delta^{\mathrm{m}}\mathbf{a})^* \\ \Delta\mathbf{a}^p = \mathbf{a}^p - \mathbf{a}^* + \Delta\mathbf{a}^* \end{cases}, \quad (67)$$

where $\Delta\mathbf{f}_{\mathrm{k,p}}^{\mathrm{lin}}$ denotes the errors of using the linear approximation of $(\mathbf{f}_{\mathrm{k}}(\mathbf{a}^{\mathrm{p}})\text{-}\mathbf{f}_{\mathrm{k}}{*})$ in the vicinity of $\mathbf{a}^{\mathrm{p}}$. In contrast to the linear case [Eq. (46)], Eq. (67) is written via differences $\Delta\mathbf{a}^{\mathrm{p}}$. Another difference is that the first equation in system (66) includes linearization errors $\Delta\mathbf{f}^{\mathrm{lin}}$. As discussed in Section 4, LSM optimization weights the contributions inversely to variances $\varepsilon_{\mathrm{k}}^2$ of errors $\Delta\mathbf{f}_{\mathrm{k}}^*$ [see Eqs. (29-31)]. Such weighting does not account for linearization errors $\Delta\mathbf{f}^{\mathrm{lin}}$ and, therefore, optimizes results only in close vicinity to the actual solution where $\Delta\mathbf{f}^{\mathrm{lin}}$ are small. Accounting for $\Delta\mathbf{f}^{\mathrm{lin}}$ can be introduced into LSM weighting by using $\varepsilon_1^2+(\varepsilon_{1,\mathrm{lin}})^2$ instead of ε_1^2 in the *Lagrange* multipliers definition [Eq. (33b)]. The value of the $\Delta\mathbf{f}_1^{lin}$ variance is not known at each point $\mathbf{a}^{\mathrm{p}}$, but can be estimated from the value of the residual, i.e. analogously to Eq. (52) one can write the following:

$$\hat{\varepsilon}_1^2 + (\hat{\varepsilon}_{1,\mathrm{lin}}(\mathbf{a}^{\mathrm{p}}))^2 \approx \frac{2\Psi(\mathbf{a}^{\mathrm{p}})}{\sum_{\mathrm{i}=1,\ldots,3}(N_{\mathbf{f}_i}) - N_{\mathbf{a}}} = \frac{2\Psi(\mathbf{a}^{\mathrm{p}})}{N_{\mathbf{f}^*} + N_{\mathbf{a}} - m}. \quad (68)$$

Using this equation, Eq. (30) can be re-written for non-linear iterations:

$$\gamma_{\mathrm{k}}\left(\mathbf{a}^{\mathrm{p}}\right) = \frac{\hat{\varepsilon}_1^2 + (\hat{\varepsilon}_{1,\mathrm{lin}}(\mathbf{a}^{\mathrm{p}}))^2}{\varepsilon_k^2} \approx \frac{2\Psi(\mathbf{a}^{\mathrm{p}})}{\varepsilon_k^2\left(\sum_{\mathrm{i}=1,\ldots,3}(N_{\mathbf{f}_i}) - N_{\mathbf{a}}\right)}. \quad (69)$$

This definition of the *Lagrange* multiplier accounts for higher linearization errors at earlier iterations. In close vicinity of the solution $\mathbf{a}'$, where $\Delta\mathbf{f}_1^{lin}$ is close to zero, Eq. (68) is reduced to Eq. (52). Hence, utilizing "adjustable" γ_{k} ($\mathrm{k} \geq 2$) in Eq. (66) improves convergence while the final solution retains the same statistical properties.

Derivations similar to those given by Eqs. (67-69) can be used even if actual *a priori* information is not available. For instance, if there is no *a priori* knowledge about magnitudes of unknowns or their correlations (smoothness), one can require such constraints on corrections $\Delta\mathbf{a}^{\mathrm{p}}$. Indeed, the restrictions on $\Delta\mathbf{a}^{\mathrm{p}} = \mathbf{a}^{\mathrm{p}} - \mathbf{a}^{\mathrm{p+1}}$ would not restrict the area of admissible solutions. For example, assuming that linearization may cause an overestimation of $\Delta\mathbf{a}^{\mathrm{p}}$, one may force $\Delta\mathbf{a}^{\mathrm{p}}$ to small values in order to retain monotonic convergence. Such constraint limits departures of $\mathbf{a}^{\mathrm{p+1}}$ from $\mathbf{a}^{\mathrm{p}}$, but it does not limit the values $\mathbf{a}^{\mathrm{p}}$. The possible negative side effect of limiting $\Delta\mathbf{a}^{\mathrm{p}}$ is a larger number of iterations if the initial guess $\mathbf{a}^0$ is taken far from the real solution. Similarly, retrieving functions $y(x)$ of $a_i = y(x_i)$, one may require smooth corrections

$\Delta y^p(x) = y^{p+1}(x) - y^p(x)$. This does not put constraints on retrieved $y(x)$ and, even if $y^0(x)$ was smooth, a large number of smooth corrections $\Delta y^p(x)$ may result in unsmooth $y^{p+1}(x)$. Thus, *Gauss-Newton* iterations can be implemented with use of constraints on $\Delta\mathbf{a}^p$ as follows:

$$\mathbf{K}_p\ \Delta\mathbf{a}^p \approx \mathbf{f}(\mathbf{a}^p) - \mathbf{f}^* + \Delta\mathbf{f}^* + \Delta\mathbf{f}_p^{\mathrm{lin}} \quad \Rightarrow \quad \begin{cases} \mathbf{K}_p\ \Delta\mathbf{a}^p \approx \mathbf{f}(\mathbf{a}^p) - \mathbf{f}^* + \Delta\mathbf{f}^* + \Delta\mathbf{f}_p^{\mathrm{lin}} \\ \mathbf{S}_{\mathrm{m}}\ \Delta\mathbf{a}^p = \mathbf{0}^* + \Delta(\Delta^{\mathrm{m}}\mathbf{a})^* \\ \Delta\mathbf{a}^p = \mathbf{0}^* + \Delta\mathbf{a}^* \end{cases} . \quad (70)$$

Here the second and third equations constrain smoothness and magnitudes of the corrections $y^p(x)$, respectively. Hence, *Gauss-Newton* iterations of Eq. (62) can be implemented as a multi-term LSM solving Eq. (70):

$$\Delta\hat{\mathbf{a}}^{p+1} = \hat{\mathbf{a}}^p - \left(\mathbf{K}_p^{\mathrm{T}}\mathbf{W}^{-1}\mathbf{K}_p + \gamma_2\mathbf{\Omega}_{\mathrm{m}} + \gamma_3\mathbf{W}_{\mathbf{a}^*}^{-1}\right)^{-1}\mathbf{K}_p^{\mathrm{T}}\mathbf{W}^{-1}\left(\mathbf{f}\left(\hat{\mathbf{a}}^p\right) - \mathbf{f}_k^*\right). \quad (71a)$$

Lagrange multipliers γ_2 and γ_3 can be determined analogously to Eq. (69) as:

$$\gamma_{\mathrm{k}}\left(\mathbf{a}^{\mathrm{p}}\right) = \frac{\hat{\varepsilon}_1^2 + (\hat{\varepsilon}_{1,\mathrm{lin}}(\mathbf{a}^{\mathrm{p}}))^2}{\varepsilon_k^2} \approx \frac{2\Psi(\mathbf{a}^{\mathrm{p}})}{\varepsilon_k^2\left(N_{\mathbf{f}^*} - N_{\mathbf{a}}\right)}. \quad (71b)$$

Following considerations of Section 4.6, the errors ε_2 and ε_3 can be established from the general knowledge of physically admissible variability of magnitudes of a_i and/or derivatives of retrieved $y^p(x)$.

Thus, the above derivation that optimizes the solution by constraining $\Delta\mathbf{a}^p$ resulting in Eq. (71), which is analogous to Eq. (64), provides additional insight to the formulating term $\gamma\ \mathbf{D}$ in the *Levenberg-Marquardt* iterations. Such an approach is employed in inversion algorithm [9] and shown to be efficient in practice for deriving aerosol properties from remote-sensing observations.

6. Possible adjustments to assumption of Normal noise

Most inversion-algorithm-optimizing solutions are based on the normal noise assumption when in the presence of random noise (see Sections 3-4). This includes even algorithms that are not based on statistical formalism, since minimization of quadratic forms is formally equivalent to assuming of Normal noise. However, in scientific literature, one can find numerous attempts of using alternative noise assumptions. Indeed, MML given by Eq. (8) does not assume this specific type of PDF and gives an optimized solution for any noise distribution, provided the assumed noise distribution is close to reality. For example, assuming that $P(f_j(\mathbf{a})\text{-}f_j^*) \sim \exp(-\left|(f_j(\mathbf{a})\text{-}f_j^*)\right|)$ leads to the *minimax* methods that differ from the LSM search for the least sum of absolute deviations. The details of implementing *minimax* and other methods based on the noise assumptions alternative to normal noise can be found in various

textbooks [3-4]. This Section discusses only a few modifications of the *Gaussian* noise assumption aimed to overcome particular difficulties in performance of the LSM.

6.1. Non-negativity constraints

As was discussed in Section 4.7, the difficulty in securing positive solutions in the retrieval of non-negative characteristics is a limitation of constrained linear inversion. This issue can be addressed by using a lognormal noise assumption in retrieval optimization [7-9]. Such assumption of lognormal noise leads to implementing inversions in logarithmic space, i.e. employing logarithmic transformation of forward model.

Retrieval of logarithms of a physical characteristic, instead of absolute values is an obvious way to avoid negative values for positively defined values. However, the literature devoted to inversion techniques tends to consider this apparently useful tactic as an artificial trick rather than a scientific approach to optimize solutions. Such misconception is probably caused by the fact that the pioneering efforts on inversion optimization by *Phillips* [11], *Tikhonov* [12] and *Twomey* [13] and many later theoretical considerations (e.g. *Hansen* [41]) were devoted to solving the *Fredholm* integral equation of the first kind, i.e. a system of linear equations produced by quadrature. The problems addressed by these methods are the retrieval of aerosol size distribution [35] or temperature profile of the atmosphere [19] by inverting spectral dependence of optical thickness. Considering optical thickness as a function of the logarithm of the aerosol concentrations or temperature profile requires replacing the initial linear equation $\mathbf{f} = \mathbf{K}\ \mathbf{a}$ by nonlinear ones $f_j = f_j(\ln a_i)$. On the face of it, such a transformation of linear problems to non-linear ones is not enthusiastically accepted by the scientific community as an optimization. On the other hand, in cases when a forward model is a nonlinear function of parameters to be retrieved (e.g., atmospheric remote sensing in cases when multiple scattering effects are significant), the retrieval of logarithms is accepted as a logical approach. Besides, as discussed in Section 4.7 non-linear *Chahine*-like iterations have proven to be efficient for inverting linear system. Rigorous statistical considerations also reveal some limitations in applying *Gaussian* functions for modeling errors in measurement of positively defined characteristics. It is well known that the curve of the normal distribution is symmetric. In other words, one may affirm that the assumption of a normal PDF is equivalent to the assumption of the principal possibility of obtaining negative results even in the case of physically nonnegative values. For such nonnegative characteristics as intensities, fluxes, etc., the choice of a log-normal distribution for describing the measurement noise seems to be more correct due to the following considerations: (i) log-normally distributed values are positively-defined; (ii) there are a number of theoretical and experimental reasons showing that for positively defined characteristics, the log-normal curve with its multiplicative errors (see [27]) is

closer to reality than normal noise with additive errors. Also, as follows from the discussion of statistical experiments [3], the lognormal distribution is best at modeling random deviations in non-negatives values. Besides, using the lognormal PDF for noise optimization does not require any revision of normal concepts and can be implemented by simple transformation of the problem to the space of normally distributed logarithms. This fact is very important from the viewpoint of both theoretical consideration and practical implementation of MML under the lognormal noise assumption. For example, due to the problem of differentiating $P(f_j(\mathbf{a})\text{-}f_j^*) \sim \exp(-\,|\,(f_j(\mathbf{a})\text{-}f_j^*)\,|\,)$, formulation of basic equations for *minimax* solutions is questionable.

Similar to the above considerations of non-negative measurements, there is a clear rational in retrieving logarithms of unknowns instead of their absolute values, e.g., $\ln(y(x_i))$ instead of $y(x_i)$, provided the retrieved characteristics are positively defined. Although, the MML does not implicitly assume a distribution of errors in the final solution, the statistical properties of the MML solution are well studied (see [27]) and, therefore can be projected in algorithm developments. In fact, according to statistical estimation theory, if PDF is normal, the MML estimates are also normally distributed. It is obvious then, that the LSM algorithm retrieving $y(x_i)$ would provide normally distributed estimates $y(x_i)$ and, therefore, it cannot provide zero probability for $y(x_i) < 0$, even if $y(x_i)$ are positively defined by nature. On the other hand, the retrieval of logarithms instead of absolute values eliminates the above contradiction because the LSM estimates of $\ln(y(x_i))$ would have a normal distribution of $\ln(y(x_i))$, i.e. a lognormal distribution of $y(x_i)$ that assures positivity of non-negative $y(x_i)$. Moreover, studies [7, 9] suggest considering the logarithmic transformation as one of the cornerstones of the practical efficiency of *Chahine's* iterative procedures. The derivations of Eqs. (55-56) from LSM formulated in logarithmic space are given in the Appendices of reference [9].

Thus, accounting for non-negativity of solutions and/or non-negativity of measurements can be implemented in the retrieval by using logarithms of unknowns ($a_i \rightarrow \ln a_i$) and/or measurements ($f_j \rightarrow \ln f_j$). In many situations, retrieval of absolute values or their logarithms is practically similar. This is because narrow lognormal or normal noise distributions are almost equivalent. For example, for small variations of non-negative value, the following relationship between Δa and $\Delta a/a$ is valid:

$$\Delta \ln a = \ln(a + \Delta a) - \ln(a) \approx \Delta a/a, \quad (\text{ if } \Delta \ln a << 1). \tag{72a}$$

Then, if only small relative variations of value a are allowed, the normal distribution of $\ln a$ is almost equivalent to the normal distribution of absolute values a. The covariance matrices of these distributions are connected as:

$$\mathbf{C}_{\ln \mathbf{a}} \approx (\mathbf{I}_\mathbf{a})^{-1}\, \mathbf{C}_\mathbf{a}\, (\mathbf{I}_\mathbf{a})^{-1}, \tag{72b}$$

where $\mathbf{I}_\mathbf{a}$ is a diagonal matrix with elements $\{\mathbf{I}_\mathbf{a}\}_{ii} = a_i$. Hence, for measurements with small relative errors, use of lognormal or normal PDFs

with covariance matrices related by Eq. (72) should give similar results. Also, since logarithmic errors can be considered approximately as relative errors, the variances $(\varepsilon_{\ln f})^2$ are unitless and, therefore, Eq. (30b) defining *Lagrange* multipliers as the variance ratio becomes particularly useful. Practical illustrations of using logarithmic transformations in inversion can be found in reference [9].

6.2. Accounting for the data redundancy

A difficulty in accounting for data redundancy is another unresloved issue in implementing optimized inversion. This issue has very high practical importance, although it is not often addressed in the literature on inversion methodologies. For example, infinite enhancement of spectral and/or angular resolutions in remote-sensing observation does not lead to accuracy improvements in retrievals above a certain level. Based on common sense this can be explained by the fact that simple increase of the number of observations $N_{\mathbf{f}}$ may lead to an increasing number of redundant measurements that do not help to improve retrievals. Theoretical considerations (e.g. in Section 3), however, do not assume any "redundant" or "useless" observations. Indeed, performing N_{f_j} straightforward repetitions of the same observation with established unchanged accuracy, from a statistical viewpoint, simply means that the variance of this particular observation f_j should decrease by factor N_{f_j}. Accordingly, the *j*-th elements of covariance matrix $\mathbf{C_f}$ should decrease and the errors of retrieved parameters [Eqs. (15), (49)] should decrease appropriately. Thus, from a theoretical viewpoint, repeating similar or even the same observation always results in some enhancements of retrieval accuracy. Such contradiction between practical experience and theoretical derivations seriously limits the efforts on estimating retrieval errors, evaluating information content of measurements and planning of optimum experiments. For the multi-term LSM approach presented here, accounting for data redundancy is also of particular importance. Indeed, individual data points from observations of the same type usually are comparable in accuracy. Therefore, it is unlikely, although possible, that inverting single-source data would not lead to a discrimination of some individual observations. In a multi-source inversion, the situation is different because an increase in the number of observations in one of several inverted sets of data would lead to an increase of the weight of this data set, even if the added observations were redundant from a practical point of view. Indeed, in the minimized quadratic form of $\Psi(\mathbf{a})$ in Eq. (29), the higher the value of the *k*-th term Ψ_k, the stronger the contribution of the *k*-th data set on the solution. Using known relationships for the χ^2 distribution, Ψ_k can be estimated as $\Psi_k \leq N_k$, i.e. the weight of the *k*-th term in Eq. (29) is proportional to the number of measurements N_k in the *k*-th data set. In order to eliminate this dependence of Ψ_k on N_k, it was suggested [9] that for redundant observations, the accuracy of a single measurement degrades as $1/N_k$ if several measurements are taken simultaneously, i.e.:

$$\varepsilon_k^2(\text{multiple}) = N_\text{k}\ \varepsilon_k^2(\text{single}), \tag{73}$$

where the term "multiple" indicates that several analogous measurements are made simultaneously. Correspondingly, Eq. (30b) can be written via accuracy of "single" measurement as follows:

$$\gamma_k = \frac{N_1\ \varepsilon_1^2(\text{single})}{N_\text{k}\ \varepsilon_k^2(\text{single})}. \tag{74}$$

This definition of γ_k makes the relative contributions of the terms $\gamma_\text{k}\ \Psi_\text{k}$ in Eq. (30a) independent of N_k, and therefore equalizes the data sets with different numbers of observation. Relationship (73) assumes that for data set with "redundant" observations ε_k increases as $\sqrt{N_k}$. Such an increase can be caused by the fact that the number of sources of random errors may increase proportionally to the number of simultaneous measurements. For example, increasing spectral and/or angular resolution in remote-sensing measurements likely results in a decrease of the quality of a single measurement due to increased complexity of the instrumentation and calibration. However, the assumption given by Eq. (73) is of intuitive character since it is not based on actual error analysis. Moreover, the developers of the instrumentation may argue justifiably that accuracy should not degrade if several measurements are taken at the same time. Therefore, it should be noted that Eq. (73) is appropriate only for data sets where actual redundancy has been achieved. Actually, the redundancy may be caused by other factors than instrumentation limits. For example, increasing angular and spectral resolution of satellite observations generally requires larger spatial integration and longer measurement time. Both these factors contribute to an increase of retrieval errors due to natural temporal and spatial variability of the atmosphere and surface.

Thus, identification of measurement redundancy in practice is a difficult effort that strongly relies on the experience of the developer. Nevertheless, it can be advisable to consider data redundancy as a practical factor that may affect retrieval. Namely, if Eq. (52) gives values much higher than the level of expected measurement errors (and retrieval errors are much higher than estimated from Eq.(49)), then it is likely that noise assumptions need to be verified. In such cases the ratios N_1/N_k can be good indications of magnitude and direction of required adjustments in ε_k^2 in order to address domination of the large inverted data sets over smaller ones. For example, assumption (73) was employed successfully in aerosol remote-sensing retrievals [9], where harmonization of the contribution of large sets of angular sky radiance measurements with much fewer observations of spectral optical thickness is beneficial. A similar principle was used in earlier studies [50].

7. Final recommendations

The considerations presented in this chapter were aimed to demonstrate that many important and well established ideas of numerical inversion can be combined and compliment each other in a single inversion methodology. Namely, it is suggested to combine all measured and *a priori* data in a single inversion procedure using the fundamental approach of MML. Under an assumption of normal noise, such an approach results in a multi-term LSM given by Eq. (32), where the contribution of each term is weighted by the values of the errors in the corresponding data set. The discussion in Sections 4-7 concludes that using LSM in the multi-term form allows fruitful connections between such well known and established techniques and methodologies as standard LSM, *Phillips-Tikhonov-Twomey* constrained inversion, and *Kalman-filter* type inversion methods, advocated in remote sensing by studies of *Rodgers* [19]. From a technical viewpoint, the derivation of a multi-term LSM is trivial and the main value of the approach presented is a deliberate consideration of various inversion aspects and approaches with the purpose of consolidating analogies and differences into a single unified concept. As a result, in addition to some generalization of inversion equations, a number of practically important conclusions and recommendations are proposed for implementing numerical inversions. For example, Section 4 suggests considering *Lagrange* multipliers as a ratio of error variances [see Eq. (30)], where variances of *a priori* constraints are related explicitly to knowledge of magnitudes of retrieved parameters. In the case of retrieval of smooth function $y(x)$, *Lagrange* multipliers can be written directly via maximum values of derivatives of $y(x)$. Section 4 also shows how smoothness constraints can be implemented in the retrieval of non-equidistantly binned functions $y(x)$ and how different types of *a priori* constraints can be employed in a single algorithm. Section 6 discusses the use of the multi-term LSM in non-linear Newtonian iterations for optimizing accuracy of the non-linear retrieval in the vicinity of a solution. Moreover, Section 6 shows that multi-term LSM can be used for implementing *Levenberg-Marquardt*-type modifications improving convergence of the non-linear iterations. Section 7 suggests modifications to normal noise assumptions to account for non-negativity of the physical values and addressing data redundancy. The lognormal noise assumption is applied for non-negative values. Under such assumptions, the MML principle results in a multi-term LSM written in logarithmic space. Also, Section 4 emphasizes distinction between two aspects of solution optimization: (i) accounting for distribution of errors in inverted data, and (ii) improving performance of mathematical inverse operations, e.g. replacing matrix inversion by other techniques. It is suggested that uniqueness of the solution should be assured by combining all available measurements and *a priori* information in a multi-term

LSM. Then, potentially advantageous mathematical techniques such as SVD, conjugated gradients, iterative search etc. can be used at the stage of solving normal Eqs. (31) for improving the performance of the technical implementation of multi-term LSM. For example, using steepest descent iterations for implementing logarithmic LSM allows (see [9]) the derivation *Chahine*-like iterations [17,14].

Finally, the derivations of the present chapter can be illustrated and summarized by a single formula written for the rather general case when the forward model is non-linear and *a priori* information on both magnitudes and smoothness of retrieved function $y(x)$ is available. If $y(x)$ needs to be retrieved from observations of two different characteristics $z_1(\lambda) = z_1(\lambda;y(x))$ and $z_2(\lambda) = z_2(\lambda;y(x))$ measured in a range of λ_i (λ can be angle, wavelength, etc.) then the optimized solution can be obtained by iterations:

$$\hat{\mathbf{a}}^p = \hat{\mathbf{a}}^{p+1} - t_p \Delta\hat{\mathbf{a}}^p, \tag{75a}$$

where $\Delta\hat{\mathbf{a}}^p$ is a solution of the normal system:

$$\left(\sum_{k=1}^{2}\gamma_k \mathbf{K}_{1,p}^{\mathrm{T}}\mathbf{W}_k^{-1}\mathbf{K}_{1,p} + \gamma_3\mathbf{\Omega}_{\mathrm{m}} + \gamma_4\mathbf{W}_{\mathbf{a}*}^{-1}\right)\Delta\hat{\mathbf{a}}^p = \\ = \sum_{k=1}^{2}\gamma_k \mathbf{K}_{k,p}^{\mathrm{T}}\mathbf{W}_k^{-1}\left(\mathbf{f}_k\left(\hat{\mathbf{a}}^p\right) - \mathbf{f}_k^*\right) + \gamma_3\mathbf{\Omega}_{\mathrm{m}}\hat{\mathbf{a}}^p + \gamma_4\mathbf{W}_{\mathbf{a}*}^{-1}\left(\hat{\mathbf{a}}^p - \hat{\mathbf{a}}^*\right). \tag{75b}$$

Here, $\mathbf{f}_{\mathrm{k}}$ is a measurement $f_{\mathrm{k}}(\lambda_i)$ and $\mathbf{a}$ is a vector of $a(x_j)$. If the measured functions $z_{\mathrm{k}}(\lambda)$ can be both positive and negative then normal noise is assumed and $f_{\mathrm{k}}(\lambda_i) = z_{\mathrm{k}}(\lambda_i)$. If $z_{\mathrm{k}}(\lambda)$ are positively defined (e.g. intensities) then the lognormal noise is expected and $f_{\mathrm{k}}(\lambda_i) = \ln(z_{\mathrm{k}}(\lambda_i))$. Similarly, if $y(x)$ can be both positive and negative then normally distributed errors are expected in retrievals and $a_i = y(x_i)$. If $y(x)$ is positively defined (e.g. concentration) then lognormal retrieval errors are expected and $a_i = \ln(y(x_i))$. Symbols $\mathbf{K}_{k,p}$ - matrices of the first derivatives

$$\left\{\mathbf{K}_{k,p}\right\}_{ji} = \left.\frac{\partial f_j(\lambda_i)}{\partial a_i}\right|_{\mathbf{a}^p}$$

calculated in the vicinity of $\mathbf{a}^p$; $\mathbf{W}_{\ldots}$ - weighting matrices defined by Eq. (30b). The smoothness matrix Ω_{m} is determined via matrices of m-th differences $\mathbf{S}_{\mathrm{m}}$ as $\Omega_{\mathrm{m}} = \mathbf{S}_{\mathrm{m}}^{\mathrm{T}}\mathbf{S}_{\mathrm{m}}$ if x_i are equidistant. If x_i are not equidistant, Ω_{m} is determined via matrices of the m-th derivatives $\mathbf{G}_{\mathrm{m}}$ as $\Omega_{\mathrm{m}} = \mathbf{G}_{\mathrm{m}}^{\mathrm{T}}\,\mathbf{W}_{\mathbf{g}*}^{-1}\mathbf{G}_{\mathrm{m}}$ (see Section 4.3). *Lagrange* multipliers γ_{k} are determined by ratios of variances, i.e.:

$$\gamma_1 = 1, \quad \text{and} \quad (\text{for } k \geq 2) \quad \gamma_k = \varepsilon_1^2/\varepsilon_k^2 \approx \hat{\varepsilon}_1^2\left(\mathbf{a}^p\right)/\varepsilon_k^2 \quad , \tag{76}$$

where $\hat{\varepsilon}_1^2(\mathbf{a}^p)$ is an estimate of ε_1^2:

$$\hat{\varepsilon}_1^2(\mathbf{a}^{\mathrm{p}}) \approx 2\Psi(\mathbf{a}^{\mathrm{p}}) \Big/ \left(\sum_{i=1,\ldots,4} N_{\mathbf{f}_i} - N_{\mathbf{a}}\right). \tag{77}$$

Here $\Psi(\mathbf{a}^{\mathrm{p}})$ denotes the value of the residual of the p-th iteration defined as

$$2\Psi(\hat{\mathbf{a}}) = \sum_{k=1}^{2} \gamma_k \left(\mathbf{f}_k(\hat{\mathbf{a}}) - \mathbf{f}_k^*\right)^{\mathrm{T}} \mathbf{W}_k^{-1} \left(\mathbf{f}_k(\hat{\mathbf{a}}) - \mathbf{f}_k^*\right) + \gamma_3 \hat{\mathbf{a}}^{\mathrm{T}} \Omega_{\mathrm{m}} \mathbf{a} + \gamma_4 \left(\hat{\mathbf{a}} - \hat{\mathbf{a}}^*\right)^{\mathrm{T}} \mathbf{W}_{\mathbf{a}^*}^{-1} \left(\hat{\mathbf{a}} - \hat{\mathbf{a}}^*\right). \tag{78}$$

The variances ε_k^2 (for $k \geq 2$) are determined before implementing iterations. ε_2^2 is the variance of the errors in the second set of measurements $z_2(\lambda)$. For the smoothness term (k =3), ε_3^2 can be implied from the knowledge of $y(x)$ m-th derivatives:

$$\varepsilon_3^* \approx \frac{\langle b_m \rangle (\Delta x)^{2m+1}}{N_{\mathbf{a}} - m} \quad \text{or} \quad \varepsilon_3^* \approx \frac{\langle b_m \rangle}{(N_{\mathbf{a}} - m)\, \Delta_m(x_1)} \ \ (\text{if } \Delta x_i \neq const), \tag{79}$$

where the first equation is for equidistant x_i, the second one is for non-equidistant x_i (Section 4.3), $\langle b_{\mathrm{m}} \rangle$ is the average norm of the of $y(x)$ m-th derivatives [Eq. (40)]. If average derivatives are unknown, $\langle b_{\mathrm{m}} \rangle$ can be implied using the m-th derivatives of the most unsmooth function $y(x)$ (see Section 4.6). For *a priori* estimates (k = 4), ε_4^2 is the variance of *a priori* estimate of a_1* and $\mathbf{W}_{\mathbf{a}^*} = \mathbf{C}_{\mathbf{a}^*}/\varepsilon_4^2$. If actual *a priori* estimates a_{i}* are not available, then a_{i}*, ε_4^2 and $\mathbf{W}_{\mathbf{a}^*}$ can be implied from the known variability ranges of a_{i} (Section 4.6).

Utilization of *Lagrange* multipliers dependent on residual $\Psi(\mathbf{a}^{\mathrm{p}})$ in Eqs. (76-77) helps to provide monotonic convergence of iterations. This operation is functionally analogous to the *Levenberg-Marquardt* method (see Section 5.4). This modification helps to provide a monotonic decrease of $\Psi(\mathbf{a}^{\mathrm{p}})$, i.e. monotonic convergence of iterations. Moreover, if no *a priori* information about the solution is available, the constraints can be applied to the *p-th* correction $\Delta\mathbf{a}^p$ instead of $\mathbf{a}^p$. Using such constraints affects only the convergence and does not bias the solution $\hat{\mathbf{a}}$. The correspondent multi-term LSM (same as Eq. (75b) with no *a priori* terms in the right part) is a full equivalent to the *Levenberg-Marquardt* method, with the only difference being that the terms added for improving convergence are clearly related with constraints on smoothness and magnitudes of $\Delta\mathbf{a}^p$ (see details in Section 5.4). The coefficient $1 \geq t_p > 0$ is used in Eq. (75a) similar to the *Levenberg-Marquardt* method. If $\Psi(\mathbf{a}^{\mathrm{p+1}}) > \Psi(\mathbf{a}^{\mathrm{p}})$, t_p should be decreased (e.g. as $t_p \rightarrow t_p/2$) until $\Psi(\mathbf{a}^{\mathrm{p}})$ is decreased.

If forward models $f_{\mathrm{k}}(\lambda_i)$ are linear, no iterations are needed and Eqs. (75) can be simplified as

$$\left(\sum_{k=1}^{2}\gamma_k \mathbf{K}_k^{\mathrm{T}}\mathbf{W}_k^{-1}\mathbf{K}_k + \gamma_3\Omega_{\mathrm{m}} + \gamma_4\mathbf{W}_{\mathbf{a}*}^{-1}\right)\hat{\mathbf{a}} = \sum_{k=1}^{2}\gamma_k \mathbf{K}_k^{\mathrm{T}}\mathbf{W}_k^{-1}\left(\mathbf{f}_k(\hat{\mathbf{a}}) - \mathbf{f}_k^*\right) + \gamma_4\mathbf{W}_{\mathbf{a}*}^{-1}\hat{\mathbf{a}}^*. \quad (80)$$

Here γ_k are given by Eq. (76), with the difference that ε_1^2 is fixed to the error variance in the first data set (k = 1). The assumed ε_1^2 should be close to the estimated $\hat{\varepsilon}_1^2(\hat{\mathbf{a}})$ obtained by Eq. (77) from the residual. A value of $\hat{\varepsilon}_1^2(\hat{\mathbf{a}})$ higher than assumed ε_1^2 indicates inconsistency in the assumptions made. One possibility is that the forward model needs corrections. Otherwise, adjustments are needed in assumptions about errors in measurements or *a priori* data. For example, in case the number of measurements N_k in the first (k = 1) and second (k = 2) sets of observations are very different, the ε_2^2 can be adjusted by a factor N_1/N_2 in order to account for data redundancy in one of the sets (see Section 6.2).

The covariance matrix of the random errors in the solution $\hat{\mathbf{a}}$ can be estimated in the linear approximation (Section 4.4):

$$\mathbf{C}_{\hat{\mathbf{a}}} = \left(\sum_{k=1}^{2}\gamma_k \mathbf{K}_k^{\mathrm{T}}\mathbf{W}_k^{-1}\mathbf{K}_k + \gamma_3\Omega_{\mathrm{m}} + \gamma_4\mathbf{W}_{\mathbf{a}*}^{-1}\right)^{-1}\hat{\varepsilon}_1^2(\hat{\mathbf{a}}). \quad (81)$$

For the non-linear case, the derivative matrix $\mathbf{K}_k$ is simulated in the vicinity of $\hat{\mathbf{a}}$.

Finally, linear equations (75b) and (79) can be solved by different methods. For example, using inverse matrices reduce Eqs. (75) and (80) to a traditional form of constrained inversion. Alternatively (Section 4.8), other numerical or computer techniques, such as, SVD, conjugated gradients, iterations, generic inversion, etc. can be used for solving Eqs. (75b) and (80).

4. Acknowledgments

I thank Alexander Sinyuk, Valery N. Shcherbakov, Gorden Videen and Ben Veihelmann for reading the chapter and providing useful comments.

References

1. S. Twomey, "Introduction to the Mathematics of Inversion in Remote Sensing and Indirect Measurements", (Elsevier, Amsterdam, 1977).
2. A.N. Tikhonov and V.Y. Arsenin, "Solution of Ill-Posed Problems" (Wiley, New York, 1977).
3. A. Tarantola, "Inverse Problem Theory: Methods for Data Fitting and Model Parameter Estimation", (Elsevier, Amsterdam, 1987).
4. W.H. Press, S.A. Teukolsky, W.T. Vetterling, B.P. Flannery, "Numerical Recipes in FORTRAN. The art of Scientific Computing", (Cambridge University Press, 1992).
5. A. Bakushinsky and A. Goncharsky, "Ill-Posed Problems: Theory and Applications" (Kluwer, Dordrecht, 1994).

6. C.D. Rodgers, "Inverse methods for atmospheric sounding: theory and practice", (World Scientific, Singapore, 2000).
7. O.V. Dubovik, T.V. Lapyonok, and S.L. Oshchepkov, Appl. Opt., **34**, 8422 (1995).
8. O. Dubovik, A. Smirnov, B. N. Holben, et al., J. Geophys. Res., **105**, 9791(2000).
9. O. Dubovik and M. D. King, J. Geophys. Res., **105**, 20673(2000).
10. R.E. Kalman, J. Basic. Engrg., 35, **82** (1960).
11. B.L. Phillips, J. Assoc. Comp. Mach, **9**, 84 (1962).
12. A.N. Tikhonov, Dokl. Akad. Nauk SSSR, **151**, 501 (1963).
13. S. Twomey, J. Assoc. Comp. Mach., **10**, 97 (1963).
14. S. Twomey, J. Comp. Phys, **18**, 188 (1975).
15. Strand, O. N., and E. R. Westwater, J. Assoc. Comput. Mach., **15**,100 (1968).
16. Strand, O. N., and E. R. Westwater, SIAM J. Numer. Anal., **5**, 287 (1968).
17. M.T. Chahine, J. Opt. Soc. Am., **12**, 1634 (1968).
18. V.F. Turchin and V.Z. Nozik, Izv. Acad. Nauk SSSR Fiz. Atmos. Okeana, **5**, 29 (1969).
19. C.D. Rodgers, Rev. Geophys. Space Phys., **14**, 609 (1976).
20. B.N. Holben, T.F. Eck, I. Slutsker et al., Remote Sens. Envir., 66, 1 (1998).
21. M.D. King, M.G. Strange, P. Leone, et al., J. Atmos. Oceanic Technol., **3**, 513 (1986).
22. D.J. Diner, J.C. Beckert, T.H. Reilly, et al., IEEE Trans. Geosci. Remote Sens., **36**, 1072 (1998).
23. Y. Sasano, M. Suzuki, T. Yokota, et al., Geophys. Res. Lett., **26**, 197 (1999).
24. P. Goloub, D. Tanre, J.L. Deuze, et al., IEEE Trans. Geosci. Remote Sens., **37**, 1586 (1999).
25. J. Chowdhary, B. Cairns, M. Mishchenko, et al., Geophys. Res. Lett., **28**, 243 (2001).
26. J. Chowdhary, B. Cairns, L.D. Travis, J. Atmos. Sci., **59**, 383 (2002).
27. W.T. Edie, D. Dryard, F.E. James, M. Roos, B. Sadoulet, "Statistical Methods in Experimental Physics", (North-Holland Publishing Company, Amsterdam, 1971)
28. C. R. Rao, "Linear Statistical Inference and Its Applications" (Wiley, New York, 1965).
29. G. A. Serber, "Linear Regression Analysis", (Wiley, New York, 1977).
30. A. Alpert, "Regression and the Moore-Penrose Pseudoiniverse" (Academic Press, New York, 1972).
31. A. N. Tikhonov, A. S. Leonov and A. G. Yagola, "Nonlinear Ill-Posed Problems", (Chapman & Hall, London 1998).
32. V.F. Turchin,V.P. Kozlov and M.S. Malkevich, Sov. Phys. Usp. Fiz.-USSR, **13**, 681 (1971).
33. A. Gelb, "Applied Optimal Estimation", (MIT Press, Cambridge, Mass., 1988).
34. D. Hartely and R. Prinn, J. Geophys. Res., **98**, 5183(1993).
35. M.D. King, D. M. Byrne, B. M. Herman et al., J. Atmos. Sci., **21**, 2153 (1978).
36. M.D. King, J. Atmos. Sci., **39**, 1356 (1982).
37. T. Nakajima, G. Tonna, R. Rao, et al., Appl. Opt., **35**, 2672 (1996).
38. D. Muller, U. Wandinger, A. Ansmann, Appl. Opt., **38**, 2346 (1999).
39. O.P. Hasekamp and J. Landgraf, J. Geophys. Res., **106**, 8077 (2001).
40. I. Veselovskii, A. Kolgotin, V. Griaznov et al., Appl. Opt., **41**, 3685 (2002).
41. P.S. Hansen, Inverse Problems, **8**, 849 (1992)
42. V.F. Turchin and L.S. Turovtse, Optika I Spectroskopiya, **36**, 280 (1974)

43. A.M. Obuhov, Izv. Acad, Sci. SSSR Geophys., 432 (1960)

44. G., Forsythe, and W. Wasow, "Finite Difference Methods for Partial Differential Equations", (Wiley, New York, 1960).

45. J.M. Ortega, "Introduction to Parallel and Vector Solution of Linear System", (Plenum Press, New York, 1988).

46. V. Krasnopolsky, L.C. Breaker, and W.H. Gemmill, J. Geophys. Res, **100**, 11,033 (1995).

47. D.E. Goldberg," Genetic algorithms in search, optimization, and machine learning", (Addison-Wesley Pub.,1989)

48. B.R. Lienert, J.N. Porter, S.K. Sharma J. Atmos. Ocean. Tech., **20**, 1403, (2003).

49. J.M. Ortega and W.C. Reinboldt, Iterative Solution of Nonlinear Equations in Several Variables, (Academic Press, New York, 1970).

50. S.L. Oshchepkov and O.V. Dubovik, J. Phys. D Appl. Phys., **26**, 728 (1993).