Article

Model Selection in Atmospheric Remote Sensing with an Application to Aerosol Retrieval from DSCOVR/EPIC, Part 1: Theory

Sruthy Sasi ¹ , Vijay Natraj ^{2,*} , Víctor Molina García ¹ , Dmitry S. Efremenko ¹ ,
Diego Loyola ¹ and Adrian Doicu ¹

¹ Remote Sensing Technology Institute, German Aerospace Center (DLR), 82234 Oberpfaffenhofen, Germany; Sruthy.Sasi@dlr.de (S.S.); Victor.MolinaGarcia@dlr.de (V.M.G.); dmitry.efremenko@dlr.de (D.S.E.); diego.loyola@dlr.de (D.L.); adrian.doicu@dlr.de (A.D.)

² Jet Propulsion Laboratory (NASA-JPL), California Institute of Technology, 4800 Oak Grove Drive, Pasadena, CA 91109, USA

* Correspondence: vijay.natraj@jpl.nasa.gov

Received: 1 October 2020; Accepted: 6 November 2020; Published: 12 November 2020

Abstract: The retrieval of aerosol and cloud properties such as their optical thickness and/or layer/top height requires the selection of a model that describes their microphysical properties. We demonstrate that, if there is not enough information for an appropriate microphysical model selection, the solution's accuracy can be improved if the model uncertainty is taken into account and appropriately quantified. For this purpose, we design a retrieval algorithm accounting for the uncertainty in model selection. The algorithm is based on (i) the computation of each model solution using the iteratively regularized Gauss–Newton method, (ii) the linearization of the forward model around the solution, and (iii) the maximum marginal likelihood estimation and the generalized cross-validation to estimate the optimal model. The algorithm is applied to the retrieval of aerosol optical thickness and aerosol layer height from synthetic measurements corresponding to the Earth Polychromatic Imaging Camera (EPIC) instrument onboard the Deep Space Climate Observatory (DSCOVR) satellite. Our numerical simulations show that the heuristic approach based on the thesolution minimizing the residual, which is frequently used in literature, is completely unrealistic when both the aerosol model and surface albedo are unknown.

Keywords: model selection; retrieval; DSCOVR/EPIC

1. Introduction

The limited information provided by satellite measurements does not allow for a complete determination of aerosol and cloud properties and, in particular, their microphysical properties. To deal with this problem, a set of models describing the microphysical properties is typically used. For example, the aerosol models implemented in the Moderate Resolution Imaging Spectroradiometer (MODIS) aerosol algorithm over land [1] consist of three spherical, fine-dominated types and one spheroidal, coarse-dominated type, while typical cloud models consist of water and ice clouds with a predefined particle shape and size. In this regard, a priori information such as the selection of a microphysical model is an essential part of the retrieval process. Standard retrieval algorithms usually ignore model uncertainty, i.e., a model is chosen from a given set of candidate models, and the retrieval is performed as if the selected model represents the true state. This approach is very risky and may result in large errors in the retrieved parameters if the model does not reflect the real scenario. An efficient way to quantify the uncertainty in model selection is the Bayesian approach and, in particular, Bayesian model selection and model averaging [2]. By model selection, we mean the

specific problem of choosing the most suitable model from a given set of candidate models. In general, model selection is not a trivial task because for a given measurement, there will be several models that fit the data equally well. These tools were used, among others, in Refs. [3,4] to study uncertainty quantification in remote sensing of aerosols from Ozone Monitoring Instrument (OMI) data.

The key quantity in Bayesian model selection is relative evidence, which is a measure of how well the model fits the measurement. This is expressed in terms of the marginal likelihood, which is the integral over the state vector space of the a priori density times the likelihood density (see below). An accurate computation of the marginal likelihood is not a trivial task because some parameters characterizing the a priori and likelihood densities, e.g., the data error variance and the a priori state variance, are not precisely known and must be estimated. Moreover, the integral over the state vector space cannot be computed analytically because the likelihood density depends on the nonlinear forward model, for which an analytical representation is not available.

In this paper, we aim to eliminate the drawbacks of the Bayesian approach by using:

1. an iteratively regularized Gauss–Newton method for computing the solution of each model and estimating the ratio of the data error variance and the a priori state variance;
2. a linearization of the forward model around the solution for integrating the likelihood density over the state vector;
3. parameter choice methods from linear regularization theory for estimating the optimal model and the data error variance.

The paper is organized as follows. In Section 2, we derive a scaled data model with white noise in order to simplify the subsequent analysis. Section 3 is devoted to a review of the Bayesian parameter estimation and model selection in order to describe the problem and clarify the nomenclature. In Section 4, we summarize the iteratively regularized Gauss–Newton method and emphasize some special features of this method, including the estimation of the ratio of the data error variance and the a priori state variance. In Section 5, we assume a linearization of the forward model around the solution and, under this assumption, extend some regularization parameter choice methods for linear problems (the maximum marginal likelihood estimation and the generalized cross-validation) for data error variance estimation and model selection. Section 6 summarizes the computational steps of the proposed retrieval algorithm. In Section 7, we apply the algorithm to the retrieval of aerosol optical thickness and aerosol layer height from the Earth Polychromatic Imaging Camera (EPIC)/Deep Space Climate Observatory (DSCOVR) measurements.

2. Data Model

Consider N_m microphysical models, and let $\mathbf{F}_m(\mathbf{x}) \in \mathbb{R}^M$ be the forward model corresponding to the model m for $m = 1, \dots, N_m$. More specifically, in our analysis, $\mathbf{F}_m(\mathbf{x})$ is the vector of the logarithms of the simulated radiances at all wavelengths λ_i , $i = 1, \dots, M$, i.e., $[\mathbf{F}_m(\mathbf{x})]_i = \ln I_m(\lambda_i, \mathbf{x})$, and $\mathbf{x} \in \mathbb{R}^N$ is the state vector encapsulating the atmospheric parameters to be retrieved, e.g., aerosol optical thickness and layer height.

The nonlinear data model under examination consists of the equations:

$$\mathbf{y}^\delta = \mathbf{y} + \delta_{\text{mes}}, \quad (1)$$

and:

$$\mathbf{y} = \mathbf{F}_m(\mathbf{x}) + \delta_{\text{aerm}} \quad (2)$$

where $\mathbf{y}^\delta \in \mathbb{R}^M$ is the measurement vector or the noisy data vector, $\mathbf{y}$ the exact data vector, $\delta_{\text{mes}} \in \mathbb{R}^M$ the measurement error vector, and $\delta_{\text{aerm}} \in \mathbb{R}^M$ the model error vector, i.e., the error due to an inadequate model. Defining the (total) data error vector by:

$$\delta_m = \delta_{\text{mes}} + \delta_{\text{aerm}},$$

the data model (1) and (2) becomes:

$$\mathbf{y}^\delta = \mathbf{F}_m(\mathbf{x}) + \delta_m. \quad (3)$$

In a deterministic setting, δ_m is characterized by the noise level Δ_m (defined as an upper bound for δ_m , i.e., $\|\delta_m\| \leq \Delta_m$), $\mathbf{x}$ is a deterministic vector, and we are faced with the solution of the nonlinear equation $\mathbf{y}^\delta = \mathbf{F}_m(\mathbf{x})$. In a stochastic setting, δ_m and $\mathbf{x}$ are random vectors, and the data model (3) is solved by means of a Bayesian approach.

Using the prewhitening technique, we transform the data model (3) into a model with white noise. For this purpose, we take:

1. δ_{mes} as a Gaussian random vector with zero mean and covariance matrix:

$$\begin{aligned} \mathbf{C}_{\text{mes}} &= \text{diag}(\sigma_{\text{mesi}}^2)_M = \sigma_{\text{mes}}^2 \mathbf{C}_{\text{mes}}, \\ \mathbf{C}_{\text{mes}} &= \text{diag}\left(\frac{\sigma_{\text{mesi}}^2}{\sigma_{\text{mes}}^2}\right)_{M'}, \quad \sigma_{\text{mes}}^2 = \frac{1}{M} \sum_{i=1}^M \sigma_{\text{mesi}}^2, \end{aligned}$$

and

2. δ_{aerm} as a Gaussian random vector with zero mean and covariance matrix:

$$\mathbf{C}_{\text{aerm}} = \sigma_{\text{aerm}}^2 \mathbf{I}_M.$$

If δ_{mes} and δ_{aerm} are independent random vectors, then δ_m is a Gaussian random vector with zero mean and covariance matrix:

$$\begin{aligned} \mathbf{C}_{\delta m} &= \mathbf{C}_{\text{mes}} + \mathbf{C}_{\text{aerm}} = \sigma_m^2 \mathbf{C}_{\delta m}, \\ \mathbf{C}_{\delta m} &= w_{\text{mes}m} \mathbf{C}_{\text{mes}} + (1 - w_{\text{mes}m}) \mathbf{I}_M, \end{aligned}$$

where:

$$\sigma_m^2 = \sigma_{\text{mes}}^2 + \sigma_{\text{aerm}}^2$$

is the variance of the data error vector and:

$$w_{\text{mes}m} = \sigma_{\text{mes}}^2 / \sigma_m^2$$

is the weighting factor giving the contribution of $\mathbf{C}_{\text{mes}}$ to the covariance matrix $\mathbf{C}_{\delta m}$. After these preliminary constructions, we introduce the scaled data model:

$$\bar{\mathbf{y}}^\delta = \bar{\mathbf{F}}_m(\mathbf{x}) + \bar{\delta}_m, \quad (4)$$

where:

$$\bar{\mathbf{y}}^\delta = \mathbf{P} \mathbf{y}^\delta, \quad \bar{\mathbf{F}}_m(\mathbf{x}) = \mathbf{P} \mathbf{F}_m(\mathbf{x}), \quad \bar{\delta}_m = \mathbf{P} \delta_m,$$

and:

$$\mathbf{P} = \mathbf{C}_{\delta m}^{-1/2} = \text{diag}\left(\left[1 - w_{\text{mes}m}\left(1 - \frac{\sigma_{\text{mesi}}^2}{\sigma_{\text{mes}}^2}\right)\right]^{-1/2}\right)_M \quad (5)$$

is the scaling matrix. Taking into account that:

$$\bar{\mathbf{C}}_{\delta m} = \mathcal{E}\{\bar{\delta}_m \bar{\delta}_m^T\} = \sigma_m^2 \mathbf{P} \mathbf{C}_{\delta m} \mathbf{P}^T = \sigma_m^2 \mathbf{I}_M,$$

we see that by means of the prewhitening technique, the Gaussian random vector $\delta_m \sim \mathcal{N}(\mathbf{0}, \mathbf{C}_{\delta m} = \sigma_m^2 \mathbf{C}_{\delta m})$ is transformed into the white noise $\bar{\delta}_m \sim \mathcal{N}(\mathbf{0}, \bar{\mathbf{C}}_{\delta m} = \sigma_m^2 \mathbf{I}_M)$. Here and in the following, the notation $\mathcal{N}(\mathbf{x}_{\text{mean}}, \mathbf{C}_x)$ stands for a normal distribution with mean $\mathbf{x}_{\text{mean}}$ and covariance matrix $\mathbf{C}_x$.

The following comments can be taken into consideration:

1. In [3], the covariance matrix $\mathbf{C}_{aerm}$ was estimated by empirically exploring a set of residuals of model fits to the measurements. Essentially, $\mathbf{C}_{aerm}$ is assumed to be in the form:

$$[\mathbf{C}_{aerm}]_{ij} = \begin{cases} \sigma_1^2 \exp[-(\lambda_i - \lambda_j)^2 / l^2], & i \neq j \\ \sigma_0^2 + \sigma_1^2, & i = j \end{cases}, \quad (6)$$

where σ_0^2 , σ_1^2 , and l (representing the non-spectral diagonal variance, the spectral variance, and the correlation length, respectively) are computed by means of an empirical semivariogram, i.e., the variances of the residual differences are calculated for each wavelength pair with the distance $d = |\lambda_i - \lambda_j|$, and a theoretical Gaussian variogram model is fitted to these empirical semivariogram values. As compared to Equation (6), in our analysis, we use the diagonal matrix approximation $\mathbf{C}_{aerm} \approx \sigma_{aerm}^2 \mathbf{I}_M$ with $\sigma_{aerm}^2 = \sigma_0^2 + \sigma_1^2$.

2. In principle, the scaling matrix $\mathbf{P}$ depends on the model error variance σ_{aerm}^2 through the weighting factor w_{mesm} . However, for $[\mathbf{F}_m(\mathbf{x})]_i = \ln I_m(\lambda_i, \mathbf{x})$, we usually have $\sigma_{mesi}^2 \approx \sigma_{mes}^2$ for all $i = 1, \dots, M$. In this case, it follows that $\mathbf{C}_{mes}$, together with $\mathbf{C}_{\delta m}$ and $\mathbf{P}$, are close to the identity matrix. This result does not mean that in our model, σ_{aerm}^2 does not play any role; σ_{aerm}^2 is included in σ_m^2 , which is the subject of an estimation process.

3. Bayesian Approach

In this section, we review the basics of Bayesian parameter estimation and model selection by following the analysis given in Refs. [2,3].

3.1. Bayesian Parameter Estimation

In statistical inversion theory, the state vector $\mathbf{x}$ is assumed to be a random vector, and in this regard, we take $\mathbf{x} \sim \mathcal{N}(\mathbf{x}_a, \mathbf{C}_x)$, where $\mathbf{x}_a$ is the a priori state vector, the best beforehand estimate of the solution, and:

$$\begin{aligned} \mathbf{C}_x &= \text{diag}(\sigma_{xi}^2)_N = \sigma_x^2 \mathbf{C}_x, \\ \mathbf{C}_x &= \text{diag}\left(\frac{\sigma_{xi}^2}{\sigma_x^2}\right)_{N'}, \quad \sigma_x^2 = \frac{1}{N} \sum_{i=1}^N \sigma_{xi}^2. \end{aligned}$$

Furthermore, the matrix $\mathbf{L}$ is defined through the Cholesky factorization:

$$\mathbf{C}_x^{-1} = \mathbf{L}^T \mathbf{L}, \text{ that is, } \mathbf{L} = \text{diag}\left(\frac{\sigma_x}{\sigma_{xi}}\right)_{N'}$$

and the parameter α is defined as:

$$\alpha = \frac{\sigma_m^2}{\sigma_x^2}. \quad (7)$$

The data model (3) gives a relation between the three random vectors $\bar{\mathbf{y}}^\delta$, $\mathbf{x}$, and δ_m , and therefore, their probability densities depend on each other. The following probability densities are relevant in Bayesian parameter estimation: (i) the a priori density $p(\mathbf{x} | m)$, which represents the conditional probability density of $\mathbf{x}$ given the model m before performing the measurement $\bar{\mathbf{y}}^\delta$, (ii) the likelihood density $p(\bar{\mathbf{y}}^\delta | \mathbf{x}, m)$, which represents the conditional probability density of $\bar{\mathbf{y}}^\delta$ given the state $\mathbf{x}$ and the model m , and (iii) the a posteriori density $p(\mathbf{x} | \bar{\mathbf{y}}^\delta, m)$, which represents the conditional probability density of $\mathbf{x}$ given the data $\bar{\mathbf{y}}^\delta$ and the model m .

The Bayes theorem of inverse problems relates the a posteriori density to the likelihood density:

$$\begin{aligned} p(\mathbf{x} | \bar{\mathbf{y}}^\delta, m) &= \frac{p(\bar{\mathbf{y}}^\delta | \mathbf{x}, m)p(\mathbf{x}, m)}{p(\bar{\mathbf{y}}^\delta, m)} \\ &= \frac{p(\bar{\mathbf{y}}^\delta | \mathbf{x}, m)p(\mathbf{x} | m)}{p(\bar{\mathbf{y}}^\delta | m)}. \end{aligned}$$

In Bayesian parameter estimation, the marginal likelihood density $p(\bar{\mathbf{y}}^\delta | m)$, defined by:

$$p(\bar{\mathbf{y}}^\delta | m) = \int p(\mathbf{x}, \bar{\mathbf{y}}^\delta | m) d\mathbf{x} = \int p(\bar{\mathbf{y}}^\delta | \mathbf{x}, m)p(\mathbf{x} | m) d\mathbf{x}, \quad (8)$$

plays the role of a normalization constant and is usually ignored. However, as we will see in the next section, this density is of particular importance in Bayesian model selection.

For $\mathbf{x} \sim N(\mathbf{x}_a, \mathbf{C}_x = \sigma_m^2(\alpha \mathbf{L}^T \mathbf{L})^{-1})$ and $\bar{\delta}_m \sim N(\mathbf{0}, \bar{\mathbf{C}}_{\delta_m} = \sigma_m^2 \mathbf{I}_M)$, the probability densities $p(\mathbf{x} | m)$ and $p(\bar{\mathbf{y}}^\delta | \mathbf{x}, m)$ are given by:

$$\begin{aligned} p(\mathbf{x} | m) &= \frac{1}{\sqrt{(2\pi\sigma_m^2)^N \det[(\alpha \mathbf{L}^T \mathbf{L})^{-1}]}} \\ &\times \exp\left[-\frac{\alpha}{2\sigma_m^2} \|\mathbf{L}(\mathbf{x} - \mathbf{x}_a)\|^2\right] \end{aligned} \quad (9)$$

and:

$$p(\bar{\mathbf{y}}^\delta | \mathbf{x}, m) = \frac{1}{\sqrt{(2\pi\sigma_m^2)^M}} \exp\left[-\frac{1}{2\sigma_m^2} \|\bar{\mathbf{y}}^\delta - \bar{\mathbf{F}}_m(\mathbf{x})\|^2\right], \quad (10)$$

respectively. The Bayes formula yields the following expression for the a posteriori density:

$$p(\mathbf{x} | \bar{\mathbf{y}}^\delta, m) \propto \exp\left[-\frac{1}{2} V_\alpha(\mathbf{x} | \bar{\mathbf{y}}^\delta, m)\right], \quad (11)$$

where the a posteriori potential $V_\alpha(\mathbf{x} | \bar{\mathbf{y}}^\delta, m)$ is defined by:

$$V_\alpha(\mathbf{x} | \bar{\mathbf{y}}^\delta, m) = \frac{1}{\sigma_m^2} \left[\|\bar{\mathbf{y}}^\delta - \bar{\mathbf{F}}_m(\mathbf{x})\|^2 + \alpha \|\mathbf{L}(\mathbf{x} - \mathbf{x}_a)\|^2 \right]. \quad (12)$$

The maximum a posteriori estimator $\hat{\mathbf{x}}_{m\alpha}^\delta$ maximizing the conditional probability density $p(\mathbf{x} | \bar{\mathbf{y}}^\delta, m)$ also minimizes the potential $V_\alpha(\mathbf{x} | \bar{\mathbf{y}}^\delta, m)$, i.e.,

$$\hat{\mathbf{x}}_{m\alpha}^\delta = \arg \min_{\mathbf{x}} V_\alpha(\mathbf{x} | \bar{\mathbf{y}}^\delta, m). \quad (13)$$

3.2. Bayesian Model Selection

The relative evidence, also known as the a posteriori probability of the model m given the measurement $\bar{\mathbf{y}}^\delta$, $p(m | \bar{\mathbf{y}}^\delta)$, is used for model comparison. By taking into consideration the Bayes theorem, this conditional probability density is defined by:

$$p(m | \bar{\mathbf{y}}^\delta) = \frac{p(\bar{\mathbf{y}}^\delta | m)p(m)}{p(\bar{\mathbf{y}}^\delta)}.$$

The mean formula:

$$p(\bar{\mathbf{y}}^\delta) = \sum_{m=1}^{N_m} p(\bar{\mathbf{y}}^\delta | m)p(m)$$

and the assumption that all models are equally likely, i.e., $p(m) = 1/N_m$, yield:

$$p(m | \bar{\mathbf{y}}^\delta) = \frac{p(\bar{\mathbf{y}}^\delta | m)}{\sum_{m=1}^{N_m} p(\bar{\mathbf{y}}^\delta | m)}. \quad (14)$$

Intuitively, the model with the highest evidence is the best among all the models involved, and in this regard, we define the maximum solution estimate as:

$$\hat{\mathbf{x}}_{\max}^\delta = \hat{\mathbf{x}}_{m^*}^\delta, \quad m^* = \arg \max_m p(m | \bar{\mathbf{y}}^\delta). \quad (15)$$

In fact, we can compare the models to see if one of them clearly shows the highest evidence, or if there are several models with comparable values of evidence. When several models can provide equally good fits to the measurement, we can use the Bayesian model averaging technique to combine the individual a posteriori densities weighted by their evidence. Specifically, using the relation:

$$\begin{aligned} p(\mathbf{x}, \bar{\mathbf{y}}^\delta) &= \sum_{m=1}^{N_m} p(\mathbf{x}, \bar{\mathbf{y}}^\delta, m) \\ &= \sum_{m=1}^{N_m} p(\mathbf{x} | \bar{\mathbf{y}}^\delta, m) p(m | \bar{\mathbf{y}}^\delta) p(\bar{\mathbf{y}}^\delta) \end{aligned}$$

on the one hand, and $p(\mathbf{x}, \bar{\mathbf{y}}^\delta) = p(\mathbf{x} | \bar{\mathbf{y}}^\delta) p(\bar{\mathbf{y}}^\delta)$, on the other hand, we are led to the Bayesian model averaging formula:

$$p_{\text{mean}}(\mathbf{x} | \bar{\mathbf{y}}^\delta) = \sum_{m=1}^{N_m} p(\mathbf{x} | \bar{\mathbf{y}}^\delta, m) p(m | \bar{\mathbf{y}}^\delta), \quad (16)$$

and, consequently, to the mean solution estimate:

$$\hat{\mathbf{x}}_{\text{mean}}^\delta = \sum_{m=1}^{N_m} \hat{\mathbf{x}}_{m\alpha}^\delta p(m | \bar{\mathbf{y}}^\delta). \quad (17)$$

The above analysis shows that in Bayesian parameter estimation and model selection, we are faced with the following problems:

Problem 1. From Equations (11)–(13), it is apparent that the computation of the estimator $\hat{\mathbf{x}}_{m\alpha}^\delta$ requires the knowledge of the parameter α , i.e., the ratio of the data error variance and the a priori state variance.

Problem 2. From Equations (15) and (17), we see that the solution estimates $\hat{\mathbf{x}}_{\max}^\delta$ and $\hat{\mathbf{x}}_{\text{mean}}^\delta$ are expressed in terms of relative evidence $p(m | \bar{\mathbf{y}}^\delta)$, which in turn, according to Equation (14), is expressed in terms of the marginal likelihood density $p(\bar{\mathbf{y}}^\delta | m)$. In view of Equation (8), the computation of the marginal likelihood density $p(\bar{\mathbf{y}}^\delta | m)$ requires the knowledge of the likelihood density $p(\bar{\mathbf{y}}^\delta | \mathbf{x}, m)$, and therefore of Equation (10), of the data error variance σ_m^2 .

Problem 3. The dependency of the likelihood density $p(\bar{\mathbf{y}}^\delta | \mathbf{x}, m)$ on the nonlinear forward model $\bar{\mathbf{F}}_m(\mathbf{x})$ does not allow an analytical integration in Equation (8).

4. Iteratively Regularized Gauss–Newton Method

In the framework of Tikhonov regularization, a regularized solution $\mathbf{x}_{m\alpha}^\delta$ to the nonlinear equation $\bar{\mathbf{y}}^\delta = \bar{\mathbf{F}}_m(\mathbf{x})$ is computed as the minimizer of the Tikhonov function:

$$\mathcal{F}_{m\alpha}(\mathbf{x}) = \|\bar{\mathbf{y}}^\delta - \bar{\mathbf{F}}_m(\mathbf{x})\|^2 + \alpha \|\mathbf{L}(\mathbf{x} - \mathbf{x}_a)\|^2, \quad (18)$$

that is,

$$\mathbf{x}_{m\alpha}^\delta = \arg \min_{\mathbf{x}} \mathcal{F}_{m\alpha}(\mathbf{x}). \quad (19)$$

In classical regularization theory, the first term on the right-hand side of Equation (18) is the squared residual; the second one is the penalty term; while α and $\mathbf{L}$ are the regularization parameter and the regularization matrix, respectively. From Equations (12) and (18), we see that $V_\alpha(\mathbf{x} | \bar{\mathbf{y}}^\delta, m) = (1/\sigma_m^2)\mathcal{F}_{m\alpha}(\mathbf{x})$, and from Equations (13) and (19), we deduce that $\hat{\mathbf{x}}_{m\alpha}^\delta = \mathbf{x}_{m\alpha}^\delta$. Thus, the maximum a posteriori estimate coincides with the Tikhonov solution, and therefore, the Bayesian parameter estimation can be regarded as a stochastic version of the method of Tikhonov regularization with an a priori chosen regularization parameter α .

In the framework of Tikhonov regularization, the computation of the optimal regularization parameter is a crucial issue. With too little regularization, reconstructions deviate significantly from the a priori, and the solution is said to be under-regularized. With too much regularization, the reconstructions are too close to the a priori, and the solution is said to be over-regularized. In the Bayesian framework, the optimal regularization parameter is identified as the true ratio of the data error variance and the a priori state variance.

Several regularization parameter choice methods were discussed in [5]. These include methods with constant regularization parameters, e.g., the maximum likelihood estimation, the generalized cross-validation, and the nonlinear L-curve method. Unfortunately, at present, there is no fail-safe regularization parameter choice method that guarantees small solution errors in any circumstance, that is for any errors in the data vector.

An improvement of the problems associated with the regularization parameter selection is achieved in the framework of the so-called iterative regularization methods. These approaches are less sensitive to overestimates of the regularization parameter, but require more iteration steps to achieve convergence. A representative iterative approach is the iteratively regularized Gauss–Newton method.

At iteration step k of the iteratively regularized Gauss–Newton method, the forward model $\bar{\mathbf{F}}_m(\mathbf{x})$ is linearized around the current iterate $\mathbf{x}_{mk}^\delta$,

$$\bar{\mathbf{F}}_m(\mathbf{x}) \approx \bar{\mathbf{F}}_m(\mathbf{x}_{mk}^\delta) + \bar{\mathbf{K}}_{mk}(\mathbf{x} - \mathbf{x}_{mk}^\delta),$$

where:

$$\bar{\mathbf{K}}_{mk} = \frac{\partial \bar{\mathbf{F}}_m}{\partial \mathbf{x}}(\mathbf{x}_{mk}^\delta)$$

is the Jacobian matrix at $\mathbf{x}_{mk}^\delta$, and the nonlinear equation $\bar{\mathbf{F}}_m(\mathbf{x}) = \bar{\mathbf{y}}^\delta$ is replaced by its linearization:

$$\bar{\mathbf{K}}_{mk}\mathbf{p} = \bar{\mathbf{y}}_{mk}^\delta, \quad (20)$$

where:

$$\mathbf{p} = \mathbf{x} - \mathbf{x}_a$$

is the step vector with respect to the a priori and:

$$\bar{\mathbf{y}}_{mk}^\delta = \bar{\mathbf{y}}^\delta - \bar{\mathbf{F}}_m(\mathbf{x}_{mk}^\delta) + \bar{\mathbf{K}}_{mk}(\mathbf{x}_{mk}^\delta - \mathbf{x}_a)$$

is the linearized data vector at iteration step k .

Since the nonlinear problem is ill-posed, its linearization is also ill-posed. Therefore, the linearized Equation (20) is solved by means of Tikhonov regularization with the penalty term $\alpha_k \|\mathbf{Lp}\|^2$, where the regularization parameters α_k are the terms of a decreasing (geometric) sequence, i.e., $\alpha_k = q\alpha_{k-1}$ with $q < 1$. Note that in the method of Tikhonov regularization, the regularization parameter α is kept constant during the iterative process. The Tikhonov function for the linearized equation takes the form:

$$\mathcal{F}_{lmk}(\mathbf{p}) = \|\bar{\mathbf{y}}_{mk}^\delta - \bar{\mathbf{K}}_{mk}\mathbf{p}\|^2 + \alpha_k \|\mathbf{Lp}\|^2,$$

and its minimizer is given by:

$$\mathbf{p}_{mk}^\delta = \bar{\mathbf{K}}_{mk}^+ \bar{\mathbf{y}}_{mk}^\delta,$$

where:

$$\bar{\mathbf{K}}_{mk}^+ = (\bar{\mathbf{K}}_{mk}^T \bar{\mathbf{K}}_{mk} + \alpha_k \mathbf{L}^T \mathbf{L})^{-1} \bar{\mathbf{K}}_{mk}^T$$

is the regularized generalized inverse at $\mathbf{x}_{mk}^\delta$. The new iterate is computed as:

$$\mathbf{x}_{mk+1}^\delta = \mathbf{x}_a + \bar{\mathbf{K}}_{mk}^+ \bar{\mathbf{y}}_{mk}^\delta.$$

For the iterative regularization methods, the number of iteration steps k plays the role of the regularization parameter, and the iterative process is stopped after an appropriate number of steps k^* in order to avoid an uncontrolled expansion of errors in the data. In fact, a mere minimization of the residual $\|\mathbf{r}_{mk}^\delta\|$, where:

$$\mathbf{r}_{mk}^\delta = \bar{\mathbf{y}}^\delta - \bar{\mathbf{F}}_m(\mathbf{x}_{mk}^\delta)$$

is the residual vector at $\mathbf{x}_{mk}^\delta$, leads to a semi-convergent behavior of the iterated solution: while the error in the residual decreases as the number of iteration steps increases, the error in the solution starts to increase after an initial decay. A widely used a posteriori choice for the stopping index k^* is the discrepancy principle. According to this principle, the iterative process is terminated after k^* steps such that:

$$\|\mathbf{r}_{mk^*}^\delta\|^2 \leq \eta \Delta_m^2 < \|\mathbf{r}_{mk}^\delta\|^2, \quad 0 \leq k < k^*,$$

with $\eta > 1$; hence, the regularized solution is $\mathbf{x}_{mk^*}^\delta$. Since the noise level cannot be estimated a priori for many practical problems arising in atmospheric remote sensing, we adopt a practical approach. This is based on the observation that the squared residual $\|\mathbf{r}_{mk}^\delta\|^2$ decreases during the iterative process and attains a plateau at approximately Δ_m^2 . Thus, if the nonlinear residuals $\|\mathbf{r}_{mk}^\delta\|$ converge to $\|\mathbf{r}_{m\infty}^\delta\|$ within a prescribed tolerance, we use the estimate:

$$\Delta_m^2 = \|\mathbf{r}_{m\infty}^\delta\|^2.$$

The above heuristic stopping rule does not have any mathematical justification, but works sufficiently well in practice.

Since the amount of regularization is gradually decreased during the iterative processes, the iteratively regularized Gauss–Newton method can handle problems that practically do not require much regularization.

The numerical experiments performed in [5] showed that at the solution $\mathbf{x}_{mk^*}^\delta$, (i) α_{k^*-1} is close to the optimal regularization parameter, and (ii) $\mathbf{x}_{mk^*}^\delta$ is close to the Tikhonov solution corresponding to the optimal regularization parameter. In this regard, we make the following assumptions:

1. $\hat{\alpha} = \alpha_{k^*-1}$ is an estimate for the optimal regularization parameter, and
2. $\mathbf{x}_{m\hat{\alpha}}^\delta = \mathbf{x}_{mk^*}^\delta$ is the minimizer of the Tikhonov function with regularization parameter $\hat{\alpha}$, $\mathcal{F}_{m\hat{\alpha}}(\mathbf{x})$.

Some consequences of the second assumption are the following. The optimality condition $\nabla \mathcal{F}_{m\hat{\alpha}}(\mathbf{x}_{m\hat{\alpha}}^\delta) = 0$, yields:

$$\bar{\mathbf{K}}_{m\hat{\alpha}}^T [\bar{\mathbf{F}}_m(\mathbf{x}_{m\hat{\alpha}}^\delta) - \bar{\mathbf{y}}^\delta] + \hat{\alpha} \mathbf{L} \mathbf{L}^T (\mathbf{x}_{m\hat{\alpha}}^\delta - \mathbf{x}_a) = 0$$

and further,

$$\mathbf{x}_{m\hat{\alpha}}^\delta = \mathbf{x}_a + \bar{\mathbf{K}}_{m\hat{\alpha}}^+ \bar{\mathbf{y}}_{m\hat{\alpha}}^\delta,$$

where $\bar{\mathbf{K}}_{m\hat{\alpha}} = \bar{\mathbf{K}}_{mk^*}$,

$$\bar{\mathbf{y}}_{m\hat{\alpha}}^\delta = \bar{\mathbf{y}}_{mk^*}^\delta = \bar{\mathbf{y}}^\delta - \bar{\mathbf{F}}_m(\mathbf{x}_{m\hat{\alpha}}^\delta) + \bar{\mathbf{K}}_{m\hat{\alpha}}(\mathbf{x}_{m\hat{\alpha}}^\delta - \mathbf{x}_a), \quad (21)$$

and:

$$\bar{\mathbf{K}}_{m\hat{\alpha}}^+ = (\bar{\mathbf{K}}_{m\hat{\alpha}}^T \bar{\mathbf{K}}_{m\hat{\alpha}} + \hat{\alpha} \mathbf{L}^T \mathbf{L})^{-1} \bar{\mathbf{K}}_{m\hat{\alpha}}^T. \quad (22)$$

Consequently, we see that:

$$\mathbf{p}_{m\hat{\alpha}}^\delta = \mathbf{x}_{m\hat{\alpha}}^\delta - \mathbf{x}_a = \bar{\mathbf{K}}_{m\hat{\alpha}}^\dagger \bar{\mathbf{y}}_{m\hat{\alpha}}^\delta \quad (23)$$

is the minimizer of the Tikhonov function:

$$\mathcal{F}_{lm\hat{\alpha}}(\mathbf{p}) = \|\bar{\mathbf{y}}_{m\hat{\alpha}}^\delta - \bar{\mathbf{K}}_{m\hat{\alpha}} \mathbf{p}\|^2 + \hat{\alpha} \|\mathbf{L} \mathbf{p}\|^2,$$

and that the residual vector of the linear equation $\bar{\mathbf{y}}_{m\hat{\alpha}}^\delta = \bar{\mathbf{K}}_{m\hat{\alpha}} \mathbf{p}$ is equal to the nonlinear residual vector at the solution, i.e.,

$$\bar{\mathbf{y}}_{m\hat{\alpha}}^\delta - \bar{\mathbf{K}}_{m\hat{\alpha}} \mathbf{p}_{m\hat{\alpha}}^\delta = \bar{\mathbf{y}}^\delta - \bar{\mathbf{F}}_m(\mathbf{x}_{m\hat{\alpha}}^\delta) = \mathbf{r}_{m\hat{\alpha}}^\delta. \quad (24)$$

In summary, in the iteratively regularized Gauss–Newton method:

1. the computation of the regularized solution $\mathbf{x}_{m\hat{\alpha}}^\delta$ depends only on the initial value α_1 and the ratio q of the geometric sequence α_k , which determine the rate of convergence, and
2. the regularization parameter at the solution is an estimate for the optimal regularization parameter, and so, for the ratio of the data error variance and the a priori state variance.

Taking these results into account, we may conclude that the iteratively regularized Gauss–Newton method gives a solution to Problem 1 of the Bayesian parameter estimation.

5. Parameter Estimation and Model Selection

In this section, we address Problems 2 and 3 related to Bayesian model selection.

To solve Problem 3, we suppose that the forward model can be linearized around the solution $\mathbf{x}_{m\hat{\alpha}}^\delta$ [6]. In other words, in the first-order Taylor expansion:

$$\bar{\mathbf{F}}_m(\mathbf{x}) = \bar{\mathbf{F}}_m(\mathbf{x}_{m\hat{\alpha}}^\delta) + \bar{\mathbf{K}}_{m\hat{\alpha}}(\mathbf{x} - \mathbf{x}_{m\hat{\alpha}}^\delta) + \bar{\mathbf{R}}_m(\mathbf{x}, \mathbf{x}_{m\hat{\alpha}}^\delta) \quad (25)$$

the remainder term $\bar{\mathbf{R}}_m(\mathbf{x}, \mathbf{x}_{m\hat{\alpha}}^\delta)$ can be neglected. As a result, the nonlinear data model $\bar{\mathbf{y}}^\delta = \bar{\mathbf{F}}_m(\mathbf{x}) + \bar{\delta}_m$ becomes the linear model:

$$\bar{\mathbf{y}}_{m\hat{\alpha}}^\delta = \bar{\mathbf{K}}_{m\hat{\alpha}} \mathbf{p} + \bar{\delta}_m, \quad (26)$$

where $\bar{\mathbf{y}}_{m\hat{\alpha}}^\delta$ is given by Equation (21), $\mathbf{p} = \mathbf{x} - \mathbf{x}_a \sim \mathcal{N}(\mathbf{0}, \mathbf{C}_x = \sigma_m^2(\hat{\alpha} \mathbf{L}^T \mathbf{L})^{-1})$ and $\bar{\delta}_m \sim \mathcal{N}(\mathbf{0}, \bar{\mathbf{C}}_{\delta m} = \sigma_m^2 \mathbf{I}_M)$. The direct consequences of dealing with the linear model (26) are the following results:

1. the a posteriori density $p(\mathbf{x} | \bar{\mathbf{y}}^\delta, m)$ can be expressed as a Gaussian distribution:

$$p(\mathbf{x} | \bar{\mathbf{y}}^\delta, m) \propto \exp \left[-\frac{1}{2} (\mathbf{x} - \mathbf{x}_{m\hat{\alpha}}^\delta)^T \hat{\mathbf{C}}_{xm}^{-1} (\mathbf{x} - \mathbf{x}_{m\hat{\alpha}}^\delta) \right],$$

where:

$$\hat{\mathbf{C}}_{xm} = \sigma_m^2 (\bar{\mathbf{K}}_{m\hat{\alpha}}^T \bar{\mathbf{K}}_{m\hat{\alpha}} + \hat{\alpha} \mathbf{L}^T \mathbf{L})^{-1}. \quad (27)$$

is the a posteriori covariance matrix, and

2. as shown in Appendix A, the marginal likelihood density $p(\bar{\mathbf{y}}^\delta | m)$ can be computed analytically; the result is:

$$\begin{aligned} p(\bar{\mathbf{y}}^\delta | m) &= \sqrt{\frac{\det(\mathbf{I}_M - \hat{\mathbf{A}}_{m\hat{\alpha}})}{(2\pi\sigma_m^2)^M}} \\ &\times \exp \left[-\frac{1}{2\sigma_m^2} \bar{\mathbf{y}}_{m\hat{\alpha}}^{\delta T} (\mathbf{I}_M - \hat{\mathbf{A}}_{m\hat{\alpha}}) \bar{\mathbf{y}}_{m\hat{\alpha}}^\delta \right], \end{aligned} \quad (28)$$

where $\hat{\mathbf{A}}_{m\hat{\alpha}} = \bar{\mathbf{K}}_{m\hat{\alpha}} \bar{\mathbf{K}}_{m\hat{\alpha}}^\dagger$ is the influence matrix at the solution $\mathbf{x}_{m\hat{\alpha}}^\delta$.

The replacement of the nonlinear data model by a linear data model will also enable us to deal with Problem 2. More specifically, for this purpose, we will adapt some regularization parameter choice

methods for linear problems, i.e., maximum marginal likelihood estimation [7–9] and generalized cross-validation [10,11], to model selection. Furthermore, estimates for the data error variance σ_m^2 delivered by these methods will be used to compute the marginal likelihood density and, therefore, the relative evidence.

In Appendix B, it is shown that:

1. in the framework of maximum marginal likelihood estimation, the data error variance can be estimated by:

$$\hat{\sigma}_{Lm}^2 = \frac{1}{M} \bar{\mathbf{y}}_{m\hat{\alpha}}^{\delta T} (\mathbf{I}_M - \hat{\mathbf{A}}_{m\hat{\alpha}}) \bar{\mathbf{y}}_{m\hat{\alpha}}^{\delta}, \quad (29)$$

and the model with the smallest value of the marginal likelihood function, defined by:

$$\lambda(m) = \frac{\bar{\mathbf{y}}_{m\hat{\alpha}}^{\delta T} (\mathbf{I}_M - \hat{\mathbf{A}}_{m\hat{\alpha}}) \bar{\mathbf{y}}_{m\hat{\alpha}}^{\delta}}{\sqrt[M]{\det(\mathbf{I}_M - \hat{\mathbf{A}}_{m\hat{\alpha}})}}, \quad (30)$$

is optimal.

2. in the framework of generalized cross-validation, the data error variance can be estimated by:

$$\hat{\sigma}_{Gm}^2 = \frac{\|\mathbf{r}_{m\hat{\alpha}}^{\delta}\|^2}{[\text{trace}(\mathbf{I}_M - \hat{\mathbf{A}}_{m\hat{\alpha}})]}, \quad (31)$$

and the model with the smallest value of the generalized cross-validation function, defined by:

$$v(m) = \frac{\|\mathbf{r}_{m\hat{\alpha}}^{\delta}\|^2}{[\text{trace}(\mathbf{I} - \hat{\mathbf{A}}_{m\hat{\alpha}})]^2}, \quad (32)$$

is optimal.

Denoting by $p_L(\bar{\mathbf{y}}^{\delta} | m)$ and $p_G(\bar{\mathbf{y}}^{\delta} | m)$ the marginal likelihoods corresponding to $\hat{\sigma}_{Lm}^2$ and $\hat{\sigma}_{Gm}^2$, respectively, we define:

1. the relative evidence of an approach based on marginal likelihood and the computation of the data error variance in the framework of the Maximum Marginal Likelihood Estimation (MLMLE) by:

$$p_{\text{MLMLE}}(m | \mathbf{y}^{\delta}) = \frac{p_L(\bar{\mathbf{y}}^{\delta} | m)}{\sum_{m=1}^{N_m} p_L(\bar{\mathbf{y}}^{\delta} | m)}, \quad (33)$$

2. the relative evidence of an approach based on the marginal likelihood and computation of the data error variance in the framework of the Generalized Cross-Validation (MLGCV) by:

$$p_{\text{MLGCV}}(m | \mathbf{y}^{\delta}) = \frac{p_G(\bar{\mathbf{y}}_m^{\delta} | m)}{\sum_{m=1}^{N_m} p_G(\bar{\mathbf{y}}_m^{\delta} | m)}. \quad (34)$$

Note that for the data error variance estimate (29), the marginal likelihood density becomes:

$$p_L(\bar{\mathbf{y}}^{\delta} | m) = \frac{c_M}{\sqrt{[\lambda(m)]^M}},$$

with:

$$c_M = \sqrt{\left(\frac{M}{2\pi}\right)^M} \exp\left(-\frac{M}{2}\right).$$

On the other hand, the statements (30) and (32) are equivalent to (cf. Equation (15)):

$$\hat{\mathbf{x}}_{\max}^{\delta} = \mathbf{x}_{m^* \hat{\alpha}}^{\delta}, \quad m^* = \arg \max_m \frac{1}{\lambda(m)}, \quad (35)$$

and:

$$\hat{\mathbf{x}}_{\max}^{\delta} = \mathbf{x}_{m^* \hat{\alpha}}^{\delta}, \quad m^* = \arg \max_m \frac{1}{v(m)}, \quad (36)$$

respectively. Deviating from a stochastic interpretation of the relative evidence and regarding this quantity in a deterministic setting merely as a merit function characterizing the solution $\mathbf{x}_{m\hat{\alpha}}^{\delta}$, we define:

1. the relative evidence of an approach based on Maximum Marginal Likelihood Estimation (MMLE) by:

$$p_{\text{MMLE}}(m | \mathbf{y}^{\delta}) = \frac{1/\lambda(m)}{\sum_{m=1}^{N_m} 1/\lambda(m)}, \quad (37)$$

2. the relative evidence of an approach based on Generalized Cross-Validation (GCV) by:

$$p_{\text{GCV}}(m | \mathbf{y}^{\delta}) = \frac{1/v(m)}{\sum_{m=1}^{N_m} 1/v(m)}. \quad (38)$$

Note that $p_{\text{MMLE}}(m | \mathbf{y}^{\delta})$ and $p_{\text{GCV}}(m | \mathbf{y}^{\delta})$ do not depend on the data error variance σ_m^2 . Note also that $\hat{\sigma}_{Gm}^2$ is defined in terms of the square residual:

$$\|\mathbf{r}_{m\hat{\alpha}}^{\delta}\|^2 = \|\bar{\mathbf{y}}_{m\hat{\alpha}}^{\delta} - \bar{\mathbf{K}}_{m\hat{\alpha}} \mathbf{P}_{m\hat{\alpha}}^{\delta}\|^2 = \|(\mathbf{I}_M - \hat{\mathbf{A}}_{m\hat{\alpha}}) \bar{\mathbf{y}}_{m\hat{\alpha}}^{\delta}\|^2,$$

while, according to Equation (29), $\hat{\sigma}_{Lm}^2$ is defined in terms of the quantity $\bar{\mathbf{y}}_{m\hat{\alpha}}^{\delta T} (\mathbf{I}_M - \hat{\mathbf{A}}_{m\hat{\alpha}}) \bar{\mathbf{y}}_{m\hat{\alpha}}^{\delta}$. The same difference exists between the numerators of the cross-validation function $v(m)$ and the marginal likelihood function $\lambda(m)$ (see Equations (32) and (30)).

6. Algorithm Description

An algorithmic implementation of the iteratively regularized Gauss–Newton method in connection with model selection is as follows:

1. compute the scaling matrix $\mathbf{P}$ by means of Equation (5) and the scaled data vector $\bar{\mathbf{y}}^{\delta} = \mathbf{P}\mathbf{y}^{\delta}$;
2. given the current iterate $\mathbf{x}_{mk}^{\delta}$ at step k , compute the forward model $\mathbf{F}_m(\mathbf{x}_{mk}^{\delta})$, the Jacobian matrix $\mathbf{K}_{mk}$, and the scaled quantities $\bar{\mathbf{F}}_m(\mathbf{x}_{mk}^{\delta}) = \mathbf{P}\mathbf{F}_m(\mathbf{x}_{mk}^{\delta})$ and $\bar{\mathbf{K}}_{mk} = \mathbf{P}\mathbf{K}_{mk}$;
3. compute the linearized data vector:

$$\bar{\mathbf{y}}_{mk}^{\delta} = \bar{\mathbf{y}}^{\delta} - \bar{\mathbf{F}}_m(\mathbf{x}_{mk}^{\delta}) + \bar{\mathbf{K}}_{mk}(\mathbf{x}_{mk}^{\delta} - \mathbf{x}_a);$$

4. compute the singular value decomposition of the quotient matrix $\bar{\mathbf{K}}_{mk} \mathbf{L}^{-1} = \mathbf{U}\mathbf{\Gamma}\mathbf{V}^T$, where $\mathbf{\Gamma} = \text{diag}(\gamma_{mi})_M$ with $\gamma_{mi} = 0$ for $i > N$ is a diagonal matrix containing the singular values γ_{mi} in decreasing order and $\mathbf{U} = [\mathbf{u}_1, \dots, \mathbf{u}_M] \in \mathbb{R}^{M \times M}$ and $\mathbf{V} = [\mathbf{v}_1, \dots, \mathbf{v}_N] \in \mathbb{R}^{N \times N}$ are orthogonal matrices containing the left and right singular column vectors $\mathbf{u}_i$ and $\mathbf{v}_i$, respectively;
5. if $k = 1$, choose $\alpha_k = \max(\sqrt{\gamma_{m1}\gamma_{mN}}, \alpha_{\min})$, where γ_{m1} and γ_{mN} are the largest and the smallest singular values, respectively; otherwise, set $\alpha_k = \max(q\alpha_{k-1}, \alpha_{\min})$;
6. compute the minimizer of the Tikhonov function for the linearized equation,

$$\mathbf{p}_{mk}^{\delta} = \sum_{i=1}^n \frac{\gamma_{mi}}{\gamma_{mi}^2 + \alpha_k} (\mathbf{u}_i^T \bar{\mathbf{y}}_{mk}^{\delta}) \mathbf{v}_i,$$

7. compute the nonlinear residual vector at $\mathbf{x}_{mk+1}^{\delta}$,

$$\mathbf{r}_{mk+1}^{\delta} = \bar{\mathbf{y}}^{\delta} - \bar{\mathbf{F}}_m(\mathbf{x}_{mk+1}^{\delta}),$$

and the residual $r_{k+1} = \|\mathbf{r}_{mk+1}^{\delta}\|^2$;

8. compute the condition number $c_k = \gamma_{m1}/\gamma_{mN}$, the scalar quantities:

$$\begin{aligned}\bar{y}_k &= \sum_{i=1}^N \frac{\alpha_{k-1}}{\gamma_{mi}^2 + \alpha_{k-1}} (\mathbf{u}_i^T \bar{\mathbf{y}}_{mk}^\delta)^2 + \sum_{i=N+1}^M (\mathbf{u}_i^T \bar{\mathbf{y}}_{mk}^\delta)^2, \\ t_k &= M - N + \sum_{i=1}^N \frac{\alpha_{k-1}}{\gamma_{mi}^2 + \alpha_{k-1}}, \\ d_k &= \prod_{i=1}^N \frac{\alpha_{k-1}}{\gamma_{mi}^2 + \alpha_{k-1}},\end{aligned}$$

and the normalized covariance matrix:

$$\hat{\mathbf{C}}_{xmk} = \hat{\mathbf{V}} \text{diag}\left(\frac{1}{\gamma_{mi}^2 + \alpha_{k-1}}\right)_N \hat{\mathbf{V}}^T,$$

where $\hat{\mathbf{V}} = \mathbf{L}^{-1} \mathbf{V}$;

9. if $r_{k+1} \geq r_k$, recompute $\mathbf{x}_{mk+1}^\delta$ by means of a step-length algorithm such that $r_{k+1} < r_k$; if the residual cannot be reduced, set $r_\infty = r_k$. and go to Step 12;
10. compute the relative decrease in the residual $\triangle_r = (r_k - r_{k+1})/r_k$;
11. if $\triangle_r > \varepsilon_R$, go to Step 1; otherwise set $r_\infty = r_{k+1}$, and go to Step 12;
12. determine k^* such that $r_{k^*} \leq \eta r_\infty < r_{k^*+1}$ for all $0 \leq k < k^*$;
13. since $r_{k^*} = \|\mathbf{r}_{m\hat{\alpha}}^\delta\|^2$, $\bar{y}_{k^*} = \bar{\mathbf{y}}_{m\hat{\alpha}}^{\delta T} (\mathbf{I}_M - \hat{\mathbf{A}}_{m\hat{\alpha}}) \bar{\mathbf{y}}_{m\hat{\alpha}}^\delta$, $t_{k^*} = \text{trace}(\mathbf{I} - \hat{\mathbf{A}}_{m\hat{\alpha}})$, and $d_{k^*} = \det(\mathbf{I}_M - \hat{\mathbf{A}}_{m\hat{\alpha}})$, compute the estimates:

$$\begin{aligned}\lambda(m) &= \frac{\bar{y}_{k^*}}{\sqrt[M]{d_{k^*}}}, \quad v(m) = \frac{r_{k^*}}{t_{k^*}^2}, \\ \hat{\sigma}_{Lm}^2 &= \frac{1}{M} \bar{y}_{k^*}, \quad \hat{\sigma}_{Gm}^2 = \frac{r_{k^*}}{t_{k^*}^2},\end{aligned}$$

the marginal likelihood:

$$p_X(\bar{\mathbf{y}}^\delta | m) = \sqrt{\frac{d_{k^*}}{(2\pi\hat{\sigma}_{Ym}^2)^M}} \exp\left(-\frac{\bar{y}_{k^*}}{2\hat{\sigma}_{Ym}^2}\right),$$

where X stands for the character strings MLMMLE and MLGCV when the character variable Y takes the values L and G, respectively, and the covariance matrix,

$$\hat{\mathbf{C}}_{xm} = \hat{\sigma}_{Ym}^2 \hat{\mathbf{C}}_{xmk^*};$$

14. compute the relative evidence $p_X(m | \mathbf{y}^\delta)$, where X stands for the character strings MLMMLE, MLGCV, MMLE, and GCV by using Equations (33), (34), (37), and (38), respectively, for those situations;
15. compute the maximum and mean solution estimates $\hat{\mathbf{x}}_{\max}^\delta$ and $\hat{\mathbf{x}}_{\text{mean}}^\delta$ by using Equations (15) and (17), respectively, and the mean a posteriori density $p_{\text{mean}}(\mathbf{x} | \mathbf{y}^\delta)$ by using Equation (16).

The control parameters of the algorithm are: (i) the ratio η of the geometric sequence of regularization parameters, (ii) the minimum acceptable value of the regularization parameter $\alpha_{\min}$, (iii) the tolerance ε_R of the residual test convergence, and (iv) the tolerance η of the discrepancy principle stopping rule.

We conclude our theoretical analysis with some comments:

1. Another estimate for the data error variance is derived in Appendix C; this is given by:

$$\hat{\sigma}_{Rm}^2 = \frac{1}{M - N} \|\mathbf{r}_{m\hat{\alpha}}^\delta\|^2. \quad (39)$$

If $\hat{\alpha} \ll \gamma_{mi}^2$ for all $i = 1, \dots, N$, we can approximate (cf. Equation (A17)):

$$\text{trace}(\mathbf{I} - \hat{\mathbf{A}}_{m\hat{\alpha}}) = M - N + \sum_{i=1}^M \frac{\hat{\alpha}}{\gamma_{mi}^2 + \hat{\alpha}}, \approx M - N,$$

and in view of Equation (31), we deduce that $\hat{\sigma}_{Rm}^2 \approx \hat{\sigma}_{Gm}^2$.

2. If a model is far from reality, it is natural to assume that the model parameter errors are large or, equivalently, that the data error variance is large. This observation suggests that, in a deterministic setting, we may define the relative evidence as:

$$p(m | \mathbf{y}^\delta) = \frac{1/\hat{\sigma}_{Ym}^2}{\sum_{m=1}^{N_m} 1/\hat{\sigma}_{Ym}^2}, \quad (40)$$

where the character variable Y takes the values R, L, or G. In this case, for $\hat{\sigma}_{Rm}^2 = \|\mathbf{r}_{m\hat{\alpha}}^\delta\|^2 / (M - N)$, the model with the smallest value of the squared residual is considered to be optimal.

7. Application to the EPIC Instrument

The Deep Space Climate Observatory (DSCOVR) flies in a Lissajous orbit about the first Earth-Sun Lagrange point (L_1), which is 1.5 million kilometers from the Earth towards the Sun. The Earth Polychromatic Imaging Camera (EPIC) is one of the two Earth observing instruments on board DSCOVR. This instrument views the entire sunlit side of the Earth from sunrise to sunset in the backscattering direction (scattering angles between 168.5° and 175.5°) with 10 narrowband filters, ranging from 317.5 to 779.5 nm [12]. This unique Earth observing geometry of EPIC compared to other instruments in Sun-synchronous orbits, that rarely view Earth at such large scattering angles [13], makes it a suitable candidate for several climate science applications including the measurement of cloud reflectivity and aerosol optical thickness and layer height [14].

We apply the algorithm proposed in Section 6 to the retrieval of aerosol optical thickness and aerosol layer height by generating synthetic measurements corresponding to the EPIC instrument. Specifically, Channels 7 and 8 in the Oxygen B-band at 680 and 687.75 nm, respectively, and Channels 9 and 10 in the Oxygen A-band at 764 and 779.5 nm, respectively, are used in the retrieval. The linearized radiative transfer model is based on the discrete ordinates method with matrix exponential and uses the correlated k -distribution method in conjunction with the principal component analysis technique to speed up the computations [15,16].

We consider the aerosol models implemented in the Moderate Resolution Imaging Spectroradiometer (MODIS) aerosol retrieval algorithm over land [1]. There are three spherical, fine-dominated model types (non-absorbing, moderately absorbing, and absorbing) and one spheroid, coarse-dominated model type (dust). These aerosol models, which depend on location and season, are the result of a cluster analysis of the climatology of almucantar retrievals from global AERONET (AERosol RObotic NETwork) measurements [17]. Each model consists of two log-normal modes (accumulated and coarse). A single log-normal mode is described by the number size distribution:

$$\frac{dN(r)}{d \ln r} = \frac{N_0}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(\ln r - \ln r_{\text{mod}})^2}{2\sigma^2}\right],$$

where r_{mod} is the modal or median radius of the number size distribution, σ is the standard deviation, and:

$$N_0 = \int_0^\infty \frac{dN(r)}{d \ln r} d \ln r$$

is the total number of particles per cross-section of the atmospheric column. Table 1 displays the log-normal size parameters and the complex refractive indices m for the four aerosol models, where, for each log-normal mode,

$$r_v = r_{\text{mod}} \exp(-3\sigma^2)$$

is the median radius of the volume size distribution:

$$\frac{dV(r)}{d \ln r} = \frac{V_0}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(\ln r - \ln r_v)^2}{2\sigma^2}\right],$$

and:

$$V_0 = \int_0^\infty \frac{4\pi r^3}{3} \frac{dN(r)}{d \ln r} d \ln r = N_0 \frac{4\pi r_{\text{mod}}^3}{3} \exp(4.5\sigma^2)$$

is the volume of particles per cross-section of the atmospheric column.

Table 1. Optical properties of the aerosol models. For each model, the first and second lines are related to the accumulated and coarse modes, respectively. The complex refractive index of dust corresponds to the wavelengths 470 (1), 550 (2), 660 (3), and 2100 (4) nm. In the simulations, the values of the refractive index at $\lambda = 660$ nm are assumed.

Model	r_v (μm)	σ	$m = (\text{Re}(m), \text{Im}(m))$	V_0 ($\mu\text{m}^3/\mu\text{m}^2$)
1 Nonabs.	$0.160 + 0.0434\tau$	$0.364 + 0.1529\tau$	$(1.42, 0.004 - 0.0015\tau)$	$0.1718\tau^{0.821}$
	$3.325 + 0.1411\tau$	$0.759 + 0.0168\tau$	$(1.42, 0.004 - 0.0015\tau)$	$0.0934\tau^{0.639}$
2 Modabs.	$0.145 + 0.0203\tau$	$0.374 + 0.1365\tau$	$(1.43, 0.008 - 0.002\tau)$	$0.1642\tau^{0.775}$
	$3.101 + 0.3364\tau$	$0.729 + 0.098\tau$	$(1.43, 0.008 - 0.002\tau)$	$0.1482\tau^{0.684}$
3 Abs.	$0.134 + 0.0096\tau$	$0.383 + 0.0794\tau$	$(1.51, 0.02)$	$0.1748\tau^{0.891}$
	$3.448 + 0.9489\tau$	$0.743 + 0.0409\tau$	$(1.51, 0.02)$	$0.1043\tau^{0.682}$
4 Dust	$0.1416\tau^{-0.052}$	$0.7561\tau^{0.148}$	$(1.48\tau^{-0.021}, 0.0025\tau^{0.132})$ (1)	$0.0871\tau^{1.026}$
			$(1.48\tau^{-0.021}, 0.002)$ (2)	
			$(1.48\tau^{-0.021}, 0.0018\tau^{-0.08})$ (3)	
			$(1.46\tau^{-0.040}, 0.0018\tau^{-0.30})$ (4)	
	2.2	$0.554\tau^{-0.052}$	$(1.48\tau^{-0.021}, 0.0025\tau^{0.132})$ (1)	$0.6786\tau^{1.057}$
			$(1.48\tau^{-0.021}, 0.002)$ (2)	
			$(1.48\tau^{-0.021}, 0.0018\tau^{-0.08})$ (3)	
			$(1.46\tau^{-0.040}, 0.0018\tau^{-0.30})$ (4)	

In our numerical analysis, the state vector is $\mathbf{x} = [\tau, H]^T$, where τ is the aerosol optical thickness and H the aerosol layer height. The true aerosol optical thicknesses to be retrieved are $\tau_t = 0.25, 0.50, 0.75, 1.0, 1.25$, and 1.5 , while for the true aerosol layer height, we take $H_t = 1.0, 1.5, 2.0, 2.5$, and 3.0 km. The a priori values, which coincide with the initial guesses, are $\tau_a = 2.0$ and $H_a = 4$ km. A Lambertian surface is assumed, and if not stated otherwise, the surface albedo is $A = 0.06$. We generate synthetic measurements by choosing the moderately absorbing aerosol model (Model 2) as the true or exact model. Specifically, we:

1. compute the radiances for the moderately absorbing aerosol model and the true values τ_t and H_t , $I_{\text{modabs}}(\lambda_i, \tau_t, H_t)$;
2. add the measurement noise δ_{Imes} ,

$$I_{\text{modabs}}^\delta(\lambda_i, \tau_t, H_t) = I_{\text{modabs}}(\lambda_i, \tau_t, H_t) + [\delta_{\text{Imes}}]_i,$$

where $\delta_{\text{Imes}} \sim \mathcal{N}(\mathbf{0}, \mathbf{C}_{\text{Imes}})$, $\mathbf{C}_{\text{Imes}} = \text{diag}(\sigma_{\text{Imesi}}^2)_M$, and:

$$\sigma_{\text{Imesi}} = \frac{I_{\text{modabs}}(\lambda_i, \tau_t, H_t)}{\text{SNR}} \text{ with SNR} = 290,$$

- which is the maximum signal-to-noise ratio according to the EPIC camera specifications and
3. in view of the approximation:

$$\begin{aligned} & \ln I_{\text{modabs}}^\delta(\lambda_i, \tau_t, H_t) \\ & \approx \ln I_{\text{modabs}}(\lambda_i, \tau_t, H_t) + \frac{[\delta_{\text{Imes}}]_i}{I_{\text{modabs}}(\lambda_i, \tau_t, H_t)}, \end{aligned}$$

identify $[\mathbf{y}^\delta]_i = \ln I_{\text{modabs}}^\delta(\lambda_i, \tau_t, H_t)$, $[\mathbf{y}]_i = \ln I_{\text{modabs}}(\lambda_i, \tau_t, H_t)$, and:

$$[\delta_{\text{mes}}]_i = \frac{[\delta_{\text{Imes}}]_i}{I_{\text{modabs}}(\lambda_i, \tau_t, H_t)},$$

implying:

$$\sigma_{\text{mesi}} = \frac{\sigma_{\text{Imesi}}}{I_{\text{modabs}}(\lambda_i, \tau_t, H_t)} = \frac{1}{\text{SNR}}.$$

The above scheme yields $\mathbf{C}_{\delta m} = \mathbf{I}_M$ and $\mathbf{P} = \mathbf{I}_M$, showing that the regularized solution does not depend on the scaling matrix. Coming to the regularization matrix, we first note that the choice $\sigma_{xi} = \varepsilon_x [\mathbf{x}_a]_i$ for some ε_x gives:

$$\mathbf{L} = \text{diag}\left(\frac{\sigma_x}{\sigma_{xi}}\right)_N = \text{diag}\left(\frac{\sqrt{\sum_{i=1}^N [\mathbf{x}_a]_i^2}}{N[\mathbf{x}_a]_i}\right)_N.$$

However, to have a better control on the amount of regularization that is applied to each component of the state vector, we introduce the weight w_{xi} of component $[\mathbf{x}]_i$, and take:

$$\mathbf{L} = \text{diag}\left(w_{xi} \frac{\sqrt{\sum_{i=1}^N [\mathbf{x}_a]_i^2}}{N[\mathbf{x}_a]_i}\right)_N.$$

If $w_{xi} \rightarrow \infty$, the component $[\mathbf{x}]_i$ is close to the a priori, while for $w_{xi} \rightarrow 0$, the component $[\mathbf{x}]_i$ is practically unconstrained. In our simulations, we used $w_\tau = w_H = 1.0$. What is left is the specification of the control parameters of the algorithm; these are $q = 0.1$, $\alpha_{\min} = 10^{-6} \gamma_{mN}$, where γ_{mN} is the smallest singular value of the quotient matrix $\bar{\mathbf{K}}_{mk} \mathbf{L}^{-1}$ at $k = 1$ (the first iteration step), $\varepsilon_R = 10^{-3}$, and $\eta = 1.05$.

To analyze the accuracy of the aerosol retrieval we consider several test examples.

Test Example 1

In the first example, we analyze the efficiency of the aerosol model selection by considering all models, i.e., we take $m = 1, 2, 3, 4$. Thus, the algorithm has the chance to identify the correct aerosol model. The solution accuracy is characterized by the relative errors:

$$\varepsilon_{\text{mean}}^\tau = \frac{|\tau_{\text{mean}} - \tau_t|}{\tau_t} \text{ and } \varepsilon_{\text{mean}}^H = \frac{|H_{\text{mean}} - H_t|}{\tau_t}$$

corresponding to (cf. Equation (17)) $\hat{\mathbf{x}}_{\text{mean}}^\delta = [\tau_{\text{mean}}, H_{\text{mean}}]$, and:

$$\varepsilon_{\text{max}}^\tau = \frac{|\tau_{\text{max}} - \tau_t|}{\tau_t} \text{ and } \varepsilon_{\text{max}}^H = \frac{|H_{\text{max}} - H_t|}{\tau_t}$$

corresponding to (cf. Equation (15)) $\hat{\mathbf{x}}_{\max}^{\delta} = [\tau_{\max}, H_{\max}]$. In Figures 1 and 2, we illustrate the variations of the relative errors with respect to τ_t and H_t , respectively, while in Table 2, we show the average relative errors:

$$\varepsilon_{\text{mean}|\tau}^{\tau} = \frac{1}{N_{\tau}} \sum_{i=1}^{N_{\tau}} \varepsilon_{\text{mean}i}^{\tau} \quad \text{and} \quad \varepsilon_{\text{mean}|\tau}^H = \frac{1}{N_{\tau}} \sum_{i=1}^{N_{\tau}} \varepsilon_{\text{mean}i}^H$$

over τ_t for $N_{\tau} = 6$, and:

$$\varepsilon_{\text{mean}|H}^{\tau} = \frac{1}{N_H} \sum_{i=1}^{N_H} \varepsilon_{\text{mean}i}^{\tau} \quad \text{and} \quad \varepsilon_{\text{mean}|H}^H = \frac{1}{N_H} \sum_{i=1}^{N_H} \varepsilon_{\text{mean}i}^H$$

over H_t for $N_H = 5$. The following conclusions can be drawn:

1. the relative errors $\varepsilon_{\text{mean}}^{\tau}$ and $\varepsilon_{\text{mean}}^H$ corresponding to the generalized cross-validation (GCV and MLGCV) are in general smaller than those corresponding to the maximum marginal likelihood estimation (MMLE and MLMMLE).
2. the relative errors $\varepsilon_{\max}^{\tau}$ and $\varepsilon_{\max}^H$ are very small, and so, we deduce that the algorithm recognizes the exact aerosol model;
3. the best method is GCV, in which case the average relative errors are $\varepsilon_{\text{mean}|\tau}^{\tau} = 0.009$, $\varepsilon_{\text{mean}|\tau}^H = 0.020$, $\varepsilon_{\text{mean}|H}^{\tau} = 0.001$, and $\varepsilon_{\text{mean}|H}^H = 0.024$;
4. the aerosol optical thickness is better retrieved than the aerosol layer height.

Figure 1. Relative errors $\varepsilon_{\text{mean}}^{\tau}$, $\varepsilon_{\max}^{\tau}$, $\varepsilon_{\text{mean}}^H$, and $\varepsilon_{\max}^H$ versus τ_t for $H_t = 3.0$ km. All four aerosol models are involved in the retrieval.

Table 2. Average relative errors for the results plotted in Figures 1 and 2. MMLE, Maximum Marginal Likelihood Estimation; GCV, Generalized-Cross Validation.

Method	Average Relative Error			
	$\varepsilon_{\text{mean} \tau}^{\tau}$	$\varepsilon_{\text{mean} \tau}^H$	$\varepsilon_{\text{mean} H}^{\tau}$	$\varepsilon_{\text{mean} H}^H$
MMLE	0.046	0.091	0.027	0.102
MLMMLE	0.009	0.062	0.022	0.059
GCV	0.009	0.020	0.001	0.024
MLGCV	0.009	0.025	0.006	0.042

Figure 2. Relative errors $\epsilon_{\text{mean}}^{\tau}$, $\epsilon_{\text{max}}^{\tau}$, ϵ_{mean}^H , and ϵ_{max}^H versus H_t for $\tau_t = 1.0$. All four aerosol models are involved in the retrieval.

Test Example 2

In the second example, we consider all aerosol models m except the exact one, i.e., we take $m = 1, 3, 4$. In this more realistic scenario, the algorithm tries to find an aerosol model that is as close as possible to the exact one. The variations of the relative errors with respect to τ_t and H_t are illustrated in Figures 3 and 4, respectively, while the average relative errors are given in Table 3. The following conclusions can be drawn:

1. the relative errors $\epsilon_{\text{mean}}^{\tau}$ and ϵ_{mean}^H corresponding to the generalized cross-validation (GCV and MLGCV) are still smaller than those corresponding to the maximum marginal likelihood estimation (MMLE and MLMMLE).
2. the relative errors $\epsilon_{\text{max}}^{\tau}$ and ϵ_{max}^H are extremely large, and so, we infer that the maximum solution estimate $\hat{\mathbf{x}}_{\text{max}}^{\delta} = [\tau_{\text{max}}, H_{\text{max}}]$ is completely unrealistic;
3. as before, the best method is GCV characterized by the average relative errors $\epsilon_{\text{mean}|\tau}^{\tau} = 0.060$, $\epsilon_{\text{mean}|\tau}^H = 0.111$, $\epsilon_{\text{mean}|H}^{\tau} = 0.022$, and $\epsilon_{\text{mean}|H}^H = 0.096$;
4. the relative errors are significantly larger than those corresponding to the case when all four aerosol models are taken into account.

Table 3. Average relative errors for the results plotted in Figures 3 and 4.

Method	Average Relative Error			
	$\epsilon_{\text{mean}\tau}^{\tau}$	$\epsilon_{\text{mean}\tau}^H$	$\epsilon_{\text{mean}H}^{\tau}$	$\epsilon_{\text{mean}H}^H$
MMLE	0.095	0.206	0.073	0.318
MLMMLE	0.180	0.210	0.139	0.348
GCV	0.060	0.111	0.022	0.096
MLGCV	0.059	0.200	0.076	0.259

Figure 3. Relative errors $\epsilon_{\text{mean}}^{\tau}$, $\epsilon_{\text{max}}^{\tau}$, ϵ_{mean}^H , and ϵ_{max}^H versus τ_t for $H_t = 3.0$ km. All aerosol models except the exact one are involved in the retrieval.

Figure 4. Relative errors $\epsilon_{\text{mean}}^{\tau}$, $\epsilon_{\text{max}}^{\tau}$, ϵ_{mean}^H , and ϵ_{max}^H versus H_t for $\tau_t = 1.0$. All aerosol models except the exact one are involved in the retrieval.

In Figure 5, we show the mean a posteriori density of τ and H computed by GCV for Test Examples 1 and 2. In Test Example 1, the a posteriori density is sharply peaked, indicating small errors in the retrieval, while in Test Example 2, the a posteriori density is wide and spread over all the aerosol models.

Figure 5. Upper panels: mean a posteriori density of τ computed by GCV for Test Examples 1 (left) and 2 (right). Lower panels: mean a posteriori density of H computed by GCV for Test Examples 1 (left) and 2 (right). The true values are $\tau_t = 1.5$ and $H_t = 3.0$ km.

Test Example 3

In the third series of our numerical experiments, we include the surface albedo in the retrieval. In fact, the surface albedo regarded as a model parameter can be (i) assumed to be known, (ii) included in the retrieval, or (iii) treated as a model uncertainty. The second situation, which leads to more accurate results, is considered here. In this case, however, the aerosol optical thickness and the surface albedo are strongly correlated (the condition number of $\bar{\mathbf{K}}_{mk}\mathbf{L}^{-1}$ at $k = 1$ is 10^3 – 10^4 times larger than for the case in which the surface albedo is not part of the retrieval). Since we are not interested in an accurate surface albedo retrieval, we chose the weight controlling the constraint of the surface albedo in the regularization matrix as $w_A = 10^3$. As $w_\tau = w_H = 1.0$, this means that the surface albedo is tightly constrained to the a priori. We use the a priori and true values $A_a = 0.06$ and $A_t = 0.063$, respectively; thus, the uncertainty in the surface albedo with respect to the a priori is 5%. The synthetic measurements are generated with the true surface albedo of 0.063.

In Figure 6, we illustrate the variations of the relative errors $\epsilon_{\text{mean}}^\tau$ and ϵ_{mean}^H with respect to τ_t and H_t , when all four aerosol models are taken into account. The average relative errors are given in Table 4. The results show that:

1. the relative errors $\epsilon_{\text{mean}}^\tau$ and ϵ_{mean}^H corresponding to generalized cross-validation (GCV and MLGCV) are in general larger than those corresponding to maximum marginal likelihood estimation (MMLE and MLMMLE);

- the best methods are MMLE and MLMMLE; the average relative errors given by MLMMLE are $\varepsilon_{\text{mean}|\tau}^{\tau} = 0.051$, $\varepsilon_{\text{mean}|\tau}^H = 0.060$, $\varepsilon_{\text{mean}|H}^{\tau} = 0.011$, and $\varepsilon_{\text{mean}|H}^H = 0.027$;
- the relative errors are significantly larger than those obtained in the first two test examples (when the surface albedo is known exactly).

Figure 7 illustrates the variations of the relative errors $\varepsilon_{\text{mean}}^{\tau}$ and $\varepsilon_{\text{mean}}^H$ with respect to τ_t and H_t , when all aerosol models except the exact one are considered. The corresponding average relative errors are given in Table 5. The results show that:

- as in the previous scenario, the relative errors $\varepsilon_{\text{mean}}^{\tau}$ and $\varepsilon_{\text{mean}}^H$ corresponding to generalized cross-validation (GCV and MLGCV) are in general larger than those corresponding to maximum marginal likelihood estimation (MMLE and MLMMLE).
- the best methods are MMLE and MLMMLE; the average relative errors delivered by MMLE are $\varepsilon_{\text{mean}|\tau}^{\tau} = 0.101$, $\varepsilon_{\text{mean}|\tau}^H = 0.113$, $\varepsilon_{\text{mean}|H}^{\tau} = 0.070$, and $\varepsilon_{\text{mean}|H}^H = 0.204$;
- the relative errors are the largest among all the test examples.

Table 4. Average relative errors for the results plotted in Figure 6.

Method	Average Relative Error			
	$\varepsilon_{\text{mean}\tau}^{\tau}$	$\varepsilon_{\text{mean}\tau}^H$	$\varepsilon_{\text{mean}H}^{\tau}$	$\varepsilon_{\text{mean}H}^H$
MMLE	0.042	0.046	0.018	0.042
MLMMLE	0.051	0.060	0.011	0.027
GCV	0.101	0.213	0.076	0.237
MLGCV	0.119	0.311	0.100	0.398

Table 5. Average relative errors for the results plotted in Figure 7.

Method	Average Relative Error			
	$\varepsilon_{\text{mean}\tau}^{\tau}$	$\varepsilon_{\text{mean}\tau}^H$	$\varepsilon_{\text{mean}H}^{\tau}$	$\varepsilon_{\text{mean}H}^H$
MMLE	0.101	0.113	0.070	0.204
MLMMLE	0.117	0.203	0.057	0.291
GCV	0.136	0.269	0.109	0.274
MLGCV	0.116	0.307	0.147	0.346

Figure 6. Relative errors $\varepsilon_{\text{mean}}^{\tau}$ and $\varepsilon_{\text{mean}}^H$ versus τ_t for $H_t = 3.0$ km (upper panels) and H_t for $\tau_t = 1.0$ (lower panels). The surface albedo is included in the retrieval and all four aerosol models are considered.

Figure 7. Relative errors $\epsilon_{\text{mean}}^{\tau}$ and ϵ_{mean}^H versus τ_t for $H_t = 3.0$ km (upper panels) and H_t for $\tau_t = 1.0$ (lower panels). The surface albedo is included in the retrieval and all aerosol models excepting the exact one are considered.

8. Conclusions

We designed a retrieval algorithm that takes into account uncertainty in model selection. The solution corresponding to a specific model is characterized by a metric called relative evidence, which is a measure of the fit between the model and the measurement. Based on this metric, the maximum solution estimate, corresponding to the model with the highest evidence, and the mean solution estimate, representing a linear combination of solutions weighted by their evidence, are introduced.

The retrieval algorithm is based on:

1. an application of the prewhitening technique in order to transform the data model into a scaled model with white noise;
2. a deterministic regularization method, i.e., the iteratively regularized Gauss–Newton method, in order to compute the regularized solution (equivalent to the maximum a posteriori estimate of the solution in a Bayesian framework) and to determine the optimal value of the regularization parameter (equivalent to the ratio of the data error and a priori state variances in a Bayesian framework);
3. a linearization of the forward model around the solution in order to transform the nonlinear data model into a linear model and, in turn, facilitate an analytical integration of the likelihood density over the state vector;
4. an extension of maximum marginal likelihood estimation and generalized cross-validation to model selection and data error variance estimation.

Essentially, the algorithm includes four selection models corresponding to:

1. the two parameter choice methods used (maximum marginal likelihood estimation and generalized cross-validation) and
2. the two settings in which the relative evidence is treated (stochastic and deterministic).

The algorithm is applied to the retrieval of aerosol optical thickness and aerosol layer height from synthetic measurements corresponding to the EPIC instrument. In the simulations, the aerosol models implemented in the MODIS aerosol algorithm over land are considered, and the surface albedo is either assumed to be known or included in the retrieval. The following conclusions are drawn:

1. The differences between the results corresponding to the stochastic and deterministic interpretations of the relative evidence are not significant.
2. If the surface albedo is assumed to be known, generalized cross-validation is superior to maximum marginal likelihood estimation; if the surface albedo is included in the retrieval, the contrary is true.
3. The errors in the aerosol optical thickness retrieval are smaller than those in the aerosol layer height retrieval. In the most realistic situation, when the exact aerosol model and surface albedo are unknown, the average relative errors in the retrieved aerosol optical thickness are about 10%, while the corresponding errors in the aerosol layer height are about 20%.
4. The maximum solution estimate is completely unrealistic when both the aerosol model and surface albedo are unknown.

Author Contributions: Conceptualization: V.N. and A.D.; Data curation: S.S.; Formal analysis: S.S. and V.M.G.; Funding acquisition: V.N., D.L. and A.D.; Investigation: S.S.; Methodology: V.N. and A.D.; Project administration: D.L. and A.D.; Supervision: V.N., D.S.E. and A.D.; Validation: V.N. and A.D.; Writing—original draft: S.S.; Writing—review & editing: S.S., V.N., D.S.E., D.L. and A.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the German Aerospace Center (DLR) and the German Academic Exchange Service (DAAD) through the program DLR/DAAD Research Fellowships 2015 (57186656), with Reference Numbers 91613528 and 91627488.

Acknowledgments: A portion of this research was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration (80NM0018D0004). V.N. acknowledges support from the NASA Earth Science US Participating Investigator program (Solicitation NNH16ZDA001N-ESUSPI).

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

In this Appendix, we compute the marginal likelihood for the linear data model (26) with $\mathbf{p} \sim \mathcal{N}(\mathbf{0}, \mathbf{C}_x = \sigma_m^2(\hat{\alpha}\mathbf{L}^T\mathbf{L})^{-1})$ and $\bar{\delta}_m \sim \mathcal{N}(\mathbf{0}, \bar{\mathbf{C}}_{\delta m} = \sigma_m^2\mathbf{I}_M)$. In this case, the a priori and likelihood densities $p(\mathbf{p} | m)$ and $p(\bar{\mathbf{y}}^\delta | \mathbf{p}, m)$ are given by:

$$p(\mathbf{p} | m) = \frac{1}{\sqrt{(2\pi\sigma_m^2)^N \det[(\hat{\alpha}\mathbf{L}^T\mathbf{L})^{-1}]}} \times \exp\left(-\frac{\hat{\alpha}}{2\sigma_m^2} \mathbf{p}^T \mathbf{L}^T \mathbf{L} \mathbf{p}\right)$$

and:

$$p(\bar{\mathbf{y}}^\delta | \mathbf{p}, m) = \frac{1}{\sqrt{(2\pi\sigma_m^2)^M}} \times \exp\left[-\frac{1}{2\sigma_m^2} (\bar{\mathbf{y}}_{m\hat{\alpha}}^\delta - \bar{\mathbf{K}}_{m\hat{\alpha}}\mathbf{p})^T (\bar{\mathbf{y}}_{m\hat{\alpha}}^\delta - \bar{\mathbf{K}}_{m\hat{\alpha}}\mathbf{p})\right],$$

respectively. Using the identity:

$$\begin{aligned} & (\bar{\mathbf{y}}_{m\hat{\alpha}}^\delta - \bar{\mathbf{K}}_{m\hat{\alpha}}\mathbf{p})^T (\bar{\mathbf{y}}_{m\hat{\alpha}}^\delta - \bar{\mathbf{K}}_{m\hat{\alpha}}\mathbf{p}) + \hat{\alpha}\mathbf{p}^T \mathbf{L}^T \mathbf{L} \mathbf{p} \\ &= (\mathbf{p} - \mathbf{p}_{m\hat{\alpha}}^\delta)^T (\bar{\mathbf{K}}_{m\hat{\alpha}}^T \bar{\mathbf{K}}_{m\hat{\alpha}} + \hat{\alpha}\mathbf{L}^T \mathbf{L}) (\mathbf{p} - \mathbf{p}_{m\hat{\alpha}}^\delta) \\ &+ \bar{\mathbf{y}}_{m\hat{\alpha}}^{\delta T} (\mathbf{I}_M - \hat{\mathbf{A}}_{m\hat{\alpha}}) \bar{\mathbf{y}}_{m\hat{\alpha}}^\delta, \end{aligned}$$

where (cf. Equation (23)) $\mathbf{p}_{m\hat{\alpha}}^\delta = \bar{\mathbf{K}}_{m\hat{\alpha}}^+ \bar{\mathbf{y}}_{m\hat{\alpha}}^\delta$ and $\hat{\mathbf{A}}_{m\hat{\alpha}} = \bar{\mathbf{K}}_{m\hat{\alpha}} \bar{\mathbf{K}}_{m\hat{\alpha}}^+$, we express the integrand in Equation (8) as:

$$\begin{aligned}
& p(\bar{\mathbf{y}}^\delta \mid \mathbf{x}, m) p(\mathbf{x} \mid m) \\
&= \frac{1}{\sqrt{(2\pi\sigma_m^2)^{N+M} \det[(\hat{\alpha}\mathbf{L}^T\mathbf{L})^{-1}]}} \\
&\times \exp\left[-\frac{1}{2\sigma_m^2}(\mathbf{p} - \mathbf{p}_{m\hat{\alpha}})^\top (\bar{\mathbf{K}}_{m\hat{\alpha}}^\top \bar{\mathbf{K}}_{m\hat{\alpha}} + \hat{\alpha}\mathbf{L}^T\mathbf{L})(\mathbf{p} - \mathbf{p}_{m\hat{\alpha}})\right] \\
&\times \exp\left[-\frac{1}{2\sigma_m^2}\bar{\mathbf{y}}_{m\hat{\alpha}}^{\delta T}(\mathbf{I}_M - \hat{\mathbf{A}}_{m\hat{\alpha}})\bar{\mathbf{y}}_{m\hat{\alpha}}^\delta\right].
\end{aligned}$$

The normalization condition:

$$\begin{aligned}
& \int \exp\left\{-\frac{1}{2}(\mathbf{p} - \mathbf{p}_{m\hat{\alpha}})^\top [\sigma_m^2(\bar{\mathbf{K}}_{m\hat{\alpha}}^\top \bar{\mathbf{K}}_{m\hat{\alpha}} + \hat{\alpha}\mathbf{L}^T\mathbf{L})^{-1}]^{-1}(\mathbf{p} - \mathbf{p}_{m\hat{\alpha}})\right\} d\mathbf{p} \\
&= \sqrt{(2\pi\sigma_m^2)^N \det[(\bar{\mathbf{K}}_{m\hat{\alpha}}^\top \bar{\mathbf{K}}_{m\hat{\alpha}} + \hat{\alpha}\mathbf{L}^T\mathbf{L})^{-1}]},
\end{aligned}$$

the identities $\det(\mathbf{A}^{-1}) = [\det(\mathbf{A})]^{-1}$ and:

$$(\bar{\mathbf{K}}_{m\hat{\alpha}}^\top \bar{\mathbf{K}}_{m\hat{\alpha}} + \hat{\alpha}\mathbf{L}^T\mathbf{L})^{-1}\hat{\alpha}\mathbf{L}^T\mathbf{L} = \mathbf{I}_N - \mathbf{A}_{m\hat{\alpha}},$$

where $\mathbf{A}_{m\hat{\alpha}} = \bar{\mathbf{K}}_{m\hat{\alpha}}^\top \bar{\mathbf{K}}_{m\hat{\alpha}}$ is the averaging kernel matrix, and the relation:

$$\det(\mathbf{I}_N - \mathbf{A}_{m\hat{\alpha}}) = \det(\mathbf{I}_M - \hat{\mathbf{A}}_{m\hat{\alpha}}),$$

yield the following expression for the marginal likelihood density:

$$\begin{aligned}
p(\bar{\mathbf{y}}^\delta \mid m) &= \sqrt{\frac{\det(\mathbf{I}_M - \hat{\mathbf{A}}_{m\hat{\alpha}})}{(2\pi\sigma_m^2)^M}} \\
&\times \exp\left[-\frac{1}{2\sigma_m^2}\bar{\mathbf{y}}_{m\hat{\alpha}}^{\delta T}(\mathbf{I}_M - \hat{\mathbf{A}}_{m\hat{\alpha}})\bar{\mathbf{y}}_{m\hat{\alpha}}^\delta\right]. \tag{A1}
\end{aligned}$$

Appendix B

In this Appendix, we present a general tool for estimating the data error variance and the optimal model. Consider the linear model (26), i.e., $\bar{\mathbf{y}}_{m\hat{\alpha}}^\delta = \bar{\mathbf{K}}_{m\hat{\alpha}}\mathbf{p} + \bar{\boldsymbol{\delta}}_m$, and a singular value decomposition of the quotient matrix $\bar{\mathbf{K}}_{m\hat{\alpha}}\mathbf{L}^{-1} = \mathbf{U}\boldsymbol{\Gamma}\mathbf{V}^\top$, where $\boldsymbol{\Gamma} = \text{diag}(\gamma_i)_N$ with the convention $\gamma_{mi}^2 = 0$ for $i > N$, $\mathbf{U} = [\mathbf{u}_1, \dots, \mathbf{u}_M]$, and $\mathbf{V} = [\mathbf{v}_1, \dots, \mathbf{v}_N]$. The covariance matrix of the data $\bar{\mathbf{y}}_{m\hat{\alpha}}^\delta$ can be computed as:

$$\begin{aligned}
\mathcal{E}\{\bar{\mathbf{y}}_{m\hat{\alpha}}^\delta \bar{\mathbf{y}}_{m\hat{\alpha}}^{\delta T}\} &= \bar{\mathbf{K}}_{m\hat{\alpha}}\mathbf{C}_x\bar{\mathbf{K}}_{m\hat{\alpha}}^\top + \bar{\mathbf{C}}_{\delta m} \\
&= \frac{\sigma_m^2}{\hat{\alpha}}\bar{\mathbf{K}}_{m\hat{\alpha}}(\mathbf{L}^T\mathbf{L})^{-1}\bar{\mathbf{K}}_{m\hat{\alpha}}^\top + \sigma_m^2\mathbf{I}_M \\
&= \mathbf{U}\boldsymbol{\Sigma}_{\bar{\mathbf{y}}}\mathbf{U}^\top
\end{aligned}$$

where:

$$\boldsymbol{\Sigma}_{\bar{\mathbf{y}}} = \text{diag}\left[\sigma_m^2\left(\frac{\gamma_{mi}^2 + \hat{\alpha}}{\hat{\alpha}}\right)\right]_M. \tag{A2}$$

Define the scaled data:

$$\bar{\mathbf{Y}}_{m\hat{\alpha}}^\delta = \mathbf{U}^T\bar{\mathbf{y}}_{m\hat{\alpha}}^\delta$$

and note that the covariance matrix of $\bar{\mathbf{Y}}_{m\hat{\alpha}}^\delta$ is the diagonal matrix $\Sigma_{\bar{\mathbf{Y}}}$, i.e.,

$$\mathcal{E}\{\bar{\mathbf{Y}}_{m\hat{\alpha}}^\delta \bar{\mathbf{Y}}_{m\hat{\alpha}}^{\delta T}\} = \Sigma_{\bar{\mathbf{Y}}}.$$

If σ_m^2 is the correct data error variance, we must have (cf. Equation (A2)):

$$\mathcal{E}\{\bar{Y}_{m\hat{\alpha}i}^{\delta 2}\} = \sigma_m^2 \left(\frac{\gamma_{mi}^2 + \hat{\alpha}}{\hat{\alpha}} \right), \quad i = 1, \dots, M,$$

where $\bar{Y}_{m\hat{\alpha}i}^\delta = [\bar{\mathbf{Y}}_{m\hat{\alpha}}^\delta]_i$. If σ_m^2 is unknown, we can find the estimate $\hat{\sigma}_m^2$ from the equations:

$$\mathcal{E}\{\bar{Y}_{m\hat{\alpha}i}^{\delta 2}\} = \hat{\sigma}_m^2 \left(\frac{\gamma_{mi}^2 + \hat{\alpha}}{\hat{\alpha}} \right), \quad i = 1, \dots, M. \quad (\text{A3})$$

However, since only one realization of the random vector $\bar{\mathbf{Y}}_{m\hat{\alpha}}^\delta$ is known, the calculation of $\hat{\sigma}_m^2$ from Equation (A3) may lead to erroneous results. Therefore, we look for another selection criterion.

Set:

$$a_i(\sigma_m^2) = \sigma_m^2 \left(\frac{\gamma_{mi}^2 + \hat{\alpha}}{\hat{\alpha}} \right) \quad (\text{A4})$$

and define the function:

$$f(\bar{\mathbf{Y}}_{m\hat{\alpha}}^\delta | m, \sigma_m^2) = \sum_{i=1}^M \left\{ \psi(a_i(\sigma_m^2)) + \psi'(a_i(\sigma_m^2)) [\bar{Y}_{m\hat{\alpha}i}^{\delta 2} - a_i(\sigma_m^2)] \right\}$$

with $\psi(a)$ being a strictly concave function, i.e., $\psi''(a) < 0$. The expected value of $f(\cdot)$ is given by:

$$\mathcal{E}\{f(\bar{\mathbf{Y}}_{m\hat{\alpha}}^\delta | m, \sigma_m^2)\} = \sum_{i=1}^M \left\{ \psi(a_i(\sigma_m^2)) + \psi'(a_i(\sigma_m^2)) [\mathcal{E}\{\bar{Y}_{m\hat{\alpha}i}^{\delta 2}\} - a_i(\sigma_m^2)] \right\},$$

hence, defining the estimate $\hat{\sigma}_m^2$ through the relation (cf. Equations (A3) and (A4)):

$$\mathcal{E}\{\bar{Y}_{m\hat{\alpha}i}^{\delta 2}\} = a_i(\hat{\sigma}_m^2), \quad (\text{A5})$$

we express $\mathcal{E}\{f(\bar{\mathbf{Y}}_{m\hat{\alpha}}^\delta | m, \sigma_m^2)\}$ as:

$$\mathcal{E}\{f(\bar{\mathbf{Y}}_{m\hat{\alpha}}^\delta | m, \sigma_m^2)\} = \sum_{i=1}^M \left\{ \psi(a_i(\sigma_m^2)) + \psi'(a_i(\sigma_m^2)) [a_i(\hat{\sigma}_m^2) - a_i(\sigma_m^2)] \right\}.$$

Then, we obtain:

$$\begin{aligned} \mathcal{E}\{f(\bar{\mathbf{Y}}_{m\hat{\alpha}}^\delta | m, \sigma_m^2)\} - \mathcal{E}\{f(\bar{\mathbf{Y}}_{m\hat{\alpha}}^\delta | m, \hat{\sigma}_m^2)\} &= \sum_{i=1}^M \left\{ \psi(a_i(\sigma_m^2)) - \psi(a_i(\hat{\sigma}_m^2)) \right. \\ &\quad \left. + \psi'(a_i(\sigma_m^2)) [a_i(\hat{\sigma}_m^2) - a_i(\sigma_m^2)] \right\}. \end{aligned} \quad (\text{A6})$$

Considering the second-order Taylor expansion around $a_i(\sigma_m^2)$,

$$\begin{aligned} \psi(a_i(\hat{\sigma}_m^2)) &= \psi(a_i(\sigma_m^2)) + \psi'(a_i(\sigma_m^2)) [a_i(\hat{\sigma}_m^2) - a_i(\sigma_m^2)] \\ &\quad + \frac{1}{2} \psi''(\xi_i) [a_i(\hat{\sigma}_m^2) - a_i(\sigma_m^2)]^2 \end{aligned}$$

for some ξ_i between $a_i(\sigma_m^2)$ and $a_i(\hat{\sigma}_m^2)$ and taking into account the fact that ψ is strictly concave, we deduce that each term in the sum (A6) is non-negative and vanishes only for $a_i(\sigma_m^2) = a_i(\hat{\sigma}_m^2)$. Thus, we have:

$$\mathcal{E}\{f(\bar{\mathbf{Y}}_{m\hat{\alpha}}^\delta|m, \sigma_m^2)\} \geq \mathcal{E}\{f(\bar{\mathbf{Y}}_{m\hat{\alpha}}^\delta|m, \hat{\sigma}_m^2)\}$$

for all σ_m^2 . In conclusion, $\hat{\sigma}_m^2$, defined through Equation (A5), is the unique global minimizer of $\mathcal{E}\{f(\bar{\mathbf{Y}}_{m\hat{\alpha}}^\delta|m, \sigma_m^2)\}$, i.e.,

$$\hat{\sigma}_m^2 = \arg \min_{\sigma_m^2} \mathcal{E}\{f(\bar{\mathbf{Y}}_{m\hat{\alpha}}^\delta|m, \sigma_m^2)\}. \quad (\text{A7})$$

Coming to the optimal model $\hat{m}$, it seems that a natural choice of $\hat{m}$ is:

$$\hat{m} = \arg \min_m \mathcal{E}\{f(\bar{\mathbf{Y}}_{m\hat{\alpha}}^\delta|m, \hat{\sigma}_m^2)\}. \quad (\text{A8})$$

However, instead of Equation (A8), we use a more general selection criterion, namely the optimal model $\hat{m}$ defined as:

$$\hat{m} = \arg \min_m F(\mathcal{E}\{f(\bar{\mathbf{Y}}_{m\hat{\alpha}}^\delta|m, \hat{\sigma}_m^2)\}), \quad (\text{A9})$$

where F is a monotonic increasing function of $\mathcal{E}\{f(\bar{\mathbf{Y}}_{m\hat{\alpha}}^\delta|m, \hat{\sigma}_m^2)\}$.

The generalized cross-validation [10,11] and maximum marginal likelihood estimation [7–9] can be obtained by a particular choice of the function $\psi(a)$.

Generalized cross-validation

For the choice:

$$\psi(a) = 1 - \frac{1}{a},$$

we obtain:

$$\mathcal{E}\{f(\bar{\mathbf{Y}}_{m\hat{\alpha}}^\delta|m, \sigma_m^2)\} = M + \sum_{i=1}^M \left[\frac{\mathcal{E}\{\bar{Y}_{m\hat{\alpha}i}^{\delta 2}\}}{a_i^2(\sigma_m^2)} - \frac{2}{a_i(\sigma_m^2)} \right]. \quad (\text{A10})$$

Since $\hat{\sigma}_m^2$ is the unique global minimizer of $\mathcal{E}\{f(\bar{\mathbf{Y}}_{m\hat{\alpha}}^\delta|m, \sigma_m^2)\}$, the condition:

$$\frac{d}{d\sigma_m^2} \mathcal{E}\{f(\bar{\mathbf{Y}}_{m\hat{\alpha}}^\delta|m, \hat{\sigma}_m^2)\} = 0, \quad (\text{A11})$$

yields:

$$\sum_{i=1}^M \left[\frac{\mathcal{E}\{\bar{Y}_{m\hat{\alpha}i}^{\delta 2}\}}{a_i^2(\hat{\sigma}_m^2)} - \frac{1}{a_i(\hat{\sigma}_m^2)} \right] = 0. \quad (\text{A12})$$

The above equation together with the relation:

$$a_i(\hat{\sigma}_m^2) = \hat{\sigma}_m^2 \frac{\gamma_{mi}^2 + \hat{\alpha}}{\hat{\alpha}}, \quad (\text{A13})$$

provide the following estimate for the data error variance:

$$\hat{\sigma}_m^2 = \frac{\sum_{i=1}^M \left(\frac{\hat{\alpha}}{\gamma_{mi}^2 + \hat{\alpha}} \right)^2 \mathcal{E}\{\bar{Y}_{m\hat{\alpha}i}^{\delta 2}\}}{\sum_{i=1}^M \frac{\hat{\alpha}}{\gamma_{mi}^2 + \hat{\alpha}}}. \quad (\text{A14})$$

On the other hand, from Equations (A10) and (A12), we obtain:

$$\mathcal{E}\{f(\bar{\mathbf{Y}}_{m\hat{\alpha}}^\delta|m, \hat{\sigma}_m^2)\} = M - \sum_{i=1}^M \frac{1}{a_i(\hat{\sigma}_m^2)},$$

and we define the function $F(\cdot)$ by:

$$F(\mathcal{E}\{f(\bar{\mathbf{Y}}_{m\hat{\alpha}}^\delta|m, \hat{\sigma}_m^2)\}) = \frac{1}{M - \mathcal{E}\{f(\bar{\mathbf{Y}}_{m\hat{\alpha}}^\delta|m, \hat{\sigma}_m^2)\}}.$$

Taking Equations (A13) and (A14) into account, we find that the function $F(\cdot)$ can be calculated as:

$$F(\mathcal{E}\{f(\bar{\mathbf{Y}}_{m\hat{\alpha}}^\delta|m, \hat{\sigma}_m^2)\}) = \frac{\sum_{i=1}^M \left(\frac{\hat{\alpha}}{\gamma_{mi}^2 + \hat{\alpha}} \right)^2 \mathcal{E}\{\bar{Y}_{m\hat{\alpha}i}^{\delta 2}\}}{\left(\sum_{i=1}^M \frac{\hat{\alpha}}{\gamma_{mi}^2 + \hat{\alpha}} \right)^2}.$$

In practice, the expectation $\mathcal{E}\{\bar{Y}_{m\hat{\alpha}i}^{\delta 2}\}$ cannot be computed since only one realization $\bar{Y}_{m\hat{\alpha}i}^\delta = \mathbf{u}_i^T \bar{\mathbf{y}}_{m\hat{\alpha}}^\delta$ is known. Therefore, we approximate $\mathcal{E}\{\bar{Y}_{m\hat{\alpha}i}^{\delta 2}\} \approx (\mathbf{u}_i^T \bar{\mathbf{y}}_{m\hat{\alpha}}^\delta)^2$ and consider the so-called generalized cross-validation function:

$$v(m) = \frac{\sum_{i=1}^M \left(\frac{\hat{\alpha}}{\gamma_{mi}^2 + \hat{\alpha}} \right)^2 (\mathbf{u}_i^T \bar{\mathbf{y}}_{m\hat{\alpha}}^\delta)^2}{\left(\sum_{i=1}^M \frac{\hat{\alpha}}{\gamma_{mi}^2 + \hat{\alpha}} \right)^2}. \quad (\text{A15})$$

Finally, using the representations:

$$\|\bar{\mathbf{y}}_{m\hat{\alpha}}^\delta - \bar{\mathbf{K}}_{m\hat{\alpha}} \mathbf{p}_{m\hat{\alpha}}^\delta\|^2 = \|\mathbf{r}_{m\hat{\alpha}}^\delta\|^2 = \sum_{i=1}^M \left(\frac{\hat{\alpha}}{\gamma_{mi}^2 + \hat{\alpha}} \right)^2 (\mathbf{u}_i^T \bar{\mathbf{y}}_{m\hat{\alpha}}^\delta)^2, \quad (\text{A16})$$

$$\text{trace}(\mathbf{I}_M - \hat{\mathbf{A}}_{m\hat{\alpha}}) = \sum_{i=1}^M \frac{\hat{\alpha}}{\gamma_{mi}^2 + \hat{\alpha}}, \quad (\text{A17})$$

we express the estimate of the data error variance as:

$$\hat{\sigma}_m^2 = \frac{\|\mathbf{r}_{m\hat{\alpha}}^\delta\|^2}{[\text{trace}(\mathbf{I}_M - \hat{\mathbf{A}}_{m\hat{\alpha}})]}, \quad (\text{A18})$$

and the generalized cross-validation function as:

$$v(m) = \frac{\|\mathbf{r}_{m\hat{\alpha}}^\delta\|^2}{[\text{trace}(\mathbf{I}_M - \hat{\mathbf{A}}_{m\hat{\alpha}})]^2}. \quad (\text{A19})$$

Maximum marginal likelihood estimation

For the choice:

$$\psi(a) = \ln a \Rightarrow \psi'(a) = \frac{1}{a},$$

we obtain:

$$\mathcal{E}\{f(\bar{\mathbf{Y}}_{m\hat{\alpha}}^\delta|m, \sigma_m^2)\} = -M + \sum_{i=1}^M \left[\frac{\mathcal{E}\{\bar{Y}_{m\hat{\alpha}i}^{\delta 2}\}}{a_i(\sigma_m^2)} + \ln a_i(\sigma_m^2) \right]. \quad (\text{A20})$$

From the minimization condition (A11), we find that $\hat{\sigma}_m^2$ satisfying the equation:

$$\sum_{i=1}^M \frac{\mathcal{E}\{\bar{Y}_{m\hat{\alpha}}^{\delta 2}\}}{a_i(\hat{\sigma}_m^2)} = M, \quad (\text{A21})$$

is given by:

$$\hat{\sigma}_m^2 = \frac{1}{M} \sum_{i=1}^M \frac{\hat{\alpha}}{\gamma_{mi}^2 + \hat{\alpha}} \mathcal{E}\{\bar{Y}_{m\hat{\alpha}}^{\delta 2}\}. \quad (\text{A22})$$

Equations (A20) and (A21) yield:

$$\mathcal{E}\{f(\bar{\mathbf{Y}}_{m\hat{\alpha}}^\delta | m, \hat{\sigma}_m^2)\} = \sum_{i=1}^M \ln a_i(\hat{\sigma}_m^2)$$

and we define the function $F(\cdot)$ by:

$$F(\mathcal{E}\{f(\bar{\mathbf{Y}}_{m\hat{\alpha}}^\delta | m, \hat{\sigma}_m^2)\}) = \exp\left[\frac{1}{M} \mathcal{E}\{f(\bar{\mathbf{Y}}_{m\hat{\alpha}}^\delta | m, \hat{\sigma}_m^2)\}\right]$$

Furthermore, using the result:

$$\sum_{i=1}^M \ln a_i(\hat{\sigma}_m^2) = \ln\left(\sum_{i=1}^M \frac{\hat{\alpha}}{\gamma_{mi}^2 + \hat{\alpha}} \mathcal{E}\{\bar{Y}_{m\hat{\alpha}}^{\delta 2}\}\right)^M - \sum_{i=1}^M \ln\left(\frac{\hat{\alpha}}{\gamma_{mi}^2 + \hat{\alpha}}\right),$$

we express $F(\cdot)$ as:

$$F(\mathcal{E}\{f(\bar{\mathbf{Y}}_{m\hat{\alpha}}^\delta | m, \hat{\sigma}_m^2)\}) = \frac{\sum_{i=1}^M \frac{\hat{\alpha}}{\gamma_{mi}^2 + \hat{\alpha}} \mathcal{E}\{\bar{Y}_{m\hat{\alpha}}^{\delta 2}\}}{\prod_{i=1}^M \left(\frac{\hat{\alpha}}{\gamma_{mi}^2 + \hat{\alpha}}\right)^{1/M}}.$$

Considering the approximation $\mathcal{E}\{\bar{Y}_{m\hat{\alpha}}^{\delta 2}\} \approx (\mathbf{u}_i^T \bar{\mathbf{y}}_{m\hat{\alpha}}^\delta)^2$, we define the maximum likelihood function $\lambda(m)$ by:

$$\lambda(m) = \frac{\sum_{i=1}^M \frac{\hat{\alpha}}{\gamma_{mi}^2 + \hat{\alpha}} (\mathbf{u}_i^T \bar{\mathbf{y}}_{m\hat{\alpha}}^\delta)^2}{\prod_{i=1}^M \left(\frac{\hat{\alpha}}{\gamma_{mi}^2 + \hat{\alpha}}\right)^{1/M}} \quad (\text{A23})$$

Using the results:

$$\bar{\mathbf{y}}_{m\hat{\alpha}}^{\delta T} (\mathbf{I}_M - \hat{\mathbf{A}}_{m\hat{\alpha}}) \bar{\mathbf{y}}_{m\hat{\alpha}}^\delta = \sum_{i=1}^M \frac{\hat{\alpha}}{\gamma_{mi}^2 + \hat{\alpha}} (\mathbf{u}_i^T \bar{\mathbf{y}}_{m\hat{\alpha}}^\delta)^2, \quad (\text{A24})$$

$$\det(\mathbf{I}_M - \hat{\mathbf{A}}_{m\hat{\alpha}}) = \prod_{i=1}^M \frac{\hat{\alpha}}{\gamma_{mi}^2 + \hat{\alpha}}, \quad (\text{A25})$$

we express the estimate for the data error variance as:

$$\hat{\sigma}_{Lm}^2 = \frac{1}{M} \bar{\mathbf{y}}_{m\hat{\alpha}}^{\delta T} (\mathbf{I}_M - \hat{\mathbf{A}}_{m\hat{\alpha}}) \bar{\mathbf{y}}_{m\hat{\alpha}}^\delta, \quad (\text{A26})$$

and the maximum likelihood function as:

$$\lambda(m) = \frac{\bar{\mathbf{y}}_{m\hat{\alpha}}^{\delta T} (\mathbf{I}_M - \hat{\mathbf{A}}_{m\hat{\alpha}}) \bar{\mathbf{y}}_{m\hat{\alpha}}^\delta}{\sqrt[M]{\det(\mathbf{I}_M - \hat{\mathbf{A}}_{m\hat{\alpha}})}}. \quad (\text{A27})$$

Appendix C

In this Appendix, we derive an estimate for the data error variance by analyzing the residual of the linear equation $\bar{\mathbf{y}}_{m\hat{\alpha}}^\delta = \bar{\mathbf{K}}_{m\hat{\alpha}} \mathbf{p}$. In terms of a singular value decomposition of the quotient matrix $\bar{\mathbf{K}}_{m\hat{\alpha}} \mathbf{L}^{-1} = \mathbf{U} \mathbf{\Gamma} \mathbf{V}^T$ with $\mathbf{\Gamma} = \text{diag}(\gamma_i)_N$, $\gamma_{mi}^2 = 0$ for $i > N$, $\mathbf{U} = [\mathbf{u}_1, \dots, \mathbf{u}_M]$, and $\mathbf{V} = [\mathbf{v}_1, \dots, \mathbf{v}_N]$, the squared norm of the residual of the linear equation $\bar{\mathbf{y}}_{m\hat{\alpha}}^\delta = \bar{\mathbf{K}}_{m\hat{\alpha}} \mathbf{p}$ is:

$$\|\bar{\mathbf{y}}_{m\hat{\alpha}}^\delta - \bar{\mathbf{K}}_{m\hat{\alpha}} \mathbf{p}_{m\hat{\alpha}}^\delta\|^2 = \sum_{i=1}^M \left(\frac{\hat{\alpha}}{\gamma_{mi}^2 + \hat{\alpha}} \right)^2 (\mathbf{u}_i^T \bar{\mathbf{y}}_{m\hat{\alpha}}^\delta)^2. \quad (\text{A28})$$

In the data model (cf. Equation (26)) $\bar{\mathbf{y}}_{m\hat{\alpha}}^\delta = \bar{\mathbf{K}}_{m\hat{\alpha}} \mathbf{p} + \bar{\delta}_m$, we set $\bar{\mathbf{y}}_{m\hat{\alpha}} = \bar{\mathbf{K}}_{m\hat{\alpha}} \mathbf{p}_{m\hat{\alpha}}^\delta$ and rewrite this equation as:

$$\bar{\mathbf{y}}_{m\hat{\alpha}}^\delta = \bar{\mathbf{y}}_{m\hat{\alpha}} + \bar{\delta}_m. \quad (\text{A29})$$

Using the relation:

$$\begin{aligned} \mathcal{E}\{(\mathbf{u}_i^T \bar{\delta}_m)(\mathbf{u}_j^T \bar{\delta}_m)\} &= \mathcal{E}\left\{\sum_{k=1}^M \sum_{l=1}^M [\bar{\delta}_m]_k [\bar{\delta}_m]_l [\mathbf{u}_i]_k [\mathbf{u}_j]_l\right\} \\ &= \sigma_m^2 \mathbf{u}_i^T \mathbf{u}_j \\ &= \sigma_m^2 \delta_{ij}, \end{aligned}$$

where δ_{ij} is the Kronecker delta function, we get:

$$\mathcal{E}\{(\mathbf{u}_i^T \bar{\delta}_m)^2\} = \sigma_m^2, \quad (\text{A30})$$

so that from Equations (A29) and (A30), we obtain:

$$\mathcal{E}\{(\mathbf{u}_i^T \bar{\mathbf{y}}_{m\hat{\alpha}}^\delta)^2\} = (\mathbf{u}_i^T \bar{\mathbf{y}}_{m\hat{\alpha}})^2 + \sigma_m^2. \quad (\text{A31})$$

Since $\bar{\mathbf{y}}_{m\hat{\alpha}} = \bar{\mathbf{K}}_{m\hat{\alpha}} \mathbf{p}_{m\hat{\alpha}}^\delta$ belongs to the range of the matrix operator $\bar{\mathbf{K}}_{m\hat{\alpha}}$, which, in turn, is spanned by the vectors $\{\mathbf{u}_i\}_{i=1}^N$, we have $\mathbf{u}_i^T \bar{\mathbf{y}}_{m\hat{\alpha}} = 0$ for $i > N$, and further (cf. Equation (A31)):

$$\mathcal{E}\{(\mathbf{u}_i^T \bar{\mathbf{y}}_{m\hat{\alpha}}^\delta)^2\} = \sigma_m^2 \text{ for } i > N. \quad (\text{A32})$$

Taking the expected value of Equation (A28) and using Equations (A31) and (A32), we get:

$$\mathcal{E}\{\|\bar{\mathbf{y}}_{m\hat{\alpha}}^\delta - \bar{\mathbf{K}}_{m\hat{\alpha}} \mathbf{p}_{m\hat{\alpha}}^\delta\|^2\} = (M - N)\sigma_m^2 + \sum_{i=1}^N \left(\frac{\hat{\alpha}}{\gamma_{mi}^2 + \hat{\alpha}} \right)^2 [(\mathbf{u}_i^T \bar{\mathbf{y}}_{m\hat{\alpha}})^2 + \sigma_m^2]$$

Now, if $\hat{\alpha} \ll \gamma_{mi}^2$ for all $i = 1, \dots, N$, we approximate:

$$\mathcal{E}\{\|\bar{\mathbf{y}}_{m\hat{\alpha}}^\delta - \bar{\mathbf{K}}_{m\hat{\alpha}} \mathbf{p}_{m\hat{\alpha}}^\delta\|^2\} \approx (M - N)\sigma_m^2,$$

and deduce that an estimate for the data error variance is:

$$\begin{aligned} \hat{\sigma}_m^2 &\approx \frac{1}{M - N} \mathcal{E}\{\|\bar{\mathbf{y}}_{m\hat{\alpha}}^\delta - \bar{\mathbf{K}}_{m\hat{\alpha}} \mathbf{p}_{m\hat{\alpha}}^\delta\|^2\} \\ &\approx \frac{1}{M - N} \|\bar{\mathbf{y}}_{m\hat{\alpha}}^\delta - \bar{\mathbf{K}}_{m\hat{\alpha}} \mathbf{p}_{m\hat{\alpha}}^\delta\|^2 \\ &= \frac{1}{M - N} \|\mathbf{r}_{m\hat{\alpha}}^\delta\|^2. \end{aligned} \quad (\text{A33})$$

References

1. Levy, R.C.; Remer, L.A.; Dubovik, O. Global aerosol optical properties and application to Moderate Resolution Imaging Spectroradiometer aerosol retrieval over land. *J. Geophys. Res. Atmos.* **2007**, *112*. [\[CrossRef\]](#)
2. Hoeting, J.A.; Madigan, D.; Raftery, A.E.; Volinsky, C.T. Bayesian Model averaging: A tutorial. *Stat. Sci.* **1999**, *14*, 382–401.
3. Määttä, A.; Laine, M.; Tamminen, J.; Veefkind, J. Quantification of uncertainty in aerosol optical thickness retrieval arising from aerosol microphysical model and other sources, applied to Ozone Monitoring Instrument (OMI) measurements. *Atmos. Meas. Tech.* **2014**, *7*, 1185–1199. [\[CrossRef\]](#)
4. Kauppi, A.; Kolmonen, P.; Laine, M.; Tamminen, J. Aerosol-type retrieval and uncertainty quantification from OMI data. *Atmos. Meas. Tech.* **2017**, *10*, 4079–4098. [\[CrossRef\]](#)
5. Doicu, A.; Trautmann, T.; Schreier, F. *Numerical Regularization for Atmospheric Inverse Problems*; Springer: Berlin/Heidelberg, Germany, 2010. [\[CrossRef\]](#)
6. Tarantola, A. *Inverse Problem Theory and Methods for Model Parameter Estimation*; Society for Industrial and Applied Mathematics (SIAM): Philadelphia, PA, USA, 2005. [\[CrossRef\]](#)
7. Patterson, H.; Thompson, R. Recovery of inter-block information when block sizes are unequal. *Biometrika* **1971**, *58*, 545–554. [\[CrossRef\]](#)
8. Smyth, G.K.; Verbyla, A.P. A conditional likelihood approach to residual maximum likelihood estimation in generalized linear models. *J. R. Stat. Soc. Ser. Methodol.* **1996**, *58*, 565–572. [\[CrossRef\]](#)
9. Stuart, A.; Ord, J.; Arnols, S. *Kendall's Advanced Theory of Statistics. Volume 2A: Classical Inference and the Linear Model*; Oxford University Press Inc.: Oxford, UK, 1999.
10. Wahba, G. Practical approximate solutions to linear operator equations when the data are noisy. *SIAM J. Numer. Anal.* **1977**, *14*, 651–667. [\[CrossRef\]](#)
11. Wahba, G. *Spline Models for Observational Data*; Society for Industrial and Applied Mathematics (SIAM): Philadelphia, PA, USA, 1990. [\[CrossRef\]](#)
12. Marshak, A.; Herman, J.; Szabo, A.; Blank, K.; Carn, S.; Cede, A.; Geogdzhayev, I.; Huang, D.; Huang, L.K.; Knyazikhin, Y.; et al. Earth Observations from DSCOVR EPIC Instrument. *Bull. Am. Meteorol. Soc.* **2018**, *99*, 1829–1850. [\[CrossRef\]](#) [\[PubMed\]](#)
13. Geogdzhayev, I.V.; Marshak, A. Calibration of the DSCOVR EPIC visible and NIR channels using MODIS Terra and Aqua data and EPIC lunar observations. *Atmos. Meas. Tech.* **2018**, *11*, 359–368. [\[CrossRef\]](#) [\[PubMed\]](#)
14. Xu, X.; Wang, J.; Wang, Y.; Zeng, J.; Torres, O.; Reid, J.; Miller, S.; Martins, J.; Remer, L. Detecting layer height of smoke aerosols over vegetated land and water surfaces via oxygen absorption bands: Hourly results from EPIC/DSCOVR in deep space. *Atmos. Meas. Tech.* **2019**, *12*, 3269–3288. [\[CrossRef\]](#)
15. García, V.M.; Sasi, S.; Efremenko, D.S.; Doicu, A.; Loyola, D. Radiative transfer models for retrieval of cloud parameters from EPIC/DSCOVR measurements. *J. Quant. Spectrosc. Radiat. Transf.* **2018**, *213*, 228–240. [\[CrossRef\]](#)
16. García, V.M.; Sasi, S.; Efremenko, D.S.; Doicu, A.; Loyola, D. Linearized radiative transfer models for retrieval of cloud parameters from EPIC/DSCOVR measurements. *J. Quant. Spectrosc. Radiat. Transf.* **2018**, *213*, 241–251. [\[CrossRef\]](#)
17. Holben, B.; Eck, T.; Slutsker, I.; Tanré, D.; Buis, J.; Setzer, A.; Vermote, E.; Reagan, J.; Kaufman, Y.; Nakajima, T.; et al. AERONET—A federated instrument network and data archive for aerosol characterization. *Remote Sens. Environ.* **1998**, *66*, 1–16. [\[CrossRef\]](#)

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).