A Bayesian approach for the retrieval of atmospheric particle properties from lidar data with uncertainty quantification

OPEN ACCESS
IOP Publishing
Inverse Problems
Inverse Problems **41** (2025) 085001 (20pp)
https://doi.org/10.1088/1361-6420/adf003

A Bayesian approach for the retrieval of atmospheric particle properties from lidar data with uncertainty quantification

Giacomo Varini[1,*], Alessia Sannino[2], Antonella Boselli[3], Riccardo Damiano[2] and Alberto Sorrentino[1]

[1] Dipartimento di Matematica, Dipartimento di Eccellenza 2023-2027, Universitá degli Studi di Genova, Genova, Italy
[2] Dipartimento di Fisica, Universitá degli Studi di Napoli Federico II, Napoli, Italy
[3] CNR-IMAA, Potenza, Italy

E-mail: giacomovarini0410@gmail.com

Received 26 February 2025; revised 9 June 2025
Accepted for publication 15 July 2025
Published 25 July 2025

Abstract

We consider the inverse problem of recovering the particle size distribution of atmospheric aerosol from backscattering and extiction coefficients at three wavelengths, as attainable using ground lidar measurements. We set up a Bayesian model composed of a collection of fixed-dimensional models, and devise a sequential Monte Carlo algorithm to sample its posterior distribution, allowing for automated model selection. As the target distribution is a complex, often multimodal distribution, we first assess the degree of reliability of the sampling procedure. We then go on to characterize the ill-posedness in terms of multimodality of the posterior distribution, in order to verify the presence of multiple alternative scenarios that are compatible with the measured data. We finally assess the accuracy of the reconstruction for practical use. Our results show that the proposed approach can successfully retrieve the particle size distribution from lidar data even under complex circumstances, and reliably characterize the uncertainty that inevitably arises due to the nature of the data.

Keywords: Bayesian approach, sequential Monte Carlo, lidar, atmospheric aerosol, particle size distribution

* Author to whom any correspondence should be addressed.

1. Introduction

In atmospheric applications, lidar systems are widely used to obtain information on geometrical, optical and microphysical properties of the aerosols along the atmospheric profile from few hundred meters to a distance of several kilometers [1–12]. To retrieve aerosol microphysical properties from lidar measurements, two inverse problems must be solved sequentially: in the first one, the measured power is used to estimate the optical extinction and backscattering parameters [13–20]; in the second one, the estimated optical parameters (at three different wavelengths) are used to obtain an estimate of the particle size distribution [21–30]; the latter can then be used to provide estimates of critical quantities, such as concentrations of particulate matter (PM) at 1 μm (PM1), 2.5 μm (PM2.5) and 10 μm (PM10). The second problem is particularly challenging due to the limited number of input variables and the fact that the microphysical parameters are derived from the optical ones through integral equations that cannot be solved analytically, the numerical solution of which leads to an ill-conditioned problem [31]. Existing approaches can be classified in two broad categories: regularization methods (e.g. [21–23]), where the particle size distribution is discretized on a large set of points and regularity of the solution is imposed using a penalty term in the regularization functional; parametric methods (e.g. [28, 30]), where the particle size distribution is modeled as the superposition of one or more components having pre-defined shape and being described by a small number of parameters. Regularization methods provide a flexible framework for the retrieval of aerosol properties, however, they typically lack quantification of uncertainty of the estimated particle size distribution; parametric methods, on the other hand, allow for such uncertainty quantification but have been so far limited by two main factors: (i) the need for manual subjective choice of the number of components and (ii) the presence of local minima that result in potential instability of the solution.

Building on previous work [30], we present Montecarlo Approximation For Automated Lidar Data Analysis (MAFALDA) a Monte Carlo algorithm that approximates the posterior distribution for a Bayesian parametric model in which the particle size distribution is represented as the superposition of log-normal components. Compared to previous work [30] the Bayesian model has been generalized to account for an *a priori* unknown number of components, as well as for *a priori* unknown complex refractive indices, i.e.: the number of components and their complex refractive indices are treated as additional random variables in the Bayesian model and are estimated from the data, allowing for a completely automated model selection. To effectively approximate the posterior distribution of this complex model we also introduce a new Monte Carlo strategy belonging to the family of Sequential Monte Carlo samplers [32] and inspired by similar solutions applied to other inverse problems [33, 34].

The paper is organized as follows: in section 2 we describe the inverse problem to be solved; in section 3 we describe MAFALDA; in section 4 we validate MAFALDA using numerical simulations and in the subsequent section 5 we apply it to two real data sets; finally, we discuss the results in section 6.

2. Retrieval of particle properties from lidar data

Lidar instruments transmit a short laser pulse toward a target and collect the backscattered light at specific wavelengths λ and as a function of the distance z to the target. This measurement can be expressed [35] in terms of the extinction coefficient $\alpha(\lambda, z)$ and the backscattering coefficient $\beta(\lambda, z)$,

$$P(\lambda,z) = \frac{C\beta(\lambda,z)}{z^2}\exp\left(-2\int_0^z \alpha(\lambda,x)\,\mathrm{d}x\right) \tag{1}$$

where P is the power of backscattered light and C is a constant that depends on the instrument characteristics. From measurements of P one can estimate the optical properties $\alpha(\lambda,z)$ and $\beta(\lambda,z)$ at different wavelength and altitudes using well known inversion procedures [13–17, 19, 20]. Once this estimate is available, it is possible to provide a characterization of the microphysical properties of the aerosol based on these indirect measurements. Typically, β is measured at three wavelengths ($\lambda = 355, 532, 1064$ nm) while α is measured at two wavelengths ($\lambda = 355, 532$ nm), giving the so-called $3\beta + 2\alpha$ standard configuration.

Assuming a spherical homogeneous particle approximation, and omitting the dependence on z (as the problem can be addressed independently at different altitudes) we can extract information about the particle size distribution $n(r)$ from the extinction and backscattering coefficients,

$$\alpha(\lambda) = \int_{r_a}^{r_b} k_\alpha(r,\lambda,m)\,n(r)\,\mathrm{d}r \tag{2}$$

$$\beta(\lambda) = \int_{r_a}^{r_b} k_\beta(r,\lambda,m)\,n(r)\,\mathrm{d}r \tag{3}$$

where r_a and r_b represent the lower and upper bounds of the particle sizes, $m = a + jb$ denotes the complex refractive index (CRI) of the target atmosphere, and $k_{\alpha/\beta}(r,\lambda,m)$ represent the integral kernels, that are well described by Mie scattering theory [31].

In this work we address the challenge of reconstructing the particle size distribution $n(r)$ based on a set of indirect measurements of α and β coefficients for different values of λ. If we consider a discretization of the variable r in R points, we can rewrite the problem in a more compact matrix form

$$y = \mathbf{K}_m n + \epsilon \tag{4}$$

where:

- $y = \left[\alpha(\lambda_1), \alpha(\lambda_2), \beta(\lambda_1), \beta(\lambda_2), \beta(\lambda_3)\right]$ represents the indirect observation;
- $n = \left[n(r_1), \ldots, n(r_R)\right]$ is the discretization of the particle size distribution;
- $\mathbf{K}_m$ is a $5 \times R$ matrix representing the discretization of the integral kernels and depends on the value of the CRI;
- $\epsilon = \left[\epsilon_1, \ldots, \epsilon_5\right]$ is the noise affecting the data.

Solving the linear inverse problem defined in equation (4) for a reasonable discretization of the variable r—tipically hundreds of points—means recovering a large number of unknowns from a relatively limited dataset. Our approach to mitigate the complexity of the problem is to assume that the particle size distribution can be approximated by a theoretical distribution that can be completely described by a small set of parameters [28, 30]. A common choice [18] is to use lognormal components or superpositions of a limited number N_c of such components. In particular, we assume that the particle size distribution can be expressed as the superposition of at most three lognormal components (i.e. $N_c \in \{1,2,3\}$), in such a way that there can be at most one component representing small (or *fine*) particles ($0.1\ \mu\text{m} < \mu < 0.2\ \mu\text{m}$), one component representing medium-size particles ($0.85\ \mu\text{m} < \mu < 2.12\ \mu\text{m}$) and one component representing large (*coarse*) particles ($2.6\ \mu\text{m} < \mu < 5.5\ \mu\text{m}$); this distinction of fine, medium and coarse particles is inspired by the classical work of Whitby [36]; the actual ranges were determined in a recent statistical analysis [30]. We can therefore say that our model is a *variable-dimension* model, i.e. a collection of fixed-dimensional models. In fact, we have a collection of seven

different fixed-dimensional models originated by all possible combinations of the three components: fine, medium and coarse (we exclude the possibility that none of these components are present). We denote by $\mathcal{M} = \{\mathcal{M}_1, \ldots, \mathcal{M}_7\}$ this collection of models, where each model $\mathcal{M}_k$ represents a different combination of components, represented with a binary notation as follows: *100* is the combination that contains only the first (fine) component, *010* the combination that contains only the second (medium) component, etc up to *111* representing the combination with all three components. Models are ordered according to the following criterion: first those with fewer components, and, with the same number of components, first those with smaller particles, from $\mathcal{M}_1$ for *100* up to $\mathcal{M}_7$ for *111*. If we denote by $\mathcal{J}_k$ the nonempty index set corresponding to model $\mathcal{M}_k$ such that that the indices represent the active components, from $\mathcal{J}_1 = \{1\}$ up to $\mathcal{J}_7 = \{1,2,3\}$, we can express the particle size distribution as follows

$$n(r) = \sum_{i \in \mathcal{J}_k} \frac{h_i}{\sqrt{2\pi} r \ln \sigma_i} \exp\left(-\frac{(\ln r - \ln \mu_i)^2}{2 (\ln \sigma_i)^2} \right). \tag{5}$$

With this expedient, the number of unknowns to be estimated is drastically reduced, but we have to sacrifice the linearity of the inverse problem that can no longer be solved with standard Tikhonov regularization. In fact, if we substitute equation (5) into equations (2) and (3), we can rewrite equation (4) as follows

$$y = \sum_{i \in \mathcal{J}_k} h_i \mathbf{K}_{m_i}(\mu_i, \sigma_i) + \epsilon \tag{6}$$

where $\mathbf{K}_{m_i}(\cdot,\cdot)$ is a non-linear function of the mean and standard deviation of each lognormal component; we notice that each log-normal component has its own CRI, as different components might represent different particle populations with differing physical properties.

Earlier work [30] considered both the refractive indices and the number (and type) of lognormal components to be known *a priori*, and tackled the problem with a Bayesian approach, using a Markov chain Monte Carlo (MCMC) method to estimate the distribution parameters. In this study we make one step forward by devising an updated Bayesian model and a sophisticated Sequential Monte Carlo algorithm such that the complex refractive indices and the model need not to be fixed *a priori*, and can be estimated from the data.

3. MAFALDA: a sequential Monte Carlo algorithm

The Bayesian framework provides a valuable tool for integrating the *a priori* information, representing our knowledge before data acquisition, with the information hidden in the collected data. The *a priori* information is encoded within the *prior* distribution, while the information contained in the data is represented by the *likelihood* function.

We have already explained in section 2 that, due to the requirement of estimating the number of lognormal components from the data, our mathematical model is a variable-dimension model represented by the collection $\mathcal{M} = \{\mathcal{M}_1, \ldots, \mathcal{M}_7\}$. Let $\Theta = \cup_{k=1}^{7} (\{k\}, \Theta_k)$ be the general state space as each model $\mathcal{M}_k$ has a n_k-dimensional vector of unknown parameters $x_k \in \Theta_k$, where n_k can take different values for different values of $k = 1, \ldots, 7$. Therefore, the unknown can be expressed as $x = (k, x_k)$ reminding ourselves from equation (6) that each component possesses 5 unknown parameters: μ_i, σ_i, h_i, a_i and b_i, where a_i and b_i are respectively the real and imaginary part of the CRI m_i and the index i represents the component. Now we can define the prior $p(x)$ conditionally on the model $\mathcal{M}_k$

$$p(x) = p(x_k | \mathcal{M}_k)\, p(\mathcal{M}_k). \tag{7}$$

We assume that individual parameters are independent of each other and therefore their joint density is the product of marginal densities. We further assume that no additional information is available and adopt a uniform prior distribution for each parameter within an appropriate interval, for continuous variables, and within a given set of values, for discrete variables. In particular, appropriate intervals for the aerosol particle size were selected based on physical considerations. In summary we define the conditional prior as

$$p\left(x_k|\mathcal{M}_k\right) = \prod_{i\in\mathcal{J}_k} \mathcal{U}\left(\mu_i\right)\mathcal{U}\left(\sigma_i\right)\mathcal{U}\left(h_i\right)\mathcal{U}\left(a_i\right)\mathcal{U}\left(b_i\right) \tag{8}$$

where $\mathcal{U}(\cdot)$ denotes both the continuous and discrete uniform distribution.

As for the prior $p(\mathcal{M}_k)$, let us assume that we do not have any *a priori* information about the properties of the PM we aim to reconstruct and consider the models as equiprobable. Therefore, we adopt as well a discrete uniform distribution on all available models in $\mathcal{M}$.

Concerning the likelihood, we make the common assumption that the data is affected by additive Gaussian noise. This assumption leads us to formulate the likelihood as follows:

$$p\left(y|x\right) = \mathcal{N}\left(y; \sum_{i\in\mathcal{J}_k} h_i \mathbf{K}_{a_i+jb_i}\left(\mu_i,\sigma_i\right), \sigma_\epsilon\right) \tag{9}$$

where $\mathcal{N}(y;\mu,\sigma)$ denotes the Gaussian distribution for the variable $y \in \mathbb{R}^5$, with mean μ and standard deviation σ.

Finally we can compute the posterior distribution according to Bayes theorem

$$p\left(x|y\right) \propto p\left(y|x\right)p\left(x\right). \tag{10}$$

Since the posterior distribution encapsulates all the information about the value of the unknown parameters given the data, it represents the solution to our inverse problem as well as the greatest strength of the Bayesian approach, that always provides a quantification of uncertainty along with the estimate.

The posterior distribution defined in equation (10) is a nontrivial function on a relatively high-dimensional space: characterizing such distributions requires a tool that is able to deal with narrow peaks and local modes. In these cases Monte Carlo methods come to use, providing a sample set that can be used to compute conditional expectations, variances and other estimators. A particular class of Monte Carlo methods, named SMC samplers, are particularly good at approximating intricate distributions thanks to their sequential nature [32]. In fact, instead of directly sampling the distribution of interest, a sequence of bridge distributions is constructed in a way that it smoothly transitions from a simple distribution to the target posterior distribution. A common choice with the Gaussian likelihood is:

$$p_n\left(x|y\right) \propto p\left(x\right)p\left(y|x\right)^{\gamma_n} \quad n = 1,\ldots,N \tag{11}$$

where the likelihood is raised to a mitigating exponent that is not fixed *a priori* but is incremented adaptively on every iteration, i.e. $0 = \gamma_1 < \ldots < \gamma_N = 1$. On the practical side, each density $p_n(x|y)$ is represented by a particle approximation $\{x_n^i, W_n^i\}_{i=1}^I$, where x_n^i are the particles (samples) and W_n^i their corresponding (normalized) weights. At stage n the algorithm propagates the particles $\{x_{n-1}^i, W_{n-1}^i\}_{i=1}^I$ so that they come to represent the intermediate target density $p_n(x|y)$. Formally, the algorithm proceeds following these steps:

(i) Draw the initial particles from $p_1(x|y) = p(x)$ and set $W_1^i = \frac{1}{I}\ \forall\ i = 1,\ldots,I$.
(ii) For $n = 2,\ldots,N$:

(a) *Reweight* the particles from stage $n-1$ by defining the incremental weights

$$w_n^i = \frac{p_n\left(y|x_{n-1}^i\right)}{p_{n-1}\left(y|x_{n-1}^i\right)} \tag{12}$$

and the normalized weights

$$W_n^i = \frac{w_n^i W_{n-1}^i}{\sum_{i=1}^I w_n^i W_{n-1}^i} \tag{13}$$

(b) *Resample* (optional) the set of particles and denote the new set by $\{\hat{x}_n^i, \frac{1}{I}\}_{i=1}^I$

(c) *Propagate* the particles $\{x_{n-1}^i, W_n^i\}_{i=1}^I$—or $\{\hat{x}_n^i, \frac{1}{I}\}_{i=1}^I$ if the resampling step is performed—via one step of a reversible-jump Metropolis-Hastings (RJMH) algorithm with stationary distribution $p_n(x|y)$

Reweighting (a) is a classic importance sampling step in which particle weights are updated to reflect stage n distribution $p_n(x|y)$.

Resampling (b) is optional; it works by probabilistically discarding particles with low weights and duplicating those with high weights, yielding a new set of equally weighted particles. This noise-injection step reduces weight degeneracy and improves the accuracy of subsequent approximations. The choice of whether or not to resample is generally determined by a rule on the value of the effective sample size

$$ESS_n = \left(\sum_{i=1}^I \left(W_n^i\right)^2\right)^{-1} . \tag{14}$$

We observe that $ESS_n = I$ if all the particles have equal weights, $ESS_n = 1$ if all the weight is concentrated on a single particle. In order to balance the trade-off between adding noise and equalizing weights, we execute the resampling step when the value of ESS_n falls below $I/2$.

Finally, propagation (c) is responsible of changing the particle values, but, compared to a classic Metropolis-Hastings step, a possible domain change is performed. As with the standard Metropolis-Hastings, a Markov chain transition from a state $x = (k, x_k)$ is realized by proposing a new state $x' = (k', x'_{k'})$ from a proposal distribution $Q(x, x')$. We recall that the standard formulation of the MH algorithm is based on the characteristic of the Markov chain to be time-reversible: moves from state x to x' are made as often as moves from x' to x with respect to the target density. This is required to ensure that the stationary distribution is the desired target distribution. This condition is enforced by setting the acceptance probability as

$$\alpha\left(x, x'\right) = \min\left\{1, \frac{p_n\left(x'|y\right) Q\left(x', x\right)}{p_n\left(x|y\right) Q\left(x, x'\right)}\right\} \tag{15}$$

where, for simplicity, we use the same symbol $p_n(\cdot|y)$ for the target density at stage n whatever the dimension of the argument. On a practical level, it translates into the sequence of two Metropolis-Hastings moves: one *within* model and one *between* models. For the within-model move we adopted a sequence of Metropolis–Hastings within Gibbs moves, each one with a Gaussian proposal, which has the pleasant property $Q_w(x_k, x'_k) = Q_w(x'_k, x_k)$. For the between-models move, instead, we define a set of transition probabilities $Q_b(k, k')$ between different state spaces such that from a state space one can only move to few others that we call neighbors; and by neighbors we mean models that share at least one component, so we can move by adding a new component or removing one. In this case, the new proposed particle is constructed according to the following criterion: the common parameters are kept unchanged while the new ones (if any) are sampled from the prior distribution.

4. Validation with synthetic data

In this section we perform a numerical validation of the proposed method with three main objectives. The first one is to assess whether the Monte Carlo approximation is sufficiently stable across different runs with different set of initial particles. The second objective is to investigate how often the estimated posterior distribution contains multiple modes: by characterizing the presence of multi-modality we assess the existence of multiple alternative scenarios that are compatible with the data. The third objective is to provide an evaluation of the average reconstruction error on the particle size distribution.

The validation is carried out using synthetically generated data, in order to have an available ground truth. We generated two synthetic datasets with two distinct goals: the first serves us to test the stability of the algorithm, while the second helps us investigate the ambiguity of the problem and the quality of our estimates. The two datasets are constructed with the following procedure:

(i) *Dataset A*: for each possible model in $\mathcal{M}$ we randomly generated ten different ground truth configurations, for a total of 70 distinct configurations. For each configuration we generate noise-free synthetic data through the forward model, and then add Gaussian noise with standard deviation equal to 5% of the signal (referred to hereafter as 5% noise for simplicity), and then again with 10% of the signal (10% noise). The inverse algorithm is then run 10 times on each dataset, giving a total of 2100 runs.

(ii) *Dataset B*: for each possible model in $\mathcal{M}$ we randomly generated 100 different ground truth configurations, for a total of 700 distinct configurations. For each configuration we generate noise-free synthetic data through the forward model, and then add 5% and 10% noise, but this time each test is carried out only once, giving again a total of 2100 distinct runs.

Tests are carried out adopting $I = 10\,000$ particles and using uninformative priors on both the models and the marginals of the parameters, as described in the previous section. The likelihood standard deviation was set equal to the noise standard deviation when the latter is non-zero; for noise-free data we set the likelihood standard deviation to 0.5% of the signal. The CRI m is discretized adopting 15 values for the real part a, linearly distributed from a minimum of 1.25 and a maximum of 1.81, and 15 values for the imaginary part b, distributed on a log-scale from a minimum of 0.0005 and a maximum of 0.146, giving 225 possible values for m. This approximation was convenient for computational reasons; indeed, keeping this variable continuous would require a tremendous computational effort since the matrix $\mathbf{K}_m$ varies significantly with m, and it is a very large, full matrix. If we allowed the algorithm to continuously explore the domain of the CRI, at each step we would have to recompute the matrix and it would be computationally unfeasible. So the algorithm can only jump between discrete values for a and b, and for every possible combination of these two values the matrix $\mathbf{K}_m$ has already been calculated and stored.

With these settings, approximating one posterior distribution with MAFALDA takes about 2 min on a MacBook Pro with M2 processor and 16 GB of RAM.

4.1. Stability of the algorithm

As the inverse problem we are tackling is ill-posed, the posterior distribution is a complex distribution in a relatively high dimensional space; Monte Carlo sampling of such distributions can be difficult, as samples may get stuck in local maxima, so that different runs may provide

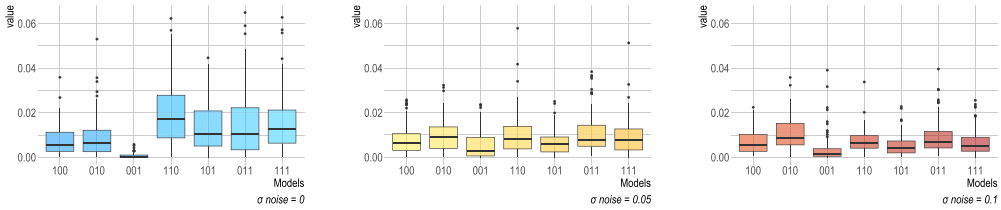

Figure 1. Stability of model inference using the one vs all test: from left to right the results with noise-free data (in blue), 5% noise (in yellow) and 10% noise (in red).

different results. As a first test we then check the stability of the algorithm, i.e. its capability of always providing the same posterior distribution as output. In order to simplify the comparison, we break the problem into several parts and choose metrics that can allow us to evaluate the distance between samples:

- *Stability of model inference*: we test whether the marginal posterior probabilities of model $\mathcal{M}$ remain stable across different runs on the same data;
- *Stability of parameter inference*, we check whether the marginal posteriors of all the parameters μ, σ, h and m, conditional on the best model, i.e. the MAP point of $\mathcal{M}$, are also stable across different runs.

We recall that the tests in this section were performed on *Dataset A*.

4.1.1. Stability of model inference. To measure the distance between samples we adopt the Kolmogorv–Smirnov statistic, a number between 0 and 1 which represents the maximum distance along the y-axis between the two cumulative distributions [37]; we will refer to it as KS statistic or KS distance. When used to compare two populations of discrete values (i.e. the labels representing the models) the value of the KS statistic corresponds to the number of labels that have to be changed from one or the other population to make them match each other. We performed two types of test:

- *one vs one*, comparing marginal distributions across all possible pairs of runs on the same data (10 runs, 45 comparisons)
- *one vs all*, comparing marginal distributions of each run with the average across 10 runs (10 comparisons)

In figure 1 we plot the empirical distribution of the Kolmogorov–Smirnov distances

$$d_{KS}^{l,k} = d_{KS}\left(p\left(\mathcal{M}^{l,k}|y\right), p\left(\overline{\mathcal{M}^{k}}|y\right)\right), \tag{16}$$

for all runs $l = 1,\dots,100$ (10 repetitions of 10 different tests) and true model cases $k = 1,\dots,7$, where $\mathcal{M}^{l,k}$ denotes the model random variable from the run l under the real model case k, and $\overline{\mathcal{M}^{k}}$ is the corresponding model random variable averaged over all runs. No significant differences are observed between this *one vs all* comparison and the analogous *one vs one* case (apart from a larger scale of distances), so the latter plot is omitted for brevity.

We observe that increasing the noise in the data decreases the variability between different runs. Since for all models and all levels of noise average values of KS statistic are less than 1-2% we can consider the algorithm to be very stable in the model choice.

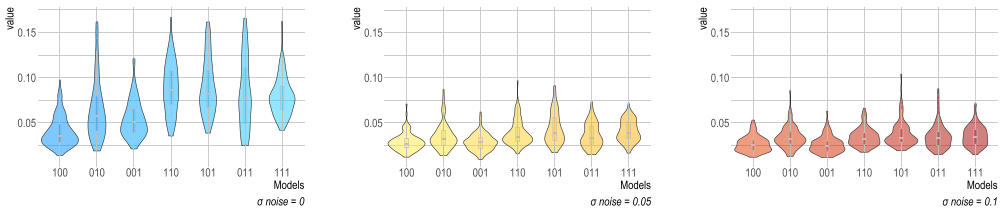

Figure 2. Stability of parameter inference using the one vs all test: from left to right the results with noise-free data (in blue), 5% noise (in yellow) and 10% noise (in red).

4.1.2. Stability of parameter inference. In figure 2 we plot the empirical distribution of the maximum Kolmogorov–Smirnov distances over parameters, defined as follows. Let $\widehat{\mathcal{M}}^{l,k}$ be the maximum *a posteriori* (MAP) model in run l for real-model case k. If $\widehat{\mathcal{M}}^{l,k}$ is the same for all l at fixed k, we keep all runs; otherwise we retain only those runs with $\widehat{\mathcal{M}}^{l,k}$ equal to the most frequent MAP (in the worst case two runs are discarded). Let $\widehat{k}$ such that $\widehat{\mathcal{M}}^{l,k} = \mathcal{M}_{\widehat{k}}$ and denote $x_{\widehat{k}} = \{\theta_j\}_{j=1}^{n_{\widehat{k}}}$, then for each pair (l,k) and each $j = 1,\dots,n_{\widehat{k}}$ we compute similarly to the model inference case

$$d_{\mathrm{KS}}^{l,k,j} = d_{\mathrm{KS}}\left(p\left(\theta_j^{l,k}|y,\mathcal{M}_{\widehat{k}}\right), p\left(\overline{\theta}_j^{k}|y,\mathcal{M}_{\widehat{k}}\right)\right). \tag{17}$$

Then we define

$$D^{l,k} = \max_{j=1,\dots,n_{\widehat{k}}} d_{\mathrm{KS}}^{l,k,j}. \tag{18}$$

The figure shows the distribution of $D^{l,k}$ across all retained runs and real-model cases.

As already observed in the previous subsection, higher levels of noise lead to reduced variability across the estimated marginal distributions. While this behavior may appear counterintuitive, it is actually very reasonable: when no or little noise is affecting the data, the posterior distribution has narrow peaks, and little changes in the parameter values lead to large changes in the KS statistics; when noise is larger, the posterior distribution has wider peaks and differences get smaller.

We notice that the KS statistics on the marginal distributions of individual parameters tend to have larger values than those obtained on the marginal distributions of models. By inspecting and comparing these marginal distributions in different runs, we observed that this larger KS distance is typically due to the distribution being multi-modal and the different approximations assigning different weights to the different modes. We also observed that when marginal distributions on individual parameters obtained by different runs do not coincide, the high probability regions (and the peaks) are typically the same but their overall probability varies. In other words, the algorithm provides multiple potential solutions with different probabilities; the potential solutions are the same each time, but their relative probabilities change in different runs.

4.2. Ambiguity of the data

Here we try to assess how complex the problem is, that is, how often the posterior distribution contains more than one mode. As we did previously to investigate the stability of the algorithm, here we also need to break the problem into two parts and separately analyze

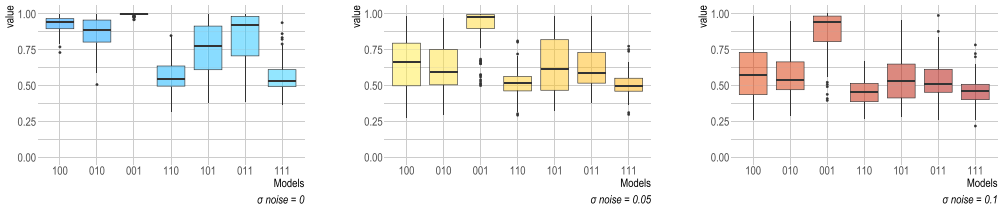

Figure 3. Weight associated with the MAP, as the model on which most of the weight is concentrated, representing ambiguity in the model choice: from left to right the results with noise-free data (in blue), 5% noise (in yellow) and 10% noise (in red).

- *Multimodality in model inference*, we check how often the algorithm assigns non-zero posterior probability to more than one model;
- *Multimodality in parameter inference*, we check whether multiple peaks are present in the posterior marginal distributions of the parameters μ, σ, h, and m.

We recall that the tests in this section were performed on *Dataset B*.

4.2.1. Multimodality in model inference. In order to provide a compact visualization of multimodality in model inference, in figure 3 we report the boxplot of the marginal probability associated to the MAP model: unimodality corresponds to the case where the probability of the MAP is one; values lower than one correspond to multimodality in model inference. We observe that multimodality in model inference increases with noise, as the likelihood becomes more tolerant to small perturbations of the data. However, such increase is not uniform across models: for instance, model *100* and model *001* have similar boxplots in absence of noise but behave differently when noise increases. This non-uniform behavior arises because the fine, medium, and coarse components contribute measurements of markedly different magnitudes. Specifically for the extinction coefficient, the coarse component yields values of smaller magnitude, the medium component yields values of intermediate magnitude, and the fine component yields values of larger magnitude. Consequently, true configurations containing only the coarse component cannot be confused with those containing medium or fine components, since the latter produce measurements that are orders of magnitude larger. Conversely, data generated by configurations with only the fine component may be mistaken for data from configurations containing both fine and coarse components, because adding the coarse component perturbs the measurements only slightly relative to the dominant fine-component signal.

As a consequence, configurations belonging to different models may produce very similar likelihoods. To confirm this fact, in figure 4 we take the results of a single run with a multimodal posterior, and plot the likelihoods of all particles grouped by models: the plot shows that the likelihoods of particles with different number of components are perfectly comparable.

4.2.2. Multimodality in parameter inference. To address this type of multimodality we need to check for multiple peaks in the marginal distribution of each parameter, and for this purpose we adopted Silverman's statistical test of unimodality vs. multimodality [38]. In figure 5 we show how often the posterior distribution is multimodal (for at least one parameter): contrary to what we saw before for multimodality in model inference, multimodality in parameter inference is observed much more in absence of noise. This is likely due to the fact that even relatively low levels of noise tend to make the posterior much flatter, reducing multimodality.

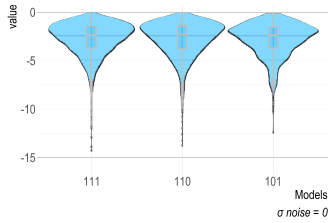

Figure 4. Log-likelihood of the particles belonging to different models at the end of an ambiguous run on a noise-free data from a ground truth with all three components: from left to right the model *111* (≈70% of the total weight), the model *110* (≈20%) and the model *101* (≈10%).

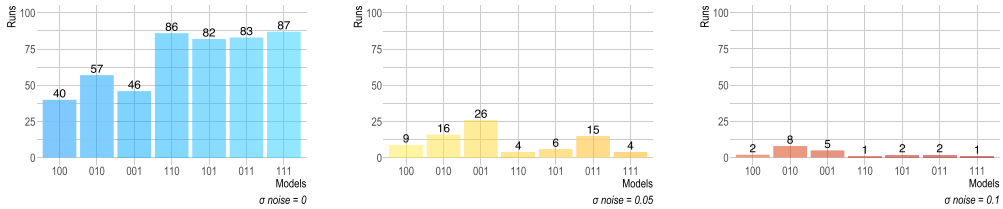

Figure 5. Number of runs in which Silverman's test detects more than one peak in the marginal posterior of at least one of the parameters: from left to right the results with noise-free data (in blue), 5% noise (in yellow) and 10% noise (in red).

Figure 6 shows an example of multimodality in a case where the data is not affected by noise and the ground truth is composed of only one fine component.

The algorithm concentrates all the probability on the correct model (there is no *multimodality in model inference*) but finds several distinct solutions for μ and h, each associated to a distinct value of the real part of the CRI: in particular, the left-most peaks in the distribution of μ and h correspond to the right-most peak in the distribution of a, and viceversa.

We also notice that the multimodality we observe in the posterior distribution of μ and h might be due to the use of discretized values for the CRI m, and would probably disappear if we used a continuous variable instead.

4.3. Quality of the estimate

The aim of this section is to investigate the performance of the algorithm in reconstructing a known ground truth starting from exact and noisy data. As before we break the problem into two parts:

- *Model selection*: we check how many times the MAP model coincides with the actual model;
- *Distribution reconstruction*: we give a measure of how far from the ground truth is our reconstruction of the particle size distribution.

We recall that the tests in this section were performed on *Dataset B*.

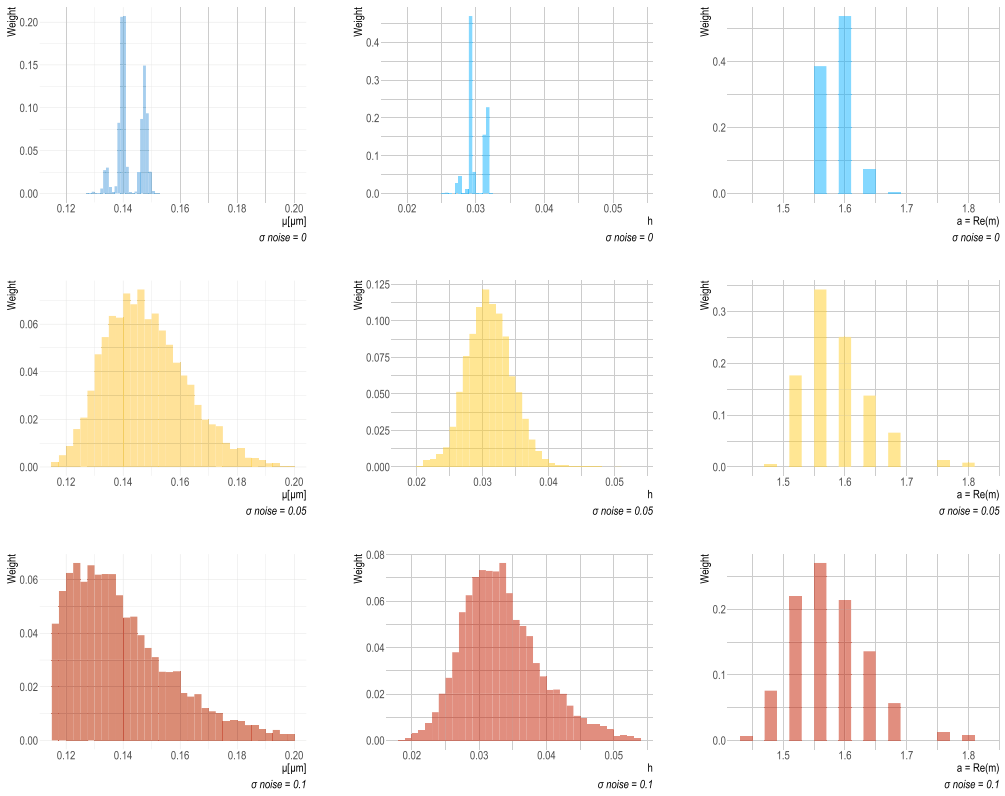

Figure 6. Example of multimodality in the marginal posteriors at the end of a run on a noise-free data from a ground truth composed of only one fine component, and comparison with noisy cases where multimodality is no longer observed in the second (5% noise) and third (10% noise) rows: from left to right the parameters μ, h and a respectively.

4.3.1. Model selection. Regarding the first part of our problem, we can assess the quality of our estimates looking at the confusion matrices in figure 7. We observe that in the absence of noise the algorithm perfectly identifies models with only one component but the accuracy worsens as the noise increases, except for model *001*. This uneven behavior is due to the fact that, as already explained in the previous Subsection, the data produced by the coarse component are considerably smaller than the data produced by the fine and medium components; therefore, when the true component is fine or medium, a coarse component is added because it helps to explain noise; on the other hand, when the true component is coarse, the order of magnitude of noise (which is added as a percentage of the true signal) is not compatible with the presence of a fine or medium component, as these would create a much stronger signal.

We also notice that the confusion matrices we observe are less symmetric than one would expect. In principle, if the model $\mathcal{M}_i$ has high posterior probability when the true model is $\mathcal{M}_j$, we expect also $\mathcal{M}_j$ to have high posterior probability when the true model is $\mathcal{M}_i$. In our results such symmetry is not always observed: for instance, in the noise-free data we have 24 cases where model *011* is estimated as *010*, but only 1 case in which the opposite is observed. This asymmetry is partly due to random effects in the data generation, and possibly partly due

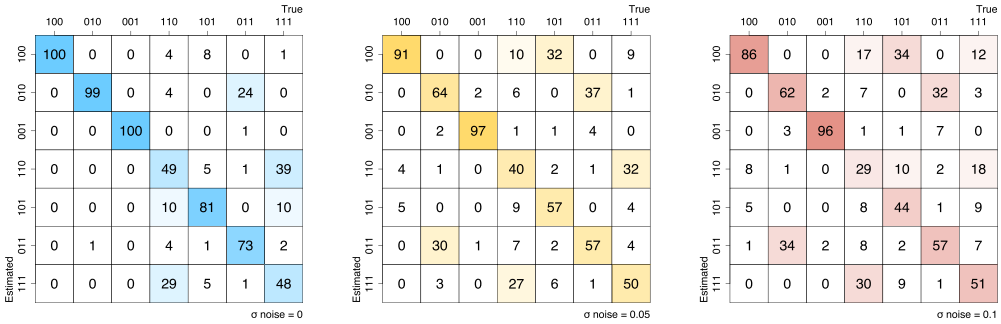

Figure 7. Confusion matrices representing how many times the model with the most weight at the end of a run (maximum *a posteriori*) coincides with the ground truth model: from left to right the results with noise-free data (in blue), 5% noise (in yellow) and 10% noise (in red).

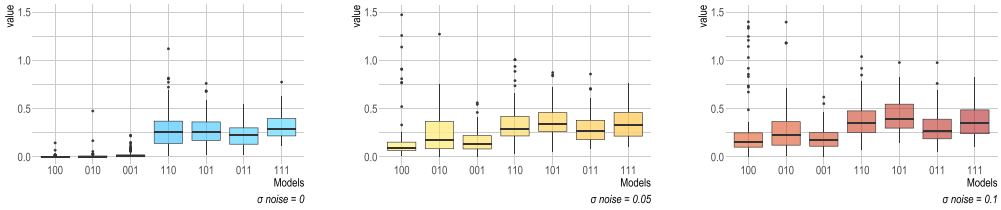

Figure 8. Reconstruction error computed as the integral of the difference between reconstructed and real distributions divided by the integral of the real one: from left to right the results with noise-free data (in blue), 5% noise (in yellow) and 10% noise (in red).

to a tendency of the algorithm to favor simpler models, despite the flat prior on models, in accordance to the celebrated Occam's razor naturally embodied by a Bayesian approach.

Finally, let us comment that the confusion matrices only concern the MAP estimate; in addition to this, we went on to check how often the true model has non-zero posterior probability (even though is not the selected model) and confirmed that this happened in all of our tests.

4.3.2. Distribution reconstruction. Now we address the second part of our problem and proceed to evaluate the quality of the reconstruction conditional on the selected model, i.e. the MAP point of $\mathcal{M}$. To quantify the reconstruction error we use the relative error, defined as follows:

$$E = \frac{\int |\widehat{n}(r) - n(r)|\,\mathrm{d}r}{\int n(r)\,\mathrm{d}r} \tag{19}$$

where $\widehat{n}(r)$ is the median particle size distribution. Figure 8 shows that the reconstruction error increases with noise. The increase is more pronounced for ground truths with only one component: in this case the error is close to zero for noise-free data and can occasionally get as high as 100% with noisy data. For more complex configurations the reconstruction error is non-negligible even with noise-free data, and grows a little further with noise.

In order to exemplify these results, in figure 9 we report the reconstructions obtained on three different configurations, for three levels of noise. In this figure the solid line represents the

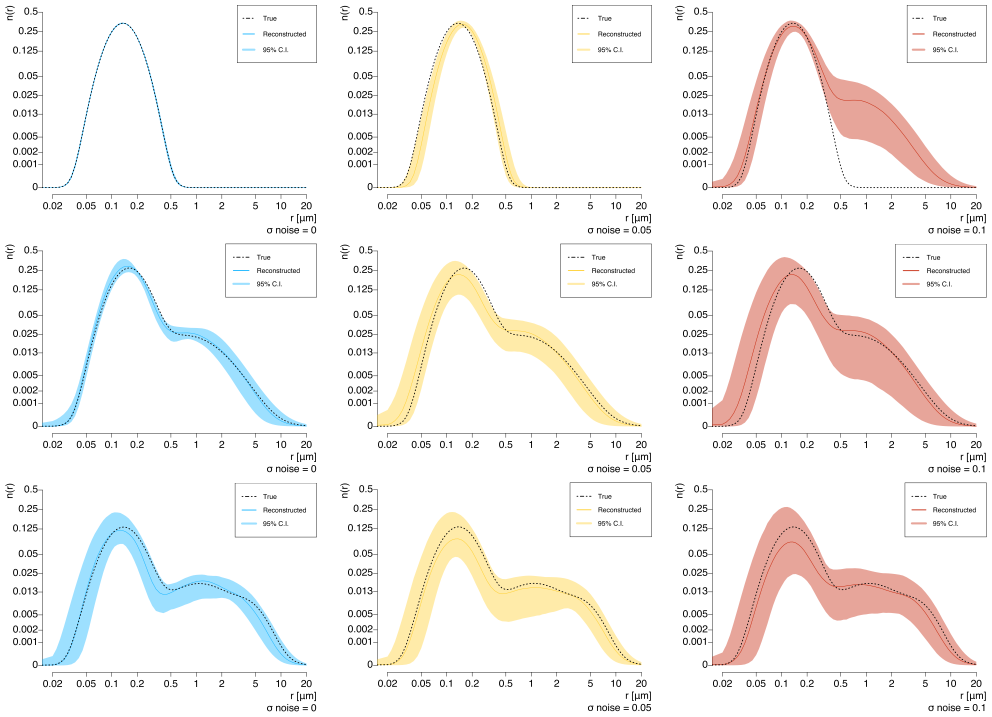

Figure 9. Representation of the true particle size distribution (black dotted line) with the reconstruction proposed by the algorithm (solid colored line) and an evaluation of its uncertainty (light-colored band). The ground truth contains one component in the first row, two components in the second row and three components in the third row. From left to right the results with noise-free data (in blue), 5% noise (in yellow) and 10% noise (in red).

median particle size distribution, the light-colored band represents the 95% credible interval, and the dotted black line the ground truth.

We observe that the reconstruction of a ground truth with a single fine component is very accurate when the data is noise-free; as noise increases the uncertainty band increases and the estimate deviates slightly from the true distribution, to the point that with 10% noise the algorithm adds a second (medium) component to account for the noise. In this latter case the estimate of the fine component remains accurate and with small uncertainty while the component that is added has a lot of variability, testifying its low impact on the data. We also notice that the 95% credible interval in this case does not contain the true distribution; this seems unpleasant, as it might suggest that the algorithm assigns zero or negligible probability to the exact solution. In fact, in this case the figure reports the credible interval for the MAP model, that is *110*; with a lower probability ($\simeq$21%), however, the algorithm also retrieves a *100* model: in figure 10 we show the credible interval for this case, that actually contains the true distribution.

When the ground truth has more than one component the uncertainty band is already fairly wide in the noise-free case; this is mostly due to the fact that we have either 10 (2 components) or 15 (3 components) unknowns and only 5 data points, leading to considerable non-uniqueness.

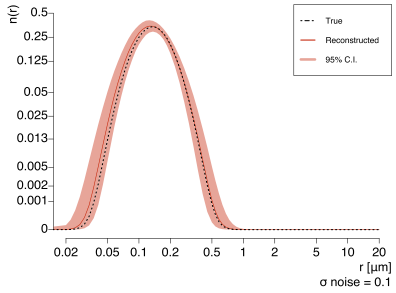

Figure 10. Median and credibility interval conditional on model *100* (with posterior probability $\simeq 21\%$) from the same posterior distribution shown in the top right corner of figure 9: in this case the true distribution falls within the credibility interval.

5. Application to experimental data

In this section we report the results of tests conducted with our algorithm on the two following sets of experimental lidar measured at the Naples (Italy) remote sensing national facility of the ACTRIS research infrastructure. Measurements refer to two events of particular aerosol load in the atmosphere:

(i) *4 September 2017*: intense fires in British Columbia (Canada) released an exceptionally high concentration of biomass burning aerosols into the atmosphere during August 2017, and these aerosols were transported by winds throughout Europe in the following weeks [39];
(ii) *25 February 2021*: a major Saharan dust outbreak and a simultaneous eruption of Mount Etna injected in the atmosphere a high concentration of dust and volcanic aerosol that were exceptionally transported towards the north [40].

In both cases the extinction and backscattering coefficients at different wavelengths (in the classical $3\beta+2\alpha$ configuration discussed in section 2) together with their corresponding uncertainties were first retrieved from lidar observations using recognized lidar retrieval procedures [13–15]. Then MAFALDA was applied to produce estimates of the particle size distribution using:

- 100 000 particles in order to guarantee stability;
- a uniform prior on the models;
- the estimated uncertainties in the optical coefficients as the noise standard deviation in the likelihood.

As a benchmark for comparison we use the reconstructions derived from the AERONET [41] Sun-sky-lunar photometer data, referred to as 'AERONET' from here on. The photometer gathers day-time and night-time radiance measurements at eight different wavelengths. By applying a set of standard algorithms [42, 43] a wide range of aerosol optical and microphysical properties—including particle size distribution—can be retrieved.

Before proceeding, it is best to make a couple of considerations:

- reconstructions from the Sun photometer data are only available during the day hours, whereas the lidar data were collected at night. As a result, small differences between our

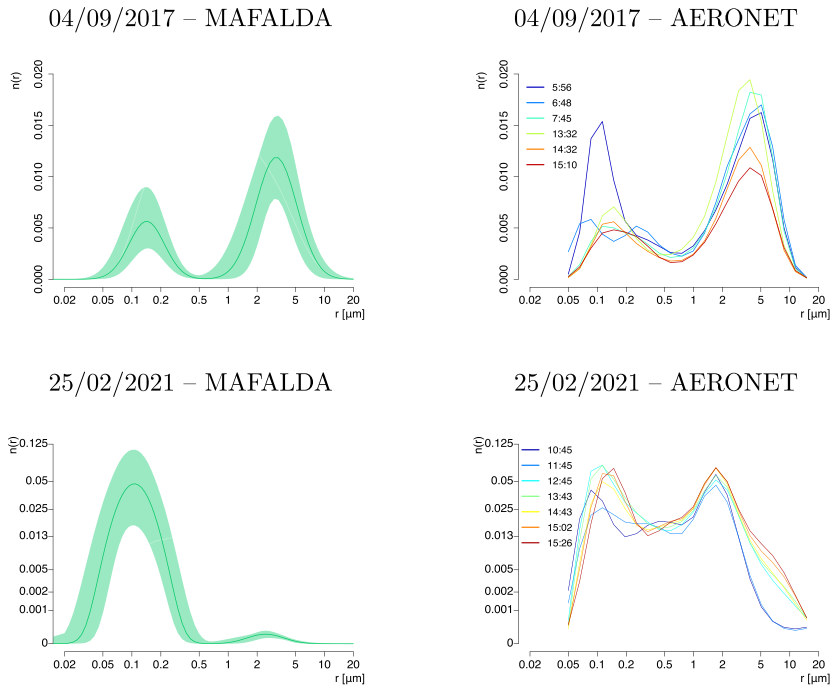

Figure 11. Test results on 2017 (first row) and 2021 (second row) data: on the left the estimated particle size distribution proposed by MAFALDA (solid green line) with an evaluation of its uncertainty (light-colored band), on the right the reconstructions proposed by AERONET at different times of the day.

distribution and AERONET's are likely to be observed, as they represent data from different times of day.

- The Sun-photometer, using solar radiation as an external light source, retrieves information on atmospheric composition across the entire investigated vertical column. The vertical distributions of aerosol concentration are derived from MERRA-2 [44] global assimilation model simulations, which incorporate CALIOP [45] aerosol vertical profile measurements. The lidar system used in this study has an atmospheric probing range of up to 30 km. During the observations, the aerosol under investigation did not extend beyond this altitude, nor were any extreme events reported that could lead to a high-density aerosol presence in the upper troposphere and stratosphere, potentially affecting the photometer's measurements

Figure 11 shows the comparison between the estimated particle size distribution provided by MAFALDA (left) and that proposed by AERONET (right). We recall that in the left plots the solid line represents the median particle size distribution while the light-colored band represents the 95% credible interval.

In particular, the first row shows that the reconstructions provided by MAFALDA and AERONET for the 2017 data are in strong agreement in correspondence with the peaks of the particle size distributions: the left peak is estimated to be lower than the right peak in both cases. The two estimates differ a little in between the two peaks: here the MAFALDA estimate is slightly lower than that provided by AERONET.

In the second row we report the results obtained on the 2021 data. At a first glance the estimate provided by MAFALDA seems quite different from that provided by AERONET. However, at a more careful look one can observe that the distribution estimated by MAFALDA contains two components, the second one being considerably smaller than the first one (notice the logarithmic scale on the vertical axis), but peaking roughly at the same value of r (about 2 μm). The main difference is then the height of this second component, which appears to be much larger in the Sun photometer data than in the lidar data. Further investigation would be needed to shed light on this difference, which may also be a consequence of the mentioned difference in recording times; however, we believe also in this case there is reasonably good agreement between the two approaches.

6. Discussion

Retrieving the particle size distribution from lidar data requires to solve a difficult inverse problem with little data and much uncertainty. In this article we presented a Bayesian model and a Monte Carlo algorithm, MAFALDA, that advances the state-of-the-art in the resolution of this problem.

The mathematical model is built on a flexible parametric framework, representing the particle size distribution as a superposition of up to three log-normal components. Unlike recent approaches in the field [30], our method allows for automated model selection, thus eliminating the need to predefine the number of components or fix the complex refractive indices in advance—drastically reducing external assumptions and subjective choices. This added flexibility enriches the posterior distribution, making it more intricate, often multimodal, and inherently challenging to approximate. To tackle this complexity, we introduced a cutting-edge Monte Carlo algorithm, belonging to the family of Sequential Monte Carlo methods: by employing multiple Markov Chains running in parallel and targeting increasingly complex distributions, these methods allow to identify and maintain multimodality in the sample set and avoid getting stuck in local maxima.

To validate the proposed approach we performed numerical tests on both synthetic and real data of various complexities containing one, two and even three distinct particle size distribution components and with various levels of noise. The measurement configuration was always the standard $3\beta+2\alpha$.

Numerical simulations on synthetic data confirmed that the posterior distribution is often multimodal, in some cases even containing substantially separated modes. Pleasantly enough the presented algorithm MAFALDA was capable of maintaining and representing these multiple modes reliably: running the algorithm multiple times on the same data provided very similar approximations to the posterior distribution, as testified by the small values of the KS distance. This implies that the algorithm does not appear to get stuck in local maxima, but has instead good mixing properties. Importantly, even when used to compute point-estimates of the particle size distribution MAFALDA performed well, providing overall relative errors often much smaller than 30%, which are considered to be more than reasonable for this problem [46].

Application of MAFALDA on two sets of experimental data also showed very good agreement with the accepted state of the art, represented in this case by reconstructions obtained from a different instrument. This confirms that MAFALDA can be reliably used on the field.

In conclusion, we believe that the proposed approach has shown to have the potential to provide reliable estimates of the particle size distribution and its uncertainty even under complicated conditions.

There are of course several ways to further improve our results. A few of them would be:

(i) overcoming Gibbs sampling and letting the algorithm propose moves for particles that simultaneously change all the parameters of a log-normal component in a correlated way. Indeed, it can be seen that there is a strong correlation: increasing values of a correspond to decreasing values of μ and h.
(ii) Exploit the conditionally linear-Gaussian structure of the data with respect to the h parameters, leading to a so-called *Rao-Blackwellization*; this can improve the performances of the algorithm [47, 48];
(iii) devise a dynamical model to combine data at different heights, in such a way to provide altitude-resolved estimates of the particle size distribution that embody continuity.

These further improvements will be the subject of future studies.

Data availability statement

The data cannot be made publicly available upon publication because no suitable repository exists for hosting data in this field of study. The data that support the findings of this study are available upon reasonable request from the authors.

Acknowledgment

The research by A S and G V was supported in part by the MIUR Excellence Department Project awarded to Dipartimento di Matematica, Universitá di Genova, CUP D33C23001110001 AS kindly acknowledges Gruppo Nazionale per il Calcolo Scientifico (GNCS) of Istituto Nazionale di Alta Matematica (INdAM) for partial financial support. A B, A S and R D have been partially supported by The European Union's Horizon 2020 Framework Program for Research and Innovation (Grant Agreement No. 654109 - ACTRIS-2 Aerosols, Clouds, and Trace Gases Research InfraStructure).

Code availability statement

The source code developed in this study is available at https://github.com/giacomovarini/mafalda.

Conflict of interest

All authors declare no conflicts of interests.

ORCID iDs

Giacomo Varini 0009-0004-1164-5020
Alberto Sorrentino 0000-0003-3457-6780

References

[1] Giannakaki E, Balis D S, Amiridis V and Zerefos C 2010 Optical properties of different aerosol types: seven years of combined Raman-elastic backscatter lidar measurements in Thessaloniki, Greece *Atmos. Meas. Tech.* **3** 569–78

[2] Lee K H and Wong M S 2018 Vertical profiling of aerosol optical properties from lidar remote sensing, surface visibility and columnar extinction measurements *Remote Sensing of Aerosols, Clouds and Precipitation* (Elsevier) pp 23–43

[3] Ritter C *et al* 2018 Microphysical properties and radiative impact of an intense biomass burning aerosol event measured over Ny-ålesund, Spitsbergen in July 2015 *Tellus B* **70** 1–23

[4] Stelitano D, Di Girolamo P, Scoccione A, Summa D and Cacciani M 2019 Characterization of atmospheric aerosol optical properties based on the combined use of a ground-based Raman lidar and an airborne optical particle counter in the framework of the hydrological cycle in the Mediterranean experiment–special observation period 1 *Atmos. Meas. Tech.* **12** 2183–99

[5] Chazette P 2020 Aerosol optical properties as observed from an ultralight aircraft over the strait of Gibraltar *Atmos. Meas. Tech.* **13** 4461–77

[6] Fernald F G 1984 Analysis of atmospheric lidar observations: some comments *Appl. Opt.* **23** 652–3

[7] Weitkamp C 2006 *Lidar: Range-Resolved Optical Remote Sensing of the Atmosphere* vol 102 (Springer)

[8] Measure R M 1984 *Laser Remote Sensing: Fundamentals and Applications* (Wiley)

[9] Flamant P H 2005 Atmospheric and meteorological lidar: from pioneers to space applications *C. R. Physique* **6** 864–75

[10] Killinger D K and Mooradian A 2013 *Optical and Laser Remote Sensing* vol 39 (Springer)

[11] Boselli A, Pisani G, Spinelli N and Wang X 2014 Laser remote sensing for environmental applications *Photonics for Safety and Security* (World Scientific) pp 175–205

[12] Dong P and Chen Q 2017 *Lidar Remote Sensing and Applications* (CRC Press)

[13] Klett J D 1981 Stable analytical inversion solution for processing lidar returns *Appl. Opt.* **20** 211–20

[14] Ansmann A, Riebesell M and Weitkamp C 1990 Measurement of atmospheric aerosol extinction profiles with a Raman lidar *Opt. Lett.* **15** 746–8

[15] Ansmann A, Riebesell M, Wandinger U, Weitkamp C, Voss E, Lahmann W and Michaelis W J A P B 1992 Combined Raman elastic-backscatter lidar for vertical profiling of moisture, aerosol extinction, backscatter and lidar ratio *Appl. Phys. B* **55** 18–28

[16] Pornsawad P, Böckmann C, Ritter C and Rafler M 2008 Ill-posed retrieval of aerosol extinction coefficient profiles from Raman lidar data by regularization *Appl. Opt.* **47** 1649–61

[17] Pornsawad P, D'Amico G, Böckmann C, Amodeo A and Pappalardo G 2012 Retrieval of aerosol extinction coefficient profiles from Raman lidar data by inversion method *Appl. Opt.* **51** 2035–44

[18] Osterloh L, Böckmann C, Nicolae D and Nemuc A 2013 Regularized inversion of microphysical atmospheric particle parameters: theory and application *J. Comput. Phys.* **237** 79–94

[19] Garbarino S, Sorrentino A, Massone A M, Sannino A, Boselli A, Wang X, Spinelli N and Piana M 2016 Expectation maximization and the retrieval of the atmospheric extinction coefficients by inversion of Raman lidar data *Opt. Express* **24** 21497–511

[20] Denevi G, Garbarino S and Sorrentino A 2017 Iterative algorithms for a non-linear inverse problem in atmospheric lidar *Inverse Problems* **33** 085010

[21] Veselovskii I, Kolgotin A, Griaznov V, Müller D, Wandinger U and Whiteman D N 2002 Inversion with regularization for the retrieval of tropospheric aerosol parameters from multiwavelength lidar sounding *Appl. Opt.* **41** 3685–99

[22] Pérez-Ramírez D, Whiteman D N, Veselovskii I, Kolgotin A, Korenskiy M and Alados-Arboledas L 2013 Effects of systematic and random errors on the retrieval of particle microphysical properties from multiwavelength lidar measurements using inversion with regularization *Atmos. Meas. Tech.* **6** 3039–54

[23] Granados-Muñoz M J *et al* 2014 Retrieving aerosol microphysical properties by lidar-radiometer inversion code (liric) for different aerosol types *J. Geophys. Res.: Atmos.* **119** 4836–58

[24] Giannakaki E, van Zyl P G, Müller D, Balis D and Komppula M 2016 Optical and microphysical characterization of aerosol layers over South Africa by means of multi-wavelength depolarization and Raman lidar measurements *Atmos. Chem. Phys.* **16** 8109–23

[25] Müller D, Böckmann C, Kolgotin A, Schneidenbach L, Chemyakin E, Rosemann J, Znak P and Romanov A 2016 Microphysical particle properties derived from inversion algorithms developed in the framework of earlinet *Atmos. Meas. Tech.* **9** 5007–35

[26] Chemyakin E, Burton S, Kolgotin A, Müller D, Hostetler C and Ferrare R 2016 Retrieval of aerosol parameters from multiwavelength lidar: investigation of the underlying inverse mathematical problem *Appl. Opt.* **55** 2188–202
[27] Ortiz-Amezcua P *et al* 2017 Microphysical characterization of long-range transported biomass burning particles from North America at three earlinet stations *Atmos. Chem. Phys.* **17** 5931–46
[28] Benavent-Oltra J A *et al* 2019 Different strategies to retrieve aerosol properties at night-time with the grasp algorithm *Atmos. Chem. Phys.* **19** 14149–71
[29] Molero F, Pujadas M and Artíñano B na 2020 Study of the effect of aerosol vertical profile on microphysical properties using grasp code with Sun/sky photometer and multiwavelength lidar measurements *Remote Sens.* **12** 4072
[30] Sorrentino A, Sannino A, Spinelli N, Piana M, Boselli A, Tontodonato V, Castellano P and Wang X 2022 A Bayesian parametric approach to the retrieval of the atmospheric number size distribution from lidar data *Atmos. Meas. Tech.* **15** 149–64
[31] Bohren C F and Huffman D R 2008 *Absorption and Scattering of Light by Small Particles* (Wiley)
[32] Moral P D, Doucet A and Jasra A 2006 Sequential Monte Carlo samplers *J. R. Stat. Soc. B* **68** 411–36
[33] Sorrentino A, Luria G and Aramini R 2014 Bayesian multi-dipole modelling of a single topography in MEG by adaptive sequential Monte Carlo samplers *Inverse Problems* **30** 045010
[34] Vivaldi V and Sorrentino A 2016 Bayesian smoothing of dipoles in magneto-/electroencephalography *Inverse Problems* **32** 045007
[35] Weitkamp C *et al* 2005 *Range-Resolved Optical Remote Sensing of the Atmosphere* vol 102 (Springer) pp 241–303
[36] Whitby K T 1978 The physical characteristics of sulfur aerosols *Sulfur in the Atmosphere* (Elsevier) pp 135–59
[37] Daniel Wayne W 1990 *Applied Nonparametric Statistics* 2nd edn (PWS–KENT Publishing Company) pp xii + 635
[38] Ameijeiras-Alonso J, Crujeiras R M and Rodríguez-Casal A 2019 Mode testing, critical bandwidth and excess mass *Test* **28** 900–19
[39] Damiano R, Amoruso S, Sannino A and Boselli A 2024 Lidar optical and microphysical characterization of tropospheric and stratospheric fire smoke layers due to Canadian wildfires passing over Naples (Italy) *Remote Sens.* **16** 538
[40] Sannino A, Amoruso S, Damiano R, Scollo S, Sellitto P and Boselli A 2022 Optical and microphysical characterization of atmospheric aerosol in the central mediterranean during simultaneous volcanic ash and desert dust transport events *Atmos. Res.* **271** 106099
[41] Holben B N *et al* 1998 Aeronet-a federated instrument network and data archive for aerosol characterization *Remote Sens. Environ.* **66** 1–16
[42] Dubovik O and King M D 2000 A flexible inversion algorithm for retrieval of aerosol optical properties from Sun and sky radiance measurements *J. Geophys. Res.: Atmos.* **105** 20673–96
[43] Holben B N *et al* 2001 An emerging ground-based aerosol climatology: aerosol optical depth from aeronet *J. Geophys. Res.: Atmos.* **106** 12067–97
[44] Gelaro R *et al* 2017 The modern-era retrospective analysis for research and applications, version 2 (merra-2) *J. Clim.* **30** 5419–54
[45] Winker D M, Hunt W H and McGill M J 2007 Initial performance assessment of caliop *Geophys. Res. Lett.* **34** 135
[46] Raut J-C and Chazette P 2009 Assessment of vertically-resolved pm 10 from mobile lidar observations *Atmos. Chem. Phys.* **9** 8617–38
[47] Sommariva S and Sorrentino A 2014 Sequential Monte Carlo samplers for semi-linear inverse problems and application to magnetoencephalography *Inverse Problems* **30** 114020
[48] Viani A, Luria G, Bornfleth H and Sorrentino A 2021 Where Bayes tweaks gauss: conditionally Gaussian priors for stable multi-dipole estimation *Inverse Problems Imaging* **15** 1099