Atmos. Chem. Phys. Discuss., 6, S3741–S3754, 2006
www.atmos-chem-phys-discuss.net/6/S3741/2006/

Atmospheric Chemistry and Physics Discussions

Interactive comment on "Implementation of a Markov Chain Monte Carlo Method to inorganic aerosol modeling of observations from the MCMA-2003 Campaign. Part I: Model description and application to the La Merced Site" *by* F. M. San Martini et al.

F. M. San Martini et al.

Received and published: 13 October 2006

- In my entire (albeit short) academic career I have never come across as thoughtful and comprehensive a review as Dr. Hellmuth's – thank you.
- Evaluation guidelines, comment 14: The reference to Hastings (Biometrika, 1970) has been added to the paper.
- Comment: *Apart from the key paper of Metropolis et al. (1953), it is sound to refer to a modern monograph, such as that of Draper (2006). It serves as a reference*

for the method... Unfortunately, it is still in preparation and, consequently, not freely accessible.

I concur that it is unfortunate that the Draper's work is still in preparation. I have addressed this deficiency by adding a reference to (Chib and Greenberg, 1995) in the manuscript and expanding the explanations of the method (see below). The reason I cite the work of Draper is because of all the treatments of MCMC I have found, Draper's is the most clear and easy to follow. Since his work was so influential in helping me understand the method, I wanted to give credit where it is due.

One clarification that should be made: the review states that "The MCMC method is used to approximate $p(\theta|Data)$ by generating random samples, from which the parameters for the probing distribution $PD(\theta|Data)$ [sic] can be estimated." The first part of the sentence is correct; indeed, the goal of the MCMC method is to approximate $p(\theta|Data)$ by generating random samples. Parameters for the probing distribution $PD(\theta^*|\theta)$, however, are not determined from the posterior distribution. Rather, the probing distribution is what guides how the Markov chain proceeds; specifically, samples from the probing distribution determine the Markov steps. This has been clarified by adding the following sentence to Section 3.4:

Samples from the probing distribution, also known as the candidate-generating (Chib and Greenberg, 1995) and jumping (Gelman et al., 1996) density, determine the proposed Markov steps.

- Comment: *I recommend to move Appendix A1 "Probing Distribution" to the non-numbered introducing paragraph of Section 3. Perhaps, it is useful to subsume the whole methodology in an own subsection. Can you regive the related parts of San Martini (2004) in more detail here?*

I have moved the Probing Distribution section from the Appendix to Section 3.4.

ACPD

6, S3741–S3754, 2006

As noted by the reviewer, the paper is already long. The following additional explanation of how the probing distribution is used to generate θ^* is therefore given here:

Probing Distribution and its Implementation to Determine the Markov Steps

We selected a multi-variate normal probing distribution with mean θ, i.e., the current position:

$$PD(\theta) \sim N_9(\theta, \Sigma) \quad (1)$$

where Σ is the covariance matrix. The task of selecting an efficient probing distribution is complicated by the fact that several of the inputs to the equilibrium model are a sum of (the unknown) random variables. Specifically, ISORROPIA requires the temperature, relative humidity, and the concentration of sodium and sulfate as well as the total ($i.e.$, particle + gas) concentration of ammonia, nitrate, and chloride (see equations 5, 6, and 7 in the manuscript).

While it is possible to formulate the problem in terms of a modified $\tilde{\theta}$, where $\tilde{\theta}$ is now the vector [T, RH, NH_3^t, NO_3^t, SO_4, Cl^t], this will lead to a Markov Chain that mixes more poorly than one where θ=[T, RH, NH_4, NO_3, SO_4, Cl, H_2O, NH_3, HNO_3, HCl],[1] because we are able to capture the covariance of the species by accounting separately for the aerosol and gas phase species. For example, when sampling the gas phase concentration of ammonia and nitric acid for the case of a dry aerosol, these concentrations are not independent: they are inversely related through the equilibrium constant $K_{NH_4NO_3}(T)$, i.e.,

$$P_{NH_3} P_{HNO_3} = K_{NH_4NO_3}(T) \quad (2)$$

Therefore, a proposed Markov step that increases both the concentration of ammonia and nitric acid (relative to the current equilibrium concentrations) will tend to be rejected, while a proposed Markov step that increases one and decreases the other according to (2) will tend to be accepted.

[1] Note that water is a calculated variable and not an input.

ACPD
6, S3741–S3754, 2006

EGU

If the elements of θ were independent, then the covariance matrix Σ would be a diagonal matrix whose off-diagonal elements are zero and whose diagonal elements are the element variances. For the case where the elements of θ are not independent, the vector of random variables θ will have covariance Σ given by:

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \cdots & \sigma_{1n} \\ \vdots & \ddots & \vdots \\ \sigma_{n1} & \cdots & \sigma_n^2 \end{pmatrix} \quad (3)$$

Σ is positive definite and symmetric, i.e., σ_{ij}=σ_{ji}. The covariance σ_{ij} of the random variables θ_i and θ_j is given by:

$$\sigma_{ij} = E\left[\left(\theta_i - E\left[\theta_i\right]\right)\left(\theta_j - E\left[\theta_j\right]\right)\right] \quad (4)$$

A positive or negative covariance indicates that for a single experiment the values of $\theta_i - E[\theta_i]$ and $\theta_j - E[\theta_j]$ tend to have the same or opposite sign.

In order to have an efficient sampling routine, we want the covariance matrix of the probing distribution (Σ) to resemble the covariance of the random variables. This is not required but merely efficient: it has been shown that the Markov Chain will converge for just about any probing distribution (Gilks et al., 1996). The more the probing distribution resembles the distribution of the random variables, the more efficient the sampling. For example, if two random variables θ_1 and θ_2 are negatively correlated, then proposed Markov steps that increase both of the variables will tend to be rejected. What is desired is a probing distribution that leads to a well-mixed chain: a probing distribution whose covariance does not resemble that of the random variables is likely to lead to a low acceptance probability and a Markov Chain that mixes poorly.

In general, slow mixing may be due to at least two reasons: the proposed Markov steps are too small so that the simulation moves very slowly through the target distribution; or the proposed Markov steps are to low-probability areas, leading the Markov chain to stand still most of the time (Gelman et al., 1996).

ACPD
6, S3741–S3754, 2006

EGU

The question now is how to estimate the covariance of the random variables θ, given that this is unknown before the analysis. The strategy used here is termed adaptive Metropolis sampling (see Gilks et al., 1998) and (Draper, Forthcoming)):

- Start with a poorly-tuned probing distribution ($e.g.$, assume the variables are independent and the variances scale with the magnitude of available observations)
- Run the MCMC simulation for a long time, with a large burn-in[2] and thin[3] the output
- Use the resulting MCMC dataset to estimate Σ
- Repeat until the estimate of Σ does not change significantly.

Gilks et al. showed that a danger of the adaptive strategy is that if adaptation is allowed to take place infinitely often, the MCMC simulation may not sample from the true target distribution (Gilks et al., 1998). Therefore, approximately 5 adaptation iterations were used in this work.

A final consideration when selecting the probing distribution concerns the magnitude of the Markov steps. If the Markov steps are too small, the chain will move too slowly through the solution space; conversely, if the Markov steps are too large, the proposed steps will try to wander far from the 'true' variable values, $i.e.$, the proposed steps will

[2]Burn-in refers to the initial Markov steps that are discarded to ensure that likelihood determined by the MCMC analysis is independent of the initial guess (see, for example, page 14 in .Gilks, W., Richardson, S. and Spiegelhalter, D., 1996. Introducing Markov Chain Monte Carlo. In: W. Gilks, S. Richardson and D. Spiegelhalter (Editors), Markov Chain Monte Carlo in Practice. Chapman & Hall, London, UK, pp. 1-21.).

[3]Thinning refers to practice of only storing every k^{th} Markov step to reduce storage requirements and help ensure that the Markov samples are not highly auto-correlated (see, for example, Raferty, A.E. and Lewis, S.M., 1992. How Many Iterations in the Gibbs Sampler? In: J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith (Editors), Bayesian Statistics 4. Oxford University Press, Oxford, U.K., pp. 763-774.).

ACPD
6, S3741–S3754, 2006

EGU

be to low-probability areas, leading to a Markov chain that mixes poorly. In the words of Draper, what we want is a chain that moves around freely, happily jumping all over the place (Draper, Forthcoming). One measure of how 'happy' the chain is jumping all over the place is the average acceptance probability: too low an acceptance probability indicates that the Markov steps may be too large, while too high an acceptance probability suggests the steps may be too small. Gelman et al. showed that for the type of probing distribution used here, the average acceptance rate should range from $\sim$44% for a univariate target distribution, and decreases to $\sim$23% for high dimensional problems (Gelman et al., 1996). Following the approach of .(Draper, Forthcoming), a scale factor κ is used, such that:

$$\Sigma = \kappa\Sigma_i \tag{5}$$

The scale parameter κ plays the role of tuning parameter, where κ is varied such that, once the covariance matrix Σ_i is estimated from the i^{th} iteration of the adaptive Metropolis sampling algorithm above, the acceptance probability is approximately equal to that suggested by (Gelman et al., 1996).

Equation (1) indicates that we wish to generate a 9-dimensional normal distribution, where the mean is the current position, and with covariance matrix Σ. Since generating this distribution is computationally inconvenient, the approach outlined in (Rao, 1992) is followed. Consider the normally distributed random variables θ_1, θ_2, ..., θ_n with known mean $\overline{\theta}$ and covariance Σ. To generate the required set of correlated random numbers θ^*, first a set of n statistically independent normally distributed random numbers W_1, W_2, ..., W_n are generated with mean $\overline{W_i}$ and variances $\sigma^2_{W_i}$.[4] The desired correlated random variables θ^* are then expressed as a linear function of the W_i:

$$\theta^*_i = a_{i1}W_1 + a_{i2}W_2 + ... + a_{in}W_n \tag{6}$$

or, in matrix notation:

$$\theta^* = AW \tag{7}$$

[4]Recall that since the W_i are independent, the covariance σ_{ij}=0 for all $i \neq j$.

ACPD
6, S3741–S3754, 2006

EGU

where A is the nxn matrix containing the a_{ij}'s of equation (6) and W is an nxn matrix where each row is the vector $W_1, W_2, \ldots, W_n$ (see equation (6)). Recall that if $X$is a normal random variable with mean μ and variance σ_x^2, and if a, b are scalars, then the random variable Y given by:

$$Y = aX + b \tag{8}$$

is also normal, with mean and variance given by:

$$E[Y] = a\mu + b \tag{9}$$

$$\sigma_Y^2 = a^2\sigma_x^2 \tag{10}$$

Therefore, from equation (6), $\overline{\theta} = A\overline{W}$, and from equation (7) we see that the covariance matrix of θ, Σ_θ, is given by:

$$\Sigma_\theta = A\Sigma_W A^T \tag{11}$$

In order to determine the matrices A and W, recall that a symmetric matrix S can be decomposed into the form of a matrix product as:

$$S = LDL^T \tag{12}$$

where L is a lower triangular matrix with ones on the main diagonal and D is a diagonal matrix. This decomposition method is a modified version of the Choleski method whose solution is well known. By comparing equations (11) and (12) we find that:

$$\Sigma_\theta = S \tag{13}$$

$$A = L \tag{14}$$

$$\Sigma_W = D \tag{15}$$

Since D is a diagonal matrix, the square root of its diagonal entries are the standard deviation of the W_i, $i.e.$,

$$\sigma_{W_i} = \sqrt{d_{ii}} \tag{16}$$

ACPD
6, S3741–S3754, 2006

EGU

In order to generate the n-element vector of normally distributed random numbers W, a further computational shortcut is taken. A vector of n independent random numbers X is generated by sampling from the standard normal $N(0,1)$ n times. The random numbers W_i are then calculated using:

$$W_i = \alpha_i X_i \tag{17}$$

Recall that X_i has mean 0 and unit variance. Since W_i is a linear function of the normal random variable X_i, we know from equation (10) that the variance of W_i is α_i^2, and therefore from equation (16) the scalars α_i are simply the standard deviations of the W_i, $i.e.$, the square root of the diagonal entries of Σ_W:

$$W_i \; = \; \sqrt{\Sigma_{ii}} X_i \tag{18}$$

The algorithm to generate the proposed Markov step given Σ is therefore:

- Calculate the modified Choleski decomposition of the probing distribution covariance matrix using equation (11)
- Sample the standard normal distribution n= 9 times and assign these values to the vector X
- Calculate the 9-element vector of independent normally distributed numbers W using equation (18)
- Calculate the proposed Markov step according to:

$$\theta^* \; = \; \theta + A \cdot W \tag{19}$$

where θ is the current position and A is the lower diagonal matrix calculated in the modified Choleski decomposition.

ACPD
6, S3741–S3754, 2006

The result of this algorithm is a symmetric probing distribution, where the mean is the current position θ and with covariance Σ.

The algorithm to estimate the covariance matrix Σ is (Draper, Forthcoming):

- Start with a Markov Chain whose Markov steps are a series of $N(0, \kappa_i\sigma_i^2)$ moves, i.e., the mean is the current position and the magnitude of the step is determined by $\kappa_i\sigma_i^2$. For the aerosol variables, temperature, and relative humidity, the σ_i^2 will scale approximately with the uncertainty of the measurements. For the gas phase variables, the range of any available previous observations may be used to estimate an upper bound on σ_i^2.
- Run the MCMC with a high burnin and thinning and use the covariance of the resulting dataset as an estimate $\Sigma_s^{\otimes}$ of Σ. If $\Sigma_s^{\otimes}$ differs significantly from $\Sigma_{s-1}^{\otimes}$, repeat.
- Run a simulation whose Markov steps are a series of $N(0,\kappa\Sigma_s^{\otimes})$ and vary κ to optimize the acceptance probability

Run the MCMC simulation with $\Sigma = \kappa\Sigma_s^{\otimes}$

- Comment: *Please itemize the principal premises of the Markovian Chain approach:*

The two key premises of the MCMC approach are:

1. The model relating the unknown variables to the observations (see Section 3.1), and
2. A probability model describing the likelihood of the observations (see Section 3.2).

ACPD

6, S3741–S3754, 2006

In addition, the following components of the MCMC method must be carefully considered:

- The selection of the prior distributions (see Section 3.3)
- The selection of the probing distribution (see Section 3.4 and response above);

Finally, in order to ensure the validity of the results, the convergence of the Markov chains must be ensured. This was accomplished using the convergence diagnostics discussed in Appendix A.

- Comment: *Is it possible to exemplarily illustrate how the proposed new θ^* and the probing $PD(\theta^*|\theta)$ distribution are generated? [This can be done verbally]*

See explanation above.

- Comment: *This is related to the question of how you arrive at a "prediction value" of, e.g., NH_3, HNO_3, etc. I suspect what you call "prediction" (denoted as "Mode", e.g., in Fig. 9), is nothing else but the modal value of the a postiori PDF according to the left hand side of Eq. 2. Is this interpretation correct?*

As noted above, the ultimate goal of the MCMC approach is to determine the posterior distribution, $p(\theta|Data)$. The goal is thus **not** to predict a single value, but a probability distribution, thus giving the range of likely values as well as their probability. In a sampling-based approach such as the MCMC method, the modal value is the most likely value. In Figure 9(a) we only show the measurement time series, the most likely value (the mode of the posterior) and the 95% confidence interval. Figure 9(b) shows an image plot of the posterior distribution. In Figure 10 we compare the most likely value predicted with the MCMC approach with the measurements.

ACPD

6, S3741–S3754, 2006

In sum, the modal value of the posterior probability density function is the most likely value of the random variable θ_i.However, the left-hand-side of equation (2) deals with the entire pdf rather than a single value (e.g., mode or expected value). As stated above, the ultimate goal of the MCMC approach is to determine the posterior probability density function. This pdf can then be used to provide a decision maker with representative values of interest (e.g., mode, 95% confidence interval, etc.).

- Comment: *This way, your "predictive values" are intrinsically diagnostic values. With respect to deterministic models, the word "prediction" colloquially refers to a future state along a trajectory in the phase space. This way, the evolution of an intrinsically predictive variable is fully determined by any conservation law with corresponding initial and boundary conditions. Measurements enter the prediction only via the determination of the initial and boundary conditions. I think, the situation here is different. You have observations, which are used the estimate the most likely "true" state.*

Although 'predict' is often understood to apply to future states (e.g., weather forecasting), it is also understood colloquially for model estimates of unobserved or unobservable quantities. Since we do not know the 'true' value of θ, we feel this is an appropriate use of the term.

- Comment: *Can you give a physical interpretation of the "acceptance rule" for the probing distribution given by Eq (3)?*

The acceptance rule given by Equation (3) is the probability of a move, i.e., if the current position of the Markov Chain is θ, then the probability that the move to the proposed θ^* is accepted in the Metropolis algorithm is determined simply by ratio of the likelihood function evaluated at the respective position (or, if this ratio is greater

ACPD
6, S3741–S3754, 2006

EGU

than unity, the step is accepted). The contribution of Hastings (1970) was to generalize the algorithm of Metropolis to allow for the use of probing distributions that are not symmetric. In Hastings' algorithm, the probability that the proposed Markov step is accepted is the product of the ratio of the likelihood function times the ratio of the probing distribution evaluated at the current position and the proposed position. In both algorithms, the probability that a move is accepted is determined by comparing the acceptance probability with U, where U is a random number generated on the interval [0,1]. By comparing the acceptance probability to a random number on [0,1] we ensure that the Markov chain is allowed to explore regions of low probability.

To understand why α has the form of equation (3) can be seen by examining the approach of Hastings (1970). We wish to construct a Markov Chain with a stationary distribution (or target distribution) π. The reversibility condition requires that:

$$\pi_i p_{ij} \;=\; \pi_j p_{ji} \tag{20}$$

where π_i is the probability of position i and p_{ij} is the probability of moving from position i to j. Assume that p_{ij} has the form (Hastings, 1970):

$$p_{ij} \;=\; q_{ij}\alpha_{ij} \quad (i \neq j)\,, \tag{21}$$

with

$$p_{ii} \;=\; 1 - \sum_{j \neq i} p_{ij} \tag{22}$$

In words, equation (21) says that the probability of moving from state i to state j is the product of the acceptance probability α_{ij} and a probing distribution q_{ij}. The question is what form should the acceptance probability take to satisfy the reversibility condition (equation (20)). For the general case, Hastings suggested the acceptance probability:

$$\alpha_{ij} \;=\; \frac{s_{ij}}{1 + \frac{\pi_i}{\pi_j}\frac{q_{ij}}{q_{ji}}} \tag{23}$$

ACPD
6, S3741–S3754, 2006

EGU

where s_{ij} is a symmetric function of i and j chosen such that 0$\leq$ α_{ij} $\leq$1 (Hastings, 1970). Substituting equation (23) into equation (21):

$$p_{ij} = \frac{q_{ij} s_{ij} (\pi_j q_{ji})}{\pi_i q_{ij} + \pi_j q_{ji}} \quad (24)$$

Similarly, the expression for p_{ji} is given by:

$$p_{ji} = \frac{q_{ji} s_{ji} (\pi_i q_{ij})}{\pi_i q_{ij} + \pi_j q_{ji}} \quad (25)$$

Multiplying equation (24) by π_i and equation (25) by π_j, and recalling that the function s_{ij} is symmetric, $i.e.$, $s_{ij} = s_{ji}$, we arrive at $\pi_i p_{ij} = \pi_j p_{ji}$, thus demonstrating that the acceptance probability given by equation (23) satisfies the reversibility condition.

Various forms have been suggested for the function s_{ij}. Barker suggested simply s_{ij}=1 (Barker, 1965), while Hastings proposed:

$$s_{ij} = \begin{cases} 1 + \frac{\pi_i q_{ij}}{\pi_j q_{ji}} & \left(\frac{\pi_j q_{ji}}{\pi_i q_{ij}} \geq 1\right) \\ 1 + \frac{\pi_j q_{ji}}{\pi_i q_{ij}} & \left(\frac{\pi_j q_{ji}}{\pi_i q_{ij}} \leq 1\right) \end{cases} \quad (26)$$

Combining equations (26) and (23) we arrive at:

$$\alpha_{ij} = \begin{cases} 1 & \left(\frac{\pi_j q_{ji}}{\pi_i q_{ij}} \geq 1\right) \\ \frac{\pi_j q_{ji}}{\pi_i q_{ij}} & \left(\frac{\pi_j q_{ji}}{\pi_i q_{ij}} \leq 1\right) \end{cases} \quad (27)$$

The π_i's comprise the distribution of steady state probabilities of the Markov Chain, i.e., in Bayesian notation $\pi_i = p(\theta_i|Data)$. Similarly, in words q_{ij} is the probability of going to state j given that the current state is state i, i.e., $q_{ij} = p(X_{n+1} = j|X_n = i) = PD(\theta^*|\theta)$, and thus equation (27) can be written as:

$$\alpha_{ij} = \min\left\{1, \frac{p(\theta^*|Data)}{p(\theta|Data)} \frac{PD(\theta|\theta^*)}{PD(\theta^*|\theta)}\right\} \quad (28)$$

ACPD
6, S3741–S3754, 2006

EGU

The algorithm of Metropolis is a special case of the more general Hastings algorithm. Metropolis suggested a symmetric probing distribution (*i.e.*, $q_{ij} = q_{ji}$, which implies that the ratio $\frac{PD(\theta|\theta^*)}{PD(\theta^*|\theta)}$ is unity) and thus the acceptance probability is given by:

$$\alpha_{ij} = \min\left\{1, \frac{p(\theta^*|Data)}{p(\theta|Data)}\right\}(q_{ij} = q_{ji}) \tag{29}$$

- Comment: *The arguments for the setup of the likelihood function are plausible. The observations X_{obs} entering the likelihood functions $p(X_{obs}|X)$ are formally "mean values". Do you use time-averaged values derived from high-resolution time series for these "mean values", and do you presume thereby already the validity of the ergodic hypothesis?*

The AMS observations are 4-minute averages; temperature and relative humidity are averaged from per minute data or, where these are unavailable, interpolated from hourly averaged data to the AMS time stamp; the TILDAS observations are available at a one second time resolution; the FTIR observations are 5-minute averages that were interpolated to the AMS time stamp. No further averaging was done when running MCMC analysis, i.e., the MCMC analysis was conducted on each 4-minute set of observations independently.

Interactive comment on Atmos. Chem. Phys. Discuss., 6, 5933, 2006.

ACPD

6, S3741–S3754, 2006

Interactive Comment