Modeling and Computation for Multivariate spatial categorical data and related theory with applications to historical ecology

By

Stephen Berg

A dissertation submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

(Statistics)

at the

University of Wisconsin–Madison

2020

Date of final oral examination: June 17, 2020

The dissertation is approved by the following members of the Final Oral Committee:

Professor Jun Zhu, Department of Statistics and Department of Entomology

Professor Murray K. Clayton, Department of Statistics and Department of Plant Pathology

Professor David Mladenoff, Department of Forest and Wildlife Ecology

Professor Yazhen Wang, Department of Statistics

Professor Ronald E. Gangnon, Department of Population Health Sciences and Department of Biostatistics and Medical Informatics

Abstract

This thesis has two main components. The first component is an analysis of the Wisconsin Public Land Survey (PLS) dataset, with the goal of identifying and mapping historical forest types. The Wisconsin Public Land Survey database describes historical forest composition at high spatial resolution, and is of interest in ecological studies of forest composition in Wisconsin just prior to significant Euro-American settlement. For such studies, it is useful to identify recurring subpopulations of tree species known as communities, but standard clustering approaches for subpopulation identification do not account for dependence between spatially nearby observations. Here, we develop and fit a latent discrete Markov random field model for the purpose of identifying and classifying historical forest communities based on spatially referenced multivariate tree species counts across Wisconsin. We show empirically for the actual dataset and through simulation that our latent Markov random field modeling approach improves prediction and parameter estimation performance. For model fitting, we introduce a new stochastic approximation algorithm, which enables computationally efficient estimation and classification of large amounts of spatial multivariate count data.

The second component of this thesis is a study of control variate methods for Markov chain Monte Carlo (MCMC) simulations. Control variates are a method used for reducing the variance of averages over samples taken from Monte Carlo or Markov chain Monte Carlo simulations. We propose new methodology for the setting of deterministic sweep sampling using $K \geq 2$ transition kernels. For the widely applicable deterministic sweep Gibbs sampler, we show that the projection properties of Gibbs

kernels lead to a statistically efficient and easy to implement control variate estimator, which has theoretical and practical benefits over competing methodology in the literature. In particular, for the data augmentation Gibbs sampler, our control variate estimator is guaranteed to achieve a smaller asymptotic variance than a widely used Rao-Blackwellization approach, typically with negligible increases in computational cost. Additionally, we provide variance reduction guarantees for a Rao-Blackwellization approach for more general Gibbs sampling settings than those in existing results. We conduct a simulation study which demonstrates that the theoretical benefits of our proposed approaches are realized in practice.

Acknowledgments

I have benefited greatly from the company and insights of many people in the Statistics Department throughout the past six years. Some of my fondest memories are from studying for the PhD qualifying exam at Memorial Library with Hyebin Song and Ting Ye during the summer of 2016. I enjoyed our discussions and excursions down State Street immensely. It was also during this intellectually stimulating period of time that my interest in statistics research solidified. I am thankful additionally for the conversations I had with my officemates Jared Huling and Mike Wurm during my time in graduate school. I particularly appreciate the perspectives and advice they shared with me, relating both to research and to surviving and thriving in Room 1570, Medical Sciences Center.

I am grateful to my thesis advisors, Professor Jun Zhu and Professor Murray Clayton. Over my time at the University of Wisconsin-Madison, they directed me toward challenging research problems, and encouraged me to develop questions and solutions myself. They deftly balanced allowing me freedom and independence to pursue my own research directions, while offering guidance and pushing me to progress forward and achieve concrete milestones. My current interests in Markov random field models and Markov chains, the topics which form the bulk of my dissertation work, can be traced to the papers Professor Zhu suggested to me on Besag's auto- models during my first project.

My parents have always supported me through encouragement and reassurance during challenging times. I wish to sincerely thank them here. I am also grateful to

Hyebin Song for her kindness, patience, and insights while writing this dissertation.

Contents

Chapter 1

Introduction

The Wisconsin Public Land Survey database describes historical forest composition at high spatial resolution, and is of interest in ecological studies of forest composition in Wisconsin just prior to significant Euro-American settlement. For such studies, it is useful to identify recurring subpopulations of tree species known as communities, but standard clustering approaches for subpopulation identification do not account for dependence between spatially nearby observations. For example, Figure 1.1 shows a map of maximum a posteriori (MAP) classifications of forest community types in Wisconsin based on PLS tree species data, which were generated from a finite mixture model fit using maximum likelihood under an assumption of spatial independence. Clearly, the resulting classifications exhibit a high degree of spatial regularity, even though neither the model fitting process nor the forest community classification process assumed or incorporated any type of spatial information.

In Chapter 2, we describe a modeling framework which enables the incorporation of spatial correlation into a finite mixture type model, and we apply our methodology in an analysis of the Wisconsin Public Land Survey dataset. Maps similar to

Figure 1.1 initially led us to consider whether it might be useful to directly include spatial information in the modeling and classification. We hypothesized that including spatial correlation in the model could lead to better classification and parameter estimation by allowing the model to share information across nearby grid cells. Chapter 2 is motivated by this consideration, and is based on work already appearing in Berg et al. [2019b].

In addition to modeling questions, namely, how best to incorporate spatial correlation in a finite mixture model framework, the work in Chapter 2 also involves interesting computational considerations. We give a brief description here. Let $\ell(\eta|y_0)$ denote a likelihood function, where y_0 corresponds to observed data, and $\eta \in \mathbb{R}^p$ is a parameter to be estimated. In many cases, including in Chapter 2, the gradient of the likelihood function involves an integral with respect to a probability density. A familiar example is the case of a canonical exponential family model. In this case, $\ell(\eta|y_0) = p(y_0|\eta) = \exp\{\eta^T T(y_0) - \xi(\eta)\}$ for a sufficient statistic $T(\cdot)$ and normalizing constant $\xi(\eta)$, and we have

$$\frac{\partial \ell}{\partial \eta} = T(y_0) - E\{T(y)|\eta\} \tag{1.1}$$

where $E\{T(y)|\eta\}$ denotes the expected value of T with respect to $p(y|\eta)$ [see, e.g. Shao, 2003] . The expectation $E\{T(y)|\eta\}$ can sometimes be computed analytically. However, in many Markov random field models, including the one in Chapter 2, the only practical way to obtain $E\{T(y)|\eta\}$ is to estimate it via a Markov chain Monte Carlo (MCMC) average [see, e.g., Younes, 1988]. In these cases, we construct an

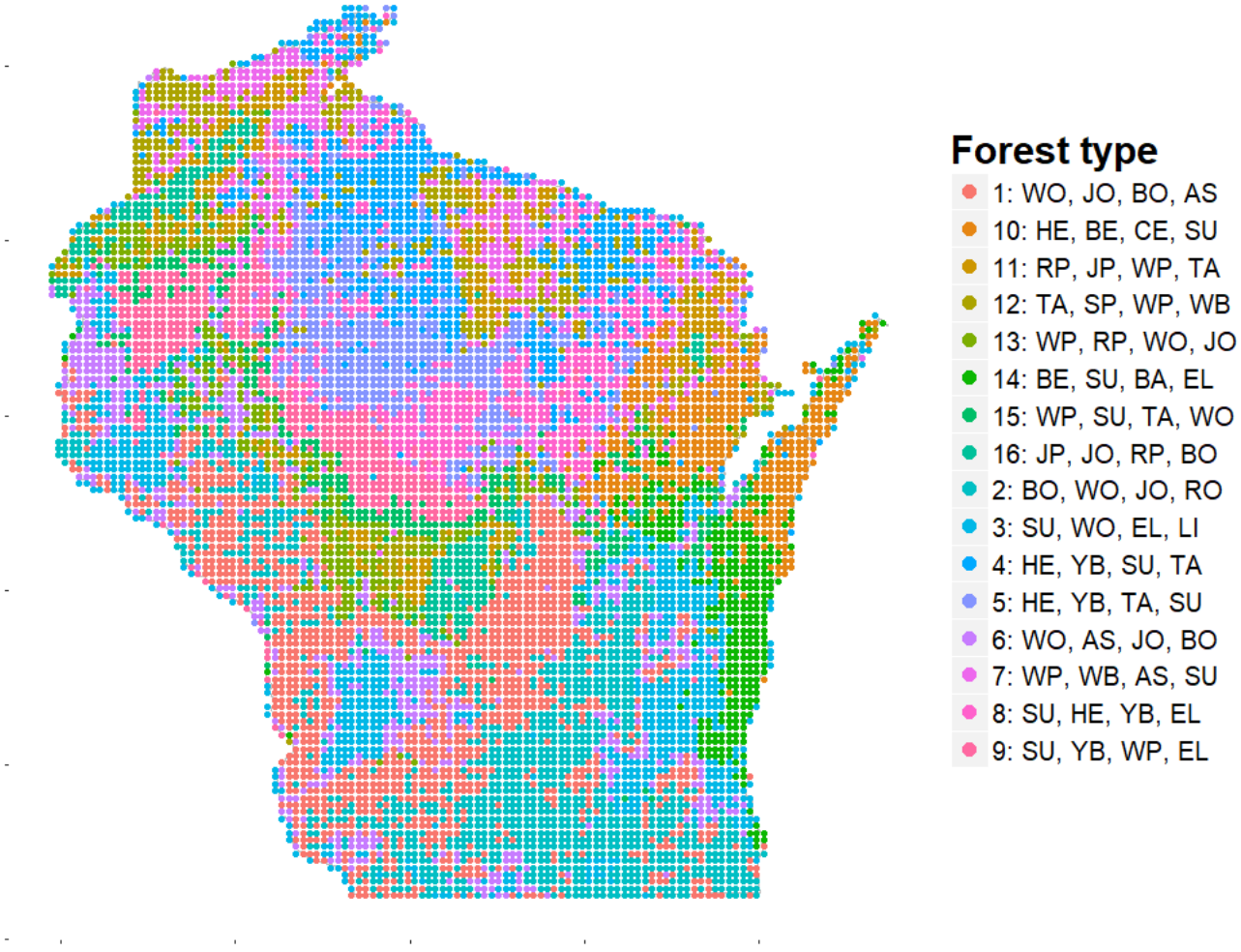


Figure 1.1: Model fitting results from a finite mixture model with 16 forest communities, fit using maximum likelihood under a spatial independence assumption. The four most common tree species within each forest community type are listed. A key to the species abbreviations is given in Table 2.2.

estimator

$$\widehat{E\{T(y)|\eta\}} = M^{-1} \sum_{t=0}^{M-1} T(y_t) \tag{1.2}$$

where $y_0, y_1, y_2, ...$ are consecutive draws from a Markov chain with invariant distribution $p(y|\eta)$. In a similar way, in Chapter 2 we are led to construct an optimization algorithm which involves MCMC estimates of the gradient, in place of the true, intractable gradient. The convergence properties of such algorithms are discussed in, e.g., Kushner and Yin [1997]. Finally, we note that performance measures for these models, even for simple metrics such as holdout likelihoods, often apparently require thoughtfully designed MCMC simulations. We negotiate these issues in Chapter 2.

Chapter 3 of this thesis relates to MCMC averages such as in (1.2). Our goal in Chapter 3, drawing motivation from MCMC stochastic approximation algorithms of the type used in Chapter 2, is to develop methodology to improve the efficiency of MCMC simulations. Bayesian statistical analyses often utilize MCMC to compute quantities such as posterior means and credible intervals, and thus MCMC for Bayesian statistics is a clear application area for the methods in Chapter 3. However, the methodology in Chapter 3 is applicable to MCMC simulations regardless of whether a Bayesian posterior distribution is involved, so in Chapter 3 we typically do not make reference to posterior distributions and other quantities specific to Bayesian statistics.

In Chapter 3, we propose new control variate methodology, with accompanying theoretical guarantees, for deterministic sweep Markov chain sampling. Control variates are a commonly used technique for reducing the variance of ordinary independent Monte Carlo averages as well as MCMC averages [see, e.g. Liu, 2008]. We preface the rigorous discussion of control variates in Chapter 3 with an informal overview. The

idea of control variates is to replace an average

$$M^{-1} \sum_{t=0}^{M-1} X_t, \tag{1.3}$$

where X_t are simulated draws from a Markov chain with stationary distribution π, with a new average of the form

$$M^{-1} \sum_{t=0}^{M-1} (X_t - W_t), \tag{1.4}$$

where we require the control variates W_t to satisfy $\lim_{M\to\infty} M^{-1} \sum_{t=0}^{M-1} W_t = 0$ almost surely. Under the condition $\lim_{M\to\infty} M^{-1} \sum_{t=0}^{M-1} W_t = 0$, the estimators (1.3) and (1.4) have the same limiting value. If the W_t are suitably chosen, then the variance of (1.4) will be reduced relative to (1.3). For example, suppose $W_t = X_t - \mu$, where μ is the expected value of X_0 with respect to the stationary measure π, that is, when $X_0 \sim \pi$. Then $M^{-1} \sum_{t=0}^{M-1} (X_t - W_t) = M^{-1} \sum_{t=0}^{M-1} \mu = \mu$. In this case, the control variate average (1.4) is exact for any finite sample size M. It is typically not possible to construct such effective control variates W_t, but this example shows that suitable W_t can lead to variance reductions. Chapter 3 deals with the construction of control variates W_t in the setting of deterministic sweep Markov chains.

While some of the calculations and regularity conditions in Chapter 3 are technical, the primary conclusions are relatively concrete. For the widely applicable deterministic sweep Gibbs sampler, we show in Chapter 3 that the projection properties of Gibbs kernels lead to a simple, novel control variate estimator, which has theoretical and practical benefits over competing methodology in the literature. In particular, for the data augmentation Gibbs sampler, our control variate estimate is guaranteed to achieve a smaller asymptotic variance than a widely used alternative approach.

Additionally, we provide variance reduction guarantees for a Rao-Blackwellization approach for more general Gibbs sampling settings than those in existing results. We conduct a simulation study which demonstrates that the theoretical benefits of our proposed approaches are realized in practice.

Chapter 2

A Latent Discrete Markov Random Field Approach to Identifying and Classifying Historical Forest Communities Based on Spatial Multivariate Tree Species Counts

2.1 Introduction

This chapter is based on joint work with myself, Jun Zhu, Murray Clayton, Monika Shea, and David Mladenoff, which is published in Berg et al. [2019b].

In this chapter, we consider analyzing historical tree species composition data and mapping forest ecological communities of keen interest in a variety of ecological disciplines, including environmental history and landscape ecology. Sound modeling and analysis of historical vegetation using novel statistical methodology is useful for multiple purposes, including to aid ecological restoration efforts by providing reference landcover information at restoration sites and to assess landscape changes over time [Schulte et al., 2002, Shea et al., 2014]. If an area is known to have historically supported a particular vegetation profile, this could indicate that restoration to the historically supported vegetation type may be more ecologically appropriate [Egan,

2005].

The historical Public Land Survey (PLS) contains informative data for studies of past forest composition. The PLS database for the state of Wisconsin is particularly noteworthy for both its spatial extent (approximately 150,000 km^2) and its high resolution (survey points at roughly half mile intervals across the entire state). The survey was initially conducted to assess land values and facilitate the sale of land, but the collated and digitized PLS data currently provide the only precise, statewide record of the natural ecosystems that were present in Wisconsin just prior to major Euro-American settlement [Schulte and Mladenoff, 2001]. The database is derived from surveyor notebooks from the original U.S. PLS, conducted across the United States from the late 1700's to the early 1900's. The Wisconsin portion of the survey was conducted over 1832–1866 [Liu et al., 2011]. Surveyors demarcated the land into square mile sections, and placed a post as a survey marker at each section corner and each half-mile point. At each survey point, the protocol required that they record several environmental characteristics, including the species of two to four "witness" trees.

Here, we consider the resulting tree species composition data from the Wisconsin PLS, and aggregate the observed tree species counts within an overlaid grid of cells for analysis. An illustration of this type of data is shown in Figure 2.1. We also consider the identification of community subpopulation structure in the PLS relating to recurring assemblages of tree species, which are described in ecological literature as forest communities [Barnes et al., 2010]. Community subpopulations are a common feature of tree species composition data such as in the PLS database. We model forest community subpopulations via the classification of each grid cell with the forest

community type most representative of that cell. Our modeling goal is two-fold. On the one hand, we would like to use tree species composition data to identify discrete assemblages of species corresponding to forest communities in the state of Wisconsin prior to the major environmental disturbances accompanying Euro-American settlement. On the other hand, we would like to classify cells in the survey region with the forest community type they most likely belong to.

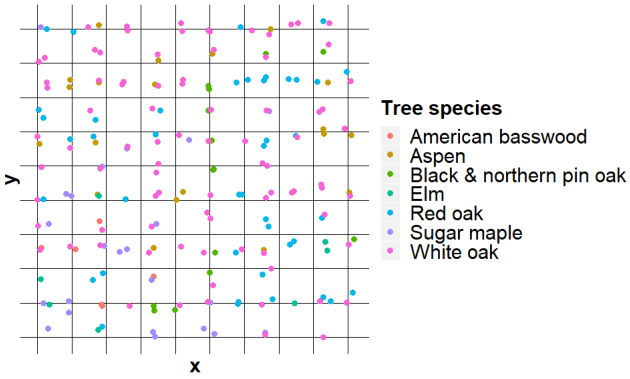


Figure 2.1: Data from a 10-km-by-10-km subregion of the Wisconsin Public Land Survey dataset. The overlaid grid cells are 1-km-by-1-km. Tree species are recorded at multiple survey points within each grid cell.

To achieve our goal of accurately modeling and mapping forest community subpopulations in the PLS survey, we develop an approach wherein forest communities across the survey region are described by discrete, spatially correlated latent variables. The observed tree counts are described by community dependent multinomial distributions. Thus, the observed tree species compositions in the PLS dataset are

assumed to result from a set of multiple underlying forest community types, which occur in a spatially correlated fashion across the survey region. Our approach allows us to describe, indirectly but flexibly, spatial correlation between observations in nearby grid cells, and also to address unobserved structure due to distinct forest community subpopulations.

Our analysis of tree species composition in the PLS dataset is unique among previous literature in that it explicitly accounts both for spatial correlation effects between nearby observations, as well as for latent forest community structure. Tree species composition in the PLS was also studied in Paciorek et al. [2016], using, for example, a latent conditional autoregressive (CAR) model to account for spatial correlation with the goal of providing estimation of tree species composition in the PLS survey region. While the posterior predictions of forest composition in Paciorek et al. [2016] capture spatial covariance between the occurrence of related tree species, these predictions do not explicitly identify or map forest communities. A dissimilarity-based clustering approach as taken in Schulte et al. [2002] allows forest communities to be identified and mapped, but this approach does not explicitly model spatial correlation in the occurrence of forest community types across the study region.

In our work, the tree species composition data come in the form of tree species count vectors, so that the data in each areal unit are multivariate, with counts of zero for absent tree species. As such, in contrast to most work in spatial clustering, our response variable is both multivariate and unordered. We do not constrain the forest community types to appear in spatially contiguous blocks. We also do not expect any ordinal relationship between the forest community types. In certain other common settings, the term "spatial cluster" may refer to a spatially contiguous block

of areal units where a response variable such as disease risk or rate is unusually high relative to other areal units. Frequently, the analysis goal in these settings is to identify "hotspots" of a disease and any associated risk factors [see, e.g., Gangnon and Clayton, 2003, Waller, 2009, Lawson, 2010]. Constraints are also sometimes imposed so that each disease rate cluster only appears in a single contiguous block of areal units [see, e.g., Knorr-Held and Raßer, 2004].

We estimate the parameters of our model via maximum likelihood (ML), and we develop a new Markov chain Monte Carlo (MCMC) stochastic approximation (SA) method to do this. Our MCMC-SA method is related to but differs from the direct expectation-maximization (EM) approach [Dempster et al., 1977]. First, instead of performing the full M-step, we take a gradient step for the Markov random field parameters, and a modified EM step for the other parameters in the model. Such EM algorithms with partial updates in the M-step are sometimes termed generalized EM algorithms [Dempster et al., 1977]. In general, performing the full M-step of the EM algorithm requires inverting between the mean parameterization and the natural parameterization of the complete data distribution [see, e.g., Fort and Moulines, 2003]. In the Markov random field setting, performing this inversion is challenging, and it requires an MCMC sampling step nested within each EM algorithm iteration, as in Forbes and Fort [2007]. We additionally apply regularization penalties to ensure that maxima of our objective function do not occur at the boundary of the parameter space [see, e.g., Städler et al., 2010, Chen, 2017, Hong et al., 2017]. Finally, in our latent Markov random field model, the spatial dependence between the latent forest community types makes computing the loglikelihood challenging, and we use a path integration approach to accurately compute loglikelihoods on holdout data (see, e.g.,

Section 6.2 in Neal 1993, or Gelman and Meng 1998).

While MCMC methods may in general be slow, our MCMC-SA method is feasible even for relatively big data like the PLS dataset due to a computationally efficient implementation of the sampling. Additionally, in the case study of the PLS dataset, we achieve significant improvement in prediction performance using our method relative to an alternative approach that does not account for spatial dependence. A simulation study further shows that our MCMC-SA method can recover the true parameters under the correct model specification and outperform some competing methodology. Though our application in this chapter focuses on identification and classification of forest communities across space, our methodology can be readily modified for use in other ecological community identifications or other settings such as medical image segmentation of tissue types.

The remainder of the chapter is organized as follows. In Section 2.2, we propose a multinomial model with a latent discrete Markov random field for the PLS dataset. In Section 2.3, we develop a maximum likelihood approach to estimate the model parameters and propose a stochastic approximation procedure to compute these estimates. In Section 2.4, we apply our model and estimation method to analyze and interpret the PLS dataset. In Section 2.5, results are presented from a simulation study. We provide an appendix containing additional technical details (Appendix A).

2.2 Model

Our observed data consist of the counts of each tree species within an overlaid grid of cells. We assume that each cell has a latent forest community type with an associated multinomial probability distribution governing the species composition for each

type of forest community. We also assume conditional independence between observed trees given the latent forest community types, which in turn are assumed to follow a Markov random field. Thus, our model is a mixture of multinomial distributions, where the types are spatially correlated.

2.2.1 Latent Model

The spatial grid of cells are assumed to be labeled with one of K possible types, in our case K different forest communities. Corresponding to each grid cell is a spatial neighborhood of adjacent grid cells. We view our approach as agnostic regarding the underlying origin of the spatial dependence in the dataset. For example, an influential environmental covariate may occur in discrete patches across a map, causing certain forest community types to appear or disappear in these areas. Additionally, local within- and between-community interactions may cause spatial patterning on the observed grid. The approach here attempts to mimic and account for the observed spatial correlation structure rather than to exactly replicate the true data generating process.

For notation, we refer to random variables with capitals, and realizations in lowercase. When referring to a probability density for a discrete random vector Z depending on a parameter vector θ, we use the shorthand $p(z|\theta)$ for $p(Z = z|\theta)$. For a vector z, we use z_i to denote the ith entry of z. For a matrix $\mathbf{A}$, we use the notation $\mathbf{A}_j$ to denote the jth column of $\mathbf{A}$, and $\mathbf{A}_{ij}$ to denote the element in the ith row and the jth column of $\mathbf{A}$. We denote the set of spatial neighbors of a cell i by the set $N(i) = \{i' : i' \text{ is a neighbor of cell } i\}$, where the neighbors are defined so that $i \notin N(i)$. We use the notation $i' \sim i$ to indicate that $i' \in N(i)$. Additionally, the

neighborhoods are assumed to be symmetric, so that if cell i is a neighbor of cell i', then cell i' is a neighbor of cell i.

Let n denote the total number of grid cells and $z \in \Omega = \{1, \ldots, K\}^n$ denote a vector of n (unobserved) forest community types. The random vector of forest community types Z is assumed to follow a Potts-type model with a vector of parameters $\eta \in \mathbb{R}^K$:

$$p(z|\eta) = \exp\left\{ \sum_{i=1}^{n} \sum_{k=1}^{K-1} \eta_k I(z_i = k) + \eta_K \sum_{i=1}^{n} \sum_{\substack{i' \in N(i) \\ i' > i}} I(z_i = z_{i'}) - \xi(\eta) \right\} \tag{2.1}$$

where z_i refers to the forest community type for cell i and

$$\xi(\eta) = \sum_{z' \in \Omega} \exp\left\{ \sum_{i=1}^{n} \sum_{k=1}^{K-1} \eta_k I(z'_i = k) + \eta_K \sum_{i=1}^{n} \sum_{\substack{i' \in N(i) \\ i' > i}} I(z'_i = z'_{i'}) \right\}$$

is a normalizing constant ensuring that $p(z|\eta)$ is a probability density [Wu, 1982]. In (2.1), for $k < K$, the parameter η_k controls the probability of the kth type relative to the baseline type K. The spatial correlation parameter η_K controls the strength of interactions between the types and when $\eta_K = 0$, the types are spatially independent across grid cells.

We define a length K vector $T(z)$ of sufficient statistics with the kth entry

$$T(z)_k = \begin{cases} \sum_{i=1}^{n} I(z_i = k); & (k < K) \\ \sum_{i=1}^{n} \sum_{\substack{i' \in N(i) \\ i' > i}} I(z_i = z'_i); & (k = K) \end{cases} \tag{2.2}$$

This allows us to rewrite the model (2.1) more succinctly as

$$p(z|\eta) = \exp\left\{ \eta^T T(z) - \xi(\eta) \right\},$$

which belongs to the exponential family with the natural parameter vector η [Shao, 2003].

For boundary conditions in lattice data models, there are several approaches to specifying the neighborhood of the cells on the boundary of the lattice. We use the so-called "free" boundary conditions, where boundary cells simply have fewer neighbors than internal cells [see, e.g., Comets and Gidas, 1992]. Other approaches attempt to ensure that each cell has the same number (usually 4, for the square lattice) of neighbors. For example, in "toroidal" boundary conditions, cells on one side or corner of the lattice are connected to cells on the opposing side or corner of the lattice.

2.2.2 Data Model

Given the forest community types, we specify our model for the conditional distribution of the observed tree species counts. For notation, we let the integer $M > 0$ denote the number of tree species in the dataset. For the PLS case study, the $M = 33$ most common species are used. We denote by $\mathbf{Y}_i$ the length M vector of tree counts in cell i, and use $\mathbf{Y} \in \mathbb{Z}^{M \times n}$ to denote the matrix of count vectors for the entire dataset. Thus, $\mathbf{Y}_{mi}$ is the count of trees of species m in cell i. We let q_i denote the total number of trees observed within the ith cell. That is, $q_i = \sum_{m=1}^{M} \mathbf{Y}_{mi}$.

We assume that each of the K forest community types is associated with a distinct multinomial distribution over the M tree species. Conditional on the latent type $Z_i = k$, the tree species of individual trees within a grid cell are assumed to be independent multinomials with sample size 1, so that the count vector $\mathbf{Y}_i$ follows a multinomial distribution with sample size q_i and species probability parameters depending on the kth forest community type. Additionally, the species of individual

trees are assumed to be independent across grid cells and thus, the counts $\mathbf{Y}_i$ are also independent across grid cells, both conditional on the latent forest community types. However, when the spatial correlation parameter $\eta_K \neq 0$, the latent types are spatially correlated, which induces spatial correlation among the tree counts $\mathbf{Y}_i$.

We parameterize the species distribution for each forest community type k using a species probability vector $\boldsymbol{\mu}_k \in \mathscr{M}$, where $\mathscr{M}$ refers to the (open) probability simplex defined by $\mathscr{M} = \{\mu = [\mu_1, \ldots, \mu_M]^T : \sum_m \mu_m = 1; \mu_m > 0, \forall m\}$. We also define the species probability matrix $\boldsymbol{\mu}$ with column vectors $\boldsymbol{\mu}_k$ by $\boldsymbol{\mu} = \begin{bmatrix}\boldsymbol{\mu}_1 & \boldsymbol{\mu}_2 & \ldots & \boldsymbol{\mu}_K\end{bmatrix} \in \mathscr{M}^K$. The $\boldsymbol{\mu}_{mk}$ element of the $\boldsymbol{\mu}$ matrix is equal to the probability that a tree in a grid cell is species m, given that the forest community type of that grid cell is k.

By the conditional independence of tree species between and within grid cells, the conditional density of the observed tree counts given the forest community types Z and the species probability matrix $\boldsymbol{\mu}$ is

$$p(\mathbf{y}|z, \boldsymbol{\mu}) \;=\; \prod_{i=1}^{n} p(\mathbf{y}_i|z_i, \boldsymbol{\mu}) = \prod_{i=1}^{n} C_i \prod_{m=1}^{M} \boldsymbol{\mu}_{m,z_i}^{\mathbf{y}_{mi}} \tag{2.3}$$

where $\boldsymbol{\mu}_{m,z_i}$ is the mth entry of column z_i of $\boldsymbol{\mu}$, and the factor $C_i = \left(\prod_{m=1}^{M} \mathbf{y}_{mi}!\right)^{-1} q_i!$ counts the number of possible ways of assigning species to each tree in the ith grid cell with the species counts $\mathbf{y}_i$.

In summary, our full data generating mechanism comprises two steps:

1. Draw the forest community types Z according to the density in (2.1).

2. Conditioning on the forest community types $Z = z$ from step 1, draw the tree species counts $\mathbf{Y}$ according to the density in (2.3).

Define $R(\mathbf{y}, z) \in \mathbb{R}^{M \times K}$ to be a matrix of statistics with the (m, k)th element $R(\mathbf{y}, z)_{mk} =$

$\sum_{i=1}^{n} \mathbf{y}_{mi} I(z_i = k)$ summarizing the total number of species m trees in the grid cells that belong to the kth type of forest community. Then, the complete data density for $(\mathbf{Y}, Z)$ is

$$\begin{aligned} p(\mathbf{y}, z|\eta, \boldsymbol{\mu}) &= p(z|\eta)p(\mathbf{y}|z, \boldsymbol{\mu}) \\ &= \exp\left\{ \eta^T T(z) - \xi(\eta) + \sum_{m=1}^{M}\sum_{k=1}^{K} \log(\boldsymbol{\mu}_{mk}) R_{mk}(\mathbf{y}, z) + \sum_{i=1}^{n} \log(C_i) \right\}, \end{aligned} \tag{2.4}$$

It is sometimes convenient to write the parameter vector η and the parameter matrix $\boldsymbol{\mu}$ using a single vector parameter θ. Conversely, we may also need to obtain η and $\boldsymbol{\mu}$ from the corresponding vector θ. Thus, we define a vectorization operator $\vec{(}A) : \mathbb{R}^{M\times K} \to \mathbb{R}^{MK}, \boldsymbol{\mu} \to \left[\boldsymbol{\mu}_1^T, \ldots, \boldsymbol{\mu}_K^T\right]^T$ for viewing the parameter matrix $\boldsymbol{\mu}$ as a vector. Then, we define $\theta \in \mathbb{R}^{K+MK}$ by $\theta = \left[\eta^T, \vec{(}\boldsymbol{\mu})^T\right]^T$. We use Θ to denote the parameter space for θ, and we use $\eta(\theta) \in \mathbb{R}^K$ and $\boldsymbol{\mu}(\theta) \in \mathbb{R}^{M\times K}$ to denote the η and $\boldsymbol{\mu}$ associated with θ. When it is clear, we simply write η or $\boldsymbol{\mu}$ rather than $\eta(\theta)$ or $\boldsymbol{\mu}(\theta)$.

2.3 Method

2.3.1 Maximum Regularized Likelihood Estimation

Here, we estimate the parameter θ via maximum likelihood. For the model described in (2.1) and (2.3), the observed data log-likelihood when $\mathbf{Y} = \mathbf{y}$ is

$$\ell(\theta) = \log\left\{ \sum_{z\in\Omega} p(\mathbf{y}, z|\theta) \right\}. \tag{2.5}$$

The consistency of the maximum likelihood estimate, $\hat{\theta} = \text{argmax}_\theta \ell(\theta)$, is shown in an increasing domain asymptotics setting, under identifiability assumptions on θ and when the estimation is constrained to a compact parameter space [Comets and Gidas, 1992].

The observed data log-likelihood (2.5) may exhibit unwanted behavior, such as maxima on the boundary of the parameter space, which is common in latent variable models. Such behavior can occur even in simple settings, such as a mixture of normal densities with component-specific variances, where it is possible to achieve an arbitrarily high likelihood by setting the mean of one of the components to a single data point, and sending the variance of that component toward 0 [see, e.g., Chen, 2017, Section 3.2]. In our work, there is apparent convergence of entries of the tree species probability matrix, $\boldsymbol{\mu}_{mk}$, to 0, which seems to occur mostly for the rarer tree species, while there is no observed convergence of components of the parameters associated with the forest community types, η, to the boundary of $\mathbb{R}^K$.

To guarantee the convergence of our estimation procedures to points within the parameter space Θ, we impose weakly informative prior penalties on the observed data log-likelihood (2.5) [see, e.g., Städler et al., 2010, Chen, 2017, Hong et al., 2017]. In particular, Kushner and Yin [1997] added "soft penalties" to ensure that the objective function is well-behaved and the iterates from a stochastic approximation procedure remain bounded, which we use here to optimize a regularized log-likelihood function:

$$\ell_{pen}(\theta) = \ell(\theta) + \rho_1(\eta) + \rho_2(\boldsymbol{\mu}), \tag{2.6}$$

where $\rho_1(\eta)$ and $\rho_2(\boldsymbol{\mu})$ are penalty functions.

For each component of η, we apply a Logistic$(0, \sigma)$ prior density with $\sigma > 0$. That is, $\rho_1(\eta) = \sum_{k=1}^{K} \log f_\sigma(\eta_k)$, where for $k = 1, \ldots, K$,

$$f_\sigma(\eta_k) = \sigma^{-1} \left[\exp\left\{\eta_k/(2\sigma)\right\} + \exp\left\{-\eta_k/(2\sigma)\right\}\right]^{-2}. \tag{2.7}$$

In the PLS case study and the simulation studies, we use $\sigma = 1$.

For each column of $\boldsymbol{\mu}$, we put a Dirichlet($\alpha 1_M$) prior with $\alpha > 1$, where 1_M is a vector of M 1's, so that

$$\rho_2(\boldsymbol{\mu}) = (\alpha - 1) \sum_{k=1}^{K} \sum_{m=1}^{M} \log(\boldsymbol{\mu}_{mk}). \tag{2.8}$$

For an integer $\alpha > 1$, $\rho_2(\boldsymbol{\mu})$ can be viewed as adding to the dataset some pseudo-data corresponding to $\alpha - 1$ grid cells for each forest community type, in which 1 tree from each of the M species is observed. We use $\alpha = 2$ as the regularization parameter.

2.3.2 Modified EM Algorithm

To optimize the penalized likelihood in (2.6), we derive a modified EM algorithm. The computations required by both the η and $\boldsymbol{\mu}$ updates involve expectations over all possible K^n configurations of the forest community types. When the spatial correlation parameter $\eta_K \neq 0$, we approximate the exact updates by a stochastic procedure, which we describe in Section 2.3.3. When the spatial correlation parameter $\eta_K = 0$, it is possible to compute the expectations exactly, and we derive the EM updates for the spatially independent model in Section A.2.1 of A.

Since the forest community types are unobserved, the problem of estimating the parameters $\theta = \left[\eta^T, \vec{(\boldsymbol{\mu})}^T\right]^T$ falls naturally into the missing data framework and the expectation-maximization (EM) algorithm is a possible solution [Dempster et al., 1977]. In each iteration of our modified EM algorithm, a surrogate function is con-

structed in the E-step, based on the current parameter value θ^{cur}:

$$\begin{aligned} Q(\theta|\theta^{cur}) &= \rho_1(\eta) + \rho_2(\boldsymbol{\mu}) + \sum_{z\in\Omega} p(z|\mathbf{y},\theta^{cur}) \log\{p(z,\mathbf{y}|\theta)\} \\ &= \left[\rho_1(\eta) + \sum_{z\in\Omega} p(z|\mathbf{y},\theta^{cur}) \log\{p(z|\eta)\}\right] \\ &+ \left[\rho_2(\boldsymbol{\mu}) + \sum_{z\in\Omega} p(z|\mathbf{y},\theta^{cur}) \log\{p(\mathbf{y}|z,\boldsymbol{\mu})\}\right] \\ &\equiv Q_1(\eta|\theta^{cur}) + Q_2(\boldsymbol{\mu}|\theta^{cur}). \end{aligned} \tag{2.9}$$

In the M-step, the parameter value θ^{new} for the next iteration is obtained by maximizing the Q-function over θ . This process is repeated iteratively by setting $\theta^{cur} = \theta^{new}$ and then maximizing the new Q-function again. Under suitable conditions, any limit point of such an EM algorithm is guaranteed to be a stationary point of the loglikelihood [Wu, 1983].

The surrogate Q-function (2.9) takes an average over the complete data loglikelihood $\log\{p(z,\mathbf{y}|\theta)\}$ with respect to the conditional distribution $p(z|\mathbf{y},\theta^{cur})$ of the types, given the observed data $\mathbf{y}$ and evaluated at θ^{cur}, whereas the penalty functions $\rho_1(\eta)$ and $\rho_2(\boldsymbol{\mu})$ remain unchanged. The Q-function (2.9) can also be shown to minorize the regularized loglikelihood (2.6), in the sense that

$$\ell_{pen}(\theta) - \ell_{pen}(\theta^{cur}) > Q(\theta|\theta^{cur}) - Q(\theta^{cur}|\theta^{cur}).$$

Thus, increasing the value of the Q-function guarantees an even greater increase in the value of the regularized loglikelihood (2.6). The implementation detail for maximizing the Q-function is given as follows.

Update η: First, we deal with the maximization of the Q_1-function in (2.9):

$$Q_1(\eta|\theta^{cur}) = \sum_{z\in\Omega} p(z|\mathbf{y},\theta^{cur}) \log\{p(z|\eta)\} + \rho_1(\eta). \tag{2.10}$$

Since $p(z|\eta)$ is in the exponential family with sufficient statistic $T(z)$, we have $\partial \log\{p(z|\eta)\}/\partial\eta = T(z) - E\{T(z')|\eta\}$ [Shao, 2003] and

$$\begin{aligned}
&\left.\frac{\partial Q_1(\eta|\theta^{cur})}{\partial\eta}\right|_{\eta=\eta^{cur}} \\
&= \left.\frac{\partial \rho_1(\eta)}{\partial\eta}\right|_{\eta=\eta^{cur}} + \sum_{z'\in\Omega} p(z'|\mathbf{y},\theta^{cur})\,[T(z') - E\{T(z)|\eta^{cur}\}] \\
&= \left.\frac{\partial \rho_1(\eta)}{\partial\eta}\right|_{\eta=\eta^{cur}} + E\{T(z)|\mathbf{y},\theta^{cur}\} - E\{T(z)|\eta^{cur}\} \qquad (2.11)
\end{aligned}$$

Thus, the gradient of $Q_1(\eta|\theta^{cur})$ has a convenient representation in terms of the conditional and marginal distributions at $\theta = \theta^{cur}$. Furthermore, it can be shown that $\left.\frac{\partial \ell_{pen}(\theta)}{\partial\eta}\right|_{\theta=\theta^{cur}} = \left.\frac{\partial Q_1(\eta|\theta^{cur})}{\partial\eta}\right|_{\eta=\eta^{cur}}$.

Finding the η which maximizes $Q_1(\eta|\theta^{cur})$ in the M-step would require inverting, at every iteration, between the exponential family natural parameter, η, and the exponential family mean parameter, $E\{T(z)|\eta\}$. For Markov random field models, this inversion would require a sequence of MCMC draws and is a challenging computational problem [Forbes and Fort, 2007]. Thus, we elect to instead use a gradient ascent update for the η component of θ:

$$\eta^{new} = \eta^{cur} + c^{-1}\frac{\partial Q_1(\eta|\theta^{cur})}{\partial\eta} = \eta^{cur} + c^{-1}\frac{\partial \ell_{pen}(\theta)}{\partial\eta}, \qquad (2.12)$$

where c is a fixed constant stepsize chosen to ensure reasonable convergence behavior.

Update $\boldsymbol{\mu}$: In contrast to η, the update for $\boldsymbol{\mu}$ has a convenient representation in terms

of the conditional distribution $p(z|\mathbf{y}, \theta^{cur})$, because by (2.3), we have

$$\begin{aligned} Q_2(\boldsymbol{\mu}|\theta^{cur}) &= \rho_2(\boldsymbol{\mu}) + \sum_{z\in\Omega} p(z|\mathbf{y}, \theta^{cur}) \log\{p(\mathbf{y}|z, \boldsymbol{\mu})\} \\ &= \rho_2(\boldsymbol{\mu}) + \sum_{z\in\Omega} p(z|\mathbf{y}, \theta^{cur}) \sum_{i=1}^{n}\sum_{k=1}^{K} \left\{\log(\boldsymbol{\mu}_k)^T \mathbf{y}_i\right\} I(z_i = k) + \sum_{i=1}^{n} \log(C_i) \\ &= \rho_2(\boldsymbol{\mu}) + \sum_{i=1}^{n}\sum_{k=1}^{K} \left\{\log(\boldsymbol{\mu}_k)^T \mathbf{y}_i\right\} P(z_i = k|\mathbf{y}, \theta^{cur}) + \sum_{i=1}^{n} \log(C_i) \\ &= \sum_{k=1}^{K} Q_2^k(\boldsymbol{\mu}_k) + \sum_{i=1}^{n} \log(C_i), \end{aligned} \tag{2.13}$$

where $\sum_{i=1}^{n} \log(C_i)$ does not depend on $\boldsymbol{\mu}$ and

$$Q_2^k(\boldsymbol{\mu}_k) = \sum_{m=1}^{M} (\alpha - 1) \log(\boldsymbol{\mu}_{mk}) + \sum_{i=1}^{n}\sum_{m=1}^{M} P(z_i = k|\mathbf{y}, \theta^{cur}) \mathbf{y}_{mi} \log(\boldsymbol{\mu}_{mk}).$$

It is shown in Section A.2.2 of Appendix A that the maximizer $\boldsymbol{\mu}^{new}$ of (2.13) has entries

$$\boldsymbol{\mu}_{mk}^{new} = \{\alpha - 1 + N_{mk}\} / \{M(\alpha - 1) + N_k\}, \tag{2.14}$$

where $N_{mk} = \sum_{i=1}^{n} P(z_i = k|\mathbf{y}, \theta^{cur}) \mathbf{y}_{mi}$ and $N_k = \sum_{m=1}^{M} N_{mk}$.

In a standard EM update for $\boldsymbol{\mu}$, we have $\boldsymbol{\mu}_k^{new} = \boldsymbol{\mu}_k^{cur} + (\boldsymbol{\mu}_k^{new} - \boldsymbol{\mu}_k^{cur})$. Here, it is more convenient to use an altered version, because $p(z|\mathbf{y}, \theta^{cur})$ is known only up to a normalizing constant and the $\boldsymbol{\mu}$ update must be approximated by MCMC. The quantities related to $p(z|\mathbf{y}, \theta^{cur})$ appear in both the numerator and denominator of (2.14) and thus, it is challenging to estimate the EM update for $\boldsymbol{\mu}$ in an unbiased fashion based only on a single draw from $p(z|\mathbf{y}, \theta^{cur})$. Thus, we propose a "short-step" for updates:

$$\widetilde{\boldsymbol{\mu}}_k^{new} = \boldsymbol{\mu}_k^{cur} + \gamma_k(\boldsymbol{\mu}_k^{new} - \boldsymbol{\mu}_k^{cur}), \tag{2.15}$$

where

$$\gamma_k = \{M(\alpha-1)+N_k\}\,/\{M(\alpha-1)+\sum_{i=1}^{n} q_i\}. \tag{2.16}$$

Since $N_k < \sum_{i=1}^n q_i$ for all $\theta \in \Theta$, we have $\gamma_k < 1$ for all θ^{cur}. On the other hand, $\gamma_k \geq \{M(\alpha-1)\}/\{M(\alpha-1)+\sum_{i=1}^n q_i\} > 0$. Thus, the $\widetilde{\boldsymbol{\mu}}^{new}$ update results from taking a shortened EM step starting from $\boldsymbol{\mu}^{cur}$. For the product $\gamma_k \boldsymbol{\mu}_k^{new}$ in (2.15), the numerator of γ_k cancels with the denominator of $\boldsymbol{\mu}_k^{new}$ in (2.14). The denominator of γ_k depends on the number of tree species M, the regularization parameter α, and the number of trees in the dataset $\sum_{i=1}^n q_i$. Thus, $\gamma_k \boldsymbol{\mu}_k^{new}$ depends on $p(z|\mathbf{y}, \theta^{cur})$ only through the numerator of the $\boldsymbol{\mu}_k^{new}$ update in (2.14), which can be estimated based on a single draw from $p(z|\mathbf{y}, \theta)$ (see Section 2.3.3).

The set $\mathscr{M}$ is convex, and from (2.14), $\boldsymbol{\mu}^{new} \in \mathscr{M}^K$. From convexity, when $\boldsymbol{\mu}^{new} \in \mathscr{M}^K$ and $\boldsymbol{\mu}^{cur} \in \mathscr{M}^K$, (2.15) implies $\widetilde{\boldsymbol{\mu}}^{new} \in \mathscr{M}^K$ as well. Additionally, the update in (2.15) preserves the ascent property of the EM algorithm. By concavity of the log function, Q_2^k is concave, so that $Q_2^k(\widetilde{\boldsymbol{\mu}}_k^{new}) \geq \gamma_k Q_2^k(\boldsymbol{\mu}_k^{new}) + (1-\gamma_k) Q_2^k(\boldsymbol{\mu}_k^{cur}) \geq Q_2^k(\boldsymbol{\mu}_k^{cur})$. When $\boldsymbol{\mu}_k^{new} \neq \boldsymbol{\mu}_k^{cur}$, the inequalities are strict. Since $Q_1(\eta|\theta^{cur}) + Q_2(\boldsymbol{\mu}|\theta^{cur})$ minorizes $\ell_{pen}(\theta)$, any increase in the value of Q_2 implies an increase in the value of $\ell_{pen}(\theta)$.

2.3.3 Stochastic Approximation

To update θ, we devise a stochastic approximation procedure $\theta^{new} = \theta^{cur} + g(\theta^{cur})$, where

$$g(\theta^{cur}) = g\left(\begin{bmatrix} \eta^{cur} \\ \vec{(}\boldsymbol{\mu}^{cur}) \end{bmatrix}\right) = \begin{bmatrix} C^{-1}\frac{\partial Q_1(\eta|\theta^{cur})}{\partial \eta}\big|_{\eta=\eta^{cur}} \\ \vec{(}\widetilde{\boldsymbol{\mu}}^{new}) - \vec{(}\boldsymbol{\mu}^{cur}), \end{bmatrix} \tag{2.17}$$

and $g(\theta)$ is to be estimated based on MCMC samples. Stochastic approximation approaches are useful when the function $g(\cdot)$ is difficult or impossible to evaluate, but

$g(\cdot)$ can be approximated by an estimate $G(\theta, \mathbf{z})$, where $\mathbf{z}$ is a random variable drawn from a distribution π_θ, such that π_θ and $G(\cdot, \cdot)$ satisfy, for each θ,

$$\int G(\theta, \mathbf{z})\pi_\theta(d\mathbf{z}) = g(\theta) \tag{2.18}$$

[see, e.g., Robbins and Monro, 1951, Benveniste et al., 1990, Kushner and Yin, 1997].

The update in (2.17) is a combination of a gradient ascent update for η and a short-step update for $\boldsymbol{\mu}$, from which an iterate sequence may be constructed in the following way. Starting from an initial parameter $\theta^{(0)}$ and initial $\mathbf{z}^{(0)}$, we obtain draws $\mathbf{z}^{(t+1)}$ from $\pi_{\theta^{(t)}}(\cdot)$ and set $\theta^{(t+1)} = \theta^{(t)} + \epsilon^{(t+1)}G(\theta^{(t)}, \mathbf{z}^{(t+1)})$. The sequence of stepsizes $\{\epsilon^{(t)}\}$ is deterministic and generally satisfies conditions such as $\epsilon^{(t)} \downarrow 0$, $\sum_{t=1}^{\infty}(\epsilon^{(t)})^2 < \infty$, and $\sum_{t=1}^{\infty}\epsilon^{(t)} = \infty$ [see, e.g., Benveniste et al., 1990, Kushner and Yin, 1997]. Here, we use $\epsilon^{(t)} = t^{-1}$.

From (2.11), the gradient of $Q_1(\eta|\theta^{cur})$, and hence the gradient of the observed loglikelihood, with respect to the η parameter can be computed based on the difference of two expectations of $T(z)$. The first expectation is taken with respect to the conditional distribution $p(z|\mathbf{y}, \theta^{cur})$ and the second expectation is taken with respect to the marginal distribution $p(z|\eta^{cur})$, while the gradient of the logistic prior $\rho_1(\eta)$ can be computed analytically. The $\boldsymbol{\mu}$ update $\widetilde{\boldsymbol{\mu}}^{new} - \boldsymbol{\mu}^{cur}$ in (2.15) can be computed by taking the expectation of the function

$$\begin{aligned} &H_\alpha(\boldsymbol{\mu}^{cur}, z)_{mk} \\ &= \frac{(\alpha-1)(1 - M\boldsymbol{\mu}_{mk}^{cur}) + \sum_{i=1}^{n} I\{z_i = k\}(\mathbf{y}_{mi} - q_i\boldsymbol{\mu}_{mk}^{cur})}{M(\alpha-1) + \sum_{i=1}^{n} q_i}, \end{aligned} \tag{2.19}$$

with respect to $p(z|\mathbf{y}, \theta)$. Thus, the update $g(\theta^{cur})$ in (2.17) can be written as an integration with respect to the density

$$\pi_{\theta^{cur}}(\mathbf{z}) = p(\mathbf{z}_1|\mathbf{y}, \theta^{cur})p(\mathbf{z}_2|\eta^{cur}), \tag{2.20}$$

where $\mathbf{z} = (\mathbf{z}_1, \mathbf{z}_2)$ denotes an ordered pair of configurations $\mathbf{z}_1, \mathbf{z}_2 \in \Omega$. From (2.20), $\mathbf{z}_1$ and $\mathbf{z}_2$ are drawn independently under $\pi_{\theta^{cur}}(\cdot)$. The probability density $\pi_{\theta^{cur}}(\cdot)$ takes as its argument an element $\mathbf{z}$ of the sample space $\Omega^2 = \Omega \times \Omega$.

Since integrals with respect to $\pi_\theta(\cdot)$ require intractable sums over all possible type configurations $\mathbf{z}_1$ and $\mathbf{z}_2$, we estimate $g(\theta)$ based on approximate draws from $\pi_\theta(\cdot)$. While it is difficult to sample directly from $p(\mathbf{z}_1|\mathbf{y}, \theta)$ and $p(\mathbf{z}_2|\eta)$ due to the spatial correlation between the types z_i, it is possible to use Markov chain transition kernels (specifically, Gibbs sampling kernels) to approximate draws from these distributions [Geman and Geman, 1984]. Due to the conditional independence between grid cells in the conditional distribution $p(\mathbf{y}|z, \boldsymbol{\mu})$, both $p(z|\mathbf{y}, \theta)$ and $p(z|\eta)$ are Markov random field densities with simple conditional distributions at each cell given the rest of the cells. Thus, it is possible to construct Gibbs sampling transition kernels $P_{1,\theta}(\cdot, \cdot) : \Omega \times \Omega \to [0, 1]$ and $P_{2,\theta}(\cdot, \cdot) : \Omega \times \Omega \to [0, 1]$ so that the stationary distributions of $P_{1,\theta}(\cdot, \cdot)$ and $P_{2,\theta}(\cdot, \cdot)$ are $p(z|\mathbf{y}, \theta)$ and $p(z|\eta)$, respectively. In order to approximately sample from π_θ, we run a Markov chain using the transition kernel $P_\theta(\mathbf{z}, \mathbf{z}') : \Omega^2 \times \Omega^2 \to [0, 1]$ defined by $P_\theta(\mathbf{z}, \mathbf{z}') = P_{1,\theta}(\mathbf{z}_1, \mathbf{z}_1')P_{2,\theta}(\mathbf{z}_2, \mathbf{z}_2')$. Detailed formulas for the Gibbs samplers are given in Section A.1.1 of Appendix A [see also, e.g., Gaetan and Guyon, 2010]. From the definition of $P_\theta(\cdot, \cdot)$, we see that the updates to the Markov chain for the conditional distribution $p(z|\mathbf{y}, \theta)$ are independent from the updates to the Markov chain for the marginal distribution $p(z|\eta)$.

In our stochastic modified EM procedure, we choose a stepsize c and define the function $G(\cdot, \cdot) : \Theta \times \Omega^2 \to \Theta$ by

$$G(\theta, \mathbf{z}) \;=\; G\left(\begin{bmatrix} \eta \\ \vec{(}\boldsymbol{\mu}) \end{bmatrix}, \mathbf{z}\right) = \begin{bmatrix} c\left\{\frac{\partial \rho_1(\eta)}{\partial \eta} + T(\mathbf{z}_1) - T(\mathbf{z}_2)\right\} \\ H_\alpha(\boldsymbol{\mu}, \mathbf{z}_1) \end{bmatrix} \tag{2.21}$$

Then, we find parameters $\hat{\theta}$ maximizing the penalized likelihood $\ell_{pen}(\theta)$ via the procedure given in Algorithm 1.

Algorithm 1: Stochastic Modified EM

Initialize parameter $\theta_0 \in \Theta$, configuration $\mathbf{z}_0 \in \Omega^2$, number of iterations T;

for *t=1* **to** T **do**

  Draw $\mathbf{z}_t \in \Omega^2$ according to $P_{\theta_{t-1}}(\mathbf{z}_{t-1}, \cdot)$

  $\epsilon_t = t^{-1}$

  $\theta_t = \theta_{t-1} + \epsilon_t G(\theta_{t-1}, \mathbf{z}_t)$

Return θ_T

Implementation details, including a discussion of the choice of the stepsize c, are given in Section A.1.2 of Appendix A.

2.4 Case Study: Historical Forest Communities based on Public Land Survey Data

The Wisconsin PLS dataset is a historical survey of trees, conducted primarily between 1832–1866 [Schulte and Mladenoff, 2001]. The dataset has been commonly used in ecological studies of forest composition prior to and concurrent with Euro-American settlement. As described in the introduction, surveyors from the PLS walked along a 1-mile-by-1-mile grid-like pattern across the state, and recorded the species of 2–4 representative trees at survey points every half-mile (Figure 2.1). The dataset is large, both in terms of the number of trees observed (328,499), distributed roughly uniformly across the state, as well as in terms of the spatial extent (145,000 square kilometers). Additionally, the tree species count data at each grid cell are highly multivariate and for small enough grid cells, most tree species counts are 0, since only 2-4 trees were observed at each survey point and the survey points are at least half a

mile away from each other.

For data analysis, a spatial grid of cells is first overlaid on the survey region (the state of Wisconsin). For each grid cell i, a count vector $\mathbf{y}_i$, of length $M = 33$ species, is constructed from the trees observed at survey points within that cell. Grid cells are not required to contain any trees. For our spatially correlated model, the forest community type probability at any cell takes into account tree information from nearby adjacent and non-adjacent grid cells containing trees. We compare three grid resolutions: 4km-by-4km, 2-km-by-2-km, and 1km-by-1km grid cells, resulting in 9,469 cells, 37,134 cells, and 146,851 cells, respectively. For each grid resolution, each cell is assumed to have a single forest community type. We use a first-order spatial neighborhood structure with up to four nearest neighbors. That is, two points with integer lattice coordinates (i, j) and (i', j') are neighbors when $|i - i'| + |j - j'| = 1$. Next, we fit the spatially correlated multinomial mixture models via the stochastic modified EM procedure in Algorithm 1. For comparison, we fit spatially independent multinomial mixture models via the standard EM algorithm, a derivation of which is given in Section A.2.1 of Appendix A.

2.4.1 Choice of K and Model Validation

We use a cross-validation procedure to determine the number of forest community types to use, and to assess the quality of the spatially correlated mixture models relative to the spatially independent mixture models. In particular, we generate a testing dataset by randomly selecting 20 percent of the trees from the full set of surveyed trees. The remaining 80 percent of the trees are placed in a training dataset. We then create training and testing datasets $\mathbf{y}_{train}$ and $\mathbf{y}_{test}$ for each grid resolution

(1km, 2km, and 4km) from these training and testing trees. We also consider five total numbers of forest community types $K = 8, 12, 16, 20$, or 24. For each combination of grid resolution and number of forest community types, we fit each of the models on the training dataset $\mathbf{y}_{train}$, starting from three random initial parameter values to mitigate the multimodality of the likelihood.

We examine two loglikelihood based measures of prediction performance, using the same training and testing datasets across models fit for different grid resolutions and numbers of forest community types to ensure the likelihoods are comparable among different models. We first compute, for each of the fitted models at each grid resolution, a holdout loglikelihood

$$\ell_{holdout}(\hat{\theta}) = \log\left\{\sum_{z\in\Omega} p(\mathbf{y}_{test}|z,\hat{\theta})p(z|\hat{\theta})\right\}. \tag{2.22}$$

We focus our model assessment on holdout loglikelihoods rather than on the errors of estimated coefficients, because the true data-generating parameters are unknown for the real data. Next, we compute a predictive loglikelihood

$$\ell_{pred}(\hat{\theta}) = \log\{p(\mathbf{y}_{test}|\mathbf{y}_{train},\hat{\theta})\} = \log\left\{\sum_{z\in\Omega} p(\mathbf{y}_{test}|z,\hat{\theta})p(z|\mathbf{y}_{train},\hat{\theta})\right\}. \tag{2.23}$$

In contrast to the holdout loglikelihood $\ell_{holdout}(\hat{\theta})$ that is marginal on the testing dataset, the predictive loglikelihood $\ell_{pred}(\hat{\theta})$ measures the quality of predictions of the testing dataset, conditional on the training dataset. Since our maps of the study area are ultimately based on the conditional distribution $p(z|\mathbf{y},\hat{\theta})$, the predictive loglikelihood is a relevant performance metric. To ensure that these likelihoods are comparable across different grid resolutions, we drop the grid-resolution dependent factors C_i in (2.3). The loglikelihoods without the constants C_i are equal to the loglikelihoods of the

individual trees, before being aggregated into counts. Unlike the tree species counts $\mathbf{y}_i$ for each cell, which vary by the grid resolution, the loglikelihood of the collection of individual trees has the same interpretation across grid resolutions.

For the spatially correlated models, the holdout loglikelihood in (2.22) is difficult to compute, and we use path integration, also known as thermodynamic integration or path sampling [Neal, 1993, Gelman and Meng, 1998]. We describe the path integration procedure in Section A.1.3 of Appendix A. The predictive loglikelihood (2.23) is also difficult to compute for the spatially correlated models. By the fact that

$$\begin{aligned}\log\{p(\mathbf{y}_{test}|\mathbf{y}_{train},\hat{\theta})\} &= \log\{p(\mathbf{y}_{train},\mathbf{y}_{test}|\hat{\theta})\} - \log\{p(\mathbf{y}_{train}|\hat{\theta})\} \\ &= \log\{p(\mathbf{y}|\hat{\theta})\} - \log\{p(\mathbf{y}_{train}|\hat{\theta})\}, \end{aligned} \tag{2.24}$$

we write $\ell_{pred}(\hat{\theta})$ as the difference between the two marginal likelihoods in (2.24) and use path integration to compute these two marginal loglikelihoods separately.

Table 2.1: Values of holdout loglikelihood ($\ell_{holdout}$) and predictive loglikelihood (ℓ_{pred}) for the Wisconsin Public Land Survey case study for either spatially independent models or the spatially correlated models, different numbers of forest community types (K), and the grid resolution (1km, 2km, or 4km), averaged over 3 runs from random initial starting parameters, and normalized by the number of trees in the testing dataset.

		$\ell_{holdout}(\hat{\theta})$			$\ell_{pred}(\hat{\theta})$		
Model	K	1km	2km	4km	1km	2km	4km
Independent	8	-2.77	-2.6	-2.37	-2.11	-2.15	-2.18
	12	-2.76	-2.58	-2.35	-2.04	-2.09	-2.13
	16	-2.76	-2.57	-2.33	-2	-2.06	-2.1
	20	-2.75	-2.57	-2.32	-1.98	-2.03	-2.08
	24	-2.75	-2.57	-2.32	-1.96	-2.02	-2.07
Spatial	8	-2.23	-2.21	-2.22	-2.03	-2.13	-2.19
	12	-2.18	-2.16	-2.2	-1.96	-2.08	-2.17
	16	-2.15	-2.15	-2.2	-1.91	-2.05	-2.15
	20	-2.15	-2.15	-2.18	-1.9	-2.03	-2.14
	24	-2.15	-2.15	-2.18	-1.88	-2.04	-2.14

Table 2.1 displays the holdout and predictive loglikelihoods obtained from the spatial and independent models for the different grid resolutions and numbers of forest community types. Intuitively, we expect it to be easier to predict the held out trees after having seen spatially nearby training trees. A comparison of the marginal and conditional loglikelihoods in Table 2.1 bears this out: the predictive loglikelihoods $\ell_{pred}(\hat{\theta})$ are always larger than the holdout loglikelihoods $\ell_{holdout}(\hat{\theta})$.

At all grid resolutions and numbers of forest community types, the spatial model performs better based on holdout loglikelihood than the corresponding spatially independent model. Additionally, the highest (best) spatially independent holdout loglikelihood is lower than the holdout loglikelihood from even the worst spatially correlated model. For the spatially independent models, the holdout loglikelihoods for models with fixed numbers of forest community types decrease as the grid resolution becomes finer, while the holdout loglikelihoods for the spatial models with fixed number of forest community types are more similar across the grid resolutions.

In contrast to the holdout loglikelihoods, the predictive loglikelihoods for both the spatially correlated and independent models improve as the grid resolution becomes finer. Additionally, the predictive loglikelihoods increase monotonically at each grid resolution as more forest community types are added to the model. The largest (best) predictive loglikelihood is obtained for a 1km spatial model with 24 forest community types. The spatially independent models sometimes achieve higher predictive loglikelihoods at the 2km and 4km grid resolutions, but the best predictive loglikelihoods out of all the models are attained by spatial models at the 1km resolution.

Finally, model fits from different initializations on the PLS dataset, where the true data generating mechanism is unknown, were qualitatively similar, with some

variability in the fitted forest communities. For a fixed number of forest community types, the correlation parameter estimates are typically similar across the grid resolutions. For example, for the 16-community models, the smallest spatial correlation parameter estimates are 1.615, 1.610, and 1.549, whereas the largest are 1.631, 1.619, and 1.596, for the grid resolutions 1km, 2km, and 4km, respectively.

2.4.2 Ecological Interpretation

After model fitting, the forest community classifications at each grid cell are determined from site-wise maximum a posteriori (MAP) estimates using Gibbs sampling. Maps of these classifications are shown in Figures 2.2–2.3, which indicate that the spatially correlated models tend to produce more spatially smooth classification maps than the spatially independent models, particularly for the smaller grid resolutions, as is expected. A key to the tree species abbreviations in these figures is given in Table 2.2.

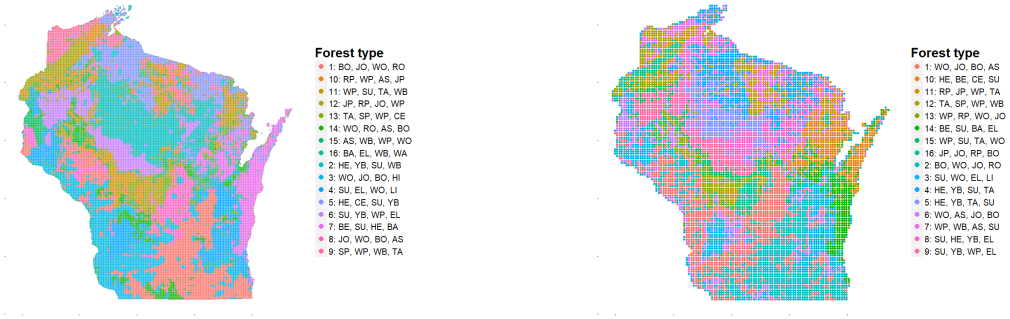
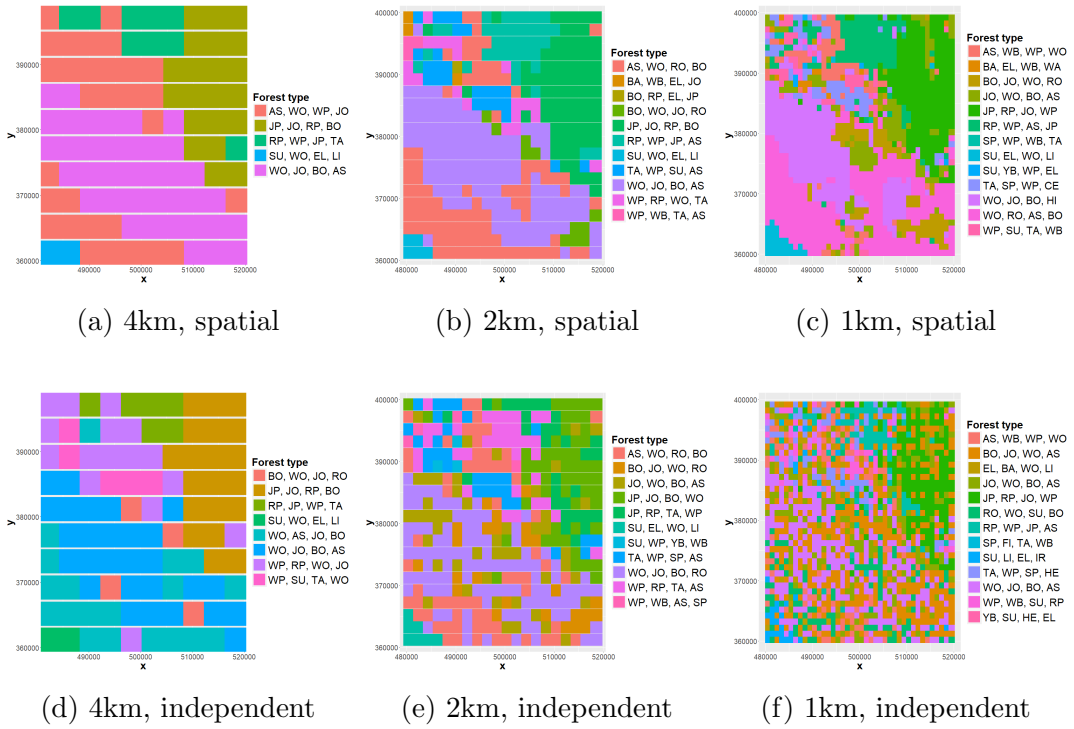


Figure 2.2: Forest community classifications for the Public Land Survey case study from the 16-community spatially correlated (left) and spatially independent (right) models with the highest holdout loglikelihoods, which occurred at the 1km and 4km grid resolutions, respectively. A key to the tree species abbreviations is given in Table 2.2.



(a) 4km, spatial (b) 2km, spatial (c) 1km, spatial

(d) 4km, independent (e) 2km, independent (f) 1km, independent

Figure 2.3: A comparison of 16-community spatial and independent models over a 40-km-by-40-km subsection of the Wisconsin survey region for grid resolutions 4km, 2km, and 1km. A key to the tree species abbreviations is given in Table 2.2.

Table 2.2: The tree species abbreviations, names, and counts for the Wisconsin Public Land Survey case study.

Abbreviation	Name	Count	Abbreviation	Name	Count
AL	Alder	100	LI	American basswood	7520
AS	Aspen	12029	RM	Red maple	1475
BA	Black ash	5957	RO	Red oak	5228
BE	American beech	7586	RP	Red pine	9925
BO	Bur oak	34065	SO	Swamp white oak	207
BU	Butternut	534	SP	Spruce	6048
BW	Black walnut	113	SU	Sugar maple	32718
CE	White cedar	8297	TA	Tamarack	19741
CH	Black cherry	454	WA	White ash	2119
CO	Eastern cottonwood	122	WB	Paper birch	11770
EL	Elm	11090	WI	Willow	346
FI	Balsam fir	4441	WM	Silver maple	550
HE	Eastern hemlock	26369	WO	White oak	33170
HI	Shagbark hickory	1198	WP	Eastern white pine	21717
IR	Ironwood	4076	YB	Yellow birch	22008
JO	Black & northern pin oak	26058	ZZ	No trees	464
JP	Jack pine	11004			

The predictive loglikelihoods for both the spatially correlated and independent

models improve as the grid resolution decreases from 4km to 1km (Table 2.1), which suggests that the tree data are more likely to come from the same forest community type within smaller grid cells, and that the larger grid cells aggregate trees from multiple forest community types. This pattern is consistent with ecological observations of forest patch size in the region [Mladenoff et al., 1993].

Table 2.3: Summaries of the 16 estimated forest community types for the Wisconsin Public Land Survey case study under the model selected based on cross validation, including the counts of grid cells which are classified as each forest community type, the top four tree species in each forest community, and the corresponding four largest estimated species probabilities. A key to the tree species abbreviations is given in Table 2.2.

Forest type	Count	Top Species	Species Probabilities
1	23174	HE, YB, SU, WB	0.352, 0.223, 0.154, 0.048
2	19373	BO, JO, WO, RO	0.72, 0.137, 0.117, 0.007
3	16121	WO, JO, BO, HI	0.478, 0.267, 0.209, 0.013
4	12380	SU, EL, WO, LI	0.282, 0.151, 0.131, 0.121
5	10746	SU, YB, WP, WB	0.342, 0.201, 0.089, 0.071
6	8427	BE, SU, HE, BA	0.39, 0.138, 0.096, 0.065
7	7370	JO, WO, BO, AS	0.643, 0.177, 0.091, 0.031
8	7219	CE, HE, TA, YB	0.283, 0.175, 0.133, 0.094
9	7075	SP, WP, WB, TA	0.179, 0.167, 0.161, 0.146
10	6013	RP, WP, AS, JP	0.489, 0.19, 0.077, 0.061
11	5720	WP, SU, TA, WB	0.66, 0.055, 0.046, 0.042
12	5506	TA, WP, SP, WB	0.821, 0.036, 0.026, 0.015
13	5420	JP, RP, JO, WP	0.753, 0.087, 0.052, 0.024
14	4865	WO, RO, AS, BO	0.484, 0.221, 0.069, 0.067
15	4256	AS, WB, WO, WP	0.62, 0.063, 0.045, 0.043
16	3186	BA, EL, WB, WA	0.236, 0.161, 0.073, 0.064

We focus our ecological interpretation on the spatially correlated model with 1km grid resolution and 16 forest community types which has the highest holdout loglikelihood out of the 1km models. This model also has the highest holdout loglikelihood for the 16 forest community models across grid resolutions. Table 2.3 summarizes the forest communities for this model and indicates that species associations within these 16 forest communities are consistent with ecological expectation for the survey

region [Curtis, 1959]. Similarly, the maps of the most likely forest community for each grid cell generally meet expectations [Curtis, 1959, Finley, 1976]. Although one of the models at the 2km grid resolution has a higher holdout loglikelihood than the model we discuss here, the difference in holdout likelihoods was small (-2.143 for the 2km model vs. -2.144 for the 1km model), while the improvement in predictive loglikelihood from 2km to 1km is more substantial (-2.04 for the 2km model, vs. -1.91 for the 1km model).

Among the oak communities, bur oak (BO) is the highest probability species in the community (forest community type 2 in Table 2.3) that is most likely to occur in the region that was historically oak savanna, mainly in topographically gentle sites [Curtis, 1959]. While all oak species in the survey region are fire-adapted, bur oak is the most fire tolerant [Peterson and Reich, 2001]. Its dominance in flatter areas could be due to increased frequencies of prairie fires passing through [Stambaugh and Guyette, 2008, Shea et al., 2014]. A more mixed oak community (forest community type 3), dominated by white oak (WO) with a high probability of black/jack oak (JO) and bur oak, was most likely to occur in a more topographically diverse, historically savanna region; the topography likely allowed for more diverse fire patterns and species assemblages [Shea et al., 2014]. The community dominated by black/jack oak (forest community type 7) had highest probability in regions with dry soils; of the oak species in Wisconsin, black and jack oak are the most drought-tolerant, so their dominance on these sites is ecologically sensible [Curtis, 1959, Shea et al., 2014]. While the other oak species likely were restricted to sunny savannas, the white oak-red oak (RO) community (forest community type 14) may have existed as a closed canopy community in southern Wisconsin. White oak and red oak are the more shade tolerant

oaks [Curtis, 1959].

The three pine species in Wisconsin occur in several communities, three of which are each dominated by the three species. The separation of the three species is expected, because while they are all associated with drier site conditions [Curtis, 1959], they are each differently adapted to drought and fire and, especially for jack pine (JP) and red pine (RP), often form monospecific stands depending on fire frequency [Burns and Honkala, 1990, Radeloff et al., 1999]. White pine (WP) has greater than 0.1 probability in the red pine dominated community (forest community type 10) as well as in a community (forest community type 9) with similar probabilities of spruce (SP), paper birch (WB), and tamarack (TA). Compared to the other pine species, white pine grows on a range of sites including those with richer soil, and has intermediate shade tolerance which allows it occur on a variety of sites and even intergrade with northern mesic forest community types [Curtis, 1959, Burns and Honkala, 1990, Fahey et al., 2012]. Given the widespread nature of white pine, it is not surprising that it has high probability of occurring in more than one community, including forest community type 9 which has species combinations that are possible on sites with recent disturbance or sites that are refuges from fire [Fahey et al., 2012].

In northern Wisconsin, mesic forest occurs on sites with rich and moist but well drained soils and is mainly dominated by eastern hemlock (HE), sugar maple (SU), yellow birch (YB), and American beech (BE) [Curtis, 1959]. The cluster results separate this forest type into four communities, and probabilities of each forest community type seem to vary geographically, depending on the range boundaries of several species [Curtis, 1959, Davis et al., 1991]. Beech dominates one community, and sugar maple and hemlock are other high-probability species in the community (forest

community type 6) which is most likely to occur east of beech's range boundary in eastern Wisconsin.

In northern Wisconsin, forest community type 8 is the most likely community, where hemlock has the highest probability along with white cedar, yellow birch, and sugar maple. White cedar (CE) is most abundant in far northern Wisconsin [Curtis, 1959]. South of that a different community is more likely to occur (forest community type 1), with high probability of hemlock, yellow birch, and sugar maple. West and south of the range of hemlock, forest community type 5 is most likely to occur; in that community, hemlock is absent and sugar maple and yellow birch dominate.

The remaining communities also align with expected forest types. In southern Wisconsin, community type 4 is southern mesic forest, which is most likely in known closed forest areas as expected [Curtis, 1959, Mladenoff et al.]. Community type 16 is wet-mesic forest in both north and south [Curtis, 1959]. Forest community type 13 is a tamarack wetland and forest community type 15 is northern dry/dry-mesic sites that are recently disturbed and dominated by aspen (AS) [Curtis, 1959].

2.4.3 Model Diagnostic and Implementation Validation

In addition to the loglikelihoods, we consider an absolute deviation measure of discrepancy between the observed and predicted proportions of tree species. To compute this measure of discrepancy, we overlay a grid of n 20-km by 20-km cells on the state of Wisconsin and compute the discrepancy

$$D = (Mn)^{-1} \sum_{i=1}^{n} \sum_{m=1}^{M} |\bar{p}_{mi} - \hat{p}_{mi}|,$$

where i indexes the 20-km by 20-km grid cells, $\bar{p}_{mi} = \mathbf{y}_{mi,test}/q_i^{test}$ denotes the empirical proportion of testing species m trees in the ith grid cell, and $\hat{p}_{mi}$ denotes the

corresponding predicted proportion under a given model. The discrepancy D measures the average difference between the observed and predicted proportions in the 20-km by 20-km grid cells. For the mixture models, the predicted species proportions for the ith grid cell $\hat{p}_{mi}$ are

$$\hat{p}_{mi} = \sum_{z \in \Omega} p(z|\mathbf{y}, \hat{\theta}) \sum_{k=1}^{K} I(z_i = k) \hat{\boldsymbol{\mu}}_{mk}$$

We compute these predicted species probabilities analytically for the spatially independent models but via MCMC for the spatially correlated models.

Table 2.4: Values of ℓ_1 discrepancy (D) on the testing dataset on a 20-km by 20-km grid for the spatially independent and dependent models with different numbers of forest community types (K) and grid resolutions (1km, 2km, 4km) in the Wisconsin Public Land Survey case study.

Model	K	1km	2km	4km
Independent	8	0.0153	0.0136	0.013
	12	0.014	0.0118	0.012
	16	0.0132	0.0111	0.0111
	20	0.013	0.0107	0.0108
	24	0.0129	0.0104	0.0103
Spatial	8	0.0125	0.0132	0.014
	12	0.011	0.0115	0.0131
	16	0.00994	0.011	0.0127
	20	0.00957	0.0106	0.0123
	24	0.00933	0.0106	0.0123

From Table 2.4, the overall differences between the predicted and observed species proportions are small, indicating good fit between the observed and predicted species proportions. The best performing models with respect to the measure D achieve an average absolute deviation of about 0.01 between the observed and predicted proportion for each of the 33 species. The deviations for the spatially correlated models decrease as the grid resolution becomes finer, in contrast to the deviations for the spatially independent models, which increase as the grid resolution becomes finer.

The overall pattern for the absolute deviations, as the number of categories and the grid resolutions change, is similar to the pattern for the predictive loglikelihoods in Table 2.1.

For the spatial models, we also investigate an intuitive approximation of $p(\mathbf{y}_{test}|\mathbf{y}_{train}, \hat{\theta})$, which allows us to validate our path integration implementation. Under the assumption

$$p(z|\mathbf{y}_{train}, \hat{\theta}) \approx \prod_{i=1}^{n} p(z_i|\mathbf{y}_{train}, \hat{\theta}),$$

we have

$$\begin{aligned} \log\{p(\mathbf{y}_{test}|\mathbf{y}_{train})\} &= \log\{\sum_{z\in\Omega} p(\mathbf{y}_{test}|z, \hat{\theta})p(z|\mathbf{y}_{train}, \hat{\theta})\} \\ &\approx \sum_{i=1}^{n} \log\left\{\sum_{k=1}^{K} p\{\mathbf{y}_{test}|z_i, \hat{\theta}\}p(z_i = k|\mathbf{y}_{train}, \hat{\theta})\right\}. \end{aligned} \tag{2.25}$$

Using this approximation combined with MCMC draws from $p(z|\mathbf{y}_{train}, \hat{\theta})$ to obtain empirical estimates of $p(z_i = k|\mathbf{y}_{train}, \hat{\theta})$, we compute an approximation of the true predictive loglikelihood $\ell_{pred}(\hat{\theta})$, denoted as $\ell_{pred}^{approx}(\hat{\theta})$. Table 2.5 suggests that the results from this approximate procedure agree very well with the results obtained via path integration in spite of the mostly different implementation details, providing evidence for the correctness of our path integral implementation.

Table 2.5: Predictive loglikelihood values on the testing dataset for the spatially correlated model, computed using path integration and the approximate method of (2.25), for the Wisconsin Public Land Survey case study with different numbers of forest community types (K) and grid resolutions (1km, 2km, 4km).

Method	K	1km	2km	4km
Path integral	8	-2.03	-2.13	-2.19
	12	-1.96	-2.08	-2.17
	16	-1.91	-2.05	-2.15
	20	-1.9	-2.03	-2.14
	24	-1.88	-2.04	-2.14
Approximate	8	-2.03	-2.14	-2.19
	12	-1.96	-2.08	-2.17
	16	-1.91	-2.05	-2.15
	20	-1.9	-2.03	-2.14
	24	-1.88	-2.04	-2.14

2.5 Simulation Study

We conduct a simulation study to evaluate the methodology applied to the PLS case study in Sections 2.2–2.4. We consider $g \times g$ grids of cells, where the grid size is $g = 50, 100, 200$, or 400 corresponding to n = 2,500, 10,000, 40,000, or 160,000 grid cells, respectively. We also consider the effect of observing larger and smaller numbers of trees within each cell, by conducting simulations at $q = 3$ or 6 trees observed per cell. For each combination of grid size (g) and number of trees per cell (q), 100 simulations are performed. There are $K = 8$ true forest community types, with associated probabilities given in the $\boldsymbol{\mu}$ matrix below where the $K = 8$ columns

of $\boldsymbol{\mu}$ each sum to 1.

$$\boldsymbol{\mu} = \begin{array}{c} \begin{array}{cccccccc} \text{Class 1} & \text{Class 2} & \text{Class 3} & \text{Class 4} & \text{Class 5} & \text{Class 6} & \text{Class 7} & \text{Class 8} \end{array} \\ \begin{bmatrix} 0.186 & 0.126 & 0.049 & 0.264 & 0.036 & 0.212 & 0.031 & 0.403 \\ 0.228 & 0.177 & 0.086 & 0.139 & 0.465 & 0.016 & 0.015 & 0.057 \\ 0.016 & 0.015 & 0.016 & 0.026 & 0.064 & 0.022 & 0.016 & 0.054 \\ 0.089 & 0.016 & 0.035 & 0.299 & 0.022 & 0.041 & 0.235 & 0.021 \\ 0.026 & 0.018 & 0.015 & 0.024 & 0.134 & 0.016 & 0.220 & 0.015 \\ 0.019 & 0.092 & 0.103 & 0.016 & 0.016 & 0.065 & 0.045 & 0.027 \\ 0.044 & 0.015 & 0.015 & 0.016 & 0.016 & 0.016 & 0.016 & 0.125 \\ 0.015 & 0.195 & 0.016 & 0.016 & 0.111 & 0.021 & 0.019 & 0.062 \\ 0.028 & 0.133 & 0.199 & 0.059 & 0.040 & 0.109 & 0.049 & 0.018 \\ 0.036 & 0.015 & 0.025 & 0.015 & 0.016 & 0.360 & 0.025 & 0.021 \\ 0.017 & 0.017 & 0.016 & 0.046 & 0.015 & 0.057 & 0.026 & 0.021 \\ 0.039 & 0.016 & 0.017 & 0.030 & 0.015 & 0.017 & 0.016 & 0.015 \\ 0.223 & 0.094 & 0.015 & 0.015 & 0.016 & 0.016 & 0.016 & 0.016 \\ 0.016 & 0.016 & 0.374 & 0.016 & 0.018 & 0.016 & 0.016 & 0.027 \\ 0.017 & 0.054 & 0.017 & 0.018 & 0.016 & 0.019 & 0.257 & 0.117 \end{bmatrix} \end{array}$$

The simulated vectors of forest community types z have density

$$p(z|\eta) = \exp\left\{\eta^T T(z) - \xi(\eta)\right\},$$

where $\eta = [-0.060, -0.055, -0.039, -0.037, -0.024, -0.057, -0.004, 1.2]^T$ and $T(z)$ is defined as in (2.2). That is, the spatial correlation parameter $\eta_K = 1.2$. Given the forest community types $Z = z$, the tree count vectors $\mathbf{Y}_i$ are independent multinomials with sample sizes $q = 3$ or $q = 6$ trees at each grid cell. Since the regularized likelihood is invariant to permutations of the mixture categories, we use the permutation of categories that minimizes the mean squared errors (MSE),

$$\text{MSE} = \sum_{k=1}^{8}\sum_{m=1}^{15}(\widehat{\boldsymbol{\mu}}_{mk} - \boldsymbol{\mu}_{mk})^2/(MK),$$

for each simulation when assessing the estimation error. The MSE for the $\boldsymbol{\mu}$ matrix are reported in Table 2.6 for the stochastic modified EM algorithm described in Algorithm

1 ("Modified EM"), in comparison to the spatially independent model fit via the EM algorithm ("Independent EM"), the ordinary stochastic gradient as described in Younes [1989] ("Ordinary SG"), and a version of stochastic gradient with differently scaled stepsizes for the η and $\boldsymbol{\mu}$ parameter ("Rescaled SG"). Implementation details for ordinary and rescaled stochastic gradient are given in Section A.2.3 of Appendix A.

Table 2.6: Simulation mean squared error (MSE) for the species probability parameter matrix $\boldsymbol{\mu}$ using different algorithms for $q = 3, 6$ simulated trees per grid cell and for different numbers of grid cells n.

Method	Trees per cell	$n = 50^2$	$n = 100^2$	$n = 200^2$	$n = 400^2$
Modified EM	$q = 3$	2e-04	4e-05	1e-05	2e-06
	$q = 6$	5e-05	1e-05	3e-06	7e-07
Independent EM	$q = 3$	4e-04	8e-05	2e-05	6e-06
	$q = 6$	2e-04	4e-05	2e-05	8e-07
Rescaled SG	$q = 3$	7e-04	5e-04	4e-04	5e-04
	$q = 6$	4e-04	4e-04	4e-04	3e-04
Ordinary SG	$q = 3$	0.003	0.003	0.003	0.003
	$q = 6$	0.002	0.002	0.002	0.002

Table 2.6 suggests that the modified EM algorithm performs best at every setting, followed by independent EM. When only $q = 3$ trees are included in each cell, the MSEs for the $\boldsymbol{\mu}$ parameter from the spatially independent model are over twice that of the spatially correlated. When $q = 6$ trees are included at each cell, the performance of the spatially correlated and independent models are more similar, although the spatially correlated model still always performs better than the spatially independent model. For both models, the MSE at each grid size is, as expected, lower when $q = 6$ trees are included than when $q = 3$ trees are included. For both the spatially correlated and independent models, the parameter estimates $\hat{\boldsymbol{\mu}}$ appear to be converging to the truth at about the rate of $\sqrt{n}$. The convergence occurs in spite of the fact that the likelihood is multimodal, while the fitting algorithms were randomly initialized.

This suggests that the estimation procedure is robust to the choice of initialization. Interestingly, the rescaled stochastic gradient performs better than ordinary stochastic gradient, but still performs worse than the independent EM algorithm.

Table 2.7: Simulation bias, variance, and mean squared error (MSE) for the spatial correlation parameter η_K using different algorithms for $q = 3, 6$ simulated trees per grid cell and for different numbers of grid cells n.

Method	Error	$n = 50^2$ $q = 3$	$n = 50^2$ $q = 6$	$n = 100^2$ $q = 3$	$n = 100^2$ $q = 6$	$n = 200^2$ $q = 3$	$n = 200^2$ $q = 6$	$n = 400^2$ $q = 3$	$n = 400^2$ $q = 6$
Modified EM	Bias	-0.002	-0.009	-0.002	-0.002	-0.001	-1e-05	-0.002	-6e-04
	Variance	9e-04	6e-04	2e-04	1e-04	4e-05	3e-05	8e-06	7e-06
	MSE	9e-04	7e-04	2e-04	1e-04	4e-05	3e-05	1e-05	7e-06
Rescaled SG	Bias	-0.03	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02
	Variance	0.002	0.001	5e-04	4e-04	2e-04	3e-04	2e-04	3e-04
	MSE	0.003	0.002	0.001	8e-04	7e-04	6e-04	6e-04	6e-04
Ordinary SG	Bias	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2
	Variance	0.005	0.006	0.005	0.006	0.005	0.007	0.005	0.005
	MSE	0.04	0.03	0.04	0.03	0.05	0.03	0.05	0.03

The spatially independent model does not include the spatial correlation parameter η_K, so that Table 2.7 compares the bias, variance, and mean squared errors for the spatial correlation parameter η_K only for the stochastic modified EM, ordinary stochastic gradient, and rescaled stochastic gradient algorithms. Again, the stochastic modified EM algorithm performs better than either rescaled stochastic gradient and ordinary stochastic gradient. This difference is particularly pronounced for the larger grid sizes. For $g = 400$ and $q = 6$, the MSE for the stochastic modified EM algorithm is approximately 100 times smaller than the MSE for the rescaled stochastic gradient algorithm. As can be seen from Table 2.7, the component of MSE due to bias for the stochastic modified EM algorithm is very small relative to the component of MSE due to variance. Additionally, the MSE decreases monotonically as the grid size increases, as well as when more trees are observed within each cell. This suggests that our algorithm accurately recovers the spatial correlation parameter in the Potts

distribution.

Since the spatially independent model does not include the spatial correlation parameter η_K, the estimates for η_k when $k < K$ (Table 2.8) for the independence model are expected to be biased relative to the true data generating η_k parameters, so that comparisons between the correlated and uncorrelated model estimates are less meaningful for these parameters. The stochastic modified EM algorithm performs best out of all the methods for every combination of grid size and number of trees per grid cell.

Table 2.8: Simulation mean squared error (MSE) for the η_k parameters when $k < K$ using different algorithms for $q = 3, 6$ simulated trees per grid cell and for different numbers of grid cells n.

Method	Trees per cell	$n = 50^2$	$n = 100^2$	$n = 200^2$	$n = 400^2$
Modified EM	$q = 3$	0.0031	0.00082	0.00017	5.9e-05
	$q = 6$	0.0031	0.00053	0.00011	2.6e-05
Independent EM	$q = 3$	0.24	0.066	0.041	0.03
	$q = 6$	0.082	0.041	0.027	0.026
Rescaled SG	$q = 3$	0.0099	0.0083	0.0089	0.0085
	$q = 6$	0.044	0.043	0.041	0.035
Ordinary SG	$q = 3$	0.62	0.66	0.69	0.61
	$q = 6$	1	1	1.2	1.1

Finally, in our simulation study, the minimum number of trees in a dataset is 7,500, while in the PLS case study, over 300,000 trees were observed. Thus, the "prior sample sizes" of trees from each forest community type are much smaller than the observed sample size, and we do not expect the prior penalties to substantially bias the estimation procedure. The simulation study results bear this out. Additionally, the η parameters are estimated in simulation with very little bias due to the regularization.

Chapter 3

Control Variates and Rao-Blackwellization for Deterministic Sweep Markov Chains

3.1 Introduction

This chapter is based on joint work with Jun Zhu and Murray Clayton, and a preprint of an earlier version appears in Berg et al. [2019a].

Markov chain Monte Carlo (MCMC) is a widely used technique for drawing samples from intractable probability distributions. In statistics, MCMC is now a standard tool in Bayesian analysis for sampling from complicated posterior distributions. The goal of MCMC is usually to approximate quantities such as $\int \pi(dx)g(x)$, where π is an intractable probability measure, and $g : X \to \mathbb{R}^d$ is a π-integrable function mapping a state space X to $\mathbb{R}^d$ for some integer $d \geq 1$. In MCMC, a Markov chain $X_0, X_1, X_2, ...$ with a stationary probability measure π is simulated for some finite number of iterations M, and $\int \pi(dx)g(x)$ is then estimated by the empirical average

$$S_M/M = M^{-1} \sum_{t=0}^{M-1} g(X_t). \tag{3.1}$$

Under suitable conditions, a central limit theorem can be shown for the estimator

S_M/M stating that

$$M^{1/2}\left\{S_M/M - \int \pi(dx)g(x)\right\} \xrightarrow{d} N(0,\Sigma) \tag{3.2}$$

as $M \to \infty$, where $\xrightarrow{d}$ denotes convergence in distribution [Meyn and Tweedie, 2009]. In this sense, S_M/M is asymptotically unbiased, and the MCMC error asymptotically comes entirely from the asymptotic variance Σ. Thus, one sensible measure of the efficiency of an MCMC estimator is the asymptotic variance Σ, which we will use in the remainder.

A variety of techniques exist for reducing the asymptotic variance Σ in (3.2) for MCMC simulations, including conditioning, control variates, and antithetic sampling [see, e.g., Liu, 2008, Robert and Casella, 2004]. We focus on control variate approaches here, although we also make connections to conditioning based approaches. In control variate approaches, mean zero random variables are added to each term of (3.1) in such a way that the variance of the sum is reduced. In approaches based on conditioning, g in (3.1) is replaced with the conditional expectation with respect to π of g given some intermediate quantity, with the hope that the resulting average has a reduced variance relative to (3.1). This procedure bears some resemblance to the classical Rao-Blackwell approach of reducing the variance of an estimator through conditioning [Rao, 1945, Blackwell, 1947], and thus the term Rao-Blackwellization is commonly used to describe techniques in which an MCMC average of a conditional expectation is taken in order to reduce the asymptotic variance Σ. However, in MCMC, unlike in classical Monte Carlo, independence does not hold and a naive conditioning approach may increase the asymptotic variance [Geyer, 1995].

The MCMC literature contains a variety of variance reduction results, espe-

cially for reversible Markov chains. For example, Casella and Robert [1996] provide variance reduction results for Markov chains resulting from the Metropolis-Hastings algorithm. A conditioning approach is used in McKeague and Wefelmeyer [2000] to obtain a variance reduction result for reversible Markov chains. In Meyn [2008], control variate methods are discussed for time-homogeneous Markov chains in the context of network models. In Douc and Robert [2011], a Rao-Blackwellization method is studied for Markov chains based on Metropolis-Hastings algorithms. In Dellaportas and Kontoyiannis [2012], a control variate method is given for reducing the variance of estimates based on reversible Markov chains. In Brosse et al. [2018], a control variate scheme is used to obtain variance reductions for certain Markov chains that can be related through a limiting process to a Langevin diffusion.

Our work here adds to the prior literature in several ways. First, deterministic sweep sampling is commonly used and more straightforward to implement than random sweep sampling, to which previous methodology applies. Thus, our proposed control variate methodology for deterministic sweep Markov chains lessens the gap between Markov chain theory and practice. As an example, our control variate methodology is applicable to deterministic sweep Gibbs samplers, whereas the control variate estimator for reversible Markov chains in Dellaportas and Kontoyiannis [2012] is applicable to random sweep Gibbs samplers but not to deterministic sweep Gibbs samplers. While we obtain several useful results for deterministic sweep Gibbs samplers, where the component transition kernels are reversible, our results also apply more generally and can be used to construct control variate estimates for deterministic sweep Markov chains composed of non-reversible Markov kernels.

Second, we propose a Rao-Blackwellization estimator for deterministic sweep

Gibbs sampling with variance reduction guarantees in broader settings than existing Rao-Blackwellization estimators. Our Rao-Blackwellization estimator in Corollary 3.2 applies to Gibbs samplers with $K \geq 2$ components. The Rao-Blackwellization estimator in Liu et al. [1994] only comes with theoretical guarantees for Gibbs samplers with $K = 2$ components. Liu et al. [1994] also require the integrand g to satisfy a relatively strong strong dependence condition which we do not require. Rao-Blackwellization for Gibbs sampling is commonly applied in practice, but theoretical justification for this approach had previously been lacking. For example, Goodfellow et al. [2013] use a Rao-Blackwellization scheme to improve the efficiency of a stochastic gradient algorithm involving Gibbs sampling, but justify their approach using the classical Rao-Blackwell theorem for independent data. Our result in Corollary 3.2 provides a more rigorous foundation for Rao-Blackwellization for deterministic sweep Gibbs samplers by showing that the conditioning leads to a smaller asymptotic variance in the Markov chain central limit theorem.

Third, for two-component Gibbs samplers, our proposed control variate methodology yields provably smaller asymptotic variances than the current state of the art for control variate and Rao-Blackwellization methods. In this setting, the asymptotic variances attained by our methodology are guaranteed to be smaller (Theorem 3.3) than those resulting from the methodology in Dellaportas and Kontoyiannis [2012], without additional computational cost. Our proposed control variate methodology also yields smaller asymptotic variances (Proposition 3.2) than the canonical Rao-Blackwellization estimate proposed in Liu et al. [1994] for the data augmentation Gibbs sampler setting. Our control variate approach will often be feasible to implement with negligible additional computational costs whenever the Liu et al. [1994]

Rao-Blackwellization approach is feasible to implement.

3.2 Notation and Setup

3.2.1 K-Component Samplers and Gibbs Kernels

We consider Markov chains $\{X_t\}_{t=0}^{\infty}$ evolving on a state space $(X, \mathscr{X})$, where X is assumed to be a complete separable metric space, and $\mathscr{X}$ is the associated Borel σ-algebra. We refer to a function $\Pi : X \times \mathscr{X} \to [0,1]$ as a probability kernel if $\Pi(\cdot, A) : X \to [0,1]$ is an $\mathscr{X}$–measurable function of x for each $A \in \mathscr{X}$, and also $\Pi(x, \cdot) : \mathscr{X} \to [0,1]$ defines a probability measure on $(X, \mathscr{X})$ for each $x \in X$. Given a probability measure λ on $(X, \mathscr{X})$, we say a probability kernel $\Pi(x, A) : X \times \mathscr{X} \to [0,1]$ is λ-stationary iff $\lambda(A) = \int \lambda(dx)\Pi(x, A)$ for all $A \in \mathscr{X}$.

We use $\mathbb{R}$ to denote the extended real line $[-\infty, \infty]$ and $\mathbb{N} = \{0, 1, 2, ...\}$ refer to the nonnegative integers. For a function $f : X \to \mathbb{R}^p$ where $p \geq 1$ and probability kernel $\Pi : X \times \mathscr{X} \to [0,1]$, we define $\Pi^0 f(x) = f(x)$, $\Pi^1 f(x) = \Pi f(x) = \int \Pi(x, dy) f(y)$, and $\Pi^t f(x) = \Pi(\Pi^{t-1} f)(x)$ for $t > 1$. We also define $\Pi^t(x, A) = \Pi^t I_A(x)$, where $I_A(\cdot) : X \to \mathbb{R}$ denotes the indicator function with $I_A(x) = 1$ for $x \in A$ and $I_A(x) = 0$ elsewhere. We define the permutation function $\sigma(\cdot) : \{1, ..., K\} \to \{1, ..., K\}$ by $\sigma(k) = k + 1$ for $k < K$ and $\sigma(k) = 1$ for $k = K$, and we define $\sigma^t(k)$ inductively by $\sigma^t(k) = k$ for $t = 0$, and $\sigma^t(k) = \sigma\{\sigma^{t-1}(k)\}$ for $t > 0$.

Deterministic sweep samplers update the Markov chain by applying the kernels Π_k in a fixed order. At time 0, the transition operator used to update the state is, without loss of generality, $\Pi_1 = \Pi_{\sigma^0(1)}$, and at time t, the transition operator is $\Pi_{\sigma^t(1)}$.

Thus, for an initial probability measure ν on $(X, \mathscr{X})$, we have

$$\begin{aligned} &P_\nu(X_0 \in A_0, ..., X_t \in A_t) \\ &= \int \nu(dx_0)\Pi_1(x_0, dx_1)...\Pi_{\sigma^t(1)}(x_{t-1}, dx_t)I(\{x_i \in A_i, \forall i\}). \end{aligned}$$

We also define $P_k^t f(x)$ by $P_k^0 f(x) = f(x)$ and $P_k^t f(x) = \left\{\Pi_{\sigma^0(k)}\Pi_{\sigma^1(k)} \cdots \Pi_{\sigma^{t-1}(k)} f\right\}(x)$ for $t \geq 1$. We define $P_k^t(x, A) = P_k^t I_A(x)$, so that P_k^t is a composition of multiple kernels rather than a repeated composition of a single kernel. Random sweep kernels update the Markov chain via the mixture kernel $Q = K^{-1}\sum_{k=1}^K \Pi_k$.

Next, we define the probability kernels used in Gibbs sampling. Define the identity map $Y : (X, \mathscr{X}) \to (X, \mathscr{X}), x \to x$ for $x \in X$. We say a probability kernel $\Pi : X \times \mathscr{X} \to [0, 1]$ is a regular conditional distribution kernel with respect to $(\mathscr{G}, \pi)$, where $\mathscr{G} \subset \mathscr{X}$ is a sub-σ-algebra of $\mathscr{X}$, whenever (1) $\Pi(x, A) = E\{I(Y \in A)|\mathscr{G}\}$ almost everywhere with respect to π (a.e. π), for each $A \in \mathscr{X}$, and (2) for π-a.e. x, $\Pi(x, A)$ is a probability measure on $(X, \mathscr{X})$. It is well-known that when X is a complete separable metric space and $\mathscr{X}$ the associated Borel σ-algebra, such regular conditional distributions Π always exist [see, e.g., Durrett, 2010].

For a measurable function $h : X \to \mathbb{R}^n$ for some $n \geq 1$, we say a probability kernel $\Pi : X \times \mathscr{X} \to [0, 1]$ is a Gibbs kernel with respect to (h, π) if Π is a regular conditional distribution kernel with respect to the σ-algebra $\sigma(h)$. Gibbs kernels have some useful properties. First, any Gibbs kernel Π with respect to (h, π) preserves π. This follows since

$$\int \pi(dx)\Pi(x, A) = \pi(A) \tag{3.3}$$

for each $A \in \mathscr{X}$, from the properties of conditional expectation. Additionally, the idempotence property $\Pi\{\Pi f\}(x) = \Pi f(x)$ holds a.e. π for each π-integrable f. Finally,

for functions f, g which are square-integrable with respect to π, we have

$$\begin{aligned}\int \pi(dx) f(x) \Pi g(x) &= \int \pi(dx) \Pi\{f \Pi g\}(x) \\ &= \int \pi(dx) \Pi f(x) \Pi g(x) = \int \pi(dx) g(x) \Pi f(x). \end{aligned} \tag{3.4}$$

The equality $\int \pi(dx) f(x) \Pi g(x) = \int \pi(dx) g(x) \Pi f(x)$ is the useful *reversibility* property. Thus from (3.4), we see that Gibbs transition kernels are reversible with respect to π. However, compositions of reversible probability kernels such as the P_k^t that arise in deterministic sweep Gibbs sampling will in general not be reversible with respect to π.

3.2.2 Control Variates and Rao-Blackwellization

In general, control variate schemes replace the estimator $M^{-1} S_M = M^{-1} \sum_{t=0}^{M-1} g(X_t)$ with an estimator of the form $M^{-1} \sum_{t=0}^{M-1} \{g(X_t) - cW_t\}$, where W_t are mean zero random variables, and c is a constant. Since the W_t are mean 0, both estimators have the same expected value. If $\{(X_t, W_t)\}_{t=0}^{M-1}$ are iid and the covariance of W_t and $g(X_t)$ is positive, then it is straightforward to check that the variance of $M^{-1} \sum_{t=0}^{M-1} \{g(X_t) - cW_t\}$ is minimized for the choice $c = \mathrm{var}(W_0)^{-1} \mathrm{cov}(W_0, g(X_0)) > 0$. However, when independence does not hold, as in MCMC, then the optimal choice of c is less straightforward, since it becomes necessary to account for correlations between terms at different time points t. The optimal choice of c in the Markov chain setting will be an important consideration for the remainder. In the Markov chain control variate estimators considered here, we consider mean zero control variates of the basic form $W_t = f(X_{t+1}) - \Pi_{\sigma^t(1)} f(X_t)$ to ensure that $M^{-1} \sum_{t=0}^{M-1} \{g(X_t) - W_t\}$ is an asymptotically unbiased estimator of $\int \pi(dx) g(x)$. Control variates of this approximate form were suggested in Henderson and Glynn [2002]. Dellaportas and Kon-

toyiannis [2012] refer to the function f as a "control variate basis function", and we adopt this terminology here.

For deterministic sweep chains $\{X_t\}_{t=0}^{\infty}$ that use probability kernel $\Pi_{\sigma^t(1)}$ to obtain X_t, we consider the three estimators (3.5), (3.6), and (3.7) below, in addition to the empirical estimator (3.1). In (3.6) and (3.7), we use $C \in \mathbb{R}^{p\times d}$ and $C_k \in \mathbb{R}^{p\times d}$ for $k = 1, ..., K$ to refer to fixed $p \times d$ matrices:

Rao-Blackwellized: (3.5)

$$M^{-1}S_M^{RB} = M^{-1}\sum_{t=0}^{M-1}\Pi_{\sigma^t(1)}g(X_t)$$

Fixed weight control variate: (3.6)

$$M^{-1}S_M^{FW} = M^{-1}\sum_{t=0}^{M-1}[g(X_t) - C^T\{f(X_{t+1}) - \Pi_{\sigma^t(1)}f(X_t)\}]$$

Variable weight control variate: (3.7)

$$M^{-1}S_M^{VW} = \sum_{t=0}^{M-1}[g(X_t) - C_{\sigma^{t+1}(1)}^T\{f(X_{t+1}) - \Pi_{\sigma^t(1)}f(X_t)\}]$$

The Rao-Blackwellized estimator (3.5) replaces each $g(X_t)$ in (3.1) with the conditional expectation $\Pi_{\sigma^t(1)}g(X_t)$. The estimator (3.5) essentially formalizes the idea that it ought to be better to replace $g(X_{t+1})$ in S_M with the conditional expectation of $g(X_{t+1})$ given X_t. The control variate estimators arise from adding mean 0 terms to the empirical estimator (3.1). We have $M^{-1}S_M^{FW} = M^{-1}S_M - M^{-1}\sum_{t=0}^{M-1} C^T\{f(X_{t+1}) - \Pi_{\sigma^t(1)}f(X_t)\}$ and $M^{-1}S_M^{VW} = M^{-1}S_M - M^{-1}\sum_{t=0}^{M-1} C_{\sigma^{t+1}(1)}^T\{f(X_{t+1}) - \Pi_{\sigma^t(1)}f(X_t)\}$. The variable weight scheme allows the control variate weight at each iteration t to

vary depending on the probability kernel being used to update the Markov chain at time t. Because the control variate and Rao-Blackwellization estimators involve the conditional expectations $\Pi_k f$ and $\Pi_k g$, respectively, it is necessary in practice for these conditional expectations to have a computationally tractable form.

We now provide some further remarks on (3.5)–(3.7). First, the Rao-Blackwellization estimator (3.5) is closely linked to the control variate estimators (3.6) and (3.7). For the choices $f = g$ and $C = I_{d\times d}$ in (3.6), we have $M^{-1}S_M^{RB} = M^{-1}S_M^{FW} - M^{-1}\{g(X_0) - g(X_M)\}$. Under reasonable conditions, the difference $M^{-1}\{g(X_0) - g(X_M)\}$ will be asymptotically negligible. Furthermore, the fixed weight control variate estimator (3.6) is a special case of the variable weight control variate estimator (3.7) with the choices $C_1 = C_2 = \cdots = C_K = C$. The fixed weight estimator (3.6) is similar to the control variate estimator proposed by Dellaportas and Kontoyiannis [2012], except that their reversible Markov chain kernel Q is replaced by the kernel $\Pi_{\sigma^t(1)}$ currently being used to update the Markov chain. Note that $\int \pi(dx)\{f - \Pi_k f\} = 0$ for $k = 1, ..., K$ when the Π_k are π stationary, and thus the control variate terms in (3.6) and (3.7) are mean 0 with respect to π. The variable weight estimator (3.7) allows a separate weight matrix $C_{\sigma(k)}$ to be used for differences involving each transition kernel Π_k.

3.3 Assumptions and Variance Reduction Results

3.3.1 Assumptions

We consider Markov chains $\{X_t\}_{t=0}^{\infty}$ with an arbitrary initial law ν on $(X, \mathscr{X})$ and time-inhomogeneous transition kernels $P_t = \Pi_{\sigma^t(1)}$. Throughout, we will take π to be a probability measure on $(X, \mathscr{X})$ which we would like to take expectations with respect to. We say a set $C \in \mathscr{X}$ is a small set with respect to a probability kernel P

if there exists an $m > 0$, and a non-trivial measure ν_m on $(X, \mathscr{X})$, such that

$$P^m(x, B) \geq \nu_m(B)$$

for all $x \in C$, $B \in \mathscr{X}$ [see, e.g., Meyn and Tweedie, 2009].

First, we make the following assumptions on the composition kernels P_k^{Kt} and the transition kernels $\Pi_1, ..., \Pi_K$.

(A.1) The kernels Π_k are π-stationary. This holds whenever at least one of (A.1a) and (A.1b) holds:

(A.1a) (Reversibility) The Π_k are reversible, so that $\langle f, \Pi_k g\rangle = \langle \Pi_k f, g\rangle$ for all square integrable functions $f, g : X \to \mathbb{R}$.

(A.1b) (Gibbs kernels) The Π_k are Gibbs kernels with respect to (h_k, π) for some set of functions $h_k : (X, \mathscr{X}) \to (\mathbb{R}^{m_k}, \mathscr{R}^{m_k})$, where $m_k \geq 1$ is an integer denoting the dimension of the range of h_k.

(A.2) (ψ-irreducibility) There exists a probability measure ψ on $(X, \mathscr{X})$ such that for each $k = 1, ..., K$ and all $A \in \mathscr{X}$ with $\psi(A) > 0$, and for all $x \in X$, there exists a positive integer $t = t(x, A, k)$ such that $P_k^{Kt}(x, A) > 0$.

(A.3) (Geometric drift) There exist small sets $C_k \in \mathscr{X}$ with respect to P_k^K, constants $\lambda_k < 1$ and $b_k < \infty$, and functions $V_k : X \to [1, \infty)$, such that for $k = 1, ..., K$,

$$P_k^K V_k(x) \leq \lambda_k V_k(x) + b_k I_{C_k}(x).$$

(A.4) (Aperiodicity) The composition kernels P_k^K are assumed to be aperiodic.

Assumption (A.1) ensures that the transition kernels Π_k preserve the stationary distribution π. Assumption (A.2) ensures that π is the unique stationary distribution for the transition kernels P_k^K for $k = 1, ..., K$ [see, e.g. Meyn and Tweedie, 2009]. Assumption (A.3) ensures that Markov chains with transition kernel P_k^K are Harris recurrent, so that for any $A \in \mathscr{X}$ with $\psi(A) > 0$, we have $P_x(\cap_{N=1}^{\infty} \cup_{k=N}^{\infty} \{X_t \in A\}) = 1$ for all $x \in A$, where P_x refers to the Markov chain law with point mass initial distribution δ_x and transition kernel P_k^K. Regarding (A.4), we say an irreducible probability kernel is aperiodic if $d = 1$ is the largest integer such that there exist sets $D_1, ..., D_d$ satisfying (1) $P_k^K(x, D_{i+1}) = 1$ for $x \in D_i$, $i = 1, ..., d-1$ and $P_k^K(x, D_1) = 1$ for $x \in D_d$, (2) $\psi\{(\cup_{i=1}^{d} D_i)^C\} = 0$, and (3) $D_1, ..., D_d$ are disjoint [see, e.g. Meyn and Tweedie, 2009]. Furthermore, our Lemma B.3 in Appendix A shows that Assumption (A.1b), when it holds, ensures (A.4) holds also.

Next, we make the following assumptions about the functions $g : X \to \mathbb{R}^d$, $f : X \to \mathbb{R}^p$, and $V_k(x)$ in (A.3):

(B.1) (Square integrability) $\int \pi(dx) V_k^2(x) < \infty$.

(B.2) $|a^T g(x)| \leq V_k(x)$ for all $a \in \mathbb{R}^d$ with $\|a\|_2 \leq 1$, where $\|x\|_2$ denotes the Euclidean norm of x. For a univariate function g, this is equivalent to assuming that $|g(x)| \leq V_k(x)$ for all x.

(B.3) $\int \pi(dx) g(x) = 0$.

(B.4) $\int \pi(dx) f^T f < \infty$ and $f^T f < \infty$ for all $x \in X$.

Assumptions (A.3), (A.4), (B.1), and (B.2) will be used to ensure certain bounds on solutions $\tilde{g}_k$ to the Poisson equations $\tilde{g}_k - P_k^K \tilde{g}_k = g - \int \pi(dx) g(x)$. Assump-

tion (B.3) is introduced for notational convenience: it allows us to write, for example, statements such as $\tilde{g}_k - P_k^K \tilde{g}_k = g$ rather than $\tilde{g}_k - P_k^K \tilde{g}_k = g - \int \pi(dx) g(x)$. For a general integrand g, the results to follow will apply to the function $\bar{g}(x) = g(x) - \int \pi(dx') g(x')$. In addition, (B.4) holds for $f = g$ when (A.1)–(A.4) and (B.1)–(B.2) hold.

3.3.2 Variance Reduction Results

In the following, we state our main variance reduction results, and defer the proofs to Appendix B.2.

Proposition 3.1. *Under (A.1)–(A.4) and (B.1)–(B.3), there exist functions $\hat{g}_k : X \to \mathbb{R}^d$ with*

$$\hat{g}_k(x) = \sum_{t=0}^{\infty} P_k^t g(x) \qquad k = 1, ..., K \tag{3.8}$$

a.e. π. The sums in the definition of $\hat{g}_k$ are absolutely convergent elementwise for π-a.e. $x \in X$, and $\int \pi(dx) \hat{g}_k^T \hat{g}_k < \infty$ for each k. Additionally, each $\hat{g}_k$ satisfies a corresponding Poisson-type equation

$$\hat{g}_k - \Pi_k \hat{g}_{\sigma(k)} = g, \quad a.e.\ \pi.$$

The Poisson equation solutions from Proposition 3.1 can be used to write each of the sums in the estimators given in (3.1) and (3.5)-(3.7) as the sum of an approximating martingale, plus a small error term. We may then obtain expressions for the asymptotic variance of these estimators by applying central limit theorems for martingales [Gordin, 1969, Hall and Heyde, 1980]. Intuitively, one can verify that $\hat{g}_k - \Pi_k \hat{g}_{\sigma(k)} = g$ by checking that each term in $\hat{g}_k$ matches a term in $\Pi_k \hat{g}_{\sigma(k)}$ except for the first term g, so that all terms besides g cancel, provided the sums and integrals can be rearranged as needed. Proposition 3.1 is proved in Appendix B.2.

We now define

$$U_k = \int \pi(dx)\{ff^T - (\Pi_k f)(\Pi_k f)^T\} \tag{3.9}$$

$$V_k = \int \pi(dx)\{f\hat{g}_{\sigma(k)}^T - (\Pi_k f)(\Pi_k \hat{g}_{\sigma(k)}^T)\} \tag{3.10}$$

as well as $U = K^{-1}\sum_{k=1}^K U_k$ and $V = K^{-1}\sum_{k=1}^K V_k$. These quantities arise in the expressions for the asymptotic variance of the control variate estimators (3.6) and (3.7). The quantities U_k and V_k can be interpreted as a conditional variance and a conditional covariance, respectively. Suppose $X_0 \sim \pi$, and that the distribution of X_1 given X_0 is $\Pi_k(X_0, \cdot)$. Then U_k is the conditional matrix of $f(X_1)$, and V_k is the conditional covariance of $f(X_1)$ and $\hat{g}_{\sigma(k)}(X_1)$.

In the remainder, for positive semidefinite matrices A and B, we say $A \geq B$ if $A - B$ is positive semidefinite. We say $A > B$ if $A - B$ is positive semidefinite with at least 1 nonzero eigenvalue. We define $\leq$ and $<$ similarly. Further, we let $A^\dagger$ denote the pseudoinverse of A. We write $N(0, \Sigma)$ for a d-dimensional multivariate normal distribution, where we allow the variance Σ to be positive semidefinite rather than strictly positive definite.

Theorem 3.1. *Assume (A.1)–(A.4) and (B.1)–(B.4). We have*

$$M^{-1/2} S_M^{VW} \xrightarrow{d} N(0, \Sigma_C)$$

. The variance Σ_C can be written as

$$\Sigma_C = \int \pi(dx) gg^T + K^{-1}\sum_{k=1}^K \sum_{t=1}^\infty \int \pi(dx)\{g(P_k g)^T + (P_k g)g^T\} \tag{3.11}$$
$$+ K^{-1}\sum_{k=1}^K C_{\sigma(k)}^T U_k C_{\sigma(k)} - C_{\sigma(k)}^T V_k - V_k^T C_{\sigma(k)},$$

and Σ_C is minimized at $C_{\sigma(k)} = \tilde{C}_{\sigma(k)}$, where $\tilde{C}_{\sigma(k)} = U_k^\dagger V_k$.

In Theorem 3.1, we establish the convergence result $M^{-1/2} S_M^{VW} \xrightarrow{d} N(0, \Sigma_C)$ for the variable weight control variate estimate in (3.7). The proof of Theorem 3.1 is in Appendix B.2. In general, the optimal weight expression $\tilde{C}_{\sigma(k)} = U_k^{\dagger} V_k$ in Theorem 3.1 appears daunting, since the V_k contain integrals involving the Poisson equation solutions $\hat{g}_{\sigma(k)}$. Corollary 3.1 below establishes a simpler form for the optimal control variate weight for the fixed weight control variate estimator $M^{-1} S_M^{FW}$ in (3.6), in the setting of Gibbs sampling.

Corollary 3.1 (Fixed weight control variates)**.** *Suppose (A.1), (A.2)–(A.3), and (B.1)–(B.4) hold. Then for the fixed weight scheme with $C_1 = \cdots = C_K = C$, we have*

$$\Sigma_C = \int \pi(dx) g g^T + K^{-1} \sum_{k=1}^{K} \sum_{t=1}^{\infty} \int \pi(dx) \{ g (P_k g)^T + (P_k g) g^T \} + C^T U C - C^T V - V^T C,$$

and Σ_C is minimized at $C = \tilde{C}$, where $\tilde{C} = U^{\dagger} V$. If (A.1b) also holds, then we have the simplified expression $V = \int \pi(dx) f g^T$.

A detailed proof of Corollary 3.1 is given in Appendix B.2. We outline the steps to obtain the simplified representation for V under (A.1b) here. First, we have $\int \pi(dx)(\Pi_k f)(\Pi_k \hat{g}_{\sigma(k)}^T) = \int \pi(dx) f \Pi_k \hat{g}_{\sigma(k)}^T$, by the idempotence and reversibility of Π_k under (A.1b). Then, we rearrange and use Proposition 3.1 to obtain

$$\begin{aligned} V &= K^{-1} \sum_{k=1}^{K} \int \pi(dx) \{ f \hat{g}_{\sigma(k)}^T - f \Pi_k \hat{g}_{\sigma(k)}^T \} \\ &= K^{-1} \sum_{k=1}^{K} \int \pi(dx) f (\hat{g}_k - \Pi_k \hat{g}_{\sigma(k)})^T = \int \pi(dx) f g^T. \end{aligned}$$

Corollary 3.1 shows that the formula for the optimal control variate weight simplifies substantially in the Gibbs sampling setting where (A.1b) holds. In general, the quantity $V = K^{-1} \sum_{k=1}^{K} f \hat{g}_{\sigma(k)}^T - (\Pi_k f)(\Pi_k \hat{g}_{\sigma(k)})^T$ depends on the Poisson equation

solutions $\hat{g}_k$ from Proposition 3.1, whereas under (A.1b), the formula for $\tilde{C}$ no longer involves the $\hat{g}_{\sigma(k)}$ explicitly. In contrast, for general transition kernels or Gibbs sampling without fixed control variate weights, the optimal weights are more complicated to obtain due to the presence of the Poisson equation solutions $\hat{g}_k$ in V (V_k).

We now establish a variance reduction result for the Rao-Blackwellized estimator (3.5), in the setting of deterministic sweep Gibbs sampling. To our knowledge, this result is new and no prior theoretical results exist for general Rao-Blackwellization schemes for deterministic sweep Gibbs sampling. We use Σ_0 to denote the variance of the ordinary empirical estimate (3.1) with control variate weight $C = 0$ in (3.6). We use $\Sigma_1 = \Sigma_{RB}$ to denote the variance of the Rao-Blackwellized estimate (3.5), which results from the choices $f = g$ and $C = I_{d\times d}$ in (3.6).

Corollary 3.2 (Rao-Blackwellized Gibbs sampling)**.** *Suppose (A.1b), (A.2)–(A.3), and (B.1)–(B.3) hold, and $f = g$. We have*

$$\Sigma_0 = K^{-1}\sum_{k=1}^{K}\left[\int \pi(dx)gg^T + \sum_{t=1}^{\infty}\int \pi(dx)\{g(P_k^t g)^T + (P_k^t g)g^T\}\right]$$

and

$$\Sigma_1 = \Sigma_0 - \int \pi(dx)gg^T - K^{-1}\sum_{k=1}^{K}\int \pi(dx)(\Pi_k g)(\Pi_k g)^T \leq \Sigma_0.$$

The result follows from collecting terms and simplifying the variance from Theorem 3.1 with $C = I_{d\times d}$ and $f = g$. We again exploit (A.1b) to use the simplified formula for V from Corollary 3.1. When $f = g$, then $V = \int \pi(dx)fg^T = \int \pi(dx)gg^T$. A detailed proof of Corollary 3.2 is given in Appendix B.2.

Corollary 3.2 shows that the asymptotic variance is always smaller for the Rao-Blackwellized average Σ_1 than for the empirical average Σ_0. Thus, for deterministic

sweep Gibbs sampling, the apparently naive Rao-Blackwellization strategy of averaging the conditional expectation of the integrand g, with respect to whichever transition kernel is being used to update X_t, leads to an improved asymptotic variance.

3.3.3 Estimating the Optimal Control Variate Weight

In order to implement the control variate estimators (3.6) and (3.7), it is necessary to choose the control variate weights C and C_k, respectively. For the fixed weight Gibbs sampler, we show that it is possible to estimate the exact optimal control variate weight. In general settings, we propose estimating an arbitrarily accurate approximation of the optimal weight. Since $\tilde{C} = U^\dagger V$ and $\tilde{C}_{\sigma(k)} = U_k^\dagger V_k$, one can compute estimates of the optimal weights via Markov chain Monte Carlo estimates of U and V or U_k and V_k.

Before we introduce our weight estimators, we recall the definitions $V_k = \int \pi(dx)\{f\hat{g}_{\sigma(k)}^T - (\Pi_k f)(\Pi_k \hat{g}_{\sigma(k)})^T\}$ and $V = K^{-1}\sum_{k=1}^K V_k$. For integers $B \geq 0$, we define

$$V_k^B = \sum_{t=0}^{B} \int \pi(dx)\left\{f(P_{\sigma(k)}^t g)^T - (\Pi_k f)(\Pi_k P_{\sigma(k)}^t g)^T\right\}$$

$$V^B = K^{-1}\sum_{k=1}^{K} V_k^B$$

as well as $C_{\sigma(k)}^B = U_k^\dagger V_k^B$ and $C^B = U^\dagger V^B$. The quantity V_k^B can be viewed as an approximation of V_k resulting from including the first $B+1$ terms in the Poisson equation solution $\hat{g}_{\sigma(k)}$, or equivalently, that drops lag-t autocovariance terms in V_k for $t > B$. From Proposition 3.1, we have the deterministic convergence results $\lim_{B\to\infty} V_k^B = V_k$ and $\lim_{B\to\infty} V^B = V$. Thus, $C^B \to \tilde{C}$ and $C_{\sigma(k)}^B \to \tilde{C}_{\sigma(k)}$ as $B \to \infty$, where $\tilde{C}$ and $\tilde{C}_{\sigma(k)}$ are optimal weights as defined in Corollary 3.1 and Theorem 3.1.

For convenience in referring to the K subchains of a K-component deterministic

sweep sampler, we define $Z(k) = \{i \in \mathbb{N} : i = k - 1 + nK \text{ for some } n \in \mathbb{N}\}$, so that $Z(k)$ is the set of integers t such that the transition kernel Π_k is used to generate X_{t+1} from X_t. That is, for $t \in Z(k)$ we have $X_{t+1}|(X_0, ..., X_t) \sim \Pi_k(X_t, \cdot)$. We also define $Z(k, y) = Z(k) \cap \{0, 1, ..., y\}$.

We define the estimators

$$\hat{U}_M = M^{-1} \sum_{t=0}^{M-1} \{f(X_{t+1}) - \Pi_{\sigma^t(1)} f(X_t)\}\{f(X_{t+1}) - \Pi_{\sigma^t(1)} f(X_t)\}^T \tag{3.12}$$

$$\hat{U}_{k,M} = M^{-1} \sum_{t \in Z(k,M-1)} \{f(X_{t+1}) - \Pi_k f(X_t)\}\{f(X_{t+1}) - \Pi_k f(X_t)\}^T \tag{3.13}$$

$$\hat{V}_M^B = M^{-1} \sum_{t=0}^{M-1} \{f(X_{t+1}) - \Pi_{\sigma^t(1)} f(X_t)\} \sum_{r=0}^{B} g(X_{t+1+r}) \tag{3.14}$$

$$\hat{V}_{k,M}^B = M^{-1} \sum_{t \in Z(k,M-1)} \{f(X_{t+1}) - \Pi_k f(X_t)\} \sum_{r=0}^{B} g(X_{t+1+r}) \tag{3.15}$$

$$\hat{V}_M^{Gibbs} = M^{-1} \sum_{t=0}^{M-1} f(X_t) g(X_t)^T \tag{3.16}$$

and propose the control variate weight estimators

$$\hat{C}_M^{Gibbs} = \hat{U}_M^\dagger \hat{V}_M^{Gibbs} \tag{3.17}$$

$$\hat{C}_M^B = \hat{U}_M^\dagger \hat{V}_M^B \tag{3.18}$$

$$\hat{C}_{k,M}^B = \hat{U}_{k,M}^\dagger \hat{V}_{k,M}^B. \tag{3.19}$$

These estimators are empirical estimators of the corresponding quantities U, U_k, V^B, etc., based on Markov chain samples of size M. Let $\Sigma_{\tilde{C}}^{VW,B}$ denote the asymptotic variance from Theorem 3.3 for the variable weight control variate estimator with weights $C_{\sigma(k)}^B$, for $k = 1, ..., K$. Further, let $\Sigma_{\tilde{C}}^{FW,B}$ denote the variance for the fixed weight control variate estimator with weight C^B, and let $\Sigma_{\tilde{C}}$ denote the variance for the fixed weight control variate estimator with the optimal weight $\tilde{C} = U^\dagger V$. For technical

reasons, we add an additional assumption. Let $\nu << \pi$ indicate that the measure ν is absolutely continuous with respect to π.

(C.1) (Absolute continuity) $\nu << \pi$ for the initial measure ν for which $X_0 \sim \nu$

We impose Assumption (C.1) in order to ensure that the generalized inverse estimates converge properly, that is, $\hat{U}_M^\dagger \to U^\dagger$ and $\hat{U}_{k,M}^\dagger \to U_k^\dagger$ almost surely. When the limiting matrices U and U_k are invertible, then Assumption (C.1) is unnecessary for the conclusions of Theorem 3.2 to hold and can be dropped.

Theorem 3.2. *Suppose (A.1)–(A.4), (B.1)–(B.4), and (C.1) hold. Then we have* $\hat{C}_M^B \overset{a.s.}{\to} C^B$, $\hat{C}_{\sigma(k),M}^B \overset{a.s.}{\to} C_{\sigma(k)}^B$, *and*

$$M^{-1/2} \sum_{t=0}^{M-1} g(X_t) - (\hat{C}_{\sigma^t(1),M}^B)^T \{f(X_{t+1}) - \Pi_{\sigma^t(1)} f(X_t)\} \overset{d}{\to} N(0, \Sigma_{\tilde{C}}^{VW,B})$$
$$M^{-1/2} \sum_{t=0}^{M-1} g(X_t) - (\hat{C}_M^B)^T \{f(X_{t+1}) - \Pi_{\sigma^t(1)} f(X_t)\} \overset{d}{\to} N(0, \Sigma_{\tilde{C}}^{FW,B})$$

Furthermore, suppose Assumptions (A.1b), (A.2)–(A.4), (B.1)–(B.4), and (C.1) hold. Then $\hat{C}_M^{Gibbs} \overset{a.s.}{\to} \tilde{C}$, *and*

$$M^{-1/2} \sum_{t=0}^{M-1} g(X_t) - (\hat{C}_M^{Gibbs})^T \{f(X_{t+1}) - \Pi_{\sigma^t(1)} f(X_t)\} \overset{d}{\to} N(0, \Sigma_{\tilde{C}})$$

The proof of Theorem 3.2 is in Appendix B.2. Theorem 3.2 shows for general Markov chains that one can replace the exact truncated control variate weights C^B and $C_{\sigma(k)}^B$ by estimates without affecting the asymptotic variance from Theorem 3.1. The asymptotic variance will in general be suboptimal for any finite truncation level B, since C^B and C_k^B will in general differ from the optimal weights $\tilde{C}$ and $\tilde{C}_{\sigma(k)}$. In

practice, the choice of B seems to be somewhat challenging, since when B is large relative to the Monte Carlo sample size M, the estimates $\hat{C}_M^B$ and $\hat{C}_{\sigma(k),M}^B$ can be expected to perform badly. However, in the Gibbs sampling setting, where Assumption (A.1b) holds, this difficulty vanishes. Theorem 3.2 shows that for Gibbs sampling, the estimator $\hat{C}_M^{Gibbs}$ leads to the control variate estimate with the optimal asymptotic variance in Corollary 3.1.

3.4 Theoretical Comparisons

3.4.1 Comparison to Liu et al. [1994]

We next compare the asymptotic variance for our control variate estimators (3.6) and (3.7) to the asymptotic variance for the Rao-Blackwellization estimator from Liu et al. [1994]. The estimator in Liu et al. [1994] applies in the so-called *data augmentation* Gibbs sampling setting, in which a two-component Gibbs sampler has components Π_1 and Π_2, and the integrand g satisfies $\Pi_k g = g$ a.e. π for at least one of the components k.

We assume the control variate basis function $f = g$, and we assume g satisfies

(D.1) (Data augmentation) $\Pi_2 g = g$ a.e. π.

(D.2) (Non-degeneracy) $\int \pi(dx) g g^T$ is a positive definite matrix.

In this setting, the estimator from Liu et al. [1994] is

$$S_M^{LWK}/M = M^{-1} \sum_{t=0}^{M-1} \Pi_1 g(X_t). \tag{3.20}$$

We remark that when $f = g$ and the data augmentation assumption (D.1) holds, then the asymptotic variance of the variable weight control variate scheme only depends

on the weight matrix C_2, but not on C_1. Thus, for the purpose of comparison, it is sufficient to consider the fixed weight control variate estimator (3.6) only.

Denote by Σ_0, Σ_1, Σ_2, Σ_{LWK}, and $\Sigma_{\tilde{C}}$ the variances obtained by the empirical estimator (3.1), the Rao-Blackwellized estimator (3.5), the fixed weight control variate scheme with $C = 2I_{d\times d}$, the conditioning estimate in (3.20) due to Liu et al. [1994], and the fixed weight control variate scheme with the optimal weight $C = \tilde{C}$, respectively. Additionally, define $A = \int \pi(dx)gg^T$ and $B = \int \pi(dx)(\Pi_1 g)(\Pi_1 g)^T$. Then we have the following result.

Proposition 3.2. *Assume $K = 2$, and that (A.1b), (A.2)–(A.3), (B.1)–(B.3), and (D.1)–(D.2) hold. Then $\Sigma_2 = \Sigma_{LWK}$, and $\Sigma_{\tilde{C}} \leq \Sigma_2 < \Sigma_1 < \Sigma_0$, with*

$$\begin{aligned}
\Sigma_{\tilde{C}} - \Sigma_2 &= -2B(A-B)^{-1}B \leq 0 \\
\Sigma_2 - \Sigma_1 &= -(A+3B)/2 < 0 \\
\Sigma_1 - \Sigma_0 &= -(B+3A)/2 < 0
\end{aligned}$$

By Proposition 3.2, we have $\Sigma_{\tilde{C}} \leq \Sigma_2 = \Sigma_{LWK}$ and thus the variance from our fixed weight control variate estimator (3.6) will be no larger than the variance from the approach of Liu et al. [1994]. When Σ_{LWK} is nonzero, then we have the strict inequality $\Sigma_{\tilde{C}} < \Sigma_{LWK}$. In addition, since $\Sigma_1 < \Sigma_0$ and $\Sigma_2 < \Sigma_1$, the variance Σ_{LWK} will be lower than the variances from both the empirical estimator as well as the Rao-Blackwellization approach in Corollary 3.2. Thus the optimal asymptotic variance of the fixed weight scheme is equal to the optimal asymptotic variance for the variable weight scheme. The proof of Proposition 3.2 is given in Appendix B.2.

In the scalar g case, the only increases in cost of the control variate estimate relative to the Liu et al. [1994] estimate are incurred in the computation and aggregation of the $g(X_t)$ values. We expect these costs will generally be smaller than the

costs of computing and aggregating the $\Pi_1 g(X_t)$, and will often be smaller than the cost of sampling additional X_t. When $d > 1$, so that g has d components, the control variate approach incurs an additional cost per iteration of d^2 operations to estimate U and V, as well as as a one-time cost of d^3 to invert U at the end of the simulation. For large d, one can consider approximate versions of the control variate scheme to reduce the cost. For example, we might take C to be a scalar and consider minimizing the asymptotic mean squared error per coefficient, which is proportional to $1^T \Sigma_C 1$. The optimal scalar choice of C for this criterion is $\tilde{C} = (1^T V 1)/(1^T U 1)$, where both the numerator and denominator can be estimated with an $O(d)$ cost per iteration by exploiting the outer product structure of U and D.

3.4.2 Comparison to Dellaportas and Kontoyiannis [2012]

We also compare the asymptotic variances resulting from deterministic and random sweep Gibbs sampling schemes with $K = 2$ components. Our control variate methodology applies to the deterministic sweep version of Gibbs sampling, whereas Dellaportas and Kontoyiannis [2012] applies only to random sweep Gibbs sampling schemes. For kernels Π_1 and Π_2, we define the random sweep kernel $Q = (\Pi_1 + \Pi_2)/2$. Additionally, we define the function

$$h = g - C^T(f - Qf). \tag{3.21}$$

Then we have the following result.

Theorem 3.3. *Suppose (A.1b)–(A.3) and (B.1)–(B.4) hold, and that the number of components $K = 2$.*

Write $\Sigma_{\tilde{C}}$ and Σ_{RS} respectively for the variances from the optimal fixed weight deterministic sweep scheme and the corresponding random sweep scheme with weight

chosen as in Dellaportas and Kontoyiannis [2012]. Then

$$\Sigma_{\tilde{C}} - \Sigma_{RS} = -(\bar{C} - \tilde{C})^T U^\dagger (\bar{C} - \tilde{C}) - \sum_{t=1}^{\infty} \int \pi(dx) \bar{h} (Q^t \bar{h})^T \leq 0.$$

where $\tilde{C}$ is the optimal weight for the deterministic sweep scheme, where $\bar{C}$ is the optimal weight for the random sweep scheme, and $\bar{h} = g - \bar{C}^T(f - Qf)$. The inequality is strict except when $\Sigma_{RS} = 0$.

Theorem 3.3 shows it is statistically more efficient to use our fixed weight, deterministic sweep control variate scheme rather than the random sweep scheme with weight as chosen in Dellaportas and Kontoyiannis [2012], for general Gibbs sampling schemes with $K = 2$ components, making our methodology useful for practical applications. For example, two component Gibbs samplers arise in applications in the common data augmentation Gibbs sampling setting [see, e.g. Robert and Casella, 2004]. Additionally, Markov random field models often contain a bipartite graph structure which leads to two-component Gibbs sampling schemes.

3.4.3 Connection Between Rao-Blackwellization and Control Variates

We further show that the Rao-Blackwellization estimator (3.5) can be viewed as an approximate control variate scheme. Suppose the control variate basis function $f = g$, and assume for simplicity that the matrices U_k in (3.9) are positive definite. Then under the assumptions of Theorem 3.1, the optimal control variate weights are $\tilde{C}_{\sigma(k)} = U_k^{-1} V_k$, where $V_k = \int \pi(dx)\{g\hat{g}_{\sigma(k)}^T - (\Pi_k g)(\Pi_k \hat{g}_{\sigma(k)})^T\}$. Now, consider using the approximation

$$\hat{g}_{\sigma(k)} = g + \Pi_{\sigma^1(k)} g + \Pi_{\sigma^1(k)} \Pi_{\sigma^2(k)} g + ... \approx g, \tag{3.22}$$

within V_k, where the infinite sum in the Poisson equation solution is truncated after a single term. Then we are left with

$$V_k \approx \int \pi(dx)\{gg^T - (\Pi_k g)(\Pi_k g)^T\} = U_k. \tag{3.23}$$

Thus, under the one term approximation of $\hat{g}_{\sigma(k)}$, we obtain $\tilde{C}_{\sigma(k)} = I_{d\times d}$ for each k. Similarly, for the fixed weight scheme, we obtain $\tilde{C} = I_{d\times d}$. But these choices of weights in (3.6) and (3.7) both lead to the estimator

$$M^{-1}\{g(X_0) - g(X_M)\} + M^{-1}\sum_{t=0}^{M-1} \Pi_{\sigma^t(1)} g(X_t),$$

which is asymptotically equivalent to the Rao-Blackwellized estimator in (3.5).

An inspection of the estimators (3.14) and (3.15) shows that setting $B = 0$ and $f = g$ can be interpreted as invoking the approximation (3.23). In particular, in the setting where $f = g$ and $B = 0$, then $\hat{U}_k \to U_k$ and $\hat{V}_k^B \to U_k$ almost surely, so that the estimators from the control variate approach with $B = 0$ will be identical to the estimates from the Rao-Blackwellized estimate (3.5). Figure 3.3 in Section 3.5 provides a numerical demonstration of this fact.

3.5 Numerical Examples

In this section, we present two concrete examples to illustrate the theory and methods developed in Sections 3.2–3.4. Since the setting of Theorem 3.1 is asymptotic, these examples provide a test of whether or not the asymptotic variance reduction properties of the control variate schemes materialize at reasonable finite sample sizes. Additionally, the examples provide some indication of the performance of the control variate schemes when the optimal control variate weight is estimated from Markov chain Monte Carlo. The first example, a bivariate normal, is relatively simple, but

provides valuable insight into the overall performance of the methods. The second example, an Ising model, is of interest in many fields.

3.5.1 Bivariate Normal

We consider a Gibbs sampling setting where π is a standard bivariate normal distribution with $\mu = [0, 0]^T$, $\sigma_1^2 = \sigma_2^2 = 1$, and correlation ρ between the two components. We take $K = 2$ and define Π_k for $k = 1, 2$ as the kernels which update states $X_t = (X_{1,t}, X_{2,t})$ by setting $X_{k,t+1} = X_{k,t}$ and drawing $X_{\sigma(k),t+1} \sim N(X_{k,t}, 1 - \rho^2)$. It can be shown that for $k = 1, 2$, the composition kernels P_k^2 satisfy Assumption (A.2) with the measure $\psi = \pi$. Additionally, Lemma B.3, combined with (A.1b) and (A.2), shows that (A.4) also holds. We verify in Lemma B.9 (Appendix A) that Assumption (A.3) also holds for P_1^2 and P_2^2 with the functions $V_1(x) = x_1^2 + rx_2^2 + 1$ and $V_2(x) = rx_1^2 + x_2^2 + 1$ for appropriately chosen $r > 0$. These V_k satisfy (B.1). Thus, Theorem 3.1 can be applied to each of the following examples.

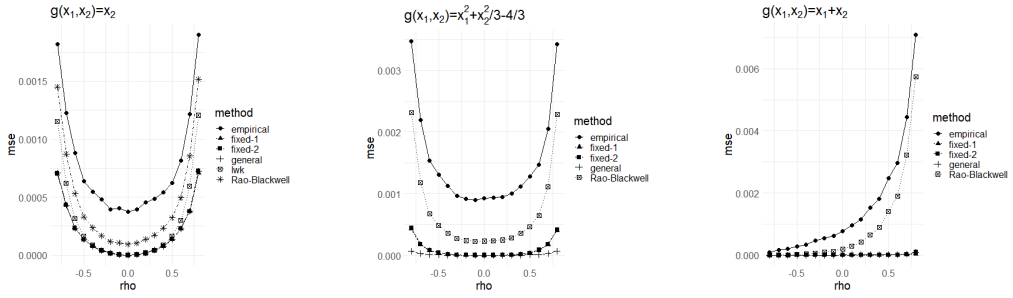
Our numerical results for the bivariate normal setting are shown in Figure 3.1. We compare the simulation mean squared error (MSE) for multiple estimators of $\int \pi(dx)g(x)$, for three different integrands g. For each example integrand g, we set the control variate basis function $f = g$. For the fixed weight control variate approach, we compare the two fixed weight control variate weight estimators (3.17) and (3.18). Empirically, the estimator (3.17) performs better than the estimator (3.18), which is expected since (3.17) requires the estimation of fewer covariance terms, and also since the estimator (3.18) estimates the truncated weight $\tilde{C}^B$ rather than the optimal weight $\tilde{C}$. We set $B = 10$. We computed MSE in each setting based on 100 simulated averages using $M = 2000$ draws, at each value of ρ.

Figure 3.1a shows a data augmented setting, where the integrand $g(x_1, x_2) = x_2$, so that g only depends x_2. Figure 3.1a compares the simulation asymptotic variances of S_M/M as the bivariate normal correlation coefficient ρ varies for the empirical (3.1), Rao-Blackwellization (3.5), fixed weight control variate (3.6), variable weight control variate (3.7), and LWK (3.20) estimators. We see that the control variate and LWK estimators outperform the empirical and Rao-Blackwell estimators, with the empirical estimator performing the worst. The LWK and control variate estimators perform similarly, although for large $|\rho|$, the control variate estimates outperform the LWK estimates. For $\rho = 0$, the LWK estimate is exactly $\Pi_1 g(x_1, x_2) = 0$ for all (x_1, x_2). Thus, at $\rho = 0$, the finite sample performance of the LWK estimate is better than the control variate estimate estimates, which accrue some error in finite samples due the estimation of $\tilde{C}$. This error vanishes asymptotically with $M^{1/2}$ normalization.

Figure 3.1b shows results for the integrand $g(x_1, x_2) = x_1^2 + x_2^2/3 - 4/3$. Since this g depends on both x_1 and x_2, the approach by Liu et al. [1994] no longer applies. Figure 3.1b compares the variances of the fixed weight and variable weight control variate estimators (3.6) and (3.7), as well as the Rao-Blackwellized and empirical estimators (3.5) and (3.1), as ρ varies. For this example, the variable weight control variate estimates outperform the fixed weight estimates. The fixed weight estimates substantially outperform the empirical and Rao-Blackwellized estimates.

Figure 3.1c shows results for the integrand $g(x_1, x_2) = x_1 + x_2$. The control variate estimates (both fixed-weight and general) attain 0 asymptotic variance, even though the empirical and Rao-Blackwellization estimates have positive asymptotic variance. This can be explained as follows. Taking the random sweep kernel $Q = (\Pi_1 + \Pi_2)/2$, we have $Qg = (1 + \rho)g/2$, so that g is an eigenfunction of Q with

eigenvalue $\lambda = (1+\rho)/2$. Therefore, taking $c = 1/(1-\lambda)$ and $f = g$ gives $g(x) - c\{g(x) - Qg(x)\} = 0$ a.e. π. Thus, the optimal random sweep Gibbs sampling scheme from Theorem 3.3 has an asymptotic variance of 0. From Theorem 3.3, we have that the optimal fixed weight deterministic sweep control variate scheme must also attain 0 asymptotic variance. Figure 3.1c demonstrates that the control variate estimates indeed achieve 0 asymptotic variance, as the MSE for the control variate estimates are nearly exactly 0 except for large ρ, where finite sample error in estimating $\tilde{C}$ causes the MSE to be just barely above 0. On the other hand, the empirical and Rao-Blackwell estimators perform much worse, particularly for larger ρ.



(a) (b) (c)

Figure 3.1: Mean squared error (MSE) for different ρ values in a bivariate normal π example, with the integrands (a) $g(x_1, x_2) = x_2$, (b) $g(x_1, x_2) = x_1^2 + x_2^2/3 - 4/3$, and (c) $g(x_1, x_2) = x_1 + x_2$. The estimator labels empirical, Rao-Blackwell, fixed-1, fixed-2, variable, and LWK correspond, respectively, to Equations (3.1), (3.5), (3.6) with weight from (3.17), (3.6) with weight from (3.18), (3.7) with weight from (3.19), and (3.20).

3.5.2 Ising Model

Next, we consider a one-parameter Ising model on an $n \times n$ square grid of cells. We take $n = 20$. The state space is $X = \{-1, 1\}^{n^2}$, where $\mathscr{X} = 2^X$ is the set of all subsets of X. The probability density function of the desired stationary measure

with respect to counting measure on $(X, \mathscr{X})$ is $\pi(x) = \exp\{\eta T(x) - \xi(\eta)\}$, where the sufficient statistic $T(x) = \sum_{i\sim j, i<j} x_i x_j$, and the notation $i \sim j$ indicates that i is a neighbor of j. Thus, the contribution from a given i,j pair with $i \sim j$ is positive when x_i and x_j are equal, and negative otherwise. The term $\xi(\eta) = \log[\sum_{x\in X} \exp\{\eta T(x)\}]$ is a log normalizing constant. We write x_{-i} for the values at all sites of x except site i. Also, we use x^i to denote the configuration x with the ith value flipped, so that $(x_i)^i = -x_i$, and $(x^i)_{-i} = x_{-i}$.

We consider deterministic sweep Gibbs samplers, as well as deterministic sweeps composed of Metropolis-type updates. We first define Gibbs sitewise kernels for each $i = 1, ..., n$ by

$$\begin{aligned}\Pi_i(x, \{x'\}) &= I(x'_{-i} = x_{-i})\pi(x')\{\pi(x) + \pi(x^i)\}^{-1} \\ &= I(x'_{-i} = x_{-i}) \exp\{\eta T(x')\} \left[\sum_{\substack{x^*\in X: \\ x^*_i \in \{-1,1\} \\ x^*_{-i} = x_{-i}}} \exp\{\eta T(x^*)\} \right]^{-1} \qquad \forall x, x' \in X.\end{aligned}$$

Each Π_i is a Gibbs kernel with respect to (h_i, π) for the coordinate projection $h_i : X \to \mathbb{R}, x \to x_{-i}$. Further, we define sitewise Metropolis kernels Q_i by

$$Q_i(x, x') = I(x' = x^i)0.9a_i(x) + I(x' = x)[0.1 + 0.9\{1 - a_i(x)\}]$$

where $a_i(x) = \min\{\pi(x^i)/\pi(x), 1\}$. Each Metropolis kernel Q_i corresponds to proposing to flip the value at the ith coordinate with probability 0.9, and then accepting any flip with probability $a_i(x)$. Note that we do not always propose to flip the value at site i. It is straightforward to show that the $Q_i(x, x')$ satisfy the reversibility condition (A.1a).

For each of the Gibbs and Metropolis sitewise update types, we consider two different types of compositions of the sitewise updates, so that in total, four Markov chain schemes are considered. First, in the raster sweep, we construct Markov chains $\{X_t\}_{t=0}^{\infty}$ using the update $\Pi_{\sigma^t(1)}$ (resp., $Q_{\sigma^t(1)}$) at each time step t, where the sites are traversed sequentially proceeding first down each column of the grid, and then across the columns in order.

We next consider a checkerboard sweep, where we partition the bipartite lattice into two components W_1 and W_2, as in Figure 3.2a, and then update each component in sequence. To update each component, we construct composition kernels

$$H_k(x, x') = \left\{ \prod_{i \notin W_k} \Pi_i \right\} (x, x') \quad k = 1, 2$$

for the Gibbs kernels and

$$J_k(x, x') = \left\{ \prod_{i \notin W_k} Q_i \right\} (x, x') \quad k = 1, 2$$

for the Metropolis kernels. Because of the lattice neighborhood structure of the sites, any ordering of the Π_i (resp., Q_i) in the compositions H_k (resp., J_k) leads to an equivalent transition kernel, and both composition kernels can be implemented using independent Bernoulli draws at every site not in W_k. For the checkerboard sweeps, we construct Markov chains $\{X_t\}_{t=0}^{\infty}$ by applying the kernel $H_{\sigma(t)}$ (resp., $J_{\sigma(t)}$) at each time t. For example, for the Gibbs-based sampler, H_1 is used to obtain X_1 from X_0, and H_2 is used to obtain X_2 from X_1. Thus, at each step, all of the cells are updated in one of the components W_k.

It is straightforward to verify that H_k itself is a Gibbs kernel with respect to (h_k, π), where $h_k : X \to \mathbb{R}^{|W_k|}, x \to x_{W_k}$ denotes the coordinate projection which ob-

tains the values in W_k. All four sweeps are irreducible with respect to the uniform probability measure on $(X, \mathscr{X})$. Additionally, taking $C = X$, $b = 1$, $V_k(x)$ to be the constant function $V_k(x) = 2 \max_{x' \in X} |T(x')|$ for $k = 1, 2$, and $g(x) = T(x) - \int \pi(dx')T(x')$ ensures (A.3) holds for chains composed of either the Gibbs and Metropolis updates. Finally, the aperiodicity condition in Assumption (A.4) holds for the composition chains P_k^K (where $K = n^2$ for the raster sweeps and $K = 2$ for the checkerboard sweeps). For both the Gibbs and Metropolis updates, this follows from the fact that

$$P_k^K(x, \{x\}) > 0, \quad \forall x \in X$$

for raster and checkerboard scans with either Gibbs or Metropolis sitewise updates. For the Gibbs sampler chains, we could alternatively have verified (A.4) by using Lemma B.3.

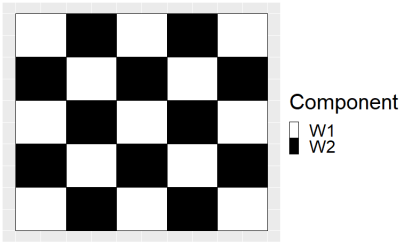
For each sweep type, we considered estimators based on 2000 sweeps through the grid. For the raster sweeps, we used $B = 5(400) = 2000$, which corresponds to lags resulting from 5 complete sweeps through the grid. For the checkerboard sweeps, we used $B = 5(2) = 10$, which also corresponds to 5 sweeps through the grid.

In Figures 3.2 and 3.3, method="fixed-1" and "fixed-2" correspond to estimating $\tilde{C}$ via (3.17) and (3.18), respectively. In Figure 3.2, MSE is computed based on the empirical average squared error of 100 estimated means, where each estimated mean used 2000 complete sweeps through the grid. That is, $M = 2000(n^2) = 800000$ for the raster sweeps, and $M = 2000(2) = 4000$ for the checkerboard sweeps.

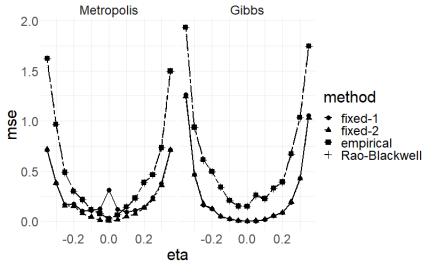
Figure 3.2 shows the MSE for the checkerboard and raster sweeps with Metropolis and Gibbs updates. Figure 3.2b shows the performance of the estimators for checkerboard sweep. Figure 3.2c shows that for the raster sweep, the variable weight

control variate estimator performs much worse than the other estimators, likely due to the fact that n^2 weights must be estimated for this scheme. Figure 3.2d shows the MSE for the raster sweep, for the estimators remaining after excluding the variable weight control variate estimate. In Figure 3.2d, the empirical and Rao-Blackwellized schemes nearly overlap for both Gibbs and Metropolis schemes. For Gibbs sampling, the fixed weight estimator based on (3.16) performs best, as expected, but the fixed weight batch estimator also performs well. For Metropolis sampling, the fixed weight estimators perform similarly, but the MSE for the batch means estimator is often smaller than for the fixed weight estimator based on (3.16).

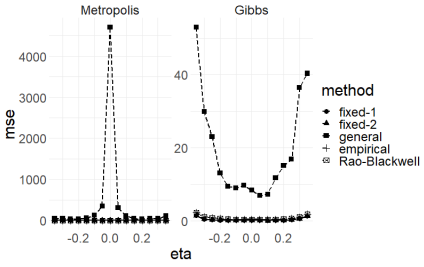
For each value of the Ising model parameter η, we estimated the true value of $\int \pi(dx)g(x)$ using a long checkerboard sweep run with 100000 complete sweeps through the grid, so that $M = 100000(2) = 200000$. We used the Rao-Blackwellized estimator to compute the means.



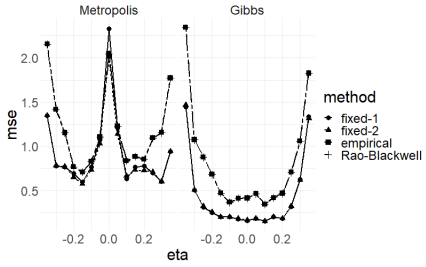
(a) Conditional independence structure for the square lattice Ising model



(b) Ising model with checkerboard sweep



(c) Ising model with raster sweep

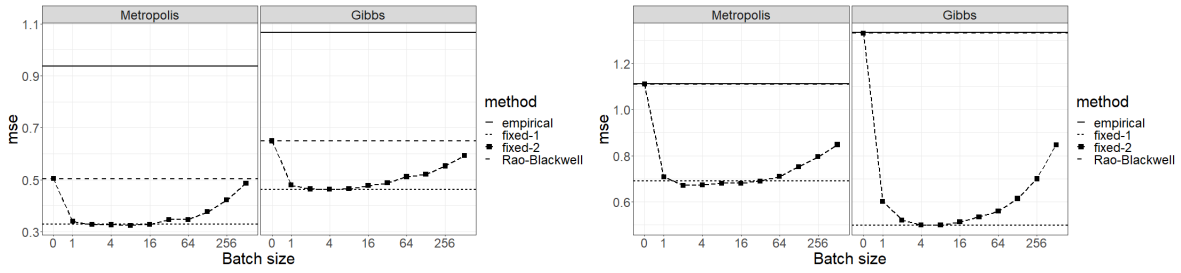


(d) Ising model with raster sweep, with variable weight control variate estimator exluded

Figure 3.2: Mean squared error (MSE) for the Ising model simulation example at different values of η, for deterministic raster and checkerboard sweeps.

We also examined the effect of the truncation level B on the performance of the various estimates (Figure 3.3). Our study allowed us to confirm three notable theoretical properties of the control variate estimators. First, for Gibbs samplers, the optimal fixed weight control variate formula based on (3.17) (horizontal line with smaller dashes) always performed better than the corresponding estimator (3.18), as expected. For the Gibbs samplers, the approximate estimator (3.18) performed best near $B = 4$, where the estimation performances were nearly identical to, but slightly worse than, the estimates using (3.17). Second, our results for each setting demonstrate empirically that setting the batch size $B = 0$ is asymptotically equivalent to the Rao-Blackwellization approach (3.5) (horizontal line with larger dashes). Third,

for the Metropolis samplers, using (3.18) with the best-performing B in each setting leads to a better control variate weight than using the fixed weight estimator (3.17) (horizontal line with smaller dashes).



(a) Checkerboard sweep

(b) Raster sweep

Figure 3.3: Mean squared error (MSE) for the fixed weight control variate method with checkerboard and raster sweeps, for the Ising model with $\eta = 0.3$, based on 100 simulated means at each batch size B, where each simulated mean used $M = 2000$ draws.

Chapter 4

Conclusions and Discussion

In Chapter 2, we modeled forest communities on a landscape via a latent Markov random field model. The spatially correlated model outperformed the spatially independent model for parameter estimation in a simulation study and for prediction on the historical Wisconsin PLS dataset. The fitted models were sensible relative to prior ecological literature, and we provided ecological interpretation of the fitted models on the PLS dataset. We also proposed a stochastic approximation procedure for jointly estimating the forest community species compositions and the spatial correlation strength in our latent Markov random field model.

In Forbes et al. [2013], which studied similar a similar model to the one in Chapter 2, the spatial correlation structure includes additional parameters to allow the interaction strength to depend on the forest types. We achieved adequate results with a single spatial correlation parameter, and leave the investigation of more sophisticated spatial correlation structures to future work. It would also be interesting to relate the forest community classifications to environmental covariates across the PLS survey area. Finally, while we provide a computationally feasible method in this work, parameter estimation for noisily observed Markov random fields is still computationally

challenging. We believe this is a promising direction for future research.

In Chapter 3, we studied control variate methods for deterministic sweep Markov chains. We proposed control variate estimators with theoretical and practical benefits relative to existing approaches in the literature. Our proposed methodology should be particularly useful in applications involving deterministic sweep Gibbs samplers, where our fixed weight control variate estimator is simple to implement, comes with rigorous theoretical guarantees, and performs well in practice.

In the future, it will be useful to investigate good choices for the control variate basis functions in practical settings. In addition, while we consider only a fixed number of control variate basis functions here, it would be of interest to study high-dimensional asymptotic settings wherein a control variate basis of suitable functions f increases in dimension to infinity along with the Markov chain sample size, similar to, e.g Mijatović et al. [2018]. A potential "Holy Grail" type goal of this approach would be to achieve, in more practical settings than currently exist, zero variance, or nearly zero variance, estimates of integrals with respect to probability distributions. Here, zero variance estimates should be contrasted with the usual "slow", $\sqrt{M}$-normalized rate of convergence that occurs in typical Monte Carlo simulations.

We add one note of caution regarding control variates. In the Markov chain setting, control variates reduce variance essentially by post-processing the Markov chain, leaving the transition kernel of the underlying chain unchanged. Thus, we expect control variate methods to be applicable primarily to Markov chains with reasonable convergence to the stationary distribution, so that asymptotic results relating to the MCMC central limit theorems are believable. Put bluntly, we do not expect control variates to be able to cure convergence difficulties resulting from an inefficient tran-

sition kernel, in which case it may take a great deal of time for asymptotic results of the kind proven in Chapter 3 to take effect. Alternative approaches which directly modify the transition kernels to improve the convergence behavior of the chain are thus a useful avenue for further research.

Appendix A

Appendix to Chapter 2

A.1 Additional Computational Details

A.1.1 Gibbs Sampling Transition Kernel $P_\theta(\mathbf{z}, \mathbf{z}')$

In this section, we define the Gibbs sampling kernel $P_\theta(\mathbf{z}, \mathbf{z}')$ for updating the label configurations $\mathbf{z}$. Recall that for $\mathbf{z} \in \Omega^2 = \Omega \times \Omega$, we have $\mathbf{z} = (\mathbf{z}_1, \mathbf{z}_2)$, where we hope to draw the label configuration $\mathbf{z}_1$ according to the conditional distribution $p(z|\mathbf{y}, \theta)$, and $\mathbf{z}_2$ according to the marginal distribution $p(z|\eta)$, so that $\mathbf{z}$ corresponds to a draw from $\pi_\theta(\mathbf{z}) = p(\mathbf{z}_1|\mathbf{y}, \theta)p(\mathbf{z}_2|\theta)$. Both $p(z|\eta)$ and $p(z|\mathbf{y}, \theta)$ are Markov random field densities with the same neighborhood structure. This can be shown by deriving (via a brief computation) the conditional probabilities $p\{\mathbf{z}_1(i)|\mathbf{y}, \mathbf{z}_1(-i), \theta\}$ and $p\{\mathbf{z}_2(i)|\mathbf{z}_2(-i), \theta\}$. For $z \in \Omega$, let $z(-i)$ refer to the values at all cells except the

ith cell. Using the notation $p_{ik} \stackrel{\Delta}{=} p(\mathbf{y}_i|z_i = k, \theta)$, we have

$$p\{\mathbf{z}_2(i) = k \big| \mathbf{z}_2(-i), \theta\}$$
$$= \begin{cases} \frac{\exp\left[\eta_k + \eta_K \sum_{i' \sim i} I\{\mathbf{z}_2(i')=k\}\right]}{\exp\left[\eta_K \sum_{i' \sim i} I\{\mathbf{z}_2(i')=K\}\right] + \sum_{k'=1}^{K-1} \exp\left[\eta_{k'} + \eta_K \sum_{i' \sim i} I\{\mathbf{z}_2(i')=k'\}\right]}; & (k < K) \\ \\ \frac{\exp\left[\eta_K \sum_{i' \sim i} I\{\mathbf{z}_2(i')=K\}\right]}{\exp\left[\eta_K \sum_{i' \sim i} I\{\mathbf{z}_2(i')=K\}\right] + \sum_{k'=1}^{K-1} \exp\left[\eta_{k'} + \eta_K \sum_{i' \sim i} I\{\mathbf{z}_2(i')=k'\}\right]}; & (k = K) \end{cases} \tag{A.1}$$

and

$$p\{\mathbf{z}_1(i) = k \big| \mathbf{y}, \mathbf{z}_1(-i), \theta\}$$
$$= \begin{cases} \frac{p_{ik} \exp\left[\eta_k + \eta_K \sum_{i' \sim i} I\{\mathbf{z}_1(i')=k\}\right]}{p_{iK} \exp\left[\eta_K \sum_{i' \sim i} I\{\mathbf{z}_1(i')=K\}\right] + \sum_{k'=1}^{K-1} p_{ik'} \exp\left[\eta_{k'} + \eta_K \sum_{i' \sim i} I\{\mathbf{z}_1(i')=k'\}\right]} \\ \quad (k < K); \\ \\ \frac{p_{iK} \exp\left[\eta_K \sum_{i' \sim i} I\{\mathbf{z}_1(i')=K\}\right]}{p_{iK} \exp\left[\eta_K \sum_{i' \sim i} I\{\mathbf{z}_1(i')=K\}\right] + \sum_{k'=1}^{K-1} p_{ik'} \exp\left[\eta_{k'} + \eta_K \sum_{i' \sim i} I\{\mathbf{z}_1(i')=k'\}\right]}; \\ \quad (k = K) \end{cases} \tag{A.2}$$

For $\mathbf{z} \in \Omega^2$, let $\mathbf{z}_j(i)$ refer to the value at cell i in the jth label configuration of $\mathbf{z}$ (recall $j \in \{1, 2\}$). We will use $\mathbf{z}_j\mathbf{z}'_j(i) \in \Omega$ to denote the configuration for which

$$(\mathbf{z}_j\mathbf{z}'_j)(i) = \begin{cases} \mathbf{z}_j(i) \text{ for } i \geq m \\ \\ \mathbf{z}'_j(i) \text{ for } i < m \end{cases}$$

The single sweep Gibbs sampler $P_\theta(\mathbf{z}, \mathbf{z}') \stackrel{\Delta}{=} P_{\theta,1}(\mathbf{z}_1, \mathbf{z}'_1)P_{\theta,2}(\mathbf{z}_2, \mathbf{z}'_2)$ is defined in terms of Gibbs sampling transition kernels $P_{\theta,1}(\cdot, \cdot) : \Omega \times \Omega \to [0, 1]$ and $P_{\theta,2}(\cdot, \cdot) : \Omega \times \Omega \to [0, 1]$. The transition kernels $P_{\theta,1}(\mathbf{z}_1, \mathbf{z}_{1'})$ and $P_{\theta,2}(\mathbf{z}_2, \mathbf{z}'_2)$ are constructed, respectively, to have $p(\mathbf{z}_1|\mathbf{y}, \theta)$ and $p(\mathbf{z}_2|\theta)$ as their stationary distributions. They are

defined in terms of *single-cell* transition kernels $P^i_{\theta,j}(\cdot,\cdot)$:

$$P_{1,\theta}(\mathbf{z}_1, \mathbf{z}'_1) \triangleq \prod_{i=1}^{n} P^i_{\theta,1}\{\mathbf{z}_1\mathbf{z}'_1(i), \mathbf{z}_1\mathbf{z}'_1(i+1)\}$$
$$P_{2,\theta}(\mathbf{z}_2, \mathbf{z}'_2) \triangleq \prod_{i=1}^{n} P^i_{\theta,2}\{\mathbf{z}_2\mathbf{z}'_2(i), \mathbf{z}_2\mathbf{z}'_2(i+1)\}$$

In turn, for $x, x' \in \Omega$, the single-cell kernels are defined by

$$P^i_{\theta,1}(x, x') \triangleq I\{x'(-i) = x(-i)\}\, p\{x'(i)|\mathbf{y}, x(-i), \theta\}$$
$$p^i_{\theta,2}(x, x') \triangleq I\{x'(-i) = x(-i)\}\, p\{x'(i)|x(-i), \theta\} \tag{A.3}$$

The conditional probabilities required by Equation A.3 are given in Equations A.1 and A.2.

A.1.2 Implementation Details for Algorithm 1

We choose the stepsize c in Algorithm 1 by finding c so that the behavior of the algorithm is reasonable: for too large a stepsize, the algorithm may oscillate between nonsensically large parameter values, while for too small a stepsize, the convergence of the algorithm is extremely slow. For the PLS data examples, we use $c = 0.02/n$, where n is the number of cells in the grid. Another useful trick is the following: rather than using the initial time $t = 1$ and stepsize sequence $\epsilon_t = 1/t$, we use the "time-shifted" sequence with initial time $t = D$ and $\epsilon_t = D/(t + D)$, for some moderately large D. This causes the stepsize to decrease much more slowly as the iterations proceed; effectively, the time shifted sequence uses a larger stepsize by a factor of D, but also starts at a higher iteration number. Some caution is in order as Younes [1988] shows that the convergence of Markov chain stochastic optimization may only be guaranteed for small enough stepsizes, but we did not experience convergence issues in our work.

For the PLS model fits, we use 8000 steps and $D = 200$, so that the shifted time at the final step is $T_{Final} = 8000 + D = 8200$.

Additionally, the Gibbs sampling step in Algorithm 1 can be made much faster for square lattice grids by exploiting the small number of spatial neighbors (≤ 4) of each cell. From Equations A.1 and A.2, the single-site conditional probabilities needed for Gibbs sampling depend only on the neighboring values. The neighboring grid cells of a given cell can be obtained quickly, and their values used to compute the necessary conditional probabilities, by using a sparse symmetric adjacency matrix representation of the lattice structure. In a sparse adjacency matrix representation, the storage format of the matrix (column-major or row-major) will determine the most efficient scheme for accessing the neighbors. In a column (row) major sparse matrix $\mathbf{A}$, it will be fastest to find the ith neighbor of site j by finding the row index of the ith nonzero entry in the jth column (row) [see, e.g., software documentation such as in Guennebaud et al., 2010, for more details].

A.1.3 Path Integration to Evaluate Loglikelihoods

For the spatially correlated models we estimate here, the holdout likelihood in Equation 2.22 of the main text is difficult to compute. It is technically possible to compute $\ell_{holdout}(\hat{\theta})$ via a Gibbs sampling average with respect to $p(z|\hat{\theta})$. However, the marginal distribution $p(z|\hat{\theta})$ will tend to put most of its mass on configurations with very low $p(\mathbf{y}_{test}|z, \hat{\theta})$, and computing Equation 2.22 by averaging over configurations obtained from Gibbs sampling of $p(z|\hat{\theta})$ is not efficient. The path integral approach we now describe is a method to estimate $\ell_{holdout}$ that overcomes the high variance of approaches based on importance sampling [Gelman and Meng, 1998, Neal, 1993].

In path integration, we aim to compute differences in marginal likelihood between two parameter settings, θ_1 and θ_2, via integrations of the form

$$\Delta(\theta_1, \theta_2) = \ell_{holdout}(\theta_2) - \ell_{holdout}(\theta_1) = \int_0^1 \frac{\partial \ell_{holdout}}{\partial \theta}^T \frac{\partial \theta(t)}{\partial t} dt \tag{A.4}$$

where $\theta(t) = (1-t)\theta_1 + t\theta_2$.

We estimate the integral in Equation A.4 using stochastic estimates of $\frac{\partial \ell_{holdout}}{\partial \theta}$. The derivative of the holdout loglikelihood with respect to η_K is $\frac{\partial \ell_{holdout}(\theta)}{\partial \eta_K} = E(T_K|\mathbf{y}_{test}, \theta) - E(T_K|\theta)$. We can approximate this derivative by the estimate $T_K(\mathbf{z}_1) - T_K(\mathbf{z}_2)$, where $\mathbf{z}$ is an approximate draw from $\pi_\theta(\mathbf{z})$ defined in Equation 2.20 in the main text. In practice, the starting point θ_1 is frequently chosen so that $\ell_{holdout}(\theta_1)$ can be easily computed. In our case, $\ell_{holdout}(\theta_1)$ can easily be computed exactly when the spatial correlation parameter η_K for θ_1 is 0.

We define $\hat{\theta}_{ind}$ to be the parameter $\hat{\theta}$ but with correlation parameter η_K set to 0 (so that the labels are marginally independent under $\hat{\theta}_{ind}$). Then we compute

$$\ell_{pen}(\hat{\theta}) = \ell_{pen}(\hat{\theta}_{ind}) + \Delta(\hat{\theta}_{ind}, \hat{\theta}) \tag{A.5}$$

The difference $\Delta(\hat{\theta}_{ind}, \hat{\theta})$ is computed using the procedure in Algorithm 2.

Algorithm 2: Path integration procedure

Initialize parameter $\theta = \hat{\theta}_{ind}$, configuration $\mathbf{z} \in \Omega^2$, number of iterations T

Set $\Delta(\hat{\theta}_{ind}, \hat{\theta}) = 0$

Set $h = 1/T$

Set $\Delta_h = h\widehat{\eta}_K$;

for *t=1* **to** T **do**

  Draw $\mathbf{z}' \in \Omega^2$ according to $P_\theta(\mathbf{z}, \cdot)$

  $\mathbf{z} = \mathbf{z}'$

  $\Delta(\hat{\theta}_{ind}, \hat{\theta}) = \Delta(\hat{\theta}_{ind}, \hat{\theta}) + \Delta_h \{T_K(\mathbf{z}_1) - T_K(\mathbf{z}_2)\}$

  $\eta(\theta)_K = \eta(\theta)_K + \Delta_h$

Return $\Delta(\hat{\theta}_{ind}, \hat{\theta})$

By the definition of $\hat{\theta}_{ind}$, only η_K differs along the $\hat{\theta}_{ind}$ to $\hat{\theta}$ path. Additionally, we note that path integration procedures are sometimes implemented by discretizing a $\theta_1 - \theta_2$ path, and then approximating the derivative $\frac{\partial \ell_{holdout}}{\partial \theta}$ at each point based on many Monte Carlo or Markov Chain Monte Carlo runs. In contrast, here we use a single draw at each iteration with new parameter values separated by small increments. The goal of this modification is to avoid the need for a burnin period at each new parameter value by slowly transitioning (over T iterations) from $\hat{\theta}_{ind}$ to $\hat{\theta}$.

A.2 EM Updates for the Spatially Independent Model, and Stochastic Gradient Updates

A.2.1 Independent EM

When the labels are spatially independent, it will be convenient to use the vector $w \in \mathbb{R}^K$ to refer to the marginal probabilities of each label type, so that $p(Z_i = k|w) =$

w_k. We will still use $\boldsymbol{\mu} \in \mathscr{M}^K$ to refer to the conditional distribution parameters and θ to refer to the joint $(w, \boldsymbol{\mu})$ parameter. Under spatial independence, the observed data density is

$$p(\mathbf{y}|\theta) = \prod_{i=1}^{n} \left(\sum_{k=1}^{K} w_k \prod_{m=1}^{M} \boldsymbol{\mu}_{mk}^{\mathbf{y}_{mi}} \right)$$

and the observed data loglikelihood is

$$\ell(\theta) = \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} w_k \prod_{m=1}^{M} \boldsymbol{\mu}_{mk}^{\mathbf{y}_{mi}} \right).$$

For our implementation of the EM algorithm for the spatially independent mixture model, we regularize the w vector via the Dirichlet prior penalty $\rho_1(w) = (\alpha - 1) \sum_{k=1}^{K} \log(w_k)$. We regularize the $\boldsymbol{\mu}$ matrix via the Dirichlet prior penalty $\rho_2(\boldsymbol{\mu}) = (\alpha - 1) \sum_{k=1}^{K} \sum_{m=1}^{M} \log(\boldsymbol{\mu}_{mk})$, which is the same $\rho_2(\boldsymbol{\mu})$ as was used in Section 2.3 for the spatially dependent EM algorithm.

We now write $\ell_{pen}(\theta) = \ell(\theta) + \rho_1(w) + \rho_2(\boldsymbol{\mu})$ for the regularized spatially independent loglikelihood.

The conditional distribution of the latent label at cell i given parameter θ and tree count vector $\mathbf{y}_i$ is

$$p(z_i = k|\mathbf{y}_i, \theta) = \frac{w_k p(\mathbf{y}_i|z_i = k, \boldsymbol{\mu}_k)}{\sum_{k'=1}^{K} w_{k'} p(\mathbf{y}_i|z_i = k', \boldsymbol{\mu}_{k'})}$$

As in the main article, we may construct the minorizing function $Q(\theta|\theta_{old}) = Q_1(w|\theta_{old}) + Q_2(\boldsymbol{\mu}|\theta_{old})$ where

$$Q_1(w|\theta_{old}) = (\alpha - 1) \sum_{k=1}^{K} \log(w_k) + \sum_{i=1}^{n} \sum_{k=1}^{K} p(z_i = k|\mathbf{y}_i, \theta^{old}) \log(w_k)$$

and

$$Q_2(\boldsymbol{\mu}|\theta_{old})$$
$$= (\alpha - 1)\sum_{m=1}^{M}\sum_{k=1}^{K}\log(\boldsymbol{\mu}_{mk}) + \sum_{n=1}^{n}\sum_{k=1}^{K}\sum_{m=1}^{M}\mathbf{y}_{mi}p(z_i = k|\mathbf{y}_i, \theta^{old})\log(\boldsymbol{\mu}_{mk})$$

It can be shown using the argument in Section A.2.2 that the entries of the maximizers w^{new} and $\boldsymbol{\mu}^{new}$ are given by

$$w_k^{new} = \frac{\alpha - 1 + \sum_{i=1}^{n} p(z_i = k|\mathbf{y}_i, \theta^{old})}{K(\alpha - 1) + \sum_{k'=1}^{K}\sum_{i=1}^{n} p(z_i = k'|\mathbf{y}_i, \theta^{old})}$$
$$\boldsymbol{\mu}_{mk}^{new} = \frac{\alpha - 1 + \sum_{i=1}^{n} p(z_i = k|\mathbf{y}_i, \theta^{old})\mathbf{y}_{mi}}{M(\alpha - 1) + \sum_{m'=1}^{M}\sum_{i=1}^{n} p(z_i = k|\mathbf{y}_i, \theta^{old})\mathbf{y}_{m'i}}$$

A.2.2 Jensen's Inequality Argument

A standard textbook result based on Jensen's inequality (see, e.g., Shao [2003] example 1.49) states that the following inequality holds for any two length M probability vectors x, y with positive entries and $\sum_{m=1}^{M} x_m = \sum_{m=1}^{M} y_m = 1$:

$$\sum_{m=1}^{M} x_m \log(x_m/y_m) \geq 0 \tag{A.6}$$

The inequality is strict except when $x = y$, where equality holds. Now, consider maximizing

$$\sum_{m=1}^{M} h_m \log(\mu_m) \tag{A.7}$$

over $\mu \in \{\mu : \sum_{m=1}^{M} \mu_m = 1, \mu_m > 0, \forall m\}$, where each h_m is required to be positive but $\sum_{m=1}^{M} h_m$ is not required to be 1. Let $H = \sum_{m=1}^{M} h_m$. Maximizing Equation A.7 over μ is equivalent to maximizing

$$\sum_{m=1}^{M} (h_m/H)\log(\mu_m) \tag{A.8}$$

over μ. By the inequality in Equation A.6, we have

$$\sum_{m=1}^{M}(h_m/H)\log\left\{\frac{(h_m/H)}{\mu_m}\right\} > 0 \tag{A.9}$$

except when $\mu_m = h_m/H$ for each m. Thus, to maximize Equation A.7, we must take $\mu_m = h_m/H$ for each m.

A.2.3 Stochastic Gradient and Rescaled Stochastic Gradient Updates

The forest community conditional densities take the form

$$\begin{aligned} p(\mathbf{y}_i|z_i = k, \boldsymbol{\mu}_k) &= \prod_{m=1}^{M} \boldsymbol{\mu}_{mk}^{\mathbf{y}_{mk}} \\ &= \exp\{\boldsymbol{\phi}_k^T \mathbf{y}_i - q_i \xi(\phi_k)\} \end{aligned}$$

where $\boldsymbol{\phi}_k \in \mathbb{R}^M$ is a vector of exponential family natural parameters corresponding to the mean parameter $\boldsymbol{\mu}_k$. Since the entries of $\boldsymbol{\mu}_k$ must sum to 1, this exponential family parameterization is not full-rank.

For the stochastic gradient update, we will take the parameter $\theta^T = \left[\eta^T, \vec{(}\boldsymbol{\phi})^T\right]$ where $\boldsymbol{\phi} \in \mathbb{R}^{M\times K}$ has columns $\boldsymbol{\phi}_k$. It is convenient to perform the gradient updates on the $\boldsymbol{\phi}$ space rather than the $\boldsymbol{\mu}$ space since the $\boldsymbol{\phi}$ space is unbounded. From standard results on exponential families, we have

$$\frac{\partial \log p(\mathbf{y}_i|z_i = k, \boldsymbol{\phi}_k)}{\partial \boldsymbol{\phi}_k} = \mathbf{y}_i - q_i \boldsymbol{\mu}_k,$$

where the relationship between the mean parameters $\boldsymbol{\mu}_k$ and the natural parameters $\boldsymbol{\phi}_k$ for exponential families was used [see, e.g., Shao, 2003].

Now, noting $p(\mathbf{y}|\theta) = \sum_{z\in\Omega} p(\mathbf{y}|z,\theta)p(z|\eta)$, we have

$$\frac{\partial \log\{p(\mathbf{y}|\theta)\}}{\partial \boldsymbol{\phi}_k} = \sum_{z\in\Omega} p(z|\mathbf{y},\theta) f_{\boldsymbol{\phi}_k}(z)$$

where $f_{\boldsymbol{\phi}_k}(z) = \sum_{i=1}^{n} I(z_i = k)(\mathbf{y}_i - q_i \boldsymbol{\mu}_k)$.

The stochastic gradient procedure follows Algorithm 1, using the update formula

$$G_{SG}(\theta, \mathbf{z}) = G_{SG}\left(\begin{bmatrix} \eta \\ \vec{(}\boldsymbol{\phi}) \end{bmatrix}, \mathbf{z}\right) = c\begin{bmatrix} \left\{\frac{\partial \rho_1(\eta)}{\partial \eta} + T(\mathbf{z}_1) - T(\mathbf{z}_2)\right\} \\ H_\alpha(\boldsymbol{\phi}, \mathbf{z}_1) \end{bmatrix}$$

where the $\boldsymbol{\phi}$ update is

$$H_\alpha(\boldsymbol{\phi}, \mathbf{z}_1)_{mk} = (\alpha - 1)(1 - M\boldsymbol{\mu}_{mk}) + \sum_{i=1}^{n} I\{\mathbf{z}_1(i) = k\}(\mathbf{y}_{mi} - q_i \boldsymbol{\mu}_{mk})$$

The gradient descent update formula is quite similar to the EM-like update formula in Equation 2.19 of the main text. However, the gradient descent update occurs on the $\boldsymbol{\phi}$ (natural parameter) scale, rather than the $\boldsymbol{\mu}$ (mean parameter) scale, and the denominator term for the stochastic gradient update is not scaled by the number of trees in the dataset. When $\sum_{i=1}^{n} q_i$ is very large, this may lead to poor behavior since the order of magnitude of the gradient components with respect to $\boldsymbol{\phi}_k$ will increase with $\sum_{i=1}^{n} q_i$, while the order of magnitude of $T(\mathbf{z}_1)$ and $T(\mathbf{z}_2)$ depends only on the number of grid cells (and not on the number of trees observed $\sum_{i=1}^{n} q_i$). In such cases it may be valuable to rescale the components of the likelihood gradient corresponding to the η and $\boldsymbol{\phi}$ parameters. In the "rescaled" stochastic gradient algorithm in the simulation study, we use

$$G_{RSG}(\theta, \mathbf{z}) = G_{RSG}\left(\begin{bmatrix} \eta \\ \vec{(}\boldsymbol{\phi}) \end{bmatrix}, \mathbf{z}\right) = c\begin{bmatrix} (1/n)\left\{\frac{\partial \rho_1(\eta)}{\partial \eta} + T(\mathbf{z}_1) - T(\mathbf{z}_2)\right\} \\ H_{\alpha,RSG}(\boldsymbol{\phi}, \mathbf{z}_1) \end{bmatrix}$$

where the $\boldsymbol{\phi}$ update is

$$H_{\alpha,RSG}(\boldsymbol{\phi}, \mathbf{z}_1)_{mk} = \frac{(\alpha - 1)(1 - M\boldsymbol{\mu}_{mk}) + \sum_{i=1}^{n} I\{\mathbf{z}_1(i) = k\}(\mathbf{y}_{mi} - q_i \boldsymbol{\mu}_{mk})}{M(\alpha - 1) + \sum_{i=1}^{n} q_i}.$$

and n is the number of grid cells. The MSE for the rescaled update are smaller than for the ordinary SG update (Tables 2.6 and 2.7). The improvement seems to result

from putting the updates for the Markov random field parameter η and the conditional distribution parameter $\boldsymbol{\phi}$ on a more similar scale. In this way, the incremental EM update in Equation 2.15 in the main article can be viewed as a sort of automatic preconditioning, although in the incremental EM algorithm it is still necessary to choose a good stepsize c for the η update.

Appendix B

Appendix to Chapter 3

The proofs of Lemmas B.1–B.9 are in Appendix A. Appendix B.2 contains (in order), the proofs of Theorem 3.1, Corollaries 3.1–3.2, Theorems 3.2–3.3, and Propositions 3.1–3.2.

B.1 Proofs of Lemmas

In Lemmas B.1–B.3, we show that the length K composition kernels P_k^K are aperiodic when the Π_k are Gibbs kernels for $k = 1, ..., K$.

Lemma B.1. *Assume (A.1b). Suppose $P_k^t I_A(x) = I_A(x)$ a.e. π for some $t > 1$. Then $P_{\sigma(k)}^{t-1} I_A = I_A$ a.e. π.*

Proof: We have

$$\begin{aligned}
&\langle I_A, I_A \rangle = \langle P_k^t I_A, P_k^t I_A \rangle = \langle P_{\sigma(k)}^{t-1} I_A, P_k^t I_A \rangle \\
&\leq \langle P_{\sigma(k)}^{t-1} I_A, P_{\sigma(k)}^{t-1} I_A \rangle^{1/2} \langle P_k^t I_A, P_k^t I_A \rangle^{1/2} \\
&= \langle P_{\sigma(k)}^{t-1} I_A, P_{\sigma(k)}^{t-1} I_A \rangle^{1/2} \langle I_A, I_A \rangle^{1/2}
\end{aligned}$$

where the first and final equalities follow since $P_k^t I_A = I_A$ a.e. π, the second equality follows from reversibility and idempotence of Π_k, and the inequality follows from the Cauchy-Schwarz inequality. Jensen's inequality gives $\langle P_{\sigma(k)}^{t-1} I_A, P_{\sigma(k)}^{t-1} I_A \rangle \leq \langle I_A, I_A \rangle$,

which from the preceding implies $\langle P^{t-1}_{\sigma(k)} I_A, P^{t-1}_{\sigma(k)} I_A \rangle = \langle I_A, I_A \rangle$. Since $\langle P^{t-1}_{\sigma(k)} I_A, P^{t-1}_{\sigma(k)} I_A \rangle = \langle I_A, I_A \rangle$, applying the Cauchy-Schwarz inequality to $P^{t-1}_{\sigma(k)} I_A$ and $P^t_k I_A$ implies $P^{t-1}_{\sigma(k)} I_A(x) = P^t_k I_A(x) = I_A(x)$ a.e. π. □

Lemma B.2 below relates the stationary measure π to the irreducibility measure ψ.

Lemma B.2. *Assume (A.1) and (A.2), and suppose $\psi(A) > 0$ for some $A \in \mathscr{X}$, and $P^K_k(x, A) = 1$ for all $x \in A$. Then $\pi(A) = 1$.*

Proof of Lemma B.2: Note that for any $A \in \mathscr{X}$,

$$\pi(A) = \sum_{t=1}^{\infty} 2^{-t} \pi(A) = \sum_{t=1}^{\infty} 2^{-t} \int \pi(dx) P^{Kt}_k I_A(x)$$

Now, suppose some set $A \in \mathscr{X}$ satisfies $\psi(A) > 0$ and $P^K_k(x, A) = 1$ for all $x \in A$. Then

$$\begin{aligned}
\pi(A) &= \sum_{t=1}^{\infty} 2^{-t} \int \pi(dx) P^{Kt}_k I_A(x) \\
&= \sum_{t=1}^{\infty} 2^{-t} \int \pi(dx) I_A P^{Kt}_k I_A(x) \\
&+ \sum_{t=1}^{\infty} 2^{-t} \int \pi(dx) I_{A^C} P^{Kt}_k I_A(x) \\
&= \pi(A) + \int \pi(dx) I_{A^C} \sum_{t=1}^{\infty} 2^{-t} P^{Kt}_k I_A(x)
\end{aligned}$$

Thus, $I_{A^C} \sum_{t=1}^{\infty} 2^{-t} P^{Kt}_k I_A(x) = 0$ a.e. π. But from ψ-irreducibility of P^K_k, the infinite sum is positive for all x. This implies $I_{A^C} = 0$ a.e. π, and thus $\pi(A) = 1$. □

Now, we finish the proof of aperiodicity.

Lemma B.3. *Under (A.1b) and (A.2), the transition kernels P^K_k are aperiodic.*

Proof: Consider some arbitrary k from $1, ..., K$. From Theorem 5.4.4 of Meyn and Tweedie [2009] and the ψ-irreducibility of P^K_k, there exists an integer d and a collection of sets $D_1, ..., D_d$ satisfying

1. $P_k^K(x, D_{i+1}) = 1$ for $x \in D_i$, $i \equiv 0, ..., d-1 \pmod{d}$

2. $\psi\{(\cup_{i=1}^d D_i)^C\} = 0$

3. $D_1, ..., D_d$ are disjoint

We show that P_k^K is aperiodic by showing that $d = 1$ is the largest integer such that 1-3 hold for a collection of sets $D_1, ..., D_d$. Suppose to the contrary that $d > 1$ for a collection of sets $D_1, ..., D_d$ satisfying 1-3. From Lemma B.2, we have $\pi(\cup_{i=1}^d D_i) = 1$. Thus, $P_k^{Kd} I_{D_i} = I_{D_i}$ a.e. π for each i. Now, $K(d-1)$ applications of Lemma B.1 imply $P_k^K I_{D_i} = I_{D_i}$ a.e. π for each i. Additionally, $\pi(D_i) > 0$ for at least one i, so for this i, $H = \{x \in D_i : P_k^K I_{D_i} = 1\}$ is non-empty. But this is a contradiction, since $P_k^K(x, D_{i+1}) = 1$ for all $x \in H$. Thus, any collection of sets $D_1, ..., D_d$ satisfying 1-3 must have $d = 1$. This proves the result. □

Lemma B.4. *Assume (A.1)–(A.4) hold. Suppose* $\Pi_k f = f$ *a.e.* π *for each* $k = 1, ..., K$, *where* $f : (X, \mathscr{X}) \to (\mathbb{R}, \mathscr{R})$. *Then* f *is constant a.e.* π.

Proof: Suppose $\Pi_k I_A = I_A$ a.e. π for each k, for some $A \in \mathscr{X}$. We first show $\pi(A) = 0$ or 1. Define a Markov chain $\{Y_t\}_{t=0}^\infty$ by the initial distribution $Y_0 \sim \pi$ and the transition kernel P_1^K. Then we have $I_A(Y_t) = I_A(Y_0)$ almost surely for all t. Now under (A.1)–(A.4), a Law of Large Numbers holds, so that $M^{-1} \sum_{t=0}^{M-1} I_A(Y_t) \overset{a.s.}{\to} \int \pi(dx) I_A(x)$. Thus, $I_A(Y_0) = \int \pi(dx) I_A(x)$ almost surely. This implies $\int \pi(dx) I_A(x) = 0$ or 1.

Now, let $\mathscr{H} = \{A \in \mathscr{F} : \Pi_k I_A = I_A \ a.e.\ \pi, \ \forall k\}$. We show $\mathscr{H}$ is a σ-field. The empty set $\phi \in \mathscr{H}$ and the state space $X \in \mathscr{H}$. Also, for any arbitrary $A \in \mathscr{H}$, we have $\pi(A) = 0$ or $\pi(A) = 1$. This implies $A^C \in \mathscr{H}$ for each $A \in \mathscr{H}$. Finally, for

$\{A_n\}_{n=1}^{\infty}$ with each $A_n \in \mathscr{H}$, we have $\pi(\cup_{n=1}^{\infty} A_n) = 0$ or 1, so that $\cup_{n=1}^{\infty} A_n \in \mathscr{H}$. Thus, $\mathscr{H}$ is a σ-field.

We have $\Pi_k f = f$ a.e. π by assumption. We now show f is $\mathscr{H}$-measurable. Let $B \in \mathscr{R}$ given. Define $A = f^{-1}(B)$. Also, define $A_k = (\Pi_k f)^{-1}(B)$ for each k. We have $I_A = I_{A_k}$ a.e. π for each k since $f = \Pi_k f$ a.e. π for each k. Thus $A = f^{-1}(B) \in \mathscr{H}$. Since B was arbitrary, f is $\mathscr{H}$ measurable.

Finally, we show that f is constant a.e. π. Without loss of generality, we assume $f \geq 0$. For general f, we may use the standard decomposition of f into positive and negative components and apply the following reasoning to each component [see, e.g., Chapter 1 of Shao, 2003]. Since f is $\mathscr{H}$-measurable, we can construct a sequence $\{f_n\}_{n=1}^{\infty}$ of $\mathscr{H}$-measurable simple functions such that $f_n \uparrow f$ pointwise. Now, let $b = \int \pi(dx) f(x)$. Then $\pi(\{x : f(x) > b\}) = \lim_{n\to\infty} \pi(\{x : f_n > b\}) = 0$, where we used $\{f_n > b\} \subset \{f_{n+1} > b\}$ for each n by monotone convergence of the f_n, as well as the fact that f_n are constant a.e. π so that $\int \pi(dx) f_n \leq b$ implies $f_n \leq b$ a.e. π. Thus, $f(x) = b$ a.e. π. This completes the proof. □

Lemma B.5. *Assume (A.1)–(A.4) hold, and suppose (B.4) holds for the function $f : X \to \mathbb{R}^p$. Then for any $a \in \mathbb{R}^p$, we have*

$$a^T \left[\sum_k \int \pi(dx) \{ f f^T - (\Pi_k f)(\Pi_k f)^T \} \right] a = 0$$

if and only if $a^T f = b$ a.e. π for some constant b.

Proof: Follows easily from Lemma B.4. □

Lemma B.6. *Let U be a symmetric positive semidefinite $p \times p$ matrix, and V be a $p \times d$ matrix such that $a^T V = 0_{1\times d}$ for any vector a with $Ua = 0$. Then for all p-vectors a,*

$$U^{\dagger} V \in \underset{C \in \mathbb{R}^{p\times d}}{\arg\min}\, a^T \{ C^T U C - C^T V - V^T C \} a$$

Proof: First, we note that when $U = 0_{p\times p}$, then $V = 0$ also, so that $a^T\{C^TUC - C^TV - V^TC\}a = 0$ for all a, for any choice of C. In particular, $a^T\{C^TUC - C^TV - V^TC\}a = 0$ for all a for $C = U^\dagger V$. Otherwise, since U is symmetric positive semidefinite, we may write $U = QDQ^T$ where Q is a $p \times r$ matrix with $r \leq p$ orthonormal columns, and D is a $r \times r$ diagonal matrix with strictly positive diagonal entries. Further, for any $C \in \mathbb{R}^{p\times d}$, we may write $C = QR + B$ where $R \in \mathbb{R}^{r\times d}$, $B \in \mathbb{R}^{p\times d}$, and $Q^TB = 0_{r\times d}$.

It can be checked that the value of the B component of C does not affect the value of $a^T(C^TUC - C^TV - V^TC)a$, so that minimizers of the form $C = QR$ exist. When $C = QR$, we have

$$C^TUC - C^TV - V^TC = X^TX - V^TQD^{-1}Q^TV \tag{B.1}$$

where $X = D^{1/2}Q^TC - D^{-1/2}Q^TV$. The second term in (B.1) does not depend on C. Now, $a^TX^TXa \geq 0$ for arbitrary C. But $U^\dagger = QD^{-1}Q^T$, so that taking $C = U^\dagger V$ gives $X = D^{1/2}Q^TQD^{-1}Q^TV - D^{-1/2}Q^TV = 0$. Thus, $a^T\{C^TUC - C^TV - V^TC\}a$ is minimized for each a whenever $C = U^\dagger V$. This completes the proof. □

In Lemma B.7–B.8, we take $K = 2$ and $Q = (\Pi_1 + \Pi_2)/2$.

Lemma B.7. *Assume (A.1b)–(A.3) and (B.1)–(B.2). Then we have* $\sum_{t=1}^\infty |Q^tg(x)|$ *is square integrable with respect to* π*, and* $\sum_{t=1}^\infty Q^tg(x) = \sum_{t=1}^\infty (P_1^t + P_2^t)g(x)$ *a.e.* π*.*

Proof: We show the result for scalar $g : X \to \mathbb{R}$. The result follows for general $g : X \to \mathbb{R}^p$ by applying the reasoning below elementwise.

Note $Q^tg = \{(\Pi_1 + \Pi_2)/2\}^tg = \sum_{i=1}^t 2^{-t}\binom{t-1}{i-1}(P_1^i + P_2^i)g$ a.e. π via idempotence

of the Π_k, so that

$$\begin{aligned}\sum_{n=1}^{M}|Q^n g(x)| &\le \sum_{n=1}^{M}2^{-n}\sum_{i=1}^{n}\binom{n-1}{i-1}\{|P_1^i g(x)|+|P_2^i g(x)|\}\\ &= \sum_{i=1}^{M}\sum_{n=1}^{M}\{|P_1^i g(x)|+|P_2^i g(x)|\}2^{-n}\binom{n-1}{i-1}I(i\le n) \qquad \text{(B.2)}\\ &\le \sum_{n=1}^{M}(|P_1^n g(x)|+|P_2^n g(x)|)\end{aligned}$$

a.e. π, where the second inequality follows because

$$\sum_{n=i}^{\infty}2^{-n}\binom{n-1}{i-1}=\sum_{r=1}^{\infty}2^{-(i-1+r)}\binom{i-2+r}{i-1}=1,$$

which itself is a well-known identity related to the pdf of a negative binomial random variable.

The Assumptions (A.1b)–(A.3) imply (A.1)–(A.4) hold, and since (B.1)–(B.2) also hold, we have $\sum_{t=1}^{\infty}|P_1^t g(x)|+|P_2^t g(x)|<\infty$ a.e. π by Proposition 3.1. Thus $\sum_{n=1}^{\infty}|Q^n g(x)|$ converges a.e. π, and

$$\begin{aligned}\sum_{n=1}^{\infty}Q^n g(x) &= \lim_{M\to\infty}\sum_{i=1}^{M}\sum_{n=1}^{M}\{P_1^i g(x)+P_2^i g(x)\}2^{-i}\binom{n-1}{i-1}I(i\le n)\\ &= \sum_{n=1}^{\infty}P_1^n g(x)+P_2^n g(x),\end{aligned}$$

a.e. π, where the final equality follows from using Fubini's Theorem and the identity $\sum_{i=n}^{\infty}2^{-i}\binom{i-1}{n-1}=1$. □

Lemma B.8. *Assume (A.1b), (A.2)–(A.3), and (B.1)–(B.4) hold. Assume* $\{X_t\}_{t=0}^{\infty}$ *is defined as in Theorem 3.1. Take* S_M *as in* (3.6) *and define* $H_M=\sum_{t=0}^{M-1}g(X_t)-C^T\{f(X_t)-Qf(X_t)\}$. *Then* $M^{-1/2}(S_M-H_M)\xrightarrow{a.s.}0$.

Proof of Lemma B.8: We prove the result assuming $X_0\sim\pi$. When $X_0\sim\nu$ for general ν, the result can be shown with a coupling argument as sketched in Theorem 3.1.

We prove the result for scalar $f, g : X \to \mathbb{R}$ and $C = c \in \mathbb{R}$. The extension to $g : X \to \mathbb{R}^d$, $f : X \to \mathbb{R}^p$, and $C \in \mathbb{R}^{p \times d}$ can be shown by applying the univariate result elementwise for each of the d elements of $M^{-1/2}(S_M - H_M)$.

For univariate f, g and scalar c, we have

$$S_M - H_M = \sum_{t=0}^{\lfloor (M-1)/2 \rfloor} c\Pi_1 f(X_{2t}) + \sum_{t=0}^{\lfloor (M-2)/2 \rfloor} c\Pi_2 f(X_{2t+1}) - \sum_{t=0}^{M-1} c\{\Pi_1 f(X_t) + \Pi_2 f(X_t)\}/2.$$

Now, when t is even, $\Pi_1 f(X_t) = \Pi_1 f(X_{t+1})$ almost surely, and similarly, $\Pi_2 f(X_t) = \Pi_2 f(X_{t+1})$ almost surely when t is odd. Thus

$$\sum_{t=0}^{\lfloor (M-1)/2 \rfloor} c\Pi_1 f(X_{2t}) - \sum_{t=0}^{M-1} c\Pi_1 f(X_t)/2 = \begin{cases} 0 & M \text{ even} \\ c\Pi_1 f(X_{M-1})/2 & M \text{ odd} \end{cases}$$

almost surely, and

$$\sum_{t=0}^{\lfloor (M-2)/2 \rfloor} c\Pi_2 f(X_{2t+1}) - \sum_{t=0}^{M-1} c\Pi_2 f(X_t)/2 = \begin{cases} -c\Pi_2 f(X_0)/2 & M \text{ odd} \\ -c\Pi_2 f(X_0)/2 + c\Pi_2 f(X_{M-1})/2 & M \text{ even} \end{cases}$$

almost surely. Thus,

$$M^{-1/2}|S_M - H_M| \leq M^{-1/2} c\{|\Pi_2 f(X_0)/2| + |\Pi_1 f(X_{M-1})/2| + |\Pi_2 f(X_{M-1})/2|\} \overset{a.s.}{\to} 0$$

as $M \to \infty$ by applying the Strong Law of Large Numbers along the $K = 2$ subchains to the function $\{(\Pi_1 + \Pi_2) f(x)/2\}^2$. □

Lemma B.9. *Consider the Gibbs kernels* Π_1 *and* Π_2 *from the bivariate normal Gibbs sampling example. Then there exists an* $r > 0$ *such that Assumption (A.3) holds for the composition kernels* P_1^2 *and* P_2^2 *with the choices* $V_1(x) = x_1^2 + rx_2^2 + 1$ *and* $V_2(x) = rx_1^2 + x_2^2 + 1$.

Proof: First, we show that the $\{X_{2n}\}_{n=0}^{\infty}$ and $\{X_{2n+1}\}_{n=1}^{\infty}$ chains are T-chains, in the sense of Meyn and Tweedie [2009]. To do this, we show that the composition kernels $P_1^2(x,\cdot)$ and $P_2^2(x,\cdot)$ are strong Feller chains. Since the kernels $P_k^2(x,\cdot)$ are aperiodic (Lemma B.3) and ψ-irreducible, this will imply from part (ii) of Theorem 6.2.5 in Meyn and Tweedie [2009] that every compact subset of $\mathbb{R}^2$ is small.

To show that $P_1^2(x,\cdot)$ is strong Feller, we check that

$$\liminf_n P_1^2(x_n, A) \geq P_1^2(x, A)$$

for any $A \in \mathscr{R}^2$ and sequence $\{x_n\}_{n=1}^{\infty}$ with $x_n = (x_{1n}, x_{2n}) \in \mathbb{R}^2$ and $x_n \to x$.

Let $A \in \mathscr{R}^2$, and suppose $\{x_n\}_{n=1}^{\infty}$ is a sequence in $\mathbb{R}^2$ with $x_n \to x^*$. We have

$$\begin{aligned}
&\int P_1^2(x_n, dx') I_A(x') \\
&= B \int \exp\{-(1-\rho^2)^{-1}(x_2' - \rho x_{1n})^2/2\} \exp\{-(1-\rho^2)^{-1}(x_1' - \rho x_2')^2/2\} I_A(x') dx_1' dx_2'
\end{aligned}$$

where $x' = (x_1', x_2')$, and the constant $B = \{2\pi(1-\rho^2)\}^{-1}$ does not depend on x_n or x'. Now, since $x_n \to x^*$, we have in particular that $x_{1n} \to x_1^*$. Thus, by continuity,

$$\begin{aligned}
&\liminf_n \exp\{-(1-\rho^2)^{-1}(x_2' - \rho x_{1n})^2/2\} \exp\{-(1-\rho^2)^{-1}(x_1' - \rho x_2')^2/2\} \\
&= \exp\{-(1-\rho^2)^{-1}(x_2' - \rho x_1^*)^2/2\} \exp\{-(1-\rho^2)^{-1}(x_1' - \rho x_2')^2/2\}
\end{aligned}$$

Therefore, from Fatou's Lemma, we have $\liminf_n P_1^2(x_n, A) \geq P_1^2(x^*, A)$, so P_1^2 is strong Feller. The proof that P_2^2 is also strong Feller is similar. Thus, all compact sets are small for P_1^2 and P_2^2 from Theorem 6.2.5 in Meyn and Tweedie [2009].

Now, consider $V_1(x) = x_1^2 + rx_2^2 + 1$ where $0 < r < (1-\rho^4)$. Take $\lambda_1 = \rho^4 + r$. Then $\Pi_2 V_1(x) = \rho^2 x_2^2 + (1-\rho^2) + rx_2^2 + 1$ and

$$\begin{aligned}\Pi_1\Pi_2 V_1(x) &= (\rho^4 + r\rho^2)x_1^2 + (1+\rho^2+r)(1-\rho^2) + 1 \\ &\leq \lambda_1 V_1(x) + (1+\rho^2+r)(1-\rho^2) + 1 - \lambda_1 - \lambda_1 r x_2^2\end{aligned}$$

Now, take $b = (1+\rho^2+r)(1-\rho^2) + 1 - \lambda_1$ and $c > 0$ such that $\lambda_1 r c^2 \geq (1+\rho^2+r)(1-\rho^2) + 1 - \lambda_1$. Then we have $P_1^2 V_1(x) = \Pi_1\Pi_2 V_1(x) \leq \lambda_1 V_1(x) + bI_C(x)$ where $C = [-c,c] \times [-c,c]$. Since C is compact, C is small, so (A.3) is satisfied for P_1^2.

Similarly, it can be shown that P_2^2 also satisfies (A.3). □

In Lemma B.10 below, take $\Pi_1, ..., \Pi_K : X \times \mathscr{X} \to [0,1]$ to be a set of transition kernels and let $\{X_t\}_{t=0}^{\infty}$ be a Markov chain with initial law ν and transition kernel $\Pi_{\sigma^t(1)}$ at time t. For positive integers m, define the Markov chain $\{Y_t^{(m)}\}_{t=0}^{\infty}$ by $Y_t^{(m)} = (X_{Kt}, ..., X_{Kt+mK-1}) \in X^{mK}$, so that each $Y_t \in X^{mK}$ contains the the $X_{t'}$ which result from m sweeps through the K kernels starting from X_{Kt}. Let $\tilde{P}$ be the transition kernel associated to Y_t. From our definition of Y_t, the initial law $\tilde{\nu}$ of the Y_t chain is $\tilde{\nu}(dy) = \nu(dy_1)\Pi_1(y_1, dy_2)\cdots\Pi_{K-1}(dy_{mK-1}, mK)$.

Lemma B.10. *Suppose Assumptions (A.1)–(A.4) hold for the kernels P_k^K and Π_k, $k = 1, ..., K$. Then $\tilde{P}$ is $\tilde{\pi}$-stationary for the measure $\tilde{\pi}(dy_1, ..., dy_{mK}) = \pi(dy_1)\Pi_1(y_1, dy_2)\Pi_2(y_2, dy_3)\cdots$ Additionally, $\tilde{P}$ is $\tilde{\psi}$-irreducible for the measure $\tilde{\psi}(dy) = \int \psi(dx_0)\Pi_K(x_0, dy_1)\Pi_1(y_1, dy_2)\cdots\Pi_{K-1}(y_{mK}$ where the integral is taken over x_0 only. Further, there exist constants $\lambda < 1$ and $b > 0$, a function $\tilde{V} : X^{mK} \to [1, \infty)$, and a $\tilde{P}$-small set $\tilde{C}$ such that the drift condition $\tilde{P}\tilde{V}(y) \leq \lambda\tilde{V}(y) + bI_{\tilde{C}}(y)$ holds. Finally, $\tilde{P}$ is aperiodic. Thus, the strong law of large numbers holds so that*

$$\lim_{M\to\infty} M^{-1} \sum_{t=0}^{M-1} g(Y_t) = \int \tilde{\pi}(dy) g(y)$$

almost surely for functions $g : X^{mK} \to \mathbb{R}$ with $\int \tilde{\pi}(dy)|g(y)| < \infty$.

Proof of Lemma B.10: The $\tilde{\pi}$-stationarity of $\tilde{P}$ follows from the π-stationarity of the Π_k, $k = 1, ..., K$ under Assumption (A.1).

Next, we show that $\tilde{P}$ is $\tilde{\psi}$-irreducible. Let a set A be given such that $\tilde{\psi}(A) > 0$, and let $y^* \in X^{mK}$ be given. We show there exists an integer t such that $\tilde{P}^t(y^*, A) > 0$. Since $\tilde{\psi}(A) > 0$, using the definition of $\tilde{\psi}$ yields the existence of constants $\epsilon_0, \epsilon_1 > 0$ and a set $B \in X$ such that $\Pi_K(x_0, dy_1)\Pi_1(y_1, dy_2) \cdots \Pi_{K-1}(y_{mK-1}, dy_{mK}) > \epsilon_0$ for all $x_0 \in B, y \in A$, and $\psi(B) > \epsilon_1$. Now, by the ψ-irreducibility of P_K^K under Assumption (A.2), $P_K^{t_0 K}(y^*_{mK}, B) > 0$ for some integer t_0. Further, observe that $\tilde{P}^t(y, X^{mK-1} \times C) = P_K^{tK}(y_{mK}, B)$ for all $y \in X^{mK}$ and $C \in \mathscr{X}$. Thus, $\tilde{P}^{t_0+m}(y^*, A) \geq P_K^{t_0 K}(y^*_{mK}, B)\epsilon_0 > 0$, so $\tilde{P}$ is $\tilde{\psi}$-irreducible.

We now verify the drift condition. Take $\lambda = \lambda_K$, $b = b_K$, $\tilde{V}(y) = V_K(y_{mK})$, and $C = X^{mK-1} \times C_K$, with λ_K, b_K, V_K, and C_K as in Assumption (A.3). Then $\tilde{P}\tilde{V}(y) = P_K^K V_K(y_{mK}) \leq \lambda_K V_K(y_{mK}) + b_K I_{C_K}(y_{mK}) = \lambda \tilde{V}(y) + b I_C(y)$. Thus, all that remains to verify the drift condition is to show that the set C is $\tilde{P}$-small. Since C_K is P_K^K-small by Assumption (A.3), there exists an integer t such that $P_K^{tK}(x, A) \geq \nu(A)$ for all $x \in C_K$, $A \in \mathscr{X}$, for some non-trivial measure ν. Let $\tilde{\nu}(dy) = \int \nu(dx_0)\Pi_K(x_0, dy_1)\Pi_1(y_1, dy_2) \cdots \Pi_{K-1}(y_{mK-1}, dy_{mK})$. Then $\tilde{P}^{t+m}(y, A) \geq \tilde{\nu}(A)$ for all $y \in C$, $A \in \mathscr{X}^{mK}$. Thus, C is $\tilde{P}$-small.

Finally, we verify that $\tilde{P}$ is aperiodic. Suppose to the contrary that for some $d > 1$ there exist disjoint $D_1, ..., D_d \in \mathscr{B}(X^{mK})$ satisfying $\tilde{P}(y, D_{\sigma(i)}) = 1$ for $y \in D_i$, and $\tilde{\psi}((\cup_{i=1}^d D_i)^C) = 0$.

By assumption, $\tilde{P}^{md+1}(y, D_{\sigma(i)}) = 1$ for $y \in D_i$, for each $i = 1, ..., d$. By the Markov property, there exists a kernel $P : X \times \mathscr{B}(X^{mK}) \to [0, 1]$, where $P(\cdot, A) : X \to [0, 1]$ is a measurable function for each $A \in \mathscr{B}(X^{mK})$ and $P(x, \cdot)$ is a probability

measure for each $x \in X$, such that $\tilde{P}^{md+1}(y, A) = P(y_{mK}, A)$ for all $y \in X^{mK}, A \in \mathscr{B}(X^{mK})$. Let $\tilde{D}_i = P(\cdot, D_{\sigma(i)})^{-1}(\{1\}) \in \mathscr{X}$, and let $\bar{D}_i = X^{mK-1} \times \tilde{D}_i$. We have $\tilde{D}_i \cap \tilde{D}_j = \phi$ for $i \neq j$ and $\bar{D}_i \cap \bar{D}_j = \phi$ for $i \neq j$. Since $\tilde{\psi}((\cup_{i=1}^d D_i)^C) = 0$, we have $\bar{D}_i = D_i$ almost everywhere with respect to $\tilde{\psi}$. Thus the $\bar{D}_i$ are an alternative collection of disjoint sets satisfying $\tilde{P}(y, \bar{D}_{\sigma(i)}) = 1$ for $y \in D_i$ and $\tilde{\psi}((\cup_{i=1}^d \bar{D}_i)^C) = 0$. But since the $\bar{D}_i$ have the form $X^{mK-1} \times \tilde{D}_i$, we have $\tilde{P}(y, \bar{D}_i) = P_K^K(y_{mK}, \tilde{D}_i)$ for each i and thus the $\tilde{D}_i$ are a collection of disjoint sets satisfying $P_K^K(x, \tilde{D}_{\sigma(i)}) = 1$ for $x \in \tilde{D}_i$. Since $P_K^K(x, \cup_{i=1}^d \tilde{D}_i) = 1$ for $x \in \cup_{i=1}^d \tilde{D}_i$, we have $\psi((\cup_{i=1}^d D_i)^C = 0$, so that P_K^K is periodic. But this is a contradiction, since P_K^K is aperiodic under Assumption (A.4).

B.2 Proofs of Theorems, Corollaries, and Propositions

Proof of Theorem 3.1:

First, we consider the case when the initial measure $\nu = \pi$. In this case, the law of $\{X_t\}_{t=0}^{\infty}$ is P_π and $X_0 \sim \pi$. To simplify notation, we will prove the result in the univariate case where $g : X \to \mathbb{R}$, $f : X \to \mathbb{R}$, and $C_k = c_k \in \mathbb{R}$. In the remainder, for notational clarity, we will use the conventions $\Pi_t = \Pi_{\sigma^t(1)}$, $c_t = c_{\sigma^t(1)}$, and $\hat{g}_t = \hat{g}_{\sigma^t(1)}$, so that

$$S_M = \sum_{t=0}^{M-1} g(X_t) - c_{t+1}\{f(X_{t+1}) - \Pi_t f(X_t)\} \tag{B.3}$$

$$= \sum_{t=0}^{M-1} \hat{g}_t(X_t) - \Pi_t \hat{g}_{t+1}(X_t) - c_{t+1}\{f(X_{t+1}) - \Pi_t f(X_t)\}$$

$$= \hat{g}_0(X_0) - \hat{g}_M(X_M) \tag{B.4}$$

$$+ \sum_{t=0}^{M-1} \hat{g}_{t+1}(X_{t+1}) - \Pi_t \hat{g}_{t+1}(X_t) - c_{t+1}\{f(X_{t+1}) - \Pi_t f(X_t)\}$$

a.e. P_π, where we are using the identity $\hat{g}_t - \Pi_t \hat{g}_{t+1} = g$ a.e. π from Proposition 3.1 in the second equality, and rearranging the sum in the third equality.

The term $U_M := \sum_{t=0}^{M-1} \hat{g}_{t+1}(X_{t+1}) - \Pi_t \hat{g}_{t+1}(X_t) - c_{t+1}\{f(X_{t+1}) - \Pi_t f(X_t)\}$ is an L_2 martingale (since $X_0 \sim \pi$, and the $\hat{g}_t$ are square integrable with respect to π from Proposition 3.1). The remainder term $\hat{g}_0(X_0) - \hat{g}_M(X_M)$ will be shown to be small using the Law of Large Numbers for Markov chains. Thus, we expect the asymptotic behavior of S_M to be similar to that of U_M, and we will apply a central limit theorem for martingales to deal with this term.

We now introduce a martingale central limit theorem, Theorem 1, which follows immediately from Theorem 3.2, Corollary 3.1 of Hall and Heyde [1980]. We use $\xrightarrow{p}$ to denote convergence in probability.

Theorem 1. *Let $\{S_{ni}, \mathscr{F}_{ni}, 1 \leq i \leq k_n, n \geq 1\}$ be a zero-mean, square integrable martingale array with differences $Y_{ni} = S_{ni} - S_{n,i-1}$ ($S_{n0} := 0$). Suppose*

1. *(conditional Lindeberg) for all $\epsilon > 0$, $\sum_{i=1}^{k_n} E\{Y_{ni}^2 I(|Y_{ni}| > \epsilon) | \mathscr{F}_{n,i-1}\} \xrightarrow{p} 0$*
2. *(converging conditional variances) $\sum_{i=1}^{k_n} E(Y_{ni}^2 | \mathscr{F}_{n,i-1}) \xrightarrow{p} \sigma^2$*

where σ^2 is a constant. Then $S_{nk_n} = \sum_i Y_{ni} \xrightarrow{d} Z$, where the R.V. Z has characteristic function $\exp(-\sigma^2 t^2/2)$.

Now, for $i > 0$ we define $D_i = \hat{g}_i(X_i) - \Pi_{i-1}\hat{g}_i(X_{i-1}) - c_i\{f(X_i) - \Pi_{i-1}f(X_{i-1})\}$, and take $k_n = n$, $\mathscr{F}_{ni} = \sigma(X_0, ..., X_i)$, and $S_{ni} = n^{-1/2}\sum_{j=1}^{i} D_j$. From these definitions, we have $\mathscr{F}_{ni} \subset \mathscr{F}_{n,i+1}$ for $1 \leq i < n$. We will verify Conditions 1 and 2 of Theorem 1 hold for $\{S_{ni}, \mathscr{F}_{ni}, 1 \leq i \leq k_n, n \geq 1\}$ defined in this way, following Section 17.4.2 of Meyn and Tweedie [2009]. In order to motivate this approach, we note that $S_{nn} = n^{-1/2}U_n$.

Now, for $k = 1, ..., K$, we define $r_k(i) = k + (i-1)K$. For $t \geq k$, we define $m_k(t) = \max\{i \in \mathbb{N} : r_k(i) \leq t\}$. For checking the conditional Lindeberg condition 1,

it is enough to show that

$$\sum_{i=1}^{m_k(n)} E\{Y_{n,r_k(i)}^2 I(|Y_{n,r_k(i)}| > \epsilon)|\mathscr{F}_{n,r_k(i)-1}\} \xrightarrow{p} 0$$

as $n \to \infty$ for each $k = 1, ..., K$. Conditions (A.1)–(A.3) imply that for $k = 1, ..., K$, the subchains $(X_{k+Kt-1})_{t=1}^{\infty}$ are Harris recurrent with stationary measure π. Therefore, the Law of Large Numbers (Theorem 17.3.2 of Meyn and Tweedie [2009]) holds for each subchain. Consider an arbitrary k. For $i \geq 1, n \geq r_k(i)$, we have

$$E\{D_{r_k(i)}^2 I(|D_{r_k(i)}| > b)|\mathscr{F}_{n,r_k(i)-1}\} = h_k^b(X_{r_k(i)-1})$$

a.e. P_π for some π-integrable function $h_k^b : X \to \mathbb{R}$. Therefore

$$\begin{aligned}
&\limsup_n \sum_{i=1}^{m_k(n)} E\{Y_{n,r_k(i)}^2 I(|Y_{n,r_k(i)}| > b)|\mathscr{F}_{n,r_k(i)-1}\} \\
&= \limsup_n n^{-1} \sum_{i=1}^{m_k(n)} E\{D_{r_k(i)}^2 I(|D_{r_k(i)}| > n^{1/2}b)|\mathscr{F}_{n,r_k(i)-1}\} \\
&\leq \limsup_n n^{-1} \sum_{i=1}^{m_k(n)} E\{D_{r_k(i)}^2 I(|D_{r_k(i)}| > b^*)|\mathscr{F}_{n,r_k(i)-1}\} \\
&\leq K^{-1} \limsup_{n\to\infty} \{m_k(n) - 1)\}^{-1} \sum_{t=1}^{m_k(n)} h_k^{b^*}(X_{r_k(i)-1}) \\
&= K^{-1} \int \pi(dx) h_k^{b^*}(x)
\end{aligned}$$

a.e. P_π for any $b^* > 0$, where the first equality follows from the definition of Y_{ni}, and the last equality follows from applying the Law of Large numbers to the subchain $\{X_{r_k(i)-1}\}_{i=1}^{\infty}$. Now, from the properties of conditional expectation, and the dominated convergence theorem, we can find a sequence $b_j \uparrow \infty$ for which $\int \pi(dx) h_k^{b_j}(x) \leq j^{-1}$

for each j. Thus, we obtain

$$\limsup_n \sum_{i=1}^{m_k(n)} E\{Y^2_{n,r_k(i)} I(|Y_{n,r_k(i)}| > b)|\mathscr{F}_{n,r_k(i)-1}\} \leq (jK)^{-1}$$

almost surely for each j, so the event

$$\left\{ \limsup_n \sum_{i=1}^{m_k(n)} E\{Y^2_{n,r_k(i)} I(|Y_{n,r_k(i)}| > b)|\mathscr{F}_{n,r_k(i)-1}\} = 0 \right\}$$
$$= \cap_j \left\{ \limsup_n \sum_{i=1}^{m_k(n)} E\{Y^2_{n,r_k(i)} I(|Y_{n,r_k(i)}| > b)|\mathscr{F}_{n,r_k(i)-1}\} \leq (jK)^{-1} \right\}$$

has probability 1. Repeating this argument for each $k = 1, ..., K$ verifies the conditional Lindeberg condition 1.

To verify the variance convergence in condition 2, we use the Law of Large Numbers on each subchain again to obtain $\sum_i E(Y^2_{ni}|\mathscr{F}_{n,i-1}) \overset{a.s.-P_\pi}{\to} \sigma^2$ where

$$\sigma^2 = K^{-1} \sum_{k=1}^{K} \int \pi(dx)\Pi_k(x, dy)[\hat{g}_{\sigma(k)}(y) - \Pi_k \hat{g}_{\sigma(k)}(x) - c_{\sigma(k)}\{f(y) - \Pi_k f(x)\}]^2$$
$$= K^{-1} \sum_{k=1}^{K} \big[\langle \hat{g}_{\sigma(k)} - c_{\sigma(k)} f, \hat{g}_{\sigma(k)} - c_{\sigma(k)} f\rangle$$
$$- \langle \Pi_k \hat{g}_{\sigma(k)} - c_{\sigma(k)} \Pi_k f, \Pi_k \hat{g}_{\sigma(k)} - c_{\sigma(k)} \Pi_k f\rangle\big] \tag{B.5}$$

The convergence in probability in Condition 2 of Theorem 1 then follows immediately from the almost sure convergence. Thus by Theorem 1, we have $S_{nn} \overset{d}{\to} Z$ where Z has characteristic function $\exp(-\sigma^2 t^2/2)$.

We now deal with the remainder term $\hat{g}_0(X_0) - \hat{g}_M(X_M) - c_0 f(X_0) + c_M f(X_M)$. Clearly, $M^{-1/2}\hat{g}_0(X_0) - c_0 f(X_0) \overset{a.s.}{\to} 0$ as $M \to \infty$. Additionally, from the Law of Large Numbers applied to each subchain,

$$\sum_{t=0}^{M-1} M^{-1}\{\hat{g}_k(X_{k+Kt-1}) - c_k f(X_{k+Kt-1})\}^2 \to \int \pi(dx)\{\hat{g}_k(x) - c_k f(x)\}^2 < \infty$$

almost surely as $M \to \infty$ for each $k = 1, ..., K$. Therefore, $M^{-1/2}\{\hat{g}_M(X_M) + c_M f(X_M)\} \xrightarrow{a.s.} 0$ also.

Applying Slutsky's Theorem, we obtain $M^{-1/2} S_M \xrightarrow{d} Z$ where Z has characteristic function $\exp(-\sigma^2 t^2/2)$.

Now, we have $\sigma^2 = K^{-1} \sum_{k=1}^{K} B_k$ where

$$B_k = \langle \hat{g}_{\sigma(k)} - c_{\sigma(k)} f, \hat{g}_{\sigma(k)} - c_{\sigma(k)} f \rangle - \langle \Pi_k \hat{g}_{\sigma(k)} - c_{\sigma(k)} \Pi_k f, \Pi_k \hat{g}_{\sigma(k)} - c_{\sigma(k)} \Pi_k f \rangle$$

$$= \langle \hat{g}_{\sigma(k)} + \Pi_k \hat{g}_{\sigma(k)}, \hat{g}_{\sigma(k)} - \Pi_k \hat{g}_{\sigma(k)} \rangle - 2c_{\sigma(k)} (\langle f, \hat{g}_{\sigma(k)} \rangle - \langle \Pi_k f, \Pi_k \hat{g}_{\sigma(k)} \rangle)$$

$$+ c^2_{\sigma(k)} (\langle f, f \rangle - \langle \Pi_k f, \Pi_k f \rangle) \quad (k = 1, ..., K).$$

Note

$$\sum_{k=1}^{K} \langle \hat{g}_{\sigma(k)} + \Pi_k \hat{g}_{\sigma(k)}, \hat{g}_{\sigma(k)} - \Pi_k \hat{g}_{\sigma(k)} \rangle$$

$$= \sum_{k=1}^{K} \langle \hat{g}_{\sigma(k)}, \hat{g}_{\sigma(k)} \rangle - \langle \Pi_k \hat{g}_{\sigma(k)}, \Pi_k \hat{g}_{\sigma(k)} \rangle$$

$$= \sum_{k=1}^{K} \langle \hat{g}_k, \hat{g}_k \rangle - \langle \Pi_k \hat{g}_{\sigma(k)}, \Pi_k \hat{g}_{\sigma(k)} \rangle$$

$$= \sum_{k=1}^{K} \langle \hat{g}_k + \Pi_k \hat{g}_k, \hat{g}_k - \Pi_k \hat{g}_{\sigma(k)} \rangle$$

$$= \sum_{k=1}^{K} \langle g, g \rangle + 2 \sum_{t=1}^{\infty} \langle g, P^t_k g \rangle$$

where the last equality used Proposition 3.1 to simplify $\hat{g}_k - \Pi_k \hat{g}_{\sigma(k)}$. Thus,

$$\sigma^2 = K^{-1} \sum_{k=1}^{K} B_k = \langle g, g \rangle + 2K^{-1} \sum_{k=1}^{K} \sum_{t=1}^{\infty} \langle g, P^t_k g \rangle$$

$$+ K^{-1} \sum_{k=1}^{K} c^2_{\sigma(k)} (\langle f, f \rangle - \langle \Pi_k f, \Pi_k f \rangle) - 2c_{\sigma(k)} (\langle f, \hat{g}_{\sigma(k)} \rangle - \langle \Pi_k f, \Pi_k \hat{g}_{\sigma(k)} \rangle)$$

We now extend to the multivariate case by the Cramer-Wold device. Let f :

$X \to \mathbb{R}^d$, $g : X \to \mathbb{R}^p$, $C_k \in \mathbb{R}^{p \times d}$, and

$$S_M = \sum_{t=0}^{M-1} g(X_t) - C_t^T f(X_t) + C_{t+1}^T \Pi_t f(X_t).$$

$a \in \mathbb{R}^d$. Define $U_k = \int \pi(dx)\{ff^T - (\Pi_k f)(\Pi_k f^T)\}$ and $V_k = \int \pi(dx)\{f\hat{g}_{\sigma(k)}^T - (\Pi_k f)(\Pi_k \hat{g}_{\sigma(k)})^T\}$. Then we have $a^T M^{-1/2} S_M \xrightarrow{d} Z$ where Z is a random variable with characteristic function $\exp(-a^T \Sigma_C a t^2/2)$, with

$$\Sigma_C = \int \pi(dx) g g^T + 2K^{-1} \sum_{k=1}^{K} \sum_{t=1}^{\infty} \int \pi(dx) g(P_k g)^T$$
$$+ K^{-1} \sum_{k=1}^{K} C_{\sigma(k)}^T U_k C_{\sigma(k)} - C_{\sigma(k)}^T V_k - V_k^T C_{\sigma(k)}.$$

Since this holds for arbitrary a, we have by the Cramer-Wold Theorem that $M^{-1/2} S_M \xrightarrow{d} Z$, where Z is a random variable with characteristic function $\exp(-t^T \Sigma_C t/2)$.

Finally, we extend from the multivariate case with initial measure π, to the multivariate case with initial measure $\nu \neq \pi$. In this case, the desired convergence in distribution can be shown to hold via a coupling argument, as in Roberts and Rosenthal [2004]. We sketch the proof here. We construct on the same probability space two Markov chains $\{X_t\}_{t=0}^{\infty}$ and $\{\tilde{X}_t\}_{t=0}^{\infty}$, with initial law $\nu \times \pi$ for $(X_0, \tilde{X}_0)$. Then, we update the chains using a joint transition kernel chosen so that

1. each chain is marginally a Markov chain with transition kernel $\Pi_t = \Pi_{\sigma^t(1)}$ at time t, and

2. $X_t = \tilde{X}_t$ for all $t > t_0$, for some random t_0, almost surely.

The aperiodicity assumption (A.4) and the geometric drift to the petite set C in Assumption (A.3) ensure such a transition kernel can be constructed.

Then, $M^{-1/2}(S_M - \tilde{S}_M) \xrightarrow{a.s.} 0$, where $S_M = \sum_{t=0}^{M-1} g(X_t) - C_{t+1}^T\{f(X_{t+1}) - \Pi_t f(X_t)\}$ and $\tilde{S}_M = \sum_{t=0}^{M-1} g(\tilde{X}_t) - C_{t+1}^T\{f(\tilde{X}_{t+1}) - \Pi_t f(\tilde{X}_t)\}$. Thus, from Slutsky's theorem, $M^{-1/2}S_M \xrightarrow{d} Z$ where Z has characteristic function $\exp(-t^T\Sigma_C t/2)$.

Now, we show that Σ_C is minimized when $C_{\sigma(k)} = U_k^\dagger V_k$. First, we show $U_k a = 0_{1\times d}$ implies $a^T V_k = 0$. To see this, note that $U_k a = 0$ implies

$$\int \pi(dx)\Pi_k(x, dy)a^T\{f(y) - \Pi_k f(x)\}\{f(y) - \Pi_k f(x)\}^T a = 0$$

so that $a^T\{f(y) - \Pi_k f(x)\} = 0$ a.e. λ_k, where λ_k is the measure on $(X^2, \mathscr{F}^2)$ defined by $\lambda_k(A \times B) = \int \pi(dx)\Pi_k(x, dy)I(x \in A, y \in B)$. In this case,

$$\begin{aligned} a^T V_k &= \int \pi(dx)a^T\{f\hat{g}_{\sigma(k)} - \Pi_k f\Pi_k\hat{g}_{\sigma(k)}^T\} \\ &= \int \lambda_k(dx \times dy)a^T\{f(y) - \Pi_k f(x)\}\{\hat{g}_{\sigma(k)}(y) - \Pi_k\hat{g}_{\sigma(k)}(x)\} \\ &= 0_{1\times d}. \end{aligned}$$

Finally, we note that Σ_C depends on $C_{\sigma(k)}$ only through the term $K^{-1}(C_{\sigma(k)}^T U_k C_{\sigma(k)} - C_{\sigma(k)}^T V_k - V_k^T C_{\sigma(k)})$. By Lemma B.6, this term is minimized when $C_{\sigma(k)} = U_k^\dagger V_k$. This completes the proof. □

Proof of Corollary 3.1: First, we obtain the simplified expression $V = K^{-1}\sum_{k=1}^K V_k = \int \pi(dx)fg^T$ for V. Under the Gibbs kernel assumption (A.1b),

$$
\begin{aligned}
V = K^{-1}\sum_{k=1}^{K} V_k &= K^{-1}\sum_{k=1}^{K}\int \pi(dx)\{f\hat{g}_{\sigma(k)}^T - \Pi_k f(\Pi_k \hat{g}_{\sigma(k)}^T)\} \\
&= K^{-1}\sum_{k=1}^{K}\int \pi(dx)\{f\hat{g}_{k}^T - \Pi_k f(\Pi_k \hat{g}_{\sigma(k)}^T)\} \\
&= K^{-1}\sum_{k=1}^{K}\int \pi(dx)\{f\hat{g}_{k}^T - f(\Pi_k \hat{g}_{\sigma(k)}^T)\} \\
&= K^{-1}\sum_{k=1}^{K}\int \pi(dx) f g^T = \int \pi(dx) f g^T
\end{aligned}
$$

where the second line rearranged the sum of the $f\hat{g}_{\sigma(k)}$ terms, and the third line used the equality $\int \pi(dx)\Pi_k f(\Pi_k \hat{g}_{\sigma(k)})^T = \int \pi(dx) f(\Pi_k \hat{g}_{\sigma(k)})^T$ from reversibility and idempotence of Π_k. The last line follows from Proposition 3.1.

In general, we have

$$
K^{-1}\sum_{k=1}^{K} C^T U_k C - V_k^T C - C^T V_K = C^T U C - C^T V - V^T C,
$$

and

$$
\begin{aligned}
\Sigma_C = &\int \pi(dx) g g^T + K^{-1}\sum_{k=1}^{K}\sum_{t=1}^{\infty}\int \pi(dx)\{g(P_k g)^T + (P_k g)g^T\} \\
&+ K^{-1}\sum_{k=1}^{K} C_{\sigma(k)}^T U_k C_{\sigma(k)} - C_{\sigma(k)}^T V_k - V_k^T C_{\sigma(k)}. \\
&= \int \pi(dx) g g^T + K^{-1}\sum_{k=1}^{K}\sum_{t=1}^{\infty}\int \pi(dx)\{g(P_k g)^T + (P_k g)g^T\} \\
&+ C^T U C - C^T V - V^T C
\end{aligned}
$$

which is the representation of Σ_C given in Corollary 3.1.

Now, we show that $Ua = 0$ implies $a^T V = 0$, so that we may apply Lemma B.6 to the term $C^T U C - C^T V - V^T C$. First, we have that $Ua = 0$ implies $a^T U a = 0$,

so from Lemma B.5, we have $Ua = 0$ implies $a^T f = b$ a.e. π. In this case $a^T V = 0$ from the same reasoning as in the proof of Theorem 3.1. Thus, Lemma B.6 shows that $C^T UC - C^T V - V^T C$ is minimized when $C = \tilde{C}$, where $\tilde{C} = U^\dagger V$. Since Σ_C depends on C only through $C^T UC - C^T V - V^T C$, we have that Σ_C is minimized at $C = \tilde{C}$. This completes the proof. □

Proof of Corollary 3.2:

We have

$$\begin{aligned}
\Sigma_1 &= \Sigma_0 + K^{-1} \sum_{k=1}^{K} C_{\sigma(k)}^T U_k C_{\sigma(k)} - C_{\sigma(k)}^T V_k - V_k^T C_{\sigma(k)} \\
&= \Sigma_0 + K^{-1} \sum_{k=1}^{K} U_k - 2V_k \\
&= \Sigma_0 - \int \pi(dx) g g^T - K^{-1} \sum_{k=1}^{K} \int \pi(dx) (\Pi_k g)(\Pi_k g)^T \leq \Sigma_0,
\end{aligned}$$

where the first equality used Theorem 3.1, and the second equality used $C_k = I_{d\times d}$ for each $k = 1, ..., K$ and the fact $f = g$. The third equality results from applying identity $K^{-1} \sum_{k=1}^{K} V_k = \int \pi(dx) = \int \pi(dx) g g^T$. The inequality holds since both integrals are of nonnegative functions, so that the subtracted integrands are nonnegative. □

Proof of Theorem 3.2:

Under the Assumptions in the statement of Theorem 3.2, we show

$$\begin{aligned}
M^{-1/2} \sum_{t=0}^{M-1} (\hat{C}^B_{\sigma^t(1),M} - \tilde{C}_{\sigma^t(1)})^T \{f(X_{t+1}) - \Pi_{\sigma^t(1)} f(X_t)\} &\xrightarrow{p} 0 \\
M^{-1/2} \sum_{t=0}^{M-1} (\hat{C}^B_{M} - \tilde{C})^T \{f(X_{t+1}) - \Pi_{\sigma^t(1)} f(X_t)\} &\xrightarrow{p} 0 \\
M^{-1/2} \sum_{t=0}^{M-1} (\hat{C}^{Gibbs}_{M} - \tilde{C})^T \{f(X_{t+1}) - \Pi_{\sigma^t(1)} f(X_t)\} &\xrightarrow{p} 0
\end{aligned}$$

The result then follows from Slutsky's Theorem and Theorem 3.1.

We give the proof for $M^{-1/2}\sum_{t=0}^{M-1}(\hat{C}_M^{Gibbs}-\tilde{C})^T\{f(X_{t+1})-\Pi_{\sigma^t(1)}f(X_t)\}\xrightarrow{p}0$. The other convergence results follow similarly. Let $U=Q_0DQ_0^T$ be an eigendecomposition of U, so that $Q\in\mathbb{R}^{p\times r}$ with $r<p$ columns and D is a diagonal matrix with positive entries on the diagonal. Let $Q_1\in\mathbb{R}^{p\times(p-r)}$ be an orthonormal basis for the orthogonal complement $Q_1^\perp$ of Q_1 in $\mathbb{R}^p$. Note $I_{p\times p}=Q_0Q_0^T+Q_1Q_1^T$ From Lemma B.5, we have $Q_1^T\{f(X_{t+1})-\Pi_{\sigma^t(1)}f(X_t))\}\neq 0$ only finitely many times under (A.1)–(A.3). Furthermore from Lemma B.5, we have for $Q_1^Tf\neq c$ for some vector c only finitely many times. Note

Proof of Theorem 3.3: First, consider the case where $g:X\to\mathbb{R}$ and $f:X\to\mathbb{R}$, and $C\in\mathbb{R}$, so that S_M is a sum of scalar terms.

From Lemma B.7, we have $\sum_{t=0}^{\infty}|Q^tg|$ is square integrable with respect to π. Thus, $\hat{h}=-Cf+\sum_{t=0}^{\infty}Q^tg$ is square integrable and satisfies the Poisson equation $\hat{h}-Q\hat{h}=h$ a.e. π.

It can be then be shown from the same martingale central limit theorem approach as in Theorem 3.1 that for the random sweep chain, we have $M^{-1/2}S_M\to Z$ where Z is a random variable with characteristic function $\exp(\Sigma_C^{RS}t^2/2)$, with $\Sigma_C^{RS}=\langle\hat{h},\hat{h}\rangle-\langle Q\hat{h},Q\hat{h}\rangle$ denoting the asymptotic variance Σ_{RS}^C for the control variate scheme with arbitrary C, in distinction to Σ_{RS} which denotes the asymptotic variance with the optimal C. We first show that $\Sigma_C^{RS}=\langle h,h\rangle+2\sum_{t=1}^{\infty}\langle h,Q^th\rangle$. We have

$$\begin{aligned}\Sigma_C^{RS}&=\langle\hat{h},\hat{h}\rangle-\langle Q\hat{h},Q\hat{h}\rangle\\&=\langle -C(f+Qf)+g+2\sum_{t=1}^{\infty}Q^tg,g-C(f-Qf)\rangle\\&=\langle h-2CQf+2\sum_{t=1}^{\infty}Q^tg,h\rangle\end{aligned}\tag{B.6}$$

where the first equality follows from the martingale CLT, and the second inequality

follows from the identify $a^2 - b^2 = (a+b)(a-b)$. Also,

$$\begin{aligned}
2\sum_{t=1}^{\infty}\langle h, Q^t h\rangle &= 2\sum_{t=1}^{\infty}\langle h, Q^t g - CQ^t(f-Qf)\rangle \\
&= -2C\sum_{t=1}^{\infty}\langle h, Q^t(f-Qf)\rangle + 2\sum_{t=1}^{\infty}\langle h, Q^t g\rangle \\
&= -2C\langle h, Qf\rangle + 2C\lim_{t\to\infty}\langle h, Q^{t+1}f\rangle + 2\sum_{t=1}^{\infty}\langle h, Q^t g\rangle .
\end{aligned}$$

Now, we define $\|f\| = \langle f, f\rangle^{1/2}$. From Lemma 2 in Burkholder and Chow [1961], we have since Q is positive and self-adjoint that there exists an idempotent, self adjoint operator $\bar{Q}$ such that $\lim_{t\to\infty}\|\bar{Q}r - Q^t r\| = 0$ for any function $r : X \to \mathbb{R}$ with $\langle r, r\rangle < \infty$. But for such a $\bar{Q}$, we have $Q\bar{Q}f = \bar{Q}f$ a.e. π, since

$$\begin{aligned}
\|Q\bar{Q}f - \bar{Q}f\| &\leq \|Q\bar{Q}f - Q^t f\| + \|Q^t f - \bar{Q}f\| \\
&\leq \|\bar{Q}f - Q^{t-1}f\| + \|Q^t f - \bar{Q}f\|
\end{aligned}$$

and $\lim_{t\to\infty}\|\bar{Q} - Q^{t-1}f\| + \|Q^t f - \bar{Q}f\| = 0$. Since $Q\bar{Q}f = \bar{Q}f$ a.e. π, we have $\Pi_1\bar{Q}f = \bar{Q}f$ a.e. π and $\Pi_2\bar{Q}f = \bar{Q}f$ a.e. π. Thus, from Lemma B.4, $\bar{Q}f$ is constant a.e. π, so that $\lim_{t\to\infty}\langle h, Q^{t+1}f\rangle = \langle h, \bar{Q}f\rangle = 0$, since $\int \pi(dx)h(x) = 0$. Therefore, $2\sum_{t=1}^{\infty}\langle h, Q^t h\rangle = -2C\langle h, Qf\rangle + 2\sum_{t=1}^{\infty}\langle h, Q^t g\rangle$, so that we may rewrite (B.6) as

$$\Sigma_C^{RS} = \langle h, h\rangle + 2\sum_{t=1}^{\infty}\langle h, Q^t h\rangle$$

We now show that $\Sigma_C = \langle h, h\rangle + \sum_{t=1}^{\infty}\langle h, Qh\rangle$ where Σ_C is the asymptotic variance in Corollary 3.1 for the fixed weight scheme. First, we observe

$$\begin{aligned}
\sum_{t=1}^{\infty}\langle h, Q^t g\rangle &= \sum_{t=1}^{\infty}\langle g, (P_1^t + P_2^t)g\rangle - C\sum_{t=1}^{\infty}\langle g, Q^t(f-Qf)\rangle \\
&= -C\langle g, Qf\rangle + C\lim_{t\to\infty}\langle g, Q^{t+1}f\rangle + \sum_{t=1}^{\infty}\langle g, (P_1^t + P_2^t)g\rangle \\
&= -C\langle g, Qf\rangle + \sum_{t=1}^{\infty}\langle g, (P_1^t + P_2^t)g\rangle
\end{aligned}$$

Thus,

$$\begin{aligned}\langle h,h\rangle + \sum_{t=1}^{\infty}\langle h, Q^t h\rangle &= \langle h,h\rangle - C\langle h, Qf\rangle - C\langle g, Qf\rangle + \sum_{t=1}^{\infty}\langle g,(P_1^t + P_2^t)g\rangle \\ &= \langle g,g\rangle + \sum_{t=1}^{\infty}\langle g,(P_1^t+P_2^t)g\rangle \\ &\quad - 2C\langle f,g\rangle + \sum_{k=1}^{2} C^2(\langle f,f\rangle - \langle \Pi_k f, \Pi_k f\rangle)\end{aligned}$$

which coincides with the asymptotic variance in Corollary 3.1.

The extension to multivariate $g : X \to \mathbb{R}^d$, $f : X \to \mathbb{R}^p$, and $C : X \to \mathbb{R}^{d\times p}$ follows via the Cramer-Wold device. We have $M^{-1/2}S_M \to Z$ where Z has characteristic function $\exp(t^T\Sigma_C^{RS}t/2)$ with

$$\Sigma_C^{RS} = \int \pi(dx)hh^T + 2\sum_{t=1}^{\infty}\int \pi(dx)h(Q^t h)^T$$

for the random sweep chain and

$$\Sigma_C = \int \pi(dx)hh^T + \sum_{t=1}^{\infty}\int \pi(dx)h(Q^t h)^T.$$

for the deterministic sweep chain. The expression in Theorem 3.3 for the difference $\Sigma_{\tilde{C}} - \Sigma_{RS}$ between the optimal variances is obtained by arithmetic. We observe that $\sum_{t=1}^{\infty}\int \pi(dx)h(Qh)^T$ is positive semidefinite since Q is a positive, self-adjoint operator and therefore has a positive, self-adjoint square root $\tilde{Q}$ with $\tilde{Q}\tilde{Q} = Q$, so that $\int \pi(dx)h(Q^t h)^T = \int \pi(dx)\tilde{Q}^t h(\tilde{Q}^t h)^T \geq 0$.

□

Proof of Proposition 3.1: We first prove the result for univariate $g : X \to \mathbb{R}$. Define $\tilde{g}_k(x) = \sum_{t=0}^{\infty} P_k^{Kt} g(x)$, for each k. By Assumption (A.4), the kernels P_k^K are aperiodic. Additionally, from Assumption (A.2), Markov chains resulting from P_k^K

are irreducible. From the geometric drift condition (A.3), Theorem 15.0.1 of Meyn and Tweedie [2009] implies

$$\sum_{t=0}^{\infty} |P_k^{Kt} g(x)| \leq RV_k(x) \tag{B.7}$$

for some $R < \infty$ and all $x \in X$, for $k = 1, ..., K$.

Recall the definition

$$\hat{g}_k(x) = \sum_{t=0}^{\infty} P_k^t g(x) \qquad k = 1, ..., K$$

We now show that $\sum_{t=0}^{\infty} |P_k^t g(x)|$ is square integrable with respect to π for each k. First, we note that it is sufficient to prove $\sum_{t=0}^{\infty} |P_k^m \{P_{\sigma^m(k)}^{Kt} g\}|$ is square integrable with respect to π for each $m = 0, ..., K-1$, since in this case

$$\sum_{t=0}^{\infty} |P_k^t g(x)| = \sum_{m=0}^{K-1} \sum_{t=0}^{\infty} |P_k^m \{P_{\sigma^m(k)}^{Kt} g\}|$$

.

Now, we have

$$\int \pi(dx) \left[P_k^m \left\{ \sum_{t=0}^{\infty} |P_{\sigma^m(k)}^{Kt} g| \right\} \right]^2 \leq \int \pi(dx) P_k^m \left[\left\{ \sum_{t=0}^{\infty} |P_{\sigma^m(k)}^{Kt} g| \right\}^2 \right]$$

$$\leq \int \pi(dx) P_k^m \{R^2 V_{\sigma^m(k)}^2\} = \int \pi(dx) R^2 V_{\sigma^m(k)}^2 < \infty$$

for each $m = 0, ..., K-1$. The first inequality follows from Jensen's inequality, the second inequality follows from (B.7) and the equality follows because the Π_k preserve the stationary probability distribution π.

We may then apply Fubini's theorem for π a.e. x to obtain

$$\int \pi(dx) \left[P_k^m \left\{ \sum_{t=0}^{\infty} |P_{\sigma^m(k)}^{Kt} g| \right\} \right]^2$$

$$= \int \pi(dx) \left[\sum_{t=0}^{\infty} P_k^m \{ |P_{\sigma^m(k)}^{Kt} g| \} \right]^2 < \infty.$$

Now, from Jensen's inequality, we have

$$\sum_{t=0}^{\infty} |P_k^m\{P_{\sigma^m(k)}^{Kt} g\}| \leq \sum_{t=0}^{\infty} P_k^m\{|P_{\sigma^m(k)}^{Kt} g|\}.$$

so that $\sum_{t=0}^{\infty} |P_k^m\{P_{\sigma^m(k)}^{Kt} g\}|$ is square integrable with respect to π, for each $m = 0, ..., K-1$.

Thus, $\sum_{t=0}^{\infty} |P_k^t g|$ is square integrable with respect to π, and also $\hat{g}_k$ is square integrable with respect to π.

Now, we verify $\hat{g}_k - \Pi_k \hat{g}_{\sigma(k)} = g$ a.e. π for each k. Since $\Pi_k \sum_{t=0}^{\infty} |P_{\sigma(k)}^t g|$ is square integrable with respect to π, we have from Fubini's theorem that

$$\Pi_k \hat{g}_{\sigma(k)} = \Pi_k \sum_{t=0}^{\infty} P_{\sigma(k)}^t g = \sum_{t=0}^{\infty} \Pi_k P_{\sigma(k)}^t g = \sum_{t=1}^{\infty} P_k^t g$$

for π a.e. x, so that $\hat{g}_k - \Pi_k \hat{g}_{\sigma(k)} = g$ for π a.e. x.

Now, for general $g : X \to \mathbb{R}^d$, we have $|a^T g| \leq V_k$ from Assumption (B.2) whenever $\|a\|_2 \leq 1$. In particular, taking a to be the vectors $e_i, i = 1, ..., d$ with e_i having 1 in the ith position and 0 elsewhere, we see from the previous reasoning that the conclusions of the Proposition still hold. We have $\hat{g}_k - \Pi_k \hat{g}_{\sigma(k)} = g$ a.e. π. Additionally $\int \pi(dx)\{\sum_{t=0}^{\infty} |P_k^t g|\}^T \{\sum_{t=0}^{\infty} |P_k^t g|\} < \infty$, so that the sum $\sum_{t=0}^{\infty} P_k^t g$ converges absolutely, elementwise, for π a.e. x, and each of the d components of $\sum_{t=0}^{\infty} |P_k^t g|$ are square integrable with respect to π. □

Proof of Proposition 3.2: First, we show that the LWK conditioning approach is, to within an asymptotically negligible term, an instance of the control variate scheme in (3.6) with $C = 2I_{d\times d}$. Note that $Qg = (\Pi_1 g + \Pi_2 g)/2 = g/2 + \Pi_1 g/2$. Define $H_M = \sum_{t=0}^{M-1} \Pi_1 g(X_t) = \sum_{t=0}^{M-1} g(X_t) - 2\{g(X_t) - Qg(X_t)\}$. Then for $S_M = \sum_{t=0}^{M-1} g(X_t) - 2\{g(X_t) - \Pi_{\sigma^t(1)} g(X_t)\}$, we have from Lemma B.8 that $M^{-1/2}(S_M -$

$H_M) \to 0$, so that $M^{-1/2}H_M$ has the same asymptotic distribution as $M^{-1/2}S_M$. Thus, $\Sigma_{\text{LWK}} = \Sigma_2$.

Since $\Pi_2 g = g$ a.e. π, we have $U = U_1/2 = (A - B)/2$. We also have $V = B$, and the term $C^T UC - V^T C - C^T V$ in the representation of Σ_C in Corollary 3.1 can be written as $C^T(A-B)C/2 - C^T A - A^T C$. Now, $\tilde{C} = U^\dagger V = 2(A-B)^{-1}A$. Substituting $C = \tilde{C}$ and $C = 2I_{d\times d}$ into (B.5) and subtracting yields

$$\begin{aligned}
&\Sigma_{\tilde{C}} - \Sigma_2 = -2A(A-B)^{-1}A - (-2A - 2B) \\
&= -2A^T(A-B)^{-1}A + 2(A-B)(A-B)^{-1}A + 2B(A-B)^{-1}(A-B) \\
&= -2B(A-B)^{-1}B \leq 0,
\end{aligned}$$

where the first equality used the symmetry of A and B, and the final inequality used the fact that $A - B$ is positive semidefinite. When B is positive definite, the inequality is strict.

Similarly,

$$\Sigma_2 - \Sigma_1 = -2(B + A) - \{(A - B)/2 - 2A\} = -(A + 3B)/2 < 0$$

since A is positive definite from Assumption (D.2) and B is positive semidefinite. Finally, $\Sigma_1 - \Sigma_0 = -(B + 3A)/2 < 0$. This completes the proof. □

Bibliography

B.V. Barnes, D.R. Zak, S.R. Denton, and S.H. Spurr. *Forest Ecology*. Wiley, fourth edition, 2010.

Albert Benveniste, Michel Métivier, and Pierre Priouret. *Adaptive Algorithms and Stochastic Approximations*, volume 22 of *Applications of Mathematics (New York)*. Springer-Verlag, Berlin., 1990.

Stephen Berg, Jun Zhu, and Murray K Clayton. Control variates and rao-blackwellization for deterministic sweep markov chains. *arXiv preprint arXiv:1912.06926*, 2019a.

Stephen Berg, Jun Zhu, Murray K. Clayton, Monika E. Shea, and David J. Mladenoff. A latent discrete Markov random field approach to identifying and classifying historical forest communities based on spatial multivariate tree species counts. *Ann. Appl. Stat.*, 13:2312–2340, 12 2019b.

David Blackwell. Conditional expectation and unbiased sequential estimation. *The Annals of Mathematical Statistics*, 18:105–110, 1947.

Nicolas Brosse, Alain Durmus, Sean Meyn, and Éric Moulines. Diffusion approximations and control variates for MCMC. *arXiv preprint arXiv:1808.01665*, 2018.

D.L. Burkholder and Y.S. Chow. Iterates of conditional expectation operators. *Proceedings of the American Mathematical Society*, 12:490–495, 1961.

Russell M. Burns and Barbara H. Honkala. *Silvics of North America*, volume 2. United States Department of Agriculture Washington, DC, 1990.

George Casella and Christian P. Robert. Rao-Blackwellisation of sampling schemes. *Biometrika*, 83:81–94, 1996.

Jiahua Chen. Consistency of the MLE under mixture models. *Statistical Science*, 32: 47–63, 2017.

Francis Comets and Basilis Gidas. Parameter estimation for Gibbs distributions from partially observed data. *The Annals of Applied Probability*, 2:142–170, 1992.

J.T. Curtis. *The Vegetation of Wisconsin: An Ordination of Wisconsin Plant Communities*. University of Wisconsin Press, 1959.

Margaret B. Davis, Mark W. Schwartz, and Kerry Woods. Detecting a species limit from pollen in sediments. *Journal of Biogeography*, 18:653–668, 1991.

Petros Dellaportas and Ioannis Kontoyiannis. Control variates for estimation based on reversible Markov chain Monte Carlo samplers. *Journal of the Royal Statistical Society: Series B*, 74:133–161, 2012.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B*, 39: 1–38, 1977.

Randal Douc and Christian P. Robert. A vanilla Rao–Blackwellization of Metropolis–Hastings algorithms. *The Annals of Statistics*, 39:261–277, 2011.

Rick Durrett. *Probability: Theory and Examples*. Cambridge University Press, Cambridge, fourth edition, 2010.

Dave Egan. *The Historical Ecology Handbook: a Restorationist's Guide to Reference Ecosystems*. Island Press, 2005.

Robert T. Fahey, Craig G. Lorimer, and David J. Mladenoff. Habitat heterogeneity and life-history traits influence presettlement distributions of early-successional tree species in a late-successional, hemlock-hardwood landscape. *Landscape Ecology*, 27: 999–1013, 2012.

Robert W. Finley. The original vegetation cover of Wisconsin. *Wisconsin DNR publication*, 1976.

F. Forbes and G. Fort. Combining Monte Carlo and mean-field-like methods for inference in hidden Markov random fields. *IEEE Transactions on Image Processing*, 16:824–837, 2007.

Florence Forbes, Myriam Charras-Garrido, Lamiae Azizi, Senan Doyle, and David Abrial. Spatial risk mapping for rare disease with hidden Markov fields and variational EM. *The Annals of Applied Statistics*, 7:1192–1216, 2013.

Gersende Fort and Eric Moulines. Convergence of the Monte Carlo expectation maximization for curved exponential families. *The Annals of Statistics*, 31:1220–1259, 2003.

Carlo Gaetan and Xavier Guyon. *Spatial statistics and modeling*. Springer, 2010.

Ronald E. Gangnon and Murray K. Clayton. A hierarchical model for spatially clustered disease rates. *Statistics in Medicine*, 22:3213–3228, 2003.

Andrew Gelman and Xiao-Li Meng. Simulating normalizing constants: from importance sampling to bridge sampling to path sampling. *Statistical Science*, 13:163–185, 1998.

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

Charles J. Geyer. Conditioning in Markov chain Monte Carlo. *Journal of Computational and Graphical Statistics*, 4:148–154, 1995.

Ian Goodfellow, Mehdi Mirza, Aaron Courville, and Yoshua Bengio. Multi-prediction deep Boltzmann machines. In *Advances in Neural Information Processing Systems*, NIPS'13, pages 548–556. Curran Associates Inc., 2013.

Mikhail Iosifovich Gordin. The central limit theorem for stationary processes. In *Doklady Akademii Nauk*, volume 188, pages 739–741. Russian Academy of Sciences, 1969.

Gaël Guennebaud, Benoît Jacob, et al. Eigen v3 sparse matrix documentation. http://eigen.tuxfamily.org, 2010.

P. Hall and C. C. Heyde. *Martingale Limit Theory and its Application*. Academic Press, New York, 1980.

Shane G. Henderson and Peter W. Glynn. Approximating martingales for variance reduction in Markov process simulation. *Mathematics of Operations Research*, 27: 253–271, 2002.

Chuan Hong, Yang Ning, Shuang Wang, Hao Wu, Raymond J. Carroll, and Yong Chen. PLEMT: A novel pseudolikelihood-based EM test for homogeneity in generalized exponential tilt mixture models. *Journal of the American Statistical Association*, 112:1393–1404, 2017.

Leonhard Knorr-Held and Günter Raßer. Bayesian detection of clusters and discontinuities in disease maps. *Biometrics*, 56:13–21, 2004.

H. Kushner and G. Yin. *Stochastic Approximation and Recursive Algorithms and Applications*. Springer-Verlag New York, 1997.

Andrew B. Lawson. Hotspot detection and clustering: ways and means. *Environmental and Ecological Statistics*, 17:231–245, 2010.

Feng Liu, David J. Mladenoff, Nicholas S. Keuler, and Lisa Schulte Moore. Broadscale variability in tree data of the historical Public Land Survey and its consequences for ecological studies. *Ecological Monographs*, 81:259–275, 2011.

Jun S. Liu. *Monte Carlo Strategies in Scientific Computing*. Springer, New York, 2008.

Jun S. Liu, Wing Hung Wong, and Augustine Kong. Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika*, 81:27–40, 1994.

Ian W. McKeague and Wolfgang Wefelmeyer. Markov chain Monte Carlo and Rao–Blackwellization. *Journal of Statistical Planning and Inference*, 85:171 – 182, 2000.

Sean Meyn. *Control Techniques for Complex Networks*. Cambridge University Press, Cambridge, 2008.

Sean Meyn and Richard L. Tweedie. *Markov Chains and Stochastic Stability*. Cambridge University Press, Cambridge, second edition, 2009.

Aleksandar Mijatović, Jure Vogrinc, et al. On the poisson equation for metropolis–hastings chains. *Bernoulli*, 24(3):2401–2428, 2018.

David J. Mladenoff, T.A. Sickley, Lisa A. Schulte, Jeanine M. Rhemtulla, and J. Bolliger. Wisconsin's land cover in the 1800s. *Wisconsin DNR publication*.

David J. Mladenoff, Mark A. White, John Pastor, and Thomas R. Crow. Comparing spatial pattern in unaltered old-growth and disturbed forest landscapes. *Ecological Applications*, 3:294–306, 1993.

Radford M. Neal. Probabilistic inference using Markov chain Monte Carlo methods. Technical report, 1993.

Christopher J. Paciorek, Simon J. Goring, Andrew L. Thurman, Charles V. Cogbill, John W. Williams, David J. Mladenoff, Jody A. Peters, Jun Zhu, and Jason S. McLachlan. Statistically-estimated tree composition for the northeastern United States at Euro-American settlement. *PLOS ONE*, 11:1–20, 2016.

David W. Peterson and Peter B. Reich. Prescribed fire in oak savanna: fire frequency effects on stand structure and dynamics. *Ecological Applications*, 11:914–927, 2001.

Volker C. Radeloff, David J. Mladenoff, Hong S. He, and Mark S. Boyce. Forest landscape change in the northwestern Wisconsin Pine Barrens from pre-European settlement to the present. *Canadian Journal of Forest Research*, 29:1649–1659, 1999.

C.R. Rao. Information and accuracy attainable in the estimation of statistical parameters. *Bulletin of the Calcutta Mathematical Society*, 37:81–91, 1945.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *Ann. Math. Statist.*, 22:400–407, 1951.

Christian P. Robert and George Casella. *Monte Carlo Statistical Methods*. Springer-Verlag, New York, second edition, 2004.

Gareth Roberts and Jeffrey Rosenthal. General state space Markov chains and MCMC algorithms. *Probability Surveys*, 1:20–71, 2004.

L.A. Schulte and D.J. Mladenoff. The original U.S. public land survey records: their use and limitations in reconstructing pre-European settlement vegetation. *Journal of Forestry*, 99:5–10, 2001.

Lisa A. Schulte, David J. Mladenoff, and Erik V. Nordheim. Quantitative classification of a historic northern Wisconsin (U.S.A.) landscape: mapping forests at regional scales. *Canadian Journal of Forest Reasearch*, 32:1616–1638, 2002.

Jun Shao. *Mathematical Statistics*. Springer Texts in Statistics. Springer-Verlag, New York, second edition, 2003.

Monika E. Shea, Lisa A. Schulte, and Brian J. Palik. Reconstructing vegetation past: pre-Euro-American vegetation for the midwest Driftless area, USA. *Ecological Restoration*, 32:417–433, 2014.

Nicolas Städler, Peter Bühlmann, and Sara Van De Geer. l1-penalization for mixture regression models. *Test*, 19:209–256, 2010.

Michael C. Stambaugh and Richard P. Guyette. Predicting spatio-temporal variability in fire return intervals using a topographic roughness index. *Forest Ecology and Management*, 254:463–473, 2008.

Lance A. Waller. Detection of clustering in spatial data. In Jan Fagerberg, David C. Mowery, and Richard R. Nelson, editors, *The SAGE Handbook of Spatial Analysis*, chapter 10, pages 299–321. Sage London, 2009.

C. F. Jeff Wu. On the convergence properties of the EM algorithm. *The Annals of Statistics*, 11:95–103, 1983.

F. Y. Wu. The Potts model. *Reviews of Modern Physics*, 54:235–268, 1982.

Laurent Younes. Estimation and annealing for Gibbsian fields. *Annales de l'Institut Henri Poincaré*, 24:269–294, 1988.

Laurent Younes. Parametric inference for imperfectly observed Gibbsian fields. *Probability Theory and Related Fields*, 82:625–645, 1989.