National Aeronautics and Space Administration

Accelerating MCMC to Operational Speeds

Principal Investigator: Amy Braverman (398); Co-Investigators: Jonathan Hobbs (398), Vijay Natraj (329), David Thompson (382), Youssef Marzouk (MIT)

Program: FY21 SURP   **Strategic Focus Area: Uncertainty Quantification**

Objectives

Our objective was to develop a Markov chain Monte Carlo (MCMC) retrieval algorithm for Earth remote sensing that is fast enough for operations. A natural approach to the geophysical retrieval problem is to invoke Bayes' Rule: if *X* and *Y* are two jointly distributed random variables (or vectors) then the posterior distribution of *X*, *P* (*X*|*Y*), is proportional to the product of the likelihood, *P*(*Y* |*X*), and the prior distribution, *P*(*X*):

$$P(X|Y) \propto P(Y\mid X)P(X).$$

P(*X*|*Y*) reflects a new description of the behavior of *X*, based on what is learned from observing *Y*. By providing this posterior *distribution* of the geophysical state given the observed spectra, uncertainty information is incorporated directly into the retrieval output.

The standard algorithm for obtaining *P* (*X*|*Y*) in remote sensing (Optimal Estimation, or "OE") [1] makes many highly restrictive assumptions. MCMC does not require these assumptions, but is computationally intensive, and generally too slow for operational use. In this project, we investigate a set of computational and theoretical modifications to standard MCMC algorithms to achieve the speed required for operational use. We used the upcoming Surface Biology and Geology (SBG) mission as a motivating example.

Background

If *X* is a true geophysical state vector and *Y* is the corresponding observed spectrum, the two are related by *Y* = *F* (*X*) + *e*, where *F* is the forward model, and *e* is an error term (Figure 1). In OE, *X* and *e* are assumed to be Gaussian, and *F* is linearized, so the posterior distribution is also Gaussian. Under these conditions, specification of the posterior distribution only requires specifying the first two central moments of that distribution. In reality *F* is not linear, and the posterior distribution is unlikely to be Gaussian. In that case, the first two central moments can be highly misleading because they ignore multi-modality, extremes, and other important characteristics of the posterior distribution necessary to draw accurate conclusions and compute accurate uncertainties in climate and Earth science problems.

MCMC [2] is a widely-used simulation-based method for obtaining posterior distributions without restrictive assumptions on their forms. However, even in simple cases MCMC can be computationally expensive, and too slow for operations. This is due, in large part, to two factors both of which are typical in remote sensing: high-dimensionality of both the spectra and the state vectors, and complex, computationally intensive forward models.

Approach and Results (1)

We used the SBG designated observable mission concept as a motivating application. SBG will retrieve 425-dimensional reflectance (state) vectors from observed radiances vectors, and we have set up a test-bed to do this via MCMC for case studies using synthetic truth derived from proxy data sources. We compared our test-bed to benchmarks of various versions of accelerated MCMC against OE. See Figure 1.

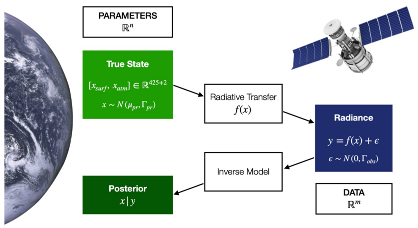
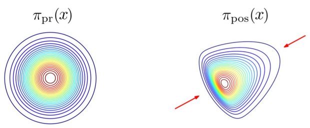


Figure 1. The SBG inference problem. MCMC is the algorithm that performs the inversion to obtain the posterior distribution of the state (x) given the observations (y).



$$\mathbb{R}^d = \underbrace{X_r}_{\pi_{\mathrm{pos}} \neq \pi_{\mathrm{pr}}} + \underbrace{X_\perp}_{\pi_{\mathrm{pos}} \approx \pi_{\mathrm{pr}}}$$

Figure 2. Likelihood-informed subspace (LIS) is a dimension reduction method applied to the state vector, x. LIS partitions the space into two. In the first, the radiances (y, which are jointly distributed with the state, x) the prior and posterior distributions are different because the data provide new information. In this figure, x is two-dimensional. The prior distribution of x is on the left and the posterior is shown on the right. Introducing the observations, y, changes the distribution differently in the two dimensions.

Approach and Results (2)

We formulated and implemented dimension reduction for both the radiance observations and the state vector using principal component analysis (PCA) and Likelihood-informed Subspaces (LIS; see Figure 2) [3], and compared their performance. We found that, 1) for any given level of dimension reduction, radiances reduced via LIS carry more information about the state (and can produce estimates with an order of magnitude less error; 2) using LIS-dimension-reduced state vectors gave better estimates of the the posterior mean and the posterior covariance matrix than when dimension was reduced by PCA (see Figure 3); and 3) MCMC converged more quickly by orders of magnitude when run on LIS-reduced states (see Figure 4).

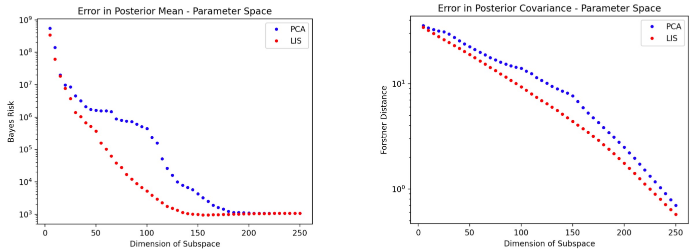


Figure 3. Errors in the posterior mean and covariance matrix estimates when dimension reduction is performed via PCA versus LIS. Bayes Risk is a matric for distance between mean vectors. Forstner distance is a distance metric for covariance matrices. This analysis was performed on a single pixel in the often-used Beckman Lawn test scene [4], and used a simple inversion based on a linear forward model.

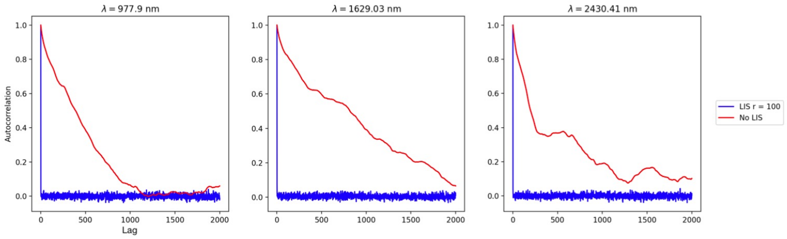


Figure 4. MCMC autocorrelation plots for three state vector wavelengths. MCMC creates a Markov Chain, which after a burn-in period, represents samples from the posterior distribution of interest. The faster the samples become uncorrelated in the chain, the faster the chain converges to the target distribution. This plot shows that using LIS dimension reduction and running the chain in the reduced space, dramatically improves convergence.

Significance

This technology will make JPL more competitive in winning new missions by providing retrieval algorithms that exploit observations more fully than do current methods. Mission studies and proposals using MCMC retrievals to demonstrate the value of new sensor technologies will have the benefit of rigorous but flexible probabilistic descriptions of the behaviors of the quantities of interest (QOI) to a much greater degree than is possible using methods that rely on Gaussian assumptions. The specific technical achievements described here represent the first concrete steps. The dimension reduction work will have near-term benefits in the application of existing OE algorithms, as well.

National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

www.nasa.gov

References:

[1] Clive Rogers, *Inverse Methods for Atmospheric Sounding,* World Scientific publishers, 2000.
[2] Andrew Gelman and Donald B. Rubin, "Inference from Iterative Simulation Using Multiple Sequences", *Statistical Science*, **7**(4) (1992), pp. 457-472, DOI: 10.1214/ss/1177011136.
[3] T. Cui, J. Martin, Y.M. Marzouk, A. Solonen, and A. Spantini, "Likelihood-informed Dimension Reduction for Nonlinear Inverse Problems", *Inverse Problems*, **30(11**) (2014), DOI: 10.1088/0266-5611/30/11/114015.
[4] D.R. Thompson et al. "Quantifying uncertainty for remote spectroscopy of surface composition," *Remote Sensing of the Environment*, **247** (2021), DOI: 10.1016/j.rse.2020.111898.

PI/Task Mgr Contact
Email: Amy.J.Braverman@jpl.nasa.gov

Clearance Number: CL#21-4742
RPC/JPL Task Number: SP20009