Accelerating Bayesian Computation in Earth Remote Sensing Problems

by

Kelvin Man Yiu Leung

Submitted to the Department of Aeronautics and Astronautics
in partial fulfillment of the requirements for the degree of

Master of Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2021

Author .....
Department of Aeronautics and Astronautics
May 18, 2021

Certified by .....
Youssef Marzouk
Professor, Aeronautics and Astronautics
Thesis Supervisor

Accepted by .....
Zoltan Spakovszky
Professor, Aeronautics and Astronautics
Chair, Graduate Program Committee

Accelerating Bayesian Computation in Earth Remote Sensing Problems

by

Kelvin Man Yiu Leung

Submitted to the Department of Aeronautics and Astronautics
on May 18, 2021, in partial fulfillment of the
requirements for the degree of
Master of Science

Abstract

Earth atmospheric remote sensing is an inverse problem that fits surface and atmospheric models to imaging spectrometer data and is critical to the analysis of the composition and biodiversity of the Earth surface. Current methods for remote sensing generally involve retrieving a point estimate of the surface reflectance and atmospheric parameters.

This thesis presents a more robust Bayesian approach to quantify the uncertainty of the retrieval, but this is computationally intractable given the high dimensionality of the problem. In many Bayesian inverse problems, however, there exists a low-dimensional likelihood-informed subspace that describes both optimal projections of the data and directions in parameter space that are most informed by the data.

In the Bayesian approach, Markov chain Monte Carlo (MCMC) is implemented within this low-dimensional subspace to increase sampling efficiency. For an example retrieval, reducing the parameter dimension by a factor of 4 increased the effective sample size of the MCMC chain by more than two orders of magnitude. This low-dimensional subspace was shown to be able to capture the key features of the posterior structure from a higher dimension. The posterior variance obtained through MCMC was also shown to better represent the uncertainty of the problem over the existing method.

Thesis Supervisor: Youssef Marzouk
Title: Professor, Aeronautics and Astronautics

Acknowledgments

This research is a collaborative project between MIT and NASA Jet Propulsion Laboratory (JPL) through the SURP program. A big thank you to everyone on the JPL team for your support and expertise throughout the past year, especially to David for helping me get acquainted with everything I needed to know with regards to remote sensing.

I also want to thank Jayanth for your continued support and flexibility for whenever I have technical questions or troubles. It has been tremendously helpful.

Last but not least, I would like to thank my advisor, Youssef, whom I got to know while working as TA during my first semester of grad school, which ultimately led me to join the group. I had very little background in this field, but I was able to get caught up quickly in the first few months through your guidance, even during Covid times.

Contents

List of Figures

Chapter 1

Introduction

1.1 Motivation

Earth atmospheric remote sensing is the acquisition of parameters on the Earth surface from satellites. Remote Visible/ShortWave InfraRed (VSWIR) imaging spectroscopy is a common remote sensing tool used to analyze the composition and biodiversity of the Earth surface. In particular, this is used in the Surface Biology and Geology (SBG) study initiated by the NASA Jet Propulsion Laboratory [10].

The goal of this remote sensing problem is to infer a set of parameters on Earth given images taken by satellites. This process is known as a retrieval. Each pixel of the image contains radiance data along a spectrum of wavelengths in the shortwave infrared and visible range from 350 to 2500 nm. A set of Earth surface parameters are inferred, with each surface parameter corresponding to the same spectrum of wavelengths as the radiance data. These parameters indicate the surface reflectance at the specified wavelength. Additional parameters describing the atmospheric conditions of the retrieval are also inferred.

Figure 1-1 shows a retrieval of a plain grass lawn at the California Institute of Technology. The imaging spectrometer records a spectrum of radiances and the surface reflectances across the same range of wavelengths is estimated.

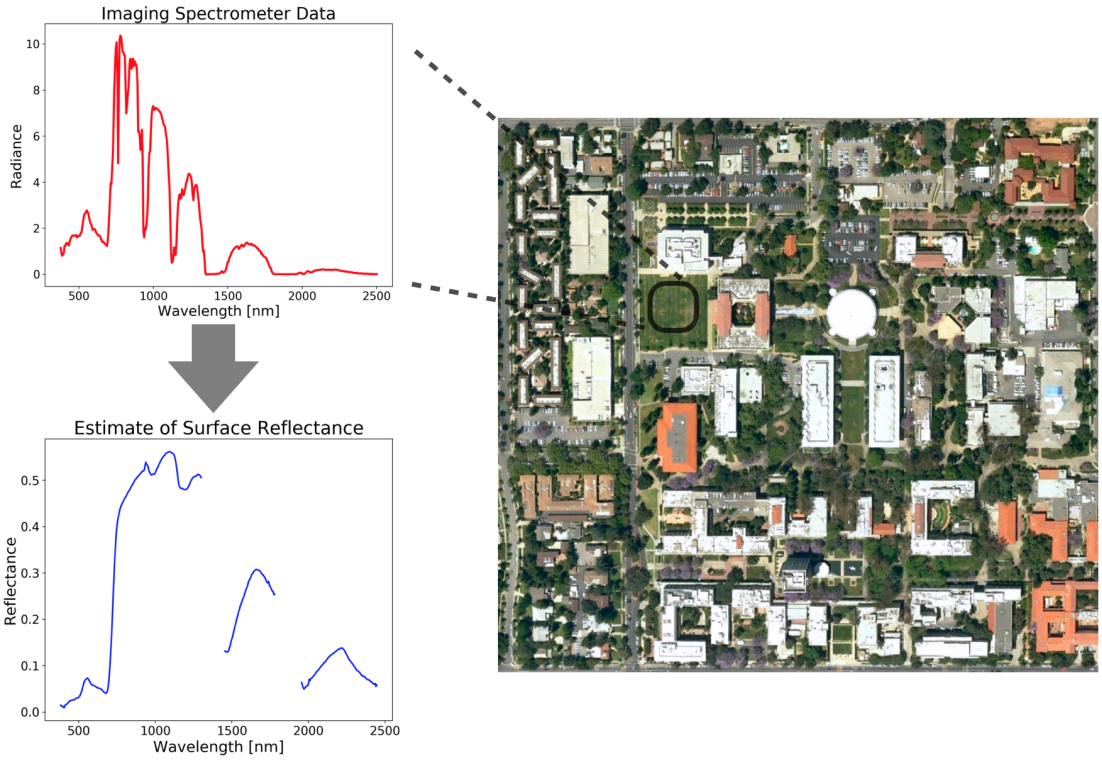


Figure 1-1: Example retrieval on a grass lawn

[13] outlines the current methodology used for remote sensing problems, referred to as optimal estimation (OE). The parameters are inferred through an optimization problem to obtain a maximum a posteriori (MAP) estimate. The covariance is estimated using the Laplace approximation as a function of the local Jacobian at the MAP estimate.

The main motivation behind this research is to improve the uncertainty quantification of the retrieval using a Bayesian approach. In the current approach, the retrieval returns a point estimate and the uncertainty is approximated using a locally linear gradient. Newer technologies being used in recent missions require exploiting as much information as possible from the observed data. The Bayesian approach returns a posterior distribution, an improvement upon a point estimate. For example, certain applications of the remote sensing problem are expected to have a non-Gaussian and multimodal posterior distribution, which cannot be characterized using the current approach.

However, the high dimensionality of this remote sensing problem poses a challenge to this approach. The efficiency of sampling algorithms deteriorate quickly with dimension, rendering this particular problem very computationally expensive. We explore methods of dimension reduction to make this approach more tractable.

1.2 Literature Review

1.2.1 Bayesian Inverse Problems

The Bayesian approach to inverse problems returns a posterior distribution conditioned on the observed data. Given radiance data collected from the imaging spectrometer, a posterior distribution of the inferred surface and atmospheric parameters can be obtained instead of a point estimate of the posterior mode. Quantities of interest such as posterior mean and posterior covariance can then be computed from this distribution. These posterior expectations are typically computed using sampling methods such as Markov chain Monte Carlo (MCMC).

Bayesian methods are based on Bayes rule, which relates the posterior distribution to the prior $\pi(x)$ and likelihood $\pi(y|x)$ distributions,

$$\pi(x|y) = \frac{\pi(y|x)\pi(x)}{\pi(y)} \propto \pi(y|x)\pi(x). \tag{1.1}$$

The setup of an inverse problem is to infer a set of parameters x given a set of observed data y modelled by

$$y = f(x) + \epsilon, \tag{1.2}$$

where $f(x)$ is the forward mapping from parameters to data and ϵ is a random variable representing the noise and model error.

A simple but common version of this problem is when the prior and likelihood are

both Gaussian,

$$x \sim \mathcal{N}(\mu_{pr}, \Gamma_{pr})$$
$$y|x \sim \mathcal{N}(0, \Gamma_{obs}),$$

where μ_{pr} and Γ_{pr} are the prior mean and covariance, and Γ_{obs} is the noise covariance. For a linear inverse problem $f(x) = Gx$, the posterior is also Gaussian with distribution

$$x|y \sim \mathcal{N}(\mu_{pos}, \Gamma_{pos}).$$

In this case, we can obtain closed form expressions for the posterior mean and covariance,

$$\mu_{pos}(y) = \Gamma_{pos}(G^\top \Gamma_{obs}^{-1} y + \Gamma_{pr}^{-1} \mu_{pr}) \tag{1.3}$$
$$\Gamma_{pos} = (H + \Gamma_{pr}^{-1})^{-1}, \tag{1.4}$$

where $H = G^\top \Gamma_{obs}^{-1} G$ is the Hessian of the negative log likelihood, or data misfit function.

When the forward model is nonlinear, the posterior distribution is generally not normally distributed and there is no closed form expression for the mean and covariance. In this case, methods such as MCMC can be used to characterize the posterior by drawing samples from the distribution.

1.2.2 Markov Chain Monte Carlo

Markov chain Monte Carlo (MCMC) is used to generate samples from a random variable with an arbitrary target distribution known up to a normalizing constant [5]. In the context of inverse problems, this target is the posterior distribution. For the remote sensing problem, we consider the Metropolis-Hastings and Adaptive Metropolis algorithms.

Metropolis-Hastings

Metropolis-Hastings is the most basic algorithm for MCMC. A Markov chain is constructed using a new proposal at every step and a mechanism that determines whether to accept or reject the proposal. This is described in Algorithm 1.

Initializing the Markov chain at $x^{(0)}$, proposed samples are drawn according to the proposal distribution $q(\cdot\,|x^{(i)})$ centered at the most recent sample $x^{(i)}$. These proposed samples are added to the chain with acceptance probability α, which is a function of the posterior density $\pi(\cdot\,|y)$. The normalizing constant is not required since they would cancel in the numerator and denominator.

Algorithm 1: Metropolis-Hastings MCMC

Initialize $x^{(0)}$

for $i = 1, \ldots, N_{samp}$ **do**

Proposal sample $z \sim q(\cdot\,|x^{(i)})$

$$x^{(i+1)} = \begin{cases} z & \text{with probability } \alpha(x^{(i)}, z) = \min\left(1, \frac{\pi(z|y)\, q(x^{(i)}|z)}{\pi(x^{(i)}|y)\, q(z|x^{(i)})}\right) \\ x^{(i)} & \text{otherwise} \end{cases}$$

end

If the proposal distribution is Gaussian, $q(\cdot\,|\,\cdot)$ in the numerator and denominator would cancel due to symmetry. The acceptance probability then becomes $\alpha(x^{(i)}, z) = \min\left(1, \frac{\pi(z|y)}{\pi(x|y)}\right)$.

Adaptive Metropolis

The Adaptive Metropolis algorithm [8] provides a way to determine the proposal distribution as more samples are added. The proposal, $q_i(\cdot\,|x^{(0)}, \ldots, x^{(i-1)})$, depends on the covariance of the previously accepted samples to better adapt to the posterior structure. This distribution is defined to be normally distributed with mean $x^{(i-1)}$

and covariance given by

$$C_i = \begin{cases} C_0 & i < i_0 \\ s_d \operatorname{cov}(x^{(0)}, \ldots, x^{(i-1)}) + s_d \epsilon I_d & i \geq i_0 \end{cases} \tag{1.5}$$

where C_0 is some initial proposal covariance defined for the first i_0 samples, cov is the sample covariance, s_d is the scaling factor that depends on dimension d, and ϵ is a small value to ensure that C_i is positive definite.

The choice of C_0 and s_d affect how quickly the proposal distribution adapts to the posterior structure and therefore the acceptance rate that leads to better mixing. As a rule of thumb, $s_d = \frac{2.38^2}{d}$ is sufficient for most implementations. From the properties of sample covariance, the proposal covariance can be computed at low cost using the update formula for $i \geq i_0$,

$$C_{i+1} = \frac{i-1}{i} C_i + \frac{s_d}{i} \big(i \bar{x}^{(i-1)} \bar{x}^{(i-1)\top} - (i+1) \bar{x}^{(i)} \bar{x}^{(i)\top} + x^{(i)} x^{(i)\top} + \epsilon I_d \big). \tag{1.6}$$

1.2.3 Methods of Dimension Reduction

Many of the problems that occur in physical applications are high-dimensional, which significantly impedes the performance of MCMC. Several methods of dimension reduction in the parameter space have been studied with the goal of accelerating MCMC in these problems, including active subspaces, certified dimension reduction, and the likelihood informed subspace.

These methods exploit problem structure to construct a low-dimensional data-informed subspace. The directions of the parameter space are ranked based on how informative the data are to the parameters. This is done using a function of the likelihood integrated over either the prior or posterior distribution. These directions are truncated at some threshold to form a basis for the low-dimensional subspace in which MCMC can be executed.

In the method of active subspaces [2], the active subspace is defined by the eigenvectors of a matrix defined using the negative log-likelihood, or data misfit function,

and integrated over the prior. In certified dimension reduction [15], the data-informed subspace is defined by the eigenvectors of a similar matrix but integrated over the posterior distribution. The method of the likelihood informed subspace [3] determines the low-dimensional subspace using a generalized eigenvalue problem involving the prior covariance and the Hessian of the negative log-likelihood. The Hessian is computed by integrating over the posterior. The likelihood informed subspace is used for dimension reduction in this research and is described in detail in Chapter 3.

1.3 Research Objectives

There are two main objectives for this research.

1. Implement a Bayesian method to solve the remote sensing inverse problem. This allows for an improved characterization of the posterior distribution. Markov chain Monte Carlo (MCMC) is used as the posterior sampling algorithm.

2. Accelerate the Bayesian method to operational speeds. Implementing MCMC on the full dimensional problem is computationally intractable. The likelihood informed subspace (LIS) is used to reduce the dimension of the parameters to increase sampling efficiency of MCMC.

1.4 Thesis Outline

This thesis is organized as follows. Chapter 2 describes the remote sensing problem in detail, including the parameters and models that are used. The current methodology is presented along with the proposed Bayesian approach. The likelihood informed subspace is introduced in Chapter 3. For dimension reduction in the parameter space, a method of performing MCMC within the subspace is presented for this remote sensing problem. Potential applications of the LIS are discussed for data space dimension reduction. For both the parameter and data spaces, the performance of the LIS is evaluated in the linear Gaussian case by comparing the posterior mean and covariance determined from the low-dimensional subspace and the full-dimensional space.

Chapter 4 presents the numerical results of the method of LIS in MCMC, including diagnostics, comparisons of the posterior distribution, and comparisons across different dimensions of the LIS.

Chapter 2

The Atmospheric Remote Sensing Problem

In the remote sensing problem, the goal is to infer a set of parameters $x \in \mathbb{R}^m$, also referred to as the state vector, given a set of observations $y \in \mathbb{R}^n$. Light reflected off the Earth surface undergoes radiative transfer through the atmosphere, which is modelled by a forward function. The satellite observes a radiance that is used to infer the surface reflectance. The current methodology is known in the remote sensing community as optimal estimation (OE). In the proposed approach, we use MCMC to obtain a full posterior distribution. Dimension reduction is also implemented for this Bayesian approach to make it computationally tractable.

Optimal estimation is implemented in the Imaging Spectrometer Optimal Fitting (Isofit) software package on Github [1]. Isofit provides a framework for fitting surface, atmosphere, and instrument models for imaging spectrometer data with flexibility in modelling choice. The parameters associated with the problem setup and the forward function are extracted from Isofit for use in the new methodology. These include the prior on the parameters, the observation noise model, and the forward model itself.

[1] https://github.com/isofit/isofit

2.1 Parameters and Data

In this particular remote sensing problem, the imaging spectrometer on the satellite observes $n = 425$ radiance values corresponding to equally spaced wavelengths from 350 to 2500 nm. The $m = 427$ parameters consist of 425 surface parameters corresponding to the same wavelengths and two additional parameters that describe the atmospheric conditions.

$$x = [x_{surf}\ x_{atm}]^\top \in \mathbb{R}^{427}, \qquad y \in \mathbb{R}^{425}$$

The two atmospheric parameters are denoted as $x_{atm} = [x_{AOD}\ x_{H2O}]$. In practice, there are many more variables in the atmosphere, but most of them can be predicted using climatology. Aerosol Optical Depth (AOD) is the atmospheric concentration of aerosols at 550 nm and is a measure of the scattering of radiation. This is also referred to as the Aerosol Optical Thickness (AOT) [13]. The second atmospheric parameter is the column precipitable water vapour (cm), which is a measure of the amount of water in a vertical column of the atmosphere. These atmospheric parameters are generally difficult to predict in practice and greatly influence the retrieved surface reflectances.

2.1.1 Prior

The prior on the parameters has the following mean and covariance structure.

$$\mu_{pr} = \begin{bmatrix} \mu_{surf} \\ \mu_{atm} \end{bmatrix}, \quad \Gamma_{pr} = \begin{bmatrix} \Gamma_{surf} & 0 \\ 0 & \Gamma_{atm} \end{bmatrix} \tag{2.1}$$

The prior on the two atmospheric parameters are independent. For this particular problem, they are fixed. The AOD parameter has prior mean 0.05 with variance 0.04, and the H2O parameter has prior mean 1.75 with variance 0.025.

The procedure that was used to determine the prior on the surface parameters is as follows. Data from libraries of over 1400 historical reflectance spectra are clus-

tered into 8 subpopulations, each represented by a multivariate Gaussian distribution. These subpopulations correspond to different types of terrain on the Earth surface that have similar characteristics, such as vegetation or aquatic environments. The distribution with the least Mahalanobis distance from the state estimate is chosen as the surface prior. This prior on the surface parameters is multivariate Gaussian of dimension $n = 425$ and is used for the retrieval and for the Bayesian approach to this remote sensing problem.

2.1.2 Forward Model

The forward model approximates radiative transfer through the Earth atmosphere. The MODTRAN 6.0 Radiative Transfer Model [1] is used to generate a lookup table for a set of reference atmospheric conditions, varying the two atmospheric parameters. The forward model then uses linear interpolation to approximate the radiative transfer given a state vector. The prior on the atmospheric parameters have variances larger than the lookup table range.

2.1.3 Noise Model

The observation uncertainty covariance is obtained using the Isofit code and accounts for instrument noise and uncertainty due to unknown parameters. The covariance matrix is the sum of these two contributions,

$$\Gamma_{obs} = \Gamma_y + K_b \Gamma_b K_b^\top, \tag{2.2}$$

where Γ_y represents randomness due to the instrument, such as photon and readout noise, Γ_b represents the uncertainty in the observation unknowns, and K_b is the Jacobian of the observations with respect to these unknowns. The instrument noise covariance Γ_y is generally a diagonal matrix with each entry having the same signal-to-noise ratio. The second term has slight off-diagonal correlations.

2.1.4 Radiance Measurements

The imaging spectrometer on the satellite observes a radiance spectrum for each pixel in the image. This radiance is a function of the parameters with added noise, and is modelled by

$$y = f(x) + \epsilon, \tag{2.3}$$

where $\epsilon \sim \mathcal{N}(0, \Gamma_{obs})$ is drawn from the noise model.

2.2 Current Methodology

Optimal estimation is currently used in many remote sensing problems. In OE, the posterior distribution is characterized by a Gaussian with mean computed using MAP estimation and covariance computed using Laplace approximation at the MAP estimate. The log prior and log likelihood are combined into the cost function

$$\chi^2(x) = \frac{1}{2}(x - \mu_{pr})^\top \Gamma_{pr}^{-1}(x - \mu_{pr}) + \frac{1}{2}(y - f(x))^\top \Gamma_{obs}^{-1}(y - f(x)), \tag{2.4}$$

which is minimized using nonlinear least squares optimization to produce the MAP estimate $\hat{x}$. The covariance is approximated using a local linearization,

$$\hat{\Gamma}_{pos} = (K^\top \Gamma_{obs}^{-1} K + \Gamma_{pr}^{-1})^{-1}, \tag{2.5}$$

where $K = \nabla f(\hat{x})$ is the Jacobian of the forward model evaluated at the MAP estimate.

2.2.1 Practical Considerations

There are certain issues in the remote sensing problem that arise in practice. We start by examining the sample retrieval shown in Figure 1-1. The wavelengths corresponding to the breaks in the retrieved reflectances are known as the deep water

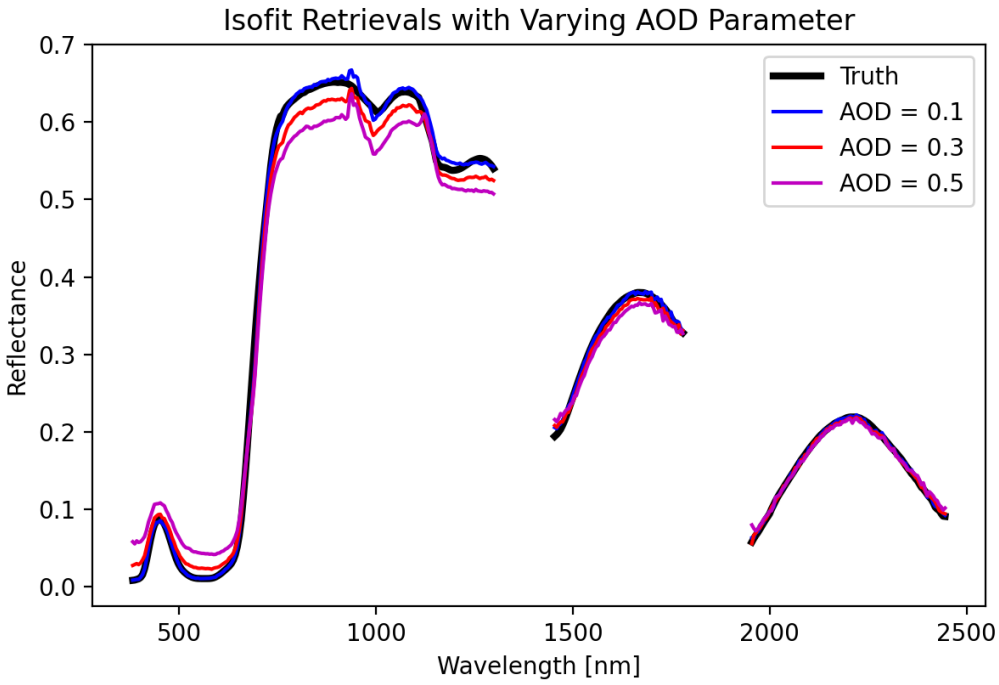


Figure 2-1: Retrievals with various AOD parameters

spectra. For these sets of wavelengths, almost all of the radiation is absorbed by water vapour, and light reflected off the Earth surface is therefore not able to reach the satellite. Most of the radiance the instrument picks up in these regions would be due to noise. In the remote sensing community, retrieved results in the deep water spectra are generally ignored and not displayed. There are four such bands of varying width present in our remote sensing problem setup: near 350 nm, 1300 nm, 1800 nm, and 2500 nm.

For higher values of the atmospheric parameters, specifically the AOD parameter, the retrieval becomes more difficult. Physically, an AOD parameter close to 0 corresponds to clear skies, and a higher AOD parameter of around 0.5 or higher corresponds to hazy skies. Haze contributes to the scattering of light in the atmosphere, which adds noise to the radiance observations. In Figure 2-1, it is evident that the retrieval is much worse with an AOD parameter equal to 0.5.

2.2.2 Improvements to Current Approach

The main issue with the current methodology is that the retrieval returns a MAP estimate, which is a point estimate rather than a full distribution. The uncertainty is then approximated using the Laplace approximation at this point estimate. While this approach would be a good estimate for a problem that is approximately linear Gaussian, this is likely not the case for a nonlinear problem. The Laplace approximation only takes into account the local linearization of the forward model. If the forward model exhibits strong nonlinearities around the MAP estimate, the Laplace approximation would be extremely sensitive to small changes in the MAP estimate.

Without a proper estimate of the uncertainty, it is difficult to justify the validity of the retrieval results. This is the main motivation for turning to a Bayesian approach for this problem.

2.3 A Bayesian Approach to Remote Sensing

Instead of a point estimate, we are interested in obtaining the full posterior distribution of the retrieval. The Bayesian approach to remote sensing involves a sampling algorithm, usually Markov chain Monte Carlo (MCMC), to produce posterior samples.

The high-level problem setup is depicted in Figure 2-2. The true state consists of surface reflectances and atmospheric parameters and is endowed with a Gaussian prior. The radiance observation is a result of the radiative transfer of surface reflectances through the atmosphere with some added noise. The objective is to determine the posterior distribution of the parameters.

Sampling from the posterior is generally much more computationally expensive than obtaining the MAP estimate through optimization. Sampling methods such as MCMC are sensitive to parameter dimension, and the main challenge of using a Bayesian method for this high-dimensional problem is to enable tractable computation. We investigate a method of dimension reduction using the likelihood informed subspace to increase the sampling efficiency of MCMC.

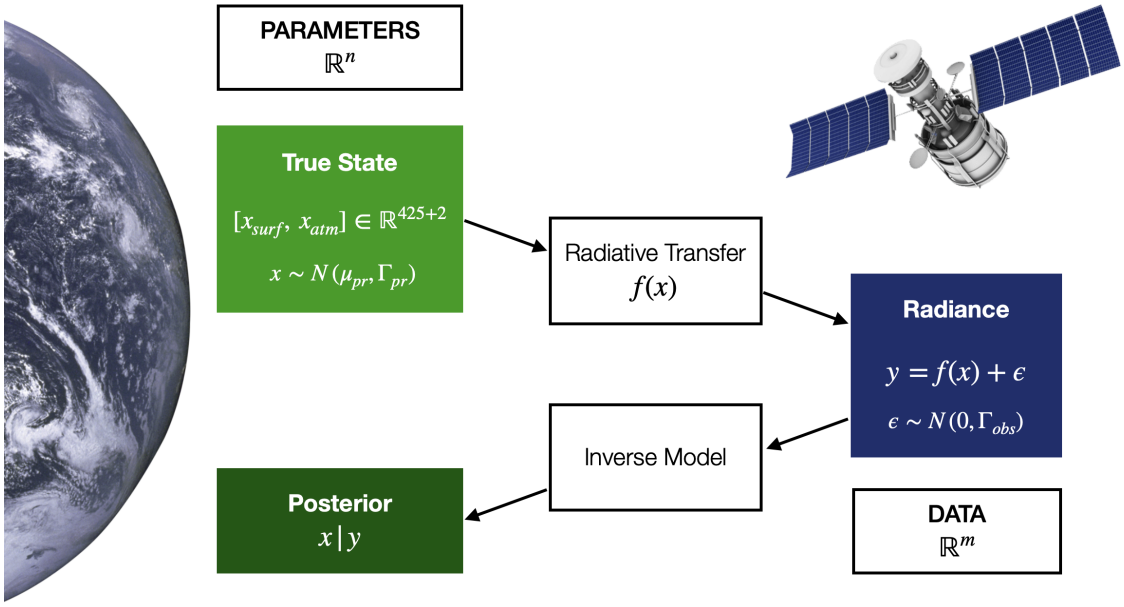


Figure 2-2: Remote sensing problem setup

2.4 Linearization of the Forward Model

If the atmospheric parameters are fixed, the retrieval of the surface parameters is mostly a linear problem. In other words, the two atmospheric parameters are the main cause of nonlinearities in the forward model.

As a result, a linearized model was developed to approximate the nonlinear radiative transfer forward model. In this research, it is used to simplify the method of the likelihood informed subspace for this particular problem. This is described in detail in Chapter 3. The linear model can also be applied in other ways. In a multifidelity approach to MCMC, it is used as the first pass in a delayed acceptance scheme. It can also be used to generate computationally efficient approximations of the forward model in a general context.

It is important to note that this linearization is not used to replace the forward model entirely. It is simply used to enhance parts of other methods using this additional model of lower fidelity. We are still interested in implementing a Bayesian algorithm for a nonlinear problem.

To create the linear model, regression with various forms of regularization were

considered, including linear least squares, Ridge, and LASSO regression. LASSO regression was chosen because the regularization term in the objective drives many of the coefficients to zero, resulting in a sparse linear operator matrix [14].

2.4.1 LASSO Regression

To perform the regression, samples $x^{(i)}$ are drawn from the prior distribution. The corresponding samples $y^{(i)}$ are generated by running the parameter samples through the forward model and adding noise.

$$y^{(i)} = f(x^{(i)}) + \epsilon^{(i)},\ i = 1 \dots N, \tag{2.6}$$

where $\epsilon^{(i)} \sim N(0, \Gamma_{obs})$. These samples are divided into two random sets of 25000 training samples and 5000 test samples.

The samples are first scaled to zero mean and unit variance

$$\tilde{x}^{(i)} = \frac{x^{(i)} - \mu_x}{\sigma_x},\ \ \tilde{y}^{(i)} = \frac{y^{(i)} - \mu_y}{\sigma_y} \tag{2.7}$$

where $\mu_x \in \mathbb{R}^m$ is the mean of the training samples, $\sigma_x \in \mathbb{R}^m$ is the square root of the variance, and $\mu_y \in \mathbb{R}^n$ and $\sigma_y \in \mathbb{R}^n$ are the corresponding mean and variance for the radiance training samples. In the remote sensing problem, the parameter dimension is $m = 427$ and the dimension of the data is $n = 425$.

In a general sense, we seek a linear operator G to approximate the forward model $f(x) \approx Gx$. Regression is performed separately for each of the n radiances. Using the scaled samples, the objective function for the scaled operator $\tilde{G}$ is given by

$$\tilde{G}_i = \arg\min_{\phi} \left\| \tilde{y}^{(i)} - \phi^\top \tilde{x}^{(i)} \right\|_2^2 + \lambda \|\phi\|_1,\ \ i = 1 \dots n, \tag{2.8}$$

where λ is the regularization parameter and $\tilde{G} = [\tilde{G}_1 \dots \tilde{G}_n]$.

The linear operator $\tilde{G}$ is computed using these scaled samples and the scaled predicted radiance is $\hat{\tilde{y}} = \tilde{G}\tilde{x}$. The predicted radiance in canonical units using this

linear operator is

$$\hat{y} = \sigma_y \tilde{G} \sigma_x^{-1}(x - \mu_x) + \mu_y. \tag{2.9}$$

2.4.2 Tuning the Regularization Parameter

The regularization parameter λ was tuned by analyzing the error from the regression. The training error is defined to be the mean squared error of the training samples and the predicted samples,

$$\epsilon_{train} = \frac{1}{N_{train}} \sum_{i=0}^{N_{train}} \left\| \tilde{y}^{(i)} - \tilde{G}\tilde{x}^{(i)} \right\|^2. \tag{2.10}$$

Similarly, the generalization error is defined using the test samples,

$$\epsilon_{gen} = \frac{1}{N_{test}} \sum_{i=0}^{N_{test}} \left\| \tilde{y}_{test}^{(i)} - \tilde{G}\tilde{x}_{test}^{(i)} \right\|^2. \tag{2.11}$$

The regularization parameter was tuned by repeating the regression using several parameters ranging from 10^{-5} to 10^{0} and analyzing the generalization error. In general, the generalization error was lowest for orders of magnitude around 10^{-3} and 10^{-2} for all n radiances, with very little variation in this range of λ. The value of $\lambda = 10^{-3}$ was therefore chosen from this analysis.

The final values of training and generalization error are plotted in Figure 2-3. Since the variables are scaled, the error can be interpreted as a percent. The generalization error is below 10 percent for much of the radiances, which is acceptable. However, there are regions where the error is more significant, such as around 1000 nm and around the deep water spectra at 1500 and 2000 nm. This is expected to be caused by the nonlinearities from the atmospheric parameters that are unable to be captured by a linear model.

Figure 2-4 compares the linearized model with the nonlinear radiative transfer model for a sample parameter "truth". For this particular example, the radiance prediction is very close to the nonlinear model.

Figure 2-3: Errors in LASSO regression

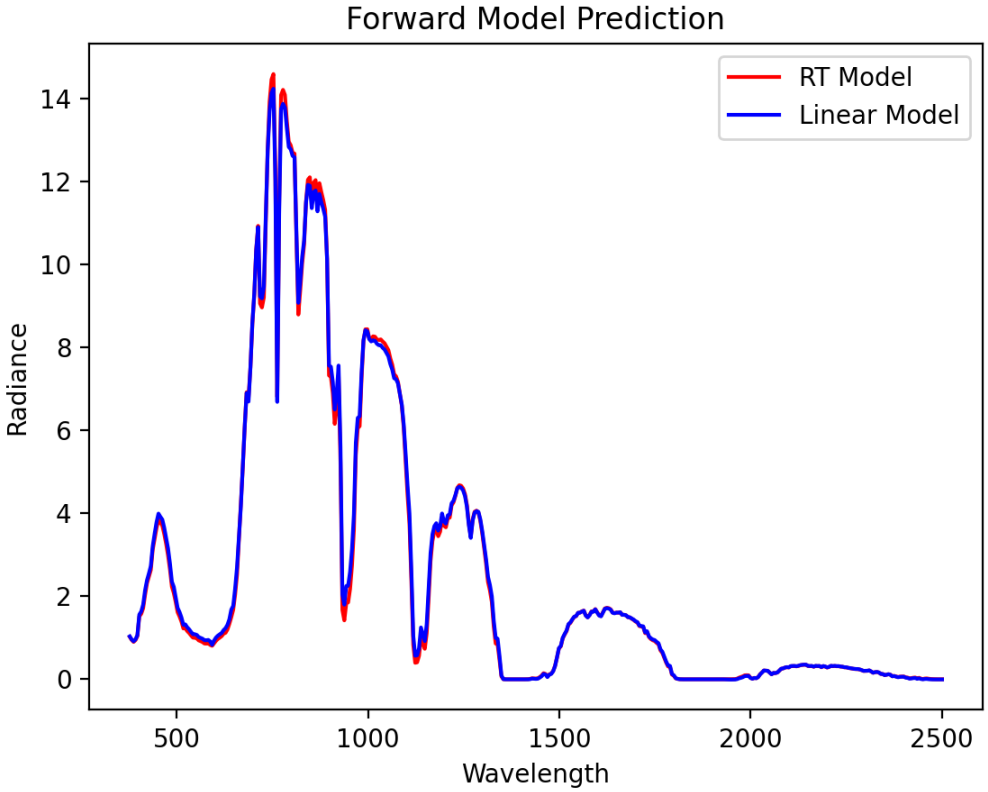


Figure 2-4: Comparison of linear and nonlinear forward models

2.4.3 Linear Operator

The structure of the linear operator $\tilde{G}$ obtained from LASSO regression is significant in understanding the remote sensing problem. The sparsity plot is shown in Figure 2-5. The horizontal axis is the index of the $m = 427$ parameters and the vertical axis corresponds to the $n = 425$ radiances. The points corresponding to the atmospheric parameters and are circled in red.

The structure of $\tilde{G}$ is consistent with the qualitative features of the forward model that are known. Except for the rows around index 200 and 300, which represent the deep water spectra, the banded structure suggests that radiances are mainly affected by parameters close in wavelength, which is consistent with the physical interpretation. Apart from small scattering effects in the atmosphere, we would not expect much mixing across wavelength channels, especially those that are far apart. The two atmospheric parameters influence most of the radiances across the entire spectrum, which also adheres to the physical interpretation of reflectances travelling through the atmosphere to reach the instrument.

Analyzing the structure not only reinforces our understanding of the forward model, but it also allows exploitation of the most prominent features. For example, the sparse nature of the matrix allows for fast approximations of the forward model when used in methods such as multifidelity MCMC while retaining the key features of the forward model.

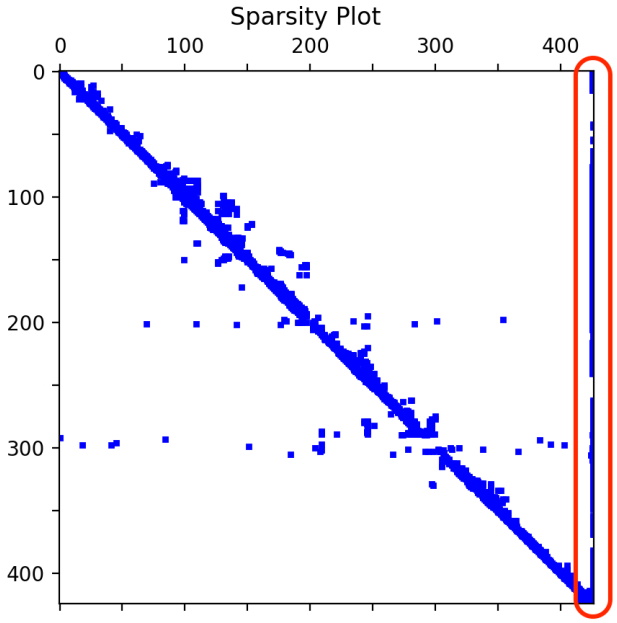


Figure 2-5: Sparsity plot of the linear operator

Chapter 3

Dimension reduction using the likelihood informed subspace

The high dimensionality of the remote sensing problem poses a challenge for the Bayesian approach. A reduction in the number of parameters can significantly increase the sampling efficiency of MCMC. T. Cui et al [3] proposes a method for dimension reduction for such inverse problems using the likelihood informed subspace. This subspace can be defined in either the parameter space or the data space [9] [6]. We focus mainly on parameter space dimension reduction so that we can sample the posterior parameters using MCMC.

In this section, the likelihood informed subspace (LIS) is introduced along with its integration with the MCMC algorithm. The performance of the LIS is evaluated using metrics for both the posterior mean and covariance. Finally, dimension reduction in the data space is discussed.

3.1 Parameter Space Dimension Reduction

We first present a conceptual overview of the likelihood informed subspace applied to the parameter space. Then we describe the general methodology of parameter space dimension reduction for nonlinear inverse problems, followed by a gradient-free approach using a linearized forward model.

3.1.1 Overview

The idea of the likelihood informed subspace is to determine a subspace in which the data is more informed than the prior. This is most prominent when the prior-to-posterior update has a low-rank structure. Generally, the variance of the posterior is reduced with respect to the prior after the data is observed because the information gained from the observation reduces the uncertainty in the parameters. However, the amount of variance reduction differs in each direction of the parameter space. The likelihood informed subspace captures the directions with the greatest reduction in variance. This concept is visually depicted in Figure 3-1.

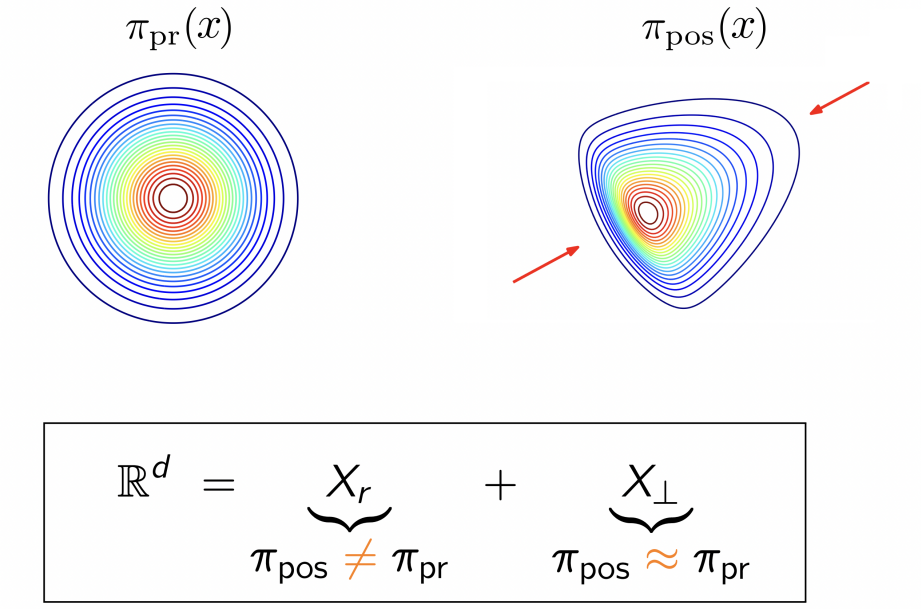


Figure 3-1: Visualization of the likelihood informed subspace

The parameter space can be split into the likelihood informed subspace, denoted by X_r, and the complementary subspace, denoted by $X_\perp$. The likelihood informed subspace contains the directions that are most informed by the data. The low-rank prior-to-posterior update affects the variance in these directions the most. In the complementary subspace, the posterior remains largely unchanged from the prior. When applied to a posterior sampling method, the idea is to divide the parameter space in this way so that sampling only needs to be performed in the low-dimensional

likelihood-informed subspace X_r. Since the posterior in the directions of the complementary subspace remain approximately equal to the prior, the samples for $X_\perp$ can simply be drawn from the prior distribution.

3.1.2 Construction of the Subspace

Consider the following Rayleigh ratio involving the Fisher information matrix $H(x) = \nabla f(x)^\top \Gamma_{obs}^{-1} \nabla f(x)$ and the prior precision matrix,

$$R(x) = \frac{x^\top H(x) x}{x^\top \Gamma_{pr}^{-1} x}. \tag{3.1}$$

The numerator is a measure of information gained from the observation in the x direction of the parameter space. This Rayleigh ratio provides a quantitative valuation of the information gained from the data relative to the prior. The likelihood informed subspace aims to determine the directions in which this Rayleigh quotient is maximized, which maximizes the information gained through observing the data. This concept arises in the form of a generalized eigenvalue problem that involves $H(x)$ and Γ_{pr}^{-1}.

The likelihood informed subspace is constructed using the eigenvectors of a generalized eigenvalue problem involving the prior and observation noise covariances. For the nonlinear case, the local generalized eigenvalue problem is given by

$$H(x)\phi_i = \lambda_i \Gamma_{pr}^{-1} \phi_i, \tag{3.2}$$

where $H(x)$ is the Hessian of the data misfit function at x and is equivalent to the Fisher information matrix. The rank-r local LIS basis is defined using the first r eigenvectors v_i.

The eigenvectors v_i corresponding to the largest eigenvalues represent the directions of the parameters in which the data is most informative in determining the parameters relative to the prior. In other words, most of the information from the data is captured in the first r directions of the parameter space.

Since the Hessian varies over x, the average Hessian over the parameter space is required to construct a global subspace. For this method, the average is computed by taking the expectation of the Hessian over a set of posterior samples $\{x^{(k)}\}, k = 1 \dots M$. This is done by first approximating the local Hessian at each posterior sample $x^{(k)}$ by writing it as an eigendecomposition truncated at some $L(k)$,

$$H(x^{(k)}) \approx \sum_{i=1}^{L(k)} \bar{\lambda}_i^{(k)} \bar{v}_i^{(k)} \bar{v}_i^{(k)\top}. \tag{3.3}$$

The global Hessian is then computed using Monte Carlo over the approximate local Hessian.

The eigenvalue problem for the global likelihood informed subspace becomes

$$\left(\frac{1}{M} \sum_{k=1}^{M} \sum_{i=1}^{L(k)} \bar{\lambda}_i^{(k)} \bar{v}_i^{(k)} \bar{v}_i^{(k)\top} \right) \phi_j = \lambda_j \Gamma_{pr}^{-1} \phi_j. \tag{3.4}$$

The global LIS basis is given by $\Phi_r = [\phi_1, \dots, \phi_r]$. The complementary basis is given by $\Phi_\perp = [\phi_{r+1}, \dots, \phi_m]$. We also define the matrices $\Theta_r = \Gamma_{pr}^{-1} \Phi_r$ and $\Theta_\perp = \Gamma_{pr}^{-1} \Phi_\perp$ that are used to transform the parameters from the canonical parameter space to the respective subspaces.

3.1.3 MCMC Sampling in the Low Dimensional LIS

Given the basis for the likelihood informed subspace, the next step is to allow MCMC to sample within this low-dimensional subspace of the parameter space. The parameters x can be split into the LIS and complementary components using the projector $\Pi_r = \Phi_r \Theta_r^\top$.

$$x = \Pi_r x + (I - \Pi_r) x \tag{3.5}$$

The matrices Θ_r and $\Theta_\perp$ are used to represent x in the lower-dimensional LIS and complementary coordinates.

$$x_r = \Theta_r^\top x, \quad x_\perp = \Theta_\perp^\top x \tag{3.6}$$

The matrices Φ_r and $\Phi_\perp$ transform these parameters back to the original parameter space. Equation 3.5 can also have the form

$$x = \Phi_r x_r + \Phi_\perp x_\perp \tag{3.7}$$

The prior can be written as a product of the priors in the LIS and complementary components.

$$\pi(x) = \pi_r(x_r)\pi_\perp(x_\perp) \tag{3.8}$$

The posterior distribution can be approximated using the likelihood conditioned on the parameters in the low dimensional subspace instead of the full parameter space.

$$\tilde{\pi}(x|y) \propto \pi(y|x_r)\pi(x) = \pi(y|x_r)\pi_r(x_r)\pi_\perp(x_\perp) \tag{3.9}$$

The LIS parameters x_r can then be sampled from the low rank posterior, and the parameters in the complementary subspace, $x_\perp$ are sampled from the complement prior.

$$x_r \sim \tilde{\pi}(x_r|y) \propto \pi(y|x_r)\pi_r(x_r) \tag{3.10}$$

$$x_\perp \sim \pi_\perp(x_\perp). \tag{3.11}$$

The full rank posterior samples are simply the sum of these two components transformed back to the original parameter space, as written in Equation 3.8.

In this way, MCMC sampling is only performed on x_r, which has dimension r instead of m. Depending on the choice of r, this can have a significant effect on the

sampling efficiency.

The process of implementing MCMC using the likelihood informed subspace is described in Algorithm 2. In practice, the samples are centered at the beginning of each MCMC chain to be equal to zero in the parameter space.

Algorithm 2: MCMC in the likelihood informed subspace

Initialize posterior sample set $\mathcal{X}_r^{(1)} = \{x^{(0)}\},\ x^{(0)} = x_{map}$

Initialize complementary sample set $\mathcal{X}_\perp^{(1)}$

for $j = 1, \ldots, J$ **do**

- Construct global LIS basis over $\mathcal{X}_r^{(j)}$ and obtain Φ_r, $\Phi_\perp$, Θ_r
- Run $L(j)$ samples of MCMC chain, $\{x_r^{(1)}, \ldots, x_r^{(L(j))}\}$
- Update posterior sample set, $\mathcal{X}_r^{(j+1)} = \mathcal{X}_r^{(j)} \cup \{\Phi_r x_r^{(1)}, \ldots, \Phi_r x_r^{(L(j))}\}$
- Obtain $L(j)$ samples from the complement prior, $\{x_\perp^{(1)}, \ldots, x_\perp^{(L(j))}\}$
- Update complementary sample set, $\mathcal{X}_\perp^{(j+1)} = \mathcal{X}_\perp^{(j)} \cup \{\Phi_\perp x_\perp^{(1)}, \ldots, \Phi_\perp x_\perp^{(L(j))}\}$

The set of posterior samples in the original parameter space is the element-wise sum of the sets $\mathcal{X}_r^{(J)}$ and $\mathcal{X}_\perp^{(J)}$.

3.2 LIS using the linearized model

The construction of the likelihood informed subspace can be simplified if the forward model is linear. A linear model eliminates the need for local gradients of the forward model. Furthermore, the process of MCMC sampling within the LIS can be simplified since samples from the posterior are no longer required to determine the LIS basis. This can be exploited in our remote sensing problem given the relatively good fit of the linear model, as determined in Chapter 2.

If the forward model is linear, $f(x) = Gx$, the Hessian simplifies to a constant for all x,

$$H = G^\top \Gamma_{obs}^{-1} G. \tag{3.12}$$

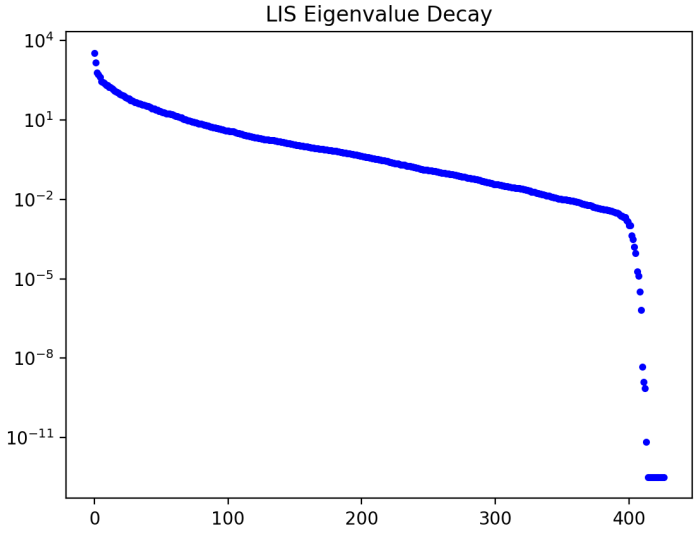


Figure 3-2: Eigenvalue decay for the LIS eigenvalue problem

The generalized eigenvalue problem can then be written as

$$G^{\top}\Gamma_{obs}^{-1}G\phi_i = \lambda_i\Gamma_{pr}^{-1}\phi_i. \tag{3.13}$$

Unlike for the general case, this eigenvalue problem only needs to be solved once since it is not a function of x. The resulting basis given by $\Phi_r = [\phi_1 \dots \phi_r]$ is the global LIS basis. Figure 3-2 plots the eigenvalue decay for our remote sensing problem using the linearized model.

Obtaining a global basis using the linearized model greatly simplifies the implementation of MCMC in the likelihood informed subspace. This is described in Algorithm 3.

Algorithm 3: MCMC in the LIS obtained from the linear model

1. Solve generalized eigenvalue problem, obtain matrices Φ_r, $\Phi_\perp$, Θ_r
2. Initialize MCMC chain $x_r^{(0)} = \Theta_r^\top x_{map}$
3. Run N_{samp} samples of the MCMC chain, $\{x_r^{(1)}, \ldots, x_r^{(N_{samp})}\}$
4. Obtain N_{samp} samples from the complement prior, $\{x_\perp^{(1)}, \ldots, x_\perp^{(N_{samp})}\}$
5. Project back to the full-dimensional parameter space

$$x^{(i)} = \Phi_r x_r^{(i)} + \Phi_\perp x_\perp^{(i)},\ i = 1, \ldots, N_{samp}$$

For this problem, we exploit the property of the forward model that it is mostly linear save for the two atmospheric parameters. Instead of computing local linear approximations of the gradients of the forward model, we use the global linear approximation of the forward model gradient. There are two main benefits to doing this.

- Since the eigenvalue problem produces a global basis for the likelihood informed subspace, the construction of this basis can be treated as a preprocessing step. Once the basis is computed, only one single MCMC chain is required to obtain all posterior samples. The basis does not need to be reconstructed based on newly obtained posterior samples.
- The resulting method for MCMC in the likelihood informed subspace is a gradient-free method. This is useful when gradients of the forward model are unavailable or expensive to compute.

3.3 Evaluating the posterior from LIS

The optimality of the likelihood informed subspace for the linear Gaussian case is proven in [12]. Although the linear model in the remote sensing problem is only used to determine the subspace and not directly used to determine the posterior, this section investigates the case for which the posterior mean and covariance are computed with the linear model using Equations 1.3 and 1.4. This is done to demonstrate

the improvement of the likelihood informed subspace over the common method of principal component analysis (PCA). The linear Gaussian assumption does not apply to remote sensing application because the nonlinear model leads to a non-Gaussian posterior distribution.

For the linear Gaussian case, the posterior covariance can be computed as a low-rank update to the prior. Using the notation in Equation 3.13, the analytical expression for the rank-r posterior covariance is

$$\Gamma_{pos}^{LIS} = \Gamma_{pr} - \sum_{i=1}^{r} \frac{\lambda_i}{\lambda_i + 1} \phi_i \phi_i^\top. \tag{3.14}$$

For a rank-zero subspace, the posterior is equal to the prior. As more directions are added in the subspace, the variance in those directions are reduced from prior to posterior given the data.

We compare the performance of LIS and PCA for a sequence of subspace dimensions with respect to the full dimensional posterior. The main difference between PCA and LIS is that while PCA identifies the principal directions of the parameter space, LIS identifies the directions in the parameter space that are most informed by the data, which leads to an improved posterior. The eigenvalue problem for PCA is set up such that the construction of the posterior covariance has the same form.

$$\big(\Gamma_{pr} - \Gamma_{pos}\big) w_i = \gamma_i w_i \tag{3.15}$$

The regular eigenvalue problem for PCA only involves the prior and posterior covariances and does not include the observation noise covariance. The eigenvalues are plotted in Figure 3-3.

Using PCA, the rank-r posterior covariance is

$$\Gamma_{pos}^{PCA} = \Gamma_{pr} - \sum_{i=1}^{r} \gamma_i w_i w_i^\top. \tag{3.16}$$

The low-rank posterior covariance is then used to determine the posterior mean

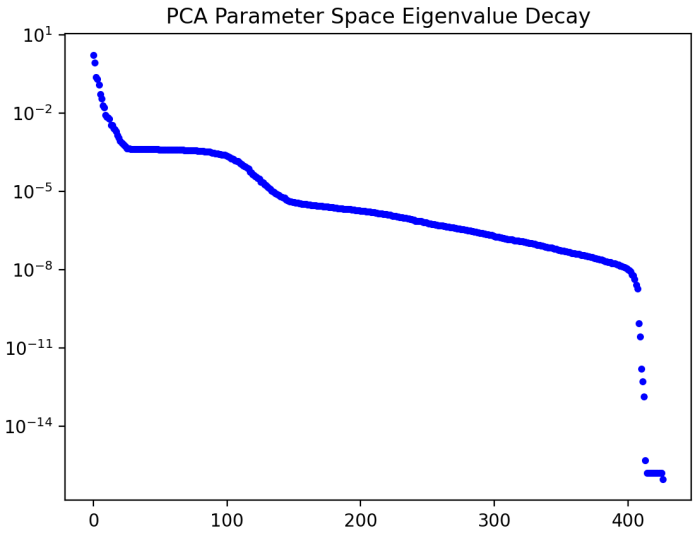


Figure 3-3: Eigenvalue decay of the parameter space PCA problem

given an instance of the data.

$$\mu_{pos}^{LIS} = \Gamma_{pos}^{LIS}(G^{\top}\Gamma_{obs}^{-1}y + \Gamma_{pr}^{-1}\mu_{pr}) \tag{3.17}$$

$$\mu_{pos}^{PCA} = \Gamma_{pos}^{PCA}(G^{\top}\Gamma_{obs}^{-1}y + \Gamma_{pr}^{-1}\mu_{pr}) \tag{3.18}$$

3.3.1 Posterior covariance

To demonstrate the improvement of LIS over PCA, we first compare the posterior covariance computed using the both low-rank subspaces with the full-rank posterior covariance as determined in Equation 1.4. We use the Forstner distance metric, which is a measure of similarity in the class of symmetric positive definite matrices.

Given two covariance matrices Γ_A and Γ_B, let (σ_i) be the sequence of eigenvalues in the generalized eigenvalue problem

$$\Gamma_A z_i = \sigma_i \Gamma_B z_i. \tag{3.19}$$

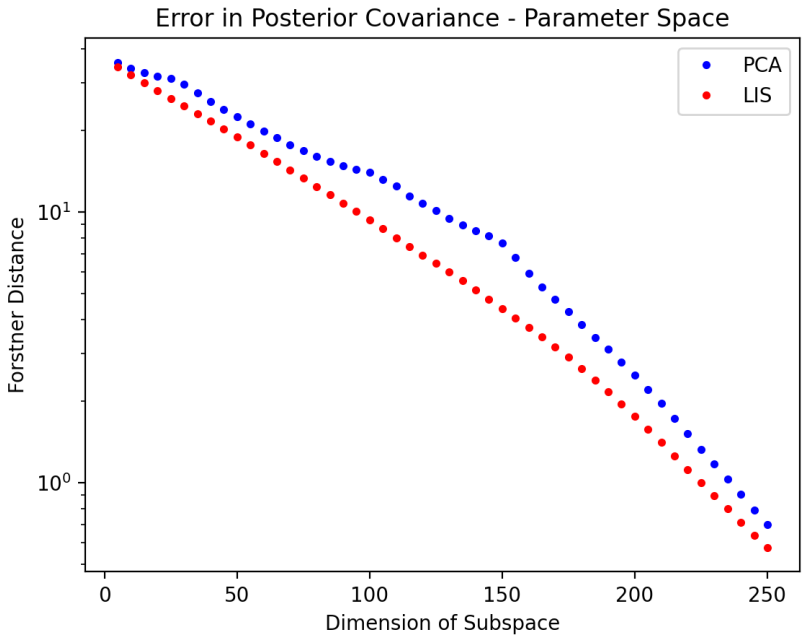


Figure 3-4: Forstner distance in posterior covariance for the parameter space

The Forstner distance is defined to be

$$d_{\mathcal{F}}(\Gamma_A, \Gamma_B) = \sqrt{\sum_i \ln^2(\sigma_i)}. \tag{3.20}$$

The comparison of the posterior covariance determined from the low-dimensional subspace and the full-dimensional posterior covariance is displayed in Figure 3-4 for dimensions 5 to 250. The error in posterior covariance computed in the likelihood informed subspace is consistently lower than in PCA.

3.3.2 Posterior mean

The performance of the low-rank posterior mean can be evaluated using the Bayes risk, which is the expected value of some loss function over the posterior. In this case, we define the loss function with respect to the true parameter x weighted by the posterior precision matrix. This is done so that a larger absolute difference in the two posterior means does not indicate a larger error if the covariance is also large.

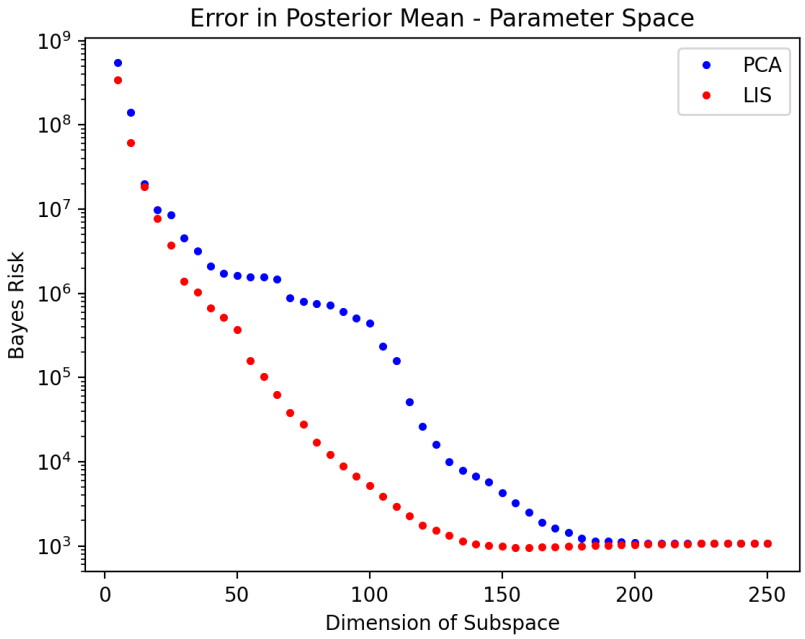


Figure 3-5: Bayes risk in posterior mean for the parameter space

We use the following expression for Bayes risk,

$$R(\mu_{pos}^r, x) = \mathbb{E}_y\left[(\mu_{pos}^r(y) - x)^\top \Gamma_{pos}^{-1}(\mu_{pos}^r(y) - x)\right], \tag{3.21}$$

where μ_{pos}^r is the low-rank posterior mean obtained from either LIS or PCA and x is the true parameter used to generate the data y. The data points were generated in the same way as the training and test samples in Chapter 2. The parameters $x^{(i)}$ are sampled from the prior, and the data $y^{(i)}$ are obtained by running the parameters through the forward model and adding noise.

The comparison of posterior means using Bayes risk is shown in Figure 3-5. Computing the posterior in the LIS subspace results in lower error up to a dimension of around 200. Note that since the posterior mean determined in the low-rank subspace is being compared to the truth instead of the full-rank subspace, the error does not approach zero as the dimension reaches full rank.

These plots demonstrate the effectiveness of the likelihood informed subspace at capturing the important information pertaining to the posterior within a low-

dimensional subspace. The LIS requires less dimensions, which is much more beneficial in terms of sampling efficiency. Although this study was only done for the linear Gaussian case, the idea can be extended to a nonlinear problem such as the remote sensing problem.

3.4 Data Space Dimension Reduction

When applied to the parameter space, the likelihood informed subspace allows for more efficient posterior sampling. This method can also be applied to the data space for other applications such as data compression. In this section, we construct the likelihood informed subspace in the data space and evaluate its performance when used to calculate the posterior mean and covariance. For simplicity, this is done only for the linear Gaussian case using the linearized forward model G.

In the data space, we solve a generalized eigenvalue problem that is the dual of the parameter space problem [9] [6]. The eigenvalues are identical, but the eigenvectors are the directions of the data space in which the data are most informative to the parameters. The eigenvalue problem is given by

$$\Gamma_y \psi_i = \lambda_i^* \Gamma_{obs} \psi_i, \tag{3.22}$$

where $\Gamma_y = G\Gamma_{pr}G^\top + \Gamma_{obs}$ is the marginal covariance of the data. The rank-r LIS basis for the data space is given by $\Psi_r = [\psi_1 \dots \psi_r]$. Note that $\lambda_i^* = \lambda_i + 1$ in this particular problem setup. If we take the eigenvalue pencil $(G\Gamma_{pr}G^\top, \Gamma_{obs})$, the eigenvalues would be (λ_i).

We again compare the likelihood informed subspace to principal component analysis in the data space. PCA in the data space involves the regular eigenvalue problem using the marginal covariance of the data

$$\Gamma_y \bar{w}_i = \bar{\gamma}_i \bar{w}_i, \tag{3.23}$$

where the bar denotes the data space as opposed to the parameter space. The PCA

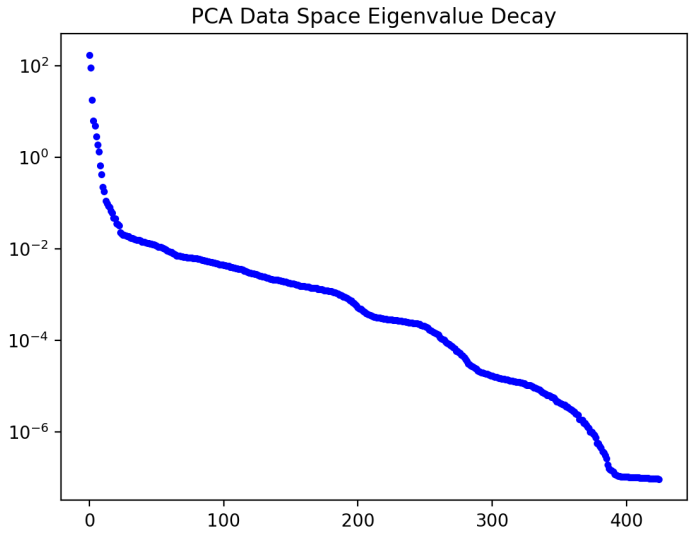


Figure 3-6: Eigenvalue decay of the data space PCA problem

basis in the data space is $\bar{W}_r = [\bar{w}_1 \dots \bar{w}_r]$. The eigenvalues are plotted in Figure 3-6.

Given the basis of the subspace, the rank-n model $y = Gx + \epsilon$ can be projected onto the rank-r subspace. This is demonstrated using the LIS basis Ψ_r and is repeated for the PCA basis $\bar{W}_r$. The rank-r model is

$$y_r = G_r x + \epsilon_r, \tag{3.24}$$

where $y_r = \Psi_r^\top y$, $G_r = \Psi_r^\top G$, and $\epsilon_r = \Psi_r^\top \epsilon$. Note that the forward model becomes $\Psi_r^\top G$ and the observation noise covariance becomes $\Gamma_{obs,r} = \Psi_r^\top \Gamma_{obs} \Psi_r$. The matrices involving the parameters are unchanged. The posterior covariance obtained using LIS is then

$$\Gamma_{pos}^{LIS} = \left(G_r \Gamma_{obs,r} G_r^\top + \Gamma_{pr}^{-1}\right)^{-1} = \left[\left(\Psi_r^\top G\right)\left(\Psi_r^\top \Gamma_{obs} \Psi_r\right)\left(\Psi_r^\top G\right)^\top + \Gamma_{pr}^{-1}\right]^{-1}. \tag{3.25}$$

The posterior mean is computed using the posterior covariance as in the parameter, but now with the low-rank forward model and observation noise covariance.

$$\mu_{pos}^{LIS} = \Gamma_{pos}^{LIS}(G_r^\top \Gamma_{obs,r}^{-1} y + \Gamma_{pr}^{-1} \mu_{pr}) \tag{3.26}$$

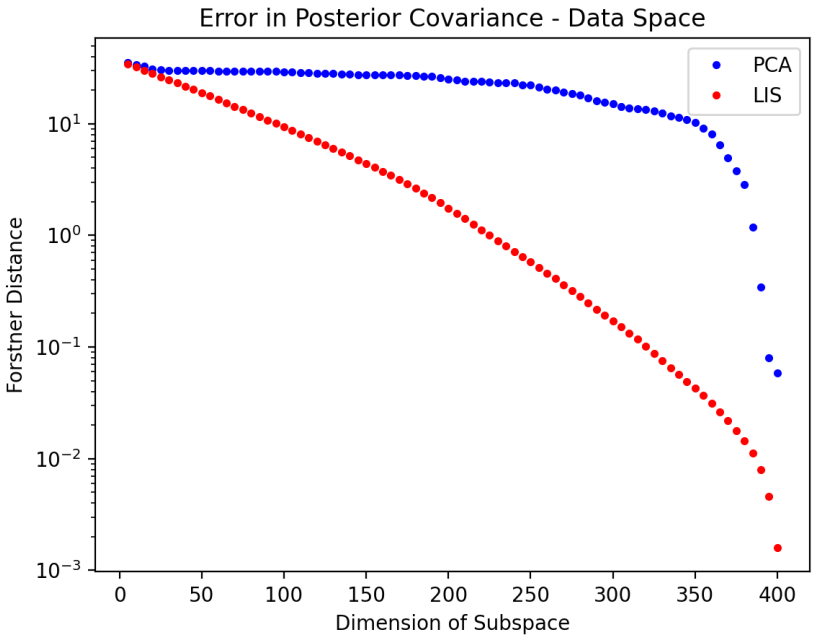


Figure 3-7: Forstner distance in posterior covariance for the data space

3.4.1 Evaluating the Posterior in the Low-Rank Data Space

A comparison of the posterior mean and covariance determined from the low-rank data subspaces was performed in the same way as for the parameter space. The Forstner distance is used to measure the error between the low-rank posterior covariance with respect to the full-dimensional covariance. The Bayes risk is used to quantify the error between the low-rank posterior mean and the true parameters used to generate the data. These plots are shown for up to dimension 400 in Figures 3-7 and 3-8.

The error in posterior covariance obtained using the likelihood informed subspace decreases much more rapidly than for PCA. In PCA, more of the information important in determining the posterior covariance is concentrated in the later dimensions.

The Bayes risk in posterior mean from data space dimension reduction follows a similar pattern to parameter space dimension reduction. For the data space, the red LIS curve flattens out at rank 250, approaching a minimum error. The PCA curve does not reach this point until rank 400.

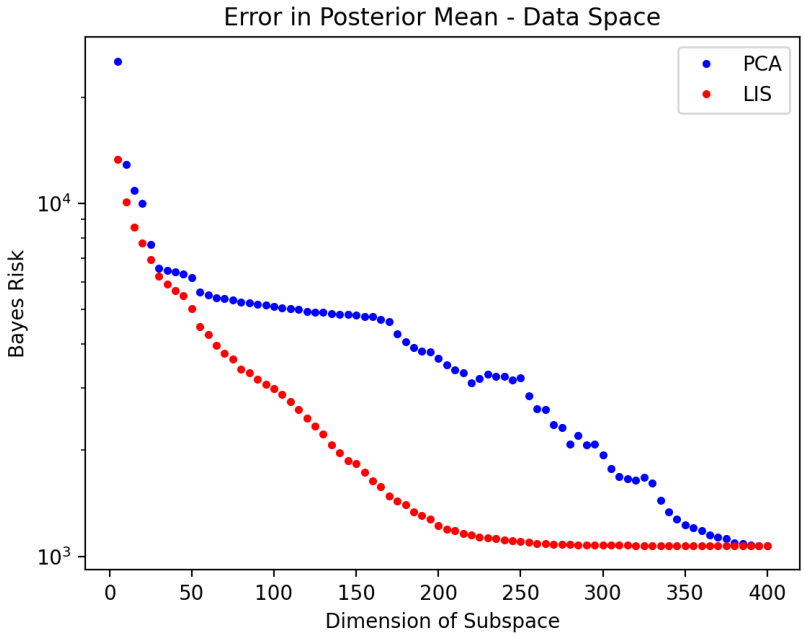


Figure 3-8: Bayes risk in posterior mean for the data space

Chapter 4

Numerical Results

This chapter presents the results of implementing MCMC using the likelihood informed subspace for the given remote sensing problem. The likelihood informed subspace was obtained using the linear model. MCMC was implemented in this subspace using the nonlinear forward model.

Most of the results in this chapter are for one specific run of MCMC. The observation was generated by running a fixed "true" state through the forward model and adding noise. This true state was defined to be the concatenation of the reflectance spectrum of the Petunia flower, obtained from the USGS Spectral Library Version 7[1], and the atmospheric parameters $x_{AOD} = 0.05$, and $x_{H2O} = 2.5$. The chain was initialized at the true state, and the Adaptive Metropolis algorithm was executed for $N = 6 \times 10^6$ samples. The first 4 million samples were discarded as burn-in to account for the non-stationarity of the first half of the chain. The dimension of the likelihood informed subspace is 100 unless specified otherwise.

We highlight several key benefits of this method over existing methods. Overall, the posterior mean is similar to those retrieved from existing remote sensing methods. However, much more information about the posterior structure is obtained. Furthermore, by implementing MCMC using the LIS, sampling efficiency is significantly increased.

[1] https://pubs.er.usgs.gov/publication/ds1035

4.1 MCMC Diagnostics

It is crucial to first check the diagnostics of the MCMC algorithm before analyzing the results. If MCMC does not mix well, for example, no conclusions can be made from the posterior samples until enough samples are included in the chain. Note that satisfactory diagnostics do not directly imply good mixing. They only indicate whether the mixing is poor.

4.1.1 Trace

The 1D trace of the MCMC samples throughout the chain is a visual indicator of stationarity. The samples should only be extracted from the stationary region after the transient part of the chain. For this multi-dimensional problem, the trace of several parameters are plotted in Figure 4-1. The chain appears to become stationary after approximately 1 million samples, so the burn-in of 4 million is more than sufficient to account for the transient region.

4.1.2 Log Posterior

The log posterior serves the same purpose as the trace plots, but summarizes the results of all parameters into a scalar value. Figure 4-2 plots the log posterior for this MCMC chain. This further confirms that the chain becomes stationary after 1 million samples.

4.1.3 Autocorrelation

Samples in the MCMC chain are not completely independent from each other. The autocorrelation quantifies the correlation of each parameter with itself at different time lags as a measure of dependence. It is also an indicator of sampling efficiency. An autocorrelation that decays slowly suggests poor mixing and many more samples are required to achieve the same result as a chain with quick autocorrelation decay.

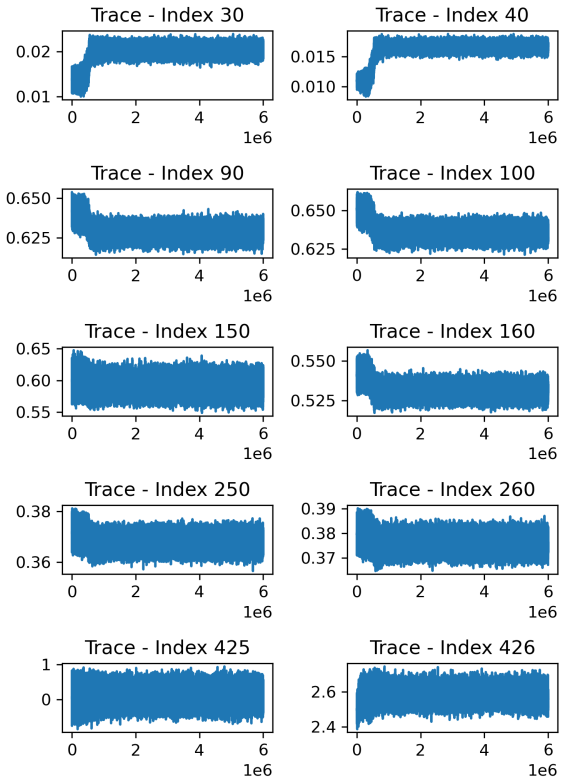


Figure 4-1: Trace plot of MCMC for LIS

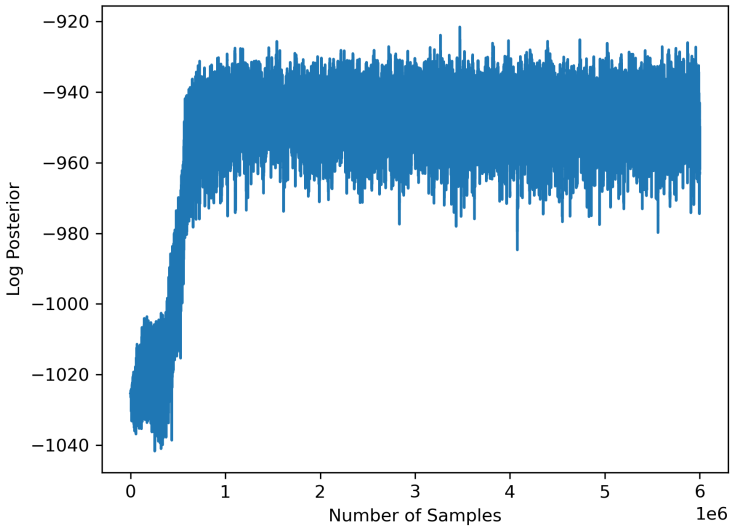


Figure 4-2: Log posterior plot of MCMC for LIS

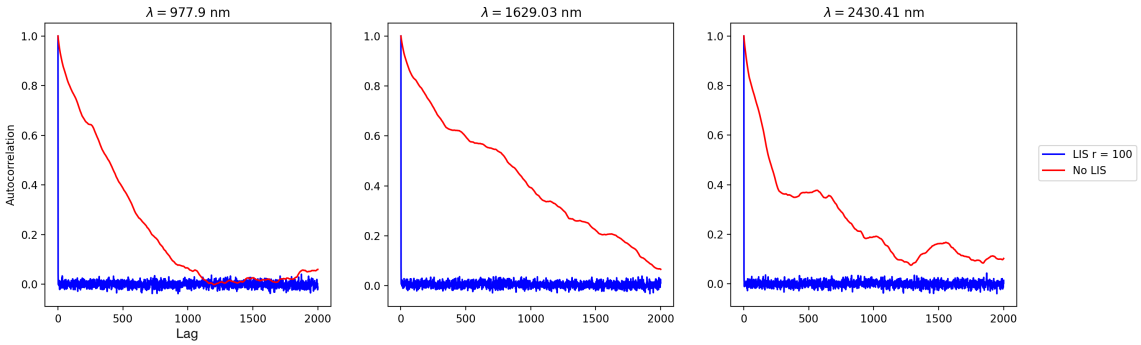


Figure 4-3: Effect of LIS on MCMC autocorrelation

Numerically, the autocorrelation of parameter X at lag k is given by

$$\rho_k = \frac{\sum_{i=1}^{N-k}(X_i - \bar{X})(X_{i+k} - \bar{X})}{\sum_{i=1}^{N}(X_i - \bar{X})^2}, \tag{4.1}$$

where $\bar{X}$ is the sample mean of X. The numerator is the covariance of the parameter with itself at k steps ahead in the chain. The denominator is the variance of the parameter.

Autocorrelation of the parameters for three different wavelengths across the spectrum are plotted in Figure 4-3 for the case with a rank-100 LIS and with MCMC in the full parameter space with no change of basis. There are two main takeaways from this series of plots. The first is that for the LIS case with dimension 100, the autocorrelation decays almost instantly, which is an indicator of satisfactory mixing in a qualitative sense. The samples are able to cover more of the parameter space throughout the chain and the results from the MCMC posterior can be trusted to a higher degree.

The second takeaway is the difference in autocorrelation decay from sampling in the low-dimensional subspace compared to sampling in the full-rank parameter space. In the full-rank space, the autocorrelation does not decay until a lag greater than 2000. This is a significant result because if a reduction in dimension from 427 to 100 leads to such a great difference in autocorrelation decay, an originally non-tractable high-dimensional problem can potentially be turned into a tractable one using this method of dimension reduction.

Table 4.1: Effective Sample Size

Wavelength	LIS	No LIS
977.9 nm	2274809	6776
1629.03 nm	269286	3521
2430.41 nm	2978107	5723

4.1.4 Effective Sample Size

Given the autocorrelation, we can compute the effective sample size (ESS), which an estimate of the true number of samples as if they were independent. The ESS is inversely proportional to the integral of autocorrelation over all lag,

$$ESS = \frac{N}{1 + 2\sum_{k=1}^{\infty} \rho_k}. \tag{4.2}$$

For the three wavelengths in Figure 4-3, the effective sample sizes for the LIS and no LIS cases are shown in Table 4.1.4. The number of independent samples is increased by at least two orders of magnitude with LIS. Without LIS, less than 0.1% of the samples are independent. Many more samples would be required, which renders MCMC computationally intractable in the canonical parameter space for this problem.

4.2 Posterior comparison

The ultimate goal of the remote sensing problem is to output a surface reflectance that can then be analyzed for biological and geological purposes. Figure 4-4 plots the surface reflectance component of the posterior mean obtained from the MAP estimate through Isofit and from MCMC. The MCMC posterior mean is very close to the MAP estimate and both are close to the truth with less than 10% error throughout the spectrum.

Figure 4-5 plots the atmospheric component of the posterior mean. The MCMC mean is close to the MAP estimate but both are far from the truth. The relative error in the AOD parameter is around 100%. In the remote sensing community, this difference is not considered too large.

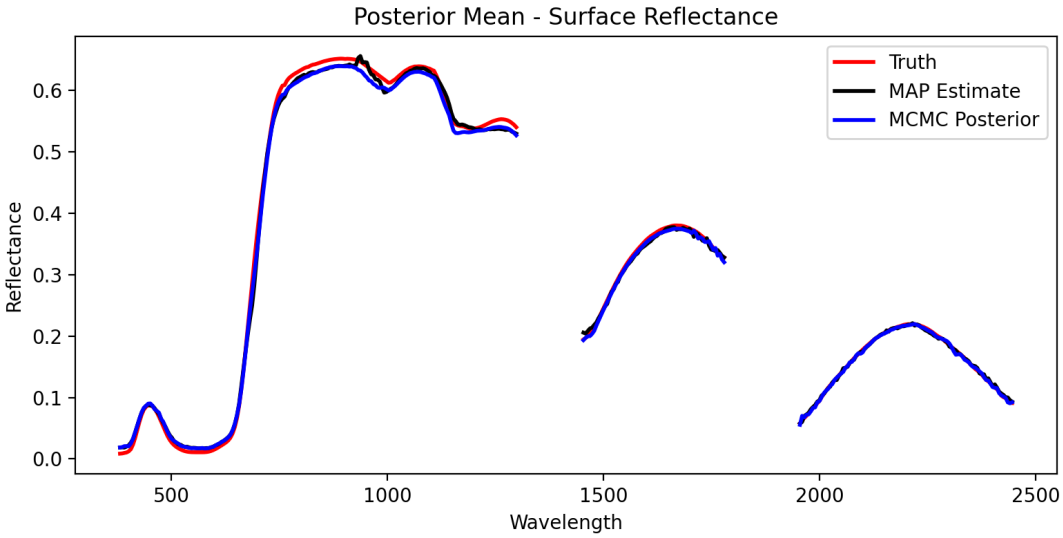


Figure 4-4: Comparison of posterior mean - surface reflectance

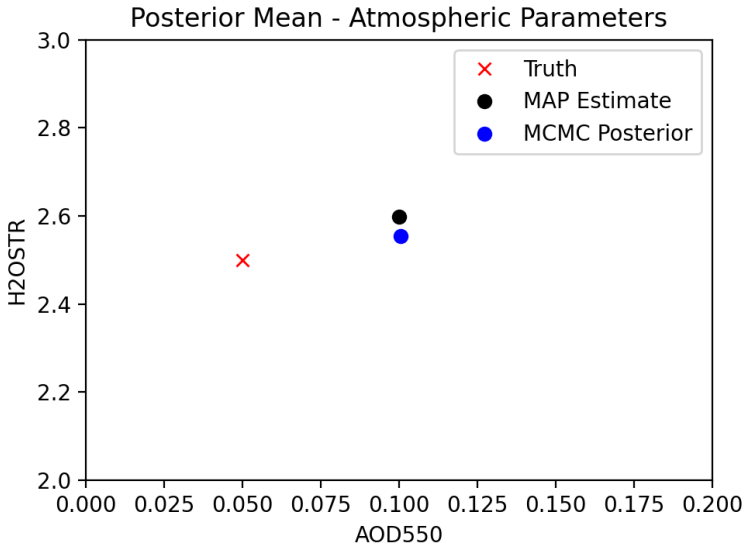


Figure 4-5: Comparison of posterior mean - atmospheric parameters

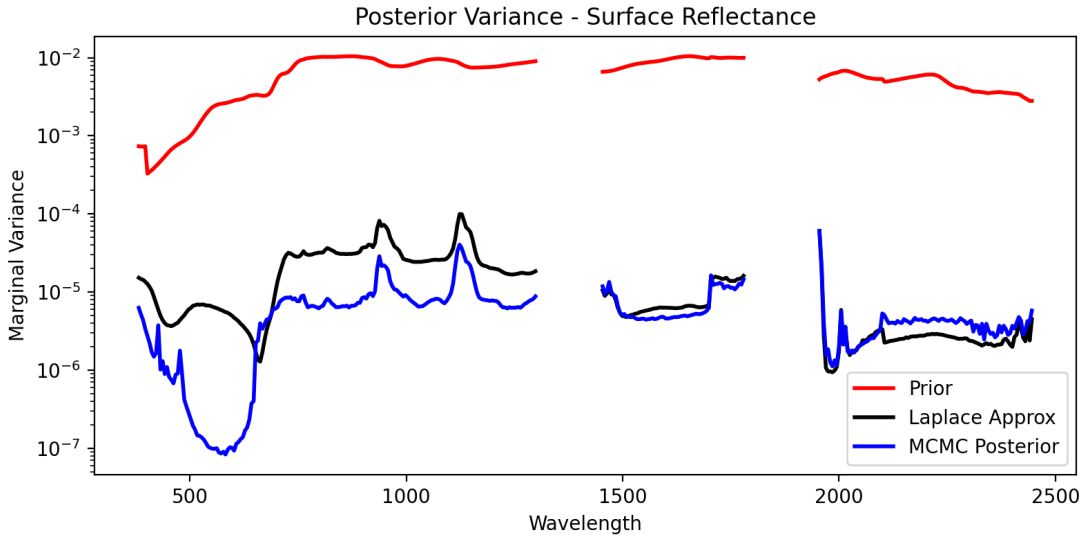


Figure 4-6: Comparison of posterior marginal variance - surface reflectance

The marginal posterior variance is plotted in Figure 4-6 for the surface parameters and in Figure 4-7 for the atmospheric parameters. The marginal prior variance is also plotted to highlight the variance reduction from prior to posterior given the data.

For the surface reflectances, the Laplace approximation and the MCMC posterior have marginal variances greatly reduced from the prior. They are approximately on the same order of magnitude. The MCMC posterior has especially low variance in the region just above 500 nm.

The atmospheric parameters demonstrate a greater difference between the Laplace approximation and the MCMC posterior, specifically for the AOD parameter. From experience with remote sensing problems, we know that the AOD parameter is difficult to estimate in general. This large uncertainty is reflected in the MCMC posterior but not in the Laplace approximation. While the Laplace approximation predicts a large drop in variance from the prior, MCMC posterior variance is approximately equal to the prior variance. This result highlights the benefit of using a Bayesian approach for this problem. By locally linearizing the Jacobian at the MAP estimate, Laplace approximation can significantly underestimate the variance. However, to understand these plots further, we need to examine the structure of the posterior.

Figure 4-8 plots the samples obtained from MCMC as well as the sample mean

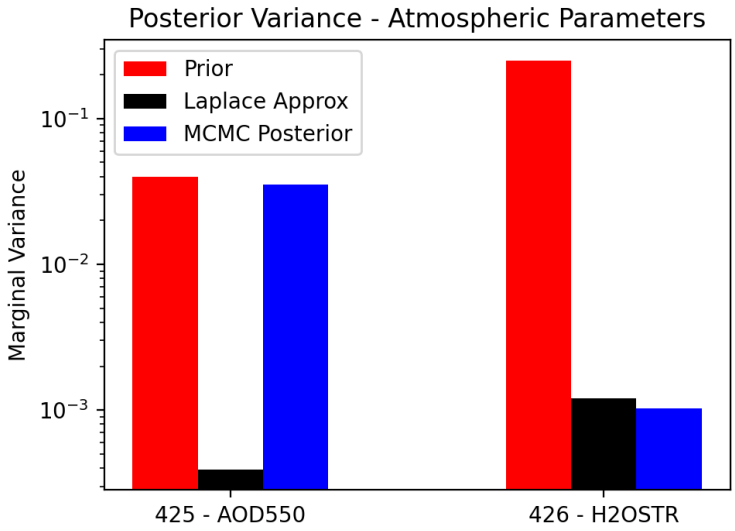


Figure 4-7: Comparison of posterior marginal variance - atmospheric parameters

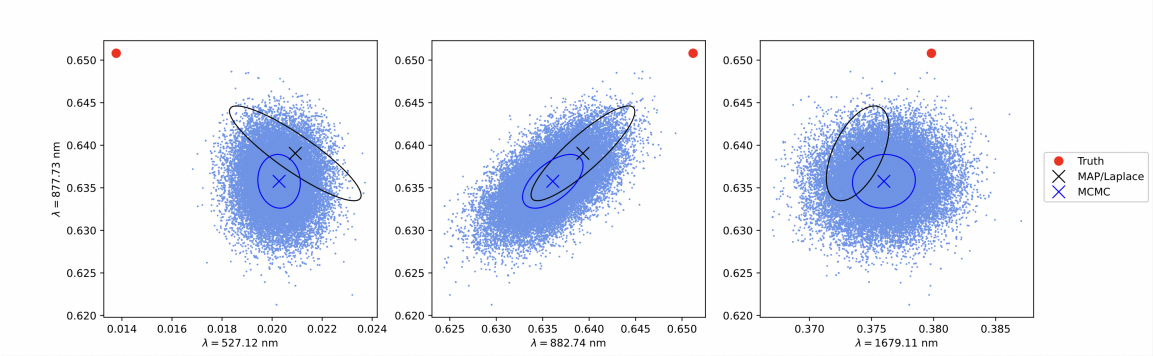


Figure 4-8: 2D marginal plots of the posterior samples

and covariance. This is compared with the MAP estimate and Laplace approximation obtained from Isofit. A point of observation is that while the Laplace approximation predicts correlations between parameters for all three cases, the parameters are only correlated when the wavelengths are close together for the MCMC posterior. Although this correlation structure is what is intuitively expected from the retrieval, no conclusions can be drawn from this qualitative observation.

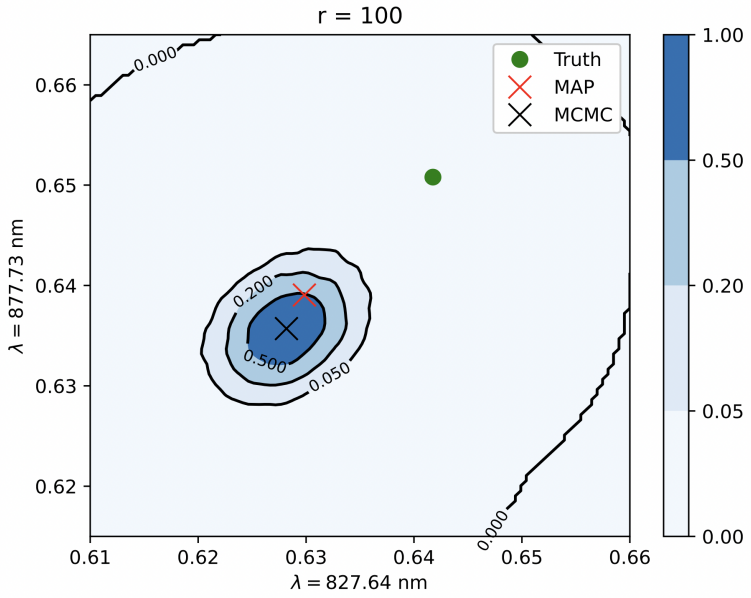


Figure 4-9: Contour plot of posterior samples at $r = 100$

4.3 Dimensions of LIS

To gain a better understanding of the importance of the dimension at which the likelihood informed subspace is truncated, the results using subspaces of two different dimensions are compared. Contour plots of the posterior density are shown for subspaces with ranks 100 and 175 in Figures 4-9 and 4-10. For both cases, the posterior mean is located near the MAP estimate. Although the posterior mean is slightly different, the covariance structure is very similar. This suggests that most of the information is contained within the lower-dimensional subspace and that an extra 75 dimensions does not contribute much to the posterior result.

To compare the sample efficiency, the trace plot for MCMC using the rank-175 subspace in Figure 4-11 can be compared to the plot for the rank-100 in Figure 4-1. An increase of 75 dimensions causes a sharp decrease in sampling efficiency. Although some parameters stabilize after 1 million samples, others such as Index 260 are still transient as the chain approaches 6 million samples. By sampling in a lower dimensional subspace, we are able to increase sampling efficiency by at least sixfold while retaining most of the important information in the posterior.

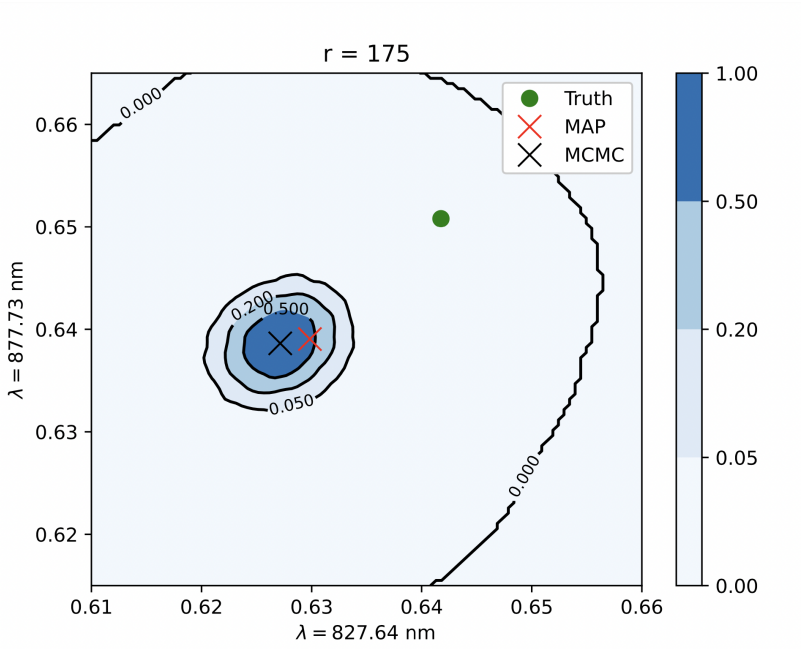


Figure 4-10: Contour plot of posterior samples at $r = 175$

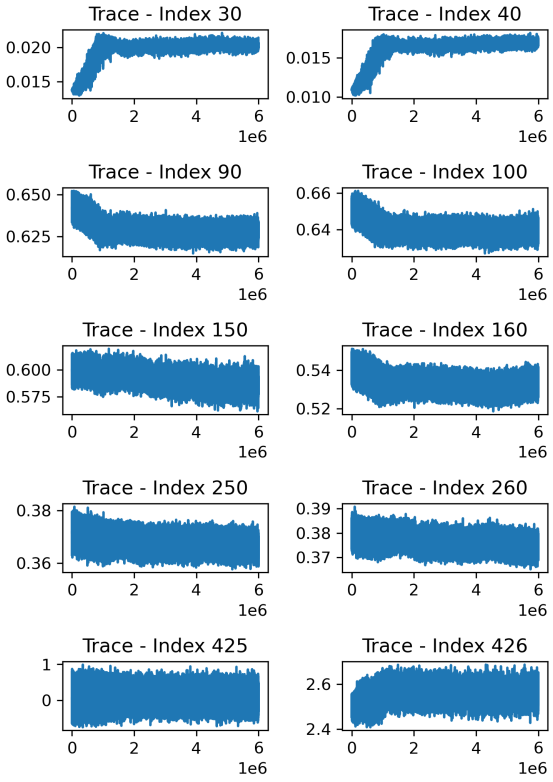


Figure 4-11: Trace plot for MCMC using LIS, $r = 175$

Chapter 5

Conclusion

In this thesis, our work on the Earth remote sensing problem is presented. Given radiance data collected from imaging spectrometers on satellites, the objective of the remote sensing problem is to infer a set of Earth surface and atmospheric parameters. The current approach for this inverse problem is to perform MAP estimation and quantify the uncertainty using a Laplace approximation. We improve upon this method by using a Bayesian approach to better quantify the uncertainty of the retrieval and to reveal more of the posterior.

The Bayesian approach consists of performing MCMC in the likelihood informed subspace. This subspace contains the directions of the parameters space that are most informed by the data. The performance of this subspace compared to the PCA benchmark was demonstrated using the Forstner distance and Bayes risk metrics for posterior mean and covariance in the linear Gaussian case. To eliminate the need for gradients of the forward model, the linearized radiative transfer model is used to determine the subspace. The methodology consists of running MCMC in this low-dimensional subspace and combining with samples from the prior in the complementary subspace to create a Markov chain in the full parameter space.

Numerical results were obtained for a true state defined to be the reflectance spectrum of a Petunia flower with atmospheric parameters $x_{AOD} = 0.05$ and $x_{H2O} = 2.5$. Six million samples were obtained using the Adaptive Metropolis algorithm. Various diagnostics were plotted for the implementation with LIS rank 100, with comparisons

to LIS rank 175 and the case with no LIS. Some of the key results of this investigation are as follows. Reducing the dimensions four-fold from 427 to 100 increases the effective sample size by more than two orders of magnitude. The covariance estimated by MCMC describes the uncertainty of the problem much better than the Laplace approximation, especially in the highly uncertain atmospheric parameters. When compared to LIS rank 175, the rank-100 LIS has a noticeable increase in sampling efficiency while maintaining a very similar posterior structure, which indicates that most of the information is captured in just 100 dimensions.

5.1 Future Work

Given the current status of the research, future work can be categorized into three stages. The first stage is to apply the current methodology to more scenarios in terms of both surface reflectance and atmospheric parameters. The analysis so far has only been done for the Petunia flower with relatively clear atmospheric conditions. Exploring different surface types such as mineral or aquatic surfaces along with expanding the range of atmospheric parameter "truths" would be of interest. Since the current methodology is expected to perform poorly in these more difficult regions, the Bayesian approach could be a more significant improvement and reveal even more about the inferred parameters for such cases.

The second stage is to implement other MCMC algorithms for the same problem to validate and potentially improve upon the results of Adaptive Metropolis. These include dimension-independent likelihood informed MCMC [4], gradient-based MCMC such as MALA [7], and multifidelity approaches to MCMC [11]. The latter approach involves a delayed acceptance scheme that exploits the computational speed of the linearized forward model to improve sampling efficiency.

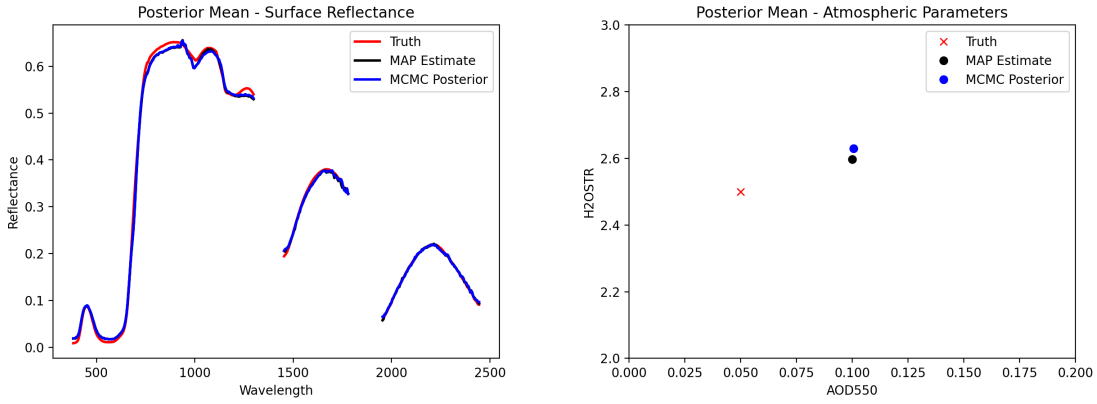
The third stage is to explore applications data compression. This could applied to goal-oriented problems such as compressing radiance data from satellites when they need to be transferred or stored in high volumes.

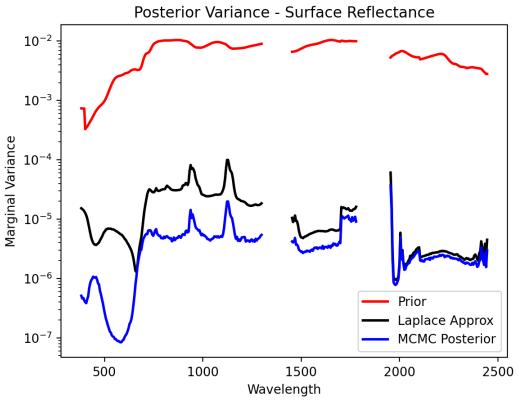
Appendix A

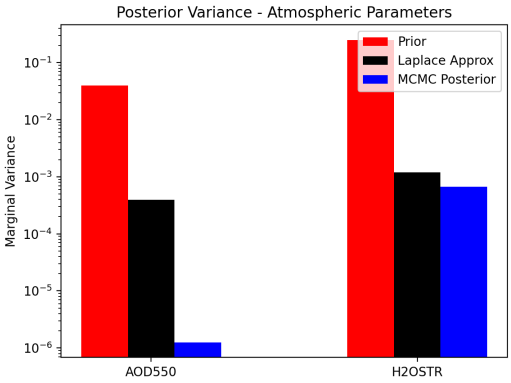
Additional MCMC results

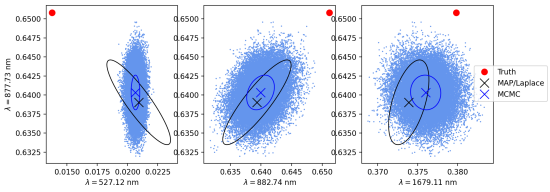
The figures in the appendix are the results of 5 runs of MCMC with varying rank and starting point of the chain, using the same setup as described in Chapter 4.

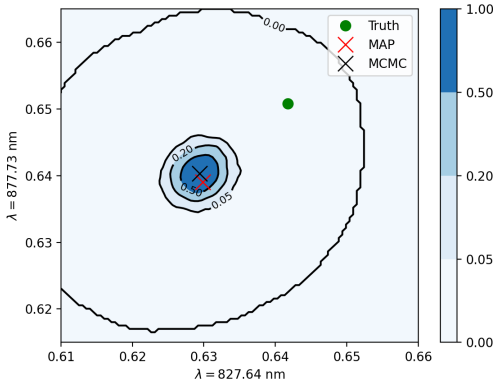
A.1 No LIS (rank 427), initialize chain at truth

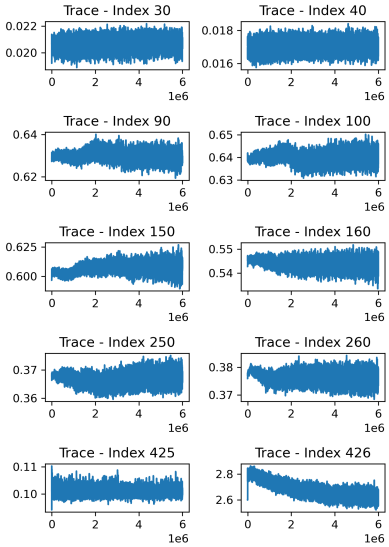



Posterior Variance - Surface Reflectance
Marginal Variance
Wavelength
Prior
Laplace Approx
MCMC Posterior

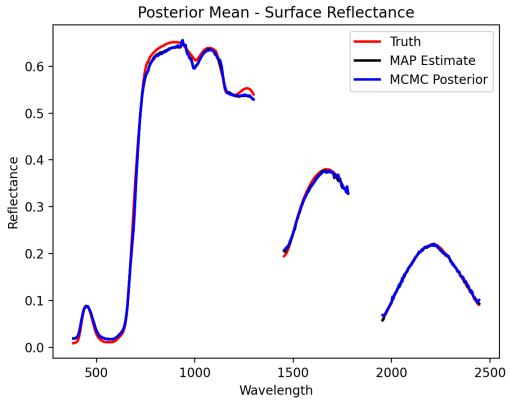

Posterior Variance - Atmospheric Parameters
Marginal Variance
Prior
Laplace Approx
MCMC Posterior
AOD550
H2OSTR

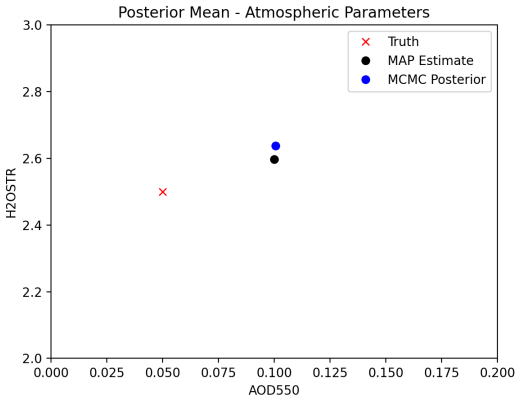

Truth
MAP/Laplace
MCMC

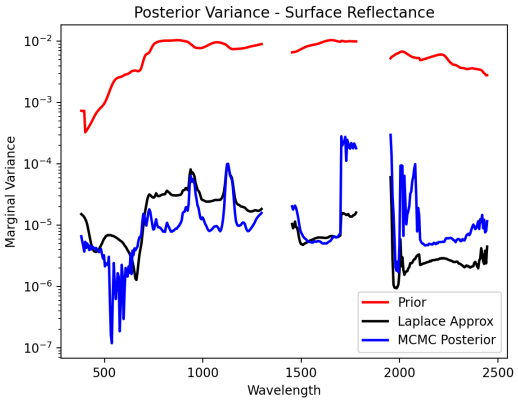

Truth
MAP
MCMC
λ = 877.73 nm
λ = 827.64 nm

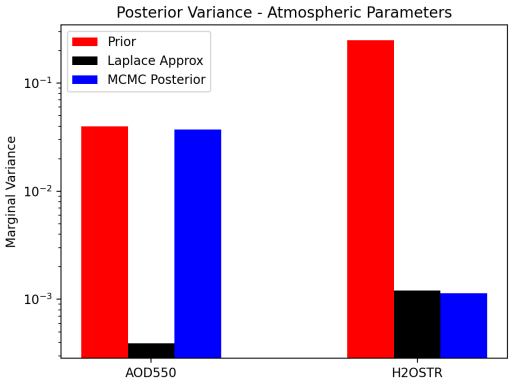

Trace - Index 30
Trace - Index 40
Trace - Index 90
Trace - Index 100
Trace - Index 150
Trace - Index 160
Trace - Index 250
Trace - Index 260
Trace - Index 425
Trace - Index 426
1e6

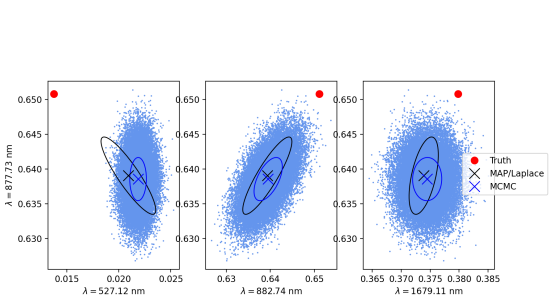
A.2 LIS rank 100, initialize chain at MAP estimate

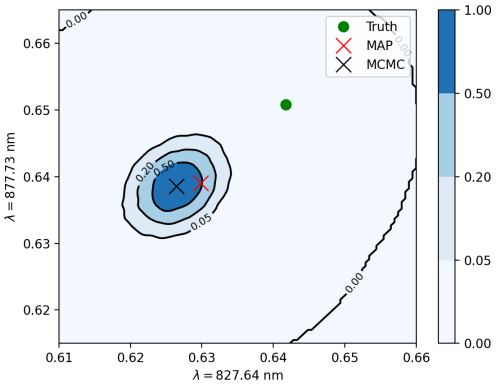


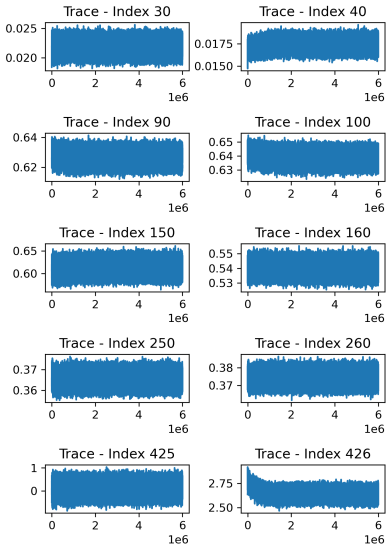




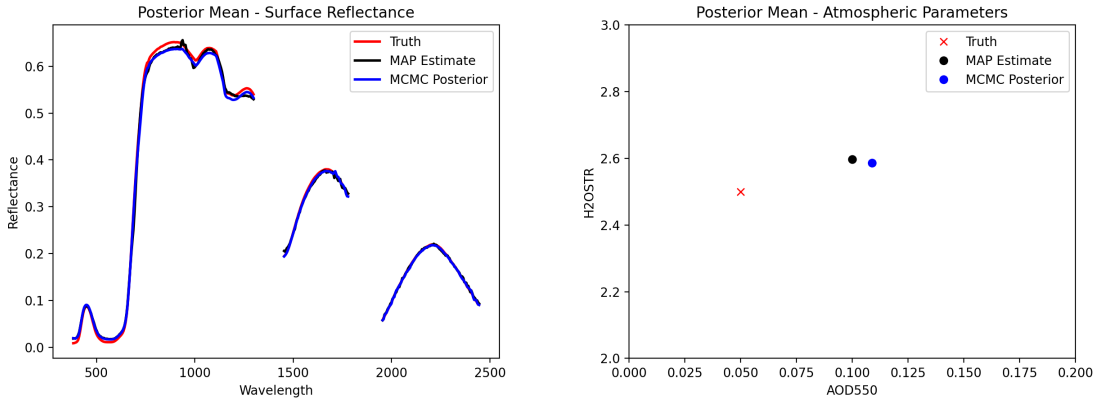


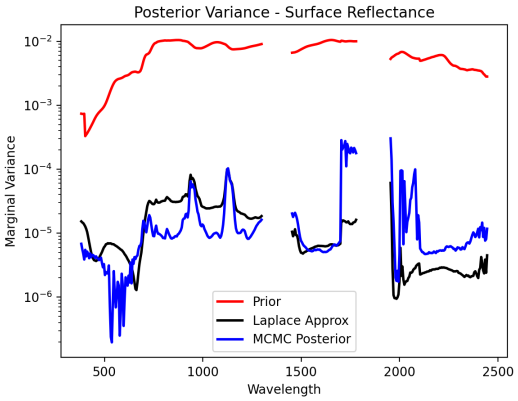


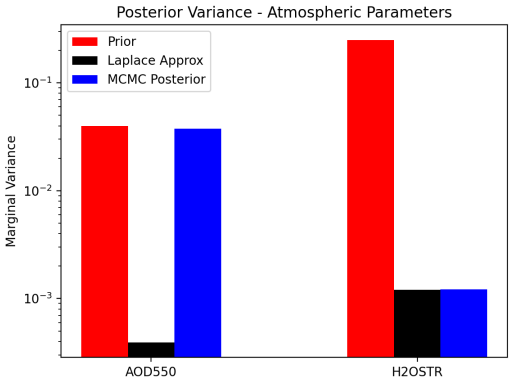


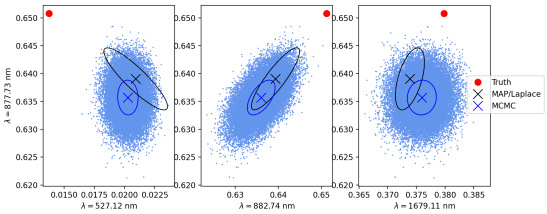


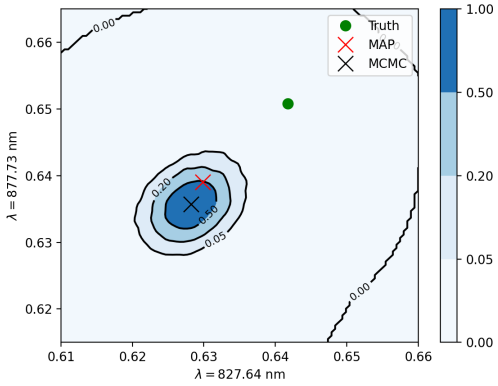
A.3 LIS rank 100, initialize chain at truth

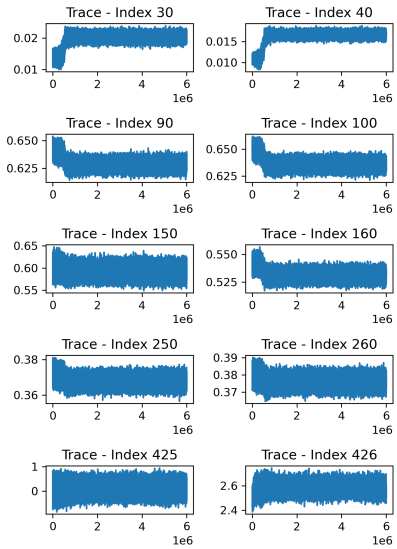



Posterior Variance - Surface Reflectance
Marginal Variance
Wavelength
Prior
Laplace Approx
MCMC Posterior

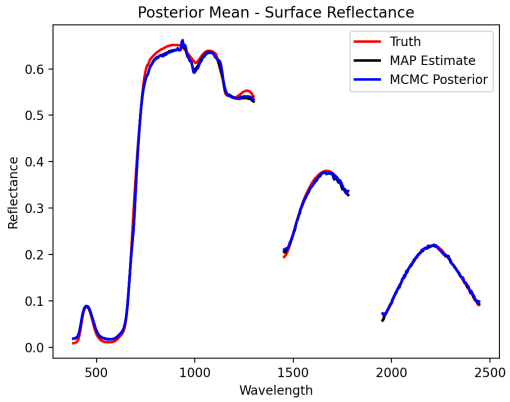

Posterior Variance - Atmospheric Parameters
Marginal Variance
Prior
Laplace Approx
MCMC Posterior
AOD550
H2OSTR

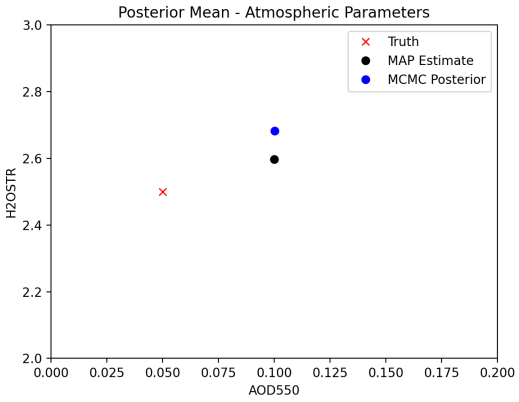

Truth
MAP/Laplace
MCMC

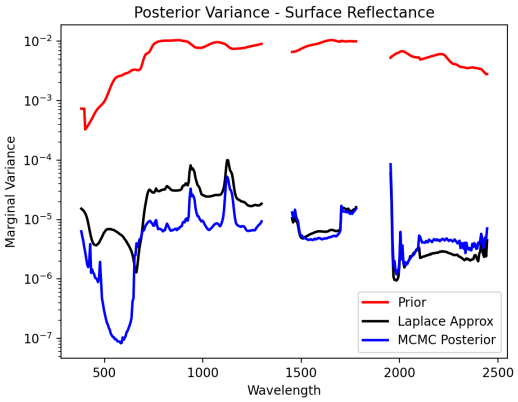

Truth
MAP
MCMC
λ = 877.73 nm
λ = 827.64 nm

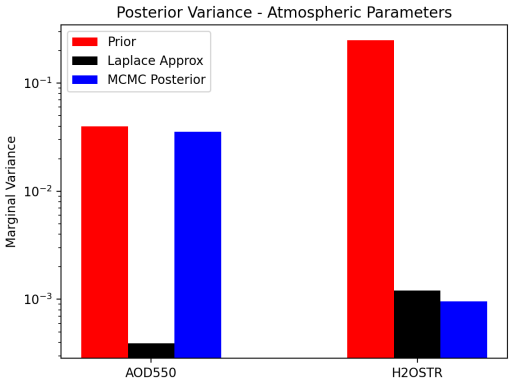

Trace - Index 30
Trace - Index 40
Trace - Index 90
Trace - Index 100
Trace - Index 150
Trace - Index 160
Trace - Index 250
Trace - Index 260
Trace - Index 425
Trace - Index 426
1e6

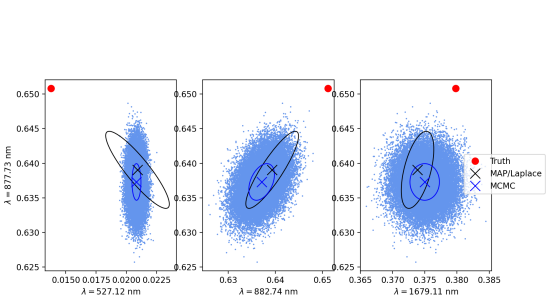
A.4 LIS rank 175, initialize chain at MAP estimate

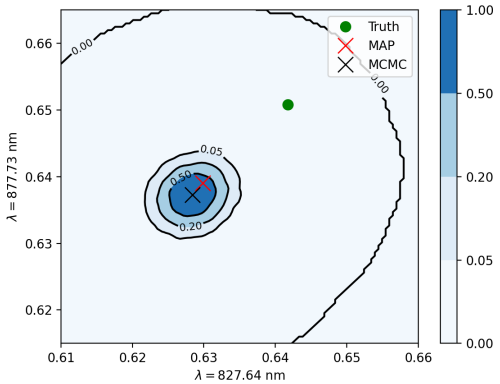


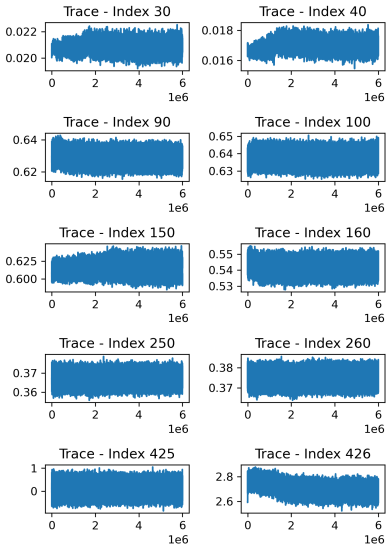




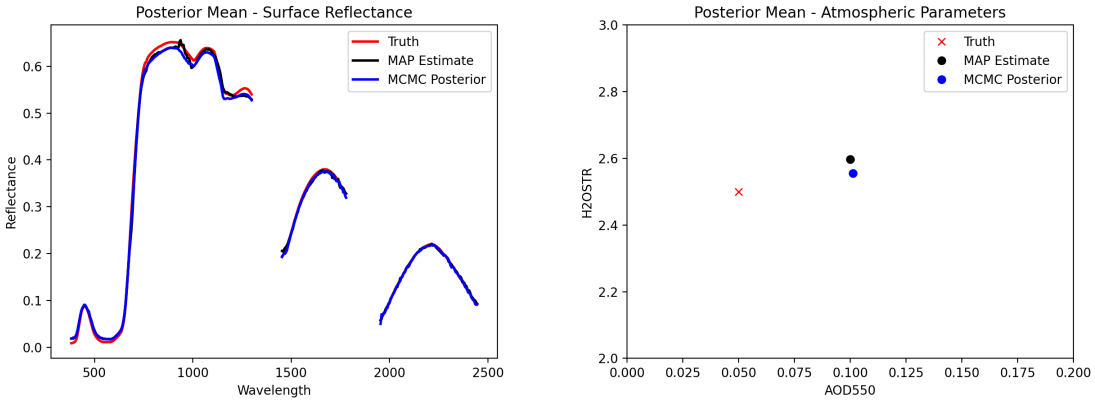


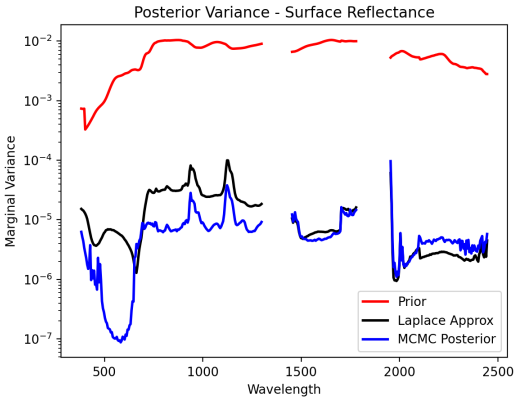


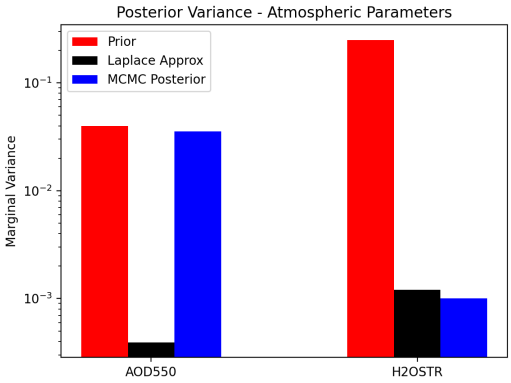


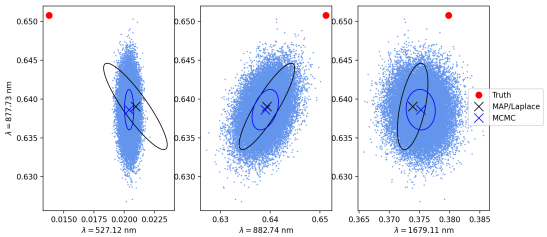


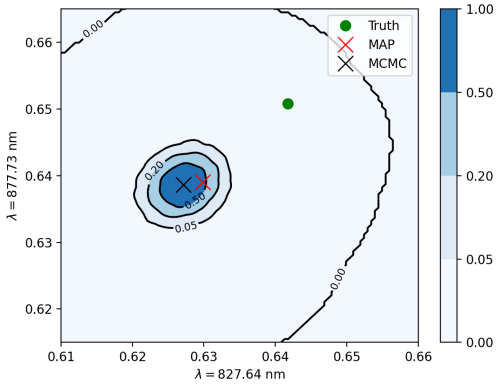
A.5 LIS rank 175, initialize chain at truth

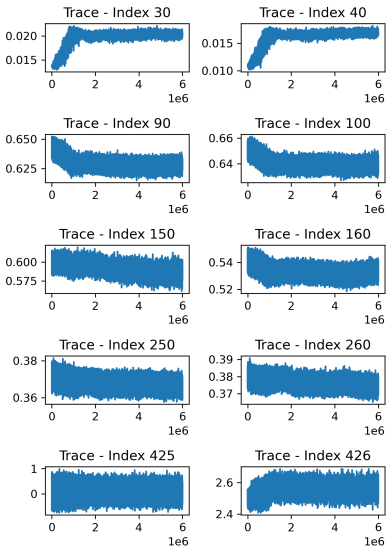



Posterior Variance - Surface Reflectance
Marginal Variance
Wavelength
Prior
Laplace Approx
MCMC Posterior


Posterior Variance - Atmospheric Parameters
Marginal Variance
Prior
Laplace Approx
MCMC Posterior
AOD550
H2OSTR


Truth
MAP/Laplace
MCMC


Truth
MAP
MCMC
λ = 877.73 nm
λ = 827.64 nm


Trace - Index 30
Trace - Index 40
Trace - Index 90
Trace - Index 100
Trace - Index 150
Trace - Index 160
Trace - Index 250
Trace - Index 260
Trace - Index 425
Trace - Index 426
1e6

Bibliography

[1] A. Berk, P. Conforti, R. Kennett, T. Perkins, F. Hawes, and J. van den Bosch. Modtran® 6: A major upgrade of the modtran® radiative transfer code. In *2014 6th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS)*, pages 1–4, 2014.

[2] Paul G. Constantine, Carson Kent, and Tan Bui-Thanh. Accelerating markov chain monte carlo with active subspaces. *SIAM Journal on Scientific Computing*, 38(5):A2779–A2805, 2016.

[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):114015, 2014.

[4] Tiangang Cui, Kody Law, and Youssef Marzouk. Dimension-independent likelihood-informed mcmc. *Journal of Computational Physics*, 304, 11 2014.

[5] W.R. Gilks, S. Richardson, and D. Spiegelhalter. *Markov Chain Monte Carlo in Practice*. Chapman & Hall/CRC Interdisciplinary Statistics. Taylor & Francis, 1995.

[6] Loïc Giraldi, Olivier P. Le Maître, Ibrahim Hoteit, and Omar M. Knio. Optimal projection of observations in a bayesian setting. *Computational Statistics Data Analysis*, 124:252–276, 2018.

[7] Mark Girolami and Ben Calderhead. Riemann manifold langevin and hamiltonian monte carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):123–214, 2011.

[8] Heikki Haario, Eero Saksman, and Johanna Tamminen. An adaptive metropolis algorithm. *Bernoulli*, 7, 2001.

[9] Jayanth Jagalur-Mohan and Youssef Marzouk. Batch greedy maximization of non-submodular functions: Guarantees and applications to experimental design, 2020.

[10] Engineering National Academies of Sciences and Medicine. *Thriving on Our Changing Planet: A Decadal Strategy for Earth Observation from Space*. The National Academies Press, Washington, DC, 2018.

[11] Benjamin Peherstorfer, Karen Willcox, and Max Gunzburger. Survey of multifidelity methods in uncertainty propagation, inference, and optimization. *SIAM Review*, 60(3):550–591, 2018.

[12] Alessio Spantini, Antti Solonen, Tiangang Cui, James Martin, Luis Tenorio, and Youssef Marzouk. Optimal low-rank approximations of bayesian linear inverse problems. *SIAM Journal on Scientific Computing*, 37(6):A2451–A2487, 2015.

[13] David R. Thompson, Vijay Natraj, Robert O. Green, Mark C. Helmlinger, Bo-Cai Gao, and Michael L. Eastwood. Optimal estimation for imaging spectrometer atmospheric correction. *Remote Sensing of Environment*, 216:355–373, 2018.

[14] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[15] Olivier Zahm, Tiangang Cui, Kody Law, Alessio Spantini, and Youssef Marzouk. Certified dimension reduction in nonlinear bayesian inverse problems. 07 2018.