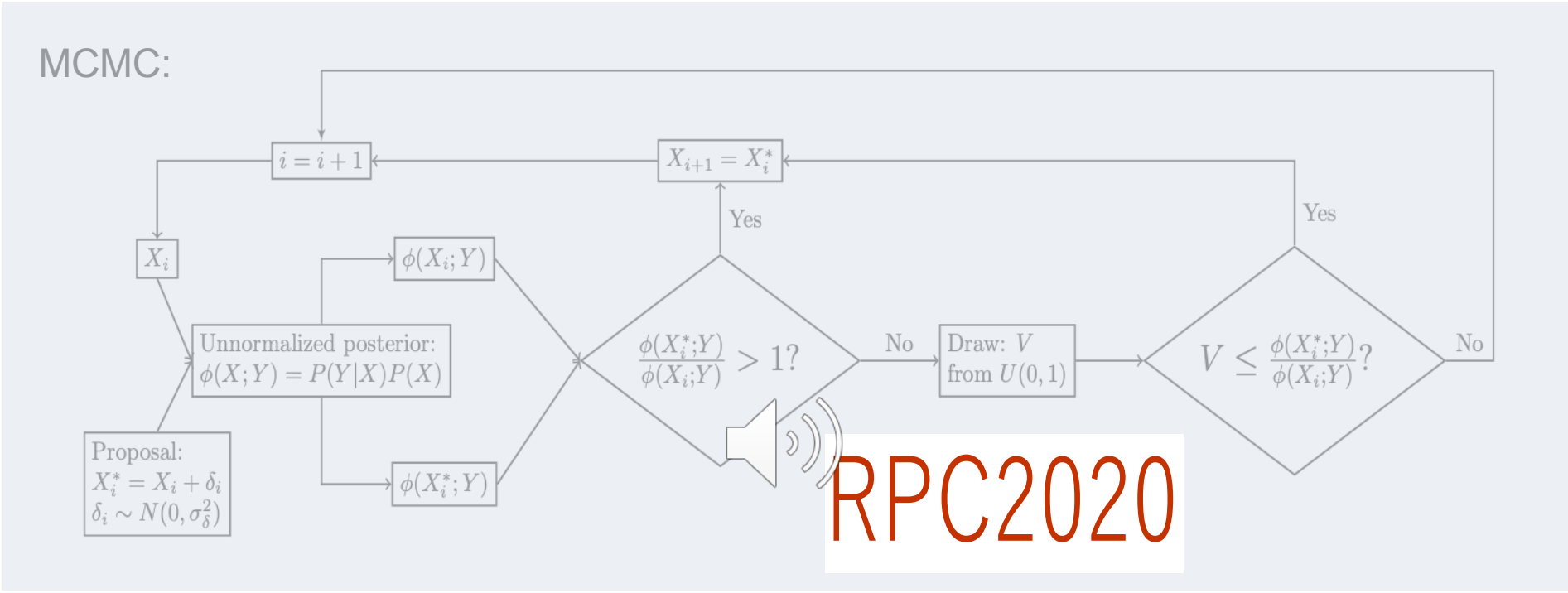


Virtual Research Presentation Conference

Accelerating MCMC to Operational Speeds

Principal Investigator: Amy Braverman (398)
Co-Is: Youssef Marzouk (MIT), Jonathan Hobbs (398), Vijay Natraj (329), and David R. Thompson (382)
Program: SURP

Assigned Presentation #RPC-119

Jet Propulsion Laboratory
California Institute of Technology

Tutorial Introduction

Abstract:

The process of inferring the true state of a remotely sensed scene from observed spectra is called a retrieval. Typical retrieval methods include least-squares fitting and application of Bayes' Rule, which uses the formula for conditional probability to obtain the probability distribution of the true state given the observed data. The Bayes retrievals are preferred because, in principal, they yield the full distribution of the state rather than simple point estimates and thus includes uncertainty information.

In practice, the most often used algorithm for Bayes retrievals (Optimal Estimation, OE) assumes all distributions to be Gaussian. A more modern alternative is Markov-chain Monte Carlo (MCMC), but it is very computationally intensive and not generally practical in operational settings. While there are many approaches to increasing MCMC speed and efficiency, none has so far achieved the kind of improvements required for operational deployment at NASA.

In this project, we explore new ways to make MCMC algorithms faster using the Surface Biology and Geology designated observable mission concept's retrieval as a guiding example. We bring advances from the Uncertainty Quantification domain to bear on this problem starting with assessments of 1) three different simple emulators (statistical models that approximate complex physical calculations) and two different dimension reduction techniques (transformations of the input data that reduce their size while sacrificing minimal information).

Problem Description

Context: Massive remote sensing data sets deliver detailed, local information on global scales, allowing scientists to test complex theories across multiple scales in time and space. To do so requires that retrievals capture complex non-Gaussian nonlinear behaviors and dependencies, and accurately quantify uncertainties in those estimates.

Advancement over current state-of-the-art: The traditional implementation of Bayes' retrievals is Optimal Estimation (OE) [1]. OE assumes that all the distributions involved in the formula,

$$P(state|obs) = \frac{P(obs|state)P(state)}{P(obs)}$$

are Gaussian. The Gaussian assumption does not necessarily hold, is rarely if ever checked, and if it fails, can lead to incorrect inferences and science conclusions. MCMC [2] methods do not require any assumption about the form of the distributions and are thus more robust. The downside is that MCMC is very slow.

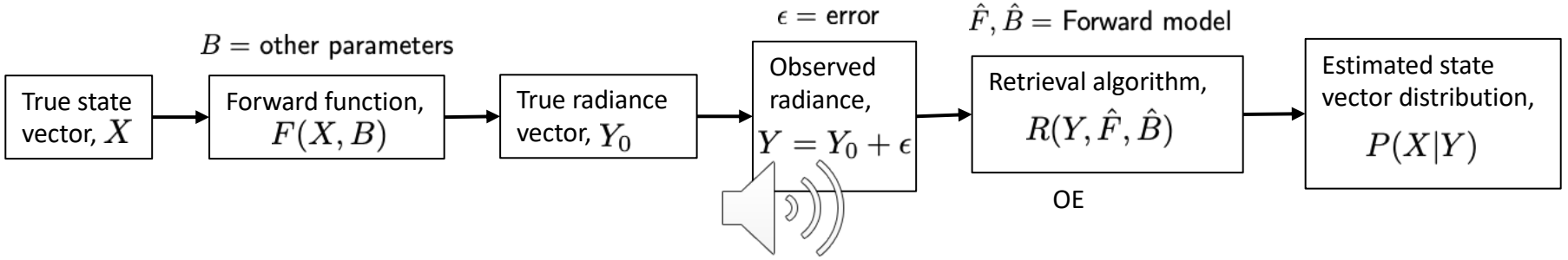
Relevance to NASA and JPL: Makes JPL more competitive in winning new missions by

1) increasing the speed, flexibility, and efficiency of probabilistic retrieval algorithms used in Earth Science

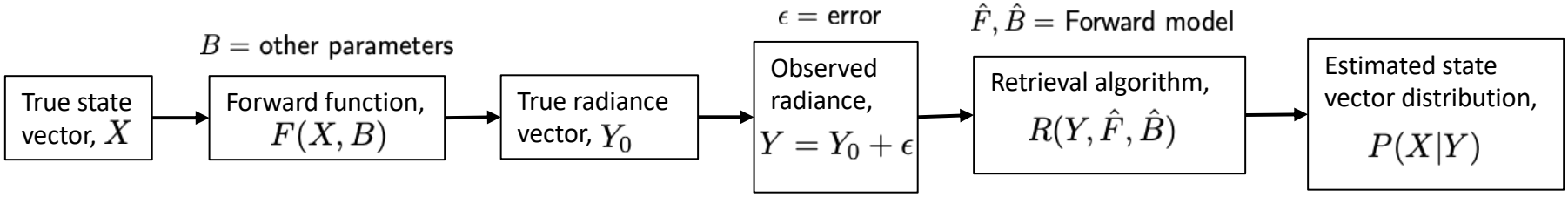
2) demonstrating the value of new sensor technologies by exploiting more of the information in their data and providing richer probabilistic descriptions of uncertainty.

Methodology

a) Observing system/experimental framework:



b) Innovation



Results

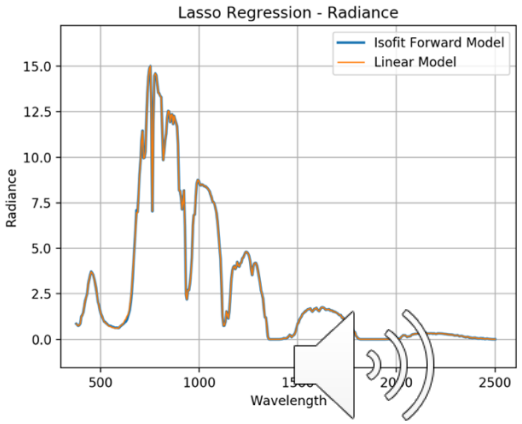
a) Completed experiments:

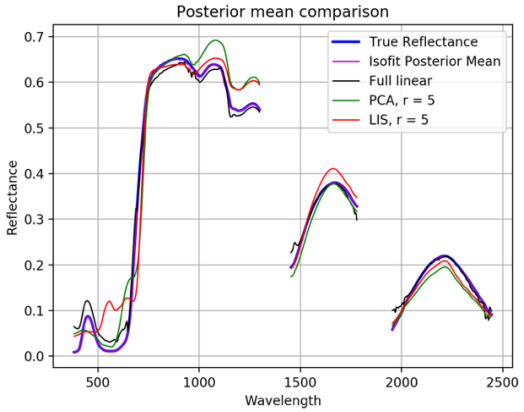
- comparing linear emulators for different error distributions
- comparing PCA vs LIS for dimension reduction for different error distributions
- first version of MCMC implementation complete

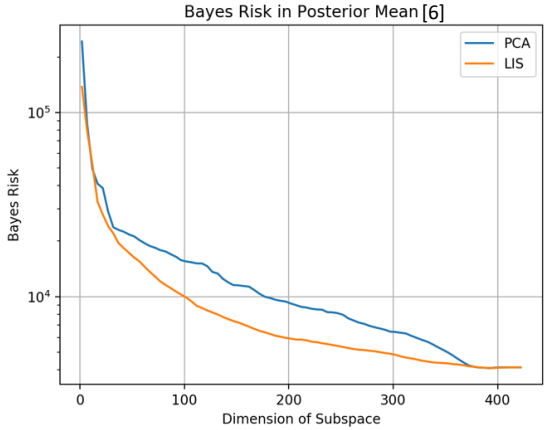
b) Significance: LASSO regression + LIS dimension reduction is most efficient among choices tested.

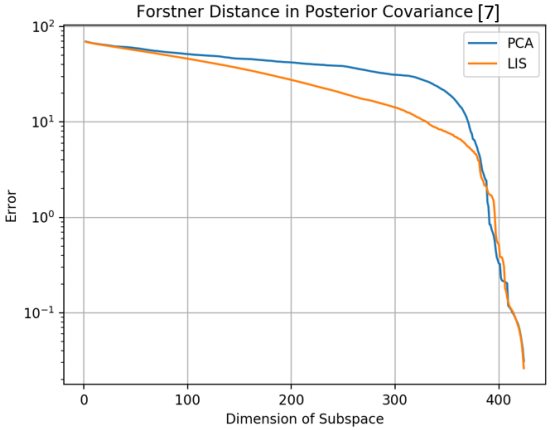
c) Next steps:

- embed forward model emulator+LIS into MCMC,
- timing studies to compare MCMC with linear emulator+LIS against ordinary MCMC and OE.









Publications and References

[1] Clive Rodgers, *Inverse Methods for Atmospheric Sounding*, World Scientific, 2000.

[2] Andrew Gelman, John Carlin, Hal Stern, and Donald Rubin, *Bayesian Data Analysis,* first edition, Chapman and Hall, 1995.

[3] A.K. Md. Ehsanes Saleh, Mohammad Arashi, and B.M. Golam Kibria, *Theory of Ridge Regression with Applications*, John Wiley and Sons, 2019.

[4] I.T. Jolliffe, *Principal Component Analysis*, Springer, 2002.

[5] Tiangang Cui, J. Martin, Y. Marzouk, A. Solonen, and Alessio Spantini, "Likelihood-informed dimension reduction for nonlinear inverse problems," *Inverse Problems,* **30**, 11, October 29, 2014.

[6] James O. Berger, *Statistical Decision Theory and Bayesian Analysis*, Second Edition, Springer, 1985.

[7] Wolfgang Forstner and Boudewijn Moonen, "A Metric for Covariance Matrices", in *Geodesy-The Challenge of the Third Millenium*, E.W. Grafarend, F.W. Krumm, and V.S. Schwarze (eds.) 2003. DOI: https://doi.org/10.1007/978-3-662-05296-9_31.