ABSTRACT

SINGH, SUSHEELA PATWARI. Bayesian Methods for Nonlinear and Discrete Data with Complex Dependence. (Under the direction of Dr. Brian J. Reich and Dr. Ana-Maria Staicu.)

Bayesian methods can be powerful tools for analyzing data with a variety of complex dependence structures. Hierarchical models are popular methods for analyzing data with dependence between observations, because they are flexible, seamlessly incorporate discrete data, and allow dependence to be specified in several layers, providing simple expressions for complicated marginal structures. When dependence occurs between model parameters rather than observations, commonly in nonlinear settings, Bayesian computational algorithms have been developed that can provide reliable samples from the posterior distribution for inference. In this dissertation, we consider both cases of dependence by developing two novel Bayesian hierarchical models for variable selection with application to high-dimensional, correlated data and then reviewing Markov chain Monte Carlo algorithms for estimating highly correlated parameters within a nonlinear regression framework.

First, we consider binary responses that are correlated both across the multivariate outcomes and spatially between observations. We develop a flexible Bayesian spike-and-slab variable selection model for presence-absence indicators that accounts for spatial dependence and cross-dependence between taxa, while reducing dimensionality in both directions. By simulation, we show that the proposed method improves variable selection, particularly for the low magnitude and low prevalence covariates that are of interest in the high-dimensional microbiome setting. We mirror the analysis of Barberán et al. (2015) and apply the proposed model and PERMANOVA, a popular distance-based method, to a fungal community found within household dust. We broadly corroborate their results that climatic and geographic variables are the main influences on fungal composition within homes, and we are able to provide additional detail about how the covariates influence individual taxa.

Second, we expand on the work of the initial model and propose analyzing ranks within samples rather than binary indicators. Ranking the outcomes within a sample simplifies comparison of composition between samples and can help to mitigate the effects of contaminated counts, while retaining structural information about the relationships between outcomes. We detail a Bayesian spike-and-slab variable selection model that is applicable to rankings of taxa derived from overdispersed, contaminated, zero-inflated counts, and we specify an extended model that addresses multivariate correlation directly using random effects. We show by simulation that our proposed model outperforms a Bayesian model for the binary response and distance-based methods with a variety of responses in nearly all cases. Our method is applied to taxa found in the stool samples of healthy adults, and we find that very few of the analyzed covariates are identified as influencing microbiome composition.

Finally, we consider the efficacy of several advanced MCMC methods for obtaining posterior

samples from highly correlated parameter spaces. In particular, we consider the problem of crystallographic structure refinement from diffraction profiles, which is typically accomplished using nonlinear least squares approaches. As the Rietveld method is the standard tool for analyzing diffraction profile data, we catalog its challenges and limitations. We present a Bayesian approach as an alternative, along with several sampling algorithms, including Gibbs sampling, variants of Metropolis sampling, Hamiltonian Monte Carlo, and Approximate Bayesian Computing. These sampling algorithms are applied to the fitting of a neutron diffraction profile from a National Institute of Standards and Technology silicon standard reference material, and their results are compared. Though several of the methods provide similar fits, we found that Delayed Rejection Adaptive Metropolis provided the best posterior sample quality in terms of effective sample size.

Bayesian Methods for Nonlinear and Discrete Data with Complex Dependence

by
Susheela Patwari Singh

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Statistics

Raleigh, North Carolina

2018

APPROVED BY:

Dr. Eric B. Laber	Dr. Krishna J. Pacifici
Dr. Brian J. Reich Co-chair of Advisory Committee	Dr. Ana-Maria Staicu Co-chair of Advisory Committee

DEDICATION

To my parents, Carol and Krishna, who have given me everything.

BIOGRAPHY

Susheela Patwari Singh was born and raised in Houston, Texas, graduating from Langham Creek High School in 2004. She completed her undergraduate work at The University of Texas at Austin, earning special honors in Economics and graduating *summa cum laude* with Bachelors of Arts in Mathematics and Economics in 2008. After graduation, Susheela moved to Philadelphia, Pennsylvania to work in Economic Research at the Federal Reserve Bank of Philadelphia. After four years, she returned home to Texas to work in Institutional Research at The University of Texas System. In 2013, she began her journey in the pursuit of postgraduate training in statistics, and received her Masters of Statistics from North Carolina State University in 2015.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my husband and partner in crime, Nikhil Singh, without whom I would be truly lost. Your humor, serenity, support, and unending patience have aided me when I needed it the most. You make me the best version of myself.

I'd also like to thank my advisors, Dr. Brian Reich and Dr. Ana-Maria Staicu, and my committee members, Dr. Eric Laber and Dr. Krishna Pacifici, for their invaluable guidance, teachings, and mentorship throughout this journey. I've been fortunate to have worked closely with Dr. Alyson Wilson, whose advice and understanding have righted my ship on more than one occasion. I am also grateful for the opportunities I've received to work with insightful collaborators including Drs. Rob Dunn, Ralph Smith, Jacob Jones, Noah Fierer, and Chris Fancher.

I am grateful to the faculty and staff in the Department of Statistics at North Carolina State University, who work incredibly hard to help prepare us to face the challenges ahead, be they academic or otherwise. In particular, I want to highlight and recognize Alison McCoy, who has been one of my biggest cheerleaders and a source of emotional support.

During my graduate studies, I have been lucky to develop wonderful friendships that provide an immensely strong foundation. Many, many thanks to Ali Miller, Meredith King, Katie Forster, Sam Morris, Kyle Roell, Matt Austin, Neal Grantham, and Eric Rose, in an inexhaustive, unordered list. You all, and others, have provided friendship, strength, levity, advice, and so much more.

Finally, I want to acknowledge the love, support, and encouragement that I've received from my family, and in particular from my sisters, Anjali and Reena. You have always believed in me, even when I couldn't believe in myself. I am eternally grateful to you all and for you all.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER

1

INTRODUCTION

Assumptions of independence upon which many statistical methods rely are often tenuous. Dependence may creep into analyses in several ways, ranging from measurement errors to complex dependence between observations. Each of these sources poses its own particular challenges to analysis. In this dissertation, we will consider the challenges of modeling observations that exhibit correlation across space and multivariate outcomes and those of parameter estimation with dependence between model parameters. This chapter provides an introduction to, and a summary of, our contributions to the analysis of problems with complex dependence.

The first contribution is the development of a model for high-dimensional binary data with spatial dependence between observations and multivariate dependence across outcomes. The proposed model performs variable selection on potential covariates and demonstrates the ability to identify low magnitude and/or low prevalence covariates that are of special interest. The second contribution builds on this framework by detailing a model that utilizes rank data as the response rather than binary data. The model based on ranks is robust to overdispersion, zero-inflation, and contamination in the underlying counts, and it displays improved power for variable selection as compared to a binary response model and several nonparametric competitors. The third contribution is a survey of advanced Markov chain Monte Carlo (MCMC) sampling techniques for parameter estimation in a nonlinear regression problem. This review of MCMC methods has been used to inform and power an open-source software package for crystallographic structure refinement. The first two contributions provide new statistical methods to address open challenges in high-dimensional

microbiome settings, while the third provides an alternative approach to an important problem in materials science that addresses many of the shortcomings of the current standard methodology.

1.1 Variable selection for dependent, discrete microbiome data

A microbiome is a diverse community of microorganisms (or microbes) such as fungi, bacteria, and viruses that occupy a specific ecological niche. These communities are found all over the planet, and interest in studying them has surged in the past few decades. With the development of high-throughput (or next-generation) sequencing technologies, the cost of studying these microbes via DNA material has steadily decreased. As a result of this newly accessible data, it is increasingly evident that microbial health plays a vital role in human health, with links to obesity and Crohn's disease (Ley et al., 2005; Dicksved et al., 2008; Turnbaugh et al., 2009), Type 2 diabetes (Qin et al., 2012), allergies (Dannemiller et al., 2014), and immune system dysfunction (Round & Mazmanian, 2009), to name a few. However, much of the focus has centered on identifying and describing the effects of a microbiome on an outcome, with relatively less attention paid to the problem of identifying which external factors may influence the microbiome itself.

This is due, at least in part, to the fact that abundance counts (or *reads*) resulting from high-throughput sequencing pose a number of challenges for standard statistical analysis. First, the total number of sequencing reads is an artifact of the sequencing process and is not comparable across samples. Thus, abundance counts are compositional, and they relay relative information rather than absolute information. Second, the high-dimensional counts are frequently zero-inflated, overdispersed, and may be contaminated by errors in the sequencing process. Third, because the counts characterize a community, there may be dependence between the taxa that reflects underlying relationships. These characteristics complicate the application of multivariate statistical techniques.

This area has been dominated by methods that reduce community composition to summary metrics such as species richness or diversity. Though these summary metrics may capture salient features of the composition, the simplification to a univariate response fails to fully exploit the complexity of community data. When the multivariate response is considered, it is often through distance-based methods, which analyze a measure of dissimilarity between samples. These distance-based methods then use permutation-based hypothesis tests to determine association between covariates and the dissimilarity between samples, crucially assuming independence between samples. Popular tools such as "ANalysis Of SIMilarities" (ANOSIM; Clarke, 1993) and "PERmutational Multivariate ANalysis Of VAriance" (PERMANOVA; Anderson, 2001; McArdle & Anderson, 2001) fall into this category. In addition to the tenuous assumption of independence across samples, known to be violated in the microbiome setting (Barberán et al., 2015), these tests are limited in interpretability. Because they partition dissimilarity between samples, they cannot provide detailed

information on *how* a covariate may influence microbiome composition or which specific taxa may be affected.

In an effort to address some of these limitations, several parametric models have been developed for the abundance counts, including those based on the Dirichlet-multinomial (Chen & Li, 2013; Wadsworth et al., 2017), negative binomial (Zhang et al., 2017), or logistic normal multinomial distributions (Xia et al., 2013; Grantham et al., 2017). Nonparametric approaches to analyzing the abundance counts are not as common, but are available. Warton (2011) utilizes a generalized linear model framework to construct a permutation-based hypothesis test for covariate effects, and Zhao et al. (2015) takes a regression kernel approach. However, these models generally do not account for correlation between samples, and several rely on optimization routines that are not feasible for large problems.

Rather than analyzing counts that are compositional and may be noisy, practitioners frequently transform the data. Though merely considering relative taxa proportions instead of counts is not a full solution in itself, the literature proposes a suite of transformations of the relative taxa proportions (e.g., logratio transformation) that remove the compositional structure so that standard multivariate statistical methods may be applied appropriately (Aitchison, 1986; Fernandes et al., 2013; Mandal et al., 2015). Alternatively, the counts are commonly transformed to binary presence-absence indicators. Some of the aforementioned methods are applicable to binary data, such as the distance-based methods and Warton's method, but they are subject to the same limitations. Shirota et al. (2017) develops a model that incorporates covariate information for presence-absence indicators, but its focus is on data fusion and prediction. Additionally, it does not account for dependence between samples or perform variable selection and testing.

In Chapter 2, we consider transforming to the multivariate binary response, which removes the compositional aspect of the data and reduces noise. We develop a nonparametric Bayesian hierarchical model with the goal of identifying covariates that influence microbiome composition while accounting for spatial dependence across samples and multivariate dependence between taxa. We show by simulation that in the presence of spatial dependence, the most popular distance-based hypothesis testing method (PERMANOVA) fails to preserve its advertised size, and the proposed model improves variable selection.

By transforming the abundance counts to presence-absence indicators, we remove much of the richness available in the data. In Chapter 3, we instead use a rank transformation on the abundance counts. As before, this transformation reduces the noise in the data, but it retains more detail about the structural relationships between taxa. Using the rank data, we develop a Bayesian spike-and-slab variable selection model with the continued goal of identifying covariates that affect composition. We present both a base model and an extension that uses data-driven basis functions to model dependence between taxa directly. In a simulation study, we show that in the presence of zero-inflation, overdispersion, and contamination in the underlying counts, the proposed model outperforms a

Bayesian variable selection model that uses the binary response and several PERMANOVA models using the binary, ranks, or counts response.

As the field continues to advance, interest in developing microbial interventions to improve outcomes will grow as well. The development of these types of interventions will necessitate more study of which factors can influence microbial communities and the mechanisms by which they do so in order to identify potential intervention targets and to efficiently allocate research efforts. The models in Chapters 2 and 3 fill gaps in the existing statistical toolbox for this kind of analysis when assumptions of independence are violated.

1.2 Posterior sampling for correlated parameter spaces

As materials scientists seek to predict and understand the properties of functional materials, knowledge of a material's crystallographic structure is vital. Diffraction analysis is a powerful tool for characterizing structural information about a particular material. Diffraction profiles are obtained when the intensity of the scattered X-rays or neutrons is measured as a function of the scattering angle. Scattering results in Bragg peaks that are each associated with a specific set of planes within a crystal structure, yielding unique curves that act as a fingerprint for a material.

The location and shape of the peaks in a diffraction profile provide critical information for the structural determination of a crystal. Ideally, each observed peak would arise from a single underlying Bragg peak, but in reality the observed peaks may consist of several Bragg peaks that overlap, making structure determination more difficult. In addition to the problem of overlapping peaks, diffraction profiles are the result of a highly nonlinear system characterized by the material parameters of interest and experimental parameters. Several of these parameters are very highly correlated, making it difficult to distinguish between their effects.

Currently, diffraction profile analysis is generally carried out via Rietveld refinement (Rietveld, 1967; Rietveld, 1969), which attempts to minimize a weighted distance between a calculated profile and the observed profile. However, Rietveld analysis is subject to specific challenges. First, it can be time consuming to perform and requires extensive knowledge of the material. Second, it can be challenging to execute, as the optimization routine requires a particular "turn-on" sequence for the parameters (Young, 1993). Finally, it can return infeasible parameter estimates and relies on ad-hoc rescaling techniques to obtain standard errors for estimates.

To address these limitations, the field is increasingly exploring Bayesian algorithms for diffraction profile analysis as an alternative to direct optimization (Bergmann & Monecke, 2011; Gagin & Levin, 2015; Fancher et al., 2016; Lesniewski et al., 2016). In Chapter 4, we adopt a Bayesian nonlinear regression model for the diffraction profiles and detail approaches to posterior sampling: univariate and joint Metropolis sampling (Metropolis et al., 1953), Delayed Rejection Adaptive Metropolis sampling (DRAM; Haario et al., 2006), Hamiltonian Monte Carlo (Duane et al., 1987; Neal, 2011),

and Approximate Bayesian Computing (Voss, 2013). We apply each method to a neutron diffraction profile from a standard reference material and find that DRAM is best suited for this particular application. This work also informed the production of an open-source software package called Quantitative Uncertainty Analysis for Diffraction (QUAD), that allows materials scientists to analyze diffraction profiles using the specified Bayesian model.

CHAPTER

2

A NONPARAMETRIC SPATIAL TEST TO IDENTIFY FACTORS THAT SHAPE A MICROBIOME

2.1 Introduction

The development and increased accessibility of high-throughput sequencing technologies have steadily decreased the cost of studying DNA (Reuter et al., 2015; Heather & Chain, 2016). This has made analysis of microbial communities found in environmental samples easier. Armed with previously cost-prohibitive data, investigators have published a flurry of work leveraging microbiome information with applications in varied fields including forensics, ecology, archeology, and public health. To date, much of this work has focused on studying abiotic and biotic factors that structure microbial communities and on identifying links between microbiome characteristics (e.g., composition or diversity) with specific outcomes. For example, studies have shown that microbiome composition can identify the source of a sample (Grantham et al., 2015), linked changes in the gut microbiome to immune system dysfunction (Round & Mazmanian, 2009), tied reduced microbial diversity to obesity (Turnbaugh et al., 2009), and connected imbalances in composition to Type 2 diabetes (Qin et al., 2012). Though there has been an increased focus on defining the characteristics and markers of "healthy" microbiome communities for various systems within the body (Human

Microbiome Project Consortium, 2012; Ravel et al., 2011), the tools to understand which factors may exert influence on microbiome composition are limited.

In this paper, we consider data from Barberán et al. (2015), which contains presence-absence indicators for over 57,000 fungal taxa based on dust samples from 1,331 homes in the contiguous United States. In addition, we have geographic, climatic, and household covariate information at each sampling location covering a wide range of explanatory variables. Our objective is to develop a testing procedure to identify covariates that influence microbiome composition that is applicable to high-dimensional, spatial, binary data and leverages the multivariate dependence between microorganisms.

Previous studies have demonstrated that a home's location, design, its occupants, and their activities, can all influence the microbiome composition present in dust within the home (Barberán et al., 2015; Kettleson et al., 2015; Dannemiller et al., 2016). These studies generally reduce the data to summary measures (e.g., richness, Shannon Diversity index) or a measurement of dissimilarity in composition between samples such as Bray-Curtis dissimilarity (Bray & Curtis, 1957). Often, investigators then test for association between environmental covariates and these summaries using nonparametric permutation-based tests, the most popular of which are "ANalysis Of SIMilarities" (ANOSIM; Clarke, 1993) and "PERmutational Multivariate ANalysis Of VAriance" (PERMANOVA; Anderson, 2001; McArdle & Anderson, 2001). A tenuous assumption of these tests is exchangeability across sampling locations; we show that violation of this assumption inflates Type I error rates. This is of particular importance in our motivating example because Barberán et al. (2015) note that nearby sampling locations exhibit more similar fungal communities than those that are far apart, and thus the assumption of exchangeability is known to be violated.

Distance-based methods are also limited in interpretability. Because they partition the pairwise distances between samples, we cannot determine precisely how a covariate affects the composition or which taxa are directly affected. In a setting where an investigator may endeavor to target an intervention at a specific taxon or group of taxa, these tests are insufficient. Techniques such as redundancy analysis and canonical correspondence analysis are commonly used tools that can allow these relationships to be specified, but they too rely on permutation-based tests with an underlying assumption of independence across sampling locations. Recently, methods addressing similar concerns have been developed for use on the compositional taxa counts (Chen & Li, 2013; Zhao et al., 2015; Grantham et al., 2017; Wadsworth et al., 2017; Wang & Zhao, 2017). However, these methods are not appropriate for binary data and do not address spatial dependence in the data. Additionally, the proposed methods in Chen & Li (2013) and Wang & Zhao (2017) rely on optimization routines that may not be suitable for problems with thousands of sample locations and tens of thousands of taxa. Grantham et al. (2017) introduces a mixed effects model that accounts for correlation between taxa, but not between sampling locations. Warton (2011) proposes a permutation-based test that analyzes the community response and is applicable to presence-absence data, but it too relies on

an assumption of spatial independence and is computationally expensive, and thus it is infeasible for large problems. Clark et al. (2017) provides a framework to unify disparate data types, including presence-absence indicators, but it does not account for spatial dependence, does not incorporate dimension reduction, and does not perform variable selection or covariate testing.

As an alternative, we propose a flexible Bayesian variable selection method that uses a spike-and-slab prior and accounts for spatial dependence between nearby samples and cross-dependence between taxa. A unique feature of microbiome data is the large number of taxa, and we exploit this feature to estimate a nonstationary spatial covariance function using data-driven basis functions (Lorenz, 1956) and to relax the normality assumption common in spatial analysis (Nelsen, 1999; Gelfand et al., 2005; Reich & Fuentes, 2007; Petrone et al., 2009; Rodríguez et al., 2010). Shirota et al. (2017) proposes a nonparametric model for presence-absence data, but their aim is prediction rather than variable selection and testing for covariate effects. We provide a global test of whether or not environmental covariates affect microbiome composition that is interpretable, reliable, and has fully characterized uncertainty. In addition, our method produces clusters of taxa and tests for covariate effects on individual taxa.

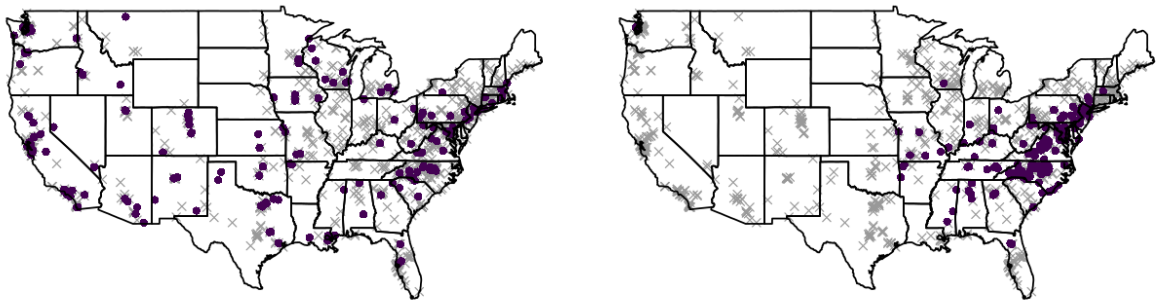
The remainder of the paper is structured as follows: in Section 2.2, we further describe the data; in Section 2.3, we detail the modeling procedure; in Section 2.4, we propose a procedure to estimate data-driven basis functions; in Section 2.5, we present a simulation study comparing our proposed method to several competitors; and in Section 2.6, we apply the proposed method to an indoor fungal community and compare our results to a previous study. Finally, we conclude with a brief summary in Section 2.7.

2.2 Motivating Data

Wild Life of Our Homes (WLOH; yourwildlife.org) is a citizen-science project focused on studying microbial diversity in and around our homes. As part of the project, participants received sampling kits and instructions specifying nine standardized locations around their homes at which samples should be taken (Dunn et al., 2013). The returned swabs were prepared using the direct PCR approach (Flores et al., 2012), which amplifies the DNA present in the samples and allows them to be sequenced and classified into Operational Taxonomic Units (OTUs). The total amount of genetic information in a sample is an artifact of the sequencing process, and as a result, the raw number of sequenced reads identified for a given OTU is not comparable across samples. Thus, rather than analyzing the read counts directly, we consider the presence-absence indicators for each taxon. This transformation to presence-absence does not entirely remove the effects of the sequencing process from the data. For example, a sample with a low total number of reads may still incorrectly consider too many taxa as absent. However, the transformation tempers the effect in most other cases.

In addition to supplying sample swabs, participants were asked to complete a questionnaire

providing details about the home's location, design features, and its occupants. Geographic and climatic information were collected based on latitude and longitude from the Climate Research Unit Time Series v3.21 Dataset (Harris et al., 2014) and the National Land Cover Database (Fry et al., 2011) for a total of over 170 covariates. From samples collected between 2012 and 2015, data was successfully sequenced for 1,331 homes spanning the 48 contiguous United States and the District of Columbia indicating the presence of 57,304 distinct fungal taxa. Of these, we focus on $m = 763$ taxa identified in Barberán et al. (2015) as being more prevalent indoors than outdoors and on a set of $p = 20$ potentially influential covariates similar to those in their analysis. The presence or absence at each sampling location for two of these taxa are mapped in Figure 2.1. In the left panel,



(a) *Trichosporon asahii* **(b)** *Perenniporia narymica*

Figure 2.1 Map of presence (purple circle) or absence (gray ×) for two primarily indoor fungal taxa at each sampling location.

Trichosporon asahii, which is commonly found living on human skin, is seen to be widespread while in the right panel, *Perenniporia narymica* is seen to occur mainly in the mid-Atlantic region. Thus, there is evidence both that there is spatial dependence underlying the presence of fungal taxa and that the strength of that dependence varies across taxa.

2.3 Nonparametric Spatial Model

Let $Y_j(\mathbf{s})$ be the binary indicator that OTU $j = 1, \ldots, m$ is present in the sample at spatial location $\mathbf{s}$. Suppose that we have a set of p covariates, $\mathbf{X}(\mathbf{s}) = [X_1(\mathbf{s}), \ldots, X_p(\mathbf{s})]$, such as those described in Section 2.2. We assume there exists a latent continuous process $Z_j(\mathbf{s})$ such that $Y_j(\mathbf{s}) = \mathbb{1}\{Z_j(\mathbf{s}) > 0\}$. The latent process is modeled as

$$Z_j(\mathbf{s}) = \beta_{j0} + \mathbf{X}(\mathbf{s})\boldsymbol{\beta}_j + e_j(\mathbf{s}), \tag{2.1}$$

where β_{j0} is an intercept and $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jp})'$ are regression coefficients that together model the probability that OTU j is present in a particular location. The final term, $e_j(\mathbf{s})$, is a multivariate spatial process with $\mathrm{E}[e_j(\mathbf{s})] = 0$ and $\mathrm{Var}[e_j(\mathbf{s})] = 1$ that models dependence not captured in the covariates between spatial locations and between OTUs. This defines a probit link for the binary responses, $\mathrm{P}\left[Y_j(\mathbf{s}) = 1 \,|\, \mathbf{X}(\mathbf{s})\right] = \Phi\left[\beta_{j0} + \mathbf{X}(\mathbf{s})\boldsymbol{\beta}_j\right]$, where Φ is the standard normal cumulative density function. The assumption that $\mathrm{Var}[e_j(\mathbf{s})] = 1$ is necessary because the covariate magnitudes are identifiable only up to the ratio of effect size to variance.

Our primary goal is to develop a test to identify factors that influence microbiome composition. A covariate influences the composition if it affects the probability that *any* of the taxa will be present in a location, and thus we test the global hypotheses

$$\mathrm{H}_{0r} : \beta_{jr} = 0 \text{ for all } j \quad \text{versus} \quad \mathrm{H}_{1r} : \beta_{jr} \neq 0 \text{ for some } j, \tag{2.2}$$

where r and j denote the covariate and OTU indices, respectively. The structure of this global test provides a means to identify an influential factor even if it affects only a small subset of the OTUs.

It remains to describe the modeling procedure for the individual components identified in (2.1). In Section 2.3.1 we specify a Bayesian variable selection model for the regression coefficients, $\boldsymbol{\beta}_j$, and in Section 2.3.2 we specify a nonparametric Bayesian model for the multivariate spatial process, $e_j(\mathbf{s})$.

2.3.1 Identifying influential covariates

We use a spike-and-slab prior for the coefficients, β_{jr}, to perform variable selection (Mitchell & Beauchamp, 1988; George & McCulloch, 1993; Kuo & Mallick, 1998). We assume that each coefficient can be written as $\beta_{jr} = \delta_{jr}\gamma_{jr}$ for an inclusion indicator, $\delta_{jr} \in \{0, 1\}$, and magnitude, $\gamma_{jr} \in \mathbb{R}$. This formulation allows us to simplify the hypotheses in (2.2) in terms of the number of OTUs for which the r^{th} covariate is included, $M_r = \sum_{j=1}^{m} \delta_{jr}$:

$$\mathrm{H}_{0r} : M_r = 0 \quad \text{versus} \quad \mathrm{H}_{1r} : M_r > 0. \tag{2.3}$$

To evaluate this, we calculate the posterior probability of the null hypothesis, $\mathrm{P}(M_r = 0 \,|\, \mathbf{Y})$, and compare to a threshold $t \in [0, 1]$. If the posterior probability of the null hypothesis is below the threshold, then the covariate is deemed influential.

Because we do not want to include the intercept in the variable selection process, we give it a separate prior $\beta_{j0} \overset{\text{iid}}{\sim} \mathrm{N}(0, \tau_0^{-1})$ with $\tau_0 \sim \mathrm{Gamma}(a_0, b_0)$. Similarly, the magnitudes have the standard conjugate formulation, $\gamma_{jr} \overset{\text{indep}}{\sim} \mathrm{N}(0, \tau_r^{-1})$ with $\tau_r \overset{\text{iid}}{\sim} \mathrm{Gamma}(a_r, b_r)$. The inclusion indicators are distributed $\delta_{jr} \overset{\text{indep}}{\sim} \mathrm{Bernoulli}(\pi_r)$, where π_r is the prior inclusion probability for the associated covariate.

The prior on π_r is chosen to induce sparsity in the coefficients such that the prior probability of the global null hypothesis in (2.3) is 0.5, reflecting no prior knowledge of whether or not a covariate is influential. In particular, the inclusion probabilities have prior density

$$\mathrm{P}(\pi_r) = \omega\left[\frac{1}{\mathrm{B}(1,\theta)}(1-\pi_r)^{\theta-1}\right] + (1-\omega), \tag{2.4}$$

a mixture of Beta$(1,\theta)$ and U$(0,1)$ distributions weighted by $\omega \in [0,1]$ and with $\theta \geq 1$. This prior has large mass on the sparse model with π_r near 0, as is common in high-dimensional Bayesian variable selection (Castillo & Vaart, 2012; Zhou et al., 2015; Ročková & George, 2016), but remains flexible enough to allow substantial probability for large values of π_r. As ω approaches 1, the prior inclusion probabilities are driven toward 0, leading to sparser coefficient vectors as in the oft used Beta$(1,\theta)$ special case, and as ω decreases to 0 the uniform component dominates and covariates will be added more readily. We can also influence the level of sparsity in the coefficients through the parameter characterizing the Beta distribution, θ. If $\theta = 1$ then the prior is simply U$(0,1)$, and the coefficient vectors will not be sparse. As θ increases, the density associated with large values of π_r decays sharply, while density associated with small values changes less drastically, leading to a steeper density curve. As a reasonable default, fix $\omega = 0.5$ and set $\theta = m^2$, where m is the number of taxa under consideration, which gives $\mathrm{P}(M_r = 0) = 0.5$ *a priori* for each covariate, as desired.

2.3.2 Capturing residual dependence

As we show in Section 2.5, properly accounting for residual dependence is necessary for valid statistical inference. To model the residual dependence in (2.1), we assume that $e_j(\mathbf{s})$ can be decomposed into a structural component, $\xi_j(\mathbf{s})$, and an independent component (or nugget), $\epsilon_j(\mathbf{s})$, such that $e_j(\mathbf{s}) = \xi_j(\mathbf{s}) + \epsilon_j(\mathbf{s})$. The structural component contributes variance $\rho \in [0,1]$, leaving the nugget distributed $\epsilon_j(\mathbf{s}) \overset{\text{iid}}{\sim} \mathrm{N}(0, 1-\rho)$ to satisfy the identifiability constraint that $\mathrm{Var}[e_j(\mathbf{s})] = 1$. We use a basis expansion model for $\xi_j(\mathbf{s})$ and write $\xi_j(\mathbf{s}) = \boldsymbol{\Psi}(\mathbf{s})\boldsymbol{\alpha}_j$, where $\boldsymbol{\Psi}(\mathbf{s}) = [\psi_1(\mathbf{s}), \ldots, \psi_L(\mathbf{s})]$ are orthogonal spatial basis functions common to all taxa and $\boldsymbol{\alpha}_j = (\alpha_{j1}, ..., \alpha_{jL})'$ are their associated loadings, for L finite or infinity. The model for the process now becomes $e_j(\mathbf{s}) = \boldsymbol{\Psi}(\mathbf{s})\boldsymbol{\alpha}_j + \epsilon_j(\mathbf{s})$.

We use a Dirichlet process prior (Ferguson, 1973) for the distribution of the loadings, which can be written as $\boldsymbol{\alpha}_j \overset{\text{iid}}{\sim} f(\boldsymbol{\alpha})$, where f is the infinite mixture

$$f(\boldsymbol{\alpha}) = \sum_{k=1}^{\infty} p_k \mathbb{1}\{\boldsymbol{\alpha} = \boldsymbol{\mu}_k\}. \tag{2.5}$$

The mixture means have priors $\boldsymbol{\mu}_k \overset{\text{iid}}{\sim} \mathrm{N}(\boldsymbol{\mu}_0, \rho\mathbf{I}_L)$, where $\boldsymbol{\mu}_0 \sim \mathrm{N}(\mathbf{0}, \tau_{\mu_0}^{-1}\mathbf{I}_L)$, $\rho \sim \mathrm{U}(0,1)$, and $\tau_{\mu_0} \sim \mathrm{Gamma}(a_{\mu_0}, b_{\mu_0})$. The mixture probabilities, p_k, are modeled using the stick-breaking representation (Sethuraman, 1994) wherein $p_1 = V_1$, $p_k = V_k \prod_{u<k}(1-V_u)$ for $k > 1$, and $V_u \overset{\text{iid}}{\sim} \mathrm{Beta}(1, D)$. This

ensures that $p_k > 0$ for all k and $\sum_{k=1}^{\infty} p_k = 1$ almost surely. Rather than fix the Dirichlet process precision parameter, we assign it an uninformative positive prior, $D \sim \text{Gamma}(a_d, b_d)$. With this infinite mixture model, our prior for the distribution of the spatial random effects, $\xi_j(\mathbf{s})$, has large support in the class of spatial processes (Gelfand et al., 2005). In practice, the infinite mixture model in (2.5) is truncated at K terms for computational purposes. That is, we assume $g_k \in \{1, ..., K\}$ for $K \leq m$ by setting $V_K = 1$, giving $f(\boldsymbol{\alpha}) = \sum_{k=1}^{K} p_k \mathbb{1}\{\boldsymbol{\alpha} = \boldsymbol{\mu}_k\}$.

The Dirichlet process prior can be viewed as a clustering model for the spatial loadings over the OTUs. If we let $g_j \in \{1, 2, \dots\}$ denote the cluster label for OTU j, then the mixture probability, p_k, can be interpreted as $\text{P}(g_j = k)$, the probability that OTU j will be assigned to cluster k. Then, given that OTU j has been assigned to cluster k, its associated spatial loading vector is the group mean for that cluster, i.e., $\boldsymbol{\alpha}_j \,|\, g_j = k$ is $\boldsymbol{\mu}_k$. In the microbiome setting, it is reasonable to believe that taxa exhibit different spatial patterns, as in Figure 2.1, and that groups of taxa will behave similarly. For example, one may expect that organisms with similar functions or that require the same nutrients might be found in close proximity to one another. This leads to a natural expectation of clustering in the spatial effects over the OTUs.

In combination with the assumptions from the previous section, the model for the latent process becomes

$$\begin{aligned} Z_j(\mathbf{s}) &= \beta_{j0} + \mathbf{X}(\mathbf{s})\boldsymbol{\beta}_j + \boldsymbol{\Psi}(\mathbf{s})\boldsymbol{\alpha}_j + \epsilon_j(\mathbf{s}) \\ &= \beta_{j0} + \sum_{r=1}^{p} X_r(\mathbf{s})\delta_{jr}\gamma_{jr} + \sum_{l=1}^{L} \psi_l(\mathbf{s})\alpha_{jl} + \epsilon_j(\mathbf{s}), \end{aligned}$$

where $\boldsymbol{\beta}_j$ captures the covariates' effect on the probability that OTU j will be present at location $\mathbf{s}$, $\boldsymbol{\Psi}(\mathbf{s})\boldsymbol{\alpha}_j$ captures residual spatial trends, and $\epsilon_j(\mathbf{s})$ are independent errors. The details of the full proposed model and its implementation, as well as a discussion of its properties, are contained in Appendix A. We also show that the covariance structure induced by our model is nonstationary in general, and that the strength of the Dirichlet process clustering controls the dependence between OTUs.

2.4 Estimating the spatial basis functions

The model detailed in Section 2.3.2 requires the construction of a set of spatial basis functions, $\boldsymbol{\Psi}(\mathbf{s})$, that are orthogonal and capable of reflecting nonstationarity. While there are several approaches available to estimate spatial basis functions from binary data (e.g., Lee et al., 2010), we follow ideas from functional principal component analysis for binary-valued functional data and estimate the basis functions as the eigenfunctions of an estimated covariance function of the spatial latent process (Hall et al., 2008; Serban et al., 2013).

Let $\mathscr{S} = \{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$ be the set of spatial locations at which the binary $Y_j(\mathbf{s})$ are observed. Our goal is to construct an estimator of the covariance of the latent process, $Z_j(\mathbf{s})$. To do so, we follow the Taylor approximation technique of Hall et al. (2008). Let $\sigma(\mathbf{s}, \mathbf{s}')$ be the covariance between $\mathbf{Z}(\mathbf{s})$ and $\mathbf{Z}(\mathbf{s}')$, which for $\mathbf{s} \neq \mathbf{s}'$ is estimated as

$$\hat{\sigma}(\mathbf{s}, \mathbf{s}') = \frac{\hat{\vartheta}(\mathbf{s}, \mathbf{s}')}{\phi\{\hat{\nu}(\mathbf{s})\}\,\phi\{\hat{\nu}(\mathbf{s}')\}}, \tag{2.6}$$

where $\phi(\cdot)$ is the standard normal density function. This is akin to equation (10) in Hall et al. (2008), where the numerator, $\vartheta(\mathbf{s}, \mathbf{s}')$, represents $\text{Cov}[\mathbf{Y}(\mathbf{s}), \mathbf{Y}(\mathbf{s}')]$, and the denominator acts as a scaling factor, with $\nu(\cdot)$ denoting the mean of the latent process.

However, the component estimators differ from Hall et al. (2008) because we cannot assume that the latent processes share a smooth mean process. In our setting, the mean process may differ across taxa or may be non-smooth due to its dependence on non-smooth covariates. We first obtain $\hat{\eta}_j(\mathbf{s})$, the predicted probability that $Y_j(\mathbf{s}) = 1$ from separate probit regressions of $\mathbf{Y}_j$ onto $\mathbf{X}$ for each taxon. Then we smooth $m^{-1}\sum_{j=1}^{m} \hat{\eta}_j(\cdot)$ over 2-D space using a bivariate kernel smoother to obtain an "average" mean process $\bar{\eta}(\cdot)$, and let $\hat{\nu}(\cdot) = \Phi^{-1}\{\bar{\eta}(\cdot)\}$, where $\Phi^{-1}(\cdot)$ is the standard normal quantile function. In order to obtain the estimated covariance of $\mathbf{Y}(\mathbf{s})$ and $\mathbf{Y}(\mathbf{s}')$, we calculate $m^{-1}\sum_{j=1}^{m}[Y_j(\mathbf{s})Y_j(\mathbf{s}') - \hat{\eta}_j(\mathbf{s})\hat{\eta}_j(\mathbf{s}')]$ and smooth these estimates using a four-dimensional kernel smoother. The resulting smoothed estimates are collected as $\hat{\vartheta}(\mathbf{s}, \mathbf{s}')$. As is typical in nonparametric statistics, the optimal bandwidths are chosen using generalized cross-validation (Craven & Wahba, 1978; Friedman et al., 2009).

Applying this procedure to the variances will result in biased estimates (Hall et al., 2008). To remove this bias, we consider a modified estimator, $\hat{\sigma}(\mathbf{s}, \mathbf{s})$, and use the intercept of the weighted linear model

$$\hat{\sigma}(\mathbf{s}, \mathbf{s}') = \beta_0 + w(\mathbf{s}, \mathbf{s}')\,\text{d}(\mathbf{s}, \mathbf{s}')\beta + \epsilon,$$

for $\mathbf{s} \neq \mathbf{s}'$ and with weights $w(\mathbf{s}, \mathbf{s}') = \exp\left[-\frac{d(\mathbf{s}, \mathbf{s}')}{d_{10}}\right]\mathbb{1}(d(\mathbf{s}, \mathbf{s}') \leq d_{10})$, where d_{10} is the distance between $\mathbf{s}$ and its 10^{th} closest neighbor for some distance measure d. In our application, we use the great-circle distance in miles.

Let $\hat{\boldsymbol{\Sigma}}$ be the initial estimate of the spatial covariance matrix with elements $\hat{\sigma}(\mathbf{s}, \mathbf{s}')$. By construction, $\hat{\boldsymbol{\Sigma}}$ is symmetric. However, to ensure that it is positive semidefinite, we consider its low rank approximation. Let $\tilde{\phi}_1(\mathbf{s}), \ldots, \tilde{\phi}_L(\mathbf{s})$ be the leading L eigenvectors of $\hat{\boldsymbol{\Sigma}}$, scaled by the square-root of their associated eigenvalues, such that they account for a specified percentage of explained variance. In our application, we use 90%. To preserve the variance structure described in Section 2.3.2 (i.e., $\text{Var}[\xi_j(\mathbf{s})] = \rho$), we need to ensure that $\sum_{l=1}^{L} \tilde{\phi}_l^2(\mathbf{s}) = 1$. If $L < n$, this will require scaling the

eigenvectors to obtain

$$\psi_l(\mathbf{s}) = \left[\frac{1}{\sum_{l=1}^{L}\tilde{\phi}_l^2(\mathbf{s})}\right]^{\frac{1}{2}}\tilde{\phi}_l(\mathbf{s}).$$

Let $\boldsymbol{\Psi} = [\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_L]$, where $\boldsymbol{\psi}_l = \{\psi_l(\mathbf{s}_1), \ldots, \psi_l(\mathbf{s}_n)\}'$ for $l = 1, \ldots, L$. After this scaling process, $\boldsymbol{\Psi}$ is no longer orthogonal on $\mathbb{R}^L$, and thus we rotate by its right singular vectors to obtain the proposed basis functions.

Now, $\boldsymbol{\Psi}$ is scaled appropriately to preserve the variance structure we require, rotated to preserve orthogonality between basis functions, and reflects the nonstationarity we expect in the data. The estimated basis functions are available only at the locations in $\mathscr{S}$, and extrapolation would be required to make spatial predictions beyond the n sample locations. However, our objective is not spatial prediction, but rather to account for the complex dependence structure at the sampling locations to give a valid global test of covariate effects.

Because of the reliance on generalized cross-validation to select the bandwidth parameter, the four-dimensional smoothing step to obtain the $\hat{\vartheta}(\mathbf{s}, \mathbf{s}')$ estimates can be prohibitively expensive. Two approaches to alleviating this burden are either to use a different method to select the bandwidth or to make the cross-validation less computationally intensive. As an example, a reasonable approach that avoids cross-validation might be to construct a variogram, identify the distance at which the correlation decays, and use that distance to set a bandwidth. Alternatively, if the data contains sampling locations that are close to one another, one could downsample the locations while approximately preserving the spatial coverage of the data. Then, generalized cross-validation can be done quickly on this smaller, representative set of locations to obtain an estimated optimal bandwidth. This latter approach is utilized in our data application in Section 2.6.

2.5 Simulation study

In this study, we consider generating data while varying the type of spatial dependence in the latent process, the existence of cross-dependence between OTUs in the latent process, the magnitude of covariate effect size, and the degree of prevalence in covariate effects, and evaluate how these factors influence the true and false positive rates of the global test in (2.3).

2.5.1 Methods

We generate data on a 15×15 grid on the unit square for a total of $n = 225$ spatial locations. For each of $m = 50$ OTUs, we draw the latent process as $\mathbf{Z}_j \sim \mathcal{N}_n\left(\mathbf{X}\boldsymbol{\beta}_j, 0.95\boldsymbol{\Sigma}_z + 0.05\mathbf{I}_n\right)$. The structure of $\boldsymbol{\Sigma}_z$ varies based on the type of spatial dependence:

(Ind) Independence: $\boldsymbol{\Sigma}_z = \mathbf{I}_n$,

(Exp) Stationary dependence: $\mathbf{\Sigma}_z$ is populated by the exponential covariance function with spatial range set such that the correlation between the two closest sites is 0.75, and

(Nonstat) Nonstationary dependence: where $\mathbf{\Sigma}_z(\mathbf{s},\mathbf{s}') = \cos(2\pi s_1)\cos(2\pi s_1') + \sin(2\pi s_2)\sin(2\pi s_2')$ for $\mathbf{s} = (s_1, s_2)$.

When the setting calls for multivariate dependence in the latent process, we assume a separable covariance function and define $\text{Cov}[Z_j(\mathbf{s}), Z_{j'}(\mathbf{s}')] = c(j,j')\mathbf{\Sigma}_z(\mathbf{s},\mathbf{s}')$, where $c(j,j') = 0.8^{|j-j'|}$ is the cross-dependence function. In reality, we do not expect a meaningful ordering of the OTUs, but this covariance is used to generate data with a reasonable range of cross-correlations. The $p = 20$ covariates are drawn from a mean-zero Gaussian process with separable covariance function $\text{Cov}[X_r(\mathbf{s}), X_{r'}(\mathbf{s}')] = c(r,r')\mathbf{\Sigma}_x(\mathbf{s},\mathbf{s}')$ where $c(r,r')$ is as above, and $\mathbf{\Sigma}_x$ is the exponential covariance with spatial range set such that the correlation between the two closest sites is 0.5.

Of the covariates, $p_0 = 6$ are influential (i.e., β_{jr} is non-zero for some j) and the remainder are unimportant for all OTUs (i.e., $\beta_{jr} = 0$ for all j). In order to examine the ability of the algorithm to detect covariate effects across prevalences and magnitudes, the influential covariates are split into 3 pairs. The first pair affects all OTUs, the second pair affects a randomly selected 50% of OTUs, and the final pair affects a randomly selected 10% of OTUs. Within each pair of non-zero coefficients, the first covariate is assigned a large magnitude of $\beta_{jr} = 0.5$, and the second is assigned a small magnitude of $\beta_{jr} = -0.25$. The randomization over taxa for prevalence is done independently so that any one OTU may have 2, 4, or 6 important covariates.

Under each of the simulation settings we generate $N = 50$ replicate datasets and fit the proposed spatial nonparametric model and several competing models:

(PERM) PERMANOVA (Anderson, 2001; McArdle & Anderson, 2001), a permutation-based hypothesis test as implemented in the `R` package `vegan 2.4-3` using Bray-Curtis dissimilarity.

(NS) Nonspatial variable selection model, i.e., $\rho = 0$.

(Mat) Parametric spatial model where $\boldsymbol{e}_j = [e_j(\mathbf{s}_1), \ldots, e_j(\mathbf{s}_n)]'$ from (2.1) is modeled using a Matérn covariance function. The smoothness has prior $\kappa \sim \text{U}(0,2)$ (Stein, 1999; Banerjee, 2005), and the range has prior $\log(\zeta) \sim \text{N}(0, \sigma_\zeta^2)$ where σ_ζ^2 is set such that the 99$^{\text{th}}$ percentile of the prior distribution for the range is the maximum observed distance.

(SNP) Proposed nonparametric spatial model using the nonstationary basis detailed in Section 2.4, with the maximum number of groups set to $K = m$.

For each of the Bayesian models (NS, Mat, and SNP), we fit the model using a special case of (2.4) where $\omega = 1$ and $\theta = m$, which simplifies the prior to $\pi_r \overset{\text{iid}}{\sim} \text{Beta}(1, m)$. This commonly used prior on the inclusion probabilities will make it more likely for π_r to be close to 0 than in the mixture

setting. Our focus is on identifying covariates that are borderline cases, i.e., factors that influence only a few taxa. The sharper cut of this simplified prior near the origin makes the sampler less likely to include these covariate spuriously. To determine sensitivity to this prior specification, we also ran the simulation using the mixture prior in (2.4) with the recommended default values. The results are qualitatively the same, with improved performance for Mat in identifying small magnitude covariates but a reduced ability to identify low prevalence covariates. The model performance for SNP is broadly unchanged. The remainder of the prior specifications are detailed in Appendix A.2. The models are run for a total of 40,000 iterations with a burn-in period of 10,000, and the posterior samples are thinned by 2. We deem the r^{th} covariate to be influential if the associated posterior probability of the null is below 0.05, i.e., $\mathrm{P}(M_r = 0 \,|\, \mathbf{Y}) < 0.05$, for the Bayesian models, or if its P-value from PERMANOVA is below 0.05.

For each dataset, we evaluate the models using true positive rate (TPR) and false positive rate (FPR), presented in Table 2.1. Let M_r^* be the indicator that the r^{th} covariate is truly influential. The true positive rate is the percent of truly influential covariates correctly classified as influential by the model for a given threshold t,

$$\mathrm{TPR}(t) = \frac{\sum_{r=1}^{p} M_r^* \mathbb{1}\{\mathrm{P}(M_r = 0 \,|\, \mathbf{Y}) < t\}}{p_0}.$$

The false positive rate is the percent of truly unimportant covariates that are incorrectly classified by the model as influential,

$$\mathrm{FPR}(t) = \frac{\sum_{r=1}^{p} (1 - M_r^*) \mathbb{1}\{\mathrm{P}(M_r = 0 \,|\, \mathbf{Y}) < t\}}{p - p_0}.$$

We also consider a "registered" true positive rate, where the threshold for each method is set to control its false positive rate at or below 0.05. In other words, for each model and simulated data set, we find the largest threshold T such that $\mathrm{FPR}(T) \leq 0.05$, and use this calibrated threshold to evaluate the model. In the case of PERMANOVA, the posterior probability of the null is replaced by the P-value. This allows us to compare the power of the methods on an even footing in Table 2.2. Finally, in Table 2.3, we consider the inclusion rate for the influential covariates for each model, broken out by magnitude of the covariate effect, small (S) or large (L), and the prevalence of the covariate effect, 100%, 50%, or 10%. The inclusion rate (IR) is defined as the proportion of the N simulation runs for which the method correctly classified the covariate as influential,

$$\mathrm{IR}(t) = \frac{1}{N} \sum_{s=1}^{N} \mathbb{1}\{\mathrm{P}\left(M_{s,r^*} = 0 \,|\, \mathbf{Y}\right) < t\},$$

for each of the $r^* = 1, \ldots, p_0$ influential covariates. As in the global results presented in Table 2.1, we use a fixed threshold of $t = 0.05$.

2.5.2 Results

Table 2.1 Summary of true positive rate (TPR), false positive rate (FPR), and average model fitting time in minutes for PERMANOVA (PERM), the nonspatial (NS), parametric Matérn (Mat), and proposed nonparametric (SNP) models.

		Dependence Between Taxa					
		Independence			Autoregressive		
Spatial Dependence	Model	TPR	FPR	Time	TPR	FPR	Time
Independence	PERM	0.62	0.05	3.75	0.49	0.06	3.68
	NS	0.38	0.00	21.59	0.38	0.01	21.58
	Mat	0.40	0.00	426.09	0.39	0.01	418.41
	SNP	0.37	0.00	34.90	0.38	0.00	35.05
Exponential	PERM	0.96	0.80	3.98	0.87	0.61	3.75
	NS	0.86	0.48	22.18	0.81	0.43	21.73
	Mat	0.54	0.04	232.48	0.51	0.04	228.38
	SNP	0.71	0.10	36.82	0.67	0.10	35.61
Nonstationary	PERM	0.81	0.49	3.74	0.81	0.49	3.88
	NS	0.91	0.43	24.01	0.90	0.47	25.27
	Mat	0.85	0.00	231.90	0.84	0.01	237.80
	SNP	0.93	0.02	39.36	0.94	0.05	40.51

As is evident in Table 2.1, in the case of no spatial dependence in the data, PERM outperforms the Bayesian models. The Bayesian tests are more conservative, but after tuning the FPR to be 0.05 (Table 2.2), they have comparable or increased true positive rates as compared to PERMANOVA. The false positive rate for PERM is well-controlled even in the face of multivariate dependence, which is reasonable given that the permutation is done at the sampling location level and thus the structure of any cross-dependence between taxa is preserved.

However, in the presence of spatial dependence, PERMANOVA fails to preserve the size of the hypothesis test and has false positive rates an order of magnitude higher than expected. This is perhaps not unexpected as the pseudo-F test is built on the assumption of exchangeability across sampling locations. Blind application of these permutation-based methods in settings where spatial independence across sampling locations is not a reasonable assumption will result in misleading conclusions.

When the data are spatially dependent, NS and PERM have high true positive rates accompanied by high false positive rates, indicating that the models favor including all covariates rather than

Table 2.2 Summary of "registered" true positive rate (TPR) and false positive rate (FPR) for PERMANOVA (PERM), the nonspatial (NS), parametric Matérn (Mat), and proposed nonparametric (SNP) models. If values are not provided, there is no threshold value or significance level that controls the false positive rate at the required level.

		Dependence Between Taxa			
		Independence		Autoregressive	
Spatial Dependence	Model	TPR	FPR	TPR	FPR
Independent	PERM	0.63	0.05	0.48	0.05
	NS	0.70	0.05	0.59	0.05
	Mat	0.70	0.05	0.59	0.05
	SNP	0.66	0.05	0.57	0.05
Exponential	Mat	0.58	0.05	0.53	0.05
	SNP	0.63	0.05	0.56	0.05
Nonstationary	Mat	0.96	0.05	0.93	0.05
	SNP	0.95	0.05	0.93	0.05

discriminating between important and unimportant factors. Therefore, we exclude these models in Table 2.3, where we present the inclusion rate for the influential covariates broken out by prevalence and magnitude for each of the models. As before, in the case of spatial independence, PERM outperforms the Bayesian models, which all perform similarly. However in the case of spatial dependence, breaking out the model performance in this way allows us to see the contrast between the Bayesian spatial models. In particular, we can see that SNP outperforms the parametric model in identifying covariates with low prevalence and/or small magnitudes, which is our primary focus. Under the exponential covariance structure, SNP picks up the low prevalence, small magnitude covariate 16-20% of the time, whereas the parametric model selects it in only 0-4% of the replications. Similarly, under the nonstationary covariance structure, the parametric model selects the covariate in only 20-28% of replications, as opposed to the 60-66% of replications for SNP.

In addition, the spatial parametric model takes 6-10× longer to fit than the other models on average, and this is a relatively small problem with only 225 locations and 50 taxa. Mat requires several inversions of an $(n \times n)$ matrix during each MCMC iteration and it is clear that this becomes computationally infeasible for problems much larger than this simulated setting. The proposed nonparametric model reduces the dimensionality of the problem for both large numbers of observations and a large number of observed taxa without sacrificing its aptitude to discern influential covariates from unimportant ones.

Table 2.3 Inclusion rate for influential covariates for PERMANOVA (PERM), the nonspatial (NS), parametric Matérn (Mat), and proposed nonparametric (SNP) models, broken out by covariate magnitude (S=Small, L=Large) and prevalence (100%, 50%, 10%).

Dependence			Covariate Prevalence and Magnitude					
Spatial	Between Taxa	Model	100% L	100% S	50% L	50% S	10% L	10% S
Ind	Ind	PERM	1.00	0.38	1.00	0.62	0.60	0.14
		NS	1.00	0.16	0.84	0.02	0.28	0.00
		Mat	1.00	0.18	0.86	0.02	0.32	0.00
		SNP	1.00	0.06	0.78	0.02	0.34	0.00
	AR(0.8)	PERM	0.92	0.26	0.98	0.26	0.42	0.12
		NS	1.00	0.14	0.76	0.08	0.32	0.00
		Mat	1.00	0.14	0.76	0.08	0.36	0.00
		SNP	1.00	0.10	0.78	0.06	0.34	0.00
Exp	Ind	Mat	1.00	0.46	0.94	0.22	0.62	0.00
		SNP	1.00	0.76	0.96	0.56	0.76	0.20
	AR(0.8)	Mat	1.00	0.38	0.94	0.18	0.50	0.04
		SNP	1.00	0.64	1.00	0.48	0.74	0.16
Nonstat	Ind	Mat	1.00	1.00	1.00	0.92	0.92	0.28
		SNP	1.00	1.00	1.00	1.00	0.98	0.60
	AR(0.8)	Mat	1.00	1.00	1.00	0.88	0.94	0.20
		SNP	1.00	1.00	1.00	0.96	1.00	0.66

2.6 Data Analysis

In light of PERMANOVA's demonstrated failure to preserve the size of the hypothesis test in the face of spatial and multivariate dependence, we revisit the analysis of Barberán et al. (2015) in which the authors determined which, if any, of a set of environmental and household covariates affect the indoor fungal community composition of homes. The covariates of interest included mean annual precipitation (MAP), mean annual temperature (MAT), net primary productivity (NPP), elevation, age of the home, number of bedrooms, number of inhabitants, female-to-male ratio of the home's inhabitants, smoking status, number of dogs/cats/birds, whether or not the home has a basement, and number of days with the windows open. Using PERMANOVA, they find that the effects of outdoor variables and geographic location are more pronounced than the household covariates, but note that the presence of a basement in the home, the age of the home, and the presence of a dog also affect the composition of the indoor fungal microbiome.

We follow the intuition of Barberán et al. (2015) and compile a similar list of covariates. In addition to those listed above, we include an indicator that the land is designated as forested, an indicator that the home is a rental unit, and the type of home (single family detached, multi-family dwelling, mobile). We replace the number of days with the windows open with the type of ventilation

(central air-conditioning, central heat, window air-conditioning). NPP was missing for 81 of the sampling locations, and when considering only indoor fungal taxa, an additional 24 sampling locations had no present taxa. These locations have been removed, leaving $n = 1{,}226$ locations and $p = 20$ covariates in the analysis.

Using both PERMANOVA and the proposed nonparametric method, we investigated each covariate's ability to affect the composition of the taxa identified as the indoor fungal microbiome. SNP was run for 80,000 total iterations, keeping the final 30,000 posterior samples. Unlike in the simulation study, the maximum number of groups is set to $K = 500 < m$. We utilized the downsampling strategy discussed in Section 2.4 to build the spatial basis functions. The first few estimated basis functions are mapped in Figure 2.2. The first several functions reflect the nonstationarity in the data, while

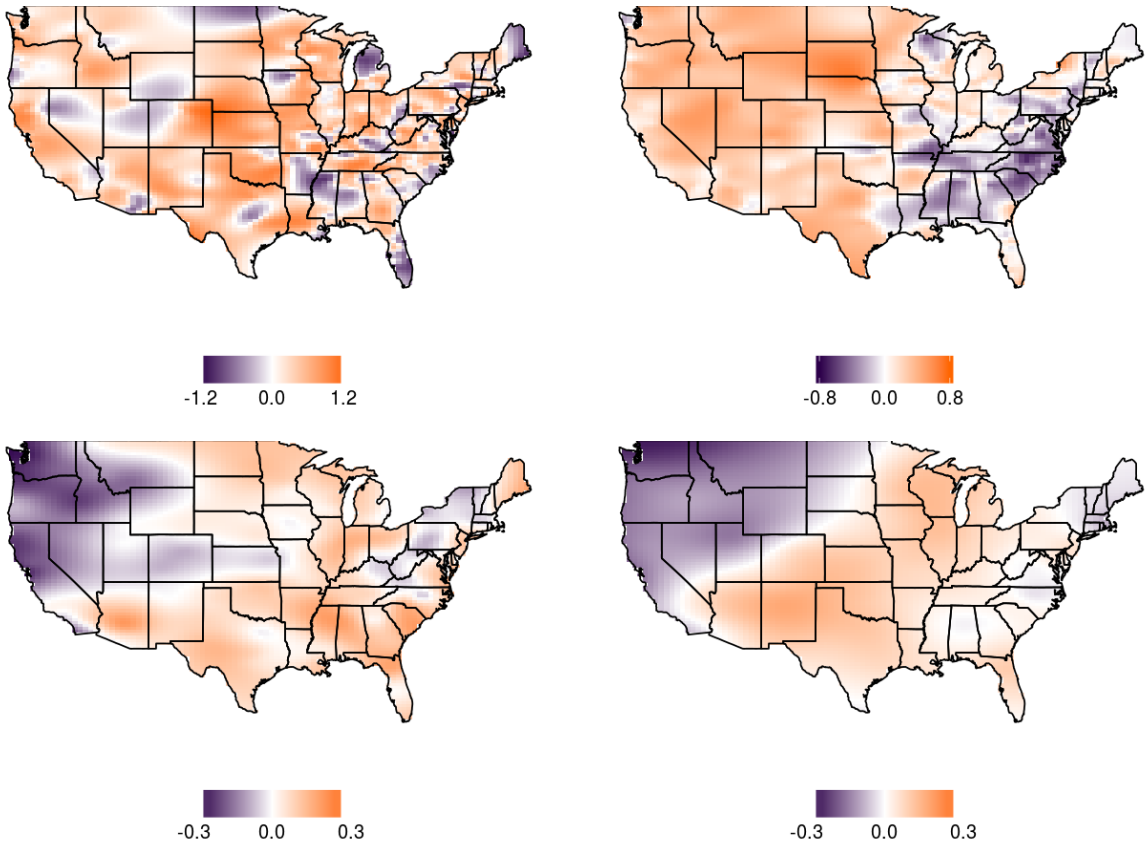


Figure 2.2 Maps of the first four spatial basis functions estimated from the WLOH data.

later basis functions reflect smooth spatial variation. Reported in Table 2.4 for each covariate are the P-value from PERMANOVA, the posterior probability of the null hypothesis, the posterior expected

number of taxa for which the covariate is selected, and a count of the number of taxa for which the associated coefficient value is positive or negative, assessed as $\sum_{j=1}^{763} \mathbb{1}\{\mathrm{P}(\beta_{jr} > 0 | \mathbf{Y}) > 0.975\}$ and $\sum_{j=1}^{763} \mathbb{1}\{\mathrm{P}(\beta_{jr} < 0 | \mathbf{Y}) > 0.975\}$, respectively, for the proposed model.

Table 2.4 Summary of variable selection results from PERMANOVA (PERM) and the proposed spatial nonparametric method (SNP). P-values are reported from PERM, and the posterior probability of the null hypothesis, the expected number of taxa for which the covariate is included, and the number of taxa for which the coefficient value is positive or negative are reported for SNP.

	PERM	SNP			
Covariate	P-value	$\mathrm{P}(\mathrm{M}_r = 0 \mid \mathbf{Y})$	$\mathrm{E}[\mathrm{M}_r \mid \mathbf{Y}]$	#Positive	#Negative
NPP	< 0.001	0.00	445	38	161
MAT	< 0.001	0.00	349	40	122
MAP	< 0.001	0.00	131	5	14
Central A/C	< 0.001	0.00	117	6	5
Multifamily dwelling	0.038	0.00	82	9	0
Forested	< 0.001	0.00	35	0	0
Elevation	< 0.001	0.00	15	0	1
Window A/C	< 0.001	0.00	15	0	1
Older home	0.078	0.00	13	0	0
Central heat	0.015	0.03	11	0	0
Basement	< 0.001	0.04	8	0	0
Number of dogs	0.152	0.05	5	0	0
Rental home	0.075	0.16	4	0	0
Number of occupants	0.016	0.43	1	0	0
Number of bedrooms	0.386	0.46	1	0	0
Mobile home	0.289	0.48	1	0	0
Smoking status	0.756	0.49	1	0	0
Percentage of females	0.735	0.51	1	0	0
Number of birds	0.627	0.51	1	0	0
Number of cats	0.558	0.76	0	0	0

Comparing the P-values from PERMANOVA and the posterior probability of the null hypothesis from SNP, we see that the two models largely agree, but we can identify several covariates that PERMANOVA includes at either the 0.05 or 0.10 significance level that would not be included in the SNP model. Given the inflated Type I error rates of the PERMANOVA test under spatial dependence in the simulation study, it seems likely that these are false positives. The proposed method is able to identify both covariates that are important to many taxa (e.g., MAT) and those that are important only to a few (e.g., whether or not a home is older). In addition, we are able to precisely describe *how* covariates influence particular taxa. For example, as one would expect, we note that most fungal taxa prefer cooler climes, but that there are some taxa that seem to thrive in the warmer temperatures. Generally, we corroborate the findings of Barberán et al. (2015) and conclude that

geographic and climatic factors are most influential to the indoor fungal microbiome composition. The household covariates that appear as influential are whether or not the home is older, the presence of a basement, whether or not the home is a multifamily dwelling, and whether or not the home has air-conditioning or central heating, all of which play a role in increasing the interaction between the indoor environment and the outdoors.

The 763 species are grouped into an estimated (posterior mean) 47 clusters. The largest clusters, based off of a k-means clustering algorithm with 47 clusters and using $1-\mathrm{P}(g_j = g_{j'})$ as the dissimilarity matrix, contain taxa that exhibit little spatial clustering and tend to be present across the country. The smaller clusters tend to group together taxa that exhibit more localized presence. For example, in Figure 2.3, the left panel displays the presence for the 100 taxa assigned to the largest cluster and the right panel displays the presence for the 3 taxa assigned to a smaller cluster.

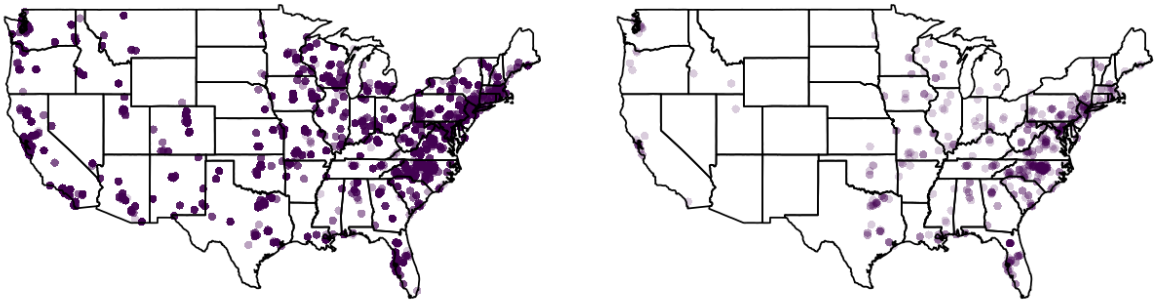


Figure 2.3 Map of presence for taxa assigned to a large cluster of 100 taxa and a small cluster of 3 taxa. A darker point indicates that a higher number of taxa are present in a location.

In as much as our results add to those of previous analyses using data from the WLOH project, it is worth commenting about the additional biological insights our approach offers. Barberán et al. (2015) found that, compared to bacteria, the composition of fungi in homes tended to be much more strongly driven by outdoor environmental conditions. In our analysis, this conclusion is even more strongly supported. The primary factors associated with differences in the composition of indoor fungi among households were those associated with climate and its effects, and nearly all (94.8%) significant associations of individual taxa with particular covariates were associations with these environmental factors.

Net Primary Productivity (NPP) was a particularly important correlate of the composition of indoor fungi. In the United States, NPP is highly correlated with forest cover, such that areas with higher NPP are almost always forests. In this light, it is perhaps not surprising that species more common in regions with high NPP were species associated with forests and dead and down wood,

including multiple taxa of the species *Xylobolus annosus*. Conversely, species that became less common under high NPP tended to be from the genera *Alternaria*, *Cladosporum*, *Aspergillus*, and *Phoma*, many of which are associated with decaying plant material. Fungi from decaying plant material, much of which is in leaf litter, might be more likely to become airborne in open habitats such as grasslands. Many species were also influenced by the direct effects of the mean annual temperature or precipitation in the region in which a house was located.

One of the few non-environmental covariates identified as influential was whether or not the home is a multifamily dwelling. Multifamily dwellings tended to favor fungi associated with human bodies or foods. These included three *Candida* taxa, *Cryptococcus oeirensis*, *Penicillium concetricum*, and the brewer's yeast (*Saccharomyces cerevisae*). Also more common in these homes were *Rhodotorula mucilagnosa* and *Cystofilobasidium capitatum*, both of which do well under stressful conditions, such as those associated with bathrooms that are frequently cleaned. The way in which a house was heated or cooled also influenced which species were present. In particular, as has been noted in smaller scale studies (Hamada & Fujita, 2002), we confirm here that houses with air conditioning tend to be more likely to have *Cladosporium* and *Penicillium* fungi, which are known to grow in air conditioning units and then spread through houses. Air conditioners were also associated with several other fungal species, including the wood rot fungus *Physisporinus vitreus*, a pattern for which the mechanistic links deserve more study.

Considering that the homes we studied differed greatly in their size, number of occupants, age, design, and much more, the fact that these variables influence so very little of fungal composition is striking. Houses, in general, favor some fungi relative to others and yet just which species appears to depend nearly exclusively on where the house is built.

2.7 Discussion

In this paper, we introduced a nonparametric Bayesian model for identifying factors that influence microbiome composition, as well as a covariance estimator amenable to high-dimensional, binary data akin to that of Hall et al. (2008). The proposed model uses spike-and-slab variable selection to identify covariates that influence the occupancy probability of even a small subset of the taxa. It also utilizes a set of orthogonal, data-driven spatial basis functions and a Dirichlet process prior over their associated loadings to cluster the OTUs into groups of taxa that exhibit similar spatial responses, allowing dimension reduction in both the number of spatial locations and the number of taxa under consideration, greatly alleviating the computational burden compared to a parametric spatial model.

We demonstrated via simulation that the proposed model outperforms a naïve nonspatial model and PERMANOVA in identifying influential covariates, and showed that violating the assumption of exchangeability of sampling locations underlying PERMANOVA leads to Type I error rates that are

not well-controlled. We also showed that the proposed model is able to better identify low prevalence and/or small magnitude covariate effects as compared to a parametric spatial competitor.

We applied our proposed model to the indoor fungal microbiome from the Wild Life of Our Homes project as identified in Barberán et al. (2015). We were able to broadly substantiate their conclusion that geography and climate are the most influential factors affecting indoor fungal communities, and we provided additional detail in describing how factors affect particular taxa rather than simply classifying factors are influential or unimportant.

This work primarily focused on the global hypothesis of whether or not a covariate influences microbiome composition as a whole. However, the model also allows for local hypothesis tests of individual covariate values, which have not been fully explored here. We discussed the application and potential of these local tests, but did not rigorously test the true and false positive rates for covariate effects on individual taxa. An additional area of focus for future work is to expedite and improve the covariance estimation process to scale with large problems.

CHAPTER

3

BAYESIAN VARIABLE SELECTION FOR HIGH-DIMENSIONAL RANK DATA

3.1 Introduction

The advancement of high throughput sequencing technologies made previously cost-prohibitive study of DNA material accessible in a variety of fields. In particular, interest in studying microbiomes, which are communities of microorganisms that occupy a specific ecological niche, has surged. To date, much of the work in the microbiome arena has focused on drawing connections and describing correlations between microbial community composition and specific outcomes. For example, childhood exposure to low fungal diversity has been linked to development of asthma (Dannemiller et al., 2014), imbalances in microbiome composition have been linked to Type 2 diabetes (Qin et al., 2012), and reduced diversity in the gut microbiome has been tied to obesity (Ley et al., 2005; Turnbaugh et al., 2009) and Crohn's disease (Dicksved et al., 2008). Recently, there has been an effort to begin identifying characteristics of "healthy" microbiomes within the body (Clemente et al., 2012; Human Microbiome Project Consortium, 2012), but fully leveraging this information requires knowledge of how to address complications in addition to identifying them. However, the availability of statistical tools to determine the effect of external factors on the composition of these communities is limited.

Microbiome data pose a number of significant analytic challenges, including high-dimensionality,

complex dependence, and non-normality. The raw abundance counts frequently display zero-inflation and overdispersion, and as an artifact of the sequencing process the counts reflect *relative* information rather than absolute information because the total amount of DNA material in a particular sample varies. Thus, standard multivariate statistical approaches are not applicable.

To overcome these complications, analyses frequently reduce the multivariate (or community) response to a summary metric (e.g., species richness, Shannon diversity index) or a measure of dissimilarity between samples, such as Euclidean or Bray-Curtis dissimilarity (Bray & Curtis, 1957). One of the most popular tools for analyzing communities takes the latter approach. "PERmutational Multivariate ANalysis Of VAriance" (PERMANOVA; Anderson, 2001; McArdle & Anderson, 2001) uses a nonparametric permutation-based test to determine which covariates affect dissimilarity between samples. These tools are applicable to many types of response variables (e.g., counts, presence-absence), but they can be difficult to interpret. These tests partition the pairwise dissimilarity between samples with respect to covariates, but do not provide clarity on *how* a covariate affects the composition or which taxa may be affected.

Recently, tools have been developed to analyze the compositional counts to address some of these concerns. Some propose transformations of the relative proportions, which are then treated using standard multivariate statistical methods (Aitchison, 1986; Fernandes et al., 2013; Mandal et al., 2015). A variety of methods have assumed a parametric model for the abundance counts, including the Dirichlet-multinomial (Chen & Li, 2013; Wadsworth et al., 2017), the negative binomial (Zhang et al., 2017), or the logistic normal multinomial (Xia et al., 2013; Grantham et al., 2017), among others. Still others have suggested nonparametric approaches, utilizing generalized linear models (Warton, 2011) or regression kernels (Zhao et al., 2015).

Rather than analyzing the counts directly, a transformation to binary presence-absence indicators is commonly made because the abundance counts in microbiome studies may be noisy. This transformation does not entirely remove the noted difficulties, but it can mitigate some of their effects. However, transforming the counts to their binary counterparts sacrifices much of the richness of the data. Instead, we propose analyzing the *ranks* of the taxa. In this way, we remove some of the noise in the raw counts without ceding all of the structure, and the data are no longer compositional. To our knowledge, this approach has not been applied to abundance counts, but data consisting of ranks are common in many applications throughout the social sciences, and many models have been developed for their analysis. Critchlow et al. (1991) provides a review of common parametric models, breaking them down into four broad categories: order statistics models, paired comparison models, distance-based models, and multistage models. Fligner & Verducci (1993) and Marden (1995) are comprehensive resources for both parametric and nonparametric approaches to rank data analysis. Maximum likelihood estimation for these models is computationally expensive for more than a handful of outcomes, and Bayesian implementations have been proposed to mitigate computational cost (Koop & Poirier, 1994; Yao & Böckenholt, 1999; Yu, 2000), to unify rankings from

a variety of sources (Johnson et al., 2002; Deng et al., 2014), and to combine rank data with other data types (Barney et al., 2015).

In this work, we develop a procedure that identifies covariates that influence microbiome composition, as quantified by taxa rankings, using a hierarchical Bayesian approach. We utilize a multivariate order statistics model for the ranks and a spike-and-slab prior for variable selection. In an extension to the base model, we detail the addition of a flexible basis function model to capture cross-dependence between taxa. This model provides global hypothesis tests of whether or not a covariate influences microbiome composition that can be used to screen many potential covariates, allowing the identification of targets for intervention or further research. Additionally, the model provides for local hypothesis tests for individual taxa which allows for a richer characterization of the mechanism through which external factors can affect microbial communities.

The remainder of the paper is laid out as follows: in Section 3.2, we introduce the motivating data from The American Gut Project; in Section 3.3, we describe the modeling procedure; in Section 3.4, we present a simulation study demonstrating the efficacy of our procedure against several competing models; in Section 3.5, we present an application of our method to a healthy subset of The American Gut Project; and we conclude with a brief summary and discussion in Section 3.6.

3.2 Motivating Data

The American Gut Project (americangut.org) is a citizen-science project with the twin goals of allowing participants to learn about their own microbes and arming researchers with publicly accessible data to study the relationships between microbes and human health. Participants in the project receive a sampling kit and instructions detailing the locations from which to take samples and how to handle samples safely. In addition to providing samples, participants are asked to fill out a survey with diet and lifestyle questions. After the samples are returned, the DNA is amplified using the direct PCR approach, sequenced, and classified into Operational Taxonomic Units (OTUs).

We focus on a subset of the data identified by The American Gut Project as the "healthy subset". These stool samples are from 3,942 adult participants of healthy weight, with no antibiotic use in the previous year, and no history of inflammatory bowel disease or diabetes. We stratify the data by sex, resulting in the removal of 88 participants without a female or male designation and an additional 31 participants with incomplete covariate information. Further, we include only OTUs that are present in at least 10% of the samples in our analysis. The final data include abundance counts for $m = 416$ taxa in $n = 3{,}823$ samples representing 2,019 female subjects and 1,804 males. A summary of the counts by sex are given in Table 3.1.

Many studies have indicated that diet has a strong influence on microbiome composition and diversity (De Filippo et al., 2010; Walker et al., 2011; Wu et al., 2011; David et al., 2014; Xu & Knight, 2015). Several other studies in both mice and humans have indicated that exercise and weight

Table 3.1 Summary of the abundance counts in a subset of the American Gut Project by sex.

		Nonzero Counts					
Sex	% Zeros	Min	Q1	Median	Mean	Q3	Max
Female	71.4%	1	4	12	98.51	47	43,940
Male	71.3%	1	4	12	98.23	47	41,049

changes can affect the gut microbiome, though the relationship is complicated (Santacruz et al., 2009; Queipo-Ortuño et al., 2013; Kang et al., 2014). As such, we consider $p = 21$ covariates, which are selected to reflect dietary differences between subjects, exercise habits, and lifestyle. We include indicators for broad dietary habits: whether or not a subject is gluten-free and whether or not a subject takes a multivitamin or probiotic. In addition, we consider more specific habits: whether or not a subject "regularly" (as opposed to "rarely" or "never") consumes fruit, dairy, poultry, red meat, seafood, vegetables, or whole grains and whether or not a subject consumes animal products treated with antibiotics, alcohol, or artificial sweeteners. There are also indicators for physical activity: whether or not a subject exercises regularly (at least 3-5 times per week) and whether or not they have gained or lost more than 10 pounds in the past year. We also include age in years, BMI, the presence of a cat in the home, the presence of a dog in the home, and whether or not the subject regularly uses cosmetics. The goal of this analysis is to identify which, if any, of these covariates influences microbiome composition for female or male subjects.

3.3 Model

Let $C_{ij} \in \{0, 1, 2, \ldots\}$ be the read count of OTU $j = 1, \ldots, m$ in sample $i = 1, \ldots, n$. Rather than model the counts directly, we retain only the indices and ranks of the nonzero counts, denoted by Y_{ij}. As is commonly argued in robust statistics, this transformation is insensitive to extreme counts such as those in Table 3.1. In particular, if $C_{ij} = 0$ then we set $Y_{ij} = 0$, and if C_{ij} is the k^{th} largest nonzero count in sample i then we set $Y_{ij} = k$. In the examples we consider, ties between nonzero counts are rare and so we simply randomize the order of the Y_{ij} to break ties. Ties could be handled more thoroughly within our framework by using slightly more elaborate orderings of the latent variables.

We adopt a multivariate order statistics model (Yao & Böckenholt, 1999; Yu, 2000) wherein the discrete data, Y_{ij}, are related to latent continuous random variables, Z_{ij}. We assume that if $Y_{ij} = 0$ then $Z_{ij} < 0$ and if $0 < Y_{ij} < Y_{ij'}$ then $0 < Z_{ij} < Z_{ij'}$. We then assume that the latent variables depend linearly on covariates $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})$, which have been centered and scaled, and write

$$Z_{ij} = \beta_{j0} + \mathbf{X}_i \boldsymbol{\beta}_j + e_{ij}, \tag{3.1}$$

where β_{j0} is an intercept, $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jp})'$ are unknown regression coefficients, and e_{ij} is mean-

zero error. We require $\text{Var}(e_{ij}) = 1$, in order to allow the scale of the regression coefficients to be identified. Identifiability of the parameters in (3.1) is discussed further in Appendix B.1.

The goal of this work is to identify covariates that affect microbiome composition. As such, we must determine whether or not the covariate affects the ranking of *any* of the m taxa. Thus, the hypothesis test for a global effect of the r^{th} covariate is

$$\text{H}_{0r} : \beta_{jr} = 0 \text{ for all } j \quad \text{versus} \quad \text{H}_{1r} : \beta_{jr} \neq 0 \text{ for some } j. \tag{3.2}$$

The modeling procedure for each of the components in (3.1) are described in the following subsections. Section 3.3.1 details the Bayesian approach to variable selection for the regression coefficients, while Section 3.3.2 presents an extension to accommodate correlation in the errors, e_{ij}. The computational details are provided in Appendices B.2 and B.3.

3.3.1 Variable Selection

Because we are interested in variable selection in an absolute sense, we adopt a spike-and-slab prior on the regression coefficients (Mitchell & Beauchamp, 1988; George & McCulloch, 1993; Kuo & Mallick, 1998). In particular, we write each coefficient as the product of an indicator variable, $\delta_{jr} \in \{0, 1\}$, and a magnitude, $\gamma_{jr} \in \mathbb{R}$, such that $\beta_{jr} = \delta_{jr}\gamma_{jr}$. The magnitudes have the standard conjugate prior distribution, $\gamma_{jr} \overset{\text{indep}}{\sim} \text{N}(0, \tau_r^{-1})$ with $\tau_r \overset{\text{iid}}{\sim} \text{Gamma}(a_\tau, b_\tau)$. The indicator variables have prior distribution $\delta_{jr} \overset{\text{indep}}{\sim} \text{Bernoulli}(\pi_r)$, with $\pi_r \sim \text{Beta}(1, m)$ (Castillo & Vaart, 2012). We do not wish to include the taxa-specific intercept terms in the variable selection procedure, so they are given the usual prior distribution, $\beta_{j0} \overset{\text{iid}}{\sim} \text{N}(0, \tau_0^{-1})$ with $\tau_0 \sim \text{Gamma}(a_0, b_0)$.

This formulation provides a straightforward framework for testing the global hypotheses in (3.2). Let $\text{M}_r = \sum_{j=1}^{m} \delta_{jr}$ for $r = 1, \ldots, p$, so that M_r represents the number of taxa for which the r^{th} covariate is included. Then, we can rewrite the hypotheses equivalently as

$$\text{H}_{0r} : \text{M}_r = 0 \quad \text{versus} \quad \text{H}_{1r} : \text{M}_r > 0. \tag{3.3}$$

We reject H_{0r} if its posterior probability, $\text{P}(\text{M}_r = 0 \,|\, \mathbf{Y})$, is less than the threshold, $t \in [0, 1]$.

3.3.2 Dependence Between Taxa

In the microbiome setting, it is plausible that there may be strong dependence between taxa. For example, several taxa may require the same nutrients to thrive and thus be generally found together. To account for dependence between taxa, we follow a similar approach to that of Chapter 2 and assume that e_{ij} can be written as the sum of two independent contributions: a structural component, ξ_{ij}, and an independent error, ϵ_{ij}. Because of the requirement that $\text{Var}(e_{ij}) = 1$, without loss of generality, we can write $\text{Var}(\xi_{ij}) = \rho \in [0, 1]$, thus leaving the independent error to be distributed

as $\epsilon_{ij} \overset{\text{iid}}{\sim} \text{N}(0, 1-\rho)$. We use an uninformative prior, $\rho \sim \text{U}(0,1)$, to reflect a lack of prior knowledge about the share of the variance contributed by the structural component.

For the structural component, we use a basis function expansion and write $\xi_{ij} = \boldsymbol{\alpha}_i' \boldsymbol{\Psi}_j$, where $\boldsymbol{\Psi}_j$ is a set of L basis functions common to all observations and $\boldsymbol{\alpha}_i = (\alpha_{i1}, \ldots, \alpha_{iL})'$ are the associated loadings for sample i. The loadings are given the conjugate prior distribution, $\alpha_{il} \overset{\text{iid}}{\sim} \text{N}(0, \rho)$. The data-driven basis functions are constructed using the rank correlation matrix of the observed ranks, Y_{ij}, using Kendall's τ, $\tilde{\boldsymbol{\Phi}}$ (Kendall, 1938; Kendall, 1945). This is an appealing choice for our rank method because for nonzero counts the rank correlation is the same as the correlation we would compute using the latent Z_{ij}, whose dependence we are trying to estimate. Let $\tilde{\phi}_1, \ldots, \tilde{\phi}_L$ be the leading L eigenvectors of $\tilde{\boldsymbol{\Phi}}$, scaled by the square-root of their associated eigenvalue. We choose L such that the included eigenvectors account for a specified percentage of explained variance, which in our implementation is 80%. To satisfy the requirement that $\text{Var}(\xi_{ij}) = \rho$, we must enforce the constraint $\sum_{l=1}^{L} \psi_{jl}^2 = 1$ for all j. If $L < m$, this will require rescaling the current estimates, so we set

$$\psi_{jl} = \left[\frac{1}{\sum_{l=1}^{L} \tilde{\phi}_{jl}^2} \right]^{\frac{1}{2}} \tilde{\phi}_{jl}.$$

Finally, we rotate by the right singular vectors to restore orthogonality and take the transpose to obtain the $(L \times m)$ collection of basis functions, $\boldsymbol{\Psi}$. In concert with the specifications from the previous section, the extended model for the latent process with random effects becomes

$$Z_{ij} = \beta_{j0} + \mathbf{X}_i \boldsymbol{\beta}_j + \boldsymbol{\alpha}_i' \boldsymbol{\Psi}_j + \epsilon_{ij} = \beta_{j0} + \sum_{r=1}^{p} X_{ir} \delta_{jr} \gamma_{jr} + \sum_{l=1}^{L} \alpha_{il} \psi_{jl} + \epsilon_{ij}.$$

3.4 Simulation Study

In this study, we consider data generated while varying four factors: (i) the level of zero-inflation, (ii) the dependence between OTUs, (iii) the magnitude of the covariate effects, and (iv) the prevalence of the covariate effects. We assess how these factors affect the overall true positive rate, false positive rate, and false discovery rate of the global hypothesis test in (3.3). To determine each model's ability to identify influential covariates, we also present the inclusion rate for each influential covariate, broken out by magnitude and prevalence.

3.4.1 Methods

For each of $n = 300$ observations, we generate $p = 20$ covariates, $\mathbf{X}_i$, as independent (over i) draws from a mean-zero Gaussian process with $\text{Cov}(X_{ir}, X_{ir'}) = 0.8^{|r-r'|}$. These covariates are then centered

and scaled. Of the 20 covariates, we randomly designate $p_0 = 6$ to be influential and the remainder to be unimportant, with $\beta_{jr} = 0$ for all j. The influential covariates are then randomized to combinations of prevalence (100%, 50%, or 10%) and coefficient magnitude. First, we select the coefficients that will have nonzero β_{jr} for 50% or 10% of the OTUs. Then, the nonzero coefficients for the selected covariate in each case are split evenly between a large positive magnitude, $\beta_{jr} = 1$, a large negative magnitude, $\beta_{jr} = -1$, a small positive magnitude, $\beta_{jr} = 0.6$, and a small negative magnitude, $\beta_{jr} = -0.6$.

The simulated counts are then generated from a zero-inflated, overdispersed Poisson distribution that has contamination in the counts as

$$C_{ij} \overset{\text{indep}}{\sim} \text{Poisson}\big(I_{ij}\big[A_{ij} + D_{ij}\exp(\eta_{ij})\big]\big), \tag{3.4}$$

where $\eta_{ij} = 1 + \sum_{r=1}^{p} X_{ir}\beta_{jr} + \theta_{ij}$, $\boldsymbol{\theta}_i \overset{\text{iid}}{\sim} \text{N}(\mathbf{0}, \mathbf{V})$ is a random effect term to induce correlation across the taxa, $I_{ij} \overset{\text{indep}}{\sim} \text{Bernoulli}(\text{expit}\{0.2\mathbf{X}_i\boldsymbol{\beta}_j + c\})$ is a zero-inflation term, $D_{ij} \overset{\text{iid}}{\sim} \text{Gamma}(10, 10)$ is an overdispersion term, and $A_{ij} \overset{\text{iid}}{\sim} \text{Gamma}(0.5, 0.01)$ is a contamination term with heavy tails. To obtain low, medium, and high levels of zero inflation, we use $c = 1.5$, $c = 0.5$, and $c = -1$, respectively. The structure of $\mathbf{V}$ determines whether or not the random effects are generated with correlation across the taxa. To induce multivariate correlation, we set $\text{Cov}(\theta_{ij}, \theta_{ij'}) = V_{jj'} = 0.8^{|j-j'|}$. In the independent case, $\theta_{ij} \equiv 0$.

Within each of $N = 50$ replications, we fit the proposed model for taxa ranks without random effects and several competing models:

(i) **B-Binary**: Bayesian binary variable selection model with a probit link function, described as the "Nonspatial variable selection (NS)" model in Chapter 2.

(ii) **B-Ranks**: Proposed Bayesian nonparametric variable selection for ranks.

(iii) **P-Binary**: PERMANOVA with the presence-absence response.

(iv) **P-Ranks**: PERMANOVA with the ranks response.

(v) **P-Counts**: PERMANOVA with the observed counts response.

Each of the PERMANOVA models is fit using Bray-Curtis dissimilarity, as implemented in the `R` package `vegan 2.4-3`. The Bayesian models are run for a total of 25,000 iterations with a burn-in period of 10,000 iterations, with convergence monitored by inspecting trace plots. To determine the threshold for deeming the r^{th} covariate influential, we follow the direct posterior probability

approach of Newton et al. (2004) and set t to be the maximum value in $[0,1]$ such that

$$\frac{\sum_{r=1}^{p} \tilde{\mathrm{P}}(M_r = 0 \mid \mathbf{Y})\, \mathbb{1}\{\tilde{\mathrm{P}}(M_r = 0 \mid \mathbf{Y}) < t\}}{1 \vee \left(\sum_{r=1}^{p} \mathbb{1}\{\tilde{\mathrm{P}}(M_r = 0 \mid \mathbf{Y}) < t\}\right)} \leq 0.10,$$

where $\tilde{\mathrm{P}}(M_r = 0 \mid \mathbf{Y})$ is the estimated posterior probability of the global null from MCMC. This data-driven threshold controls the false discovery rate at 10%. The standard significance level of 0.05 is used for the PERMANOVA tests.

We use average true positive rate (TPR), false positive rate (FPR), false discovery rate (FDR), and area under the ROC curve (AUC; Hanley & McNeil, 1982) to assess overall performance for each model, as presented in Table 3.2 along with the average computing time in minutes. The true positive rate (or recall) for a given threshold, t, is the percentage of truly influential covariates that are identified as such by the model,

$$\mathrm{TPR}(t) = \frac{\sum_{r=1}^{p} M_r^* \mathbb{1}\{\mathrm{P}(M_r = 0 \mid \mathbf{Y}) < t\}}{p_0},$$

where $M_r^* = \mathbb{1}\{r^{\text{th}}$ covariate is truly influential$\}$. The false positive rate is the percentage of truly unimportant covariates that the model identifies as influential,

$$\mathrm{FPR}(t) = \frac{\sum_{r=1}^{p} (1 - M_r^*) \mathbb{1}\{\mathrm{P}(M_r = 0 \mid \mathbf{Y}) < t\}}{p - p_0}.$$

The false discovery rate is the percentage of covariates that the model identifies as influential that are truly unimportant,

$$\mathrm{FDR}(t) = \frac{\sum_{r=1}^{p} (1 - M_r^*) \mathbb{1}\{\mathrm{P}(M_r = 0 \mid \mathbf{Y}) < t\}}{1 \vee \left(\sum_{r=1}^{p} (1 - M_r^*) \mathbb{1}\{\mathrm{P}(M_r = 0 \mid \mathbf{Y}) < t\}\right)},$$

where the denominator is adjusted to avoid division by 0. The area under the ROC curve is calculated using the `R` package `pROC 1.9-1` using the M_r^* indicators as the responses and $\tilde{p}_r = 1 - \tilde{\mathrm{P}}(M_r = 0 \mid \mathbf{Y})$ as the predicted probabilities. Note that AUC values are not computed for the PERMANOVA models because the method does not provide a probability of the null hypothesis.

Of particular interest in the microbiome setting are low magnitude and low prevalence covariates. Thus, in addition to overall performance, we are interested in performance for individual covariates. Figure 3.1 displays the inclusion rate (IR) for each influential covariate denoted by its prevalence (100%, 50%, 10%) and its magnitude, large (L) or small (S). The inclusion rate represents

the proportion of the N replications for which each covariate was correctly identified as influential in each model.

$$\text{IR}(t) = \frac{1}{N}\sum_{s=1}^{N} \mathbb{1}\{\text{P}^{(s)}(M_{r^*} = 0 \,|\, \mathbf{Y}) < t\},$$

where $\text{P}^{(s)}(M_r = 0 \,|\, \mathbf{Y})$ denotes the posterior probability of the null for the r^{th} covariate in replication s, and $r^* = 1, \ldots, p_0$ indexes the influential covariates.

Finally, Table 3.3 presents the true positive rate, false positive rate, and false discovery rate for each method when we consider using a "registered" threshold to find the rejection region. This threshold is calculated to control the overall false positive rate for the method at or below 0.05 within each setting. In order to apply this process to the PERMANOVA models as well as the Bayesian models, the threshold is applied to the P-values rather than the posterior probabilities of the null. Though not practical to use to develop a decision rule because it requires knowledge of the truth, this registered threshold facilitates model comparison because it corrects for models that may have tendencies to be too conservative or liberal in practice.

3.4.2 Results

By comparing the two groups of columns in Table 3.2, we can see that the models seem robust to the presence of cross-correlation between the taxa because the model performance is similar between the two groups. In the case of PERMANOVA, this is expected, because resampling is done at the observation level, and thus any dependence between taxa is preserved. Across all settings, the data-driven threshold controls the false discovery rate for B-Ranks close to the nominal value of 0.10, though it fails to do so for B-Binary. The threshold seems to be too liberal in the binary case, which perhaps reflects poor separation between the null and non-null covariates.

In the case of low zero-inflation, the proposed model outperforms the competitors handily. The binary models struggle relative to the models based on counts and ranks, though the Bayesian model does well in comparison to PERMANOVA version. We also note that the model with counts is outperformed by the ranks models, likely because the ranks are less affected by contamination. These results hold in the case of medium zero-inflation as well, with performance of the ranks and counts models beginning to wane with the reduced information. Finally, in the case of high zero-inflation, performance for all models degrades, with high false discovery rates for nearly all models other than B-Ranks. However, the proposed Bayesian model for ranks still edges out the competitors, despite its reduced power as the proportion of zeros increases. After registering the false positive rates in Table 3.3, the proposed model emerges as the clear winner across all settings.

Inspecting Figure 3.1, we immediately see that the Bayesian models outperform the PERMANOVA models in all settings, i.e., the purple lines are generally higher than the orange ones. Similarly, the ranks models, represented by solid lines, are usually the highest within each color group. In the low

Table 3.2 Summary of average computing time in minutes, true positive rate (TPR), false positive rate (FPR), false discovery rate (FDR), and area under the ROC curve (AUC) for the Bayesian binary variable selection model (B-Binary), the proposed Bayesian ranks variable selection model (B-Ranks), PERMANOVA with the binary response (P-Binary), ranks (P-Ranks), and the observed counts (P-Counts).

		Dependence Between Taxa									
		Independence					Autoregressive				
Zero-inflation	Model	Time	TPR	FPR	FDR	AUC	Time	TPR	FPR	FDR	AUC
Low	B-Binary	7.18	0.63	0.12	0.29	0.82	7.19	0.60	0.11	0.27	0.81
(c=1.5)	B-Ranks	7.34	0.75	0.05	0.13	0.93	7.37	0.75	0.05	0.13	0.92
	P-Binary	1.06	0.33	0.05	0.21		1.06	0.31	0.06	0.28	
	P-Ranks	1.05	0.58	0.04	0.10		1.05	0.55	0.05	0.15	
	P-Counts	1.05	0.49	0.04	0.13		1.05	0.53	0.04	0.12	
Medium	B-Binary	6.97	0.65	0.12	0.28	0.84	6.97	0.61	0.15	0.34	0.82
(c=0.5)	B-Ranks	7.47	0.67	0.05	0.13	0.90	7.46	0.65	0.04	0.13	0.90
	P-Binary	1.06	0.36	0.04	0.20		1.06	0.34	0.05	0.19	
	P-Ranks	1.05	0.55	0.04	0.12		1.05	0.52	0.06	0.17	
	P-Counts	1.05	0.47	0.05	0.17		1.05	0.54	0.05	0.15	
High	B-Binary	7.08	0.55	0.11	0.30	0.80	7.08	0.60	0.12	0.29	0.82
(c=-1)	B-Ranks	7.90	0.44	0.03	0.12	0.84	7.91	0.48	0.04	0.12	0.85
	P-Binary	1.06	0.25	0.04	0.20		1.06	0.29	0.06	0.26	
	P-Ranks	1.05	0.21	0.04	0.27		1.05	0.26	0.05	0.23	
	P-Counts	1.05	0.34	0.04	0.18		1.05	0.43	0.04	0.15	

zero-inflation cases in the top row, the proposed model, B-Ranks, is either the best at recovering each of the influential covariates, or it is a very close second. As the level of zero-inflation increases, this dominance dissipates, but the proposed model is still very competitive. In the medium zero-inflation settings in the second row, B-Binary seems to be closer to B-Ranks in performance, but this is partially due to an overly liberal threshold, as is evident from the elevated false positive and false discovery rates in Table 3.2. Finally, B-Binary appears to perform best across the board in the high zero-inflation setting, but this is again tempered by the high false positive and false discovery rates. Excluding B-Binary, the proposed model does very well among the remaining models across all settings, matching or outperforming the competitors, particularly for the lower magnitude and prevalence covariates of interest.

3.5 Data Analysis

Figure 3.2 displays the empirical distribution function of the correlation between taxa in The American Gut Project data for the female and male subgroups. The null distribution, obtained by permutation, is given in gray. The data do not seem to display evidence of strong dependence between

Table 3.3 Summary of true positive rate (TPR), false positive rate (FPR), and false discovery rate (FDR) using the "registered" threshold to control overall FPR below 5% for the Bayesian binary variable selection model (B-Binary), the proposed Bayesian ranks variable selection model (B-Ranks), PERMANOVA with the binary response (P-Binary), ranks (P-Ranks), and the observed counts (P-Counts).

		Dependence Between Taxa					
		Independence			Autoregressive		
Zero-inflation	Model	TPR	FPR	FDR	TPR	FPR	FDR
Low	B-Binary	0.50	0.05	0.17	0.47	0.05	0.16
(c=1.5)	B-Ranks	0.75	0.05	0.12	0.77	0.05	0.11
	P-Binary	0.33	0.05	0.21	0.26	0.05	0.24
	P-Ranks	0.59	0.05	0.14	0.55	0.05	0.14
	P-Counts	0.51	0.05	0.14	0.54	0.05	0.13
Medium	B-Binary	0.50	0.05	0.16	0.50	0.05	0.17
(c=0.5)	B-Ranks	0.67	0.05	0.13	0.69	0.05	0.13
	P-Binary	0.39	0.05	0.22	0.34	0.05	0.19
	P-Ranks	0.55	0.05	0.14	0.52	0.05	0.15
	P-Counts	0.47	0.05	0.17	0.52	0.05	0.15
High	B-Binary	0.39	0.05	0.19	0.45	0.05	0.18
(c=-1)	B-Ranks	0.50	0.05	0.16	0.54	0.05	0.14
	P-Binary	0.28	0.05	0.23	0.25	0.05	0.22
	P-Ranks	0.22	0.05	0.30	0.26	0.05	0.22
	P-Counts	0.36	0.05	0.19	0.43	0.05	0.16

taxa, as the observed empirical distribution function does not seem to meaningfully deviate from the null distribution. Nonetheless, we fit the suite of models from the simulation study as well as the Bayesian models with random effects, denoted by the appellation "RE", to each subgroup of the data.

For the Bayesian models, we run the MCMC samplers for 60,000 iterations each, discarding the first 10,000 iterations as burn-in. Convergence is diagnosed by inspecting trace plots. Table 3.4 and Table 3.5 present the P-values from PERMANOVA using the binary, ranks, and counts responses and the posterior probability of the null obtained for the binary and ranks responses using the Bayesian models for the female and male subgroups, respectively. Note that the values of 1 are the result of rounding for presentation purposes.

Inspecting Table 3.4, we see that all three of the PERMANOVA models identify alcohol use as a significant factor, with P-Counts also identifying BMI as significant using the standard significance level of 0.05. However, only the P-value for whether or not a subject consumes alcohol in the P-Binary model remains significant after adjusting for multiple comparisons using a Bonferroni correction. B-Binary demonstrates separation between probiotic use, consumption of alcohol, and consumption

Table 3.4 Summary of the P-values from PERMANOVA and the posterior mean probability of the global null from the Bayesian models when applied to The American Gut Project data for females.

	PERMANOVA P-value			$P(M_r = 0 \mid \mathbf{Y})$			
Covariate	Binary	Ranks	Counts	Binary	Binary RE	Ranks	Ranks RE
Age	0.97	0.87	0.61	0.83	0.97	0.99	0.98
BMI	0.81	0.95	0.03	0.80	0.97	0.97	0.95
Uses alcohol	0.001	0.01	0.04	0.27	0.83	0.88	0.99
Presence of a cat	0.48	0.30	0.78	0.79	0.96	0.99	0.97
Antibiotic fed products	0.48	0.34	0.57	0.76	0.96	0.99	0.97
Presence of a dog	0.71	0.54	0.17	0.69	0.98	0.99	0.98
Uses cosmetics	0.20	0.83	0.15	0.79	0.96	0.99	0.92
Artificial sweeteners	0.98	0.76	0.69	0.83	0.95	0.97	0.98
Gluten-free	0.96	0.89	0.78	0.83	0.96	0.96	0.97
Uses multivitamin	0.08	0.18	0.52	0.63	0.95	0.95	0.77
Uses probiotic	0.24	0.30	0.62	0.26	0.96	0.95	0.97
Fruit	0.57	0.61	0.56	0.70	0.97	0.64	0.43
Dairy	0.82	0.66	0.75	0.81	0.96	0.97	0.98
Poultry	0.90	0.99	0.96	0.78	0.97	0.96	0.99
Red meat	0.78	0.89	0.71	0.81	0.95	0.96	0.95
Seafood	0.72	0.56	0.29	0.79	0.96	0.99	0.93
Vegetables	0.76	0.99	0.71	0.78	0.97	0.95	0.95
Whole grains	0.46	0.80	0.88	0.14	0.98	0.92	0.92
Regularly exercises	0.22	0.22	0.91	0.70	0.92	0.98	0.99
Weight gain in past year	0.79	0.97	0.89	0.81	0.97	0.98	0.97
Weight loss in past year	0.22	0.73	0.14	0.76	0.96	0.97	0.97

of whole grains versus the remainder of the covariates, but these are not deemed significant using the data-driven threshold defined in Section 3.4. With the addition of the random effects in B-Binary RE, this separation disappears entirely. Similary, B-Ranks and B-Ranks RE displays some separation between the posterior probability of the null for consumption of fruit and the other covariates, but there is not enough evidence identify it as influential. Using the binary response, the contribution of the structural component to the variance is estimated as $\tilde{\rho} = 0.951$, with a 95% credible interval of $(0.949, 0.952)$. In B-Ranks RE, it is estimated as $\tilde{\rho} = 0.626$, with a 95% credible interval of $(0.622, 0.630)$.

For males, all three of the PERMANOVA models identify consumption of whole grains as significant, with P-Binary and P-Ranks also identifying consumption of fruit as significant. As before, most of these results do not hold up after applying a Bonferroni correction, with only the P-value for consumption of whole grains in the P-Binary model remaining significant. Both of the Bayesian models fit to the binary response display evidence that consumption of vegetables influences microbiome composition, though the posterior probability for B-Binary RE is directly on the threshold while the posterior probability for B-Binary is well within the rejection region. The Bayesian models fit to the rank response do not identify any influential covariates. The contribution of the structural component to the variance using the binary response in B-Binary RE is estimated as $\tilde{\rho} = 0.951$, with

Table 3.5 Summary of the P-values from PERMANOVA and the posterior mean probability of the global null from the Bayesian models when applied to The American Gut Project data for males.

	PERMANOVA P-value			$P(M_r = 0 \mid \mathbf{Y})$			
Covariate	Binary	Ranks	Counts	Binary	Binary RE	Ranks	Ranks RE
Age	0.53	0.50	0.84	0.80	0.92	0.98	0.96
BMI	0.17	0.18	0.35	0.72	0.92	0.97	0.95
Uses alcohol	0.57	0.32	0.48	0.79	0.89	0.98	0.95
Presence of a cat	0.67	0.34	0.40	0.70	0.89	0.93	0.88
Antibiotic fed products	0.53	0.26	0.28	0.74	0.89	0.92	0.90
Presence of a dog	0.73	0.66	0.27	0.74	0.89	0.96	0.96
Uses cosmetics	0.71	0.17	0.25	0.79	0.89	0.97	0.97
Artificial sweeteners	0.09	0.14	0.25	0.58	0.89	0.97	0.92
Gluten-free	0.93	0.83	0.44	0.79	0.91	0.97	0.96
Uses multivitamin	0.59	0.86	0.26	0.81	0.90	0.96	1.00
Uses probiotic	0.44	0.80	0.16	0.80	0.90	0.91	1.00
Fruit	0.03	0.05	0.06	0.61	0.92	0.73	0.95
Dairy	0.45	0.39	0.67	0.77	0.88	0.98	0.95
Poultry	0.47	0.24	0.09	0.74	0.90	0.97	0.99
Red meat	0.95	0.98	0.67	0.80	0.90	0.95	0.96
Seafood	0.62	0.76	0.56	0.84	0.91	0.98	0.97
Vegetables	0.61	0.77	0.94	0.03	0.12	0.95	0.93
Whole grains	0.002	0.004	0.03	0.69	0.88	0.94	0.98
Regularly exercises	0.94	0.79	0.77	0.79	0.88	0.96	0.96
Weight gain in past year	0.48	0.67	0.60	0.82	0.87	0.96	0.68
Weight loss in past year	0.39	0.95	0.43	0.76	0.88	0.96	0.99

a 95% credible interval of $(0.950, 0.953)$. Using the rank response, it is estimated as $\tilde{\rho} = 0.438$, with a 95% credible interval of $(0.433, 0.443)$. Interestingly, the estimated contribution is similar across the subgroups using the binary response, but there is a large difference between the females and males when considering the rank response.

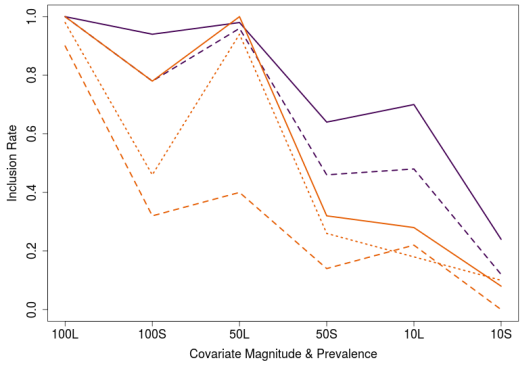
3.6 Discussion

In this paper, we introduced a Bayesian hierarchical model to identify covariates that influence microbiome composition using taxa rankings. The transformation to rankings allows for the retention of much of the information in taxa abundance, while removing some of the challenges posed by analyzing abundance counts directly. We detailed a base model that adopts a multivariate normal order statistics model for the ranks and a spike-and-slab prior for the regression coefficients, allowing for a global hypothesis test of whether or not a given covariate affects composition. In addition, we presented an extension of our model that uses data-driven basis functions in a random effects framework to capture cross-dependence between taxa, if the counts or ranks are highly correlated.

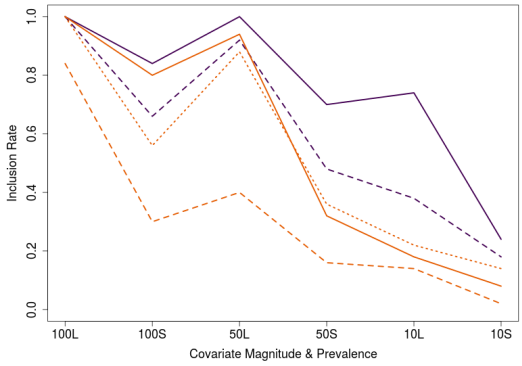
By simulation of data with zero-inflation, overdispersion, and contamination, we showed that in settings designed to resemble microbiome data, the proposed model based on ranks outperforms a Bayesian variable selection model using a binary response and PERMANOVA using either the binary, ranks, or count response in terms of overall variable selection across several levels of zero-inflation. In the case of the small magnitude and/or low prevalence covariates that are of particular interest in the microbiome setting, the proposed model represents a marked improvement over the competitors.

We applied the proposed model to stool samples in the "healthy subset" of The American Gut Project. We found that after adjusting for multiple comparisons using a Bonferroni correction, PERMANOVA applied to the binary response identified consumption of alcohol and consumption of whole grains as significant influences on microbiome composition for females and males, respectively. The Bayesian variable selection model from Chapter 2 identified consumption of vegetables as a significant influence on the microbiome composition of males. Overall, the majority of the models did not identify any of the covariates as influential after controlling for false discoveries using either a Bonferroni correction or a data-driven threshold.

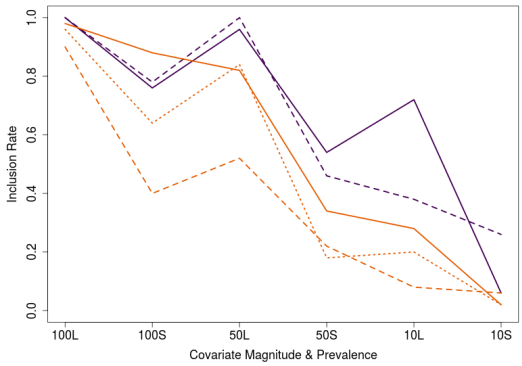
In both subgroups, the estimated share of the variance contributed by the structural component was surprisingly high using either the binary or ranks response. Based on Figure 3.2, we would expect the estimated $\tilde{\rho}$ values to be close to 0. However, when considering the binary response, $\tilde{\rho} \approx 0.95$ for both groups, and when considering the rank response, $\tilde{\rho} \approx 0.63$ and $\tilde{\rho} \approx 0.44$ for females and males, respectively. Based on simulation testing, we have seen similarly inflated estimates of ρ in the presence of zero-inflation and contamination in the counts. Thus, it is possible that the random effects models are more sensitive to these factors than the model assuming independence. Further study of the precise conditions in which the random effects models struggle is necessary.



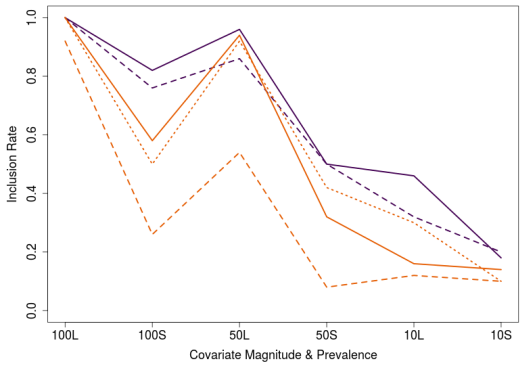
(a) Low zero-inflation, Independence



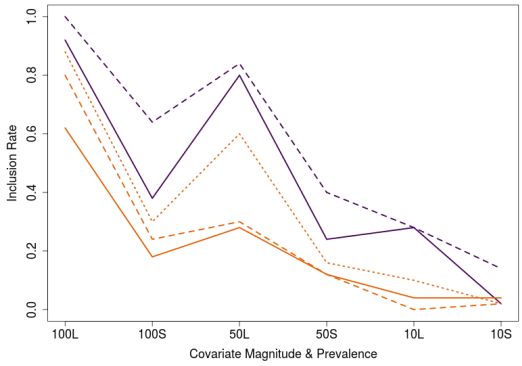
(b) Low zero-inflation, AR(0.8)



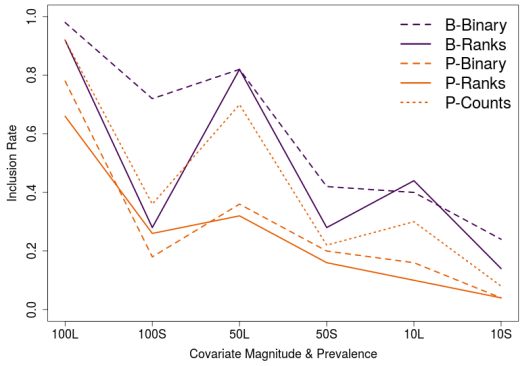
(c) Medium zero-inflation, Independence



(d) Medium zero-inflation, AR(0.8)

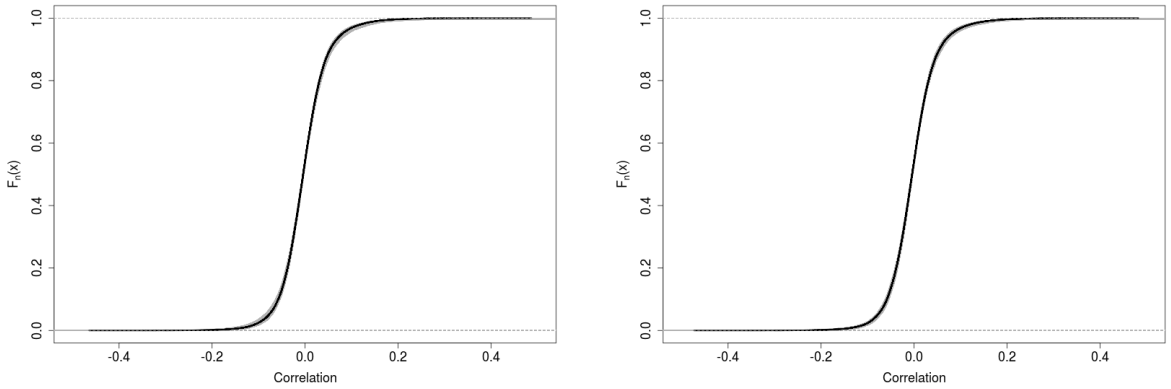


(e) High zero-inflation, Independence



(f) High zero-inflation, AR(0.8)

Figure 3.1 Inclusion rates for influential covariates by prevalence (100%, 50%, 10%) and magnitude (L=Large, S=Small) for the Bayesian binary variable selection model (B-Binary), the proposed Bayesian ranks variable selection model (B-Ranks), PERMANOVA with the binary response (P-Binary), ranks (P-Ranks), and the observed counts (P-Counts).



(a) Females **(b)** Males

Figure 3.2 Empirical cumulative density function for correlation between taxa in The American Gut Project data for the female and male subgroups, with the null distribution given in gray.

CHAPTER

4

A SURVEY OF MCMC ALGORITHMS FOR DIFFRACTION PATTERN ANALYSIS

4.1 Introduction

Knowledge of a material's crystallographic structure is a prerequisite for predicting and understanding properties of functional materials. To that end, diffraction techniques are powerful tools used by materials scientists to study atomic structures. Diffraction techniques generate diffraction profiles that provide structural information about the material of interest, such as the size and shape of the unit cell and the position of the atoms. Common techniques include X-ray and neutron diffraction, where X-rays or neutrons are scattered from the electron clouds of atoms or an atom's nucleus, respectively. Differences in the scattering properties of X-rays and neutrons provide complementary information. The scattering of the X-rays or neutrons results in constructive or destructive interference, where constructive interference gives rise to Bragg peaks (or *reflections*) that correspond to scattering from a set of crystal planes. Diffraction patterns are typically measured as a function of the scattering angle (2θ), yielding patterns as illustrated in Figure 4.1. Often, single crystals of a suitable size cannot be readily obtained for many materials because they are prohibitively difficult to grow. As a result, *powder diffraction* techniques were developed to measure powder samples made up of small, randomly oriented crystallites.

In this chapter, we illustrate the application of several algorithms to the fitting of neutron

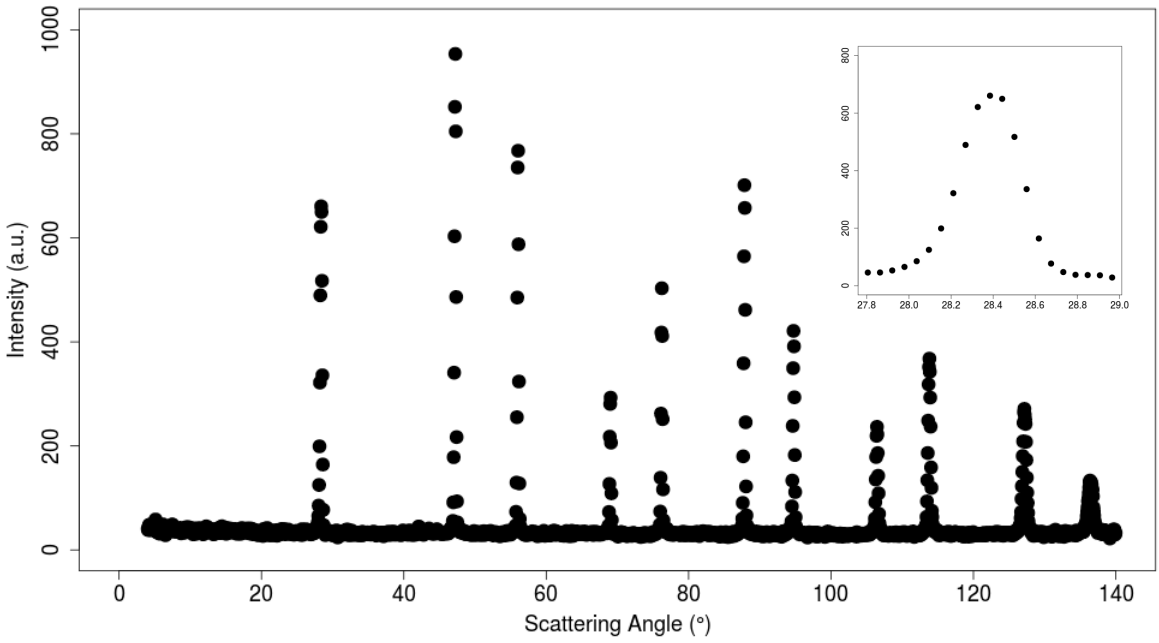


Figure 4.1 Angular dependent neutron diffraction profile from a NIST silicon standard reference material.

diffraction data. A neutron diffraction pattern of a National Institute of Standards and Technology (NIST) silicon standard reference material (SRM 640c) was measured on the HB2A High Resolution Neutron Powder Diffractometer (Garlea et al., 2010) at the High Flux Isotope Reactor at Oak Ridge National Laboratory using a wavelength of approximately 1.54 Å. The data are measured at $n = 2{,}345$ scattering angles, evenly spaced between 4° and 140°, for one hour. Figure 4.1 displays a scatterplot of observed intensity versus scattering angle. Inset in the figure is a close-up of the first observed peak, located at a scattering angle of approximately 28.38°.

Whereas diffuse scattering near Bragg peaks contains information about defects and disorder in a material, critical information about the structure is contained within the peaks themselves. The peaks in a diffraction profile may arise from a single Bragg peak or may consist of multiple peaks that are overlapping, a problem which is exacerbated in low symmetry and mixed-phase materials. This problem is the result of the compression of three-dimensional information about the crystal structure into one-dimensional diffraction patterns. The inability to resolve peaks complicates the identification of the crystal structure or symmetry, and can lead to an incorrect structure determination. Another difficulty of analyzing diffraction patterns is that multiple parameters may contribute to the same profile feature. For example, peak width may be influenced by both crystallite size and

microstrain, and their effects may be difficult to distinguish.

Early manual deconvolution methods for analyzing diffraction profiles were limited in their application to poorly resolved diffraction peaks and complicated crystal structures. These limitations led to the development of new approaches for structure determination, such as the Rietveld method (Rietveld, 1967; Rietveld, 1969). The Rietveld method is one of several methods for refining crystallographic parameters that proceed by minimizing the squared distance between a calculated profile pattern and the observed pattern. In this work, we limit our focus to the Rietveld method because it is the current standard approach for analyzing diffraction profiles. Many software packages are available to calculate theoretical diffraction patterns, given a set of crystallographic parameters, such as Materials Analysis Using Diffraction (MAUD; Lutterotti, 2010) or General Structure and Analysis System II (GSAS-II; Toby & Von Dreele, 2013). In the examples that follow, we will rely on GSAS-II to calculate diffraction patterns. Specifically, we employ GSAS-II within a variety of Bayesian samplers as in Fancher et al. (2016).

Rietveld analysis is a powerful tool, but it has limitations. The Rietveld approach can be very time consuming to perform, is highly sensitive to the order in which parameters are optimized, relies heavily on the practitioner's knowledge of their material, can return infeasible parameter estimates, and relies on ad-hoc rescaling techniques to quantify uncertainty. As an alternative, techniques from Bayesian statistics are being explored for diffraction profile analysis. Bayesian statistics allows a user to easily incorporate expert knowledge, prevents infeasible parameter values, and provides fully specified uncertainty.

In both Rietveld analysis and the Bayesian methods to be detailed, the observed intensities, Y_i, are modeled as a function of the intensity due to the material, $f(2\theta_i \,|\, \boldsymbol{\alpha})$, and the intensity due to background scattering, $b(2\theta_i \,|\, \boldsymbol{\gamma})$. In particular,

$$Y_i = f(2\theta_i \,|\, \boldsymbol{\alpha}) + b(2\theta_i \,|\, \boldsymbol{\gamma}) + e_i, \tag{4.1}$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_q)$ are the structural and instrumental parameters characterizing the material and experiment, $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_B)$ are the parameters governing the background intensity, and $e_1, \ldots, e_n$ are independent errors. Powder diffraction data exhibits heteroscedastic errors that are proportional to the intensity; i.e., the observed intensities show more variability for large intensity values than for small. As a result, the individual errors are modeled using a Normal distribution with mean zero and variance $\text{Var}(e_i) = \sigma_i^2$. The subscript i on the variance term, σ_i^2, reflects the fact that each observed intensity may have a different variance value. Thus, the full model for the observed intensities is $Y_i \overset{\text{indep}}{\sim} \text{N}\{f(2\theta_i \,|\, \boldsymbol{\alpha}) + b(2\theta_i \,|\, \boldsymbol{\gamma}), \sigma_i^2\}$, where $i = 1, \ldots, n$ indexes the observations and $\text{N}(\mu, \sigma^2)$ denotes the Normal distribution with mean μ and variance σ^2. The likelihood function,

$\mathscr{L}(\cdot)$, and log-likelihood function, $\ell(\cdot)$, that characterize this model are given by

$$\mathscr{L}(Y_1,\ldots,Y_n\,|\,\boldsymbol{\alpha},\boldsymbol{\gamma},\sigma_i^2)=\prod_{i=1}^{n}\phi\{Y_i;\,f(2\theta_i\,|\,\boldsymbol{\alpha})+b(2\theta_i\,|\,\boldsymbol{\gamma}),\sigma_i^2\}, \tag{4.2}$$

and

$$\begin{aligned}\ell(Y_1,\ldots,Y_n\,|\,\boldsymbol{\alpha},\boldsymbol{\gamma},\sigma_i^2)&=\sum_{i=1}^{n}\ln\big[\phi(Y_i\,|\,\boldsymbol{\alpha},\boldsymbol{\gamma},\sigma_i^2)\big]\\&=-\frac{n}{2}\ln(2\pi)-\sum_{i=1}^{n}\left[\ln(\sigma_i)+\frac{1}{2\sigma_i^2}\{Y_i-f(2\theta_i\,|\,\boldsymbol{\alpha})-b(2\theta_i\,|\,\boldsymbol{\gamma})\}^2\right],\end{aligned} \tag{4.3}$$

where $\phi\{\cdot;\mu,\sigma^2\}$ is the probability density function of the $\mathrm{N}(\mu,\sigma^2)$ distribution.

Despite the fact that the observed intensities are counts, the Normal distribution approximation to the Poisson distribution is reasonable in this setting because the observed intensity counts are large. For example, the minimum observed value in the neutron diffraction data is 21.94. In the descriptions that follow, it may be useful to refer to the set of structural, instrumental, and background parameters as $\boldsymbol{\vartheta}=(\boldsymbol{\alpha},\boldsymbol{\gamma})$. This group of parameters has $p=q+B$ elements, and is commonly called the set of *adjustable* parameters in diffraction profile analysis.

The remainder of this chapter provides an introduction to the Rietveld method in Section 4.2, and a Bayesian approach to structure refinement in Section 4.3. We outline the Bayesian modeling framework for diffraction profile analysis, introduce various sampling methods, and apply each to diffraction profile fitting. In Section 4.4, we compare the Bayesian approaches and finally in Section 4.5, we summarize.

4.2 Rietveld Refinement

The Rietveld method is a crystal structure refinement method that was developed by H. Rietveld to address the problem of poorly resolved diffraction peaks (Rietveld, 1967; Rietveld, 1969). In this method, all observed intensity is assigned to individual Bragg reflections that may overlap, and the refinement is carried out by refining the crystal structure and the parameters describing peak positions and shapes together. This allows for feedback between the improvement of the structural knowledge and improvement of the allocation of observed intensity. Other methods of structure refinement treat these two processes as separate procedures, and thus feedback during refinement is not present. The use of profile intensities, rather than the previously used method of integrated intensities[1], reduces the loss of information contained in overlapping peaks.

[1] Integrated intensity methods attempt to match the area underneath the peak rather than the full peak shape.

Rietveld refinement proceeds by minimizing a weighted least squares criterion

$$\text{WSSE} = \sum_{i=1}^{n} w_i (Y_i - Y_i^c)^2, \tag{4.4}$$

where $Y_i^c = f(2\theta_i \,|\, \boldsymbol{\alpha}) + b(2\theta_i \,|\, \boldsymbol{\gamma})$ is the calculated intensity from a software package (Young, 1993). The weights, w_i, may be defined in several ways. The simplest definition defines $w_i = Y_i^{-1}$. The intuition behind this approach relies on the mean-variance relationship of the Poisson distribution. In particular, for a parameter r_i, if $Y_i \sim \text{Poisson}(r_i)$ then $\text{E}[Y_i] = \text{Var}(Y_i) = r_i$. Ideally, one would weight the observations by the inverse of their variance, such that observations with high variability are downweighted and observations with low variability are upweighted. However, because the mean-variance parameter, r_i, is unknown, the observation value itself is used as an estimate of r_i. In subsequent calculations, we utilize the weights calculated by GSAS-II, as outlined in the *GSAS Technical Manual* (Larson & Von Dreele, 2004). Refinements are carried out until the "best fit" is obtained between the observed powder diffraction pattern and the calculated pattern based on the model. The set of parameter values that minimizes this criterion is the same as the set that maximizes the log-likelihood function in (4.3). To see this, substitute for Y_i^c in (4.4) and notice that with respect to $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$, $\text{WSSE} \propto \sum_{i=1}^{n} w_i \{Y_i - f(2\theta_i \,|\, \boldsymbol{\alpha}) - b(2\theta_i \,|\, \boldsymbol{\gamma})\}^2$. Similarly, $\ell(Y_1, \ldots, Y_n \,|\, \boldsymbol{\alpha}, \boldsymbol{\gamma}) \propto -\sum_{i=1}^{n} \sigma_i^{-2} \{Y_i - f(2\theta_i \,|\, \boldsymbol{\alpha}) - b(2\theta_i \,|\, \boldsymbol{\gamma})\}^2$.

It is essential to accurately describe the background intensity in a Rietveld refinement. GSAS-II provides several options for fitting the background of powder diffraction data including linear interpolation, cosine Fourier series, or Chebyschev polynomials. We limit our discussion to the Chebyschev polynomial approach. In this case, the background intensity is computed as

$$b(2\theta_i \,|\, \boldsymbol{\gamma}) = \sum_{l=1}^{B} \gamma_l \, \nu_{l-1}(2\theta_i), \tag{4.5}$$

where B is the total number of polynomials in the model, the ν_l are known functions of the scattering angles, and the γ_l are unknown coefficients. The ν_l are Chebyschev polynomials of the first kind, as detailed in the *GSAS Technical Manual*, and the values of γ_l are determined by least squares. In general, it is best to use the smallest number of background terms that will give a satisfactory fit, because too many terms will affect the quality of other parts of the least squares model (Larson & Von Dreele, 2004).

4.2.1 Nonlinear Least Squares Algorithms

There are several methods available for nonlinear least squares fitting, but most software packages for Rietveld refinement use the Gauss-Newton algorithm to find parameters that minimize (4.4) (Izumi, 1993). The algorithm determines the change in the parameter values, $\Delta\vartheta$, using a set of

normal equations

$$\mathbf{C}\Delta\boldsymbol{\vartheta} = \mathbf{N}_c, \tag{4.6}$$

where $\mathbf{C}$ is a $(p \times p)$ coefficient matrix, and both $\Delta\boldsymbol{\vartheta}$ and $\mathbf{N}_c$ are vectors of length p. $\mathbf{C}$ and $\mathbf{N}_c$ are functions of the partial derivatives, with respect to the adjustable parameters, of the profile model and minimization criterion, respectively. These quantities are known, and are defined by equations (17) and (18) in Izumi (1989). Most structure refinement programs evaluate $\Delta\boldsymbol{\vartheta}$ as $\mathbf{C}^{-1}\mathbf{N}_c$, though it should be noted that the required matrix inversion can be computationally intensive for large p. At time t, the subsequent set of parameter values is obtained by

$$\boldsymbol{\vartheta}^{(t+1)} = \boldsymbol{\vartheta}^{(t)} + d\Delta\boldsymbol{\vartheta}, \tag{4.7}$$

with $d = 2^{-m}$ for $m = 0, 1, 2, 3, 4$. d is the variable damping factor, which is initially set at 1, i.e., $m = 0$. If (4.4) does not decrease with $\boldsymbol{\vartheta}^{(t+1)}$, then d is decreased and $\boldsymbol{\vartheta}^{(t+1)}$ is recalculated from (4.7). Unfortunately, the range of convergence for this method is narrow, and refinements often converge to a local minimum (Howard & Preston, 1989).

Though the Gauss-Newton algorithm is the standard implementation, there are several other optimization algorithms that may be more suitable in specific settings. When the coefficient matrix, $\mathbf{C}$, is not positive definite, the modified Marquardt method is used. This algorithm is effective for highly nonlinear functions, or when the starting parameter values are far from their true values (Izumi, 1989). The conjugate direction method avoids the computational cost of Newton methods while accelerating the convergence rate as compared to steepest descent methods, and is well-suited for solving problems with highly correlated parameters (Powell, 1964).

Regardless of the optimization algorithm used, the Rietveld method has a relatively small radius of convergence of the refined parameters, which makes good starting values of particular importance (Dinnebier & Müller, 2013). One approach for obtaining good initial values is combining Rietveld refinement with a global optimization algorithm such as grid search, simulated annealing, or genetic algorithms. Global optimization methods are often stochastic in nature, and these approaches are sometimes referred to as *global Rietveld refinements* in the context of crystallography (Shankland, 2004). These methods are able to escape from local minima that can lead to incorrect structures, and in principle, can locate the global minimum from any starting point (Caliandro et al., 2008).

4.2.2 Uncertainty Quantification

Uncertainty about parameter estimates in Rietveld refinement is most often described using an estimated standard deviation (e.s.d.). For the j^{th} parameter, it is calculated as

$$\sigma_j = \left[C_{jj}^{-1} \frac{\sum_{i=1}^{n} w_i (Y_i - Y_i^c)^2}{n - p + \eta_c} \right]^{1/2} \qquad \text{for } j = 1, \ldots, p, \tag{4.8}$$

where C_{jj}^{-1} is the j^{th} diagonal element of the inverse coefficient matrix, $\mathbf{C}^{-1}$, and η_c is the number of constraints applied (Young, 1993). While e.s.d.'s reflect precision of refined parameters, they provide little information about the accuracy of the parameters (Scott, 1983). The e.s.d. is the minimum possible probable error, calculated with the assumption that counting statistics are the only source of error (McCusker et al., 1999). Because it is difficult to collect diffraction data that is entirely free from systematic error, e.s.d.'s do not provide a true impression of the accuracy of the structures (Cheetham, 2002). The actual probable errors will be larger than the calculated e.s.d.'s. For example, it has been suggested that e.s.d.'s may be underestimated by at least a factor of two (Scott, 1983; Sakata & Cooper, 1979). It has even been found that analysis of a diffraction profile with the same structural model can produce different e.s.d.'s when carried out in different Rietveld refinement programs (McCusker et al., 1999). The ambiguity in the significance of the e.s.d. in Rietveld refinements illustrates the need for a better method of uncertainty quantification.

Refinement strategies are also shown to play a large role in the outcome of a Rietveld refinement. In the 1990s, the Commission on Powder Diffraction of the International Union of Crystallography performed studies that compared currently used Rietveld refinement software, instruments, and data collection methods, and examined the effects of various refinement protocols on the accuracy and precision of the parameters obtained through a Rietveld refinement round robin (Hill, 1992; Hill & Cranswick, 1994). Participants were provided with the same neutron and X-ray data sets, yet obtained large variations in values of goodness of fit criteria and in various parameters, such as the atomic positions (Hill, 1992). This study also demonstrated that Rietveld analyses often lead to e.s.d.'s that are smaller than the observed variations in parameter values between repeated refinements. This highlights the possibility of high precision, but low accuracy, of the Rietveld method. Participants were also provided with a sample of m-ZrO_2 and collected X-ray and neutron data sets from different instruments, which demonstrated the effect of the instrument and data collection methodology on the refinement results (Hill & Cranswick, 1994). The results from these round robin tests led to a series of recommendations for data collection and refinement strategies, such as the guidelines reported by McCusker et al. (1999).

In addition, the order in which parameters are refined has a large effect on the accuracy and

reliability of a Rietveld refinement. A one-by-one turn-on sequence is used in the Rietveld method to facilitate optimization. Rietveld software packages can select which parameters or groups of parameters will be refined in each run. A suggested parameter turn-on sequence is provided by Young (1993). Generally, uncorrelated parameters should be turned on first. Turning on non-linear, unstable parameters too early in the sequence will lead to a poor refinement. A systematic turn-on sequence also provides an effective tool for identifying parameters that cause problems with the refinement, such as the occurrence of divergence between the observed and calculated profiles.

4.2.3 Criteria of Fit

Within structure refinement, there are a variety of criteria of fit used to judge whether the "best fit" has been obtained. R-factors are one such set of criteria used to indicate the quality of a least squares refinement. The weighted profile R-factor (R_{wp}),

$$R_{wp} = \sqrt{\frac{\text{WSSE}}{\sum_{i=1}^{n} w_i Y_i^2}}, \tag{4.9}$$

includes the residual being minimized in the numerator, making it the most meaningful R-factor because it reflects the refinement progress (Young, 1993). The expected R-factor (R_{exp}) is given by

$$R_{exp} = \sqrt{\frac{n-p}{\sum_{i=1}^{n} w_i Y_i^2}}. \tag{4.10}$$

This is considered the "best possible R_{wp}" (Toby, 2006). Other R-factors include the R-pattern factor (R_p) and R-Bragg factor (R_B).

Another criterion of fit is the goodness of fit χ^2, which is defined as $\chi^2 = (R_{wp}/R_{exp})^2$. A χ^2 value that is much larger than 1 may indicate the use of a poor model, or that a false minimum has been reached. Alternatively, $\chi^2 < 1$ suggests that the quality of the data does not justify the number of parameters in the model (Prince, 1993). Note that χ^2 should never be less than 1.

There is no established rule of thumb for determining values of R_{wp} and χ^2 that indicate that a refinement is reliable. These criteria are measures of not only the fit of the diffraction peak profiles, but also the fit of the background. If the background intensity is large, even an invalid structural model may yield relatively small values of χ^2 and R_{wp} (McCusker et al., 1999). On the other end of the spectrum, high values of χ^2 and R_{wp} may be obtained when data is collected to a very high precision because the influence of minor imperfections in the fit is increased, even though the obtained model will be improved (Toby, 2006).

In addition to evaluating these numerical criteria of fit, it is critical to consider whether or not the

crystal structure model itself is reasonable, because these criteria do not always accurately represent the fit quality. Graphical criteria of fit provide different information about the quality of the model than numerical criteria. It is important to visually evaluate the observed and calculated profiles, and the difference plots obtained from a refinement. These can provide clear indication of the cause of high R_{wp} values. For instance, a secondary phase is easy to observe in a difference plot. It must be noted that even when a refinement result yields low R-factors, a smooth difference curve, and a chemically sensible structure, it is possible that the refined crystal structure may still be incorrect. Buchsbaum and Schmidt demonstrate that the Rietveld refinement of γ-quinacridone using the crystal structure of β-quinacridone yields an apparent reasonable fit (Buchsbaum & Schmidt, 2007). This example shows that a "successful" Rietveld refinement may still be wrong.

4.3 Bayesian Approaches

As an alternative to the optimization approach discussed in the previous section, Bayesian statistical methods are being applied to crystallographic refinement with increasing frequency (Bergmann & Monecke, 2011; Gagin & Levin, 2015; Fancher et al., 2016; Lesniewski et al., 2016). Bayesian statistics expresses uncertainty about a parameter value using probability distributions (Gelman et al., 2014). The prior distribution quantifies beliefs about the parameter before having access to data, allowing expert knowledge to be incorporated into an analysis. In particular, we incorporate expert knowledge by setting upper and lower bounds for the value of a particular material or instrument parameter, as well as a probability distribution within those bounds. In the present work, unless otherwise specified, each element of $\boldsymbol{\alpha}$ is given a uniform prior distribution, $\alpha_j \sim \mathrm{U}(l_j, u_j)$. Table 4.1 lists the prior bounds for each parameter in the examples to follow.

Table 4.1 Prior bounds for the material and instrument parameters

Parameter	Lower Bound	Upper Bound
Microstrain (%*100)	0	1200
Crystallite Size (μm)	0	1.5
Wavelength (Å)	1.53	1.55
Axial divergence	0	0.5
U	200	300
V	-400	-250
W	125	225
2θ Offset (°)	-0.10	0.10
Scale	1000	2000

Because Bayesian methods for analysis of diffraction data have not been widely employed

previously, we present our implementation and evaluation of these techniques in the present section. For the purposes of the examples to follow, the Bayesian model has not yet been fully specified. The overall form of the model is given by (4.1), GSAS-II is responsible for modeling the intensity due to the material, and the prior distributions for the material and instrument parameters have been defined in Table 4.1. However, it remains to specify the distribution of the errors, the model for the background intensity, and the priors for its associated parameters. The background scattering is modeled using a basis function expansion,

$$b(2\theta_i \,|\, \boldsymbol{\gamma}) = \mathbf{B}(2\theta_i)\boldsymbol{\gamma} = \sum_{l=1}^{B} \gamma_l \mathrm{B}_l(2\theta_i), \tag{4.11}$$

where $\mathbf{B}(2\theta_i)$ is a B-vector of cubic B-splines[2] evaluated at $2\theta_i$, and $\boldsymbol{\gamma} \in \mathbb{R}^B$ are the associated coefficients. The coefficients have the standard prior distribution, $\boldsymbol{\gamma} \sim \mathrm{N}_B(\mathbf{0}, \tau_b^{-1}\boldsymbol{I})$, where $\tau_b \sim \mathrm{Gamma}(0.1, 0.1)$ controls the precision of the distribution. Finally, because powder diffraction data exhibits heteroscedastic errors, as described in the introduction, we assume that

$$e_i \sim \mathrm{N}(0, \tau_e^{-1}[1 + \tilde{Y}_i]), \tag{4.12}$$

where $\tau_e \sim \mathrm{Gamma}(0.1, 0.1)$ controls the precision and $\tilde{Y}_i$ are the smoothed values of Y_i smoothed on $2\theta_i$ using a local regression (or loess) function (Hastie et al., 2009).

Obtaining samples from the posterior distribution, $\pi(\cdot)$, for the parameters in the model detailed above can be obtained in many ways. Because we have specified conjugate prior distributions, posterior samples for $\boldsymbol{\gamma}$ and τ_e are easily obtained using Gibbs sampling (Geman & Geman, 1984; Gelfand & Smith, 1990). In the sections to follow, we will focus on the more complicated process of obtaining posterior samples for the adjustable parameters, $\boldsymbol{\vartheta}$. We will detail several Markov chain Monte Carlo (MCMC) algorithms as well as an Approximate Bayesian Computing (ABC) approach. Within each section, we also present example results obtained from using the current algorithm to fit the neutron data introduced in Section 4.1. To facilitate comparison across algorithms, we will run each sampler for 100,000 total iterations with the first 10,000 iterations discarded as burn-in. The initial values are provided by a preliminary Rietveld analysis. Within each section, we will present trace plots for the crystallite size, wavelength, Caglioti parameter V, and scale parameters to diagnose convergence. These parameters are chosen because crystallite size and V are difficult to estimate and are highly correlated with other parameters, the wavelength is crucial to correctly modeling peak locations, and scale is relatively well-behaved. In addition to inspecting trace plots, we will diagnose convergence using P-values from the Geweke diagnostic (Geweke, 1992). Not every model will exhibit ideal convergence for all parameters during the runtime that we have set, but

[2] One could use another set of basis functions of their choice, e.g., Chebyschev polynomials.

pre-determining the chain length allows us to compare the algorithms fairly.

4.3.1 Metropolis Algorithm

The Metropolis algorithm (Metropolis et al., 1953) is the algorithm most commonly used when the posterior distribution of a target parameter does not have a known distributional form. In its simplest form, the Metropolis algorithm is a univariate method, meaning that each parameter, $\alpha_j^{(t)}$, is updated in sequence during each iteration. In the next section, we will discuss a more robust version of the algorithm that updates the parameters jointly, but for the moment the algorithm proceeds by sampling a candidate parameter value, α_j^*, from a neighborhood around the current value. The size and shape of this neighborhood are defined by the proposal distribution, which is generally taken to be a Normal distribution centered at $\alpha_j^{(t)}$ and with variance $\sigma_{M,j}^2$. These variances are tuning parameters, discussed in more detail below.

Once a candidate value has been proposed, the algorithm chooses either to accept the candidate and set $\alpha_j^{(t+1)} = \alpha_j^*$ or to reject the candidate and leave $\alpha_j^{(t+1)} = \alpha_j^{(t)}$ unchanged. In particular, the algorithm accepts the candidate value with probability

$$\mathrm{P}(\alpha_j^*, \alpha_j^{(t)}) = \min\left(1, \frac{\pi\left(\alpha_j^* \mid \cdots\right)}{\pi\left(\alpha_j^{(t)} \mid \cdots\right)}\right), \tag{4.13}$$

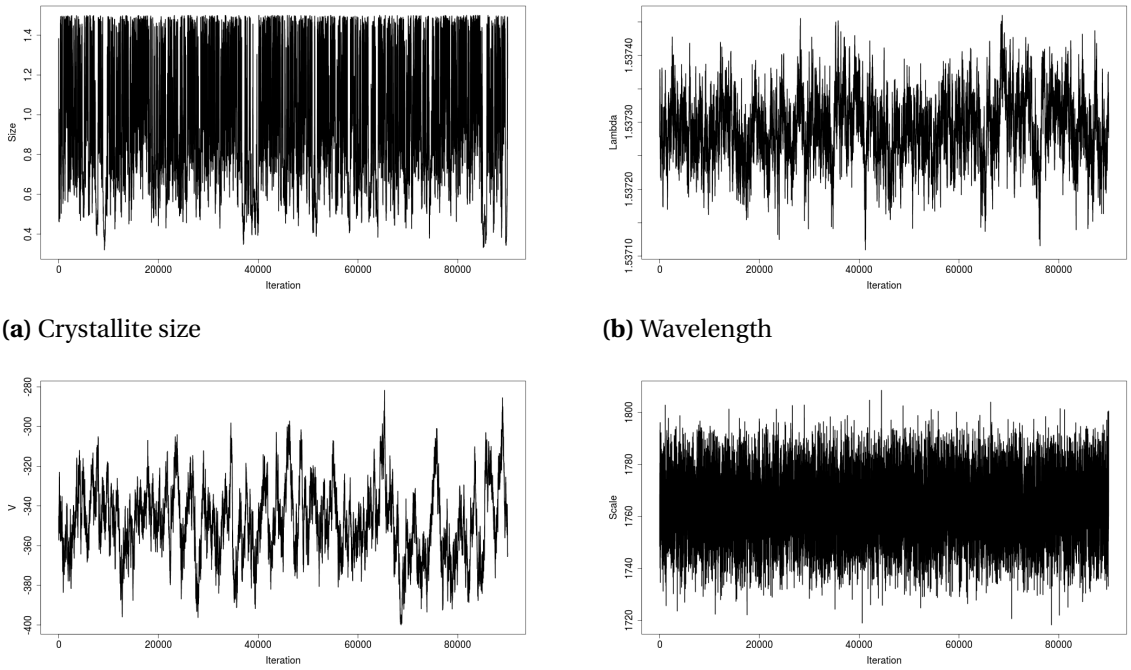
assuming that the proposal distribution is symmetric, as we have. By defining the acceptance probability in this way, the algorithm always accepts a candidate value if it has a higher associated posterior probability, and it will occasionally accept a candidate value despite its lower associated posterior probability. If a proposed candidate results in a much smaller posterior probability as compared to the current value, the acceptance probability will be close to 0, but if the reduction is relatively small, the acceptance probability will be close to 1.

A driver of the Metropolis algorithm's continued popularity is its simple implementation, but even this algorithm requires some tuning by the user. For example, the acceptance rate for each parameter, calculated as the proportion of candidate values that are accepted in a given number of iterations, should be monitored to ensure that it falls in a reasonable range, generally $[0.3, 0.5]$ (Gelman et al., 2014). If the acceptance rate for parameter j is too high, then convergence will be slow and the associated proposal distribution variance, $\sigma_{M,j}^2$, should be increased. Conversely, an acceptance rate that is too low is not exploring the parameter space, and the associated proposal distribution variance should be decreased.

Despite its wide usage, the algorithm does not work well in all settings. Because the parameters are updated individually, the algorithm can struggle when strong relationships between parameters exist. When parameters are correlated, the algorithm's one-at-a-time updating scheme can lead to slow convergence. This can be partly ameliorated by choosing a proposal distribution that closely

resembles the expected posterior distribution, when possible. In addition, the Metropolis algorithm is prone to becoming stuck in local optima. There are many extensions of this basic algorithm that address these concerns (and more), a few of which we will discuss later.

We fit the data in Figure 4.1 using a one-at-a-time Metropolis sampler, which took 627.52 minutes of computation time to complete. Trace plots for the selected parameters are shown in Figure 4.2. Upon inspection of the trace plots, there is no apparent discernible upward or downward trajectory to



(a) Crystallite size

(b) Wavelength

(c) V

(d) Scale

Figure 4.2 Trace plots for selected parameters from a one-at-a-time Metropolis sampler.

the sampled parameter values, and thus it seems reasonable to believe that the chain has converged. P-values from Geweke's diagnostic offer additional support, as no parameter has an associated P-value below the standard 0.05 threshold. However, there is a clear distinction between the trace plot for scale, which is well-converged, and those of the wavelength and V parameters. The trace plot for wavelength seems to exhibit a sinusoidal pattern to the samples, and the trace plot for V is more disjointed than is ideal. Though the trace plot for crystallite size is dense and does not exhibit a strong pattern, the plot seems to indicate that the posterior samples are being limited by the upper bound of the prior distribution.

Figure 4.3 displays a scatter plot of the posterior sampled values for two of the Caglioti parameters,

U and V, from the one-at-a-time Metropolis sampler with the prior bounds for both parameters marked by dashed lines. The one-at-a-time sampler will select all points in the box created by the

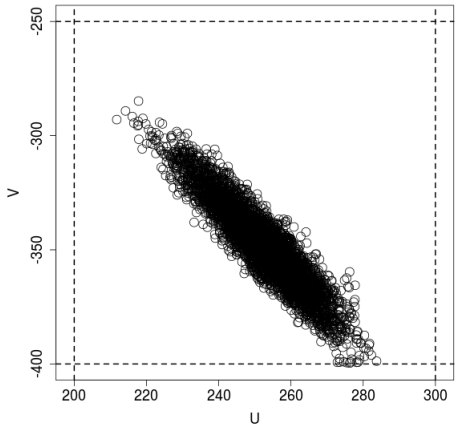


Figure 4.3 Scatter plot of posterior sampled values for the Caglioti parameters U and V using the one-at-a-time Metropolis sampler. The prior bounds for the parameters are noted by dashed lines.

prior bounds with equal probability, despite the fact that there is a clear region in which sampled values are being accepted. Accounting for the correlation between the parameters, as we do in the next section, allows the algorithm to propose candidates that are more likely to be accepted, because it understands the geometry underlying the relationships between the parameters.

4.3.2 Joint Metropolis

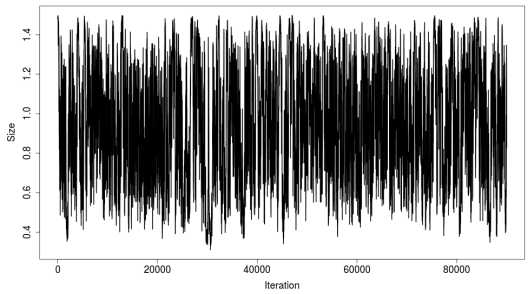
When parameters are highly correlated, the fact that the standard Metropolis algorithm updates parameters separately becomes problematic. Specifically, the sampler will converge to the target distribution very slowly because the algorithm is proposing candidates inefficiently. As an alternative, the parameters can be updated jointly. The difference between this joint Metropolis algorithm and the univariate Metropolis algorithm is that the candidate parameter *vector*, $\boldsymbol{\alpha}^*$, is drawn from a neighborhood around the current parameter *vector*, $\boldsymbol{\alpha}^{(t)}$. Akin to before, this neighborhood is generally taken to be a multivariate Normal distribution with covariance matrix $\sigma_M^2\mathbf{V}$. Ideally, $\mathbf{V}$ reflects the relationship between the parameters. There are several approaches to obtain such an estimate; e.g., a least squares estimate, a preliminary run of another MCMC algorithm, or expert consultation. As in the case of standard Metropolis, the scaling term, σ_M^2, is tuned to maintain a reasonable acceptance rate. Given a proposed candidate vector, the acceptance probability is

computed as

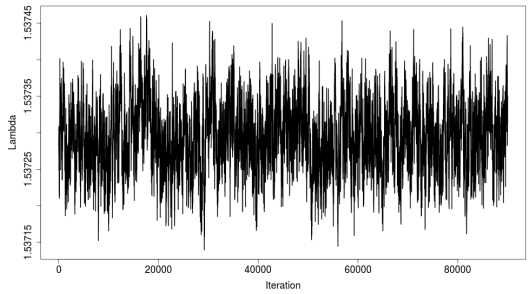
$$P(\boldsymbol{\alpha}^*, \boldsymbol{\alpha}^{(t)}) = \min\left(1, \frac{\pi(\boldsymbol{\alpha}^* | \cdots)}{\pi(\boldsymbol{\alpha}^{(t)} | \cdots)}\right), \tag{4.14}$$

which is nearly identical to (4.13), but the parameters are evaluated together rather than individually. This sampling scheme has been successfully employed to refine adjustable parameters for X-ray diffraction data (Fancher et al., 2016).

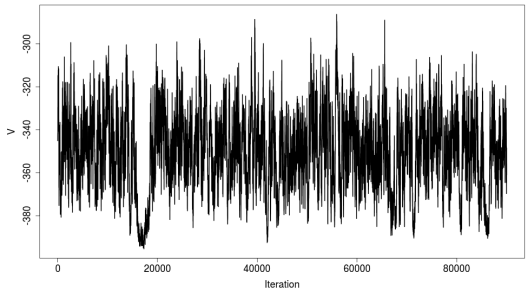
We again fit the neutron diffraction data, this time using a joint Metropolis sampler. The proposal covariance matrix, $\mathbf{V}$, is obtained from a preliminary run and the scale is initialized to $\sigma_M^2 = 0.01$. GSAS-II provides a covariance matrix for the Rietveld estimates that could also be used, but we found that using the posterior covariance from a preliminary run provided a better proposal distribution. The joint Metropolis algorithm took 178.82 minutes to complete. Scanning the selected trace plots



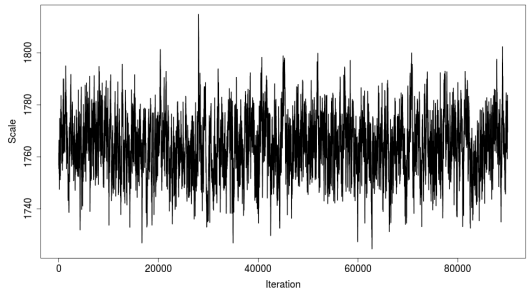
(a) Crystallite size



(b) Wavelength



(c) V



(d) Scale

Figure 4.4 Trace plots for selected parameters using a joint Metropolis sampler.

in Figure 4.4, it is immediately evident from the plots for wavelength and V that perhaps the burn-in period for this algorithm should be extended to 30,000 iterations from the original 10,000. The P-values from Geweke's diagnostic are above the 0.05 threshold, but in reporting results in Section 4.4, we will consider the first 30,000 iterations as burn-in. In comparison to the previous section,

the sinusoidal pattern is less evident in the trace plot for the wavelength, and the trace plot for V is more dense. The trace plot for scale, while not as well-converged as in the previous case, is still well-behaved. It would seem that accounting for the correlation between the parameters results in an overall improvement in the convergence properties of the sampler.

4.3.3 Delayed Rejection Adaptive Metropolis

Delayed Rejection Adaptive Metropolis (DRAM) (Haario et al., 2006) combines two powerful modifications to the joint Metropolis algorithm. Delayed rejection provides a technique for a second candidate to be proposed in the case that the first candidate is rejected in the initial Metropolis algorithm (Tierney & Mira, 1999; Green & Mira, 2001). This process can help the algorithm avoid getting stuck at a particular solution by increasing the number of accepted candidates. Adaptive Metropolis allows the proposal distribution to be tuned as the algorithm proceeds based on the covariance of the accepted candidates (Haario et al., 1999; Haario et al., 2001). In theory, this adjustment should lead the algorithm to propose candidates more efficiently as the proposal distribution adapts to reflect successful candidates.

At the start of a two-stage DRAM algorithm, prior knowledge is used to specify an initial covariance matrix for the proposal distribution, $\mathbf{V}^{(0)}$. Then, at fixed intervals during the sampler, the proposal distribution covariance is updated using an adaptation step

$$\mathbf{V}^{(t)} = s_d \operatorname{Cov}\{\boldsymbol{\alpha}^{(0)}, \boldsymbol{\alpha}^{(1)}, \ldots, \boldsymbol{\alpha}^{(t-1)}\} + s_d\, c\, \boldsymbol{I}_q, \tag{4.15}$$

where $s_d = {}^{2.4^2}/_{q}$ depends only on the dimension of the parameter space, $c > 0$ is a small constant, and $\boldsymbol{I}_q$ is the $(q \times q)$ identity matrix. At the first stage of iteration t, DRAM proceeds just as in the joint Metropolis algorithm, where the candidate, $\boldsymbol{\alpha}_1^*$, is drawn from a multivariate Normal distribution centered at the current value, $\boldsymbol{\alpha}^{(t)}$, with covariance matrix $\mathbf{V}^{(t)}$. The candidate is evaluated as in (4.14). If the candidate is accepted, then $\boldsymbol{\alpha}^{(t+1)} = \boldsymbol{\alpha}_1^*$, and the sampler proceeds to the next iteration. However, if the candidate is rejected, then a second candidate is proposed. The second stage candidate, $\boldsymbol{\alpha}_2^*$, is drawn from its own proposal distribution. In more complex examples, this second stage proposal may depend on the previously rejected first stage candidate. In the simplest case, the second stage candidate is again drawn from a multivariate Normal distribution centered at the current value, but the second stage covariance is $\mathbf{V}_2^{(t)} = \kappa \mathbf{V}^{(t)}$, where $0 < \kappa < 1$ is a scaling parameter (Green & Mira, 2001). The second stage acceptance probability is

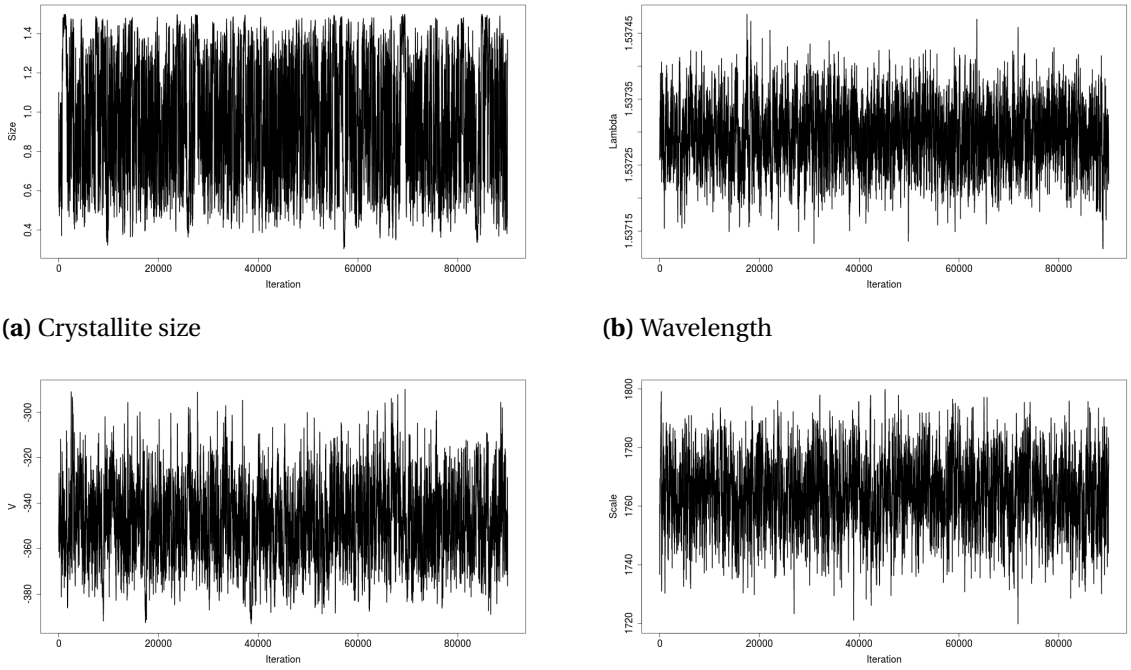
$$\mathrm{P}(\boldsymbol{\alpha}_2^*, \boldsymbol{\alpha}_1^*, \boldsymbol{\alpha}^{(t)}) = \min\left(1, \frac{\pi\{\boldsymbol{\alpha}_2^*\}\phi(\boldsymbol{\alpha}_1^*; \boldsymbol{\alpha}_2^*, \mathbf{V}^{(t)})[1-\omega\{\boldsymbol{\alpha}_2^*, \boldsymbol{\alpha}_1^*\}]}{\pi\{\boldsymbol{\alpha}^{(t)}\}\,\phi(\boldsymbol{\alpha}_1^*; \boldsymbol{\alpha}^{(t)}, \mathbf{V}^{(t)})[1-\omega\{\boldsymbol{\alpha}^{(t)}, \boldsymbol{\alpha}_1^*\}]}\right), \tag{4.16}$$

where $\phi(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the multivariate Normal probability density function with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ and $\omega\{\boldsymbol{a}, \boldsymbol{b}\} = \min\{1, {}^{\pi(\boldsymbol{a})}/_{\pi(\boldsymbol{b})}\}$. By shrinking the neighborhood from which a second stage

candidate is proposed, the algorithm can balance exploring the parameter space in the first stage with more conservative adjustments in the second. Here, we have outlined a two-stage algorithm, but the method is applicable to an arbitrary number of stages.

DRAM has two primary tuning parameters, the scaling parameter (or shrinkage factor), κ, and the number of iterations between adaptation steps for the covariance, N_0. When the scaling parameter is small, the second stage proposal distribution will be a very small neighborhood around the current value, and the acceptance rate will be high. As κ approaches 1, the second stage proposal distribution will be similar to the first stage's proposal distribution, and the acceptance rate will be lower. The length of the adaptation interval, $N_0 > 0$, can be relatively freely chosen. When the interval is long, the effects of the adaptation will be felt more slowly, but convergence of the algorithm is improved if the adaptation is not done at each iteration (Haario et al., 2001; Haario et al., 2006).

In the implementation for this example, we set $\kappa = 0.1$, $c = 0.0001$, $N_0 = 500$ iterations, and $\mathbf{V}^{(0)} = 0.05\boldsymbol{I}_q$. The DRAM sampler took 199.27 minutes to run. The trace plots in Figure 4.5 seem to indicate the the algorithm has converged, and P-values from Geweke's diagnostic are all above the 0.05 threshold, providing additional evidence for convergence. Using the DRAM algorithm, none of



(a) Crystallite size

(b) Wavelength

(c) V

(d) Scale

Figure 4.5 Trace plots for selected parameters using the DRAM sampler.

the parameters exhibit any clear patterns, either in trajectory or in the sampled values themselves; e.g., the sinusoidal pattern evident in Figure 4.2(b) has disappeared.

4.3.3.1 Hamiltonian Monte Carlo

As is the case with GSAS-II, commercial crystallography packages may provide gradient information for the parameters of interest. In that case, an algorithm called Hamiltonian (or *Hybrid*) Monte Carlo (HMC) is an attractive option (Duane et al., 1987; Neal, 2011). By taking advantage of the additional information provided by the gradient, Hamiltonian Monte Carlo can offer improved parameter space exploration and convergence speed as compared to a joint Metropolis algorithm.

As indicated by its name, Hamiltonian Monte Carlo borrows its motivation from physics. Imagine a surface with varying heights and a frictionless puck that slides over the surface. At any time, the dynamics of the puck are described by both its current position and its momentum. If the puck is on a flat portion of the surface, it moves at a constant speed proportional to its momentum. If the puck encounters an uphill climb in the surface, its momentum will allow it to climb the surface for some time before beginning to slide back down the slope.

In the context of MCMC, the puck's current position is the current value of parameter, $\boldsymbol{\alpha}^{(t)}$, and its potential energy is the negative log of the posterior distribution for the current parameters of interest, $U(\boldsymbol{\alpha}^{(t)}) = -\ell_\pi(\boldsymbol{\alpha}^{(t)} | \cdots)$. Now, parameter values with higher associated posterior probabilities will have *lower* associated potential energy, and thus the system still prefers those values. "Momentum" variables, $\mathbf{v} = (v_1, \ldots, v_q)$, are added to the system artificially. The momentum variables are drawn from a mean-zero multivariate Normal distribution with covariance matrix $\mathbf{M}$, where $\mathbf{M}$ is usually taken to be an identity matrix. The addition of these momentum variables can help the algorithm escape local optima and can encourage exploration of the parameter space.

Hamilton's equations are a set of differential equations that determine how position and momentum change over time. The Hamiltonian function is akin to a total energy function and is defined as

$$H(\boldsymbol{\alpha}, \mathbf{v}) = U(\boldsymbol{\alpha}) + K(\mathbf{v}), \tag{4.17}$$

where the kinetic energy, $K(\mathbf{v})$, is usually defined as $K(\mathbf{v}) = \frac{1}{2}\mathbf{v}'\mathbf{M}^{-1}\mathbf{v}$. For computational purposes, the solution to Hamilton's equations must be approximated. There are several approaches to doing so, but a common technique is leapfrog integration. During a leapfrog step, each parameter and

momentum variable pair, $\{(\alpha_j, v_j)\}_{j=1}^{q}$, are updated as

$$v_j^{(t+\epsilon/2)} = v_j^{(t)} - \frac{\epsilon}{2} \frac{\partial U(\boldsymbol{\alpha}^{(t)})}{\partial \alpha_j} \tag{4.18a}$$

$$\alpha_j^{(t+\epsilon)} = \alpha_j^{(t)} + \epsilon \frac{v_j^{(t+\epsilon/2)}}{m_j} \tag{4.18b}$$

$$v_j^{(t+\epsilon)} = v_j^{(t+\epsilon/2)} - \frac{\epsilon}{2} \frac{\partial U(\boldsymbol{\alpha}^{(t+\epsilon)})}{\partial \alpha_j}, \tag{4.18c}$$

where $\epsilon > 0$ is a small stepsize and m_j is the j^{th} diagonal element of $\mathbf{M}$. This process is repeated to obtain a trajectory of any length.

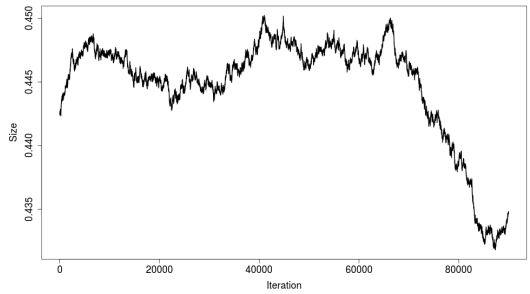
Unlike the algorithms previously discussed, there are *two* steps during each iteration of HMC. Before proposing and evaluating a candidate, a new set of momentum variables, $\mathbf{v}^{(t)}$, are drawn from their associated distribution. Then, a trajectory of length L is simulated using Hamiltonian dynamics, as defined in (4.18), starting from the current state, $(\boldsymbol{\alpha}^{(t)}, \mathbf{v}^{(t)})$. The proposed candidate is the last state in the trajectory, $(\boldsymbol{\alpha}^*, \mathbf{v}^*)$. Rather than the acceptance probability being based solely on the posterior distribution, the probability is adjusted to use the Hamiltonian function

$$\mathrm{P}(\boldsymbol{\alpha}^*, \boldsymbol{\alpha}^{(t)}) = \min\left(1, \exp\left[-H(\boldsymbol{\alpha}^*, \mathbf{v}^*) + H(\boldsymbol{\alpha}^{(t)}, \mathbf{v}^{(t)})\right]\right). \tag{4.19}$$

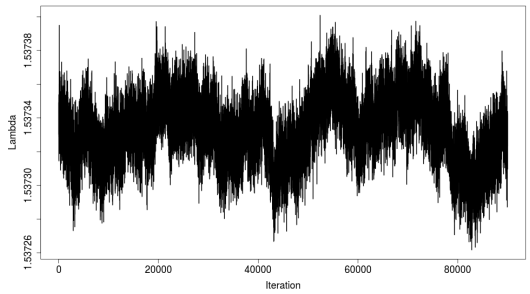
Because of the need to evaluate the gradient several times per iteration and the need to simulate a trajectory to propose a candidate, each individual iteration of HMC will be slow compared to the simpler algorithms. However, the algorithm makes up for this deficiency by using the gradient to propose smarter candidates, leading to faster convergence.

When working well, HMC can be a very effective solution, but tuning is more difficult than in the previously discussed methods. In the simplest case where $\mathbf{M}$ is taken to be the identity matrix, the trajectory length, L, and the step size, ϵ, must be tuned. As usual, the step size is tuned to maintain a reasonable acceptance rate, but choosing an appropriate trajectory length can be more complicated. Neal (2011) provides a detailed discussion of tuning, and is an excellent reference for developing a well-tuned algorithm.

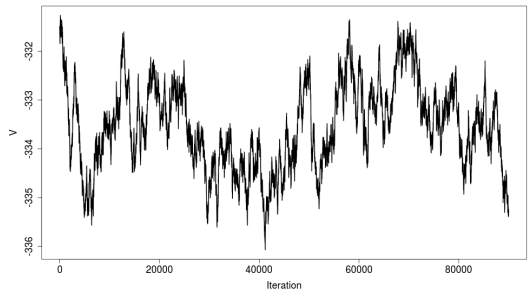
We apply HMC to the neutron diffraction data using a trajectory length of $L = 8$ steps. During each iteration, we draw a step size, $\epsilon^{(t)}$, from a uniform distribution, $\mathrm{U}(8 \times 10^{-7}, 1.2 \times 10^{-6})$, with the center of the distribution chosen to maintain a reasonable acceptance rate. Drawing a new step size at each iteration can help the algorithm avoid becoming stuck with a step size that leads to unstable trajectories. We also account for correlation between parameters by setting $\mathbf{M}$ equal to the inverse of the posterior covariance matrix from a preliminary run. The method took 1,515.45 minutes to complete, and trace plots for the selected parameters are presented in Figure 4.6. Clearly, this



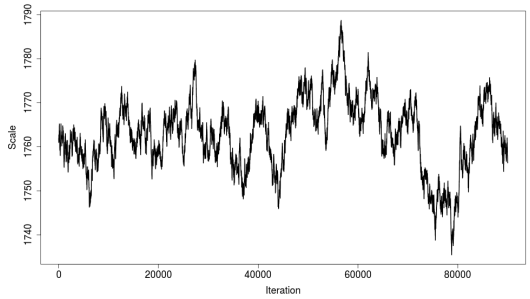
(a) Crystallite size



(b) Wavelength



(c) V



(d) Scale

Figure 4.6 Trace plots for selected parameters using the HMC sampler.

sampler is not properly tuned and has not converged. Indeed, the P-values for Geweke's diagnostic are below the 0.05 threshold for several parameters, providing further evidence that the method has not converged. We tried varying the trajectory length and using different estimates for the posterior covariance matrix, but were unable to find a suitable combination of tuning parameters. An extension of HMC, called the No U-Turns Sampler (NUTS; Hoffman & Gelman, 2014), has been developed to address some of these tuning difficulties. NUTS automatically determines an appropriate trajectory length by implementing a stopping rule that does not allow the algorithm to reverse its progression. The displayed results represent the best combination of settings we were able to find for the base Hamiltonian algorithm. Because this sampler has not converged, we will exempt its results from the criteria of fit comparisons in Section 4.4, though we will still include HMC in our discussion of sample quality.

4.3.4 Approximate Bayesian Computing

When the statistical model can be easily used to generate data, a technique called Approximate Bayesian Computing (ABC) can be advantageous (Voss, 2013). In fact, ABC can even be used in cases where it is difficult or impossible to write down the posterior distribution directly. Rather

than using MCMC to develop a chain of sampled values and diagnosing convergence, ABC draws a set of parameter values from their prior distribution, generates data from this draw, and decides whether or not the generated data is "close to" the observed data. This process is repeated over many replications to obtain a collected sample from the posterior distribution. Therefore, the same ends are achieved as in MCMC.

One of the limitations of ABC is that it can suffer from inefficiency as the number of parameters increases, a problem that is exacerbated by diffuse or uninformative prior distributions. Because we are uninterested in directly modeling the background intensity, we introduce a variation of the model in (4.1) in which we will marginalize over the background parameters. Specifically, the model for the observations becomes

$$Y_i \sim \mathrm{N}\big(f\{2\theta_i \,|\, \boldsymbol{\alpha}\}, \mathbf{B}(2\theta_i)\boldsymbol{\Omega}\mathbf{B}'(2\theta_i) + \tau_e^{-1}[1+\tilde{Y}_i]\big), \tag{4.20}$$

where $\boldsymbol{\Omega}$ is a diagonal matrix with elements representing the prior variance of the basis function coefficients. In the examples to follow, we set $\boldsymbol{\Omega} = 10\boldsymbol{I}_B$. This allows the model to account for the background intensity without needing to draw values for the $\boldsymbol{\gamma}$ parameters explicitly. Thus, we need only draw candidate values for the material and instrument parameters, $\boldsymbol{\alpha}^*$, and the scale of the precision, τ_e^*.

At the start of each replication, the candidate parameters, $\boldsymbol{\alpha}^*$ and τ_e^*, are drawn from their associated prior distributions. Because ABC suffers from inefficiencies when using uninformative priors, such as the uniform distributions from the MCMC framework, we will adjust the prior distributions for this section. From a preliminary Rietveld analysis, we have Rietveld estimates, $\hat{\alpha}_j$, and their associated standard errors, $\hat{\sigma}_j$, for each of the parameters. We adjust the prior distribution so that each of the parameters of interest is independently drawn from $\alpha_j^* \overset{\text{indep}}{\sim} \mathrm{N}(\hat{\alpha}_j, 25\hat{\sigma}_j^2)$, truncated to lie within the intervals specified in Table 4.1. The candidate precision, τ_e^*, is drawn from a standard log-normal distribution. Then, candidate data, $Y_1^*, \ldots, Y_n^*$, is generated from (4.20).

Both the observed data, $Y_1, \ldots, Y_n$, and the candidate data are represented by a statistic (or set of statistics), S, which is a numerical summary of the data. The appropriate summary statistic will be dependent on the specific problem, but should quantify all of the salient information about the data. Ideally, if $S(Y_1^*, \ldots, Y_n^*)$ is exactly equal to $S(Y_1, \ldots, Y_n)$, then the current set of candidate parameter values is accepted as one of the posterior samples. However, in practice, obtaining an exact match is prohibitively unlikely. Thus, a set of candidate parameters is accepted if the difference between the two summary statistics is below some threshold; i.e., the candidate is accepted if $|S(Y_1^*, \ldots, Y_n^*) - S(Y_1, \ldots, Y_n)| < \delta$, for some $\delta > 0$.

The quality of the collected posterior samples is highly dependent on both the summary statistic employed in the analysis and the definition of "close enough", δ, which we will discuss in turn. A general prescription for the summary metric is to use a sufficient statistic, but so long as a metric

captures all of the salient features of the data or problem at hand, this may be relaxed. Tuning the threshold, δ, is of crucial importance in ABC. The goal is to provide samples that are as close to the observed data as possible, but there is an inherent trade off between the quality and quantity of the collected samples. For threshold values that are too close to 0, the algorithm may be unable to collect any samples at all. Thus, careful consideration must be given to the details of the problem.

Because there are software packages available to calculate a diffraction profile from a set of parameter values, generating data for ABC from (4.20) is straightforward, making diffraction analysis a prime candidate for this approach. ABC is relatively simple to implement and provides for good exploration of the parameter space, because the candidate parameter values are independent across replications and thus there is no risk of getting stuck in local optima. However, ABC can also exhibit highly inefficient behavior, taking millions of replications to collect even a small number of posterior samples.

In our crystallography problem, we start by removing the background intensity from each of the observed and candidate data sets. We obtain the residuals from a linear regression analysis of each data set on the basis functions and then smooth each set of residuals using a local regression (loess) smoother. We denote the smoothed residuals from the observed data as $R_1, \ldots, R_n$ and the smoothed residuals from the candidate data as $R_1^*, \ldots, R_n^*$. Then, the statistic used to compare the candidate and observed data is

$$S(Y_1^*, \ldots, Y_n^*, Y_1, \ldots, Y_n) = \sum_{i=1}^{n} \frac{(R_i^* - R_i)^2}{\tilde{Y}_i}, \tag{4.21}$$

which is a weighted sum of squared differences between the smoothed candidate residuals and the smoothed observed residuals. We determine the threshold, δ, by simulation. In particular, we generated 1,000 data sets while holding the parameter values fixed at their Rietveld estimates, calculated the value of S for each using the first generated data set as the observed data, and set δ as the 95^{th} percentile of these collected values resulting in a threshold value of 1482. The intuition behind this simulation procedure is that because the parameter values are held fixed, the 1,000 collected values of S should represent the error due to the random variation. Within the context of ABC, this indicates that if a calculated statistic is below this threshold value, the error could plausibly be due to random variation, and thus the candidate parameter values should be accepted.

Implemented as described above, we ran ABC for a total of 18,000 iterations, resulting in 3,583.63 minutes of computation time. Of the parameter draws, 1,059 (5.88%) were accepted. It should be noted that because the parameter draws are completely independent over iterations, this problem is embarrassingly parallel, and the computation time could be cut drastically if this feature is exploited.

4.4 Discussion

Using the posterior mean estimates from each of the converged Bayesian algorithms from the previous section, we used GSAS-II to generate a calculated diffraction profile. Table 4.2 summarizes the computation time and several criteria of fit for each algorithm, Table 4.3 summarizes the posterior mean estimates and 95% credible intervals for each parameter, and Figure 4.8 presents the posterior density estimates for the Bayesian models along with the Rietveld point estimates as vertical lines with ± 2 e.s.d.'s given as error bars[3]. Recall that because HMC did not converge, its results are omitted from this discussion. From Figure 4.8, we can see that the posterior density estimates from ABC differ widely from those produced by the MCMC methods, which are largely consistent with one another. This indicates that our implementation of ABC may not be collecting sampled values that are truly representative of the posterior distribution. As discussed in Section 4.3.4, the efficacy of ABC is highly dependent on the definition of the summary statistic and on the value of the selected threshold. In this case, simply decreasing the threshold value does not seem to improve the posterior density estimate significantly. Thus, it is possible that the summary statistic defined in (4.21) is not, in fact, close enough to a sufficient statistic, and that we need to update the method. It is worth nothing that had we run ABC on its own, this determination would have been difficult to reach. Because it is possible that our collected samples are not from the posterior distribution, we exclude ABC from Tables 4.2 and 4.3. The Bayesian methods, and particularly the MCMC methods, give very

Table 4.2 Computation time and criteria of fit for the Bayesian algorithms presented in Section 4.3 and the Rietveld method.

Algorithm	Time (mins)	WSSE	R_{wp}(%)	χ^2
One-at-a-time Metropolis	627.52	2,538.74	6.6919	1.00
Joint Metropolis	178.82	2,537.82	6.6907	1.00
DRAM	199.27	2,538.12	6.6911	1.00
Rietveld	–	2,985.18	7.2565	1.17

similar estimates for the parameters, and thus the criteria of fit values are comparable. Because the methods give similar estimates, we display the difference curve and fitted profile from only one of the MCMC methods, DRAM specifically, in Figure 4.7. Based on this graphical criteria of fit, our model produces a good fit when applied to the neutron data from Figure 4.1.

From panel (b) in Figure 4.8, we can see that the posterior density for crystallite size is relatively uniform on the interval $[0.5, 1.5]$. Recall that the prior distribution for this parameter is uniform on the interval $[0, 1.5]$, and thus the data are not providing much additional information to aid in

[3]The e.s.d. for crystallite size in Figure 4.8(b) is small, but present.

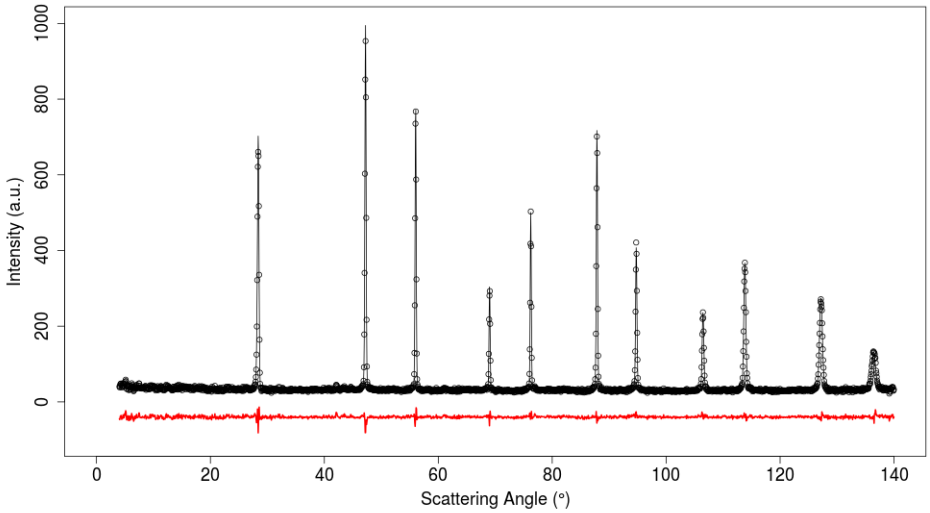


Figure 4.7 Difference curve and fitted profile from the DRAM algorithm applied to the neutron diffraction profile data.

estimating this parameter. This is likely due to the fact that both microstrain and crystallite size affect the same feature of a diffraction profile, namely the peak width. The two parameters are not fully unidentifiable, but there are methods such as Williamson-Hall analysis (Williamson & Hall, 1953) that can help to account for the relationship between them. Figure 4.9 displays the pairwise scatter plots of the posterior samples from DRAM for each pair of parameters within the set of material and instrument parameters. These scatter plots help visualize the relationships between parameters. Plots that do not show a clear relationship indicate relative independence between the associated parameters. We can clearly see the relationship between microstrain and crystallite size that we discussed previously in the first panel of Figure 4.9. There are also strong relationships between the wavelength, axial divergence, and 2θ offset, as well as between the Caglioti parameters.

The various Bayesian algorithms differentiate themselves from one another in terms of computation time, as evident in Table 4.2, and in the quality of the collected samples. Because MCMC methods rely on a Markov chain, the sampled values are not independent. The autocorrelation function, denoted $\rho(h)$, measures the correlation between sampled parameter values that are h iterations apart (sometimes referred to as h-lag autocorrelation). Ideally, the collected samples are completely independent and $\rho(h) \approx 0$ for all h, but this is not realistic in general. A related, one-number summary is the effective sample size (ESS) for each parameter. The effective sample

Table 4.3 Estimated posterior mean and 95% credible interval (CI) for each parameter using each of the Bayesian algorithms from Section 4.3.

	One-at-a-time		Joint		DRAM	
Parameter	Mean	95% CI	Mean	95% CI	Mean	95% CI
Microstrain	786.84	[498.0, 995.7]	787.03	[483.2, 992.3]	787.82	[522.3, 987.3]
Crystallite size	0.96	[0.45, 1.47]	0.95	[0.44, 1.46]	0.95	[0.45, 1.46]
Wavelength	1.54	[1.54, 1.54]	1.54	[1.54, 1.54]	1.54	[1.54, 1.54]
Axial divergence	0.09	[0.09, 0.10]	0.09	[0.09, 0.10]	0.09	[0.09, 0.10]
U	252.27	[229.8, 272.5]	253.11	[233.6, 273.2]	252.59	[233.0, 271.7]
V	-347.97	[-382.3, -310.5]	-349.58	[-383.2, -315.6]	-348.54	[-379.6, -315.5]
W	183.22	[167.2, 198.4]	183.95	[169.4, 198.2]	183.50	[169.0, 197.2]
2θ Offset	0.07	[0.06, 0.07]	0.07	[0.06, 0.07]	0.07	[0.06, 0.07]
Scale	1763.75	[1741.8, 1785.3]	1764.25	[1742.6, 1787.3]	1764.11	[1742.0, 1785.7]

size of $N_{samples}$ collected posterior samples is defined as

$$\text{ESS} = \frac{N_{samples}}{1+2\sum_{h=1}^{\infty}\rho(h)}, \tag{4.22}$$

and it measures the equivalent number of independent samples contained within a correlated set of MCMC samples. Broadly, it quantifies the quality of the information in the MCMC samples, and ESS values close to $N_{samples}$ are ideal. In Table 4.4, we catalog the effective sample size for each parameter using each of the MCMC methods from Section 4.3. Note that ABC produces independent samples, so it is not included in the table. As suggested by the trace plots in the individual algorithm

Table 4.4 Effective sample size (ESS) for each parameter using each of the MCMC algorithms.

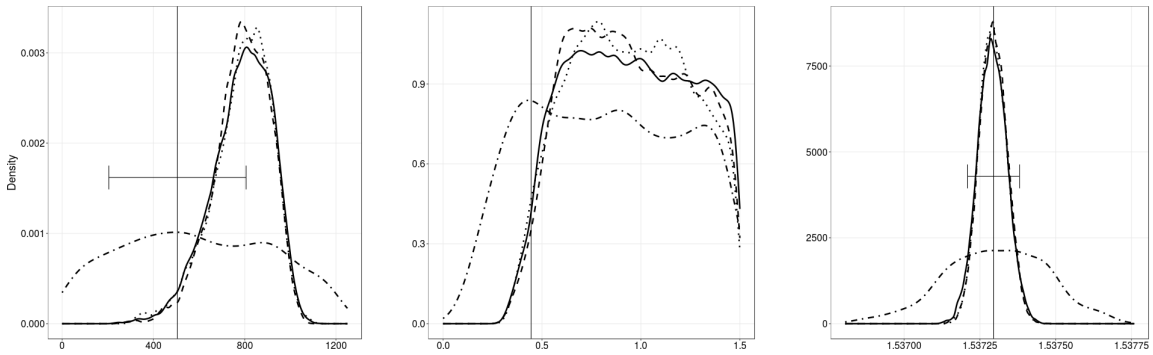
Parameter	One-at-a-time	Joint	HMC	DRAM
Microstrain	486.85	382.81	2.95	810.81
Crystallite size	586.32	399.06	2.92	705.28
Wavelength	398.53	406.45	164.46	1,029.23
Axial divergence	385.48	465.11	562.46	1,082.36
U	147.30	316.56	7.18	837.31
V	104.48	306.43	24.35	823.00
W	101.71	301.62	8.40	846.53
2θ Offset	387.00	379.94	15.32	896.20
Scale	5,388.97	385.41	18.32	929.62
Minimum	101.71	301.62	2.92	705.28

sections, the DRAM algorithm provides the “best” collection of posterior samples, with the joint

Metropolis algorithm slightly outperforming the one-at-a-time Metropolis algorithm. The effective sample sizes for HMC reinforce its lack of convergence, with single digit ESS values for several parameters. It should be noted that our purpose in this chapter is to contrast the efficacy of a variety of Bayesian sampling algorithms. An effective sample size of ≈1,000 is likely too low to obtain parameter estimates with high precision for a particular application when fitting any of these algorithms.

4.5 Summary

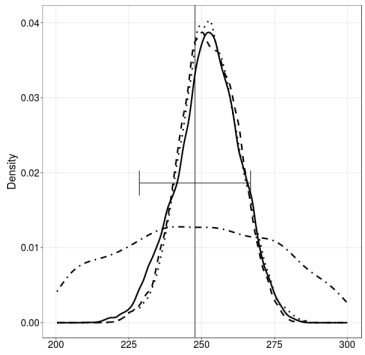
In this chapter, we have provided an introduction to the Rietveld method and provided a discussion of its implementation in Rietveld software packages. The shortcomings of the numerical criteria used to evaluate a Rietveld fit were also summarized, illustrating the need for an alternative method with more quantifiable uncertainty. To that end, we discussed a Bayesian framework for diffraction profile analysis and noted that a Bayesian approach allows for seamless integration of expert knowledge, does not rely on a complex optimization order, and provides fully quantified uncertainty in the form of a full posterior distribution for each parameter of interest. We detailed several Markov chain Monte Carlo sampling methods, as well as an alternative method of gathering posterior samples called Approximate Bayesian Computing, and applied them to the profile fitting problem using neutron diffraction data from a NIST silicon standard reference material. The DRAM sampler outperformed the other MCMC sampling algorithms based on sampling quality, though the methods returned similar fits. The Hamiltonian Monte Carlo sampler did not converge despite several attempts at tuning the algorithm. Approximate Bayesian Computing proved to be a more time-consuming approach, but that issue can be ameliorated by employing parallel computing. In all, we have shown that Bayesian approaches to diffraction profile analysis provide an attractive alternative to traditional least squares approaches like the Rietveld method but that results will vary based on the Bayesian sampling method employed.



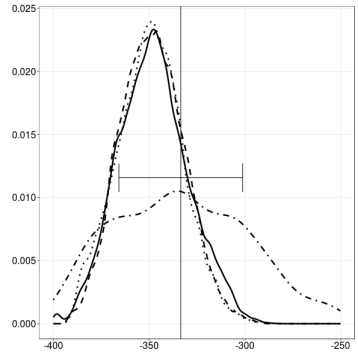
(a) Microstrain

(b) Crystallite Size

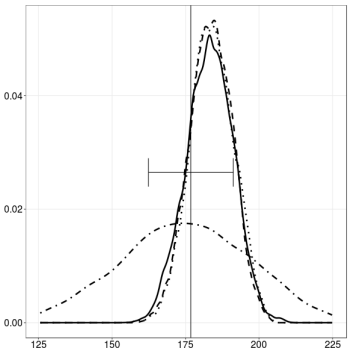
(c) Wavelength



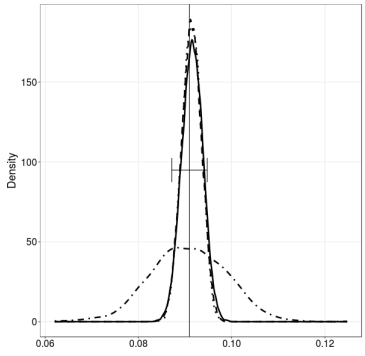
(d) U



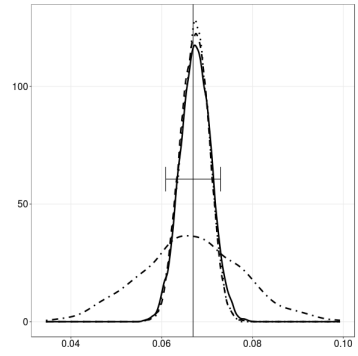
(e) V



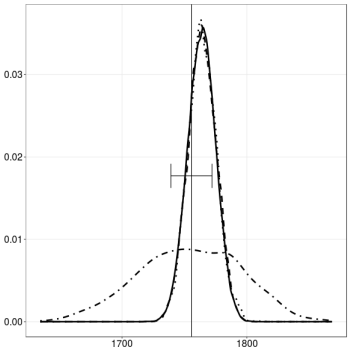
(f) W



(g) Axial Divergence



(h) 2θ Offset



(i) Scale

Figure 4.8 Estimated posterior density for each parameter using each of the algorithms in Section 4.3. One-at-a-time Metropolis is a solid line, Joint Metropolis is a dashed line, DRAM is a dotted line, and ABC is dot-dashed line. The Rietveld estimate ± 2 e.s.d.'s is also given.

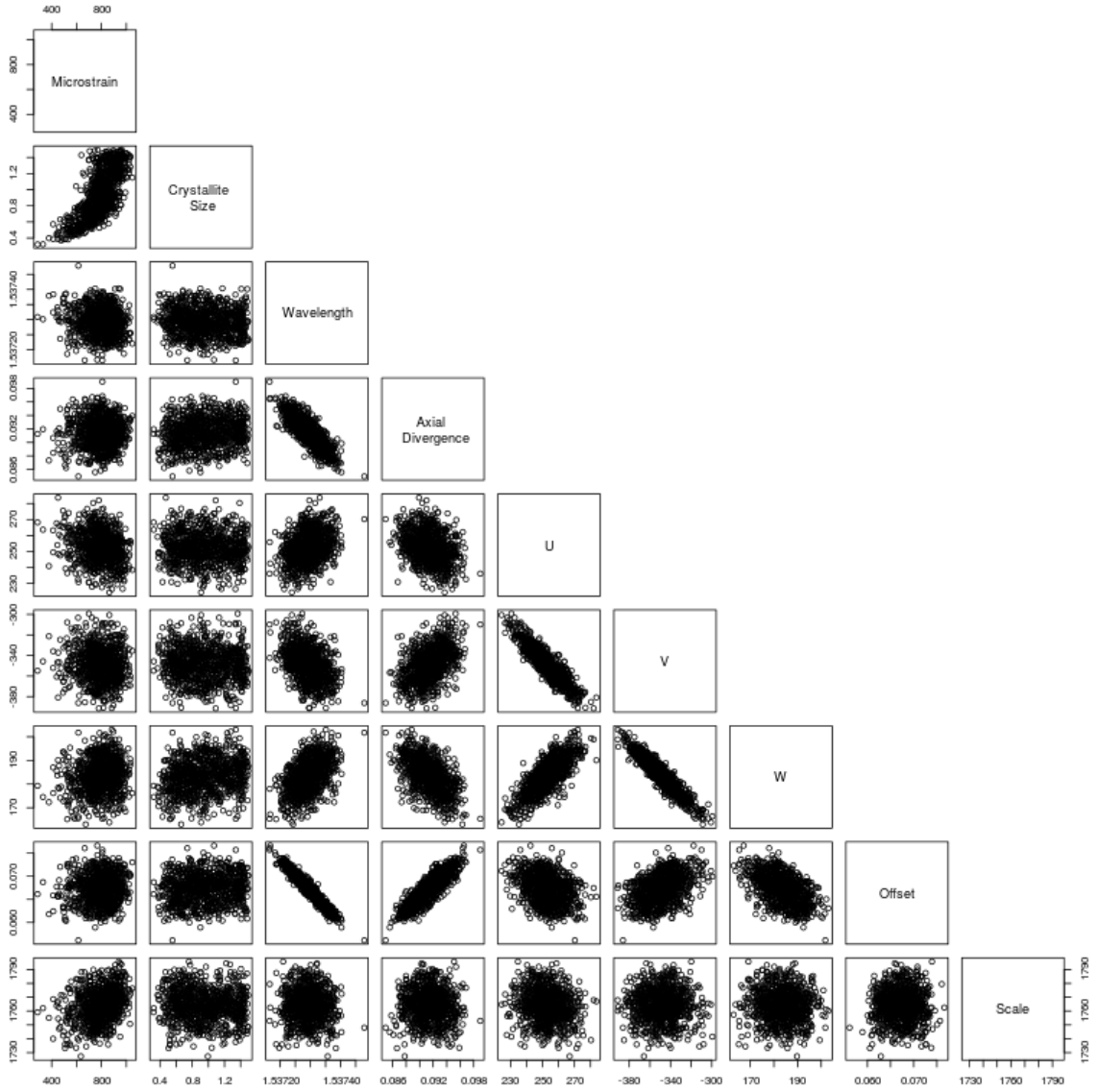


Figure 4.9 Pairwise scatter plots of the posterior samples for each of the instrument and material parameters of interest using the DRAM sampler.

BIBLIOGRAPHY

Aitchison, J (1986). *The Statistical Analysis of Compositional Data*. London: Chapman & Hall.

Anderson, M. J. (2001). "A new method for non-parametric multivariate analysis of variance". *Austral Ecology* **26**.1, pp. 32–46.

Antoniak, C. E. (1974). "Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems". *The Annals of Statistics* **2**.6, pp. 1152–1174.

Banerjee, S. (2005). "On geodetic distance computations in spatial modeling". *Biometrics* **61**.2, pp. 617–625.

Barberán, A. et al. (2015). "The ecology of microscopic life in household dust". *Proceedings of the Royal Society of London B: Biological Sciences* **282**.1814, pp. 212–220.

Barney, B. J. et al. (2015). "Joint Bayesian modeling of binomial and rank data for primate cognition". *Journal of the American Statistical Association* **110**.510, pp. 573–582.

Bergmann, J & Monecke, T (2011). "Bayesian approach to the Rietveld refinement of Poisson-distributed powder diffraction data". *Journal of Applied Crystallography* **44**.1, pp. 13–16.

Bray, J. R. & Curtis, J. T. (1957). "An ordination of the upland forest communities of southern Wisconsin". *Ecological Monographs* **27**.4, pp. 325–349.

Buchsbaum, C & Schmidt, M. U. (2007). "Rietveld refinement of a wrong crystal structure". *Acta Crystallographica Section B: Structural Science* **63**.6, pp. 926–932.

Caliandro, R et al. (2008). "Crystal Structure Determination". *Powder Diffraction: Theory and Practice*. Ed. by Dinnebier, R. E. & Billinge, S. J. L. Cambridge: The Royal Society of Chemistry, pp. 227–265.

Castillo, I. & Vaart, A. van der (2012). "Needles and straw in a haystack: posterior concentration for possibly sparse sequences". *The Annals of Statistics* **40**.4, pp. 2069–2101.

Cheetham, A. K. (2002). "Structure determination from powder diffraction data: an overview". *Structure Determination from Powder Diffraction Data*. Ed. by David, W. I. F. et al. New York: Oxford University Press, pp. 13–28.

Chen, J. & Li, H. (2013). "Variable selection for sparse Dirichlet-multinomial regression with an application to microbiome data analysis". *The Annals of Applied Statistics* **7**.1, pp. 418–442.

Clark, J. S. et al. (2017). "Generalized joint attribute modeling for biodiversity analysis: median-zero, multivariate, multifarious data". *Ecological Monographs* **87**.1, pp. 34–56.

Clarke, K. R. (1993). "Non-parametric multivariate analyses of changes in community structure". *Australian Journal of Ecology* **18**.1, pp. 117–143.

Clemente, J. C. et al. (2012). “The impact of the gut microbiota on human health: an integrative view”. *Cell* **148**.6, pp. 1258–1270.

Craven, P. & Wahba, G. (1978). “Smoothing noisy data with spline functions”. *Numerische Mathematik* **31**.4, pp. 377–403.

Critchlow, D. E. et al. (1991). “Probability models on rankings”. *Journal of Mathematical Psychology* **35**.3, pp. 294–318.

Dannemiller, K. C. et al. (2014). “NextâĂŘgeneration DNA sequencing reveals that low fungal diversity in house dust is associated with childhood asthma development”. *Indoor Air* **24**.3, pp. 236–247.

Dannemiller, K. C. et al. (2016). “Influence of housing characteristics on bacterial and fungal communities in homes of asthmatic children”. *Indoor Air* **26**.2, pp. 179–192.

David, L. A. et al. (2014). “Diet rapidly and reproducibly alters the human gut microbiome”. *Nature* **505**.7484, pp. 559–563.

De Filippo, C. et al. (2010). “Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa”. *Proceedings of the National Academy of Sciences* **107**.33, pp. 14691–14696.

Deng, K. et al. (2014). “Bayesian aggregation of order-based rank data”. *Journal of the American Statistical Association* **109**.507, pp. 1023–1039.

Dicksved, J. et al. (2008). “Molecular analysis of the gut microbiota of identical twins with Crohn’s disease”. *The ISME Journal* **2**.7, pp. 716–727.

Dinnebier, R. & Müller, M. (2013). “Modern Rietveld Refinement, A Practical Guide”. *Modern Diffraction Methods*. Ed. by Mittemeijer, E. J. & Welzel, U. Weinheim, Germany: Wiley-VCH Verlag GmbH & Co. KGaA, pp. 27–60.

Dorazio, R. M. (2009). “On selecting a prior for the precision parameter of Dirichlet process mixture models”. *Journal of Statistical Planning and Inference* **139**.9, pp. 3384–3390.

Duane, S. et al. (1987). “Hybrid Monte Carlo”. *Physics Letters B* **195**.2, pp. 216–222.

Dunn, R. R. et al. (2013). “Home life: factors structuring the bacterial diversity found within and between homes”. *PLoS ONE* **8**.5, e64133.

Escobar, M. D. & West, M. (1995). “Bayesian density estimation and inference using mixtures”. *Journal of the American Statistical Association* **90**.430, pp. 577–588.

Fancher, C. M. et al. (2016). “Use of Bayesian Inference in Crystallographic Structure Refinement via Full Diffraction Profile Analysis”. *Scientific Reports* **6**, p. 31625.

Ferguson, T. S. (1973). “A Bayesian analysis of some nonparametric problems”. *The Annals of Statistics* **1**.2, pp. 209–230.

Fernandes, A. D. et al. (2013). “ANOVA-like differential expression (ALDEx) analysis for mixed population RNA-Seq”. *PLoS One* **8**.7.

Fligner, M. A. & Verducci, J. S., eds. (1993). *Probability Models and Statistical Analyses for Ranking Data*. New York: Springer-Verlag.

Flores, G. E. et al. (2012). “A direct PCR approach to accelerate analyses of human-associated microbial communities”. *PLoS ONE* **7**.9, e44563.

Friedman, J. et al. (2009). *The Elements of Statistical Learning*. Vol. 2. New York: Springer.

Fry, J. A. et al. (2011). “Completion of the 2006 national land cover database for the conterminous United States”. *Photogrammetric Engineering and Remote Sensing* **77**.9, pp. 858–864.

Gagin, A. & Levin, I. (2015). “Accounting for unknown systematic errors in Rietveld refinements: a Bayesian statistics approach”. *Journal of Applied Crystallography* **48**.4, pp. 1201–1211.

Garlea, V. O. et al. (2010). “The high-resolution powder diffractometer at the High Flux Isotope Reactor”. *Applied Physics A* **99**.3, pp. 531–535.

Gelfand, A. E. & Smith, A. F. (1990). “Sampling-based approaches to calculating marginal densities”. *Journal of the American Statistical Association* **85**.410, pp. 398–409.

Gelfand, A. E. et al. (2005). “Bayesian nonparametric spatial modeling with Dirichlet process mixing”. *Journal of the American Statistical Association* **100**.471, pp. 1021–1035.

Gelman, A. et al. (2014). *Bayesian Data Analysis*. Boca Raton: CRC Press.

Geman, S. & Geman, D. (1984). “Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images”. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **PAMI-6**.6, pp. 721–741.

George, E. I. & McCulloch, R. E. (1993). “Variable selection via Gibbs sampling”. *Journal of the American Statistical Association* **88**.423, pp. 881–889.

Geweke, J. (1992). “Evaluating the Accuracy of Sampling-Based Approaches to the Calculation of Posterior Moments”. *Bayesian Statistics 4*. Oxford: Oxford University Press, pp. 169–193.

Grantham, N. S. et al. (2015). “Fungi identify the geographic origin of dust samples”. *PLoS ONE* **10**.4, e0122605.

Grantham, N. S. et al. (2017). “MIMIX: a Bayesian mixed-effects model for microbiome data from designed experiments”. Manuscript in review.

Green, P. J. & Mira, A. (2001). "Delayed Rejection in Reversible Jump Metropolis-Hastings". *Biometrika* **88**.4, pp. 1035–1053.

Haario, H. et al. (1999). "Adaptive proposal distribution for random walk Metropolis algorithm". *Computational Statistics* **14**.3, pp. 375–395.

— (2001). "An adaptive Metropolis algorithm". *Bernoulli* **7**.2, pp. 223–242.

Haario, H. et al. (2006). "DRAM: efficient adaptive MCMC". *Statistics and Computing* **16**.4, pp. 339–354.

Hall, P. et al. (2008). "Modelling sparse generalized longitudinal observations with latent Gaussian processes". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70**.4, pp. 703–723.

Hamada, N. & Fujita, T (2002). "Effect of air-conditioner on fungal contamination". *Atmospheric Environment* **36**.35, pp. 5443–5448.

Hanley, J. A. & McNeil, B. J. (1982). "The meaning and use of the area under a receiver operating characteristic (ROC) curve". *Radiology* **143**.1, pp. 29–36.

Harris, I et al. (2014). "Updated high-resolution grids of monthly climatic observations – the CRU TS3.10 Dataset". *International Journal of Climatology* **34**.3, pp. 623–642.

Hastie, T. et al. (2009). *The Elements of Statistical Learning*. Vol. 2. New York: Springer Series in Statistics.

Heather, J. M. & Chain, B. (2016). "The sequence of sequencers: the history of sequencing DNA". *Genomics* **107**.1, pp. 1–8.

Hill, R. J. (1992). "Rietveld refinement round robin. I. Analysis of standard X-ray and neutron data for $PbSO_4$". *Journal of Applied Crystallography* **25**.5, pp. 589–610.

Hill, R. J. & Cranswick, L. M. D. (1994). "Rietveld refinement round robin. II. Analysis of monoclinic ZrO_2". *Journal of Applied Crystallography* **27**.5, pp. 802–844.

Hoffman, M. D. & Gelman, A. (2014). "The No-U-Turn Sampler: adaptively Setting Path Lengths in Hamiltonian Monte Carlo". *Journal of Machine Learning Research* **15**.1, pp. 1593–1623.

Howard, S. A. & Preston, K. D. (1989). "Profile Fitting of Powder Diffraction Patterns". *Reviews in Mineralogy, Vol. 20: Modern Powder Diffraction*. Ed. by Bish, D. L. & Post, J. E. Washington: Mineralogical Society of America, pp. 217–276.

Human Microbiome Project Consortium (2012). "Structure, function and diversity of the healthy human microbiome". *Nature* **486**.7402, pp. 207–214.

Izumi, F. (1993). “Rietveld analysis programs RIETAN and PREMOS and special applications”. *The Rietveld Method*. Ed. by Young, R. A. New York: Oxford University Press, pp. 236–253.

Izumi, F. (1989). “RIETAN: a Software Package for the Rietveld Analysis and Simulation of X-Ray and Neutron Diffraction Patterns”. *The Rigaku Journal* **6**.1, pp. 10–20.

Johnson, V. E. et al. (2002). “Bayesian analysis of rank data with application to primate intelligence experiments”. *Journal of the American Statistical Association* **97**.457, pp. 8–17.

Kang, S. S. et al. (2014). “Diet and exercise orthogonally alter the gut microbiome and reveal independent associations with anxiety and cognition”. *Molecular Neurodegeneration* **9**.1, p. 36.

Karhunen, K. (1947). *Über lineare Methoden in der Wahrscheinlichkeitsrechnung*. Vol. 37. Universitat Helsinki.

Kendall, M. G. (1938). “A new measure of rank correlation”. *Biometrika* **30**.1-2, pp. 81–93.

— (1945). “The treatment of ties in ranking problems”. *Biometrika* **33**.3, pp. 239–251.

Kettleson, E. M. et al. (2015). “Key determinants of the fungal and bacterial microbiomes in homes”. *Environmental Research* **138**, pp. 130–135.

Koop, G & Poirier, D. J. (1994). “Rank-ordered logit models: an empirical analysis of Ontario voter preferences”. *Journal of Applied Econometrics* **9**.4, pp. 369–388.

Kuo, L. & Mallick, B. (1998). “Variable selection for regression models”. *Sankhyā: The Indian Journal of Statistics, Series B* **60**.1, pp. 65–81.

Larson, A. C. & Von Dreele, R. B. (2004). *General Structure Analysis System (GSAS)*. Report LAUR 86-748. Los Alamos National Laboratory.

Lee, S. et al. (2010). “Sparse logistic principal components analysis for binary data”. *The Annals of Applied Statistics* **4**.3, pp. 1579–1601.

Lesniewski, J. E. et al. (2016). “Bayesian method for the analysis of diffraction patterns using BLAND”. *Journal of Applied Crystallography* **49**.6, pp. 2201–2209.

Ley, R. E. et al. (2005). “Obesity alters gut microbial ecology”. *Proceedings of the National Academy of Sciences* **102**.31, pp. 11070–11075.

Lorenz, E. N. (1956). “Empirical orthogonal functions and statistical weather prediction”.

Lutterotti, L. (2010). “Total pattern fitting for the combined size-strain-stress-texture determination in thin film diffraction”. *Nuclear Instruments and Methods in Physics Research Section B: Beam Interactions with Materials and Atoms* **268**.3, pp. 334–340.

Mandal, S. et al. (2015). “Analysis of composition of microbiomes: a novel method for studying microbial composition”. *Microbial Ecology in Health and Disease* **26**.

Marden, J. I. (1995). *Analyzing and Modeling Rank Data*. New York: Chapman & Hall.

McArdle, B. H. & Anderson, M. J. (2001). “Fitting multivariate models to community data: A comment on distance-based redundancy analysis”. *Ecology* **82**.1, pp. 290–297.

McCusker, L. B. et al. (1999). “Rietveld refinement guidelines”. *Journal of Applied Crystallography* **32**.1, pp. 36–50.

Metropolis, N. et al. (1953). “Equation of state calculations by fast computing machines”. *The Journal of Chemical Physics* **21**.6, pp. 1087–1092.

Mitchell, T. J. & Beauchamp, J. J. (1988). “Bayesian variable selection in linear regression”. *Journal of the American Statistical Association* **83**.404, pp. 1023–1032.

Murugiah, S. & Sweeting, T. (2012). “Selecting the precision parameter prior in Dirichlet process mixture models”. *Journal of Statistical Planning and Inference* **142**.7, pp. 1947–1959.

Navarro, D. J. et al. (2006). “Modeling individual differences using Dirichlet processes”. *Journal of Mathematical Psychology* **50**.2, pp. 101–122.

Neal, R. M. (2011). “MCMC using Hamiltonian dynamics”. *Handbook of Markov Chain Monte Carlo*. Ed. by Brooks, S. et al. Vol. 2. New York: CRC Press, pp. 113–162.

Nelsen, R. B. (1999). *An Introduction to Copulas*. New York: Springer.

Newton, M. A. et al. (2004). “Detecting differential gene expression with a semiparametric hierarchical mixture method”. *Biostatistics* **5**.2, pp. 155–176.

Petrone, S. et al. (2009). “Hybrid Dirichlet mixture models for functional data”. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **71**.4, pp. 755–782.

Powell, M. J. D. (1964). “An efficient method for finding the minimum of a function of several variables without calculating derivatives”. *The Computer Journal* **7**.2, pp. 155–162.

Prince, E (1993). “Mathematical aspects of Rietveld refinement”. *The Rietveld Method*. Ed. by Young, R. A. New York: Oxford University Press, pp. 43–54.

Qin, J. et al. (2012). “A metagenome-wide association study of gut microbiota in type 2 diabetes”. *Nature* **490**.7418, pp. 55–60.

Queipo-Ortuño, M. I. et al. (2013). “Gut microbiota composition in male rat models under different nutritional status and physical activity and its association with serum leptin and ghrelin levels”. *PLoS ONE* **8**.5, e65465.

Ravel, J. et al. (2011). "Vaginal microbiome of reproductive-age women". *Proceedings of the National Academy of Sciences* **108**.Supplement 1, pp. 4680–4687.

Reich, B. J. & Fuentes, M. (2007). "A multivariate semiparametric Bayesian spatial modeling framework for hurricane surface wind fields". *The Annals of Applied Statistics* **1**, pp. 249–264.

Reuter, J. A. et al. (2015). "High-throughput sequencing technologies". *Molecular Cell* **58**.4, pp. 586–597.

Rietveld, H. M. (1967). "Line profiles of neutron powder-diffraction peaks for structure refinement". *Acta Crystallographica* **22**.1, pp. 151–152.

— (1969). "A Profile Refinement Method for Nuclear and Magnetic Structures". *Journal of Applied Crystallography* **2**.2, pp. 65–71.

Ročková, V. & George, E. I. (2016). "The spike-and-slab lasso". *Journal of the American Statistical Association*. (in press).

Rodríguez, A. et al. (2010). "Latent stick-breaking processes". *Journal of the American Statistical Association* **105**.490, pp. 647–659.

Round, J. L. & Mazmanian, S. K. (2009). "The gut microbiota shapes intestinal immune responses during health and disease". *Nature Reviews Immunology* **9**.5, pp. 313–323.

Sakata, M & Cooper, M. J. (1979). "An analysis of the Rietveld refinement method". *Journal of Applied Crystallography* **12**.6, pp. 554–563.

Santacruz, A. et al. (2009). "Interplay between weight loss and gut microbiota composition in overweight adolescents". *Obesity* **17**.10, pp. 1906–1915.

Scott, H. G. (1983). "The estimation of standard deviations in powder diffraction Rietveld refinements". *Journal of Applied Crystallography* **16**.2, pp. 159–163.

Serban, N. et al. (2013). "Multilevel cross-dependent binary longitudinal data". *Biometrics* **69**.4, pp. 903–913.

Sethuraman, J. (1994). "A constructive definition of Dirichlet priors". *Statistica Sinica* **4**.2, pp. 639–650.

Shankland, K (2004). "Global Rietveld Refinement". *Journal of Research of the National Institute of Standards and Technology* **109**.1, pp. 143–154.

Shirota, S. et al. (2017). "Spatial joint species distribution modeling using Dirichlet processes".

Stein, M. L. (1999). *Interpolation of Spatial Data: Some Theory for Kriging*. New York: Springer.

Tierney, L. & Mira, A. (1999). "Some adaptive Monte Carlo methods for Bayesian inference". *Statistics in Medicine* **18**.18, pp. 2507–2515.

Toby, B. H. (2006). "*R* factors in Rietveld analysis: how good is good enough?" *Powder Diffraction* **21**.1, pp. 67–70.

Toby, B. H. & Von Dreele, R. B. (2013). "GSAS-II: the genesis of a modern open-source all purpose crystallography software package". *Journal of Applied Crystallography* **46**.2, pp. 544–549.

Turnbaugh, P. J. et al. (2009). "A core gut microbiome in obese and lean twins". *Nature* **457**.7228, pp. 480–484.

Voss, J. (2013). "Beyond Monte Carlo". *An Introduction to Statistical Computing: A Simulation-based Approach*. Vol. 1. Chichester: John Wiley & Sons, pp. 181–211.

Wadsworth, W. D. et al. (2017). "An integrative Bayesian Dirichlet-multinomial regression model for the analysis of taxonomic abundances in microbiome data". *BMC Bioinformatics* **18**.1, p. 94.

Walker, A. W. et al. (2011). "Dominant and diet-responsive groups of bacteria within the human colonic microbiota". *The ISME Journal* **5**, pp. 220–230.

Wang, T. & Zhao, H. (2017). "A Dirichlet-tree multinomial regression model for associating dietary nutrients with gut microorganisms". *Biometrics*. (in press).

Warton, D. I. (2011). "Regularized sandwich estimators for analysis of high-dimensional data using generalized estimating equations". *Biometrics* **67**.1, pp. 116–123.

Williamson, G. K. & Hall, W. H. (1953). "X-ray line broadening from filed aluminium and wolfram". *Acta Metallurgica* **1**.1, pp. 22–31.

Wu, G. D. et al. (2011). "Linking long-Term dietary patterns with gut microbial enterotypes". *Science* **334**.6052, pp. 105–108.

Xia, F. et al. (2013). "A logistic normal multinomial regression model for microbiome compositional data analysis". *Biometrics* **69**.4, pp. 1053–1063.

Xu, Z. & Knight, R. (2015). "Dietary effects on human gut microbiome diversity". *British Journal of Nutrition* **113**, S1–S5.

Yao, G. & Böckenholt, U. (1999). "Bayesian estimation of Thurstonian ranking models based on the Gibbs sampler". *British Journal of Mathematical & Statistical Psychology* **52**, pp. 79–92.

Young, R. A. (1993). "Introduction to the Rietveld Method". *The Rietveld Method*. Ed. by Young, R. A. New York: Oxford University Press, pp. 1–38.

Yu, P. L. H. (2000). "Bayesian analysis of order-statistics models for ranking data". *Psychometrika* **65**.3, pp. 281–299.

Zhang, X. et al. (2017). “Negative binomial mixed models for analyzing microbiome count data”. *BMC Bioinformatics* **18**.

Zhao, N. et al. (2015). “Testing in microbiome-profiling studies with MiRKAT, the microbiome regression-based kernel association test”. *The American Journal of Human Genetics* **96**.5, pp. 797–807.

Zhou, J. et al. (2015). “Bayesian factorizations of big sparse tensors”. *Journal of the American Statistical Association* **110**.512, pp. 1562–1576.

APPENDICES

APPENDIX

A

A NONPARAMETRIC SPATIAL TEST TO IDENTIFY FACTORS THAT SHAPE A MICROBIOME

A.1 Model properties

With the assumptions from Section 2.3, the model for the latent process is

$$\begin{aligned} Z_j(\mathbf{s}) &= \beta_{j0} + \mathbf{X}(\mathbf{s})\boldsymbol{\beta}_j + \boldsymbol{\Psi}(\mathbf{s})\boldsymbol{\alpha}_j + \epsilon_j(\mathbf{s}) \\ &= \beta_{j0} + \sum_{r=1}^{p} X_r(\mathbf{s})\delta_{jr}\gamma_{jr} + \sum_{l=1}^{L} \psi_l(\mathbf{s})\alpha_{jl} + \epsilon_j(\mathbf{s}). \end{aligned}$$

Conditionally on the cluster labels, g_j, the induced covariance between OTUs is

$$\mathrm{Cov}\left[Z_j(\mathbf{s}), Z_{j'}(\mathbf{s}') \,|\, g_j, g_{j'}\right] = \begin{cases} \rho \sum_{l=1}^{L} \psi_l(\mathbf{s})\psi_l(\mathbf{s}') & \text{if } g_j = g_{j'} \\ 0 & \text{if } g_j \neq g_{j'}, \end{cases}$$

for $j \neq j'$, which is nonstationary in general. With L sufficiently large, we can approximate any spatial covariance function by appealing to the Karhunen-Loéve theorem if the basis functions $\psi_l(\mathbf{s})$ are orthonormal (Karhunen, 1947). Marginally over the cluster labels, the induced covariance is

$$\text{Cov}\left[Z_j(\mathbf{s}), Z_{j'}(\mathbf{s}')\right] = 2\rho\varphi\left[\sum_{l=1}^{L}\psi_l(\mathbf{s})\psi_l(\mathbf{s}')\right],$$

where $j \neq j'$. The probability that two OTUs are from the same cluster, $\varphi = \sum_{k=1}^{\infty} p_k^2$, controls the dependence between OTUs in this separable, spatial, multivariate covariance function. If p_k is large for only a few clusters then φ will be close to 1, and the OTUs will partition into a small number of clusters leading to strong dependence. Otherwise, if the p_k values are smaller and more uniform, indicating weaker groupings, then φ will be close to 0.

Figure A.1 displays the empirical distribution of $\sum_{k=1}^{200} p_k^2$ for several values of the Dirichlet process precision parameter, D. For small values of D, the process favors fewer clusters reflected in

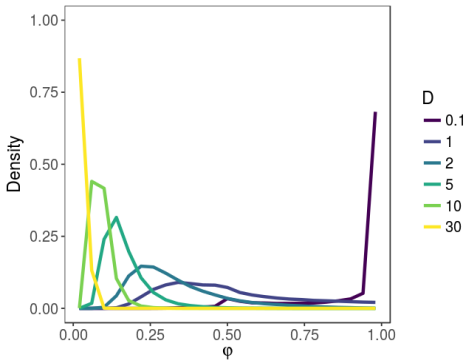


Figure A.1 Empirical distribution of $\varphi = \sum_{k=1}^{200} p_k^2$ for several values of the Dirichlet precision parameter, D.

$\sum_{k=1}^{200} p_k^2$ closer to 1. For large values of D, we see the reverse. We truncate the number of clusters at 200 for an example case of 1,000 taxa because $\text{E}[K \mid m, D] \approx D\log[(m+D)/D]$ (Antoniak, 1974), where K is the total number of groups created from m taxa. Empirically, the median value for the maximum of D based on its prior, discussed further in the following section, is 77 and thus a reasonable value for the maximum number of clusters is roughly 200.

A.2 Computing details

Recall that $i = 1,\ldots,n$ indexes the sampling locations, $j = 1,\ldots,m$ indexes the taxa, $r = 1,\ldots,p$ indexes the covariates, $l = 1,\ldots,L$ indexes the basis functions, and $k = 1,\ldots,K$ indexes the clusters for the Dirichlet process, which are capped at $K = \min(m, 500)$ for computational purposes. The full proposed model is

$$
\begin{aligned}
&Y_j(\mathbf{s}_i) \,|\, Z_j(\mathbf{s}_i) = \mathbb{1}\{Z_j(\mathbf{s}_i) > 0\} \\
&Z_j(\mathbf{s}_i) \,|\, \mathbf{X}, \beta_{j0}, \boldsymbol{\beta}_j, \boldsymbol{\alpha}_j, \rho \overset{\text{indep}}{\sim} \mathrm{N}\left(\beta_{j0} + \mathbf{X}(\mathbf{s}_i)\boldsymbol{\beta}_j + \boldsymbol{\Psi}(\mathbf{s}_i)\boldsymbol{\alpha}_j, 1-\rho\right) \\
&\boldsymbol{\alpha}_j \,|\, g_j = k, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K = \boldsymbol{\mu}_k \\
&\boldsymbol{\mu}_k \,|\, \boldsymbol{\mu}_0, \rho \overset{\text{iid}}{\sim} \mathrm{N}_L\left(\boldsymbol{\mu}_0, \rho \mathbf{I}_L\right) \\
&\boldsymbol{\mu}_0 \,|\, \tau_{\mu_0} \sim \mathrm{N}_L\left(\mathbf{0}, \tau_{\mu_0}^{-1}\mathbf{I}_L\right),\ \tau_{\mu_0} \sim \text{Gamma}(a_{\mu_0} = 0.1, b_{\mu_0} = 0.1) \\
&\rho \sim \mathrm{U}(0,1) \\
&\mathrm{P}(g_j = k) = p_k = V_k \prod_{u<k}(1-V_u) \text{ for } k > 1 \text{ and } p_1 = V_1 \\
&V_u \,|\, D \overset{\text{iid}}{\sim} \text{Beta}(1, D) \text{ for } u = 1,\ldots,K-1 \text{ and } V_K = 1 \\
&D \sim \text{Gamma}(a_d = 0.1, b_d = 0.1) \\
&\beta_{j0} \,|\, \tau_0 \overset{\text{iid}}{\sim} \mathrm{N}(0, \tau_0^{-1}),\ \tau_0 \sim \text{Gamma}(a_0 = 0.1, b_0 = 0.1) \\
&\beta_{jr} \,|\, \delta_{jr}, \gamma_{jr} = \delta_{jr}\gamma_{jr} \\
&\delta_{jr} \,|\, \pi_r \overset{\text{indep}}{\sim} \text{Bernoulli}(\pi_r) \\
&\mathrm{P}(\pi_r \,|\, \omega, \theta) = \omega\left[\frac{1}{\mathrm{B}(1,\theta)}(1-\pi_r)^{\theta-1}\right] + (1-\omega) \text{ for } \omega \in [0,1] \text{ and } \theta \geq 1, \text{ fixed} \\
&\gamma_{jr} \,|\, \tau_r \overset{\text{indep}}{\sim} \mathrm{N}(0, \tau_r^{-1}),\ \tau_r \overset{\text{iid}}{\sim} \text{Gamma}(a_r = 1, b_r = 2.7)
\end{aligned}
$$

We have chosen to follow the approach of allowing $a_d, b_d \to 0$ by setting them to small values (Escobar & West, 1995; Navarro et al., 2006). Recently, alternative approaches have been developed that attempt to correct for pitfalls wherein learning about D is difficult and therefore inference is sensitive to its prior specification in small sample problems (Dorazio, 2009; Murugiah & Sweeting, 2012). However, these approaches are not feasible for high-dimensional problems because of a reliance on unsigned Stirling numbers of the first kind in Dorazio (2009) or on an extensive performance study in Murugiah & Sweeting (2012), which are computationally demanding, if not impossible.

The remaining hyperprior parameters are chosen as the standard uninformative values, with

the exception of a_r and b_r. In this setting, the usual values of 0.1 caused numerical instability within matrix inversions. To resolve this, the values $a_r = 1$ and $b_r = 2.7$ were chosen to closely match the Gamma(0.1, 0.1) distribution while restricting the maximum induced variance slightly to improve computational stability. An alternative solution is to assume that the variance of the magnitudes is the same for all covariates in the prior, i.e., $\tau_1 = \ldots = \tau_r = \tau$, and use the standard prior $\tau \sim \text{Gamma}(0.1, 0.1)$. Both options correct the numerical instability and based on simulation testing, selecting either option is effective for variable selection. Thus, unless there is outside expertise to suggest otherwise, we recommend the parsimonious option and suggest fitting the models using a common variance, where $\gamma_{jr} \,|\, \tau \overset{\text{iid}}{\sim} \text{N}(0, \tau^{-1})$, and $\tau \sim \text{Gamma}(a_\gamma = 0.1, b_\gamma = 0.1)$. In the implementation details to follow, we give the update for the more complex case, but it should be stated that we use the common variance option as the default value in our implementation and that we utilized this simplification in all of the results presented in the body of the paper.

Posterior samples are drawn using Markov chain Monte Carlo (MCMC) with convergence monitored by inspecting trace plots. Most model parameters can be updated via Gibbs sampling, with the exception of the variance of the structural component of the residual dependence, ρ, which is updated using the Metropolis algorithm.

Metropolis sampling for ρ

The log posterior distribution for ρ, conditional on all other parameters, is given by

$$\ell(\rho \,|\, \cdots) \propto -\frac{nm}{2}\log(1-\rho) - \frac{LK}{2}\log(\rho) - \frac{1}{2\rho}\sum_{k=1}^{K}\left\|\boldsymbol{\mu}_k - \boldsymbol{\mu}_0\right\|_2^2 - \frac{1}{2(1-\rho)}\sum_{j=1}^{m}\left\|\mathbf{Z}_j - \beta_{j0}\mathbf{1}_n - \mathbf{X}\boldsymbol{\beta}_j - \boldsymbol{\Psi}\boldsymbol{\alpha}_j\right\|_2^2.$$

Thus, ρ cannot be updated using Gibbs sampling and instead requires a Metropolis update. Because ρ is bound by the interval $[0, 1]$, we use the logit transformation and work with the continuous variable $\text{logit}(\rho) = \log\left(\frac{\rho}{1-\rho}\right)$. At each iteration, we propose a candidate, $\text{logit}(\rho^*) \sim \text{N}\{\text{logit}(\rho), \sigma_M^2\}$, where σ_M^2 is adapted within the burn-in period to maintain an acceptance rate $\in [0.3, 0.7]$.

Gibbs sampling

All other model parameters are drawn from their full conditional distributions using Gibbs sampling. The full conditional distributions are as follows:

$$Z_j(\mathbf{s}_i) \,|\, \cdots \sim \text{TN}\{\beta_{j0} + \mathbf{X}(\mathbf{s}_i)\boldsymbol{\beta}_j + \boldsymbol{\Psi}(\mathbf{s}_i)\boldsymbol{\alpha}_j, (1-\rho); (l_{ij}, u_{ij})\},$$

where $\text{TN}\{\mu, \sigma^2; (a, b)\}$ denotes the $\text{N}(\mu, \sigma^2)$ distribution truncated to lie in the interval (a, b),

and $(l_{ij}, u_{ij}) = (0, \infty)$ if $Y_j(\mathbf{s}_i) = 1$ or $(l_{ij}, u_{ij}) = (-\infty, 0)$ if $Y_j(\mathbf{s}_i) = 0$,

$$\beta_{j0} \,|\, \cdots \sim \mathrm{N}\left(\frac{1}{n + (1-\rho)\tau_0} \sum_{i=1}^{n} \left[Z_j(\mathbf{s}_i) - \mathbf{X}(\mathbf{s}_i)\boldsymbol{\beta}_j - \boldsymbol{\Psi}(\mathbf{s}_i)\boldsymbol{\alpha}_j \right], \frac{1-\rho}{n + (1-\rho)\tau_0} \right),$$

$$\tau_0 \,|\, \cdots \sim \mathrm{Gamma}\left(a_0 + \frac{m}{2}, b_0 + \frac{1}{2}\sum_{j=1}^{m} \beta_{j0}^2 \right),$$

$$\mathrm{P}\left(g_j = k \,|\, \cdots\right) = \frac{p_k \mathrm{P}(\mathbf{Z}_j \,|\, \boldsymbol{\alpha}_j = \boldsymbol{\mu}_k)}{\sum_{c=1}^{K} p_c \mathrm{P}(\mathbf{Z}_j \,|\, \boldsymbol{\alpha}_j = \boldsymbol{\mu}_c)},$$

$V_u \,|\, \cdots \sim \mathrm{Beta}(1 + n_k, D + n_{>k})$ for $u = 1, \ldots, K-1$,

where $n_k = \sum_{j=1}^{m} \mathbb{I}\{g_j = k\}$ and $n_{>k} = \sum_{j=1}^{m} \mathbb{I}\{g_j > k\}$

$$D \,|\, \cdots \sim \mathrm{Gamma}\left(a_d + K - 1, b_d - \sum_{u=1}^{K-1} \log(1 - V_u) \right),$$

$$\boldsymbol{\mu}_k \,|\, \cdots \sim \mathrm{N}_L\left(\left[\frac{n_k}{1-\rho}\boldsymbol{\Psi}'\boldsymbol{\Psi} + \frac{1}{\rho} \right]^{-1} \left[\frac{1}{\rho}\boldsymbol{\mu}_0 + \frac{1}{1-\rho}\boldsymbol{\Psi}' \sum_{j:g_j=k} \mathbf{Z}_j - \beta_{j0}\mathbf{1}_n - \mathbf{X}\boldsymbol{\beta}_j \right], \left[\frac{n_k}{1-\rho}\boldsymbol{\Psi}'\boldsymbol{\Psi} + \frac{1}{\rho} \right]^{-1} \right),$$

if $n_k > 0$, otherwise $\boldsymbol{\mu}_k$ is drawn from the prior distribution,

$$\boldsymbol{\mu}_0 \,|\, \cdots \sim \mathrm{N}_L\left(\frac{1}{K + \rho\tau_{\mu_0}} \sum_{k=1}^{K} \boldsymbol{\mu}_k, \frac{\rho}{K + \rho\tau_{\mu_0}} \right),$$

$$\tau_{\mu_0} \,|\, \cdots \sim \mathrm{Gamma}\left(a_{\mu_0} + \frac{L}{2}, b_{\mu_0} + \frac{1}{2}\boldsymbol{\mu}_0'\boldsymbol{\mu}_0 \right),$$

$$\boldsymbol{\gamma}_j \,|\, \cdots \sim \mathrm{N}_p\left(\left[\frac{1}{1-\rho}\mathbf{X}_{*j}'\mathbf{X}_{*j} + \mathbf{T}_\gamma \right]^{-1} \frac{1}{1-\rho}\mathbf{X}_{*j}' \left[\mathbf{Z}_j - \beta_{j0}\mathbf{1}_n - \boldsymbol{\Psi}\boldsymbol{\alpha}_j \right], \left[\frac{1}{1-\rho}\mathbf{X}_{*j}'\mathbf{X}_{*j} + \mathbf{T}_\gamma \right]^{-1} \right),$$

where $\mathbf{T}_\gamma = \mathrm{diag}\{\tau_1, \ldots, \tau_p\}$, $\boldsymbol{\Lambda}_j = \mathrm{diag}\{\delta_{j1}, \ldots, \delta_{jp}\}$, and $\mathbf{X}_{*j} = \mathbf{X}\boldsymbol{\Lambda}_j$,

$$\tau_r \,|\, \cdots \sim \mathrm{Gamma}\left(a_r + \frac{m}{2}, b_r + \frac{1}{2}\sum_{j=1}^{m} \gamma_{jr}^2 \right),$$

$$\delta_{jr} \,|\, \cdots \sim \mathrm{Bernoulli}\left(\pi_{jr}^*\right)$$

where $\mathrm{logit}(\pi_{jr}^*) = \log(\pi_r) - \log(1 - \pi_r)$

$$- \frac{1}{2(1-\rho)} \sum_{i=1}^{n} \left[Z_j(\mathbf{s}_i) - \beta_{j0} - \boldsymbol{\Psi}(\mathbf{s}_i)\boldsymbol{\alpha}_j - \sum_{q \neq r} X_q(\mathbf{s}_i)\delta_{jq}\gamma_{jq} - X_r(\mathbf{s}_i)\gamma_{jr} \right]^2$$

$$+ \frac{1}{2(1-\rho)} \sum_{i=1}^{n} \left[Z_j(\mathbf{s}_i) - \beta_{j0} - \boldsymbol{\Psi}(\mathbf{s}_i)\boldsymbol{\alpha}_j - \sum_{q \neq r} X_q(\mathbf{s}_i)\delta_{jq}\gamma_{jq} \right]^2,$$

$$\mathrm{P}(\pi_r \mid \cdots) = W_r \mathrm{Beta}(1+M_r, \theta+m-M_r) + (1-W_r)\mathrm{Beta}(1+M_r, 1+m-M_r)$$

where $\mathrm{B}(a,b)$ is the Beta function, $M_r = \sum_{j=1}^{m} \delta_{jr}$,

and $W_r = \dfrac{\omega\theta\,\mathrm{B}(1+M_r, \theta+m-M_r)}{\omega\theta\,\mathrm{B}(1+M_r, \theta+m-M_r) + (1-\omega)\mathrm{B}(1+M_r, 1+m-M_r)}$.

APPENDIX

B

BAYESIAN VARIABLE SELECTION FOR HIGH-DIMENSIONAL RANK DATA

B.1 Identifiability of Parameters

To begin, recall that the response of interest is a set of ranks, $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{im})$, where $Y_{ij} \in \{1, \ldots, m\}$. These ranks are invariant to monotonic shifts that affect all m taxa equally. We assume that these ranks depend on a set of latent counts, Z_{ij}, that in turn depend linearly on a set of p covariates, $X_{i1}, \ldots, X_{ip}$, which have been centered and scaled, and a set of taxa-specific intercepts, $\beta_{10}, \ldots, \beta_{m0}$. To assess identifiability, we assume at the outset that $\beta_{j0}, \beta_{jr}, e_{ij} \overset{\text{iid}}{\sim} (0,1)$, where $W \sim (0,1)$ indicates that $\mathrm{E}[W] = 0$ and $\mathrm{Var}(W) = 1$. By initially considering linear combinations of these parameters, we will derive the necessary constraints for identifiability in this framework. In the most general form, we write

$$\begin{aligned} Z_{ij} &= [a_0 + b_0\beta_{j0}] + \sum_{r=1}^{p} X_{ir}[a_r + b_r\beta_{jr}] + \sigma e_{ij} \\ &= a_0 + b_0\beta_{j0} + \sum_{r=1}^{p} [a_r X_{ir} + b_r X_{ir}\beta_{jr}] + \sigma e_{ij} \end{aligned}$$

$$= \underbrace{\left[a_0 + \sum_{r=1}^{p} a_r X_{ir} \right]}_{\text{No } j} + \underbrace{\left[b_0 \beta_{j0} \right]}_{\text{Depends on } j} + \underbrace{\left[\sum_{r=1}^{p} b_r X_{ir} \beta_{jr} + \sigma e_{ij} \right]}_{\text{Depends on } i \text{ and } j}$$

$$= \left[a_0 + \sum_{r=1}^{p} a_r X_{ir} \right] + \left[b_0 \beta_{j0} \right] + \sigma \left[\sum_{r=1}^{p} \frac{b_r}{\sigma} X_{ir} \beta_{jr} + e_{ij} \right]. \tag{B.1}$$

From the final term in (B.1), in order for the scale of β_{jr} coefficients to be identified, we see that one of the two multiplicative terms, σ and b_r, needs to be constrained. Without loss of generality, we set $\sigma = 1$. Thus, the expression for Z_{ij} becomes

$$Z_{ij} = \left[a_0 + \sum_{r=1}^{p} a_r X_{ir} \right] + \left[b_0 \beta_{j0} \right] + \left[\sum_{r=1}^{p} b_r X_{ir} \beta_{jr} + e_{ij} \right] \tag{B.2}$$

Note that because we are concerned with the *ranks* of the m-dimensional count vectors, $\mathbf{Z}_i$, the "overall intercept" includes any terms that do not depend on j, as these terms will affect only the magnitude of the values and not their relative ranks to one another. The overall intercept in the first term of (B.2) affects the values of $Z_{i1}, \ldots, Z_{im}$, but does not affect their relative ranks. As such, we must constrain this term to be 0 by setting $a_0 = 0$ and $a_r = 0$ for all $r = 1, \ldots, p$. Without this requirement, two different latent process vectors could give rise to the same ranked response vector. Therefore, the final identifiable expression for the latent process is

$$Z_{ij} = b_0 \beta_{j0} + \sum_{r=1}^{p} b_r X_{ir} \beta_{jr} + e_{ij}, \tag{B.3}$$

corresponding to the requirement that the covariates be centered and scaled, $\beta_{j0} \overset{\text{iid}}{\sim} (0, b_0^2)$, $\beta_{jr} \overset{\text{iid}}{\sim} (0, b_r^2)$, and $e_{ij} \sim (0, 1)$. Because we enforce these constraints in the prior distributions outlined in Section 3.3, all of the parameters in this model are identifiable.

B.2 Computing Details for Base Model

Recall that $i = 1, \ldots, n$ indexes observations, $j = 1, \ldots, m$ indexes taxa, $r = 1, \ldots, p$ indexes covariates. Let C_{ij} be the read count for taxon j in sample i. Then $Y_{ij} = k$, or equivalently $R_{ik} = j$, indicates

that in sample i, taxon j has the k^{th} largest value of C_{ij}. The proposed model is

$$Y_{ij} = \begin{cases} 0 & \text{if } Z_{ij} < 0 \\ 1 + \sum_{k \neq j} \mathbb{1}\{Z_{ik} < Z_{ij}\} & \text{if } Z_{ij} > 0 \end{cases}$$

$$Z_{ij} \,|\, \mathbf{X}, \beta_{j0}, \boldsymbol{\beta}_j \overset{\text{indep}}{\sim} \text{N}\left(\beta_{j0} + \mathbf{X}_i \boldsymbol{\beta}_j, 1\right) \text{ with } Z_{iR_{i1}} \geq \ldots \geq Z_{iR_{im}}$$

$$\beta_{j0} \,|\, \tau_0 \overset{\text{iid}}{\sim} \text{N}(0, \tau_0^{-1}),\ \tau_0 \sim \text{Gamma}(a_0 = 0.1, b_0 = 0.1)$$

$$\beta_{jr} \,|\, \delta_{jr}, \gamma_{jr} = \delta_{jr} \gamma_{jr}$$

$$\delta_{jr} \,|\, \pi_r \overset{\text{indep}}{\sim} \text{Bernoulli}(\pi_r),\ \pi_r \overset{\text{iid}}{\sim} \text{Beta}(a_\pi = 1, b_\pi = m)$$

$$\gamma_{jr} \,|\, \tau_r \overset{\text{indep}}{\sim} \text{N}(0, \tau_r^{-1}),\ \tau_r \overset{\text{iid}}{\sim} \text{Gamma}(a_\tau = 0.1, b_\tau = 0.1)$$

Posterior samples for this model are drawn using Markov Chain Monte Carlo (MCMC) sampling, with convergence monitored by inspecting trace plots. All parameters in the detailed model can be updated using Gibbs sampling (Geman & Geman, 1984; Gelfand & Smith, 1990). The full conditional distributions are as follows:

$$Z_{ij} \,|\, \cdots \sim \text{TN}\left(\beta_{j0} + \mathbf{X}_i \boldsymbol{\beta}_j,\, 1;\, (l_{ij}, u_{ij})\right)$$

where $\text{TN}\{\mu, \sigma^2; (a, b)\}$ denotes the $\text{N}(\mu, \sigma^2)$ distribution truncated to lie in the interval (a, b), and $(l_{ij}, u_{ij}) = (Z_{R_{iY_{ij}-1}}, Z_{R_{iY_{ij}+1}})$ with $Z_{R_{iY_{ij}-1}} > 0$ if $C_{ij} > 0$ and $(l_{ij}, u_{ij}) = (-\infty, 0)$ if $C_{ij} = 0$,

$$\beta_{j0} \,|\, \cdots \sim \text{N}\left(\frac{\sum_{i=1}^n Z_{ij} - \mathbf{X}_i \boldsymbol{\beta}_j}{n + \tau_0},\, \frac{1}{n + \tau_0}\right),$$

$$\tau_0 \,|\, \cdots \sim \text{Gamma}\left(a_0 + \frac{m}{2},\, b_0 + \frac{1}{2} \sum_{j=1}^m \beta_{j0}^2\right),$$

$$\delta_{jr} \,|\, \cdots \sim \text{Bernoulli}(\pi_{jr}^*)$$

$$\text{where } \text{logit}(\pi_{jr}^*) = \log(\pi_r) - \frac{1}{2} \sum_{i=1}^n \left[Z_{ij} - \beta_{j0} - \sum_{k \neq r} X_{ik} \delta_{jk} \gamma_{jk} - X_{ir} \gamma_{jr} \right]^2$$

$$- \log(1 - \pi_r) + \frac{1}{2} \sum_{i=1}^n \left[Z_{ij} - \beta_{j0} - \sum_{k \neq r} X_{ik} \delta_{jk} \gamma_{jk} \right]^2,$$

$$\pi_r \,|\, \cdots \sim \text{Beta}(a_\pi + \text{M}_r,\, b_\pi + m - \text{M}_r) \text{ where } \text{M}_r = \sum_{j=1}^m \delta_{jr},$$

$$\gamma_j \,|\, \cdots \sim \mathrm{N}\Big(\Big[\mathbf{X}'_{*j}\mathbf{X}_{*j} + \mathbf{T}_\gamma\Big]^{-1} \Big[\mathbf{X}'_{*j}\big(\mathbf{Z}_j - \beta_{j0}\mathbf{1}_n\big)\Big], \Big[\mathbf{X}'_{*j}\mathbf{X}_{*j} + \mathbf{T}_\gamma\Big]^{-1} \Big)$$

where $\mathbf{T}_\gamma = \mathrm{diag}\{\tau_1, \ldots, \tau_p\}$, $\mathbf{\Lambda}_j = \mathrm{diag}\{\delta_{j1}, \ldots, \delta_{jp}\}$ and $\mathbf{X}_{*j} = \mathbf{X}\mathbf{\Lambda}_j$,

$$\tau_r \,|\, \cdots \sim \mathrm{Gamma}\left(a_\tau + \frac{m}{2},\, b_\tau + \frac{1}{2}\sum_{j=1}^{m} \gamma_{jr}^2 \right).$$

B.3 Computing Details for Random Effects Model

To capture cross-dependence in the abundance counts, we incorporate random effects into the base model. In addition to the notation established in Appendix B.2, let $l = 1, \ldots, L$ index basis functions. The extended model is

$$Y_{ij} = \begin{cases} 0 & \text{if } Z_{ij} < 0 \\ 1 + \sum_{k \neq j} \mathbb{1}\{Z_{ik} < Z_{ij}\} & \text{if } Z_{ij} > 0 \end{cases}$$

$$Z_{ij} \,|\, \mathbf{X}, \mathbf{\Psi}, \beta_{j0}, \boldsymbol{\beta}_j, \boldsymbol{\alpha}_i, \rho \overset{\text{indep}}{\sim} \mathrm{N}\big(\beta_{j0} + \mathbf{X}_i\boldsymbol{\beta}_j + \boldsymbol{\alpha}_i'\mathbf{\Psi}_j, 1-\rho\big) \text{ with } Z_{iR_{i1}} \geq \ldots \geq Z_{iR_{im}}$$

$$\beta_{j0} \,|\, \tau_0 \overset{\text{iid}}{\sim} \mathrm{N}(0, \tau_0^{-1}),\ \tau_0 \sim \mathrm{Gamma}(a_0 = 0.1,\, b_0 = 0.1)$$

$$\boldsymbol{\beta}_{jr} \,|\, \delta_{jr}, \gamma_{jr} = \delta_{jr}\gamma_{jr}$$

$$\delta_{jr} \,|\, \pi_r \overset{\text{indep}}{\sim} \mathrm{Bernoulli}(\pi_r),\ \pi_r \overset{\text{iid}}{\sim} \mathrm{Beta}(a_\pi = 1,\, b_\pi = m)$$

$$\gamma_{jr} \,|\, \tau_r \overset{\text{indep}}{\sim} \mathrm{N}(0, \tau_r^{-1}),\ \tau_r \overset{\text{iid}}{\sim} \mathrm{Gamma}(a_\tau = 0.1,\, b_\tau = 0.1)$$

$$\alpha_{il} \,|\, \rho \overset{\text{iid}}{\sim} \mathrm{N}\big(0, \rho\big),\ \rho \sim \mathrm{U}(0,1)$$

As in the base model, MCMC is used to obtain posterior samples. However, in this model, the posterior distribution for ρ does not take the form of a known distribution. Therefore, the update step for ρ is accomplished using the Metropolis algorithm (Metropolis et al., 1953).

Gibbs sampling

The full conditional distributions for all model parameters other than ρ are as follows:

$$Z_{ij} \,|\, \cdots \sim \mathrm{TN}\big(\beta_{j0} + \mathbf{X}_i\boldsymbol{\beta}_j + \boldsymbol{\alpha}_i'\mathbf{\Psi}_j,\, 1-\rho; (l_{ij}, u_{ij})\big)$$

where $\mathrm{TN}\{\mu, \sigma^2; (a,b)\}$ denotes the $\mathrm{N}(\mu, \sigma^2)$ distribution truncated to lie in the interval (a,b), and $(l_{ij}, u_{ij}) = (Z_{R_{iY_{ij}-1}}, Z_{R_{iY_{ij}+1}})$ with $Z_{R_{iY_{ij}-1}} > 0$ if $C_{ij} > 0$ and $(l_{ij}, u_{ij}) = (-\infty, 0)$

if $C_{ij}=0$,

$$\beta_{j0}\,|\cdots\sim \mathrm{N}\left(\frac{\sum_{i=1}^{n} Z_{ij}-\mathbf{X}_i\beta_j-\boldsymbol{\alpha}_i'\boldsymbol{\Psi}_j}{n+(1-\rho)\tau_0},\ \frac{1-\rho}{n+(1-\rho)\tau_0}\right),$$

$$\tau_0\,|\cdots\sim \mathrm{Gamma}\left(a_0+\frac{m}{2},\ b_0+\frac{1}{2}\sum_{j=1}^{m}\beta_{j0}^2\right),$$

$$\delta_{jr}\,|\cdots\sim \mathrm{Bernoulli}(\pi_{jr}^*)$$

$$\text{where } \mathrm{logit}(\pi_{jr}^*)=\log(\pi_r)-\frac{1}{2(1-\rho)}\sum_{i=1}^{n}\left[Z_{ij}-\beta_{j0}-\boldsymbol{\alpha}_i'\boldsymbol{\Psi}_j-\sum_{k\neq r}X_{ik}\delta_{jk}\gamma_{jk}-X_{ir}\gamma_{jr}\right]^2$$

$$-\log(1-\pi_r)+\frac{1}{2(1-\rho)}\sum_{i=1}^{n}\left[Z_{ij}-\beta_{j0}-\boldsymbol{\alpha}_i'\boldsymbol{\Psi}_j-\sum_{k\neq r}X_{ik}\delta_{jk}\gamma_{jk}\right]^2,$$

$$\pi_r\,|\cdots\sim \mathrm{Beta}(a_\pi+\mathrm{M}_r,\ b_\pi+m-\mathrm{M}_r) \text{ where } \mathrm{M}_r=\sum_{j=1}^{m}\delta_{jr},$$

$$\boldsymbol{\gamma}_j\,|\cdots\sim \mathrm{N}\left(\left[\frac{1}{1-\rho}\mathbf{X}_{*j}'\mathbf{X}_{*j}+\mathbf{T}_\gamma\right]^{-1}\left[\frac{1}{1-\rho}\mathbf{X}_{*j}'\left(\mathbf{Z}_j-\beta_{j0}\mathbf{1}_n-\boldsymbol{\alpha}'\Psi_j\right)\right],\left[\frac{1}{1-\rho}\mathbf{X}_{*j}'\mathbf{X}_{*j}+\mathbf{T}_\gamma\right]^{-1}\right)$$

where $\mathbf{T}_\gamma=\mathrm{diag}\{\tau_1,\ldots,\tau_p\}$, $\boldsymbol{\Lambda}_j=\mathrm{diag}\{\delta_{j1},\ldots,\delta_{jp}\}$ and $\mathbf{X}_{*j}=\mathbf{X}\boldsymbol{\Lambda}_j$,

$$\tau_r\,|\cdots\sim \mathrm{Gamma}\left(a_\tau+\frac{m}{2},\ b_\tau+\frac{1}{2}\sum_{j=1}^{m}\gamma_{jr}^2\right),$$

$$\boldsymbol{\alpha}_i\,|\cdots\sim \mathrm{N}\left(\left[\frac{1}{1-\rho}\boldsymbol{\Psi}\boldsymbol{\Psi}'+\frac{1}{\rho}\mathbf{I}_L\right]^{-1}\left[\frac{1}{1-\rho}\boldsymbol{\Psi}(\mathbf{Z}_i-\boldsymbol{\beta}_0-\mathbf{X}_i\boldsymbol{\beta})'\right],\left[\frac{1}{1-\rho}\boldsymbol{\Psi}\boldsymbol{\Psi}'+\frac{1}{\rho}\mathbf{I}_L\right]^{-1}\right)$$

where $\boldsymbol{\beta}_0=(\beta_{10},\ldots,\beta_{m0})$.

Metropolis sampling for ρ

Conditional on all other parameters, the log posterior distribution function for ρ is given by

$$\ell(\rho\,|\cdots)\propto-\frac{nm}{2}\log(1-\rho)-\frac{nL}{2}\log(\rho)-\sum_{i=1}^{n}\sum_{j=1}^{m}\left[Z_{ij}-\beta_{j0}-\mathbf{X}_i\boldsymbol{\beta}j-\boldsymbol{\alpha}_i'\boldsymbol{\Psi}_j\right]^2-\frac{1}{2\rho}\sum_{i=1}^{n}\sum_{l=1}^{L}\alpha_{il}^2.$$

Because ρ is constrained to the interval $[0,1]$, we use a logit transformation and draw the candidate parameter as $\mathrm{logit}(\rho^*)\sim \mathrm{N}\{\mathrm{logit}(\rho),\sigma_M^2\}$, where σ_M^2 controls the spread of the proposal distribution and is adapted to maintain an acceptance rate in $[0.3,0.5]$ during the burn-in period.