ENSEMBLE MARKOV CHAIN MONTE CARLO METHOD FOR ASSESSING UNCERTAINTIES OF AEROSOL PROPERTIES FROM MULTI-WAVELENGTH LIDAR MEASUREMENTS

Benjamin R. Herman, Barry Gross, Fred Moshary, and Samir Ahmed

City College of New York, Electrical Engineering Dept., Convent Ave. at 140th Street, New York, NY 10031 USA,
Email: herman@ee.ccny.cuny.edu, gross/moshary/ahmed@ccny.cuny.edu

ABSTRACT

An ensemble Markov chain Monte Carlo method of assessing uncertainty of aerosol properties from lidar measurements of extinction and backscatter is presented. The method applies the Metropolis-Hastings algorithm to an ensemble of Markov chains. Candidates are drawn from a hybrid random walk/independence sampler random generator. The independence sampler is formed by analyzing the ensemble partway along the chain to find regions of interests in which to generate random candidates. Convergence to the target Bayesian posterior probability density function (PDF) was found to be greatly expedited by also including sampling from the prior PDF. The Kolmogoroff-Smirnov test was applied to the evolving ensemble to verify convergence.

1. INTRODUCTION

The body of research on retrieving aerosol size distributions from backscatter and extinction coefficients estimated from multi-wavelength lidar and sun photometer measurements has been oriented towards devising algorithms that produce point estimates of size distribution along with complex index of refraction. Uncertainty has usually been assessed from the point of view of how much the retrieval deviates from the true aerosol [1]. However, it has been shown that using this approach to assess uncertainty *a posteriori* results in uncertainty assessments that significantly differ from Bayesian assessments, and that these discrepancies arise when any of three non-ideal conditions are present [2]. These conditions are: 1) non-linear relationship between aerosol model parameters and coefficients, 2) non-uniform prior weighting functions, and 3) dependency of error statistical properties on underlying optical coefficients. Non-linearity arises when using parameterized aerosol size distribution models (e.g. log-normal). Even in linearly represented size distributions there is still non-linear relationship between optical coefficients and the index of refraction.

Non-uniformity in the prior weighting function is always present in the form of constraints on valid aerosol model parameters due to the unrealizability of negative aerosol loadings. Furthermore a non-uniform prior weighting function with higher weight at smoother size distributions is an implicit assumption in the Twomey-Tikhonov regularization used by Veselovskii et al. [1] and Müller et al. [3].

The third condition arises from the nature of estimating aerosol backscatter from the elastic lidar backscatter signal. The Fernald inverse formulation that is used to retrieve aerosol backscatter requires a molecular backscatter profile, an aerosol extinction-to-backscatter ratio profile, and a value for the molecular backscatter fraction at an aerosol scarce altitude, as inputs to the inverse formula. Retrieval error that arises from inexact values of these parameters is known to be a function of the backscatter coefficient [4].

A method of uncertainty assessment of the median radius and geometric standard deviation log-normal parameters and complex index of refraction using simple computational integrations over individual parameter coordinates has been demonstrated [2], however the application of this approach would become unmanageable if a multi-mode aerosol model were to be used due to an increase in the dimension of the problem. Furthermore, it is not always possible to retrieve a unique reconstruction of the underlying size distribution with the limited number of coefficients that can typically be measured with a multi-wavelength lidar system. Therefore Monte Carlo methods are desirable. The idea is to represent uncertainty by an ensemble of random outcomes of aerosol model parameters that are statistically congruent with their Bayesian posterior probability density function (PDF). Uncertainty in macroscopic aerosol properties such as optical properties that are not directly retrieved from lidar or volume, number, or surface concentration can be assessed in the form of estimated variances, histograms or estimated cumulative distribution functions (CDFs) derived from their values computed from the ensemble members. Markov chain Monte Carlo (MCMC) methods offer a way of assessing uncertainty that does not need any approximations to deal with the three non-ideal conditions previously mentioned and are increasingly being used in atmospheric statistical inverse problems [5].

2. METHODOLOGY

A basic description of a Markov Chain is that it is a sequence of random vectors (RVs) where the probability function for each RV in terms of all previous RVs in the chain is dependent only on the RV immediately preceding it and is identical for each RV in the chain [6]. This process is described by a transition kernel $K(\boldsymbol{\varphi},\boldsymbol{\theta})$. The kernel is the conditional PDF for the next RV in the chain, $\boldsymbol{\Phi}$, given that the previous RV, $\boldsymbol{\Theta} = \boldsymbol{\theta}$. Thus if $\boldsymbol{\Theta}$ has a PDF of $f_n(\boldsymbol{\theta})$, the PDF of $\boldsymbol{\Phi}$ is given by

$$f_{n+1}(\boldsymbol{\varphi}) = \int K(\boldsymbol{\varphi},\boldsymbol{\theta}) f_n(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} \quad . \qquad (1)$$

The goal of MCMC in this application is to find a kernel so that given an initial PDF, the PDFs of subsequent RVs converge to a target PDF, $\pi(\boldsymbol{\theta})$, as the RVs are further down the chain. In this case π would be the posterior PDF of the parameters,

$$\boldsymbol{\theta} = (N_1, \bar{r}_1, \sigma_1, \cdots, N_M, \bar{r}_M, \sigma_M), \qquad (2)$$

of a multi mode log-normal size distribution given by

$$n(r) = \sum_{i=1}^{M} \frac{N_i}{(2\pi)^{\frac{1}{2}} \ln(\sigma_i) r} \exp\left(-\frac{\ln(r/\bar{r}_i)^2}{2\ln(\sigma_i)^2} \right), \qquad (3)$$

where M is the number of modes and $(N_i, \bar{r}_i, \sigma_i)$ are the number concentration, median radius, and geometric standard deviation (GSD) .

A suitable kernel is given by the Metropolis–Hastings algorithm whose essential descriptions as described in chapter 1 of Ref. [7] are summarized in this paragraph. The generalized Metropolis–Hastings algorithm is implemented as follows: One begins with an initial outcome of $\boldsymbol{\Theta}_0$. At each link in the chain a candidate outcome, $\boldsymbol{\Phi}$, is obtained with a random number generator. The candidate outcome may then be accepted as the next outcome in the chain with a probability of

$$a(\boldsymbol{\Phi},\boldsymbol{\Theta}) = \min\left\{ 1, \frac{\pi(\boldsymbol{\Phi}) q(\boldsymbol{\Theta} \mid \boldsymbol{\Phi})}{\pi(\boldsymbol{\Theta}) q(\boldsymbol{\Phi} \mid \boldsymbol{\Theta})} \right\}, \qquad (4)$$

where $q(\boldsymbol{\varphi}|\boldsymbol{\theta})$ represents the conditional PDF for generating $\boldsymbol{\varphi}$ when $\boldsymbol{\Theta}_n = \boldsymbol{\theta}$. If the candidate is accepted then $\boldsymbol{\Theta}_{n+1}:=\boldsymbol{\Phi}$, otherwise the candidate is rejected and $\boldsymbol{\Theta}_{n+1}:= \boldsymbol{\Theta}_n$. The remarkable thing about this algorithm is that it has been proven that $f_n(\boldsymbol{\theta})$ always converges to $\pi(\boldsymbol{\theta})$ as $n \to \infty$ regardless of the shape of q if all regions of the domain of $\boldsymbol{\Theta}$ have a non-zero probability of being generated (see chapters 3 and 4 of Ref. [7]).

We develop an ensemble MCMC method that utilizes a devised form of the Metropolis-Hastings algorithm to generate samples of aerosol model parameters that are consistent with the statistical properties of their Bayesian posterior uncertainty probability density function. The algorithm is applied to an ensemble of parallel chains and is implemented with a hybrid random walk/independence sampling candidate generator (CG) to expedite convergence especially when local PDF maxima are present. The ensemble begins with an initial distribution and the chain is first executed with random walk candidate generation. Partway along the chain the ensemble is analyzed by finding the ensemble member with the maximum PDF value and then finding subsequent members that maximize a suitability parameter,

$$\begin{aligned} W(\boldsymbol{\Theta}) &= \pi(\boldsymbol{\Theta})^m \\ &\times \min\{ \left|\boldsymbol{\Theta} - \boldsymbol{\Theta}_{\mathrm{opt},i}\right| : i \in [1..N_{\mathrm{opt.\,set}}] \}, \end{aligned} \qquad (5)$$

where $\boldsymbol{\Theta}_{\mathrm{opt},\,i}$ represent the optimal parameters found so far. The parameter m is included so that more or less weight can be given to distances from members relative to PDF values. This member set in used to form an independence CG which randomly selects one of the optimal members and generates a sample centered around it. Once the independence CG is formed it remains without modification for subsequent links in the ensemble. We also employ a variation in which samples may also be drawn from the prior PDF, which is uniform in this case. The independence CG is merged with the random walk generator to form the hybrid CG. The merging is accomplished by random selection of walk or independence candidate generation. The PDF for the CG as a whole is then

$$q(\boldsymbol{\varphi} \mid \boldsymbol{\theta}) = p_W q_W(\boldsymbol{\varphi} \mid \boldsymbol{\theta}) + \sum_{i=1}^{N_g} p_i q_i(\boldsymbol{\varphi}), \qquad (6a)$$

with p_i satisfying

$$\sum_{i=1}^{N_g} p_i = 1 - p_W, \qquad (6b)$$

where p_w is the probability that a random step is chosen, $q_w(\boldsymbol{\varphi}|\boldsymbol{\theta})$ is the random step PDF, $q_i(\boldsymbol{\varphi})$ is the PDF corresponding to the i^{th} independent CG, N_g is the number of CGs, and p_i is the probability of sampling from $q_i(\boldsymbol{\varphi})$.

This method is similar to the *adaptive Metropolis* algorithm in Ref. [5], but in this case adaptation done only once and is based on an analysis of an ensemble of outcomes at the same link in many chains rather than at many links of one chain. By this way the ensemble of chains is representative of the evolution of the PDF as it is repeatedly transformed by the transition kernel that characterizes the Markov chain, and the ensemble can be monitored for convergence. The ensemble members in the last link of the chains are taken to be the output samples to be used to assess the uncertainty of aerosol properties. The ensemble MCMC method is also similar to the genetic algorithm applied to the retrieval of aerosol size distributions from extinction measurements by Lienert et al. [8]; however the ensemble MCMC method has a well defined probabilistic characterization.

3. RESULTS

In this study we applied the ensemble MCMC method to assessing uncertainty of the size parameters under the assumption of a single mode aerosol consisting of spherical particles with an index of refraction of 1.5 - 0*i*. Although these are not realistic assumptions, our purpose in this paper is to illustrate the methodology and the issues pertaining to it. A technique for separating the number concentration from **Θ** to compute a posterior PDF of only the size parameters from ratios of optical coefficient measurements has been formulated [2], which we employ. An example case was studied by applying the method to synthetic measurements of extinction at 355 and 532nm and backscatter at 355, 532, and 1064nm, all with 10% error, resulting from a size distribution with a median radius of 0.3μm and a GSD of 1.6. Figure 1 shows the state of the ensemble at link 50 in the upper left corner at which it is analyzed to form a hybrid CG. The results of further evolution of the ensemble without hybridization are shown in the upper right corner of Fig. 1. Results from two different hybrid generators are shown. Hybrid-B employs prior PDF sampling while hybrid-A does not. In both hybrids the probability of choosing walk generation was 90%. In hybrid-B the probability of choosing prior PDF sampling was 2%.

Two relevant features are demonstrated. One is that the problem of multiple local posterior PDF maxima which is endemic in non-linear inverse problems [8] is addressed by independence candidate generation. The cluster of particle size distributions located around $(\bar{r},\sigma) = (0.1\mu m, 1.2)$ is shown by the hybrid CG to be of little relevance. Comparing the ensemble states from the hybrid-A and hybrid-B CGs shows the principle of the detailed balance which explains convergence to the target PDF. In the hybrid-B CG, candidates are generated from the uniform prior PDF with a probability of 2%, yet the effect is that the ensemble members that are spread out end up being teleported to the relevant region in the $(\bar{r},\sigma)$ domain. This can be explained from the acceptance formula of Eq. 4. The term $q(\mathbf{\Theta}|\mathbf{\Phi})$ in the numerator results in low acceptance probability when there is little probability of returning from the generated candidate back to the neighborhood of the current sample.

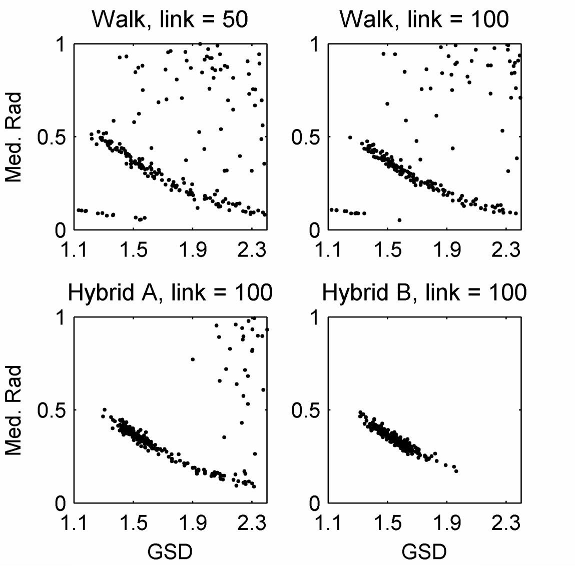

Figure 1. Ensemble state scatter plots

To demonstrate that the ensemble does converge to the target Bayesian posterior PDF, we compare cumulative distribution functions (CDFs) in median radius and GSD estimated from the ensemble to their numerically computed CDFs. The CDFs are expressed as

$$\mathrm{F}_X(x) = \int_{X_{\min}}^{x} \int_{Y_{\min}}^{Y_{\max}} \pi_{X,Y}(x',y')\,\mathrm{d}y'\,\mathrm{d}x', \qquad (7)$$

with X representing the log-normal distribution parameter that is being considered, either $\bar{r}$ or σ, and Y representing the other parameter. The CDFs estimated from the ensemble is expressed as,

$$\hat{\mathrm{F}}_X(x) = N\{X \le x\}/N_{\mathrm{total}}, \qquad (8)$$

where $N\{X \le x\}$ represents the number of ensemble members with the distribution parameter under consideration less than or equal to x and N_{total} represents the total number of ensemble members.

The Kolmogoroff-Smirnov test is a standard criterion to determine if an ensemble of single variable samples could have been drawn from a specific probability space with a specific CDF in that variable [9]. The test statistic is given by

$$\alpha_X = 2\exp\left(-2N_{\mathrm{total}} \max\left|\hat{\mathrm{F}}_X(X) - \mathrm{F}_X(X)\right|^2\right). \qquad (9)$$

This formula is an approximation, good for α < .25, to a random variable that would be uniformly distributed if the samples were drawn according to F_X. Hence sample ensembles that have $\alpha << 1$ are considered not to have been drawn from the probability space in question.

Figure 2 shows the evolution of α for median radius and GSD along the links of the ensemble of chains. Both values of α rise above 0.1 at link 118 and above 0.2 at link 228. Also it is noteworthy that the ensemble drifts into poorly representative states as can be seen around links 300 and 900. This effect is more pronounced in an example case with synthetic measurements resulting from a size distribution with a median radius of 0.6μm and a GSD of 1.6 as can be seen in figure 3.

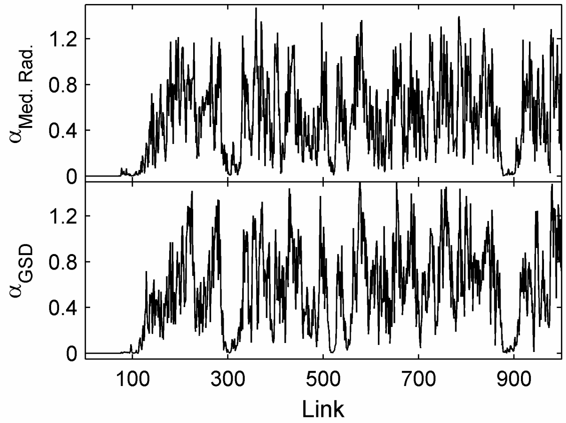

Figure 2. Evolution of Kolmogoroff-Smirnov test statistic for synthetic measurements generated from $(\bar{r},\sigma) = (0.3\mu m, 1.6)$.

4. CONCLUSION

We have demonstrated the use of the Metropolis-Hastings MCMC algorithm applied to an ensemble of samples representative of the uncertainty described by a Bayesian PDF in a 2-dimensional aerosol parameter domain. Our motivation for using this simple model was that it is one that we are familiar with and we could readily make comparisons with our past work. This is only a first step toward the development of an algorithm that could be used for assessing uncertainty of aerosol properties when aerosol is modeled as multi-modal with an unknown index of refraction. We are currently working on the construction of different correlated multidimensional random vector generation schemes to allow for tuned candidate generation since we anticipate that the simpler generation methods used in this study would be inadequate in higher dimensional models.

ACKNOWLEDGEMENTS

This work is supported by the research projects of NOAA #NA17AE1625 and NASA #NCC-1-03009.

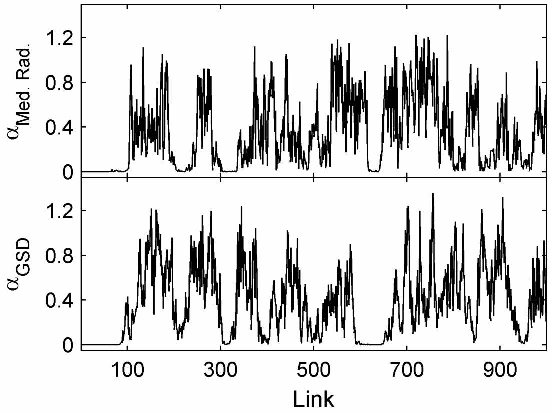

Figure 3. Evolution of Kolmogoroff-Smirnov test statistic for synthetic measurements generated from $(\bar{r},\sigma) = (0.6\mu m, 1.6)$.

REFERENCES

[1] Veselovskii I., A. Kolgotin, V. Griaznov, D. Müller, U. Wandinger, and D. N. Whiteman, 2002: Inversion with regularization for the retrieval of tropospheric aerosol parameters from multiwavelength lidar sounding *Applied Optics*, **41**, pp. 3685-3699.

[2] Herman B. R., B. Gross, F. Moshary, and S. Ahmed, 2008: Bayesian assessment of uncertainty in aerosol size distributions and index of refraction retrieved from multi-wavelength lidar measurements, *Applied Optics*, accepted for publication.

[3] Müller D., U. Wandinger, and A. Ansmann, 1999: Microphysical particle parameters from extinction and backscatter lidar data by inversion with regularization: Theory, *Applied Optics*, **38**, pp. 2346-2357.

[4] Sasano Y., E. V. Browell, and S. Ismail, 1985: Error caused by using a constant extinction/backscattering ratio in the lidar solution, *Applied Optics*, **24**, pp. 3929-3932

[5] Tamminen J., 2004: Validation of nonlinear inverse algorithms with Markov chain Monte Carlo method *Journal of Geophysical Research*, **109**, pp. D19303.

[6] Bremaud P., 1999: *Markov Chains: Gibbs Fields, Monte Carlo Simulation, and Queues*, Springer Science+Business Media Inc.

[7] Gilks W. R., S. Richardson, and D. J. Spiegelhalter (editors), 1996: *Markov Chain Monte Carlo in Practice*, Chapman & Hall.

[8] Lienert B. R., J. N. Porter, and S. K. Sharma, 2001: Repetitive genetic inversion of optical extinction data, *Applied Optics*, **40**, pp. 3476-3482.

[9] Papoulis A., 1991: *Probability, Random Variables, and Stochastic Processes* 3^{rd} Edition, McGraw-Hill.