UNIVERSITY OF CALIFORNIA

Los Angeles

Large-scale Inference of Correlation between Complex Biological Traits

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Biostatistics

by

Zhenyu Zhang

2022

ABSTRACT OF THE DISSERTATION

Large-scale Inference of Correlation between Complex Biological Traits

by

Zhenyu Zhang
Doctor of Philosophy in Biostatistics
University of California, Los Angeles, 2022
Professor Marc A. Suchard, Chair

Inferring dependencies between complex biological traits while accounting for evolutionary relationships among specimens is of great scientific interest, yet remains infeasible when trait and specimen counts grow large. I aim to develop a scalable Bayesian inference framework to assess correlation between complex traits along the evolutionary tree relating the specimens and informed by molecular sequences. To accommodate discrete and continuous traits, I posit a phylogenetic multivariate probit model that uses a latent variable framework. Posterior computation under this model requires integrating many latent variables, or equivalently making many computationally expensive draws from a high-dimensional multivariate truncated normal distribution (MTN). To tackle this challenge, I propose an inference scheme that exploits 1) representative cutting-edge Markov chain Monte Carlo (MCMC) methods including the bouncy particle sampler (BPS), the Markovian Zigzag sampler (ZZ), and the Zigzag Hamiltonian Monte Carlo (Zigzag-HMC) that can simultaneously sample all truncated normal dimensions, and 2) novel dynamic programming strategies that reduce the cost of likelihood and gradient evaluations for all three samplers to linear in sample size. Compared to the previous best practices that employ multiple-try rejection sampling, my

approach achieves an order-of-magnitude speedup, allowing us to tackle previously unworkable large-scale problems. In an application with 535 HIV-1 viruses and 24 traits that necessitates sampling from a 11,235-dimensional MTN, my method makes it possible to examine the conditional dependencies between 21 immune escape mutations and 3 virulence measurements. In a second application I study the evolution of influenza H1N1 glycosylations using around 900 viruses. Lastly, I extend the phylogenetic probit model to incorporate categorical traits and demonstrate its use to investigate *Aquilegia* flower and pollinator coevolution. In summary, the contribution of this dissertation is two-fold. First, I develop a state-of-the-art solution for the long-standing problem in Bayesian phylogenetics — learning correlation among complex biological traits with joint tree modeling. Second, further empirical and theoretical investigation of BPS, ZZ, and Zigzag-HMC yield insight into the differences and similarities between these recently developed MCMC samplers. As Zigzag-HMC outperforms the other two on MTNs, I also implement this approach in a standalone R package, aiming to provide a general efficient tool for high-dimensional MTN simulation.

The dissertation of Zhenyu Zhang is approved.

Hua Zhou

Sudipto Banerjee

Rita M. Cantor

Marc A. Suchard, Committee Chair

University of California, Los Angeles

2022

To my parents

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

Four chapters of this dissertation resulted in or are anticipated to result in co-authored publications. Chapter 3 is a version of Zhenyu Zhang, Akihiko Nishimura, Paul Bastide, Xiang Ji, Rebecca P Payne, Philip Goulder, Philippe Lemey, and Marc A Suchard. Large-scale inference of correlation among mixed-type biological traits with phylogenetic multivariate probit models. *The Annals of Applied Statistics* (2021) 15.1, pp. 230–251 (DOI: 10.1214/20-AOAS1394). Chapter 4 is a version of Zhenyu Zhang, Akihiko Nishimura, Trovão Nídia, Joshua L. Cherry, Andrew J Holbrook, Xiang Ji, Philippe Lemey, and Marc A. Suchard. Accelerating Bayesian inference of dependency between complex biological traits. *Under review.* Chapter 5 contains modified text and figures from Akihiko Nishimura, Zhenyu Zhang, and Marc A. Suchard. Hamiltonian zigzag sampler got more momentum than its Markovian counterpart: Equivalence of two zigzags under a momentum refreshment limit. *Under review.* Chapter 6 is a version of Zhenyu Zhang, Akihiko Nishimura, Andrew Chin, and Marc A. Suchard. hdtg: An R package for high-dimensional truncated normal simulation. *Under review.*

First I want to thank my adviser Marc Suchard, who has, in every possible way, exceeded my previous idea of what makes a great mentor. Marc supports me in all aspects of an academic life, whether it is scientific writing, giving conference talks, or research collaboration. He advises, inspires, and helps me to realize my potentials. It is only when looking back that I realize how he has been dedicatedly helping me to achieve my personal and professional goals at each stage of my PhD. With his guidance I become a better learner, thinker, writer, and speaker. I was given full trust from day one, but whenever I need help, Marc is there. When I first started coding in BEAST Marc would spend two hours looking through the code with me — just one of many examples of his mentoring efforts. Marc has the best interests of his mentees at heart all the time and genuinely cares about our success and well-being.

Besides guiding me through the difficulties in research, Marc makes sure I get the resources and support I need. He makes every effort to help me build my professional networks and regularly sends me to conferences. My PhD journey has not always been smooth, but Marc has been a supportive and caring mentor throughout. He understands studying in a foreign country is not easy and encourages me to go home in winter and summer breaks. During the COVID-19 pandemic when I was living in isolation, at the end of our weekly meeting Marc always asked if I need any help from him, which is so heartwarming to hear. Having Marc as my PhD adviser is one of the best things ever happened to me. I admire Marc for his profound knowledge, humor, wisdom, but mostly, for his heart of gold.

I am grateful to my committee members, Hua Zhou, Sudipto Banerjee, and Rita Cantor for giving me so much of their valuable time and expert advice. They provide new perspectives and fresh ideas I would never think of on my own, and inspire me to continuously improve my work. What I learned from Hua's scentific computing course and Sudipto's Bayesian inference course are fundemental to this dissertation. I also thank Janet Sinsheimer for her thoughtful questions and her kind encouragement when I started my PhD.

I thank Akihiko Nishimura who is like my second mentor. It is Aki who first brought my attention to the cutting-edge MCMC methods which become the key algorithms in this dissertation. Aki has answered endless questions from me during his postdoc and after he started his faculty position. None of my works is possible without his advice on the relevant mathematics and efficient implementation. Aki is also a trustworthy and warm friend who would hide a chocolate bar under my keyboard when I faced difficulties in life.

I thank everyone in Marc's group with whom my PhD journey is much more enjoyable. I thank Gabe Hassler and Alex Fisher for taking the phylogenetics adventure with me in the past five years. I thank Xiang Ji and Andrew Holbrook for sharing their rich knowledge and helping me editing manuscripts. I thank Max Tolkoff who encouraged me a lot when I first joined the group. I thank Jianxiao Yang for the soothing conversations we have. I also thank Yeesuk Kim and Paul Bastide for bringing joy when they visited us. After over a year

working from home, I am super grateful to get to know our lovely new members Fan Bu, Andy Magee, Kathik G, Faaizah Arshad, Kelly Li, and Yucai Shao. I know I will miss this team terribly. I also want to thank my awesome collaborators outside UCLA. In particular to Philippe Lemey, for his great input to the phylogenetic applications in this dissertation and exceeding kindness to me throughout my PhD.

I am grateful for the financial support I received, which allowed me to concentrate on my research without worrying about money. I was supported by research funds from Marc, as well as UCLA university fellowship and dissertation year fellowship. I am thankful for the biostatistics department's efforts to help students secure funding.

I thank professors in the biostatistics department for their well-taught courses from which I obtained vast amount of knowledge and skills in statistics. I particularly thank Robert Weiss for recommending me to Marc back in 2017. I thank professor Cumberland for his considerate advice when I entered graduate school as a master's student. Additionally, I thank Roxy Naranjo, our student affairs officer, for patiently helping with all kinds of logistics during my six years at UCLA.

I was fortunate to meet many trusty friends at UCLA. Thank you Leiwen, Wenxi, Rui, Xinya, Qi, Xinyu for the fun experiences in LA. Thank you Jane for being my great roommate and friend. Thank you Nan for being by my side during and after your days at UCLA. Thank you Shuang for being so cheerful and thoughtful. Thank you Sam for our amazing walks with Henry the dog. Thank you Sarah for always keeping me in mind. Thank you Ian for the relaxing conversations during the last phase of our PhD. Thank you Minyan for sharing our first year in the US. Thank you Zizhao for the fun badminton time. I particularly thank my loyal friends Di and Liu who brought me so much comfort and delight during the past three years. I cannot imagine doing my PhD without you.

I thank my old friends who always check on me when we are thousands of miles away. Thank you Xiaotong for being my friend since kindergarten. My happy childhood with you is the foundation for many years ahead. Thank you Yajie for sharing with me our journey

from teenage years to now and beyond. Thank you Yiling, Yazhi, Xinyi for our chat about everything studying abroad.

I would like to thank Ram, who was super supportive and taught me essential statistical consulting skills during my first working experience at UCLA. I thank Ole who kindly hosted me as a visiting student. That was a wonderful summer in Jutland. I thank Yang, Ruoqing, Kristy, Kevin at Amyris for making my internship enjoyable and fruitful. I thank Buzz for his greatest guitar teaching.

I thank my partner Hefei for understanding me the way nobody else can. Thank you for supporting my wildest dreams, and how lucky I am to explore this incredible world with you! I am indebted to everyone in my family for their wholehearted support. I cannot wait to get back and see you all. Finally, with my deepest gratitude, I dedicate this dissertation to my parents who love me unconditionally. Thank you mom and dad. You have been my source of power and determination since the very first day I came to the world.

VITA

2012–2016	B.E. Biomedical Engineering, Peking University, Beijing, China
2016–2018	M.S. Biostatistics, University of California, Los Angeles
2016–2017	Student consultant, Department of Medicine Statistics Core University of California, Los Angeles
2019 summer	Graduate Summer Research Mentorship Program University of California, Los Angeles
2018–2021	University Fellowship, University of California, Los Angeles
2018–present	Graduate student researcher, Department of Human Genetics University of California, Los Angeles
2021–present	Dissertation Year Fellowship, University of California, Los Angeles

PUBLICATIONS

Accepted peer-reviewed journals

Zhenyu Zhang, Akihiko Nishimura, Paul Bastide, Xiang Ji, Rebecca P Payne, Philip Goulder, Philippe Lemey, and Marc A Suchard (2021). "Large-scale inference of correlation among mixed-type biological traits with phylogenetic multivariate probit models". In: *The Annals of Applied Statistics* 15.1, pp. 230–251

Xiang Ji, Zhenyu Zhang, Andrew Holbrook, Akihiko Nishimura, Guy Baele, Andrew Rambaut, Philippe Lemey, and Marc A Suchard (2020). "Gradients do grow on trees: a linear-time O(N)-dimensional gradient for statistical phylogenetics". In: *Molecular biology and evolution* 37.10, pp. 3047–3060

Alexander A Fisher, Xiang Ji, Zhenyu Zhang, Philippe Lemey, and Marc A Suchard (2021a). "Relaxed random walks at scale". In: *Systematic Biology* 70.2, pp. 258–267

Gabriel W Hassler, Andrew Magee, Zhenyu Zhang, Guy Baele, Philippe Lemey, Xiang Ji, Mathieu Fourment, and Marc A Suchard (2023). "Data Integration in Bayesian Phylogenetics". In: *Annual Review of Statistics and Its Application*

Preprints under review

Zhenyu Zhang, Akihiko Nishimura, Nídia S Trovão, Joshua L Cherry, Andrew J Holbrook, Xiang Ji, Philippe Lemey, and Marc A Suchard (2022a). "Hamiltonian zigzag accelerates large-scale inference for conditional dependencies between complex biological traits". In: *arXiv preprint arXiv:2201.07291*

Akihiko Nishimura, Zhenyu Zhang, and Marc A Suchard (2021). "Hamiltonian zigzag sampler got more momentum than its Markovian counterpart: Equivalence of two zigzags under a momentum refreshment limit". In: *arXiv preprint arXiv:2104.07694*

Zhenyu Zhang, Andrew Chin, Akihiko Nishimura, and Marc A Suchard (2022b). "hdtg: An R package for high-dimensional truncated normal simulation". In: *under review*

CHAPTER 1

Introduction

1.1 Literature review and study objectives

Phylogenetics is the study of the evolutionary history relating individuals or groups of organisms and plays a central role in understanding the process of evolution and the interpretation of biological information. In the 1960s, several pioneer groups launched the inference of phylogenetic trees, or phylogenies, from molecular sequences. Cavalli-Sforza and Edwards introduced the parsimony and likelihood methods for inferring phylogenies (Cavalli-Sforza and Edwards, 1967; Edwards and Cavalli-Sforza, 1965). Fitch and Margoliash (1967) developed the first distance matrix method for tree construction. In the past decades, thanks to the explosively accumulating molecular data, greatly improved computing power and sophisticated statistical methods, phylogenetics continues to experience significant growth and has found use in nearly all branches in biology (Lyubetsky, Piel, and Quandt, 2014). Modern phylogenetics has two major goals: to reconstruct the evolutionary tree among species and to investigate the mechanisms of the evolutionary process giving rise to these species (Yang, 2006). A variety of tree-reconstruction methods exist, and among the most popular ones are the unweighted pair-group method using arithmetic averages (Sokal and Sneath, 1963, UPGMA), maximum parsimony (Fitch, 1971), maximum likelihood (Felsenstein, 1981) and Bayesian methods (Li, Pearl, and Doss, 2000; Rannala and Yang, 1996; Sinsheimer, Lake, and Little, 1996). Bayesian inference is a flexible and versatile tool in phylogenetics, as it can incorporate various modeling assumptions and at the same time provide a treatment for uncertainty. This is especially important when we are interested in some evolutionary pro-

cess happening on the tree, but high uncertainty lingers about the tree itself. Under a joint Bayesian framework we can average over the tree space, test evolutionary hypotheses without conditioning on a single tree and therefore reduce bias (Suchard, Weiss, and Sinsheimer, 2001).

A tree structure is an intuitive device to describe an evolutionary history of molecular sequences that are inherited in a vertical fashion from parent to progeny. With increasing sample sizes and model complexity, however, posterior inference for highly structured tree-based models tends to be computationally expensive. Therefore, much recent effort in Bayesian phylogenetics has focused on developing efficient inference frameworks. To name a few, Pybus et al. (2012) and Cybis et al. (2015) propose tree "traversals" to integrate out the internal and root node trait values of biological phenotypes analytically, Hassler et al. (2021) and Tolkoff et al. (2018) develop an efficient phylogenetic factor analysis approach for learning correlations among high-dimensional traits, Ji et al. (2020, 2021) and Fisher et al. (2021a,b) use scalable Hamiltonian Monte Carlo on general sequence substitution and trait evolution models, and Hassler et al. (2020) propose a fast method to marginalize missing traits. Of equal importance to the statistical development stand scientific software that provide efficient implementations of these methods as well as user-friendly interfaces. There exist a variety of packages specialized in Bayesian phylogenetic inference, such as the Bayesian Evolutionary Analysis by Sampling Trees (Suchard et al., 2018, BEAST), MrBayes (Ronquist et al., 2012), and RevBayes (Höhna et al., 2016).

The main goal of this dissertation is to develop a scalable inference framework to learn correlation between complex biological traits observed on evolutionarily related taxa, while simultaneously estimating the potentially unknown phylogeny. Here "complex" means that the trait values can be continuous or discrete — a situation commonly seen in applications. This is an important yet unsolved problem. To describe complex trait evolution I develop a phylogenetic probit model based on the latent liability model (Cybis et al., 2015) for its great utility and flexibility. Then I develop an inference toolbox for posterior computation under

this model and implement all developed methods in BEAST. Alternative approaches for complex traits on unknown trees are limited. Phylogenetic regression models (Grafen, 1989) assume a known fixed tree and their logistic extensions (Ives and Garland, 2009) take a single binary trait as the regression outcome. On the other hand, for continuous traits, comparative methods (Felsenstein, 1985) scale well on random trees (Pybus et al., 2012; Tung Ho and Ané, 2014). Likewise, continuous-time Markov chain-based models (Lewis, 2001; Pagel, 1994) are popular for multiple binary traits, but restrictively assume independence between traits given the tree.

Although my methodological development is driven by real-world phylogenetics applications (Chapter 3 and 4), I recognize that this development also delivers advances in statistical computing beyond phylogenetics. So a side goal of my dissertation is to further explore this broader contribution (Chapter 5 and 6).

1.2 Dissertation structure

I organize my dissertation around four projects in Chapters 3, 4, 5, 6 that together support the thesis objectives yet all lead to individual manuscripts for publication. The project chapters can be read as such and, therefore, please forgive the slight inconsistency in notation and repetition of material between chapters. Placed before these independent chapters sits Chapter 2 that provides necessary background in phylogenetics and statistics to help to situate the projects.

To describe the evolution of mixed-type traits, Chapter 3 introduces a phylogenetic multivariate probit model by assuming latent parameters for binary outcome dimensions at the tips of an unknown tree informed by molecular sequences. The focus of Chapter 3 is on the inference challenge under the phylogenetic probit model. When fitting this model to a large data set, the computational bottleneck is to repeatedly sample the latent variables from a high-dimensional multivariate truncated normal (MTN) with a possibly random cor-

relation structure. For this task, I develop a new inference approach that exploits 1) the bouncy particle sampler (Bouchard-Côté, Vollmer, and Doucet, 2018, BPS) based on piecewise deterministic Markov processes (PDMP) to simultaneously sample all truncated normal dimensions, and 2) novel dynamic programming that reduces the cost of likelihood and gradient evaluations for BPS to linear in sample size. Compared to the previous best practices (Cybis et al., 2015), BPS achieves a 74× speed-up in terms of minimal effective sample size (ESS) on a 11,235-dimensional MTN from an HIV data set with 535 taxa and 24 mixed-type traits, making it possible to estimate the across-trait correlation with an explicit tree modeling.

In Chapter 4, I develop a yet more efficient inference scheme for the phylogenetic probit model to tackle larger problems. As the number of specimens grows, the BPS method fails to reliably characterize conditional dependencies between traits. To resolve this limitation, I employ the recent Zigzag Hamiltonian Monte Carlo (Nishimura, Dunson, and Lu, 2020, Zigzag-HMC) that can utilize the same linear-order gradient evaluation method developed in Chapter 3. I demonstrate that Zigzag-HMC better explores the parameter space than BPS and so it samples from a high-dimensional MTN more efficiently. Zigzag-HMC also enjoys another strong advantage in that it allows joint transition kernels for highly correlated latent variables and correlation matrix elements. On the same HIV application as in Chapter 3, Zigzag-HMC yields a further 5-fold speedup compared to BPS and makes it possible to learn partial correlations between candidate viral mutations and virulence. I demonstrate the improved scalability of this approach in studying the evolution of influenza H1N1 glycosylations on around 900 viruses. For broader applicability, I extend the phylogenetic probit model to incorporate categorical traits, and demonstrate its use to study *Aquilegia* flower and pollinator co-evolution.

I test another cutting edge PDMP-based sampler, the Markovian zigzag sampler (Bierkens, Fearnhead, Roberts, et al., 2019, ZZ) on MTNs from the phylogenetic probit model to see if it is a better choice than BPS or Zigzag-HMC. Zigzag-HMC turns out to be more efficient

and therefore becomes the final choice in Chapter 3. In Chapter 5 my collaborator Akihiko Nishimura and I recognize an intriguing connection between the two zigzag samplers that helps to explain their different behavior. Nishimura, Zhang, and Suchard (2021) theoretically proves an equivalence of the two methods under certain limits. This result is consistent with our observation that Zigzag-HMC outperforms ZZ when the dependency among parameters increases.

The impressive efficiency of Zigzag-HMC on MTNs in the phylogenetic applications suggests its broader use. Therefore, in Chapter 6 I introduce the `hdtg` R package for efficient MTN simulations. MTN simulation arises in various statistical applications yet remains a challenging problem in high dimensions. There is no available software package that can generate samples from an arbitrary MTN with thousands of dimensions and this limitation prevents researchers from developing methods that exploit sampling from a large MTN. Besides Zigzag-HMC, the current best algorithms for MTN simulation are the minimax tilting accept-reject sampler (Botev, 2017, MT) and the harmonic Hamiltonian Monte Carlo (Pakman and Paninski, 2014, Harmonic-HMC). However, the scale limit of minimax tilting is often only a few hundred dimensions, and there is no efficient implementation of Harmonic-HMC. I aim to bridge this gap with the `hdtg` package that implements both Zigzag-HMC and Harmonic-HMC. I compare the efficiency between Zigzag-HMC, Harmonic-HMC, and MET on MTNs with various correlation structures. Zigzag-HMC outperforms the other two in most cases with a dimension > 1000. I also provide some practical guidance on method choice for MTN simulation and illustrate the usage of `hdtg` functions.

Finally, in Chapter 7 I discuss what new knowledge and insights my doctoral projects have contributed, and point out future directions in studying complex trait evolution.

CHAPTER 2

Background

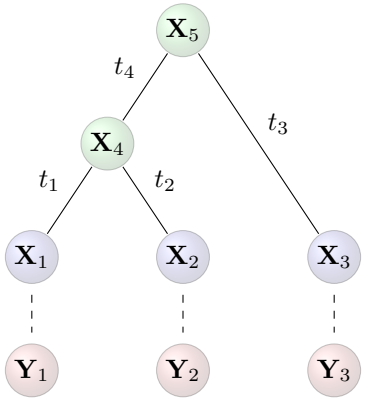
2.1 Trait evolution framework

Here I introduce the basic trait evolution framework used throughout Chapters 3 and 4. The phylogeny $\mathscr{F} = (\mathbb{V}, \mathbf{t})$ is a directed, bifurcating acyclic graph with a set of nodes $\mathbb{V}$ and branch lengths $\mathbf{t}$. The node set $\mathbb{V}$ contains N tip nodes ("taxa") and $N-1$ internal nodes including the root. The branch lengths $\mathbf{t} = (t_1, \ldots, t_{2N-2})$ denote the time distance from every node to its single parent (except the root). On the ith taxon ($i = 1, \ldots, N$), we observe P continuous or discrete traits $\mathbf{Y}_i$. To jointly model the evolution of these complex traits, I build a phylogenetic probit model that extends the popular threshold model for binary traits (Cybis et al., 2015; Felsenstein, 2005, 2011). This model assumes $2N-1$ latent variables $\mathbf{X}_i$ for all nodes and those at tree tips decide the observed trait values. The latent variables themselves follow a Brownian diffusion along the tree (Felsenstein, 1985). I then specify a mapping function from $\mathbf{X}_i$ to $\mathbf{Y}_i$ that accommodates continuous, binary, and categorical trait types (Section 3.2.1 and 4.2.1). Figure 2.1 visualizes such a trait evolution framework on a 3-tip example tree.

2.2 Bayesian inference and Markov chain Monte Carlo

I take a Bayesian approach for every inference task in this dissertation. Bayesian inference derives the posterior distribution of model parameters $\boldsymbol{\theta}$ given all observed data $\mathbf{Y}$ according

Figure 2.1: An example tree with $N = 3$ tips (purple) and 2 internal nodes (green). I denote the latent variable and observed trait on the ith node with $\mathbf{X}_i$ and $\mathbf{Y}_i$, respectively, and t_i is the branch length from node i to its parent node.



to Bayes' theorem

$$p(\boldsymbol{\theta} \mid \mathbf{Y}) = \frac{p(\mathbf{Y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})}{p(\mathbf{Y})}. \tag{2.1}$$

Note that $\boldsymbol{\theta}$ includes all unknown model parameters. In phylogenetics applications, these parameters define the tree topology and branch lengths, as well as any evolutionary process described by the model. Data $\mathbf{Y}$ may contain sequence data, phenotypic trait data, and relevant geographic information. In Chapter 3 and 4, I follow a 5-step process to perform Bayesian analysis:

1. Identify the observed data $\mathbf{Y}$ and scientific questions.
2. Build a probabilistic model that gives the likelihood $p(\mathbf{Y} \mid \boldsymbol{\theta})$.
3. Specify prior distributions $p(\boldsymbol{\theta})$.
4. Learn the posterior $p(\boldsymbol{\theta} \mid \mathbf{Y})$.
5. Interpret the results.

Except for special cases where priors and likelihoods are conjugate, an analytical posterior is not available, and it requires either a sampling- or approximation-based approach

to characterize $p(\boldsymbol{\theta} \mid \mathbf{Y})$. Markov chain Monte Carlo (MCMC) is one of the most widely used sampling approaches for Bayesian inference, and the methodology development in this dissertation focuses on novel and efficient MCMC methods to tackle inference challenges in studying complex trait evolution.

In short, MCMC algorithms construct a Markov chain whose stationary distribution is the posterior distribution of interest. There are various ways to construct such a Markov chain, including the classical Metropolis-Hastings algorithm (Metropolis et al., 1953) and Gibbs sampling (Geman and Geman, 1984), as well as the more recent Hamiltonian Monte Carlo (Neal, 2011, HMC) and MCMC based on piecewise deterministic Markov processes (Davis, 1984; Fearnhead et al., 2018, PDMPs). HMC- and PDMP-based MCMC avoid the "random-walk behavior" that hampers many MCMC methods by utilizing gradient information to better explore the parameter space. However, none of the current MCMC methods is readily applicable to the bottleneck of sampling the conditional posterior of latent variables under the phylogenetic probit model, and this limitation motivates me to develop the inference machinery in Chapter 3 and 4. As HMC- and PDMP-based MCMC are the key algorithms in my development, I give a brief overview of them in the next sections.

2.3 Hamiltonian Monte Carlo

Hamiltonian Monte Carlo (Neal, 2011, HMC) is a state-of-the-art general purpose sampler. HMC only requires evaluations of the log-density and its gradient, yet is capable of sampling efficiently from complex high-dimensional distributions (Gelman et al., 2013). In order to sample a d-dimensional parameter $\boldsymbol{x} = (x_1, \ldots, x_d)$ from the target distribution $\pi(\boldsymbol{x})$, HMC introduces an auxiliary *momentum* variable $\boldsymbol{p} = (p_1, \ldots, p_d) \in \mathbb{R}^d$ and samples from the product density $\pi(\boldsymbol{x}, \boldsymbol{p}) = \pi(\boldsymbol{x})\pi(\boldsymbol{p})$ by numerically discretizing the Hamiltonian dynamics

$$\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t} = \nabla K(\boldsymbol{p}), \quad \frac{\mathrm{d}\boldsymbol{p}}{\mathrm{d}t} = -\nabla U(\boldsymbol{x}), \tag{2.2}$$

where $U(\boldsymbol{x}) = -\log \pi(\boldsymbol{x})$ and $K(\boldsymbol{p}) = -\log \pi(\boldsymbol{p})$ are the *potential* and *kinetic energy*. The sum of $U(\boldsymbol{x})$ and $K(\boldsymbol{p})$ is the *Hamiltonian* which stays invariant over time. In each HMC iteration, we first draw $\boldsymbol{p}$ from its marginal distribution $\pi(\boldsymbol{p}) \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, a standard Gaussian and then approximate (2.2) from time $t = 0$ to $t = T$ by $L = \lfloor T/\epsilon \rfloor$ steps of the *leapfrog* update with stepsize ϵ (Leimkuhler and Reich, 2004):

$$\boldsymbol{p} \leftarrow \boldsymbol{p} + \frac{\epsilon}{2}\nabla_{\boldsymbol{x}} \log \pi(\boldsymbol{x}), \quad \boldsymbol{x} \leftarrow \boldsymbol{x} + \epsilon \boldsymbol{p}, \quad \boldsymbol{p} \leftarrow \boldsymbol{p} + \frac{\epsilon}{2}\nabla_{\boldsymbol{x}} \log \pi(\boldsymbol{x}). \tag{2.3}$$

The end state is a valid *Metropolis* proposal that one accepts or rejects according to the standard acceptance probability formula (Hastings, 1970; Metropolis et al., 1953).

By virtue of the properties of Hamiltonian dynamics, the HMC proposals generated above can be far away from the current state yet be accepted with high probability. Good performance of HMC depends critically on well-calibrated choices of L and ϵ. In this dissertation I automate these choices via the stochastic optimization approach of Andrieu and Thoms (2008) and the *No-U-Turn* (NUTS) algorithm of Hoffman and Gelman (2014) that have been shown to achieve performance competitive with manually optimized HMC. The key algorithm in Chapter 4, Zigzag-HMC, is a less explored version of HMC where the momentum components have independent Laplace distributions and the resulting Hamiltonian trajectory possesses a zigzag shape. Since Zigzag-HMC is based on the reversible *Hamiltonian zigzag dynamics* (Nishimura, Dunson, and Lu, 2020; Nishimura, Zhang, and Suchard, 2021), it can also utilize NUTS to avoid manual tuning.

2.4 Piecewise deterministic Monte Carlo

The recent development of new MCMC algorithms based on piecewise deterministic Markov processes (PDMPs) has attracted an explosion of interest (Dunson and Johndrow, 2020). Examples include the bouncy particle sampler (Bouchard-Côté, Vollmer, and Doucet, 2018, BPS), the Zig-Zag sampler (Bierkens and Duncan, 2017; Bierkens, Fearnhead, Roberts,

et al., 2019, ZZ), and the Boomerang sampler (Bierkens et al., 2020). Unlike traditional MCMC methods that simulate a discrete-time Markov chain, PDMP-based MCMC simulates a continuous-time Markov process designed to have the target distribution as its stationary distribution. These conceptually new methods have demonstrated promising efficiency in large-scale problems and my work in Chapter 3 to 5 are among the first examples where they contribute to tackle challenging inference tasks in real-world applications (Nishimura, Zhang, and Suchard, 2021; Zhang et al., 2021, 2022a).

Following Fearnhead et al. (2018), I briefly introduce PDMPs and their use in building continuous-time MCMC algorithms. A PDMP is a continuous-time stochastic process (Davis, 1984). One can visualize a d-dimensional PDMP in terms of an imaginary particle moving in the $\mathbb{R}^d$ space, whose state at time t is $(\boldsymbol{x}_t, \boldsymbol{v}_t)$, where $\boldsymbol{x}_t = \left(x_t^1, \ldots, x_t^d\right)$ denotes the particle's position at time t and $\boldsymbol{v}_t = \left(v_t^1, \ldots, v_t^d\right)$ is the velocity. The PDMP trajectory consists of random velocity changing events and deterministic dynamics between the events. The following three quantities define a PDMP:

(i) *The deterministic dynamics.* Between events, the trajectory follows a deterministic path described by the ordinary differential equation:

$$\frac{\mathrm{d}x_t^i}{\mathrm{d}t} = v_t^i, \quad i = 1, \ldots, d. \tag{2.4}$$

(ii) *The event rate.* Velocity changing events will happen at a rate $\lambda\left(\boldsymbol{x}_t, \boldsymbol{v}_t\right)$ that depends on the current state $(\boldsymbol{x}_t, \boldsymbol{v}_t)$ and the target distribution. In other words, the probability that an event happens during interval $[t, t+h]$ is $\lambda\left(\boldsymbol{x}_t\right) h + o(h)$.

(iii) *The transition at events.* The velocity will change at each event. If an event happens at time τ, we have $\boldsymbol{v}_\tau = q\left(\boldsymbol{x}_{\tau-}, \boldsymbol{v}_{\tau-}\right)$, where $\boldsymbol{v}_{\tau-}$ and $\boldsymbol{x}_{\tau-}$ are the position and velocity immediately before the event and the transition kernel $q(\cdot, \cdot)$ specifies how they determine $\boldsymbol{v}_\tau$.

A PDMP-based MCMC algorithm first specifies these characteristic quantities. Then one can simulate the underlying PDMP via:

(1) Given the current state $(\boldsymbol{x}_t, \boldsymbol{v}_t)$, simulate the next event time τ.

(2) Compute $(\boldsymbol{x}_{\tau-}, \boldsymbol{v}_{\tau-})$, the state right before the event.

(3) Update the state at time τ so $\boldsymbol{v}_\tau \leftarrow q\left(\boldsymbol{x}_{\tau-}, \boldsymbol{v}_{\tau-}\right)$ and $\boldsymbol{x}_\tau \leftarrow \boldsymbol{x}_{\tau-}$.

(4) Go back to Step 1 with the current state being $(\boldsymbol{x}_\tau, \boldsymbol{v}_\tau)$.

Consequently, one MCMC iteration involves simulating the process for an arbitrary total time duration T and the end position makes the MCMC sample. Some differences between BPS, ZZ and the Boomerang sampler include 1) the between-event velocity for BPS and ZZ stays constant while that for Boomerang is time-varying. 2) Besides the target-informed events in (ii), BPS and the Boomerang sampler also incorporate random velocity refreshment events to ensure ergodicity of the process, and ZZ does not require such velocity refreshments.

CHAPTER 3

Large-scale inference of correlation among mixed-type biological traits with phylogenetic multivariate probit models

3.1 Introduction

Phylogenetics stands as a key tool in assessing rapidly evolving pathogen diversity and its impact on human disease. Important taxonic examples include RNA viruses, such as influenza and human immunodeficiency virus (HIV). Pathogens sampled from infected individuals are implicitly correlated with each other through their shared evolutionary history, often described through a phylogenetic tree that one reconstructs by sequencing the pathogen genomes. Drawing inference about concerted changes within multiple measured pathogen and host traits along this history leads to highly structured models. These models must simultaneously entertain and adjust for the across-taxon correlation and the between-trait correlation that characterizes the trait evolutionary process, leading to high computational burden. This burden arises from the need to integrate over the unobserved trait process and possible uncertainty in the history. This burden grows more challenging as the sample size, both in terms of number of taxa N and number of traits P, increases and, especially, when traits are of mixed-type, including both continuous quantities and discrete outcomes. Here, even best current practices (Cybis et al., 2015) fail to provide reliable estimates for emerging biological problems due to high computational complexity.

To jointly model continuous and binary trait evolution along an unknown tree, we adopt and extend the popular phylogenetic threshold model for binary traits (Felsenstein, 2005, 2011) with a long tradition in statistical genetics (Wright, 1934). This model assumes that unobserved continuous latent parameters for each tip taxon in the tree determine the observed binary traits according to a threshold. The latent parameters themselves arise from a Brownian diffusion along the tree (Felsenstein, 1985). The correlation matrix of the diffusion process informs correlation between latent parameters that map to concerted changes between binary traits. Here one interprets the latent parameters as the combined effect of all relevant genetic factors that influence the binary traits after adjusting for the shared evolutionary history.

As in Cybis et al. (2015), we extend the threshold model to include continuous traits by treating them as directly observed dimensions of the latent parameters. We recognize an identifiability issue in Cybis et al. (2015) and address this limitation with specific constraints on the diffusion covariance. We arrive at a mixed-type generalization of the multivariate probit model (Chib and Greenberg, 1998) that allows us to jointly model continuous and binary traits. We call this the phylogenetic multivariate probit model. Similar strategies for mixed-type data that assume latent processes underlying discrete data are commonly employed in various domain fields, including the biological and ecological sciences (Clark et al., 2017; Irvine, Rodhouse, and Keren, 2016; Schliep and Hoeting, 2013), optimal design (Fedorov, Wu, and Zhang, 2012), and computer experiments (Pourmohamad, Lee, et al., 2016). The observed outcomes can also be conveniently clustered (Dunson, 2000; Murray et al., 2013). Likewise, our phylogenetic probit model is easily extendable to categorical and ordinal data (Cybis et al., 2015).

Bayesian inference for the phylogenetic multivariate probit model involves, however, repeatedly sampling latent parameters from an NP dimensional truncated normal distribution, with N being the number of taxa and P the number of traits. To attempt this, Cybis et al. (2015) use Markov chain Monte Carlo (MCMC) based on a multiple-try rejection sampler.

The sampler has a computational complexity of $\mathcal{O}(NP^2)$ to update P dimensions of the latent parameters for just one taxon within a Gibbs cycle. Hence, to touch all dimensions, the resulting cost is $\mathcal{O}(N^2P^2)$. Further, since only a small portion of the latent parameter dimensions are updated per rejection-sample, the resulting MCMC chain is highly auto-correlated, hurting efficiency.

To overcome this limitation, we develop a scalable approach to sample from the multivariate truncated normal by combining the recently developed bouncy particle sampler (Bouchard-Côté, Vollmer, and Doucet, 2018, BPS) and an extension of the dynamic programming strategy by Pybus et al. (2012). BPS samples from a target distribution by simulating a Markov process with a piecewise linear trajectory. The simulation generally requires solving a one-dimensional optimization problem within each line segment. When sampling from a truncated normal, however, this optimization problem can be solved via a single log-density gradient evaluation. In the phylogenetic multivariate probit model, a direct evaluation of this gradient requires $\mathcal{O}(N^2P + NP^2)$ computation. By extending the dynamic programming strategy of Pybus et al. (2012) for diffusion processes on trees, we reduce this computational cost to $\mathcal{O}(NP^2)$ — a major practical gain as $N \gg P$ in most applications. Compared to the current practice, our BPS sampler achieves superior mixing rate, allowing us to attack previously unworkable problems.

We apply this Bayesian inference framework to assess correlation between HIV-1 *gag* gene immune-escape mutations and viral virulence, the pathogen's capacity to cause disease. By adjusting for the unknown evolutionary history that confounds our epidemiologically collected data, we identify significant correlations that closely match with the biological experimental literature and increase our understanding of the underlying molecular mechanisms of HIV.

3.2 Modeling

3.2.1 Phylogenetic multivariate probit model for mixed-type traits

Consider N biological taxa, each with P trait measurements. These measurements partition as $\mathbf{Y} = \{y_{ij}\} = \left[\mathbf{Y^b}, \mathbf{Y^c}\right]$ with $\mathbf{Y^b}$ being an $N \times P_b$ matrix of P_b binary traits and $\mathbf{Y^c}$ an $N \times P_c$ matrix of P_c continuous traits, where $P = P_b + P_c$. We assume that $\mathbf{Y}$ arises from a partially observed multivariate Brownian diffusion process along a phylogenetic tree $\mathscr{F}$. The tree $\mathscr{F} = (\mathbb{V}, \mathbf{t})$ is a directed, bifurcating acyclic graph with a set of nodes $\mathbb{V}$ and branch lengths $\mathbf{t}$. The node set $\mathbb{V}$ contains N degree-1 tip nodes, $N-2$ internal nodes of degree 3, and one root node of degree 2. The branch lengths $\mathbf{t} = (t_1, \ldots, t_{2N-2})$ denote the distance in real time from each node to its parent (Figure 3.1, left). The tree $\mathscr{F}$ is either known or informed by molecular sequence alignment $\mathbf{S}$ (Suchard et al., 2018).

We associate each node i in $\mathscr{F}$ with a latent parameter $\mathbf{X}_i \in \mathbb{R}^P$ for $i = 1, \ldots, 2N-1$. A Brownian diffusion process characterizes the evolutionary relationship between latent parameters, such that $\mathbf{X}_i$ is multivariate normal (MVN) distributed,

$$\mathbf{X}_i \sim \mathcal{N}\left(\mathbf{X}_{\mathrm{pa}(i)}, t_i \mathbf{\Omega}\right), \tag{3.1}$$

centered at its parent node value $\mathbf{X}_{\mathrm{pa}(i)}$ with across-trait, per-unit-time, $P \times P$ variance matrix $\mathbf{\Omega}$ that is shared by all branches along $\mathscr{F}$.

At the tips of $\mathscr{F}$, we collect the $N \times P$ matrix $\mathbf{X} = \{x_{ij}\} = [\mathbf{X}_1, \ldots, \mathbf{X}_N]^T$ and map it to the observed traits through the function

$$y_{ij} = g(x_{ij}) = \begin{cases} \mathrm{sign}(x_{ij}), & j = 1, \ldots P_b, \\ x_{ij}, & j = P_b + 1, \ldots, P, \end{cases} \tag{3.2}$$

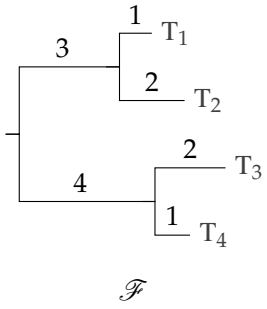
where $\mathrm{sign}(x_{ij})$ takes the value 1 on positive values and -1 on negative values. As a result, latent parameters at the tips and a threshold (that we set to zero without loss of generality) determine the corresponding binary traits, and continuous traits can be seen as directly

observed.

Turning our attention to the joint distribution of tip latent parameters $\mathbf{X}$, we can integrate out $\mathbf{X}_{N+1}, \ldots, \mathbf{X}_{2N-1}$ by assuming a conjugate prior on the tree root, $\mathbf{X}_{2N-1} \sim \mathcal{N}\left(\boldsymbol{\mu}_0, \tau_0^{-1}\boldsymbol{\Omega}\right)$ with prior mean $\boldsymbol{\mu}_0$ and prior sample size τ_0. Then $\mathbf{X}$ follows a matrix normal distribution

$$\mathbf{X} \sim \mathrm{MTN}_{NP}(\mathbf{M}, \boldsymbol{\Upsilon}, \boldsymbol{\Omega}), \tag{3.3}$$

where $\mathbf{M} = (\boldsymbol{\mu}_0, \ldots, \boldsymbol{\mu}_0)^T$ is an $N \times P$ mean matrix and the across-taxa tree covariance matrix $\boldsymbol{\Upsilon} = \mathbf{V}(\mathscr{T}) + \tau_0^{-1}\mathbf{J}$ (Pybus et al., 2012). The tree diffusion matrix $\mathbf{V}(\mathscr{T})$ is a deterministic function of $\mathscr{T}$ and $\mathbf{J}$ is an $N \times N$ matrix of all ones, such that the term $\tau_0^{-1}\mathbf{J}$ comes from the integrated-out tree root prior. Figure 3.1 illustrates how the tree structure determines $\mathbf{V}(\mathscr{T})$: the diagonals are equal to the sum of branch lengths from tip to root, and the off-diagonals are equal to the branch length from root to the most recent common ancestor of two tips. Combining equations (3.2) and (3.3) enables us to write down the



$\mathscr{T}$

$$\mathbf{V}(\mathscr{T}) = \begin{array}{c} \begin{array}{cccc} \mathrm{T}_1 & \mathrm{T}_2 & \mathrm{T}_3 & \mathrm{T}_4 \end{array} \\ \begin{pmatrix} 4 & 3 & 0 & 0 \\ 3 & 5 & 0 & 0 \\ 0 & 0 & 6 & 4 \\ 0 & 0 & 4 & 5 \end{pmatrix} \end{array} \begin{array}{l} \mathrm{T}_1 \\ \mathrm{T}_2 \\ \mathrm{T}_3 \\ \mathrm{T}_4 \end{array}$$

Figure 3.1: A 4-taxon phylogenetic tree $\mathscr{T}$ with tips (T_1, T_2, T_3, T_4) and their corresponding tree diffusion matrix $\mathbf{V}(\mathscr{T})$.

augmented likelihood of $\mathbf{X}$ and $\mathbf{Y}$ through the factorization

$$p(\mathbf{Y}, \mathbf{X} \mid \boldsymbol{\Upsilon}, \boldsymbol{\Omega}, \boldsymbol{\mu}_0, \tau_0, g) = p(\mathbf{Y} \mid \mathbf{X}) p(\mathbf{X} \mid \boldsymbol{\Upsilon}, \boldsymbol{\Omega}, \boldsymbol{\mu}_0, \tau_0), \tag{3.4}$$

where $p(\mathbf{Y} \mid \mathbf{X}) = \mathbb{I}(\mathbf{Y} \mid \mathbf{X}, g)$, the indicator function that takes the value 1 if $\mathbf{X}$ are consistent with the observations $\mathbf{Y}$ and 0 otherwise.

3.2.2 Decomposition of trait-covariance to account for varying data scales

The previous work of Cybis et al. (2015) uses a conjugate Wishart prior on $\mathbf{\Omega}^{-1}$ for computational convenience. However, there are two problems with the Wishart prior. First, with mixed-type data, it leaves the model not parameter-identifiable. For a binary trait, we only know the sign of its latent parameter; the absolute value is arbitrary. Consider a latent parameter x_{ij} and its marginal trait variance $\mathbf{\Omega}_{jj}$, the jth diagonal element of $\mathbf{\Omega}$. If we scale them to kx_{ij} and $\mathbf{\Omega}_{jj}/k$ by any positive number k, then according to (3.3), the likelihood remains unchanged. Therefore, we need to fix the marginal variances for latent parameters underlying binary traits. On the other hand, continuous traits can be seen as directly observed latent parameters, and their marginal trait variances depend on the potentially differing rates of change along $\mathscr{F}$ and should be inferred from the data. A Wishart prior on $\mathbf{\Omega}^{-1}$ does not allow such distinct constraints on the marginal variances for binary and continuous traits. The second problem with the Wishart prior is that strong dependencies exist among correlations and their joint distribution is considerably different from uniform (Tokuda et al., 2011). Without knowing the true correlation structure, these prior assumptions may not be appropriate. Hence, we favor a noninformative, uniform prior on the correlation matrix.

We solve the above problems by decomposing $\mathbf{\Omega}$ into an across-trait correlation matrix and standard deviations, with a jointly uniform prior on the correlation matrix. Specifically, we decompose $\mathbf{\Omega} = \mathbf{DRD}$, where $\mathbf{R}$ is the $P \times P$ correlation matrix and $\mathbf{D}$ is a diagonal matrix with elements $D_{ii} = 1$, for $i = 1, \ldots, P_b$ and $D_{ii} = \sigma_i > 0$ for $i = P_b + 1, \ldots, P$. We use the prior of Lewandowski, Kurowicka, and Joe (LKJ) on the positive-definite correlation matrix $\mathbf{R}$ (Lewandowski, Kurowicka, and Joe, 2009a), with density

$$\text{LKJ}(\mathbf{R}|\eta) = c(\eta)\det(\mathbf{R})^{\eta-1}, \tag{3.5}$$

where $\eta > 0$ is a shape parameter and $c(\eta)$ is the normalizing constant. When $\eta = 1$, the

LKJ prior implies a uniform distribution over all correlation matrices of dimension P. For the diagonal standard deviation matrix $\mathbf{D}$, we assume independent log normal priors on the variances σ_i^2 for $i = P_b + 1, \cdots, P$ with mean 0 and variance 1 on the log scale. We describe how to carry out the posterior inference under this prior in Section 3.3.2. There exists other methods for specifying a prior distribution on $\mathbf{DRD}$. For example, Huang, Wand, et al. (2013) use half-t distributions on standard deviations and achieve marginally uniform correlations. We prefer log normal priors over half-t because the latter has non-zero probability density for a zero standard deviation. If one favors half-t standard deviations or marginally uniform correlations, our approach easily adapts to the prior in Huang, Wand, et al. (2013).

3.3 Inference

Primary scientific interest lies in the across-trait correlation matrix $\mathbf{R}$. We integrate out the nuisance parameters by sampling from the joint posterior

$$p(\mathbf{R}, \mathbf{D}, \mathbf{X}, \mathscr{F} \,|\, \mathbf{Y}, \mathbf{S}) \propto p(\mathbf{Y} \,|\, \mathbf{X}) \;\times\; p(\mathbf{X} \,|\, \mathbf{R}, \mathbf{D}, \mathscr{F}) \times \\ p(\mathbf{R}, \mathbf{D}) \;\times\; p(\mathbf{S} \,|\, \mathscr{F}) \;\times\; p(\mathscr{F}) \tag{3.6}$$

via a random-scan Gibbs scheme (Liu, Wong, and Kong, 1995), and drop the posterior's dependence on the hyper-parameters $(\boldsymbol{\Upsilon}, \boldsymbol{\mu}_0, \tau_0, g)$ to ease notation. The joint posterior factorizes because sequences $\mathbf{S}$ only affect the parameters of primary interest through $\mathscr{F}$, since we assume $\mathbf{S}$ to be conditionally independent of other parameters given $\mathscr{F}$.

Within the Gibbs scheme, we alternatively update $\mathbf{X}$, $(\mathbf{R}, \mathbf{D})$ and $\mathscr{F}$ from their full conditionals, taking advantage of the conditional independence structure. We construct $p(\mathbf{S} \,|\, \mathscr{F})$ from a continuous-time Markov chain evolutionary model (Suchard, Weiss, and Sinsheimer, 2001) that describes nucleotide substitutions along the branches of $\mathscr{F}$ that give rise to $\mathbf{S}$. We assume a typical tree prior $p(\mathscr{F})$ based on a coalescent process (Kingman, 1982) and

adopt a random-scan mixture of effective Metropolis-Hastings transition kernels (Suchard et al., 2018) to update parameters that define $\mathscr{F}$. For more details on tree sampling and tree priors choices, we refer interested readers to Suchard et al. (2018). This section focuses on overcoming the scalability bottleneck of updating $\mathbf{X}$ from an NP-dimensional truncated normal distribution by combining BPS with dynamic programming strategy. We also describe how we deploy Hamiltonian Monte Carlo (HMC) to update $(\mathbf{R}, \mathbf{D})$ to accommodate the non-conjugate prior on $\boldsymbol{\Omega} = \mathbf{DRD}$.

3.3.1 BPS for updating high-dimensional latent parameters

BPS is a non-reversible "rejection-free" sampler originally introduced in the computational physics literature by Peters and de With (2012) for simulating particle systems. Bouchard-Côté, Vollmer, and Doucet (2018) later adopted the algorithm with modifications to better suit statistical applications. BPS explores a target distribution $p(\mathbf{x})$ by simulating a piecewise deterministic Markov process. The simulated particle follows a piecewise linear trajectory, with its evolution governed by the landscape of the *energy* function $U(\mathbf{x}) := -\log p(\mathbf{x})$. To respect the target distribution, classical Monte Carlo algorithms first propose a move, then either accept or reject it such that a move towards areas of low probability or, equivalently, of high energy, is more likely to be rejected than one towards areas of high probability. On the other hand, BPS modifies its particle trajectory via a Newtonian elastic collision against the energy gradient, thereby avoiding wasteful rejected moves.

BPS is an efficient sampler for log-concave target distributions in general, with the additional ability to account for parameter constraints by treating them as "hard-walls" against which the particle bounces. Of particular interest to us is the fact that, when the target distribution is a truncated MVN, the critical computation for BPS implementation is multiplying the precision matrix of the unconstrained MVN by an arbitrary vector. So BPS becomes an especially efficient approach when one can carry out these matrix-vector operations quickly. In our application, the tree diffusion process only defines the covariance,

not the precision. But fortunately, the structured Brownian diffusion process enables us to efficiently compute the precision-vector products without costly matrix inversion. BPS also allows us to condition on a subset of dimensions that correspond to the continuous traits without extra computation. We begin with an overview of BPS following Bouchard-Côté, Vollmer, and Doucet (2018) and describe how to incorporate parameter constraints (Bierkens et al., 2018); the subsequent sections describe how to optimize the implementation when sampling from a truncated MVN.

3.3.1.1 BPS overview

To sample from the target distribution $p(\mathbf{x})$, BPS simulates a particle with position $\mathbf{x}(t)$ and velocity $\mathbf{v}(t)$ for time $t \geq 0$, initialized from $\mathbf{v}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and a given $\mathbf{x}_0$ at time $t = 0$. Over time intervals $t \in [t_k, t_{k+1}]$, the particle follows a piecewise linear path with velocity $\mathbf{v}(t) = \mathbf{v}_k$ and position $\mathbf{x}(t) = \mathbf{x}_k + (t - t_k)\mathbf{v}_k$. An inhomogeneous Poisson process governs the inter-event times $s_{k+1} = t_{k+1} - t_k$ with rate

$$\lambda(\mathbf{x}(t), \mathbf{v}_k) = \max\left\{0, \langle \mathbf{v}_k, \nabla U(\mathbf{x}(t))\rangle\right\}, \tag{3.7}$$

where $\langle \cdot, \cdot \rangle$ denotes an inner product.

When the target density is log-concave and differentiable, $U(\mathbf{x})$ is convex, so one can conveniently simulate the Markov process. We describe how to simulate the process for a pre-specified amount of time $t_{\text{total}} > 0$, and the mapping $\mathbf{x}_0 \to \mathbf{x}(t_{\text{total}})$ defines a Markov transition kernel with $p(\mathbf{x})$ as the stationary density:

1. Solve a one-dimensional optimization problem to find

$$s_{\min} = \operatorname*{argmin}_{s \geq 0} U(\mathbf{x}_{k-1} + s\mathbf{v}_{k-1}) \text{ and } U_{\min} = U(\mathbf{x}_{k-1} + s_{\min}\mathbf{v}_{k-1}). \tag{3.8}$$

2. Draw $T \sim \text{Exp}(1)$, an exponential random variable with rate 1, and solve for the next

inter-event time s_k, the minimal root of

$$U(\mathbf{x}_{k-1} + s_k \mathbf{v}_{k-1}) - U_{\min} = T \text{ and } s_k > s_{\min}. \tag{3.9}$$

3. Update $(\mathbf{x}, \mathbf{v})$ as

$$\mathbf{x}_k \leftarrow \mathbf{x}_{k-1} + s_k \mathbf{v}_{k-1}, \quad \mathbf{v}_k \leftarrow \mathbf{v}_{k-1} - 2\frac{\langle \mathbf{v}_{k-1}, \nabla U(\mathbf{x}_k)\rangle}{\|\nabla U(\mathbf{x}_k)\|^2}\nabla U(\mathbf{x}_k). \tag{3.10}$$

4. Stop if $\sum_{j=1}^{k} s_j \geq t_{\text{total}}$ and return $\mathbf{x}(t_{\text{total}}) = \mathbf{x}_{k-1} + (t_{\text{total}} - t_{k-1})\mathbf{v}_{k-1}$ where $t_{k-1} = \sum_{j=1}^{k-1} s_j$, otherwise repeat Steps 1 - 3.

Steps 1-4 form one conditional update by BPS inside a Gibbs scheme. They are the same as the basic BPS algorithm in Bouchard-Côté, Vollmer, and Doucet (2018), except that we do not include velocity refreshment as random Poisson events. Since we use BPS for conditional updates, we resample the velocity from $\mathcal{N}(\mathbf{0}, \mathbf{I})$ at the beginning of every BPS iteration. BPS without velocity refreshment is known to suffer from reducible behavior when applied to an isotropic multivariate normal distribution (Bouchard-Côté, Vollmer, and Doucet, 2018). Our velocity resampling already avoids this reducibility issue, and so we opt not to incorporate further refreshment inside the transition kernel. As long as the entire chain remains irreducible, Peskun-Tierney theory for non-reversible MCMC suggests that adding further events only reduces the efficiency (Andrieu and Livingstone, 2019; Bierkens and Duncan, 2017).

When the target distribution is constrained to some region $\mathbf{x} \in D$, the bounce events are caused not only by the gradient $\nabla U(\mathbf{x})$ but also by the domain boundary ∂D. We call these bounces "gradient events" and "boundary events" respectively. Whichever occurs first is the actual bounce. More precisely, we define the boundary event time $s_{\text{bd},k}$ as

$$s_{\text{bd},k} = \inf_{s>0} \{\mathbf{x}_{k-1} + s\mathbf{v}_{k-1} \notin D\}. \tag{3.11}$$

Then the bounce time is given by $s_k = \min\{s_{\mathrm{bd},k}, s_{\mathrm{gr},k}\}$, where $s_{\mathrm{gr},k}$ denotes the gradient event time of (3.9). If $s_{\mathrm{bd},k} < s_{\mathrm{gr},k}$, we have a boundary bounce and the position is updated as in (3.10) while the velocity is updated as

$$\mathbf{v}_k \leftarrow \mathbf{v}_{k-1} - 2\langle \mathbf{v}_{k-1}, \boldsymbol{\nu} \rangle \boldsymbol{\nu}, \tag{3.12}$$

where $\boldsymbol{\nu} = \boldsymbol{\nu}(\mathbf{x}_k)$ is a unit vector orthogonal to the boundary at $\mathbf{x}_k \in \partial D$.

3.3.1.2 BPS for truncated MVNs

We now describe how the BPS simulation simplifies when the target density is a d-dimensional truncated MVN of the form

$$\mathbf{x} \sim \mathcal{N}(\mathbf{m}, \boldsymbol{\Sigma}) \text{ subject to } \mathbf{x} \in D = \{\mathrm{sign}(\mathbf{x}) = \mathbf{y}\} \text{ for } \mathbf{y} \in \{\pm 1\}^d. \tag{3.13}$$

Importantly, we can implement BPS so that, besides basic and computationally inexpensive operations, it relies solely on matrix-vector multiplications by the precision matrix $\boldsymbol{\Phi} = \boldsymbol{\Sigma}^{-1}$. Moreover, under the orthant constraint $\{\mathrm{sign}(\mathbf{x}) = \mathbf{y}\}$, we can handle a bounce against the boundary in a particularly efficient manner, only requiring access to a column of $\boldsymbol{\Phi}$.

We start with gradient events and then describe how to find boundary event times. Now $U(\mathbf{x}) = -\log p(\mathbf{x}) = \frac{1}{2}(\mathbf{x} - \mathbf{m})^\intercal \boldsymbol{\Phi}(\mathbf{x} - \mathbf{m}) + C$ where constant C does not depend on $\mathbf{x}$, therefore

$$U(\mathbf{x} + s\mathbf{v}) = \frac{1}{2}\langle \mathbf{v}, \boldsymbol{\varphi}_{\mathbf{v}} \rangle s^2 + \langle \mathbf{v}, \boldsymbol{\varphi}_{\mathbf{x}} \rangle s + \frac{1}{2}\langle \mathbf{x} - \mathbf{m}, \boldsymbol{\varphi}_{\mathbf{x}} \rangle + C$$
$$\text{where } \boldsymbol{\varphi}_{\mathbf{v}} = \boldsymbol{\Phi}\mathbf{v} \text{ and } \boldsymbol{\varphi}_{\mathbf{x}} = \boldsymbol{\Phi}(\mathbf{x} - \mathbf{m}) = \nabla U(\mathbf{x}). \tag{3.14}$$

The solution to the optimization problem (3.8) is given by

$$\begin{aligned} s_{\min} &= \max\left\{0, -\langle \mathbf{v}, \boldsymbol{\varphi}_{\mathbf{x}} \rangle / \langle \mathbf{v}, \boldsymbol{\varphi}_{\mathbf{v}} \rangle\right\}, \\ U_{\min} &= \frac{1}{2}\langle \mathbf{v}, \boldsymbol{\varphi}_{\mathbf{v}} \rangle s_{\min}^2 + \langle \mathbf{v}, \boldsymbol{\varphi}_{\mathbf{x}} \rangle s_{\min} + \frac{1}{2}\langle \mathbf{x} - \mathbf{m}, \boldsymbol{\varphi}_{\mathbf{x}} \rangle + C. \end{aligned} \tag{3.15}$$

It follows from (3.14) that the gradient event time in (3.9) coincides with the larger root of the quadratic equation $as^2 + bs + c = 0$ with

$$a = \frac{1}{2}\langle \mathbf{v}, \boldsymbol{\varphi}_{\mathbf{v}} \rangle, \;\; b = \langle \mathbf{v}, \boldsymbol{\varphi}_{\mathbf{x}} \rangle, \;\; \text{and} \;\; c = -\frac{1}{2}\langle \mathbf{v}, \boldsymbol{\varphi}_{\mathbf{v}} \rangle s_{\min}^2 - \langle \mathbf{v}, \boldsymbol{\varphi}_{\mathbf{x}} \rangle s_{\min} - T,$$

so

$$s_{\text{gr}} = \frac{-b + \sqrt{b^2 - 4ac}}{2a}.$$

When a gradient event takes place, the position and velocity are updated according to (3.10) with

$$\nabla U(\mathbf{x} + s\mathbf{v}) = \boldsymbol{\varphi}_{\mathbf{x}+s\mathbf{v}} = \boldsymbol{\Phi}(\mathbf{x} - \mathbf{m}) + s\boldsymbol{\Phi}\mathbf{v} = \boldsymbol{\varphi}_{\mathbf{x}} + s\boldsymbol{\varphi}_{\mathbf{v}}. \tag{3.16}$$

Note that $\boldsymbol{\varphi}_{\mathbf{x}+s\mathbf{v}}$ can be computed by an element-wise addition of $\boldsymbol{\varphi}_{\mathbf{x}}$ and $s\boldsymbol{\varphi}_{\mathbf{v}}$, rather than the expensive matrix-vector operation $\mathbf{x} + s\mathbf{v} \to \boldsymbol{\Phi}(\mathbf{x} + s\mathbf{v})$.

The orthant boundary is given by $\cup_i \{x_i = 0\}$. When $\text{sign}(x_i) = \text{sign}(v_i)$, where x_i and v_i denotes the i-th coordinate of particle position and velocity, the particle is moving away from the i-th coordinate boundary $\{x_i = 0\}$ and thus never reaches it. Otherwise, the coordinate boundary is reached at time $s = |x_i/v_i|$. Hence s_{bd} can be expressed as

$$s_{\text{bd}} = |x_{i_{\text{bd}}}/v_{i_{\text{bd}}}|, \;\; i_{\text{bd}} = \text{argmin}_{i \in I} |x_i/v_i| \;\; \text{for} \;\; I = \{i : x_i v_i < 0\}.$$

When a boundary event takes place, the particle bounces against the plane orthogonal to the standard basis vector $\boldsymbol{\nu} = \boldsymbol{e}_{i_{\text{bd}}}$. As the updated velocity takes the form $\mathbf{v}^* \leftarrow \mathbf{v} - 2v_{i_{\text{bd}}}\boldsymbol{e}_{i_{\text{bd}}}$,

we can save computational cost of simulating the next line segment by realizing that

$$\boldsymbol{\varphi}_{\mathbf{v}^*} = \boldsymbol{\Phi}\mathbf{v}^* = \boldsymbol{\varphi}_{\mathbf{v}} + 2v^*_{i_{\text{bd}}}\boldsymbol{\Phi}\boldsymbol{e}_{i_{\text{bd}}} \quad \text{where} \quad v^*_{i_{\text{bd}}} = -v_{i_{\text{bd}}}. \tag{3.17}$$

In other words, we can compute $\boldsymbol{\varphi}_{\mathbf{v}^*}$ by simply extracting the i_{bd}-th column of $\boldsymbol{\Phi}$ and updating $\boldsymbol{\varphi}_{\mathbf{v}}$ with an element-wise addition. This avoids the expensive matrix-vector operation $\mathbf{v}^* \to \boldsymbol{\Phi}\mathbf{v}^*$.

Algorithm 1 describes BPS implementation for truncated MVNs based on the discussion above, with the most critical calculations optimized. Within each line segment, $\boldsymbol{\varphi}_{\mathbf{x}}$ and $\boldsymbol{\varphi}_{\mathbf{v}}$ once efficiently computed (Section 3.3.1.3) can be re-used throughout. In our application the observed continuous traits correspond to fixed dimensions in $\mathbf{x}$, so we slightly modify the BPS such that it can sample from a conditional truncated MVN. Specifically, we partition $\mathbf{x} = (\mathbf{x}_b, \mathbf{x}_c)$ by latent ($\mathbf{x}_b$) and observed dimensions ($\mathbf{x}_c$), with the aim to generate samples from the conditional distribution $p(\mathbf{x}_b \,|\, \mathbf{x}_c)$ (details in Appendix 3.8.1). We choose the tuning parameter t_{total} based on a heuristic that works well in practice (Section 3.8.1).

3.3.1.3 Dynamic programming strategy to overcome computational bottleneck

A straight implementation of BPS remains computationally challenging, as computing $\boldsymbol{\varphi}_{\mathbf{x}}$ and $\boldsymbol{\varphi}_{\mathbf{v}}$ in Algorithm 1 involves a high-dimensional matrix inverse when the model is parameterized in terms of $\boldsymbol{\Sigma}$. From (3.3) and the equivalence between matrix normal and multivariate normal distributions, to sample latent parameters $\mathbf{X}$ from their conditional posterior, the target distribution (3.13) specifies as $\mathbf{x} = \text{vec}\,(\mathbf{X})$, $\mathbf{m} = \text{vec}\,(\mathbf{M})$, $\boldsymbol{\Sigma} = \boldsymbol{\Omega} \otimes \boldsymbol{\Upsilon}$, and $\mathbf{y} = \text{vec}\,(\mathbf{Y})$, where $\text{vec}\,(\cdot)$ is the vectorization that converts an $N \times P$ matrix into an $NP \times 1$ vector and $\otimes$ denotes the Kronecker product. A naive matrix inverse operation $\boldsymbol{\Sigma}^{-1} = \boldsymbol{\Omega}^{-1} \otimes \boldsymbol{\Upsilon}^{-1}$ has an intimidating complexity of $\mathcal{O}(N^3 + P^3)$. If we have a fixed tree,

Algorithm 1 Bouncy particle sampler for multivariate truncated normal distributions

Require: t_{total}, initial value for $\mathbf{x}$

$\mathbf{v} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$

$\boldsymbol{\varphi}_{\mathbf{x}} \leftarrow \boldsymbol{\Phi}(\mathbf{x} - \mathbf{m})$ ▷ $\boldsymbol{\varphi}_{\mathbf{x}} = \nabla U(\mathbf{x})$ is the gradient of energy

while $t_{\text{total}} > 0$ **do**

▷ compute reused quantities once

  if previous bounce is a boundary event at coordinate i **then**

    $\boldsymbol{\varphi}_{\mathbf{v}} \leftarrow \boldsymbol{\varphi}_{\mathbf{v}} + 2v_i \boldsymbol{\Phi} \boldsymbol{e}_i$

  else

    $\boldsymbol{\varphi}_{\mathbf{v}} \leftarrow \boldsymbol{\Phi}\mathbf{v}$ ▷ the expensive step

  end if

  $\varphi_{\mathbf{v},\mathbf{x}} \leftarrow \mathbf{v}^{\intercal}\boldsymbol{\varphi}_{\mathbf{x}}, \varphi_{\mathbf{v},\mathbf{v}} \leftarrow \mathbf{v}^{\intercal}\boldsymbol{\varphi}_{\mathbf{v}}$

▷ find gradient event time

  $s_{\min} \leftarrow \max\{0, -\varphi_{\mathbf{v},\mathbf{x}}/\varphi_{\mathbf{v},\mathbf{v}}\}$

  $T \sim \text{Exp}(1)$

  $a \leftarrow \frac{1}{2}\varphi_{\mathbf{v},\mathbf{v}},\ b \leftarrow \varphi_{\mathbf{v},\mathbf{x}},\ c \leftarrow -\frac{1}{2}s_{\min}^2\varphi_{\mathbf{v},\mathbf{v}} - s_{\min}\varphi_{\mathbf{v},\mathbf{x}} - T$

  $s_{\text{gr}} \leftarrow (-b + \sqrt{b^2 - 4ac})/(2a)$

▷ find truncation event time at coordinate i

  $s_{\text{bd}} \leftarrow \text{argmin}_i\, x_i/v_i,$ for i with $x_i v_i < 0$

▷ bounce happens

  $s \leftarrow \min\{s_{\text{gr}}, s_{\text{bd}}, t_{\text{total}}\}$

  $\mathbf{x} \leftarrow \mathbf{x} + s\boldsymbol{v},\ \boldsymbol{\varphi}_{\mathbf{x}} \leftarrow \boldsymbol{\varphi}_{\mathbf{x}} + s\boldsymbol{\varphi}_{\mathbf{v}}$

  if $s = s_{\text{bd}}$ **then**

    $v_i \leftarrow -v_i$

  else if $s = s_{\text{gr}}$ **then**

    $\mathbf{v} \leftarrow \mathbf{v} - (2\langle \mathbf{v}, \boldsymbol{\varphi}_{\mathbf{x}} \rangle / \|\boldsymbol{\varphi}_{\mathbf{x}}\|^2)\, \boldsymbol{\varphi}_{\mathbf{x}}$

  end if

  $t_{\text{total}} \leftarrow t_{\text{total}} - s$

end while

such that $\boldsymbol{\Upsilon}^{-1}$ is known, the typical computation proceeds via

$$\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\mathbf{m}) = \left(\boldsymbol{\Omega}^{-1}\otimes\boldsymbol{\Upsilon}^{-1}\right)(\mathbf{x}-\mathbf{m}) = \operatorname{vec}\left(\boldsymbol{\Upsilon}^{-1}(\mathbf{X}-\mathbf{M})\,\boldsymbol{\Omega}^{-1}\right), \tag{3.18}$$

with a cost $\mathcal{O}(N^2P + NP^2)$. When the tree is random, the $\mathcal{O}(N^3)$ cost to get $\boldsymbol{\Upsilon}^{-1}$ seems unavoidable. However, we show that even with a random tree, evaluating $\boldsymbol{\varphi}_{\mathbf{x}}$ and $\boldsymbol{\varphi}_{\mathbf{v}}$ can be $\mathcal{O}(NP^2)$. We use conditional densities to evaluate these products (Proposition 1) and obtain all conditional densities simultaneously via a dynamic programming strategy that avoids explicitly inverting $\boldsymbol{\Upsilon}$.

Proposition 1. *Given joint variance matrix* $\boldsymbol{\Sigma}$ *and vectorized latent data* $\mathbf{x}$*, the energy gradient* $\nabla U(\mathbf{x})$ *is*

$$\boldsymbol{\varphi}_{\mathbf{x}} = \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\mathbf{m}) = \begin{pmatrix} \mathbf{Q}_1(\mathbf{X}_1-\boldsymbol{\mu}_1) \\ \vdots \\ \mathbf{Q}_N(\mathbf{X}_N-\boldsymbol{\mu}_N) \end{pmatrix}, \tag{3.19}$$

where $\boldsymbol{\mu}_i$ *and* $\mathbf{Q}_i$ *are the mean and the precision matrix of the distributions* $p(\mathbf{X}_i \,|\, \mathbf{X}_{(i)})$ *for* $i = 1, \ldots, N$*, and* $p(\mathbf{X}_i \,|\, \mathbf{X}_{(i)})$ *is the conditional distribution of latent parameters at one tree tip given those of all the other tips.*

Proof. $\mathbf{x} \sim \mathcal{N}(\mathbf{m}, \boldsymbol{\Sigma})$, so $p(\mathbf{X}_i \,|\, \mathbf{X}_{(i)})$ are also multivariate normal. Note that

$$\frac{\partial}{\partial \mathbf{x}}\left[\log p(\mathbf{x})\right] = -\frac{1}{2}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\mathbf{m}). \tag{3.20}$$

Likewise, $\frac{\partial}{\partial \mathbf{x}}\left[\log p(\mathbf{x})\right] = \left(\frac{\partial}{\partial \mathbf{X}_1}\left[\log p(\mathbf{x})\right], \ldots, \frac{\partial}{\partial \mathbf{X}_N}\left[\log p(\mathbf{x})\right]\right)^T$ with

$$\begin{aligned}\frac{\partial}{\partial \mathbf{X}_i}\left[\log p(\mathbf{x})\right] &= \frac{\partial}{\partial \mathbf{X}_i}\left[\log p(\mathbf{X}_i \,|\, \mathbf{X}_{(i)}) + \log p(\mathbf{X}_{(i)})\right] \\ &= \frac{\partial}{\partial \mathbf{X}_i}\left[\log p(\mathbf{X}_i \,|\, \mathbf{X}_{(i)})\right] \\ &= -\frac{1}{2}\mathbf{Q}_i\left(\mathbf{X}_i - \boldsymbol{\mu}_i\right).\end{aligned} \tag{3.21}$$

Equating (3.20) and (3.21) completes the proof. □

In Proposition 1, the partition is by taxon, but we can generalize to any arbitrary partitioning of the dimensions. By replacing $\mathbf{x} - \mathbf{m}$ with $\mathbf{v}$ (or $\boldsymbol{e}_i$), we achieve a similar result for $\boldsymbol{\varphi}_{\mathbf{v}}$ (or $\boldsymbol{\Phi}\boldsymbol{e}_i$). Given $\boldsymbol{\mu}_i$ and $\mathbf{Q}_i$, the $\mathcal{O}(NP^2)$ matrix-vector operation $\mathbf{v}^* \to \boldsymbol{\Phi}\mathbf{v}^*$ based on Proposition 1 is generally required for updating $\boldsymbol{\varphi}_{\mathbf{v}^*}$, but for boundary bounces, we can exploit (3.17) and update $\boldsymbol{\varphi}_{\mathbf{v}^*}$ in $\mathcal{O}(NP)$. For the conditional posterior distribution in our HIV application (Section 3.4), boundary bounces occur far more frequently than gradient ones and thus the efficient update via (3.17) leads to further significant speed-up.

Fortunately, we are able to efficiently compute $\boldsymbol{\mu}_i$ and $\mathbf{Q}_i$ through a dynamic programming strategy that recursively traverses the tree (Pybus et al., 2012) and enjoys a complexity of $\mathcal{O}(NP)$. Here we give the results and omit the derivatives found in Pybus et al. (2012) and Cybis et al. (2015).

The recursive traversals visit every node first in post-order (child → parent) and then again in pre-order (parent → child) to calculate partial data likelihoods that lead to $\boldsymbol{\mu}_i$ and $\mathbf{Q}_i$. The post-order traversal begins at a tip and ends at the root, while pre-order starts at the root and reaches every tip. The following results are in terms of the node triplets (i, j, k) where $\text{pa}(i) = \text{pa}(j) = k$ as in Figure 3.2. We define $\lfloor i \rfloor$ as the tree tips that are descendants to or include ("below") node i and $\lceil i \rceil$ as the tree tips that are not descendants to ("above") node i.

During the post-order traversal, the partial likelihoods of the data $\mathbf{X}_{\lfloor i \rfloor}$ given latent $\mathbf{X}_i$

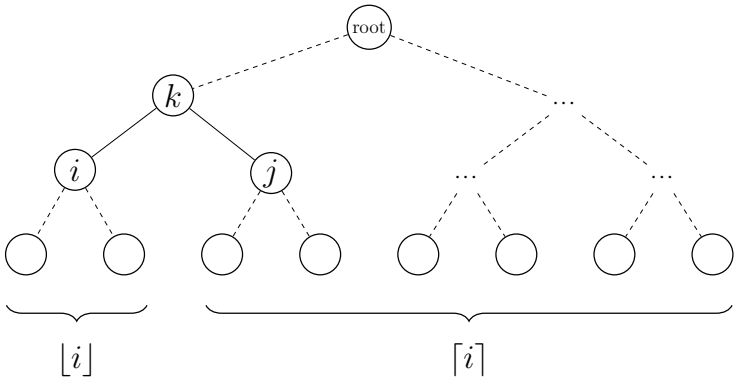


Figure 3.2: A sample tree to illustrate post- and pre- traversals for efficiently computing $p(\mathbf{X}_i \,|\, \mathbf{X}_{(i)})$. In the triplet (i, j, k), parent node k has two children i and j. We group the tip nodes into two disjoint and exhaustive classes: $\lfloor i \rfloor$ = tree tips that are descendants to or include node i and $\lceil i \rceil$ = tree tips that are not descendants to i.

is proportional to a MVN density of $\mathbf{X}_i$, in terms of a post-order mean $\mathbf{m}_i$ and variance $v_i\boldsymbol{\Omega}$ (Pybus et al., 2012), that is,

$$p(\mathbf{X}_{\lfloor i \rfloor} \,|\, \mathbf{X}_i) \propto \mathrm{MVN}(\mathbf{X}_i; \mathbf{m}_i, v_i\boldsymbol{\Omega})\,. \tag{3.22}$$

We re-employ these quantities shortly in the pre-order traversal. At the tree tips, $\mathbf{m}_i = \mathbf{X}_i$ and the variance scalar $v_i = 0$. For internal nodes,

$$\begin{aligned} \mathbf{m}_k &= v_k \left[(v_i + t_i)^{-1} \mathbf{m}_i + (v_j + t_j)^{-1} \mathbf{m}_j \right], \text{ with} \\ v_k &= \left[(v_i + t_i)^{-1} + (v_j + t_j)^{-1} \right]^{-1}. \end{aligned} \tag{3.23}$$

Similarly, for the pre-order traversal, we calculate the conditional density of $\mathbf{X}_i$ at node i given the data above it,

$$p(\mathbf{X}_i \,|\, \mathbf{X}_{\lceil i \rceil}) \propto \mathrm{MVN}(\mathbf{X}_i; \boldsymbol{\mu}_i, w_i\boldsymbol{\Omega})\,, \tag{3.24}$$

in terms of a pre-order mean $\boldsymbol{\mu}_i$ and variance $w_i\boldsymbol{\Omega}$. Starting from the root where $w_{2N-1} = \tau_0^{-1}$

and $\boldsymbol{\mu}_{2N-1} = \boldsymbol{\mu}_0$, the traversal proceeds via

$$\begin{aligned} \boldsymbol{\mu}_i &= w_i^* \left[(v_j + t_j)^{-1} \mathbf{m}_j + w_k^{-1} \boldsymbol{\mu}_k \right], \text{ with} \\ w_i^* &= \left[(v_j + t_j)^{-1} + w_k^{-1} \right]^{-1}, \text{ and} \\ w_i &= w_i^* + t_i. \end{aligned} \tag{3.25}$$

When reaching the tips where $\lceil i \rceil = (i)$, we obtain both the desired conditional mean $\boldsymbol{\mu}_i$ and precision $\mathbf{Q}_i = (w_i \boldsymbol{\Omega})^{-1}$.

For both pre- and post-order traversals, at each node we require $\mathcal{O}(P)$ elementary operations to obtain the mean vector and variance scalar; so, visiting all the nodes costs $\mathcal{O}(NP)$. With $\boldsymbol{\mu}_i$ and $\mathbf{Q}_i$ for $i = 1, \ldots, N$ ready in hand, the computation in (3.19) remains $\mathcal{O}(NP^2)$.

3.3.2 Hamiltonian Monte Carlo for updating trait covariance components

The across-trait covariance components $\mathbf{R}$ and $\mathbf{D}$ have complex and high-dimensional full conditional distributions, with no obvious structure to admit sampling via specialized algorithms. We therefore rely on HMC (Neal, 2011) to sample $p(\mathbf{R}, \mathbf{D} \mid \mathbf{X}, \mathscr{F})$ (see Section 2.3 for HMC details). We automate HMC tuning via the stochastic optimization approach of Andrieu and Thoms (2008) and the *No-U-Turn* algorithm of Hoffman and Gelman (2014). Because HMC applies most conveniently to a distribution without parameter constraints, we map $\mathbf{R}$ and $\mathbf{D}$ to an unconstrained space using standard transformations (Stan Development Team, 2018).

3.4 Application on HIV immune escape

3.4.1 Background

As a rapidly evolving RNA virus, HIV-1 has established extensive genetic diversity that researchers classify into different major groups and, for HIV-1 group M, into different sub-

types (Hemelaar, 2012). Such diversity implies that phenotypic traits can vary remarkably among strains circulating in different patients. Differences in viral virulence and their determinants, together with host factors, may explain the large variability in disease progression rates among patients. On the host side, human leukocyte antigen (HLA) class I alleles are important determinants of immune control that are known to be associated with differential HIV disease outcomes, with particular HLA alleles offering considerable protective effect (Goulder and Walker, 2012). An interesting phenomenon is that HIV-1 can evolve to escape the HLA-mediated immune response, but the responsible escape mutations may compromise fitness and hence reduce viral virulence (Nomura et al., 2013; Payne et al., 2014). Identifying these mutations and their effect on virulence while controlling for the evolutionary relationships among the viruses that spread in populations with heterogeneous HLA backgrounds represents a particular challenge. Here, we address this by estimating the posterior distribution of across-trait correlation while controlling for the unknown viral evolutionary history.

We analyze a data set of $N = 535$ aligned HIV-1 *gag* gene sequences collected from 535 patients in Botswana and South Africa between 2003 and 2010 (Payne et al., 2014). Both countries are severely affected by the subtype C variant of HIV-1 group M. Each sequence is associated with a known sampling date and phenotypic measurements, including $P_c = 3$ continuous traits that are replicative capacity (RC), viral load (VL), and cluster of differentiation 4 (CD4) cell count. An increasing VL and a decreasing CD4 count in the asymptomatic stage characterize a typical HIV infection; RC is a viral fitness measure obtained by an assay that, in this case, assesses the growth rate of recombinant viruses containing the patient-specific *gag-protease* gene relative to a control virus (Payne et al., 2014). We further link each sequence with $P_b = 21$ binary traits, including the presence/absence of candidate HLA-associated escape mutations at 20 different amino acid positions in the *gag* protein, and another binary trait for the country of sampling (Botswana or South Africa). In cases where ambiguous nucleotide states in a codon prevent the determination of presence/absence of

escape mutations, we encode binary trait states as unobserved (ranging from 0.2% to 21% across taxa) and set them as unbounded dimensions in the truncated normal distribution sampled by BPS.

3.4.2 Correlation among traits

We revisit the original study questions in Payne et al. (2014) concerning the extent to which HLA-driven HIV adaptation impacts virulence in both Botswana and South Africa populations. Differences in HIV adaptation and virulence may arise from the fact the HIV epidemic in Botswana precedes that in South Africa, leaving more time for the virus to adapt to protective HLA alleles. Our approach employing a Bayesian inference framework based on the phylogenetic multivariate probit model, is substantially different from Payne et al. (2014) as they did not control for the shared evolutionary history between samples. For this $N = 535, P_b = 21, P_c = 3$ data set, after fitting the phylogenetic multivariate probit model, we obtain posterior samples for parameters that are of scientific interest. For MCMC convergence assessment, we run the chain until the minimal effective sample size (ESS) across all dimensions of $\mathbf{X}$, $\mathbf{R}$ and $\mathbf{D}$ is above 200. This takes about 10^7 individual transition kernel applications under our random-scan Gibbs scheme (iterations) and 30 hours on an Amazon EC2 c5.large instance, and we discard the first 10% of the samples as burn-in. As a further diagnostic, we execute five independent chains and confirm that the potential scale reduction statistic $\hat{R}$ for all correlation elements fall within range [1, 1.04], well below the standard convergence criterion of 1.1 (Gelman, Rubin, et al., 1992). We implement the method in the software BEAST (Suchard et al., 2018), and provide the data set and source code in the online supplementary material.

The heat map in Figure 3.3 depicts significant across-trait correlation determined by a 90% highest posterior density (HPD) interval that does not contain zero. We mainly focus on the last 4 rows that relate to questions addressed by Payne et al. (2014), e.g. difference in HLA escape mutations between the two countries and correlation between escape muta-

tions and infection traits (VL and CD4 count) as well as replicative capacity. We identify

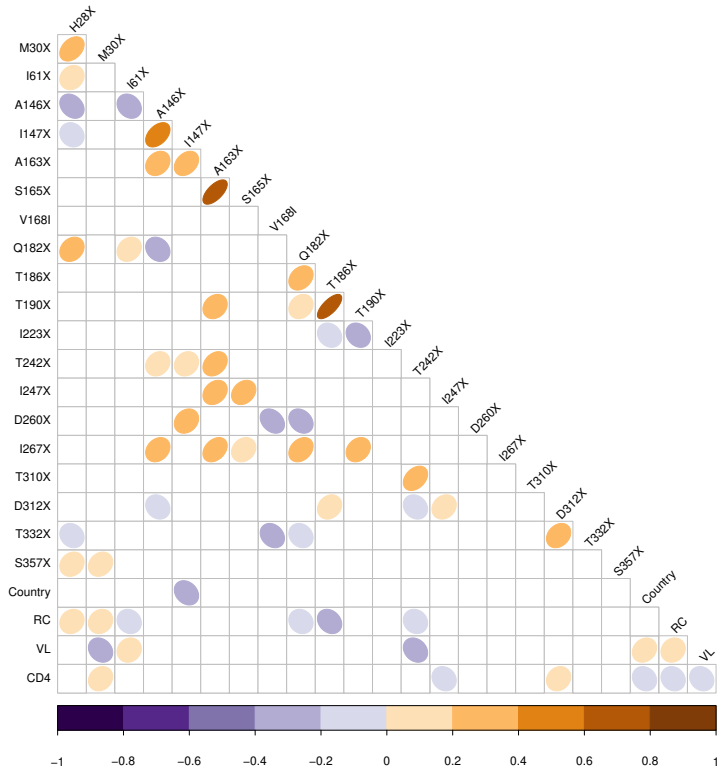


Figure 3.3: Significant across-trait correlation with $< 10\%$ posterior tail probability and their posterior mean estimates (in color). HIV *gag* mutations are named by the wild type amino acid state, the amino acid site number according to the standard reference genome (HXB2), and the amino acid 'escape' state that is any other amino acid or a deletion ('X') in almost all cases. Country = sample region: 1 = South Africa, -1 = Botswana; RC = replicative capacity; VL = viral load; CD4 = CD4 cell count.

one escape mutation I147X being significantly more prevalent in Botswana as indicated by its negative correlation with South Africa. Located at the amino-terminal position of an HLA-B57-restricted epitope ('ISW9'), variation at *gag* residue 147 is known to be associated with expression of B57 (Draenert et al., 2004). It is worth noting that three of the four escape mutations that correlate negatively with RC (I61X, Q182X and T242X) have a higher frequency in Botswana and may therefore have contributed to the lower RC found in Botswana by Payne et al. (2014). Interestingly, the negative effect on RC we estimate for two mutations finds clear confirmation in experimental testing: in vitro experiments provide

evidence for a reduction in RC by T242X (Martinez-Picado et al., 2006; Song et al., 2012) and T186X is also found to greatly impair RC (Huang et al., 2011).

Our analysis recovers the expected inverse correlation between CD4 count and RC or VL, as well as the positive correlation between RC and VL (Prince et al., 2012), confirming that more virulent viruses result in faster disease progression. Also, South Africa is associated with higher VL and lower CD4, suggesting that the South African cohort may comprise individuals with more advanced disease, even though the two cohorts are closely matched in age (Payne et al., 2014). This is somewhat at odds with the original study that also finds a higher VL for South Africa, but at the same time a higher CD4 count for patients from this country. Such differences are likely to arise from controlling or not for the phylogeny.

The remaining significant correlation between escape mutations (row 1 to 19 in Figure 3.3) can be considered as epistatic interactions, some of which are strongly positive. For example, we find a strong positive correlation between T186X and T190X. The former represents an escape mutation for HLA-B*81-mediated immune responses and has been reported to be strongly correlated with reduced virus replication (Huang et al., 2011; Wright et al., 2010), as also reflected in the negative correlation between this mutation and RC. In fact, Wright et al. (2012) show T186X requires T190I (or Q182X, also positively correlated with T186X, Figure 3.3) to partly compensate for this impaired RC. The other strong positive correlation between A163X and S165X has also been found to be a case of a compensatory mutation, with S165N partially compensating for the reduced viral RC of A163G (Crawford et al., 2007). The same holds true for the positive correlation between A146X and I147X, with I147L partially compensating the fitness cost associated with the escape mutation A146P (Troyer et al., 2009).

3.4.3 Tree inference

Figure 3.4 reports the maximum clade credibility tree from the posterior sample. The tree maximizes the sum of posterior clade probabilities. The posterior mean tree height is roughly

30 years; so with the most recent samples from 2010, we date the common ancestor of all viruses back to around 1980, consistent with the beginning of this epidemic.

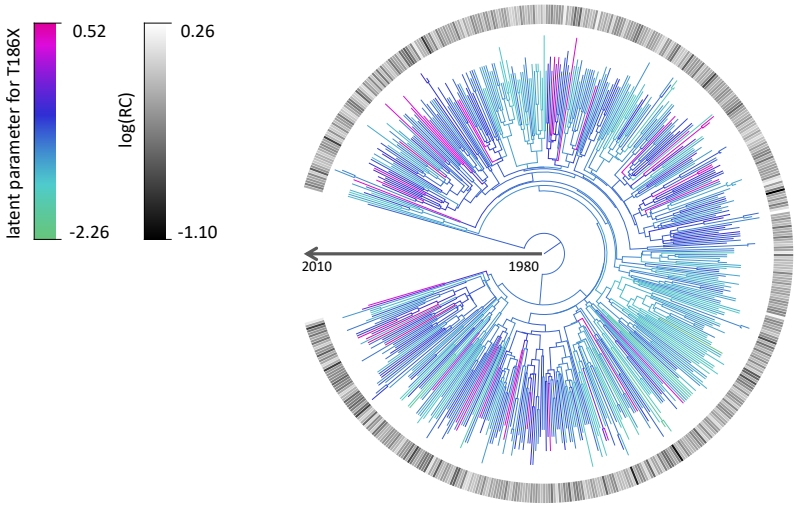


Figure 3.4: The maximum clade credibility tree with branches colored by the posterior mean of the latent parameter corresponding to mutation T186X. Outer circle shows log(RC) in gray scale.

3.5 Efficiency comparison and goodness-of-fit test

3.5.1 Efficiency comparison

To compare efficiency of BPS with the multiple-try rejection sampling in Cybis et al. (2015), we run both samplers on the whole data set ($N = 535, P = 24$) and a subset with $P = 8$ including the three continuous traits, and fix the tree and across-trait covariance at the same values from preliminary runs. The efficiency criterion is per unit-time ESS across all NP latent parameters. BPS outperforms rejection sampling to a greater extent as P increases. For $P = 24$, BPS yields a 74× increase in terms of the minimum ESS and an 11× increase for the median ESS (Table 3.1). This order-of-magnitude improvement is more clear in Figure 3.5. Because rejection sampling only updates one taxon per iteration, some latent parameters rarely change their values (Figure 3.6). As a result, the minimum ESS of

multiple-try rejection sampling is much lower than BPS which simultaneously updates all latent dimensions.

Table 3.1: Efficiency comparison between the bouncy particle sampler (BPS) and multiple-try rejection sampling in terms of minimum and median of effective sample size (ESS) per hour run-time. We report ESS values and their standard deviations (SD) across five independent simulations.

	$P = 8$		$P = 24$	
ESS/hr (SD)	min	median	min	median
BPS	5392 (411)	20596 (271)	282 (20)	1468 (11)
Rejection	237 (20)	4707 (25)	3.8 (0.1)	137 (0.7)
Speed-up	23×	4.4×	74×	11×

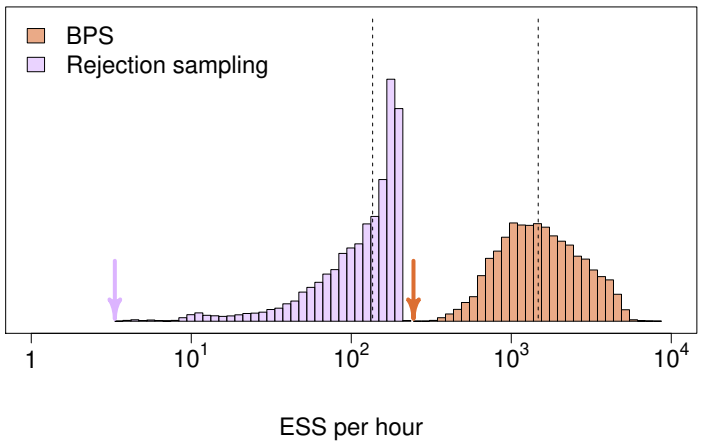


Figure 3.5: A representative histogram of ESS across latent parameters, sampled by BPS or rejection sampling in one hour run-time. Arrows and dashed lines denote the minimum and median ESS ($N = 535, P = 24$).

3.5.2 Model goodness-of-fit

We compare the phylogenetic probit model fit to reduced models that do not include phylogenetic correction. This comparison not only allows us to assess goodness-of-fit of the

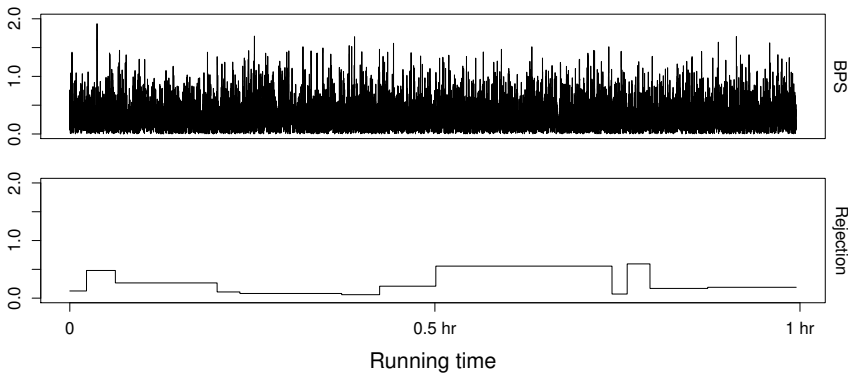


Figure 3.6: Trace plot of the latent parameter with the least ESS by rejection sampling (bottom) and trace plot of the same latent parameter sampled by BPS (top) for an one hour run-time. BPS and rejection sampling run 1.1×10^4 and 2.6×10^5 iterations, respectively $(N = 535, P = 24)$.

phylogenetic probit model, but also tests whether explicit tree modeling is necessary in practice. The two reduced models both assume independence among virus samples such that the across-taxa tree covariance $\boldsymbol{\Upsilon}$ is diagonal. The first "dated star" model incorporates varying viral sampling time information such that $\boldsymbol{\Upsilon}$ has diagonal elements equal to the time distance from virus sample date to the root date fixed, without loss of generality, to 1980. To understand the star-moniker, phylogeneticists often use a "star-tree" in which all branch lengths between internal nodes equal 0 to represent independent samples. The second "ultrametric star" model, assumes that all taxa have traits that are identically distributed so $\boldsymbol{\Upsilon}$ is an identity matrix.

For each of the three models, we assess out-of-sample prediction by repeatedly splitting up the HIV data into a training set used to build each model, and a test set to evaluate the prediction. Across the 21 binary traits for all taxa, we hold out $n_t = 21 \times 535 \times 20\%$ of the observations, build the model and then estimate the posterior probability r_h for $h = 1, \ldots n_t$ that held-out trait h equals its observed value.

We summarize performance through quantiles of the score $\log r_h$ to measure accuracy, and a higher score represents better prediction (Table 3.2). The phylogenetic probit model commands higher scores compared to the two reduced models and we conclude that joint

tree modeling through the phylogenetic probit model leads to better data fit.

Table 3.2: Prediction accuracy in out-of-sample logarithmic score. We report the score quantiles and their standard deviations (SD) across five independent MCMC simulations with 20% randomly held-out binary traits.

Log score quantiles (SD)	25%	50 %	75 %
Phylogenetic probit model	-0.441 (0.007)	-0.128 (0.003)	-0.029 (0.003)
Dated star model	-0.599 (0.014)	-0.187 (0.005)	-0.050 (0.006)
Ultrametric star model	-0.592 (0.011)	-0.187 (0.003)	-0.052 (0.006)

3.6 Discussion

We present an efficient Bayesian inference framework to learn the correlation among mixed-type traits across a large number of taxa, while jointly inferring the phylogenetic tree through sequence data. Our approach significantly improves upon Cybis et al. (2015) in both modeling and inference. Better modeling comes from the decomposition of across-trait covariance matrix $\mathbf{\Omega} = \mathbf{DRD}$ that keeps the generalized probit model identifiable and allows a jointly uniform LKJ prior on $\mathbf{R}$. Compared to the convenient but restrictive Wishart prior that causes mixing problems for sampling $\mathbf{\Omega}^{-1}$ and $\mathbf{X}$, this decomposition facilitates correlation inference among continuous traits and latent parameters (Appendix Figure A.2). Our main contribution lies in an efficient inference framework, specifically, an optimized BPS to sample latent parameters from a high-dimensional truncated normal distribution. In contrast to the "one-taxon-at-a-time" design in Cybis et al. (2015), BPS jointly updates all dimensions therefore reducing auto-correlation among MCMC samples. The most expensive steps involved are matrix-vector multiplications by the precision matrix $\mathbf{\Phi} = \mathbf{\Sigma}^{-1}$. In our case, the tree precision matrix is unknown and getting it by matrix inversion is notoriously $\mathcal{O}(N^3)$. Thanks to the insight in Proposition 1, we circumvent this obstacle by utilizing a dynamic programming strategy and obtain the desired matrix-vector products in $\mathcal{O}(NP^2)$. BPS also

enjoys an advantage especially important for mixed-type traits. That is, we can simply "mask out" the fixed continuous traits when sampling latent parameters for binary traits. Whereas the rejection sampling in Cybis et al. (2015) has to calculate the conditional distribution of latent dimensions given continuous traits at each tip. This cost-free "masking" technique to condition on a subset of dimensions exploits properties of normal distributions and can be shared with other dynamics-based sampler, like HMC. Taking all of these points together, the optimized BPS provides a huge gain in efficiency.

Naturally, BPS may also be an efficient choice in situations where $\mathbf{\Phi}$ itself has special structures that facilitate quick matrix-vector multiplication. For example, inducing precision matrices that are sparse or composed of sparse components is a common strategy for analyzing large spatial data (Heaton et al., 2019). Methods like the nearest neighbor Gaussian process (Datta et al., 2016), integrated nested Laplace approximations (Rue, Martino, and Chopin, 2009), and multi-resolution approximation of Gaussian processes (Katzfuss, 2017) all achieve computational efficiency from sparsity in $\mathbf{\Phi}$. Whether BPS would be useful in these scenarios, especially with mixed-type data, is an interesting topic for future research.

Our application provides important information on the complex association between HLA-driven HIV *gag* mutations and virulence that was previously assessed by experimental and epidemiological studies. To our best knowledge, this is the first study to examine essential HIV virus-host interactions while explicitly modeling the phylogenetic tree. Our setup is also different from the original study (Payne et al., 2014) in that we attempt to identify correlation between individual epitope escape mutations, virulence, and country of sampling, instead of considering all mutations together or grouping them with particular HLA types (e.g. HLA-B*57/58:01). While the latter may increase power to detect population-level differences in escape mutation frequencies, our approach allows us to pinpoint particular mutations contributing to virulence. Good consistency between the mutations that we associate with reduced RC and literature reports on virological assays suggests that our approach may complement or help in prioritizing experimental testing, and therefore further assist in the

battle against HIV-1. Our method contributes to a general framework to assess correlation among mixed-type traits in virology, but also more broadly in evolutionary biology.

One future improvement lies in the prior choice on across-trait correlation. The LKJ prior works well for our $N = 535, P = 24$ data set, as it is noninformative as desired, and correlation elements are well-mixed through No-U-Turn HMC. Under this choice, we view correlations with 90% HPD intervals not covering zero as significant. We can adjust this decision threshold based on resource availability for follow-up experimental studies. However, with much larger P and when only a small portion of the observed traits are truly involved in the underlying biology, it becomes vital to control for false positive signals, and one may favor a systematic solution. For example, it may be preferable to put a shrinkage-based prior on $\mathbf{R}$ that shrinks individual elements towards zero. Ideas like the graphical lasso prior (Wang et al., 2012) and factor models with shrinkage prior on the loading elements (Bhattacharya and Dunson, 2011) are potential directions to explore.

Lastly, as understanding the relationship among mixed-type variables is a common question in different fields, our method suits a large class of problems beyond evolutionary biology. The optimized BPS sampler through dynamic programming serves as an efficient inference tool for any multilevel (hierarchical) model (Gelman, 2006) with an additive covariance structure on a directed acyclic graph (Figure 3.1). The tree variance matrix $\boldsymbol{\Upsilon}$ that we use to describe the covariation of shared evolutionary history also arises from other kinds of relationships. For example, additive covariance includes pedigree-based or genomic relationship matrices in animal breeding (Mrode, 2014; Vitezica, Varona, and Legarra, 2013) and distance matrices decided by geographical locations in infectious disease research (Barbu et al., 2013). Intriguingly, our dynamic programming strategy also provides a way to invert the $N \times N$ tree variance matrix $\boldsymbol{\Upsilon}$ in $\mathcal{O}(N^2)$ by piecing together the products $\boldsymbol{\Upsilon}^{-1}\boldsymbol{e}_i$ for $i = 1, \ldots, N$. While this seems likely a well-known result, we have failed to find precedence in the literature. Finally, the phylogenetic probit model can be generalized to categorical and ordinal data, which will only add to its broad applicability.

3.7 Acknowledgments

We thank Oliver Pybus for useful discussions on an earlier version of the data set analyzed here. The research leading to these results has received funding from the European Research Council under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 725422 - ReservoirDOCS). The Artic Network receives funding from the Wellcome Trust through project 206298/Z/17/Z. PB acknowledges support by the Research Foundation – Flanders ('Fonds voor Wetenschappelijk Onderzoek – Vlaanderen', 12Q5619N and V434319N). MAS acknowledges support through NSF grant DMS 1264153 and NIH grants R01 AI107034 and U19 AI135995. PL acknowledges support by the Research Foundation – Flanders ('Fonds voor Wetenschappelijk Onderzoek – Vlaanderen', G066215N, G0D5117N and G0B9317N).

3.8 Appendices

3.8.1 BPS details

BPS modification for conditional truncated MVNs. Here we consider modifying the BPS to incorporate fixed dimensions that are the observed, continuous traits in our mixed-type model. We partition $\mathbf{x} = (\mathbf{x}_b, \mathbf{x}_c)$ by latent and observed dimensions and then generate samples from the conditional distribution $p(\mathbf{x}_b \,|\, \mathbf{x}_c)$. To make progress, we parameterize $p(\mathbf{x}_b \,|\, \mathbf{x}_c)$ in terms of $p(\mathbf{x})$ with partitioned mean $\mathbf{m} = (\mathbf{m}_b, \mathbf{m}_c)$ and precision matrix

$$\boldsymbol{\Sigma}^{-1} = \begin{bmatrix} \boldsymbol{\Phi}_{bb} & \boldsymbol{\Phi}_{bc} \\ \boldsymbol{\Phi}_{cb} & \boldsymbol{\Phi}_{cc} \end{bmatrix}. \tag{3.26}$$

With a similarly partitioned velocity $\mathbf{v} = (\mathbf{v}_b, \mathbf{v}_c)$, the distribution $p(\mathbf{x}_b \,|\, \mathbf{x}_c)$ carries potential energy

$$U_{b\,|\,c}(\mathbf{x}_b + t\mathbf{v}_b) = \frac{t^2}{2}\mathbf{v}_b^\intercal \boldsymbol{\Phi}_{bb}\mathbf{v}_b + t\mathbf{v}_b^\intercal \boldsymbol{\Phi}_{bb}(\mathbf{x}_b - \mathbf{m}_{b\,|\,c}) + C, \tag{3.27}$$

where constant C does not depend on t. The conditional mean $\mathbf{m}_{b\,|\,c} = \mathbf{m}_b - \mathbf{\Phi}_{bb}^{-1}\mathbf{\Phi}_{bc}(\mathbf{x}_c - \mathbf{m}_c)$, so

$$U_{b\,|\,c}(\mathbf{x}_b + t\mathbf{v}_b) = \frac{t^2}{2}\mathbf{v}_b^\intercal \mathbf{\Phi}_{bb}\mathbf{v}_b + t\mathbf{v}_b^\intercal\left[\mathbf{\Phi}_{bb}(\mathbf{x}_b - \mathbf{m}_b) + \mathbf{\Phi}_{bc}(\mathbf{x}_c - \mathbf{m}_c)\right] + C. \quad (3.28)$$

This expression is equivalent to masking out the dimensions of $\mathbf{v}$ in (3.14) that corresponds to $\mathbf{x}_c$ via the vector $\tilde{\mathbf{v}} = (\mathbf{v}_b, \mathbf{0})$. To be explicit, we rewrite (3.28) as

$$U_{b\,|\,c}(\mathbf{x}_b + t\mathbf{v}_b) = \frac{t^2}{2}\tilde{\mathbf{v}}^\intercal \mathbf{\Phi}\tilde{\mathbf{v}} + t\tilde{\mathbf{v}}^\intercal\mathbf{\Phi}(\mathbf{x} - \mathbf{m}) + C. \quad (3.29)$$

Therefore, adding this masking operation for $\mathbf{v}, \boldsymbol{\varphi}_{\mathbf{x}}, \boldsymbol{\varphi}_{\mathbf{v}}$ in Lines 1, 2, 5, 7 in Algorithm 1 allows sampling from the conditional truncated MVN $p(\mathbf{x}_b \,|\, \mathbf{x}_c)$ without any additional cost.

Tuning t_{total} for BPS. The total simulation time t_{total} for the Markov process is a tuning parameter in Algorithm 1. If t_{total} is too small, the particle does not travel far enough from the initial position, leading to high auto-correlation among MCMC samples. On the other hand, an unnecessarily large t_{total} would waste computational efforts without any substantial gain in mixing rate. To achieve best computational efficiency, therefore, one would like to choose a t_{total} just large enough that $\mathbf{x}(t_{\text{total}})$ is effectively independent of $\mathbf{x}(0)$. To help find such t_{total} for BPS applied to truncated MVNs, we develop a heuristic based on the following observations.

At stationarity, the BPS has a velocity distributed as $\mathcal{N}(\mathbf{0}, \mathbf{I})$. In other words, we have $\mathbf{v}(t) \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ for all $t \geq 0$ if starting from stationarity. In particular, the velocity along any unit vector $\mathbf{u}$ would be distributed as $\langle \mathbf{v}(t), \mathbf{u}\rangle \sim \mathcal{N}(0, 1)$, so that $\mathbb{E}|\langle \mathbf{v}(t), \mathbf{u}\rangle| = \sqrt{2/\pi}$. Now, the motion of the particle along $\mathbf{u}$ is given by $\langle \mathbf{x}(t), \mathbf{u}\rangle = \langle \mathbf{x}(0), \mathbf{u}\rangle + \int_0^t \langle \mathbf{v}(s), \mathbf{u}\rangle \mathrm{d}s$. At the same time, for a MVN with covariance $\mathbf{\Sigma}$, its high density region has a diameter proportional to $\sqrt{\lambda_{\max}}$, where $\lambda_{\max}$ denotes the largest eigenvalue of $\mathbf{\Sigma}$. Therefore, in order

to allow the particle to travel across the high density region, we would like it to move a distance proportional to $\sqrt{\lambda_{\max}}$, that is, $|\int_0^{t_{\text{total}}} \langle \mathbf{v}(s), \mathbf{u} \rangle \, \mathrm{d}s| \propto \sqrt{\lambda_{\max}}$.

Since BPS is designed to suppress the random-walk behavior of more traditional MCMC algorithms (Peters and de With, 2012), we expect the motion of the particle along $\mathbf{u}$ not to change its direction frequently. Or equivalently, we expect the velocity along $\mathbf{u}$, given by $\langle \mathbf{v}(t), \mathbf{u} \rangle$, not to change its sign frequently. When there is no change in $\langle \mathbf{v}(t), \mathbf{u} \rangle$ during $[0, t_{\text{total}}]$, we would have $|\int_0^{t_{\text{total}}} \langle \mathbf{v}(s), \mathbf{u} \rangle \, \mathrm{d}s| = \int_0^{t_{\text{total}}} |\langle \mathbf{v}(s), \mathbf{u} \rangle| \, \mathrm{d}s$. This, combined with the observation that $\mathbb{E}|\langle \mathbf{v}(t), \mathbf{u} \rangle| = \sqrt{2/\pi}$ at stationarity, suggest that roughly, the particle moves an average distance of $\sqrt{2/\pi}$ during one unit of time. We so conjecture that there is a choice of travel time $t_{\text{total}} \propto \sqrt{\lambda_{\max}}$ that achieves $|\int_0^{t_{\text{total}}} \langle \mathbf{v}(s), \mathbf{u} \rangle \, \mathrm{d}s| \propto \sqrt{\lambda_{\max}}$ and good mixing. This heuristic applies to a truncated MVN when assuming its high density region diameter is comparable to that of the untruncated MVN. We find that BPS performance is not overly sensitive to a specific choice of t_{total}. After preliminary runs (Table A.1), we choose $t_{\text{total}} = 0.01\sqrt{\lambda_{\max}}$ for our $N = 535, P = 24$ application, as it yields the maximum median effective sample size (ESS) per hour run-time.

Table A.1: Effective sample size per hour run-time (ESS/hr) of latent parameters sampled by BPS with different t_{total}. We fix the tree and use the No-U-Turn sampler to sample the across-trait covariance matrix. With $t_{\text{total}} = 0.01\sqrt{\lambda_{\max}}$, the minimum, 5%, and 50% percentile of ESS/hr are either larger or close to those with other t_{total} values compared.

	t_{total}		
ESS/hr percentile	$5 \times 10^{-3}\sqrt{\lambda_{\max}}$	$10^{-2}\sqrt{\lambda_{\max}}$	$10^{-1}\sqrt{\lambda_{\max}}$
min	72	68	27
5%	227	428	357
50%	515	1050	885

3.8.2 Identifiability issue with a Wishart prior

We examine differences between assuming an LKJ + log normal priors on **DRD** and a Wishart prior on $\mathbf{\Omega}^{-1}$. For the Wishart case, we set the degree of freedom equal to $P +$

1, so each correlation marginally follows a uniform distribution on $[-1, 1]$ (Gelman et al., 2013), and the Normal-Wishart conjugacy yields easy Gibbs sampling for $\mathbf{\Omega}^{-1}$. Without constraining the marginal variance of any latent dimension, the Wishart prior leaves the model not parameter-identifiable and causes mixing problems, even with a small $P = 8$ (Figure A.2).

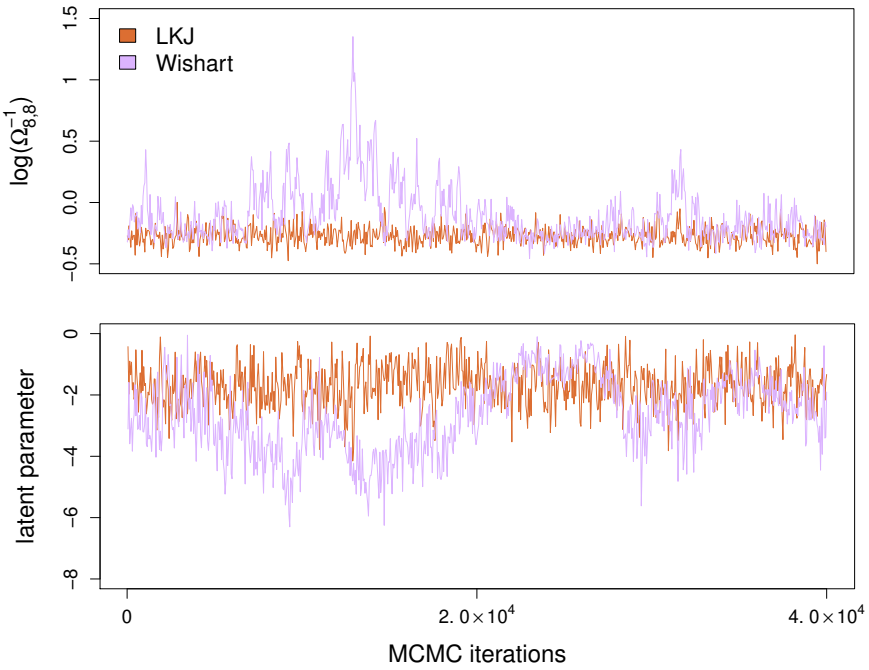


Figure A.2: Trace plot of a representative $\mathbf{\Omega}^{-1}$ element (top) in log scale and the latent parameter with the least ESS when assuming a Wishart prior on $\mathbf{\Omega}^{-1}$ (bottom).

CHAPTER 4

Accelerating Bayesian inference of dependency between complex biological traits

4.1 Introduction

An essential goal in evolutionary biology is to understand the associations between traits observed within biological samples, or *taxa*, ranging from plants and animals to microorganisms and pathogens such as human immunodeficiency virus (HIV) and influenza. This task is difficult because taxa are implicitly correlated through their shared evolutionary history often described with a reconstructed phylogenetic tree. Here, tree tips correspond to the taxa themselves, and internal nodes are their unobserved ancestors. Inferring across-trait covariation requires a highly structured model that can explicitly describe the tree structure and adjust for across-taxa covariation. Phylogenetic models do exactly this but are computationally challenging because one must integrate out unobserved ancestor traits while accounting for uncertainties arising from tree estimation. The computational burden increases when taxon and trait counts grow large and becomes worse when traits include continuous and discrete quantities. Zhang et al. (2021) show that their phylogenetic multivariate probit model provides a promising tool to learn correlations among complex traits at scale when combined with an efficient inference scheme that achieves order-of-magnitudes efficiency gains over the previous best approach (Cybis et al., 2015). Zhang et al. (2021) demonstrate their method on a data set with $N = 535$ HIV viruses and $P = 24$ traits that requires sampling from a truncated normal distribution with more than 11,000 dimensions.

In this work, we significantly advance performance compared to (Zhang et al., 2021) and solve more challenging problems including the (a) inference of across-trait partial correlations that present clues for potential causal pathways and (b) integration of complex traits with categorical outcomes.

To jointly model complex traits, the phylogenetic probit model assumes discrete traits arise from continuously valued latent variables that follow a Brownian diffusion along the tree (Cybis et al., 2015; Felsenstein, 1985; Zhang et al., 2021). Assuming latent processes is a common strategy for modeling complex data and it finds uses across various fields (Clark et al., 2017; Fedorov, Wu, and Zhang, 2012; Irvine, Rodhouse, and Keren, 2016; Pourmohamad, Lee, et al., 2016; Schliep and Hoeting, 2013). For N taxa and P continuous or binary traits, Bayesian inference for the phylogenetic probit model involves repeatedly sampling latent variables from their conditional posterior, an $(N \times P)$-dimensional truncated normal distribution. For this task, Zhang et al. (2021) develop a bouncy particle sampler (BPS) (Bouchard-Côté, Vollmer, and Doucet, 2018) augmented with an efficient dynamic programming approach that speeds up the most expensive step in the BPS implementation. However, BPS suffers from a major limitation — it does not allow joint sampling of the latent variables $\mathbf{X}$ and the trait correlation $\mathbf{R}$. Zhang et al. (2021) use a separate Hamiltonian Monte Carlo sampler (Neal, 2011, HMC) to infer $\mathbf{R}$ and update the two sets of parameters alternately within a random-scan Gibbs scheme (Liu, Wong, and Kong, 1995). Since $\mathbf{X}$ and $\mathbf{R}$ are highly correlated by model assumption, the Gibbs scheme hurts efficiency.

Our solution utilizes a state-of-the-art Markov chain Monte Carlo (MCMC) method called Zigzag-HMC (Nishimura, Dunson, and Lu, 2020). Zigzag-HMC can take advantage of the same $\mathcal{O}(N)$ gradient evaluation strategy advanced by Zhang et al. (2021), yet allows a joint update of $\mathbf{X}$ and $\mathbf{R}$ through differential operator splitting (Nishimura, Dunson, and Lu, 2020; Strang, 1968) which generalizes the previously proposed split HMC framework based on Hamiltonian splitting (Neal, 2011; Shahbaba et al., 2014). The joint sampling scheme greatly improves the mixing of elements in $\mathbf{R}$ and thus provides reliable estimates

of across-trait partial correlations that describe the conditional dependence between any two traits, free of confounding from other traits in the model. As seen in our applications, these conditional dependencies provide insights into potential causal pathways driven by real biological processes.

We apply our methodology to three real-world examples. First, we re-evaluate the HIV evolution application in Zhang et al. (2021) and identify HIV-1 *gag* immune-escape mutations linked with virulence through strong conditional dependence relationships. Our findings closely match with the experimental literature and indicate a general pattern in the immune escape mechanism of HIV. Second, we examine the influenza H1N1 glycosylation pattern across different hosts and detect strong conditional dependencies between glycosylation sites closely related to host switching. Finally, we investigate how floral traits of *Aquilegia* flower attract different pollinators, for which we generalize the phylogenetic probit model to accommodate a categorical pollinator trait.

4.2 Methods

4.2.1 Complex trait evolution

We describe biological trait evolution with the phylogenetic multivariate probit model following Zhang et al. (2021) and extend it to categorical traits as in Cybis et al. (2015). Consider N taxa on a phylogenetic tree $\mathscr{F} = (\mathbb{V}, \mathbf{t})$ that is a directed, bifurcating acyclic graph. We either know the tree *a priori* or infer it from a molecular sequence alignment $\mathbf{S}$ (Suchard et al., 2018). The node set $\mathbb{V}$ of size $2N-1$ contains N tip nodes, $N-2$ internal nodes and one root node. The branch lengths $\mathbf{t} = (t_1, \ldots, t_{2N-2})$ denote the child-parent distance in real time. We observe P traits of complex for each taxon. The trait data $\mathbf{Y} = \{y_{ij}\} = \left(\mathbf{Y^c}, \mathbf{Y^b}\right)$ partition as $\mathbf{Y^c}$, an $N \times P_c$ matrix of continuous traits and $\mathbf{Y^b}$, an $N \times P_b$ matrix of discrete ones. For each node i in $\mathscr{F}$, we assume a d-dimensional latent variable $\mathbf{X}_i \in \mathbb{R}^d$, $i = 1, \ldots, 2N-1$, where $d = P_c + \sum_{j=1}^{P_b} (m_j - 1)$ and m_j is the number

of classes for the jth discrete trait. To relate latent variables to observed discrete traits, we assume a threshold model for binary traits and a choice model for traits with more than two classes. For a categorical trait y_{ij}, the possible classes are $\{c_1, \ldots, c_{m_j}\}$ with the reference class being c_1. Multiple latent variables $x_{ij}i, j', \ldots, x_{ij}i, j' + m_j - 2$ decide the value of y_{ij}. We summarize the mapping from $\mathbf{X}$ to $\mathbf{Y}$ as

$$y_{ij} = \begin{cases} x_{ij}, & \text{if } y_{ij} \text{ is continuous,} \\ \text{sign}(x_{ij}ij), & \text{if } y_{ij} \text{ is binary,} \\ c_1, & \text{if } y_{ij} \text{ is categorical and } M = 0, \\ c_m, & \text{if } y_{ij} \text{ is categorical}, m > 1, \text{and } M = x_{ij}i, j' + m - 2 > 0, \end{cases} \tag{4.1}$$

where $M = \max(x_{ij}i, j', \ldots, x_{ij}i, j' + m_j - 2)$ and $\text{sign}(x_{ij})$ returns the value 1 on positive values and -1 on negative values. This data augmentation strategy is a common choice to model categorical data (Albert and Chib, 1993). As a side note, for continuous y_{ij} the corresponding x_{ij} is observed, and so $\mathbf{X}_i$ is actually a partially latent vector. Since in our applications only a small fraction of y_{ij} is continuous, we omit "partial" to ease the notation.

The latent variables follow a multivariate Brownian diffusion process along $\mathscr{T}$ such that $\mathbf{X}_i$ distributes as a multivariate normal (MVN)

$$\mathbf{X}_i \sim \mathcal{N}\left(\mathbf{X}_{\text{pa}(i)}, t_i\mathbf{\Omega}\right), i = 1, ..., 2N - 2, \tag{4.2}$$

where $\mathbf{X}_{\text{pa}(i)}$ is the parent node value and the $d \times d$ covariance matrix $\mathbf{\Omega}$ describes the across-trait association. The intuition behind $t_i\mathbf{\Omega}$ is that the further away a child node is from its parent node (larger t_i), the bigger difference between their node values. Assuming a conjugate root prior $\mathbf{X}_{2N-1} \sim \mathcal{N}\left(\boldsymbol{\mu}_0, \tau_0^{-1}\mathbf{\Omega}\right)$ with prior mean $\boldsymbol{\mu}_0$ and prior sample size τ_0, we can analytically integrate out latent variables on all internal nodes. Marginally, then, the

$N \times d$ tip latent variables $\mathbf{X}$ have the matrix normal (MTN) distribution

$$\mathbf{X} \sim \text{MTN}_{Nd}(\mathbf{M}, \boldsymbol{\Upsilon}, \boldsymbol{\Omega}), \tag{4.3}$$

where $\mathbf{M} = (\boldsymbol{\mu}_0, \ldots, \boldsymbol{\mu}_0)^T$ is an $N \times d$ mean matrix and the across-taxa covariance matrix $\boldsymbol{\Upsilon}$ equals $\mathbf{V}(\mathscr{F}) + \tau_0^{-1}\mathbf{J}$ (Pybus et al., 2012). The tree $\mathscr{F}$ determines the diffusion matrix $\mathbf{V}(\mathscr{F})$ and $\tau_0^{-1}\mathbf{J}$ comes from the integrated-out tree root prior, where $\mathbf{J}$ is an all-one $N \times N$ matrix. The augmented likelihood of $\mathbf{X}$ and $\mathbf{Y}$ factorizes as

$$p(\mathbf{Y}, \mathbf{X} \,|\, \boldsymbol{\Upsilon}, \boldsymbol{\Omega}, \boldsymbol{\mu}_0, \tau_0) = p(\mathbf{Y} \,|\, \mathbf{X}) p(\mathbf{X} \,|\, \boldsymbol{\Upsilon}, \boldsymbol{\Omega}, \boldsymbol{\mu}_0, \tau_0), \tag{4.4}$$

where $p(\mathbf{Y} \,|\, \mathbf{X}) = 1$ if $\mathbf{X}$ are consistent with $\mathbf{Y}$ according to Equation (4.1) and 0 otherwise. Following Zhang et al. (2021), we decompose $\boldsymbol{\Omega}$ as $\mathbf{DRD}$ such that $\mathbf{R}$ is the $d \times d$ correlation matrix and $\mathbf{D}$ is a diagonal matrix with marginal standard deviations. Importantly, since discrete traits only inform the sign or ordering of their underlying latent variables, certain elements of $\mathbf{D}$ must be set as a fixed value to ensure that the model is parameter-identifiable. Zhang et al. (2021) demonstrate the necessity of this $\mathbf{DRD}$ decomposition, which also allows a non-informative prior (Lewandowski, Kurowicka, and Joe, 2009a, LKJ) on $\mathbf{R}$. For goodness-of-fit of the phylogenetic probit model we refer interested readers to Zhang et al. (2021) where the explicit tree modeling leads to a significantly better fit.

4.2.2 A novel inference scheme

We sample from the joint posterior to learn the across-trait correlation $\mathbf{R}$

$$\begin{aligned} p(\mathbf{R}, \mathbf{D}, \mathbf{X}, \mathscr{F} \,|\, \mathbf{Y}, \mathbf{S}) \propto p(\mathbf{Y} \,|\, \mathbf{X}) \;\times\; p(\mathbf{X} \,|\, \mathbf{R}, \mathbf{D}, \mathscr{F}) \;\times \\ p(\mathbf{R}, \mathbf{D}) \;\times\; p(\mathbf{S} \,|\, \mathscr{F}) \;\times\; p(\mathscr{F}), \end{aligned} \tag{4.5}$$

where we drop the dependence on hyper-parameters $(\boldsymbol{\Upsilon}, \boldsymbol{\mu}_0, \tau_0)$ to ease notation. We then specify the priors $p(\mathbf{R}, \mathbf{D})$ and $p(\mathscr{F})$ as in Zhang et al. (2021). Assuming $p(\mathbf{R}, \mathbf{D}) = p(\mathbf{R})p(\mathbf{D})$ and an LKJ prior on $\mathbf{R}$, we set independent log normal priors on $\mathbf{D}$ diagonals that correspond to discrete traits, and assume a typical coalescent tree prior on $\mathscr{F}$ (Kingman, 1982). Zhang et al. (2021) use a random-scan Gibbs (Liu, Wong, and Kong, 1995) scheme to alternately update $\mathbf{X}$, $\{\mathbf{R}, \mathbf{D}\}$ and $\mathscr{F}$ from their full conditionals (Suchard et al., 2018). They sample $\mathbf{X}$ from an Nd-dimensional truncated normal distribution with BPS and deploy the standard HMC based on Gaussian momentum (Hoffman and Gelman, 2014) to update $\{\mathbf{R}, \mathbf{D}\}$. Instead, we simulate the joint Hamiltonian dynamics on $\{\mathbf{X}, \mathbf{R}, \mathbf{D}\}$ by combining novel Hamiltonian zigzag dynamics on $\mathbf{X}$ (Nishimura, Zhang, and Suchard, 2021) and traditional Hamiltonian dynamics on $\{\mathbf{R}, \mathbf{D}\}$. This strategy enables an efficient joint update of the two highly-correlated sets of parameters. We first describe how Zigzag-HMC samples $\mathbf{X}$ from a truncated normal and then detail the joint update of $\{\mathbf{X}, \mathbf{R}, \mathbf{D}\}$.

4.2.2.1 Zigzag-HMC for truncated multivariate normals

We outline the main ideas behind HMC (Neal, 2011) before describing Zigzag-HMC as a version of HMC based on *Hamiltonian zigzag dynamics* (Nishimura, Dunson, and Lu, 2020; Nishimura, Zhang, and Suchard, 2021). In order to sample a d-dimensional parameter $\boldsymbol{x} = (x_1, \ldots, x_d)$ from the target distribution $\pi(\boldsymbol{x})$, HMC introduces an auxiliary *momentum* variable $\boldsymbol{p} = (p_1, \ldots, p_d) \in \mathbb{R}^d$ and samples from the product density $\pi(\boldsymbol{x}, \boldsymbol{p}) = \pi(\boldsymbol{x})\pi(\boldsymbol{p})$ by numerically discretizing the Hamiltonian dynamics

$$\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t} = \nabla K(\boldsymbol{p}), \quad \frac{\mathrm{d}\boldsymbol{p}}{\mathrm{d}t} = -\nabla U(\boldsymbol{x}), \tag{4.6}$$

where $U(\boldsymbol{x}) = -\log \pi(\boldsymbol{x})$ and $K(\boldsymbol{p}) = -\log \pi(\boldsymbol{p})$ are the potential and kinetic energy. In each HMC iteration, we first draw $\boldsymbol{p}$ from its marginal distribution $\pi(\boldsymbol{p}) \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, a standard Gaussian and then approximate (4.6) from time $t = 0$ to $t = \tau$ by $L = \lfloor \tau/\epsilon \rfloor$ steps

of the *leapfrog* update with stepsize ϵ (Leimkuhler and Reich, 2004):

$$\boldsymbol{p} \leftarrow \boldsymbol{p} + \frac{\epsilon}{2}\nabla_{\boldsymbol{x}} \log \pi(\boldsymbol{x}), \quad \boldsymbol{x} \leftarrow \boldsymbol{x} + \epsilon \boldsymbol{p}, \quad \boldsymbol{p} \leftarrow \boldsymbol{p} + \frac{\epsilon}{2}\nabla_{\boldsymbol{x}} \log \pi(\boldsymbol{x}). \tag{4.7}$$

The end state is a valid *Metropolis* proposal that one accepts or rejects according to the standard acceptance probability formula (Hastings, 1970; Metropolis et al., 1953).

Zigzag-HMC differs from standard HMC insofar as it posits a Laplace momentum $\pi(\boldsymbol{p}) \propto \prod_i \exp(-|p_i|), i = 1, \ldots, d$. The Hamiltonian differential equations now become

$$\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t} = \operatorname{sign}(\boldsymbol{p}), \quad \frac{\mathrm{d}\boldsymbol{p}}{\mathrm{d}t} = -\nabla U(\boldsymbol{x}), \tag{4.8}$$

and the velocity $\boldsymbol{v} := \mathrm{d}\boldsymbol{x}/\mathrm{d}t \in \{\pm 1\}^d$ depends only on the sign of $\boldsymbol{p}$ and thus remains constant until one of p_i's undergoes a sign change (an "event"). To understand how the Hamiltonian zigzag dynamics (4.8) evolve over time, one must investigate when such events happen. Before moving to the truncated MVN, we first review the event time calculation for a general $\pi(\boldsymbol{x})$ following Nishimura, Zhang, and Suchard (2021). Let $\tau^{(k)}$ be the kth event time and $\left(\boldsymbol{x}\left(\tau^{(0)}\right), \boldsymbol{v}\left(\tau^{(0)}\right), \boldsymbol{p}\left(\tau^{(0)}\right)\right)$ is the initial state at time $\tau^{(0)}$. Between $\tau^{(k)}$ and $\tau^{(k+1)}$, $\boldsymbol{x}$ follows a piecewise linear path and the dynamics evolve as

$$\boldsymbol{x}\big(\tau^{(k)} + t\big) = \boldsymbol{x}\big(\tau^{(k)}\big) + t\boldsymbol{v}\big(\tau^{(k)}\big), \quad \boldsymbol{v}\big(\tau^{(k)} + t\big) = \boldsymbol{v}\big(\tau^{(k)}\big), \quad t \in \big[0, \tau^{(k+1)} - \tau^{(k)}\big), \tag{4.9}$$

and

$$p_i\big(\tau^{(k)} + t\big) = p_i\big(\tau^{(k)}\big) - \int_0^t \partial_i U\left[\boldsymbol{x}\big(\tau^{(k)}\big) + s\boldsymbol{v}\big(\tau^{(k)}\big)\right] ds \quad \text{for } i = 1, \ldots, d. \tag{4.10}$$

Therefore we can derive the $(k+1)$th event time

$$\tau^{(k+1)} = \tau^{(k)} + \min_i t_i, \quad t_i = \min_{t>0}\left\{p_i\big(\tau^{(k)}\big) = \int_0^t \partial_i U\big[\boldsymbol{x}(\tau^{(k)}) + s\boldsymbol{v}(\tau^{(k)})\big] ds\right\}, \tag{4.11}$$

and the dimension causing this event is $i^* = \text{argmin}_i\, t_i$. At the moment of $\tau^{(k+1)}$, the i^*th velocity component flips its sign

$$v_{i^*}\big(\tau^{(k+1)}\big) = -v_{i^*}\big(\tau^{(k)}\big), \quad v_j\big(\tau^{(k+1)}\big) = v_j\big(\tau^{(k)}\big) \text{ for } j \neq i^*. \tag{4.12}$$

Then the dynamics continue for the next interval $\big[\tau^{(k+1)}, \tau^{(k+2)}\big)$.

We now consider simulating the Hamiltonian zigzag dynamics for a d-dimensional truncated MVN defined as

$$\boldsymbol{x} \sim \mathcal{N}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right) \text{ subject to } \boldsymbol{x} \in \{\text{map}(\boldsymbol{x}) = \mathbf{y}\}, \tag{4.13}$$

where $\mathbf{y} \in \mathbb{R}^P$ is the complex data, $\text{map}(\cdot)$ is the mapping from latent variables $\boldsymbol{x}$ to $\mathbf{y}$ as in Equation (4.1), $\boldsymbol{x} \in \mathbb{R}^d$ and $d \geq P$. In this setting, we have $\nabla U(\boldsymbol{x}) = \boldsymbol{\Sigma}^{-1}\boldsymbol{x}$ whenever $\boldsymbol{x} \in \{\text{map}(\boldsymbol{x}) = \mathbf{y}\}$. Importantly, this structure allows us to simulate the Hamiltonian zigzag dynamics exactly and efficiently (Nishimura, Zhang, and Suchard, 2021). We handle the constraint $\text{map}(\boldsymbol{x}) = \mathbf{y}$ with a technique from Neal (2011) where the constraint boundaries embody "hard walls" that the Hamiltonian zigzag dynamics "bounce" against upon impact. To distinguish different types of events, we define *gradient events* arising from solutions of Equation (4.11), *binary events* arising from hitting binary data boundaries and *categorical events* arising from hitting categorical data boundaries.

We first consider how to find the gradient event time. Starting from a state $(\boldsymbol{x}, \boldsymbol{v}, \boldsymbol{p})$, by plugging in $\nabla U(\boldsymbol{x}) = \boldsymbol{\Sigma}^{-1}\boldsymbol{x}$ to Equation (4.11), we can calculate the gradient event time s_{gr} by first solving d quadratic equations

$$\boldsymbol{p} = t\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) + \frac{t^2}{2}\boldsymbol{\Sigma}^{-1}\boldsymbol{v}, \tag{4.14}$$

and then taking the minimum among all positive roots of Equation (4.14). When $\pi(\boldsymbol{x})$ is a truncated MVN arising from the phylogenetic probit model, we exploit the efficient gradient

evaluation strategy in Zhang et al. (2021) to obtain $\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$ and $\boldsymbol{\Sigma}^{-1}\boldsymbol{v}$ without the notorious $\mathcal{O}(d^3)$ cost to invert $\boldsymbol{\Sigma}$.

Next, we focus on the binary and categorical events. We partition $\boldsymbol{x}$ into three sets: $S_{\text{cont}} = \{x_i : x_i \text{ is for continuous data}\}$, $S_{\text{bin}} = \{x_i : x_i \text{ is for binary data}\}$, and $S_{\text{cat}} = \{x_i : x_i \text{ is for categorical data}\}$. Since latent variables in S_{cont} are fixed, we "mask" them out following Zhang et al. (2021). Starting from a state $(\boldsymbol{x}, \boldsymbol{v}, \boldsymbol{p})$, a binary event happens at time s_{bd} when the trajectory first reaches a binary boundary at dimension i_{b}

$$s_{\text{bd}} = |x_{i_{\text{b}}}/v_{i_{\text{b}}}|\,,\ i_{\text{b}} = \operatorname{argmin}_{i\in I_{\text{bin}}} |x_i/v_i| \ \text{ for } \ I_{\text{bin}} = \{i : x_i v_i < 0 \text{ and } x_i \in S_{\text{bin}}\}. \tag{4.15}$$

Here, we only need to check the dimensions satisfying $x_i v_i < 0$, i.e., those for which the trajectory is heading towards the boundary. At time s_{bd}, the trajectory bounces against the binary boundary, and so the i_{b}th velocity and momentum element both undergo an instantaneous flip $v_{i_{\text{b}}} \leftarrow -v_{i_{\text{b}}}$, $p_{i_{\text{b}}} \leftarrow -p_{i_{\text{b}}}$, while other dimensions stay unchanged.

Finally, we turn to categorical events. Suppose that a categorical trait $y_j = c_m$ belongs to one of n possible classes, and $x_1, x_2, \ldots, x_{n-1}$ the underlying latent variables. Equation (4.1) specifies the boundary constraints. If $m = 1$, the $n-1$ latent variables must be all negative, which poses the same constraint as if they were for $n-1$ binary traits, therefore we can solve the event time using Equation (4.15). If $m > 1$, we must check when and which two dimensions first violate the order constraint $x_{ij}m - 1 = \max(x_{ij}1, \ldots, x_{ij}n - 1) > 0$. With the dynamics starting from $(\boldsymbol{x}, \boldsymbol{v}, \boldsymbol{p})$, the categorical event time s_{c}^j is given by

$$\begin{aligned} s_{\text{c}}^j = |(x_{m-1} - x_{i_{\text{c}}})/(v_{m-1} - v_{i_{\text{c}}})|\,,\ i_{\text{c}} = \operatorname{argmin}_{i\in I_{\text{cat}}} |(x_{m-1} - x_i)/(v_{m-1} - v_i)|\,, \\ \text{for } \ I_{\text{cat}} = \{i : v_{m-1} < v_i \text{ and } x_i \in S_{\text{cat}}\}, \end{aligned} \tag{4.16}$$

when $x_{ij}i_{\text{c}}$ reaches $x_{ij}m - 1$ and violates the constraint. To identify i_{c} we only need to check dimensions with $v_{m-1} < v_i$ where the distance $x_{ij}m - 1 - x_{ij}i$ is decreasing. At s_{c}^j, the two dimensions involved ($m-1$ and i_{c}) bounce against each other such that $v_{m-1} \leftarrow -v_{m-1}$,

$v_{i_c} \leftarrow -v_{i_c}$, $p_{m-1} \leftarrow -p_{m-1}$, $p_{i_c} \leftarrow -p_{i_c}$. Note s_c^j is for a single y_j and we need to consider all categorical data to find the actual categorical event time $s_c = \min_j s_c^j$.

We now present the dynamics simulation with all three event types included, starting from a state $(\boldsymbol{x}, \boldsymbol{v}, \boldsymbol{p})$ with $\boldsymbol{x} \in \{\text{map}(\boldsymbol{x}) = \mathbf{y}\}$:

1. Solve s_{gr}, s_{bd}, s_c using Equations (4.14), (4.15) and (4.16) respectively.

2. Determine the actual (first) event time $t = \min\{s_{\text{gr}}, s_{\text{bd}}, s_c\}$ and update $\boldsymbol{x}$ and $\boldsymbol{p}$ as in Equations (4.9) and (4.10) for a duration of t.

3. Make instantaneous velocity and momentum sign flips according to the rules of the actual event type, then go back to Step 1.

Based on the above discussion, Algorithm 2 describes one iteration of Zigzag-HMC on truncated MVNs where we simulate the Hamiltonian zigzag dynamic for a pre-specified duration t_{total}. For a truncated MVN arising from the phylogenetic probit model, we adopt the dynamic programming strategy of Zhang et al. (2021) to speed up the most expensive gradient evaluation step in line 3 and reduce its cost from $\mathcal{O}(N^2 d + N d^2)$ to $\mathcal{O}(N d^2)$. In brief, this strategy avoids explicitly inverting $\boldsymbol{\Upsilon}$ by recursively traversing the tree (Pybus et al., 2012) to obtain N conditional densities that directly translate to the desired gradient.

4.2.2.2 Jointly updating latent variables and across-trait covariance

The $N \times d$ latent variables and $d \times d$ across-trait covariance are highly correlated with each other, so individual Gibbs updates can be inefficient. The posterior conditional of $\mathbf{X}$ is truncated normal and thus allows for the efficient Hamiltonian zigzag simulation as described in Section 4.2.2.1. The conditional distribution for covariance components $\mathbf{R}$ and $\mathbf{D}$ has no such special structure, so we map them to an unconstrained space and deploy Hamiltonian dynamics based on Gaussian momentum. We use a standard mapping of $\mathbf{R}$ elements to real numbers (Stan Development Team, 2018) that first transforms $\mathbf{R}$ to canonical partial corre-

Algorithm 2 Zigzag-HMC for multivariate truncated normal distributions

function HzzTMVN($\boldsymbol{x}, \boldsymbol{p}, t_{\text{total}}$)
  $\boldsymbol{v} \leftarrow \text{sign}(\boldsymbol{p})$
  $\boldsymbol{\varphi_x} \leftarrow \boldsymbol{\Phi}(\boldsymbol{x} - \boldsymbol{\mu})$
  $t_{\text{remain}} \leftarrow t_{\text{total}}$
  while $t_{\text{remain}} > 0$ **do**
    ▷ find gradient event time s_{gr}
    $\boldsymbol{a} \leftarrow \boldsymbol{\varphi_v}/2, \boldsymbol{b} \leftarrow \boldsymbol{\varphi}_x, \boldsymbol{c} \leftarrow -\boldsymbol{p}$
    $s_{\text{gr}} \leftarrow \min_i \{\text{minPositiveRoot}(a_i, b_i, c_i)\}$ ▷ "minPositiveRoot" defined below
    ▷ find binary boundary event time
    $s_{\text{bd}} \leftarrow \min_i x_i/v_i$, for i with $x_i v_i < 0$ and $x_i \in S_{\text{bin}}$
    ▷ find categorical boundary event time, n_c = number of categorical traits
    for $j = 1, \ldots, n_c$ **do**
      $s_{\text{c}}^j \leftarrow \min_i |(x_{k-1} - x_{i_c})/(v_{k-1} - v_i)|$ for i with $v_{k-1} < v_i$ and $x_i \in S_{\text{cat}}$
    end for
    $s_{\text{c}} \leftarrow \min_j s_{\text{c}}^j$
    ▷ the actual event happens at time s
    $s \leftarrow \min\{s_{\text{gr}}, s_{\text{bd}}, s_{\text{c}}, t_{\text{remain}}\}$
    $\boldsymbol{x} \leftarrow \boldsymbol{x} + s\boldsymbol{v}$, $\boldsymbol{p} \leftarrow \boldsymbol{p} - s\boldsymbol{\varphi_x} - s^2\boldsymbol{\varphi_v}/2$, $\boldsymbol{\varphi_x} \leftarrow \boldsymbol{\varphi_x} + s\boldsymbol{\varphi_v}$
    if a gradient event happens at i_g **then**
      $v_{i_g} \leftarrow -v_{i_g}$
    else if a binary boundary event happens at i_{b} **then**
      $v_{i_{\text{b}}} \leftarrow -v_{i_{\text{b}}}$, $p_{i_{\text{b}}} \leftarrow -p_{i_{\text{b}}}$
    else if a categorical boundary event happens at $i_{\text{c1}}, i_{\text{c2}}$ **then**
      $v_{i_{\text{c1}}} \leftarrow -v_{i_{\text{c1}}}, v_{i_{\text{c2}}} \leftarrow -v_{i_{\text{c2}}}, p_{i_{\text{c1}}} \leftarrow -p_{i_{\text{c1}}}, p_{i_{\text{c2}}} \leftarrow -p_{i_{\text{c2}}}$
    end if
    $\boldsymbol{\varphi_v} \leftarrow \boldsymbol{\varphi_v} + 2v_i\boldsymbol{\Phi}\boldsymbol{e}_i$
    $t_{\text{remain}} \leftarrow t_{\text{remain}} - s$
  end while
return $\boldsymbol{x}, \boldsymbol{p}$
end function

* minPositiveRoot(a_i, b_i, c_i) returns the minimal positive root of the equation $a_i x^2 + b_i x + c = 0$, or else returns $+\infty$ if no positive root exists.

lations (CPC) that fall in $[-1, 1]$ and then apply the Fisher transformation to map CPC to the real line. We then construct the joint update of latent variables and covariance via differential operator splitting (Nishimura, Dunson, and Lu, 2020; Strang, 1968) to approximate the joint dynamics of Laplace-Gauss mixed momenta.

We denote the two concatenated sets of parameters $\mathbf{X}$ and $\{\mathbf{R}, \mathbf{D}\}$ as $\boldsymbol{x} = (\boldsymbol{x}_{\mathrm{G}}, \boldsymbol{x}_{\mathrm{L}})$ with momenta $\boldsymbol{p} = (\boldsymbol{p}_{\mathrm{G}}, \boldsymbol{p}_{\mathrm{L}})$, where indices G and L refer to Gaussian or Laplace momenta. The joint sampler updates $(\boldsymbol{x}_G, \boldsymbol{p}_G)$ first, then $(\boldsymbol{x}_L, \boldsymbol{p}_L)$, followed by another update of $(\boldsymbol{x}_G, \boldsymbol{p}_G)$. This symmetric splitting ensures that the simulated dynamics is reversible and hence constitute a valid *Metropolis* proposal mechanism (Nishimura, Dunson, and Lu, 2020). The LG-STEP function in Algorithm 3 describes the process of simulating the joint dynamics for time duration 2ϵ via the analytical Hamiltonian zigzag dynamics for $(\boldsymbol{x}_L, \boldsymbol{p}_L)$ and the approximate leapfrog dynamics (4.7) for $(\boldsymbol{x}_G, \boldsymbol{p}_G)$. Because $\boldsymbol{x}_G$ and $\boldsymbol{x}_L$ can have very different scales, we incorporate a tuning parameter, the step size ratio r, to allow different step sizes for the two dynamics. To approximate a trajectory of the joint dynamics from $t = 0$ to $t = \tau$, we apply the function LG-STEP $m = \lfloor \tau / 2\epsilon \rfloor$ times, and accept or reject the end point following the standard acceptance probability formula (Hastings, 1970; Metropolis et al., 1953). We call this version of HMC based on Laplace-Gauss mixed momenta as *LG-HMC* and describe one iteration of LG-HMC in Algorithm 3 where the inputs include the joint potential function $U(\boldsymbol{x}_G, \boldsymbol{x}_L)$. We use LG-HMC to update $\{\mathbf{X}, \mathbf{R}, \mathbf{D}\}$ as a Metropolis-within-Gibbs step of our random-scan Gibbs scheme. The overall sampling efficiency strongly depends on m, the step size ϵ and the step size ratio r, so it is preferable to auto-tune all of them. Appendix 4.6 provides an empirical method to automatically tune r. We utilize the no-U-turn algorithm to automatically decide the trajectory length m (Hoffman and Gelman, 2014) and call the resulting algorithm *LG No-U-Turn Sampler* (LG-NUTS). We adapt the step size ϵ with primal-dual averaging to achieve an optimal acceptance rate (Hoffman and Gelman, 2014).

Algorithm 3 One LG-HMC iteration

function LG-HMC($\boldsymbol{x}_G, \boldsymbol{x}_L, \boldsymbol{p}_G, \boldsymbol{p}_L, U, m, \epsilon, r$)

▷ Record the initial state

  $\boldsymbol{x}_G^0 \leftarrow \boldsymbol{x}_G, \boldsymbol{x}_L^0 \leftarrow \boldsymbol{x}_L, \boldsymbol{p}_G^0 \leftarrow \boldsymbol{p}_G, \boldsymbol{p}_L^0 \leftarrow \boldsymbol{p}_L$

  for $i = 1, \ldots, m$ **do**

    $\boldsymbol{x}_G, \boldsymbol{x}_L, \boldsymbol{p}_G, \boldsymbol{p}_L \leftarrow$ LG-STEP($\boldsymbol{x}_G, \boldsymbol{x}_L, \boldsymbol{p}_G, \boldsymbol{p}_L, \epsilon, r$)

  end for

▷ Calculate the acceptance probability a, where K_G and K_L denote the kinetic energy based on Gaussian or Laplace momentum and $\|\cdot\|_1$, $\|\cdot\|_2$ are the L^1 and L^2 norm.

  $K_G^0 \leftarrow \left(\|\boldsymbol{p}_G^0\|_2\right)^2/2,\ K_L^0 \leftarrow \|\boldsymbol{p}_L^0\|_1$

  $K_G \leftarrow \left(\|\boldsymbol{p}_G\|_2\right)^2/2,\ K_L \leftarrow \|\boldsymbol{p}_L\|_1$

  $a \leftarrow \min\{1, \exp\left[U(\boldsymbol{x}_G^0, \boldsymbol{x}_L^0) - U(\boldsymbol{x}_G, \boldsymbol{x}_L) + K_G^0 + K_L^0 - K_G - K_L\right]\}$

▷ Accept or reject

  $u \leftarrow$ one draw from uniform$(0, 1)$

  if $u < a$ **then**

    return $\boldsymbol{x}_G, \boldsymbol{x}_L, \boldsymbol{p}_G, \boldsymbol{p}_L$

  else

    return $\boldsymbol{x}_G^0, \boldsymbol{x}_L^0, \boldsymbol{p}_G^0, \boldsymbol{p}_L^0$

  end if

end function

function LG-STEP($\boldsymbol{x}_G, \boldsymbol{x}_L, \boldsymbol{p}_G, \boldsymbol{p}_L, \epsilon, r$)

  $\boldsymbol{x}_G, \boldsymbol{p}_G \leftarrow$ LeapFrog($\boldsymbol{x}_G, \boldsymbol{p}_G, \epsilon$)

  $\boldsymbol{x}_L, \boldsymbol{p}_L \leftarrow$ HzzTMVN($\boldsymbol{x}_G, \boldsymbol{p}_G, r\epsilon$)

  $\boldsymbol{x}_G, \boldsymbol{p}_G \leftarrow$ LeapFrog($\boldsymbol{x}_G, \boldsymbol{p}_G, \epsilon$)

  return $\boldsymbol{x}_G, \boldsymbol{x}_L, \boldsymbol{p}_G, \boldsymbol{p}_L$

end function

function LeapFrog($\boldsymbol{x}_G, \boldsymbol{p}_G, \epsilon$)

  $\boldsymbol{p}_G \leftarrow \boldsymbol{p}_G + \frac{\epsilon}{2}\nabla_{\boldsymbol{x}_G} \log p(\boldsymbol{x})$

  $\boldsymbol{x}_G \leftarrow \boldsymbol{x}_G + \epsilon \boldsymbol{p}_G$

  $\boldsymbol{p}_G \leftarrow \boldsymbol{p}_G + \frac{\epsilon}{2}\nabla_{\boldsymbol{x}_G} \log p(\boldsymbol{x})$

  return $\boldsymbol{x}_G, \boldsymbol{x}_L$

end function

4.3 Results

We demonstrate the superior efficiency of our joint inference scheme on learning dependency between traits under the phylogenetic probit model, as compared to the state-of-the-art BPS. To illustrate the broad applicability of our method, we detail three real-world applications and discuss the scientific findings. In Section 4.3.1 we apply our method to the HIV virulence application of Zhang et al. (2021). The improved efficiency (shown in Section 4.3.2) allows us to estimate the across-trait partial correlation with adequate effective sample size (ESS) and to reveal the conditional dependence among traits of scientific interest. We use the same HIV data set to demonstrate that LG-HMC and LG-NUTS outperform BPS (Section 4.3.2), followed by two more LG-NUTS applications on influenza (Section 4.3.3) and *Aquilegia* flower (Section 4.3.4) evolution. We conclude this section with MCMC convergence criteria and timing results (Section 4.3.5).

4.3.1 HIV immune escape

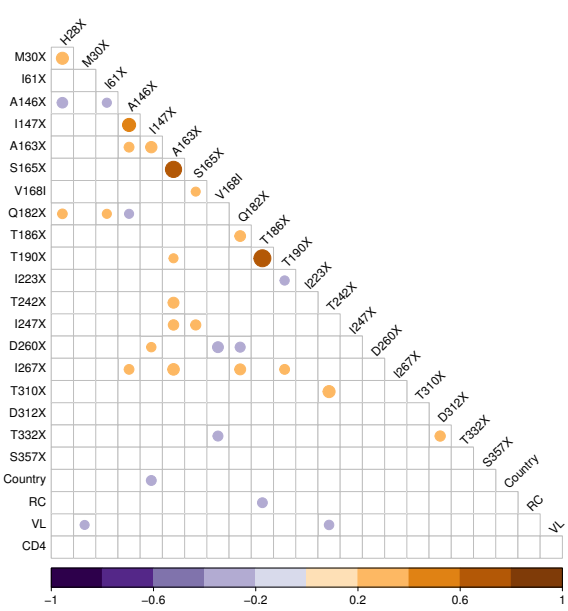
In the HIV evolution application of Zhang et al. (2021), a main scientific focus lies on the association between HIV-1 immune escape mutations and virulence, the pathogen's ability to cause disease. The human leukocyte antigen (HLA) system is predictive of the disease course as it plays an important role in the immune response against HIV-1. Through its rapid evolution, HIV-1 can acquire mutations that aid in escaping HLA-mediated immune response, but the escape mutations may reduce its fitness and virulence (Nomura et al., 2013; Payne et al., 2014). Zhang et al. (2021) identify HLA escape mutations associated with virulence while controlling for the unknown evolutionary history of the viruses. However, Zhang et al. (2021) interpret their results based on the across-trait correlation $\mathbf{R}$ which only informs marginal associations that can remain confounded. Now armed with a more efficient inference method, we are able to focus on the across-trait partial correlation matrix $\mathbf{P} = \{r_{ij}\}$ that indicates the conditional dependency between two interested traits without

confounding from other factors. We obtain $\mathbf{P}$ by transforming the inferred $\mathbf{\Omega}$ through

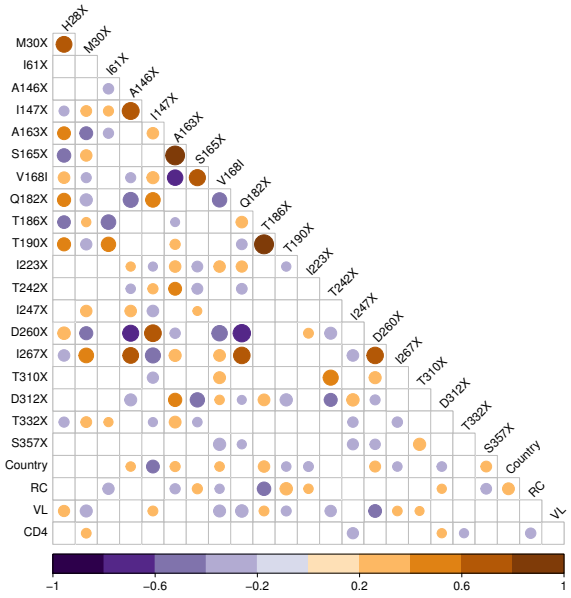
$$\mathbf{\Omega}^{-1} = \mathbf{\Omega}^{-1} = \{p_{ij}\}, \quad r_{ij} = -\frac{p_{ij}}{\sqrt{p_{ii}p_{jj}}}. \tag{4.17}$$

The data contain $N = 535$ aligned HIV-1 *gag* gene sequences collected from 535 patients between 2003 and 2010 in Botswana and South Africa (Payne et al., 2014). Each sequence is associated with 3 continuous and 21 binary traits. The continuous virulence measurements are replicative capacity (RC), viral load (VL) and cluster of differentiation 4 (CD4) cell count. The binary traits include the existence of HLA-associated escape mutations at 20 different amino acid positions in the *gag* protein and another trait for the sampling country (Botswana or South Africa). Figure 4.1 depicts across-trait correlations and partial correlations with posterior medians > 0.2 (or < -0.2). Compared to correlations (Figure 4.1a), we observe more partial correlations with greater magnitude (Figure 4.1b). They indicate conditional dependencies among traits after removing effects from other variables in the model, helping to explore the causal pathway. For example, we only detect a negative conditional dependence between RC and CD4. In other words, holding one of CD4 and RC as constant, the other does not affect VL, suggesting that RC increases VL via reducing CD4. The fact that RC is not found to share a strong conditional dependence with VL may be explained by the strong modulatory role of immune system on VL. Only when viruses with higher RC also lead to more immune damage, as reflected in the CD4 count, higher VL may be observed as a consequence of less suppression of viral replication. As such, our findings are in line with the demonstration that viral RC impacts HIV-1 immunopathogenesis independent of VL (Claiborne et al., 2015).

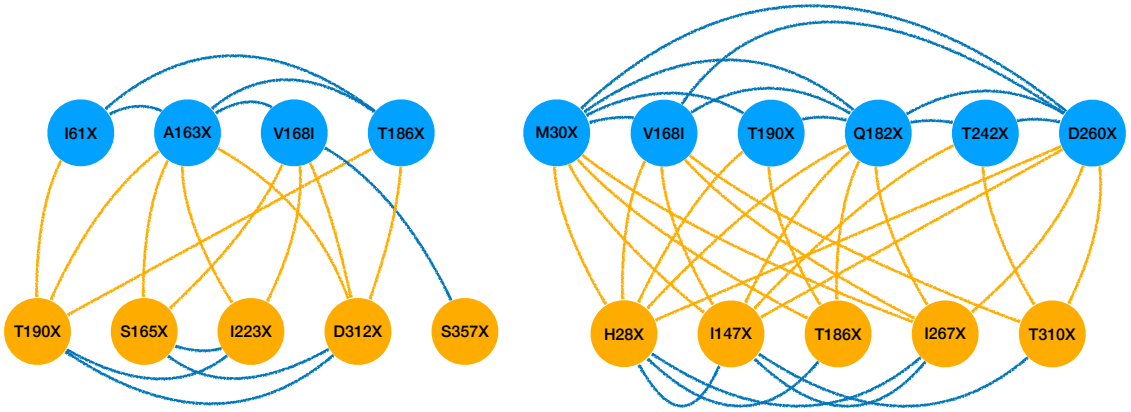
The partial correlation also helps to decipher epistatic interactions and how the escape mutations and potential compensatory mutations affect HIV-1 virulence. For example, we find a strong positive partial correlation between T186X and T190X. Studies have shown that T186X is highly associated with reduced VL (Huang et al., 2011; Wright et al., 2010) and it



(a) Correlation



(b) Partial correlation



RC VL

(c) Conditional dependencies among mutations

Figure 4.1: (a) Across-trait correlation and (b) partial correlation with a posterior median > 0.2 or < -0.2 (in color). HIV *gag* mutation names start with the wild type amino acid state, followed by the amino acid site number according to the HXB2 reference genome and end with the amino acid as a result of the mutation ('X' means a deletion). Country = sample region: 1 = South Africa, -1 = Botswana; RC = replicative capacity; VL = viral load; CD4 = CD4 cell count. (c) Conditional dependencies between HIV-1 immune escape mutations that affect RC or VL. Node and edge color indicates whether the dependence is positive (orange) or negative (blue).

requires T190I to partly compensate for this impaired fitness so the virus stays replication competent (Wright et al., 2012). The negative conditional dependence between T186X and RC and the positive conditional dependence between T190I and RC are consistent with this experimental observation. In contrast, with the strong positive association between T186X and T190, the marginal association fails to identify their opposite effects on RC. Another pair of mutations that potentially shows a similar interaction is H28X and M30X, which have a positive and negative partial correlation with VL, respectively. These mutations have indeed been observed to co-occur in *gag* epitopes from longitudinally followed-up patients (Olusola, Olaleye, and Odaibo, 2020). Figure 4.1b keeps all the other compensatory mutation pairs in Figure 4.1a such as A146X-I147X and A163X-S165X that find confirmation in experimental studies (Crawford et al., 2007; Troyer et al., 2009).

More generally, when considering the viral trait RC and the infection trait VL, for which their variation are to a considerable extent attributable to viral genetic variation (Blanquart et al., 2017), we reveal an intriguing pattern. As in Figure 4.1c, when two escape mutations impair virulence, and there is a conditional dependence between them, it is always negative. When two mutations have opposing effects on these virulence traits, the conditional dependence between them (if present) is almost always positive, with one exception of the negative effect between V168I and S357X. For example, T186X and I61X both have a negative impact on RC and the negative effect between them suggests that their additive, or even potentially synergistic, impact on RC is inhibited. Moreover, they appear to benefit from a compensatory mutation, T190X, which has been corroborated for the T186X-T190X pair at least as reported above. Also for VL, the conditional dependence between mutations that both have a negative impact on this virulence trait is consistently negative. Several of these individual mutations may benefit from H28X as a compensatory mutation, as indicated by the positive effect between pairs that include this mutation, and as suggested above for H28X - M30X. This illustrates the extent to which escape mutations may have a negative impact on virulence and the need to evolve compensatory mutations to restore it. We note

that our analysis is not designed to recover compensatory mutations at great length as we restrict it to a limited set of known escape mutations, while mutations on many other sites may be compensatory. In fact, our analysis suggests that some of the considered mutations may be implicated in immune escape due to their compensatory effect rather than a direct escape benefit.

4.3.2 Efficiency gain from the new inference scheme

We demonstrate that the joint update of latent variables $\mathbf{X}$ and the covariance matrix $\mathbf{\Omega}$ significantly improve inference efficiency. Table 4.1 compares the performance of four sampling schemes on the HIV immune escape example (described in Section 4.3.1) with $N = 535, P_b = 21, P_c = 3$. We choose our efficiency criterion to be the per run-time ESS for the across-trait correlation $\mathbf{R} = \{\sigma_{ij}\}$ and partial correlation $\mathbf{P} = \{r_{ij}\}$ that are of chief scientific interest. BPS and Zigzag-HMC only update $\mathbf{X}$ and we use the standard NUTS transition kernel (i.e. standard HMC combined with no-U-turn algorithm) for the $\mathbf{\Omega}$ elements. LG-HMC employs the joint update of $\mathbf{X}$ and $\mathbf{\Omega}$ described in Section 4.2.2.2. LG-NUTS additionally employs the No-U-Turn algorithm to decide the number of steps and a primal-dual averaging algorithm to calibrate the step size. We set the same t_{total} for BPS and Zigzag-HMC for a fair comparison. To tune LG-HMC, we first supply it with an optimal step size ϵ learned by LG-NUTS, then decide the number of steps $m = 100$ as it gives the best performance among the choices (10, 100, 1000). As reported in Table 4.1, it is indeed harder to infer partial correlations than correlations and jointly updating $\mathbf{X}$ and $\mathbf{\Omega}$ largely eliminates this problem. BPS loses to the three other samplers and LG-HMC performs the best in terms of ESS for r_{ij}, yielding a 5× speed-up. Without the joint update of $\mathbf{X}$ and $\mathbf{\Omega}$, Zigzag-HMC is only slightly more efficient than BPS. While a formal theoretical analysis is beyond the scope of this work, we provide an empirical explanation for the different performances of BPS and Zigzag-HMC in Appendix 4.6. Compared to the manually optimized LG-HMC, LG-NUTS has a slightly lower efficiency likely because the No-U-Turn algorithm

requires simulating trajectory both forward and backward to maintain reversibility and this process incurs additional steps (Hoffman and Gelman, 2014). In practice, we recommend using the tuning-free LG-NUTS.

Table 4.1: Efficiency comparison among different sampling schemes. Efficiency is in terms of minimal effective sample size (ESS) per running hour (hr) for correlation and partial correlation matrix elements σ_{ij} and r_{ij}. We report median values across 3 independent simulations and all numbers are relative to the minimal per-hr ESS of r_{ij} using BPS ($= 1^*$).

	min ESS/hr	
Sampler	σ_{ij}	r_{ij}
BPS	4.0	1^*
Zigzag-HMC	9.4	1.6
LG-HMC	5.1	5.0
LG-NUTS	5.3	4.2

4.3.3 Glycosylation of Influenza A virus H1N1

Influenza A viruses of the H1N1 subtype currently circulate in birds, humans, and swine (Song et al., 2008; Trovão and Nelson, 2020; Webster et al., 1992), where they are responsible for substantial morbidity and mortality (Boni et al., 2013; Ma, 2020). The two surface glycoproteins hemagglutinin (HA) and neuraminidase (NA) interact with a cell surface receptor and so their characteristics largely affect virus fitness and transmissibility. Mutations in the HA and NA, particularly in their immunodominant head domain, sometimes produce glycosylations that shield the antigenic sites against detection by host antibodies and so help the virus evade antibody detection (Daniels et al., 2003; Hebert et al., 1997; Östbye et al., 2020; Skehel et al., 1984). On the other hand, glycosylation may interfere with the receptor binding and also be targeted by the innate host immunity to neutralize viruses. Therefore there must be an equilibrium between competing pressures to evade immune detection and maintain virus fitness (Lin et al., 2020; Tate et al., 2014). The number of glycosylations that leads to this balance is expected to vary in host species experiencing different strengths of

immune selection. Despite decades of tracking IAVs evolution in humans for vaccine strain selection and recent expansions of zoonotic surveillance, the evolvability and selective pressures on the HA and NA have not been rigorously compared across multiple host species. Here, we examine the conditional dependence between host type and multiple glycosylation sites by estimating the posterior distribution of across-trait partial correlation while jointly inferring the IAVs evolutionary history.

We use hemagglutinin (H1) and neuraminidase (N1) sequence data sets for influenza A H1N1 produced by Trovão et al. as described in Trovão et al. (2022). We scan all H1 and N1 sequences to identify potential N-linked glycosylation sites, based on the motif Asn-X-Ser/Thr-X, where X is any amino acid other than proline (Pro) (Mellquist et al., 1998). We then set a binary trait for each sequence encoding for the presence or absence of glycosylations at a particular amino acid site. We keep sites with a glycosylation frequency between 20% and 80% for our analysis. This gives six sites in H1 and four sites in N1. We include another binary trait for the host type being mammalian (human or swine) or avian, so the sample sizes are $N = 964, P = 7$ (H1) and $N = 896, P = 5$ (N1).

The six H1 glycosylation sites consist of three pairs that are physically close (63/94, 129/163, and 278/289, see Figure 4.2). Sites 63 and 94 are particularly close to each other, though distances will vary slightly with sequence. A negative conditional dependence suggests glycosylation at two close sites may be harmful for the virus (63/94 and 278/289) while a positive effect between two sites suggests a potential benefit (63/129 and 94/278). We detect a negative conditional dependence between mammalian host and glycosylation site 94 and 289. Avian viruses have a stronger tendency to have site 289 glycosylated (Figure 4.2). In N1, glycosylations are more strongly correlated than H1 (Figure 4.3). Two pairs of glycosylation sites have a positive conditional dependency in between (50/68 and 50/389) and two pairs (44/68 and 68/389) have a negative one. We omit a structural interpretation since all sites but 389 are located in the NA stalk, for which no protein structure is available. There is a positive conditional dependence between mammalian host and glycosylations at

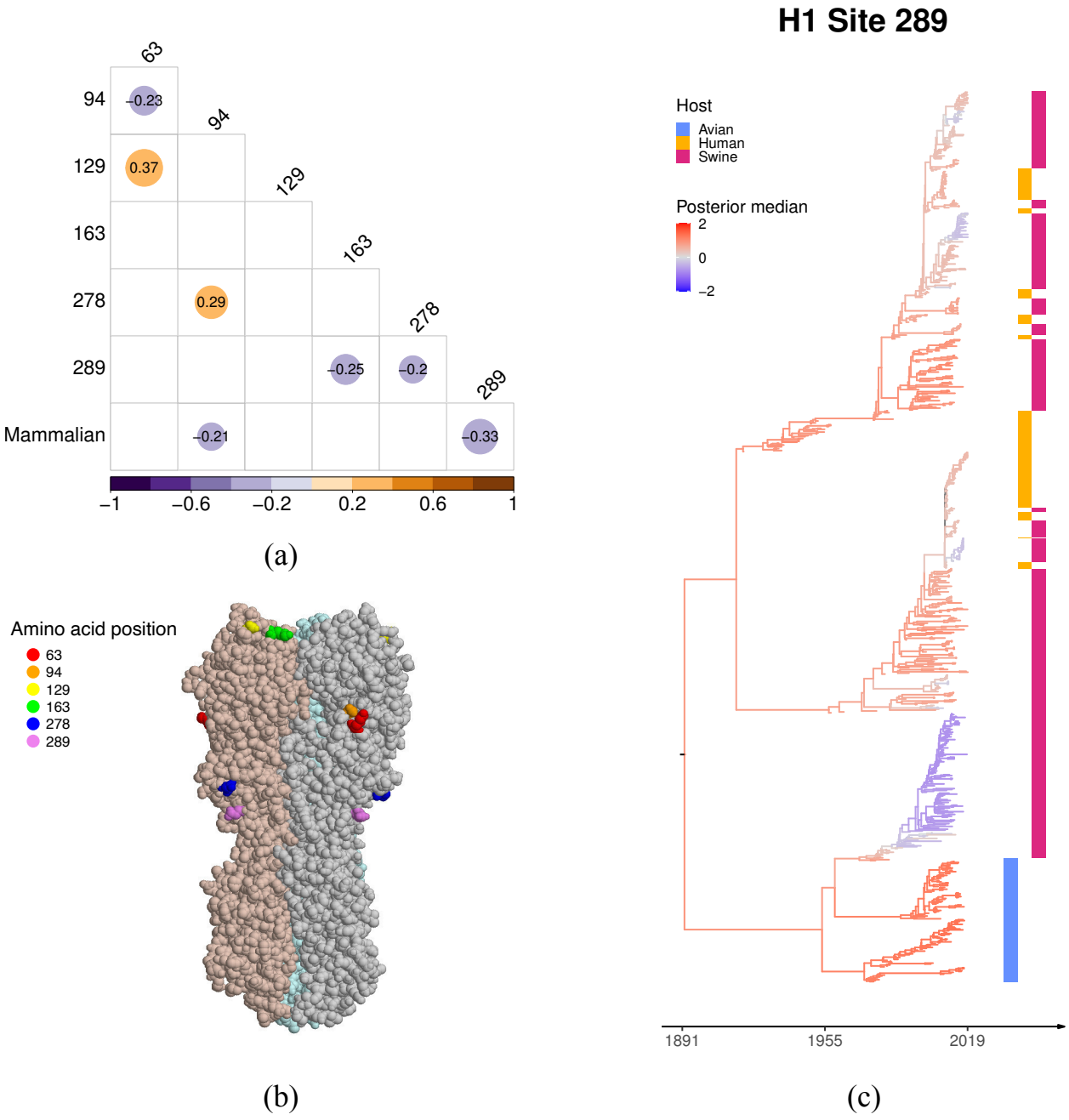


Figure 4.2: (a) Across-trait partial correlation among H1 glycosylation sites and host type with a posterior median > 0.2 or < -0.2 (in color and number). (b) HA structure of a 2009 H1N1 influenza virus (PDB entry 3LZG) with six glycosylation sites highlighted. Site 278 and 289 are in the stalk domain and all others are in the head domain. (c) The maximum clade credibility (MCC) tree with branches colored by the posterior median of the latent variable underlying H1 glycosylation site 289. The heatmap on the right indicates the host type of each taxon.

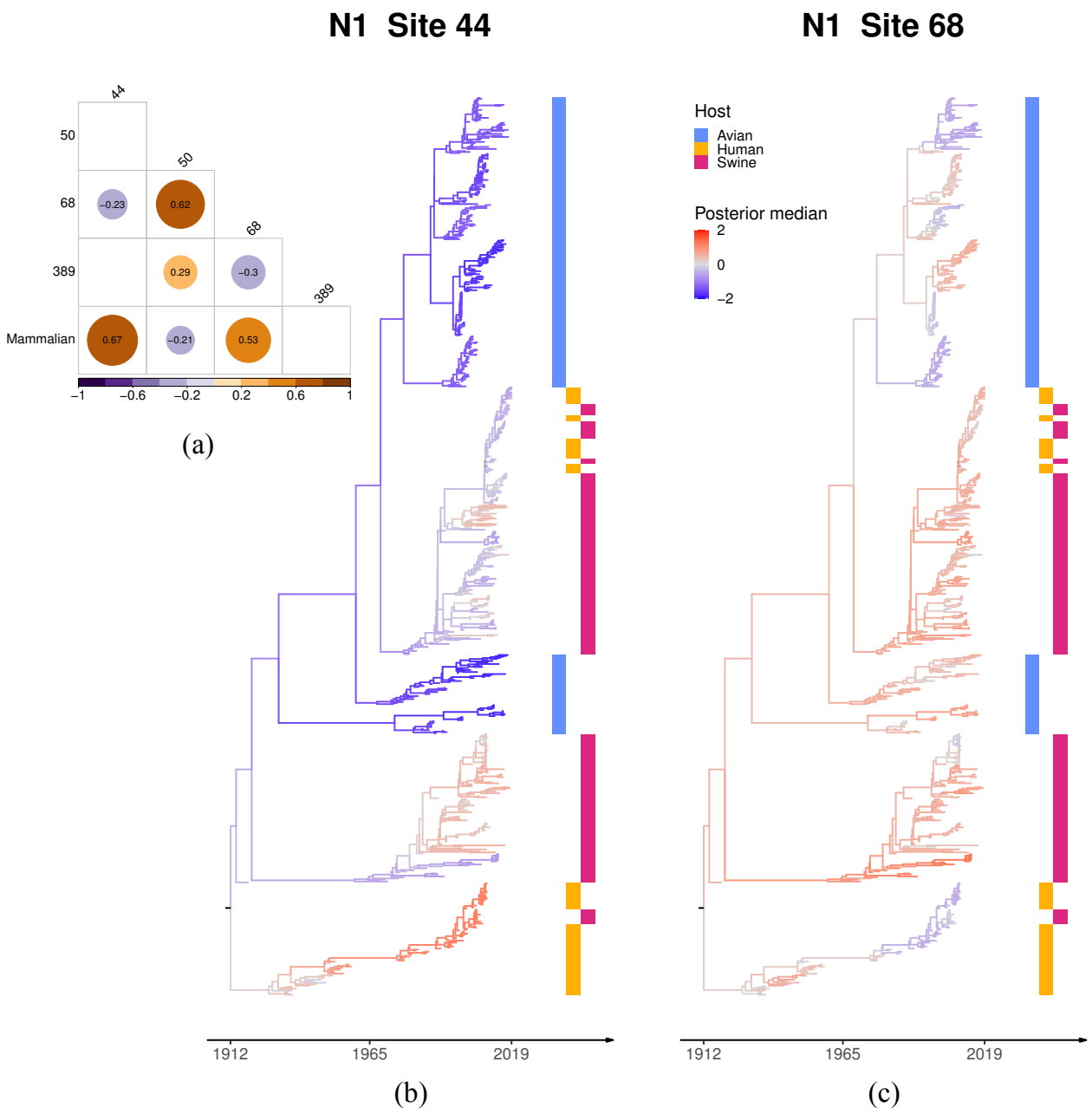


Figure 4.3: (a) Across-trait partial correlation among N1 glycosylation sites and host type with a posterior median > 0.2 or < -0.2 (in color and number). (b)(c) The maximum clade credibility (MCC) tree with branches colored by the posterior median of the latent variable underlying N1 glycosylation site 44 and 68.

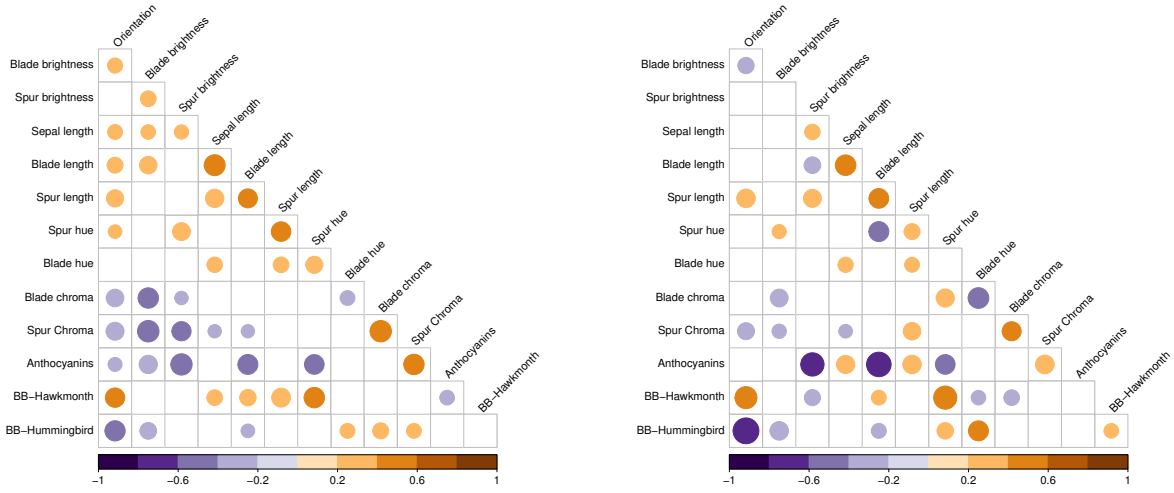
sites 44 and 68. None of the avian lineages has glycosylation site 44 while most swine and some human lineages have it. Similarly, glycosylation at site 68 is present in most swine and human lineages but only in avian lineages circulating in wild birds, not those in poultry.

4.3.4 *Aquilegia* flower and pollinator co-evolution

Reproductive isolation allows two groups of organisms to evolve separately, eventually forming new species. For plants, pollinators play an important role in reproductive isolation (Lowry et al., 2008). We examine the relationship between floral phenotypes and the three main pollinators for the columbine genus *Aquilegia*: bumblebees, hummingbirds, and hawk moths (Whittall and Hodges, 2007). Here, the pollinator species represents a categorical trait with three classes and we choose bumblebee with the shortest tongue as the reference class. Figure 4.4 provides the across-trait correlation and partial correlation. Compared to a similar analysis on the same data set that only looks at correlation or marginal association (Cybis et al., 2015), partial correlation controls confounding and indicates the conditional dependencies between pollinators and floral phenotypes that can bring new insights.

For example, we observe a positive marginal association between hawk moth pollinator and spur length but no conditional dependence between them. The marginal association matches with the observation that flowers with long spur length have pollinators with long tongues (Rosas-Guerrero et al., 2014; Whittall and Hodges, 2007). The absence of a conditional dependence makes intuitive sense because hawk moth's long tongue is not likely to stop them from visiting a flower with short spurs when the other floral traits are held constant. In fact, researchers observe that shortening the nectar spurs does not affect hawk moth visitation (Fulton and Hodges, 1999). Similarly, the positive partial correlation between orientation and hawk moth also finds experimental support. The orientation trait is the angle of flower axis relative to gravity, in the range of (0, 180). A small orientation value implies a pendent flower whereas a large value represents a more upright flower (Hodges et al., 2002). Due to their different morphologies, hawk moths prefer upright flowers while

hummingbirds tend to visit pendent ones. Making the naturally pendent *Aquilegia formosa* flowers upright increases hawk moth visitation (Hodges et al., 2002). These results suggest that partial correlation may have predictive power for results from carefully designed experiments with controlled variables.



(a) Correlation (b) Partial correlation

Figure 4.4: Across-trait correlations and partial correlations with posterior medians > 0.2 or < -0.2 (in color). BB = bumblebee.

4.3.5 MCMC setup and convergence assessment

We run all simulations on a node equipped with AMD EPYC 7642 server processors. For every MCMC run, the minimal effective sample size (ESS) across all dimensions of $\mathbf{X}$ and $\mathbf{P}$ after burn-in is above 100. As another diagnostic, for our two large-scale applications (Section 4.3.3 and 4.3.1) we run three independent chains and confirm the potential scale reduction statistic $\hat{R}$ for all partial correlation elements falls between [1, 1.03], below the common criterion of 1.1 (Gelman, Rubin, et al., 1992). To reach a minimal ESS = 100 across all $\mathbf{P}$ elements, the post burn-in run-time and number of MCMC transition kernels applied for the joint inference are 21 hours and 1.3×10^6 (HIV-1), 113 hours and 7.9×10^7 (H1), 76 hours and 1.4×10^8 (N1). These run-times suggest the difficulty of our large-scale

inference tasks where besides the main challenge of sampling $\{\mathbf{X}, \mathbf{R}, \mathbf{D}\}$, updating the many tree parameters with Metropolis-Hastings transition kernels also takes a large number of iterations.

4.4 Discussion

Learning how different biological traits interact with each other from many evolutionarily related taxa is a long-standing problem of scientific interest that sheds light on various aspects of evolution. Towards this goal, we develop a scalable solution that significantly improves inferential efficiency compared to established state-of-the-art approaches (Cybis et al., 2015; Zhang et al., 2021). Our novel strategy enables learning across-trait conditional dependencies that are more informative than the previous marginal association based analyses. This approach provides reliable estimates of across-trait partial correlations for large problems, on which the established BPS-based method struggles. In two large-scale analyses featuring HIV-1 and H1N1 influenza, the improved efficiency allows us to infer conditional dependencies among traits of scientific interest and therefore investigate some of the most important molecular mechanisms underlying the disease. In addition, our approach incorporates automatic tuning, so that the most influential tuning parameters automatically adapt to the specific challenge the target distribution presents. Finally, we extend the phylogenetic probit model to include categorical traits and illustrate its use in examining the co-evolution of *Aquilegia* flower and pollinators.

We leverage the cutting-edge Zigzag-HMC (Nishimura, Dunson, and Lu, 2020) to tackle the exceedingly difficult computational task of sampling from a high-dimensional truncated normal distribution in the context of the phylogenetic probit model. Zigzag-HMC proves to be more efficient than the previously optimal approach that uses the BPS (Section 4.3.2), especially when combined with differential operator splitting to jointly update two sets of parameters $\mathbf{X}$ and $\boldsymbol{\Omega}$ that are highly correlated. The improved efficiency allows us to obtain

reliable estimates of the conditional dependencies among traits. In our applications, we find that these conditional dependencies better describe trait interactions than do the marginal associations. It is worth mentioning that another closely related sampler, the Markovian zigzag sampler (Bierkens, Fearnhead, Roberts, et al., 2019), or MZZ, may also be appropriate for this task but provides lower efficiency than Zigzag-HMC (Nishimura, Zhang, and Suchard, 2021). While Zigzag-HMC is a recent and less explored version of HMC, BPS and MZZ are two central methods within the piecewise deterministic Markov process literature that have attracted growing interest in recent years (Dunson and Johndrow, 2020; Fearnhead et al., 2018). Intriguingly, the most expensive step of all three samplers is to obtain the log-density gradient, and the same linear-order gradient evaluation method (Zhang et al., 2021) largely speeds it up.

We now consider limitations of this work and the future directions to which they point. First, the phylogenetic probit model does not currently accommodate a directional effect among traits since it only describes pairwise and symmetric correlations. However, the real biological processes are often not symmetric but directional, where it is common that one reaction may trigger another but not the opposite way. A model allowing directed paths is preferable since it better describes the complicated causal network among multiple traits. Graphical models with directed edges (Lauritzen, 1996) are commonly used to learn molecular pathways (Benedetti et al., 2017; Neapolitan, Xue, and Jiang, 2014), but challenges remain to integrate these methods with a large and randomly distributed phylogenetic tree. Toward this goal, one may construct a continuous-time Markov chain to describe how discrete traits evolve (O'Meara, 2012; Pagel, 1994), but with P binary traits the transition rate matrix grows to the astronomical size 2^P. Second, though our method achieves the current best inference efficiency under the phylogenetic probit model, there is still room for improvement. In the influenza glycosylation example, we use a binary trait indicating the host being either avian or mammal (human or swine), instead of setting a categorical trait for host type. In fact, we choose not to use a three-class host type trait because it

causes poor mixing for the partial correlation elements. We suspect two potential reasons for this. First, according to our model assumptions for categorical traits (Equation 4.1), the latent variables underneath the same trait are very negatively correlated, leading to a more correlated and challenging posterior. Second, in our specific data sets, the glycosylation sites tend to be similar in human and swine viruses, further increasing the correlation among posterior dimensions. One potential solution is to de-correlate some latent variables by grouping them into independent factors using phylogenetic factor analysis (Hassler et al., 2021; Tolkoff et al., 2018). Finally, one may consider a logistic or softmax function to map latent variables to the probablity of a discrete trait. This avoids the hard truncations in the probit model but also adds another layer of noise. It requires substantial effort to develop an approach that overcomes the above limitations while supporting efficient inference at the scale of applications in this work.

4.5 Acknowledgments

We thank Kristel Van Laethem for useful discussion about HIV replicative capacity, CD4 counts and viral load. ZZ, PL and MAS are partially supported by National Institutes of Health grant R01 AI153044. MAS and PL acknowledge support from the European Research Council under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 725422 - ReservoirDOCS) and from the Wellcome Trust through project 206298/Z/17/Z (The Artic Network). JLC is supported by the intramural research program of the National Library of Medicine, National Institutes of Health. AH is supported by NIH grant K25AI153816. This work uses computational and storage services provided by the Hoffman2 Shared Cluster through the UCLA Institute for Digital Research and Education's Research Technology Group. The opinions expressed in this article are those of the authors and do not reflect the view of the National Institutes of Health, the Department of Health and Human Services, or the United States government.

4.6 Appendices

Auto-tuning of r. We describe a simple heuristic to auto-tune the step size ratio r on the fly. Let $\boldsymbol{\Sigma}_G$ and $\boldsymbol{\Sigma}_L$ be the covariance matrices for $\boldsymbol{x}_G$ and $\boldsymbol{x}_L$ respectively, then their minimal eigenvalues $\lambda_{\min,G}$ and $\lambda_{\min,L}$ describe the variance magnitude in the most constrained direction. Intuitively, for both HMC and Zigzag-HMC, the step size should be proportional to the diameter of this most constrained density region, which is $\sqrt{\lambda_{\min,G}}$ or $\sqrt{\lambda_{\min,L}}$. Therefore we propose a choice of $r = \frac{\sqrt{\lambda_{\min,L}}}{\sqrt{\lambda_{\min,G}}}$, assuming the two types of momenta lead to similar travel distance during one unit time. It is straightforward to check this assumption. At stationarity, HMC has a velocity $\boldsymbol{v}_G \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, so its velocity along any unit vector $\boldsymbol{u}$ would be distributed as $\langle \boldsymbol{v}_G, \boldsymbol{u} \rangle \sim \mathcal{N}(0, 1)$, and the travel distance $\mathbb{E}|\langle \boldsymbol{v}_G, \mathbf{u} \rangle| = \sqrt{2/\pi}$. For Zigzag-HMC, as $\langle \boldsymbol{v}_L, \boldsymbol{u} \rangle$ does not follow a simple distribution, we estimate $\mathbb{E}|\langle \boldsymbol{v}_L, \mathbf{u} \rangle|$ by Monte Carlo simulation and it turns out to be ≈ 0.8, close to $\sqrt{2/\pi}$.

We test this intuitive choice of r on a subset of the HIV data in Zhang et al. (2021) with 535 taxa, 5 binary and 3 continuous traits. We calculate the optimal $r = \frac{\sqrt{\lambda_{\min,L}}}{\sqrt{\lambda_{\min,G}}} \approx 2.5$ with $\boldsymbol{\Sigma}_G$ and $\boldsymbol{\Sigma}_L$ estimated from the MCMC samples. Clearly, r has a significant impact on the efficiency as a very small or large r leads to lower ESS (Table 4.2). Also, an r in the order of our optimal value generates the best result, so we recommend this on-the-fly automatic tuning $r = \frac{\sqrt{\lambda_{\min,L}}}{\sqrt{\lambda_{\min,G}}}$ (Table 4.2).

Table 4.2: Minimal effective sample size (ESS) per running hour (hr) for partial correlation matrix elements r_{ij} with different r ($N = 535, P_b = 5, P_c = 3$). ESS values report medians across 3 independent simulations.

	ESS/hr	
r	min	median
0.1	32	266
1	106	771
10	**118**	**855**
100	25	110

Zigzag-HMC explores the energy space more efficiently than BPS. In our experience, BPS tends to generate samples with high auto-correlation between their respective energy function evaluations $-\log \pi(\boldsymbol{x})$. In other words, it slowly traverses the target distribution's energy contours even when the marginal dimensions all appear to demonstrate good mixing. A similar behavior has also been reported by Bouchard-Côté, Vollmer, and Doucet, 2018, who introduce a velocity refreshment to address the issue. As we demonstrate below, however, even velocity refreshments cannot fully remedy BPS's slow-mixing on the energy space.

We apply BPS and Zigzag-HMC to a 256-dimensional standard normal truncated to the positive orthant (all $x_i > 0$). We run both samplers for 2000 iterations where per-iteration travel time is one unit time interval and repeat the experiments for 10 times with varying initial values. For BPS we include Poisson velocity refreshments to avoid reducible behavior and set the refreshment rate to an optimal value 1.4 (Bierkens, Kamatani, and Roberts, 2018). At every iteration we refresh Zigzag-HMC's momentum by redrawing it from the marginal Laplace distribution. Both samplers have no problem sampling from the target distribution and the minimal ESS across all dimensions are 158 ± 25 (mean $\pm$ SD) for BPS and 207 ± 21 for Zigzag-HMC, estimated from the last 1000 samples of the MCMC chains across 10 runs. As a sanity check, the average sample mean and variance are $(0.800, 0.365)$ for BPS and $(0.798, 0.363)$ for Zigzag-HMC, close to the analytical values — the univariate marginal distribution of our truncated standard normal is a truncated normal with mean $2/\sqrt{2\pi} \approx 0.798$ and variance $1 - 2/\pi \approx 0.363$ (Cartinhour, 1990).

However, Zigzag-HMC returns a clear win over BPS in the mixing of joint density (Figure 4.5). The sampling inefficiency for $-\log \pi(\boldsymbol{x})$ is less of a problem if one only needs to sample from a truncated normal with a fixed covariance matrix, but we are keenly interested in sampling the covariance matrix as a target of scientific interest. In this context, inefficient traversal across energy contours harms the sampling efficiency for all model parameters (Section 4.3.2).

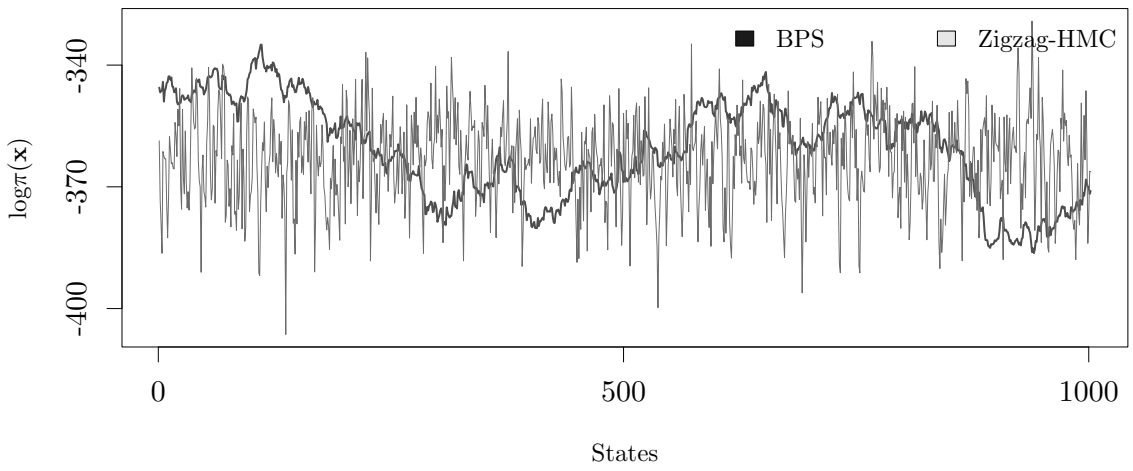


Figure 4.5: Trace plot of the log density of a 256-dimensional truncated standard normal sampled by BPS and Zigzag-HMC for 1000 MCMC iterations.

We can provide an intuition for BPS's slow movement in energy space. Assume the d-dimensional parameter at the tth MCMC iteration is $\boldsymbol{x}(t) = (x_1(t), \ldots, x_d(t)) \in \mathbb{R}^d$, $t = 1, \ldots, T$, with T being the total number of iterations. For a truncated standard normal, its log density $\log \pi(\boldsymbol{x}) \propto \sum_i^d x_i^2$, and a high auto-correlation suggests $\log \pi(\boldsymbol{x})$ changes little between successive iterations, that is, the squared jumping distances

$$J_D = \left[\sum_i^d x_i^2(t+1) - \sum_i^d x_i^2(t) \right]^2, \quad t = 0, \ldots, T-1$$

are small. We then decompose J_D into two components

$$\begin{aligned} J_D &= J_1 + J_2, \\ J_1 &= \sum_i^d \left[x_i^2(t+1) - x_i^2(t) \right]^2, \\ J_2 &= \sum_{j \neq k}^d \left[x_j^2(t+1) - x_j^2(t) \right] \left[x_k^2(t+1) - x_k^2(t) \right], \; t = 0, \ldots, T-1, \end{aligned} \tag{4.18}$$

where J_1 measures the sum of the marginal travel distances and J_2 the covariance among them. We compare J_D, J_1 and J_2 between BPS and Zigzag-HMC in the aforementioned ex-

periments. Clearly seen in Table 4.3, BPS yields a much lower J_D than Zigzag-HMC because its J_2 is largely negative, suggesting strong negative correlation among the coordinates.

Table 4.3: Squared jumping distance (J_D) of $\log \pi(\boldsymbol{x})$ sampled by the bouncy particle sampler (BPS) and Zigzag Hamiltonian Monte Carlo (Zigzag-HMC). We report the empirical mean of J_1 and J_2 in their means and standard deviations (SD) across ten independent simulations with $T = 2000$ samples. Both samplers have a per-iteration travel time 1.

	BPS		Zigzag-HMC	
Quantity	mean	SD	mean	SD
J_D	9	0.4	560	13.9
J_1	558	18.4	564	2.2
J_2	-549	18.3	-4	13.8

CHAPTER 5

Hamiltonian zigzag got more momentum than its Markovian counterpart

5.1 Introduction

MCMC based on continuous-time, non-reversible processes are fundamentally new methods (Fearnhead et al., 2018), among which the two best studied examples are the bouncy particle sampler (Bouchard-Côté, Vollmer, and Doucet, 2018; Deligiannidis, Bouchard-Côté, and Doucet, 2019, BPS) and the Zigzag sampler (Bierkens, Fearnhead, Roberts, et al., 2019; Bierkens, Roberts, Zitt, et al., 2019, ZZ). Following the successful application of BPS in Chapter 3, we naturally consider if ZZ would be more efficient for posterior inference under the phylogenetic probit model.

One known issue with BPS is its near-reducible behavior in the absence of frequent velocity refreshment (Bouchard-Côté, Vollmer, and Doucet, 2018; Fearnhead et al., 2018). In case of a high-dimensional independent normal, BPS achieves optimal performance when velocity refreshment accounts for 78% of all the velocity changes (Bierkens, Kamatani, and Roberts, 2018). However, such frequent velocity refreshment can lead to "random-walk behavior", hurting computational efficiency (Andrieu and Livingstone, 2019; Fearnhead et al., 2018; Neal, 2011). ZZ on the other hand is provably ergodic without velocity refreshment (Bierkens, Fearnhead, Roberts, et al., 2019). In addition to BPS and ZZ, we also explore another Zigzag-HMC sampler that is a version of HMC (Neal, 2011; Nishimura, Dunson, and Lu, 2020). How these samplers perform in practice stands as a critical research area

(Bierkens, Fearnhead, Roberts, et al., 2019; Dunson and Johndrow, 2020; Fearnhead et al., 2018). Early empirical results, while informative, remain limited to low-dimensional examples (Bierkens, Fearnhead, Roberts, et al., 2019; Bierkens, Kamatani, and Roberts, 2018). Deligiannidis et al. (2021) shows the first position and velocity component of BPS particle converges weakly to a randomized Hamiltonian Monte Carlo, but offers little insight about why a randomized HMC appears as the limit.

In this chapter we focus on ZZ and Zigzag-HMC that are based on *Markovian zigzag* (MZZ) and *Hamiltonian zigzag* dynamics (HZZ), respectively. The two types of dynamics both have a "zig-zag" shaped trajectory and this intriguing similarity inspires us to investigate if there is a connection between the two methods. Indeed, we uncover a remarkable connection between MZZ and HZZ — in the limit of increasingly frequent momentum refreshments, HZZ converges strongly to MZZ (Nishimura, Zhang, and Suchard, 2021). This theoretical insight suggests that Zigzag-HMC may outperform ZZ on target distributions with highly correlated parameters. As in Section 5.3, Zigzag-HMC indeed demonstrates superior efficiency on highly-correlated synthetic MTNs as well as an 11,235-dimensional MTN from the phylogenetic application in Zhang et al. (2021).

5.2 Similarity between ZZ and Zigzag-HMC

Details of Zigzag-HMC can be found in Chapter 4 so here we briefly describe the Markovian zigzag process behind ZZ (Bierkens, Fearnhead, Roberts, et al., 2019). Recall that a PDMP specifies the velocity changing event rate, transition at events, and the deterministic dynamics between events (Section 2.4). To facilitate the comparison between two samplers, we adopt the same notation for Zigzag-HMC as in 4.2.2.1. Let $\tau^{(k)}$ be the kth event time and $\left(\boldsymbol{x}\left(\tau^{(0)}\right), \boldsymbol{v}\left(\tau^{(0)}\right)\right)$ with $\boldsymbol{v} \in \{\pm 1\}^d$ is the initial state at time $\tau^{(0)}$. Between $\tau^{(k)}$ and

$\tau^{(k+1)}$, $\boldsymbol{x}$ follows a piecewise linear path

$$\boldsymbol{x}\big(\tau^{(k)}+t\big)=\boldsymbol{x}\big(\tau^{(k)}\big)+t\boldsymbol{v}\big(\tau^{(k)}\big),\quad \boldsymbol{v}\big(\tau^{(k)}+t\big)=\boldsymbol{v}\big(\tau^{(k)}\big),\quad t\in\big[0,\tau^{(k+1)}-\tau^{(k)}\big), \tag{5.1}$$

for $t\geq 0$ until the next velocity switch event which occurs with Poisson rate

$$\lambda_i(\boldsymbol{x},\boldsymbol{v})=[v_i\partial_i U(\boldsymbol{x})]^+:=\max\{0,v_i\partial_i U(\boldsymbol{x})\}. \tag{5.2}$$

Therefore the $(k+1)$th event time is

$$\tau^{(k+1)}=\tau^{(k)}+\min_i t_i,\quad t_i=\min_{t>0}\left\{-\log u_i=\int_0^t\big[v_i(\tau^{(k)})\partial_i U(\boldsymbol{x}(\tau^{(k)})+s\boldsymbol{v}(\tau^{(k)}))\big]^+\,\mathrm{d}s\right\},$$
$$\text{for } u_i\sim\mathrm{Unif}(0,1). \tag{5.3}$$

The event causes an instantaneous velocity sign change at dimension $i^*=\mathrm{argmin}_i\, t_i$ such that

$$v_{i^*}(\tau^{(k+1)})=-v_{i^*}(\tau^{(k)}),\quad v_j(\tau^{(k+1)})=v_j(\tau^{(k)})\ \text{ for }\ j\neq i^*.$$

Then the position $\boldsymbol{x}$ continues its linear path as in (5.1) with the updated velocity.

We now compare the Markovian and Hamiltonian zigzag dynamics. By adding in a velocity term, we could rewrite (4.11) as

$$t_i=\min_{t>0}\left\{|p_i\big(\tau^{(k)}\big)|=\int_0^t v_i(\tau^{(k)})\partial_i U\big[\boldsymbol{x}(\tau^{(k)})+s\boldsymbol{v}(\tau^{(k)})\big]\,\mathrm{d}s\right\}. \tag{5.4}$$

As one may notice, (5.3) and (5.4) appear to be very similar. Aside from the positive part sign ($^+$), the only difference is $-\log u_i$ vs. $|p_i\big(\tau^{(k)}\big)|$ on the left side of the equation within braces. By construction these two variables are equivalent in distribution

$$-\log u_i\overset{d}{=}|p_i\big(\tau^{(k)}\big)|\sim\mathrm{Exp}\,(1)\,. \tag{5.5}$$

In fact, for Hamiltonian zigzag, if we periodically refresh the momentum *magnitude* by redrawing $|p_i(\tau^{(k)})|$ from Exp (1) while keeping its sign, the dynamics converge to Markovian zigzag when refreshment frequency $\to +\infty$. See Nishimura, Zhang, and Suchard (2021) for a rigorous proof.

What does this connection tell us? Here is an intuitive interpretation. The momentum variable of HZZ contains both magnitude and direction where the latter decides its velocity. MZZ has no apparent concept of momentum but only a velocity constrained to $\{\pm 1\}^d$. The equivalence between HZZ and MZZ under infinitely frequent momentum refreshment suggests that MZZ is like HZZ with a "partial" momentum, since HZZ "forgets" its momentum magnitude at the refreshment. Therefore the fully retained momentum information may allow the original HZZ (without momentum refreshment) to travel longer distance and so better explore the parameter space when strong dependency exists. Figure 5.1 visualizes such a case where HZZ has traversed a high-density region while MZZ is still slowly diffusing away from the initial position.

5.3 Two zigzags over multivariate truncated normal

We now compare the performance of ZZ and Zigzag-HMC on a variety of MTNs, where analytical solutions for MZZ and HZZ are available. ZZ is tuning-free in this case, while Zigzag-HMC requires periodic momentum refreshments $p_i \sim \text{Laplace}\,(\text{scale} = 1)$ to ensure ergodicity and the integration time T in-between refreshments remains a user-specified input. Fortunately, the reversible HZZ can take advantage of the no-U-turn algorithm (NUTS) of Hoffman and Gelman (2014) to automatically determine an effective T, and we call the resulting sampler Zigzag-NUTS. This way, we only need to supply a base integration time ΔT and NUTS will then identify an appropriate $T = 2^k \Delta T$, where integer k is the height of the binary searching tree at which the trajectory exhibits a U-turn behavior for the first time. Nishimura, Zhang, and Suchard (2021) provide an empirically optimal choice $\Delta T = 0.1\lambda_{\max}^{1/2}$

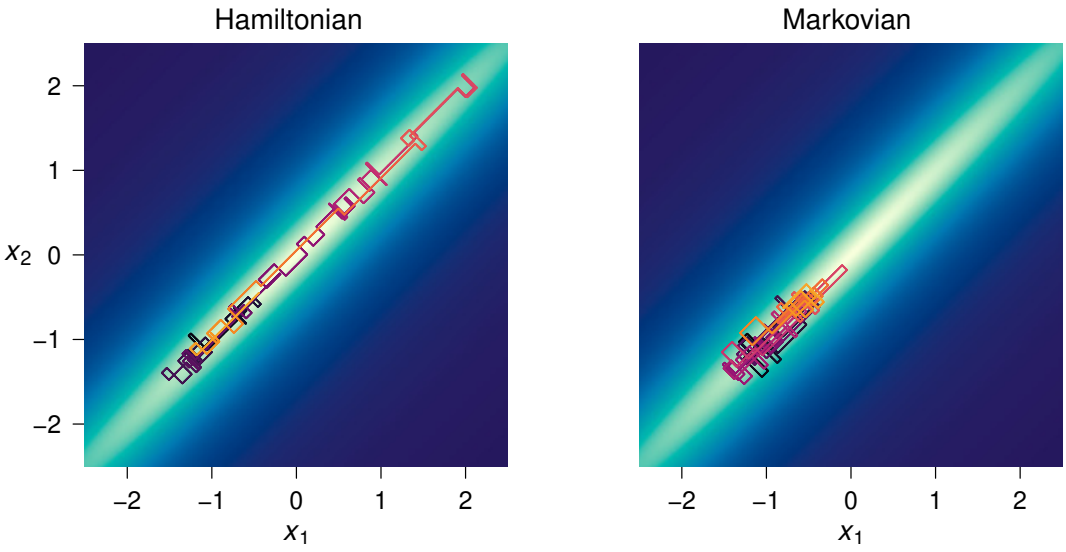


Figure 5.1: Trajectories of the first two position coordinates of Hamiltonian zigzag without momentum refreshment (left) and Markovian zigzag (right). The target is a 1,024-dimensional normal distribution, corresponding to a stationary lag-1 auto-regressive process with auto-correlation 0.99 and unit marginal variances. Both dynamics are simulated for 10^5 linear segments, starting from the same position $x_i = -1$ for all i and same random velocity. The line segment colors change from darkest to lightest as the dynamics evolve.

where $\lambda_{\max}$ is the minimal eigenvalue of the covariance matrix $\mathbf{\Sigma}$. We also include Zigzag-HMC without NUTS in the comparison and choose $T = \sqrt{2}\lambda_{\max}^{1/2}$ as it yields the optimal effective sample size (ESS) in the majority of cases.

We first consider MTN targets with compound symmetric covariance

$$\mathrm{Var}\,(x_i) = 1, \quad \mathbf{\Sigma}\,(x_i, x_j) = \rho \in [0, 1) \;\; \text{for} \;\; i \neq j, \tag{5.6}$$

and all $x_i > 0$ for $i = 1, \ldots, d$. We test the algorithms under $\rho = 0, 0.9, 0.99$ and two dimensions $d = 256$ and $1,024$. Table 5.1 summarizes the results. As x_i are exchangeable, we calculate ESS for the first coordinate and that along the principal eigenvector of $\mathbf{\Sigma}$ since HMC typically struggles most in sampling from the least constrained direction (Neal, 2011). As predicted, Zigzag-HMC demonstrates increasingly superior performance over ZZ as the correlation increases, delivering 4.5 to 4.7-fold gains in relative ESS at $\rho = 0.9$ and 40 to 54-fold gains at $\rho = 0.99$. The efficiency gain is generally greater at the higher dimensions.

Table 5.1: ESS per computing time — relative to that of Markovian zigzag sampler under the compound symmetric MTN targets. We test the algorithms under different dimensions and correlation values. ESS are calculated along the first coordinate and along the principal eigenvector of $\mathbf{\Sigma}$, each shown under the labels "x_1" and "PC".

Compound symmetric	Relative ESS per time				
	$\rho = 0$	$\rho = 0.9$		$\rho = 0.99$	
	x_1	x_1	PC	x_1	PC
Case: $d = 256$					
ZZ	1	1	1	1	1
Zigzag-NUTS	0.64	4.5	4.6	41	40
Zigzag-HMC	5.5	46	66	180	180
Case: $d = 1{,}024$					
Zigzag-NUTS	0.57	4.7	4.5	54	54
Zigzag-HMC	5.6	56	85	300	300

We then consider a real-world 11,235-dimensional MTN target from the Bayesian phylogenetic multivariate probit model (see Chapter 3 for the model and posterior details). On this real-world posterior, HZZ again outperforms its Markovian counterpart with a 6.5-fold increase in the minimum ESS across the coordinates and 19-fold increase in the ESS along the principal eigenvector of the 11,235 dimensional covariance matrix (Table 5.2). Apparently, the joint structure, truncation, and high-dimensionality together make for a complex target, which HZZ can explore more efficiently by virtue of its full momentum.

Table 5.2: Relative ESS per computing time under the phylogenetic probit posterior ($d =$ 11,235).

Phylogenetic probit	Relative ESS per time	
	min	PC
ZZ	1	1
Zigzag-NUTS	6.5	19

CHAPTER 6

hdtg: An R package for high-dimensional truncated normal simulation

6.1 Introduction

Sampling from a multivariate truncated normal (MTN) distribution is a recurring problem in many statistical applications. The MTN distribution of a d-dimensional random vector $\boldsymbol{x} \in \mathbb{R}^d$ has the form

$$\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \text{ with } \boldsymbol{l} \leq \boldsymbol{x} \leq \boldsymbol{u} \text{ bounded}, \tag{6.1}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the mean vector and covariance matrix, and $\boldsymbol{l}, \boldsymbol{u} \in \mathbb{R}^d$ denote the lower and upper truncation bounds. MTNs arise in various context including probit and tobit models (Albert and Chib, 1993; Tobin, 1958), latent Gaussian models (Bolin and Lindgren, 2015), copula regression (Pitt, Chan, and Kohn, 2006), spatial models (Baltagi, Egger, and Kesina, 2018; Tsionas and Michaelides, 2016; Zareifard and Khaledi, 2021), Bayesian metabolic flux analysis (Heinonen et al., 2019), and many others. When the dimension d is small, a standard rejection sampler (Geweke, 1991; Kotecha and Djuric, 1999) works well and is a common choice. However, simulation from a larger MTN with hundreds or thousands of correlated dimensions remains a computational challenge. Work towards this goal include harmonic Hamiltonian Monte Carlo (Pakman and Paninski, 2014, Harmonic-HMC), rejection sampling based on minimax (saddle point) exponential tilting (Botev, 2017, MET), and the most recent Zigzag Hamiltonian Monte Carlo (Nishimura, Dunson, and Lu, 2020; Nishimura, Zhang, and Suchard, 2021, Zigzag-HMC) methods.

The MET method provides independent samples but can suffer from low acceptance rates and becomes impractical with $d > 100$, except in special cases like when the MTN has a strongly positive correlation structure (Botev, 2017). Both Harmonic-HMC and Zigzag-HMC are Markov chain Monte Carlo (MCMC) approaches that generate correlated samples, but can nonetheless be highly efficient and scale to thousands or more dimensions. To our knowledge, however, there is no general-purpose implementation of either method; the `tmg` package provided by Pakman and Paninski (2014) is no longer available on CRAN, and Zhang et al. (2022a) implement Zigzag-HMC for their phylogenetics applications in the specialized software BEAST (Suchard et al., 2018). Therefore, we have developed the `hdtg` R package for efficient MTN simulation. The package implements tuning-free Zigzag-HMC and Harmonic-HMC. We provide performance comparisons among these two methods and a MET implementation from the `TruncatedNormal` package (Botev and Belzile, 2021). In most of the test cases with $d > 100$, Harmonic-HMC and Zigzag-HMC outperform MET. We then conclude with some empirical guidance on which method to use in different scenarios.

6.2 Algorithm

We begin by briefly introducing Harmonic-HMC and Zigzag-HMC, both of which are variants of HMC, an effective proposal generation mechanism exploiting the properties of Hamiltonian dynamics (Neal, 2011). Harmonic-HMC and Zigzag-HMC follow the same general framework. To sample $\boldsymbol{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d$ from the target distribution $\pi(\boldsymbol{x})$, the HMC variants introduce an auxiliary *momentum* variable $\boldsymbol{p}$ and define an augmented target distribution $\pi(\boldsymbol{x}, \boldsymbol{p}) = \pi(\boldsymbol{x})\pi(\boldsymbol{p})$ in the joint space. They then propose the next state by first re-sampling the momentum variable from its marginal and then simulating the solution of Hamiltonian dynamics governed by the differential equations

$$\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t} = \nabla K(\boldsymbol{p}), \quad \frac{\mathrm{d}\boldsymbol{p}}{\mathrm{d}t} = -\nabla U(\boldsymbol{x}), \tag{6.2}$$

where $U(\boldsymbol{x}) = -\log \pi(\boldsymbol{x})$ and $K(\boldsymbol{p}) = -\log \pi(\boldsymbol{p})$ are referred to as *potential* and *kinetic* energies. The dynamics are simulated for a pre-set time duration T and the end state constitutes a valid Metropolis proposal to be accepted or rejected according to the standard formula (Hastings, 1970; Metropolis et al., 1953).

The most common versions of HMC use the momentum distribution $\pi(\boldsymbol{p}) \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and rely on the leapfrog integrator to numerically solve (6.2), as its solutions are analytically intractable in general settings. Harmonic-HMC takes advantage of the fact that (6.2) admits analytical solutions when the target $\pi(\boldsymbol{x})$ is an MTN. The solution follows independent harmonic oscillations along the principal components of the covariance/precision matrix (Pakman and Paninski, 2014); we thus refer to the algorithm as Harmonic-HMC. Truncation boundaries are handled via elastic "bounces" against hard "potential energy walls" (Neal, 2011). We refer interested readers to Pakman and Paninski (2014) for details on Harmonic-HMC.

Zigzag-HMC differs from the common HMC versions in that it deploys a Laplace momentum (Nishimura, Dunson, and Lu, 2020; Nishimura, Zhang, and Suchard, 2021)

$$\pi(\boldsymbol{p}) \propto \prod_i \exp\left(-|p_i|\right), i = 1, \ldots, d. \tag{6.3}$$

The Hamiltonian dynamics then become

$$\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t} = \operatorname{sign}(\boldsymbol{p}), \quad \frac{\mathrm{d}\boldsymbol{p}}{\mathrm{d}t} = -\nabla U(\boldsymbol{x}), \tag{6.4}$$

where $\operatorname{sign}(p_i)$ returns 1 if p_i is positive and -1 otherwise. Because the velocity $\mathrm{d}\boldsymbol{x}/\mathrm{d}t \in \{\pm 1\}^d$ remains constant until one of the p_i flips its sign, the trajectory of these Hamiltonian dynamics has a zigzag pattern, hence the name Zigzag-HMC. The zigzag dynamics also admit analytical solutions under an MTN target and can handle the truncation in the same manner as in Harmonic-HMC. We refer interested readers to Nishimura, Zhang, and Suchard (2021) and Zhang et al. (2022a) for Zigzag-HMC algorithm details, including how to determine the

time of a momentum sign change and of a bounce against truncation boundaries.

The simulation duration T, i.e. how long Hamiltonian dynamics is simulated for each proposal generation, critically affects efficiencies of both Harmonic and Zigzag-HMC. For Harmonic-HMC, Pakman and Paninski (2014) suggest setting $T = \pi/2$; when using this fixed T, however, we observe inefficiencies in some of our examples in Section 6.4 due to Hamiltonian dynamics' periodic behaviors (Neal, 2011). We therefore randomize the duration T, as recommended by Neal (2011), and draw it from a uniform distribution on $[\pi/8, \pi/2]$. For Zigzag-HMC, we adopt the choice $T = \sqrt{2}\lambda_{\text{min}}^{-1/2}$ based on the heuristics of Nishimura, Zhang, and Suchard, 2021, where λ_{min} is the minimal eigenvalue of the precision matrix $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$. We compute λ_{min} using the Lanczos algorithm (Demmel, 1997) as in the `mgcv` package (Wood, 2017). We further implement the no-U-turn algorithm (NUTS) of Hoffman and Gelman (2014) to automatically determine the integration time. With NUTS, we only need to pick a base integration time ΔT which we set to $0.1\lambda_{\text{min}}^{-1/2}$ as recommended by Nishimura, Zhang, and Suchard (2021).

6.3 Using hdtg

The `hdtg` package allows users to draw MCMC samples from an MTN with fixed or random mean and covariance/precision matrix. In our current implementation, Zigzag-HMC accepts the most commonly seen element-wise truncations as in Equation (6.1) while Harmonic-HMC can handle a more general constraint

$$(\mathbf{F}\boldsymbol{x} + \mathbf{g})_i \geq 0, \text{ for } i = 1, \ldots, m. \tag{6.5}$$

Here the $m \times d$ matrix $\mathbf{F}$ and m-dimensional vector $\mathbf{g}$ specify the truncations and $(\cdot)_i$ denotes the ith vector element. As an example, one may use the following code to generate 1,000 samples from a 10-dimensional MTN with zero mean and an identity covariance matrix truncated to the positive orthant:

```
# set the random seed
set.seed(1)

# draw MTN samples using Zigzag-HMC
samplesZHMC <- zigzagHMC(n = 1000, mean = rep(0, 10), prec = diag(10),
                         init = rep(0.1, 10), lowerBounds = rep(0, 10),
                         upperBounds = rep(Inf, 10))

# draw MTN samples using Harmonic-HMC
samplesHHMC <- harmonicHMC(n = 1000, mean = rep(0, 10),
                         choleskyFactor = diag(10), precFlg = TRUE,
                         init = rep(0.1, 10), F = diag(10), g = rep(0, 10))
```

The arguments are:

- `n`: number of samples.
- `mean`: a d-dimensional mean vector.
- `prec`: the precision matrix.
- `init`: a vector of the initial value that must satisfy all constraints.
- `lowerBounds`: a d-dimensional vector specifying the lower bounds.
- `upperBounds`: a d-dimensional vector specifying the upper bounds.
- `choleskyFactor`: upper triangular matrix $\mathbf{U}$ from Cholesky decomposition of precision or covariance matrix into $\mathbf{U}^T\mathbf{U}$.
- `precFlg`: whether `choleskyFactor` is from precision (`TRUE`) or covariance matrix (`FALSE`).
- `F`: the $\mathbf{F}$ matrix.
- `g`: the $\mathbf{g}$ vector.

With a random $\boldsymbol{\mu}$ or $\boldsymbol{\Omega}$, one can simply call `zigzagHMC` or `harmonicHMC` and pass the updated $\boldsymbol{\mu}$ and $\boldsymbol{\Omega}$ as arguments. But a more efficient usage of Zigzag-HMC exists. `zigzagHMC` calls the function `createEngine` (or `createNutsEngine` if using NUTS) to create a C++ object that sets up truncation boundaries and SIMD (single instruction-stream, multiple data-stream) vectorization. Therefore, we can avoid repeated calls of `createEngine` by reusing the C++ object, as in the following example where the 10-dimensional target MTN has a random mean and precision:

```
set.seed(1)
n <- 1000
d <- 10
samples <- array(0, c(n, d))
# initialize MTN mean and precision
m <- rnorm(d, 0, 1)
prec <- rWishart(n = 1, df = d, Sigma = diag(d))[,,1]

# call createEngine once
engine <- createEngine(dimension = d, lowerBounds = rep(0, d),
upperBounds = rep(Inf, d), seed = 1, mean = m, precision = prec)

HZZtime <- sqrt(2) / sqrt(min(mgcv::slanczos(A = prec, k = 1,
kl = 1)[[‘values’]]))

currentSample <- rep(0.1, d)
for (i in 1:n) {
        m <- rnorm(d, 0, 1)
        prec <- rWishart(n = 1, df = d, Sigma = diag(d))[,,1]
        setMean(sexp = engine$engine, mean = m)
        setPrecision(sexp = engine$engine, precision = prec)
        currentSample <- getZigzagSample(position = currentSample,
        nutsFlg = F, engine = engine, stepZZHMC = HZZtime)
        samples[i, ] <- currentSample
}
```

6.4 Efficiency comparison and method choice

To assess the performance of Harmonic-HMC, Zigzag-HMC and MET, we compare them on MTNs with a variety of correlation structures. The three examples are: 1) MTNs with its covariance matrix $\mathbf{\Sigma}$ drawn from the uniform LKJ distribution (Lewandowski, Kurowicka, and Joe, 2009b) as implemented in the `rlkjcorr` function from package `trialr` (Brock, 2020); 2) MTNs with a compound symmetric covariance matrix such that $\mathbf{\Sigma}_{i,i} = 1$ and $\mathbf{\Sigma}_{i,j} = 0.9$ for $i \neq j$; and 3) a real-world MTN that arises as a posterior conditional distribution in a statistical phylogenetics model of HIV evolution (Zhang et al., 2021, 2022a). For simplicity, we assume the truncation $x_i > 0$ for $i = 1, \ldots, d$ in the first two examples. For the HIV example, the truncation is determined by the signs of observed binary biological features.

We now specify our comparison criteria and the rationale behind them. A more efficient MCMC algorithm takes shorter time to achieve a certain effective sample size (ESS). For all three samplers considered, we compare their run-time to obtain the first one or 100 effectively independent samples (t_1 and t_{100}). We include both t_1 and t_{100} because t_{100} reflects a practical run-time for simulation from a fixed MTN and t_1 better captures the pre-processing overhead that remains relevant in cases where $\mathbf{\Sigma}$ is random. Recall that the main pre-processing costs are the Cholesky decomposition of $\mathbf{\Sigma}$ or $\mathbf{\Omega}$ (Harmonic-HMC), calculating the minimal precision matrix eigenvalue $\lambda_{\min}$ (Zigzag-HMC), and solving the minimax optimization problem (MET). Therefore we have

$$\begin{aligned} t_1 &= t_0 + c \\ t_{100} &= t_0 + 100c, \end{aligned} \tag{6.6}$$

where t_0 and c are the pre-processing time required for each $\mathbf{\Sigma}$ update and the average run-time per one effective sample. For simulation from a fixed MTN, t_0 is a one-time cost and so t_{100} serves as a better efficiency criterion. When $\mathbf{\Sigma}$ is random (e.g. the second example in Section 6.3), if $\mathbf{\Sigma}$ changes its value k times, the total run-time to obtain one effective sample

for each $\mathbf{\Sigma}$ is kt_1 and so the t_1 criterion would be more informative.

For Harmonic-HMC and Zigzag-HMC, we estimate the ESS using the `coda` package (Plummer et al., 2006) and define n_1 as the average number of MCMC iterations required for one effectively independent sample. We approximate n_1 by $L/\text{ESS}_{\text{min}}$, where ESS_{min} is the minimal ESS across all dimensions and L is the chain length. We fix $n_1 = 1$ for MET as it generates independent samples. Therefore c in Equation (6.6) equals the average time to complete n_1 iterations after pre-processing. Table 6.1 reports our efficiency comparison in terms of t_1 and t_{100}. We run each test on a quad-core Intel i7 4 GHZ equipped machine with 32GB of memory.

Table 6.1: Efficiency comparison of Harmonic-HMC, Zigzag-HMC, Zigzag-HMC with NUTS (Zigzag-NUTS), and MET sampling approaches across three example correlation structures. We report t_1 and t_{100} (in seconds), the run-time to obtain one or 100 effective samples. In some cases MET takes more than two hours to generate 100 effective samples so the results are not shown. We benchmark each test for three replications and report the average run-time. Bold numbers are column minimums in each test.

		$d = 100$		400		800		1600	
		t_1	t_{100}	t_1	t_{100}	t_1	t_{100}	t_1	t_{100}
LKJ	Harmonic-HMC	**0.004**	**0.34**	**0.17**	**13**	**0.95**	**82**	16	1567
	Zigzag-HMC	0.028	1.8	0.37	20	2.1	136	15	1098
	Zigzag-NUTS	0.029	1.3	0.39	20	1.7	94	**13**	**975**
	MET	4.3	42						
$CS_{0.9}$	Harmonic-HMC	**0.001**	**0.026**	**0.009**	**0.18**	**0.056**	**0.84**	**0.40**	**4.8**
	Zigzag-HMC	0.010	0.63	0.33	29	1.8	147	10	895
	Zigzag-NUTS	0.035	3.2	1.3	129	6.9	689	20	1759
	MET	0.13	0.20	5.1	5.7	39	40	296	302
HIV	Harmonic-HMC	**0.008**	0.74	0.23	20	1.7	137	22	2185
	Zigzag-HMC	0.013	**0.65**	**0.22**	14	**0.98**	**40**	**3.9**	**225**
	Zigzag-NUTS	0.020	1.0	0.30	19	1.3	69	10	626
	MET	0.060	0.084	2.7	**3.5**	22	**40**		

The efficiency of all three methods strongly depends on the correlation structure. MET fails to generate 100 effectively independent samples within two hours in a few higher di-

mensional tests, while Harmonic-HMC and Zigzag-HMC/NUTS enjoy a $t_{100} < 3600$ seconds across all tests. In the LKJ example, Zigzag-HMC/NUTS become more efficient than Harmonic-HMC when d reaches 1600. Zigzag-HMC and Zigzag-NUTS tend to share similar performance. While Zigzag-NUTS is the most efficient choice for the LKJ test ($d = 1600$), Zigzag-HMC wins the test on an MTN from the HIV example ($d = 800, 1600$). On the other hand, when $\mathbf{\Sigma}$ is compound symmetric with a high correlation of 0.9, Harmonic-HMC consistently outperforms the other methods. When MET does function for a target MTN, its t_{100} is close to t_1, as solving the initial minimax optimization problem takes most of its run-time.

In practice, we recommend running a quick efficiency comparison to decide which method to use. Nevertheless we provide some general guidance on method choice for high-dimensional MTN simulation:

- If $d \leq 100$ or the correlation structure is strongly positive, use MET or Harmonic-HMC. Harmonic-HMC may run faster but MET has the advantage of generating independent samples.
- For all other cases, Zigzag-HMC/NUTS is presumably more efficient, although Harmonic-HMC may outperform them when $d < 1000$.
- It is always worth trying MET which is free of MCMC convergence concerns. Since our simulation only examines a few correlation structures, it is possible that MET can handle other large MTNs.

A final point that needs consideration is that Zigzag-HMC/NUTS require $\mathbf{\Omega}$ and if only $\mathbf{\Sigma}$ is available, the method first inverts $\mathbf{\Sigma}$. This is a one-time operation and likely negligible cost when $\mathbf{\Sigma}$ is constant. The approaches does become expensive if $\mathbf{\Sigma}$ is random, as the $\mathcal{O}(d^3)$ inversion is necessary for each value of $\mathbf{\Sigma}$. In practice, statistical models may be parameterized in terms of $\mathbf{\Sigma}$ (Lachaab et al., 2006; Molstad, Hsu, and Sun, 2021) or $\mathbf{\Omega}$

(Baltagi, Egger, and Kesina, 2018; Lehnert et al., 2019; Li, McComick, and Clark, 2020). Harmonic-HMC carries a similar limitation since it requires a $\mathcal{O}(d^3)$ Cholesky decomposition of $\mathbf{\Sigma}$ or $\mathbf{\Omega}$, whichever is provided. Therefore, when d is large and the target MTN has a random correlation structure, one may favor Zigzag-HMC/NUTS over Harmonic-HMC especially if a closed-form $\mathbf{\Omega}$ is at hand.

6.5 Conclusion

This article introduces the `hdtg` package oriented for efficient MTN simulation. In most of our high-dimensional tests the implemented Harmonic-HMC and Zigzag-HMC algorithms outperform the current best approach available in the `TruncatedNormal` package. To our best knowledge, `hdtg` is the first tool that can generate samples from an arbitrary MTN with thousands of dimensions. We discuss the usage of functions and provide practical suggestions on method choice. We expect to see future large-scale statistical applications utilizing the efficiency of `hdtg`.

CHAPTER 7

Discussion

7.1 Achieved research goals

This dissertation aims to develop efficient statistical methods to learn the interaction between complex biological traits from many evolutionarily related taxa. This long-standing problem is of great scientific interest as researchers are often interested in how one trait affects another, as between-trait interactions can shed light on various aspects of evolution across all sorts of organisms such as infectious disease pathogens, animals, and plants. Limited by computational burdens, no previous approach can infer correlation between complex traits with explicitly modeling or adjusting for the phylogenetic tree at the scale of applications in this dissertation. My development mainly focuses on the posterior computation under the phylogenetic probit model (Cybis et al., 2015; Zhang et al., 2021) where the computational bottleneck is to sample from a high-dimensional MTN with a random covariance structure. The advances in Chapter 3 and 4 push the scale limit to at least hundreds of taxa and > 20 traits. While the BPS method in Chapter 3 achieves order-of-magnitude speedup compared to the previous best approach (Cybis et al., 2015), the Zigzag-HMC sampler in Chapter 4 largely outperforms BPS and stands as the current state-of-the-art inference framework for complex trait evolution. I implement all the developed methods in the widely used BEAST software (Suchard et al., 2018) to make them more accessible for researchers. I also create the standalone R package `hdtg` that is the current most efficient tool for sampling from an arbitrary large MTN.

7.2 Advances in methodology

It takes two main ingredients to tackle the MTN simulation challenge — cutting-edge MCMC methods like BPS, Zigzag-HMC, or ZZ, and the novel linear time gradient evaluation method developed in Chapter 3. These samplers well represent recent developments in the MCMC literature (Bierkens, Fearnhead, Roberts, et al., 2019; Bouchard-Côté, Vollmer, and Doucet, 2018; Nishimura, Dunson, and Lu, 2020), covering two of the most promising MCMC designs that are continuous-time, non-reversible based methods (Fearnhead et al., 2018) and Hamiltonian Monte Carlo (Neal, 2011). One essential feature shared by BPS, Zigzag-HMC, and ZZ is that with a MTN target, we can analytically simulate their dynamics so avoid the acceptance-rejection step which hurts efficiency. Nevertheless, without the linear time gradient evaluation, none of these samplers is applicable in the phylogenetic probit model context, as we would have to repeatedly invert a huge covariance matrix whose dimension easily exceeds 10,000. Fortunately, by utilizing the tree traversal strategy in Pybus et al. (2012), I develop a $\mathcal{O}(NP^2)$ gradient evaluation method to circumvent the need for matrix inversion, making all three considered samplers suitable for phylogenetic probit model posterior sampling. As discussed in Chapter 4 and 5, Zigzag-HMC turns out to be the best choice not only because it is more efficient on a strongly correlated large MTN, but also because it allows joint updates of latent variables and across-trait correlation, which is not feasible with BPS and ZZ. In the end, Zigzag-HMC allows us to infer conditional dependencies between traits of scientific interest and therefore investigate some of the most important evolutionary molecular mechanisms.

One advantage of the BPS and Zigzag-HMC methods is they are tuning-free. This is of practical importance as manual tuning MCMC step sizes on high-dimensional targets is nontrivial and a poorly tuned MCMC can cause slow burn-in and poor mixing. I develop a heuristic that uses the maximum and minimal eigenvalues of the target density's covariance matrix to quantify the "diameter" of the most and least constrained direction, and then

choose an appropriate step size proportional to this diameter (Section 3.8.1 and 4.6). The other part of the tuning solution is the *No-U-Turn* algorithm of Hoffman and Gelman (2014) with the primal-dual averaging method (Nesterov, 2009) to decide the base step-size on the fly.

Encouraged by the good performance of Zigzag-HMC, I develop the `hdtg` R package for efficient MTN simulation. MTN simulation is commonly seen in a variety of statistical applications yet there is no implementation that can efficiently sample from a high-dimensional MTN with an arbitrary and potentially random covariance structure. Besides Zigzag-HMC, `hdtg` also implements the harmonic Hamiltonian Monte Carlo (Harmonic-HMC) by Pakman and Paninski (2014), another HMC sampler specialized for sampling MTNs. Harmonic-HMC is not suitable for posterior inference under the phylogenetic probit model as it requires an expensive Cholesky decomposition of the covariance or precision matrix. But without frequent covariance changes, it performs well and in some cases outperforms Zigzag-HMC (Section 6.4). To date, the statistical literature contains few examples involving large MTN simulation, likely due to the lack of a good computational tool, and researchers have developed alternative methods (Chakraborty, Ou, and Dunson, 2021; Souris, Bhattacharya, and Pati, 2018). I hope `hdtg` can become one useful tool for statisticians.

7.3 Scientific insight

The large-scale phylogenetic applications in Chapter 3 and 4 feature two important infectious disease pathogens, HIV-1 and H1N1 influenza. With help from my collaborators who are experts in virus evolution, I interpret the inference results for their scientific meaning. For the HIV-1 application in Chapter 3, I mainly focus on the across-trait correlation which provides information on the complex association between HLA-driven HIV *gag* mutations and virulence (Section 3.4.2). Thanks to the improved efficiency of the Zigzag-HMC method developed in Chapter 4, I can then infer conditional dependencies between traits, which

provide insight into potential causal pathways driven by real biological processes. For example, the conditional dependence in RC-CD4-VL suggests that RC increases VL via reducing CD4, a potential pathway that the marginal correlation fails to identify. I also reveal an intriguing pattern that the conditional dependence between two escape mutations impairing virulence is almost always negative, and if two mutations have opposing effects on virulence, the conditional dependence between them is almost always positive. This illustrates that under selection pressure, the viruses generally require multiple compensatory mutations to restore virulence impaired by immune escape mutations. In the H1N1 application, I reveal how the tendency of having certain H1 and N1 glycosylation sites varies in different hosts. And the conditional dependence between glycosylation sites helps to identify glycosylation pairs that are physically close and may be harmful or beneficial for the virus (Section 4.3.3). The *Aquilegia* flower and pollinator co-evolution application in Section 4.3.4 also highlights that the across-trait conditional dependence may well predict results obtained from carefully designed experiments with controlled variables. In summary, the inference framework with phylogenetic probit model may complement or help in prioritizing experimental testing, and further assist in understanding many important evolutionary processes.

7.4 Limitations and future directions

I now consider the limitations of the methods developed in this dissertation and the future directions to which they point. First, the standard Brownian diffusion assumes that latent parameters at a child node have the same mean with its parent node, so we cannot detect if certain node has a mean "shift" such that all of its descendant nodes tend to have a higher or lower value of the corresponding trait. One way to adjust the Brownian diffusion model is to allow every branch to have its own mean shift but this breaks model identifiability (Gill et al., 2017), as different branch shifts can give the same likelihood. Gill et al. (2017) develop a relaxed directional random walk model (RDRW) that restricts the mean shifts

while retaining identifiability. For a tree triplet, they assume two possible cases — either both child branches inherit the shift from their parent branch, or one child branch gets a new value while the other inherits the parent shift. However, the RDRW model forces shifts to be inherited by at least one of the two child branches. This is more for keeping identifiability and not a natural assumption, and we can imagine cases where two branches take very different shifts. Therefore a model between the arbitrary branch shift model and RDRW would be more biologically realistic.

Second, the phylogenetic probit model only describes pairwise and symmetric correlations, while the real biological processes are often not symmetric but directional. Although there is no phylogenetic approach for learning a more expressive regulation network among traits, methods like graphical models (Lauritzen, 1996) commonly used to learn molecular pathways (Benedetti et al., 2017; Dobra et al., 2004; Neapolitan, Xue, and Jiang, 2014) are worth exploring. Additionally, after obtaining the posterior distribution of across-trait (partial) correlations, I use empirical thresholds like the high posterior density (HPD) interval coverage or posterior median to identify correlations that are more likely from real interactions. In cases where there is a large number of traits and only a few of them are involved in the underlying biology, it is essential to control false positive signals and one would prefer a systematic solution like putting a shrinkage-based prior on the correlations. Extensive simulations would be required to choose a post-inference approach that achieves high sensitivity and low false positive rates.

Finally, in both BEAST and the `hdtg` Package, I implement Zigzag-HMC for MTNs with independent truncations, except that the latent parameters for categorical traits possess truncation intervals depending on multiple dimensions (Section 4.2.1). Despite the fact that independent truncations are the most commonly seen cases for MTNs, a more general MTN class would provide greater modeling flexibility. For example, a MTN subject to linear inequalities applies to Bayesian splines for inferring positive functions (Pakman and Paninski, 2014). In the phylogenetic context, a MTN constrained to a polygon region can be

used to model the geographical distribution of infectious disease pathogens or climate niche information of birds.

Bibliography

Albert, James H and Siddhartha Chib (1993). "Bayesian analysis of binary and polychotomous response data". In: *Journal of the American statistical Association* 88.422, pp. 669–679.

Andrieu, Christophe and Samuel Livingstone (2019). "Peskun-Tierney ordering for Markov chain and process Monte Carlo: beyond the reversible scenario". In: *arXiv preprint arXiv:1906.06197*.

Andrieu, Christophe and Johannes Thoms (2008). "A tutorial on adaptive MCMC". In: *Statistics and Computing* 18.4, pp. 343–373.

Baltagi, Badi H, Peter H Egger, and Michaela Kesina (2018). "Generalized spatial autocorrelation in a panel-probit model with an application to exporting in China". In: *Empirical Economics* 55.1, pp. 193–211.

Barbu, Corentin M et al. (2013). "The effects of city streets on an urban disease vector". In: *PLoS computational biology* 9.1, e1002801.

Benedetti, Elisa et al. (2017). "Network inference from glycoproteomics data reveals new reactions in the IgG glycosylation pathway". In: *Nature communications* 8.1, pp. 1–15.

Bhattacharya, Anirban and David B Dunson (2011). "Sparse Bayesian infinite factor models". In: *Biometrika*, pp. 291–306.

Bierkens, Joris and Andrew Duncan (2017). "Limit theorems for the zig-zag process". In: *Advances in Applied Probability* 49.3, pp. 791–825.

Bierkens, Joris, Paul Fearnhead, Gareth Roberts, et al. (2019). "The zig-zag process and super-efficient sampling for Bayesian analysis of big data". In: *The Annals of Statistics* 47.3, pp. 1288–1320.

Bierkens, Joris, Kengo Kamatani, and Gareth O Roberts (2018). "High-dimensional scaling limits of piecewise deterministic sampling algorithms". In: *arXiv preprint arXiv:1807.11358*.

Bierkens, Joris, Gareth O Roberts, Pierre-André Zitt, et al. (2019). "Ergodicity of the zigzag process". In: *The Annals of Applied Probability* 29.4, pp. 2266–2301.

Bierkens, Joris et al. (2018). "Piecewise deterministic Markov processes for scalable Monte Carlo on restricted domains". In: *Statistics & Probability Letters* 136, pp. 148–154.

Bierkens, Joris et al. (2020). "The boomerang sampler". In: *International conference on machine learning*. PMLR, pp. 908–918.

Blanquart, François et al. (2017). "Viral genetic variation accounts for a third of variability in HIV-1 set-point viral load in Europe". In: *PLoS Biol* 15.6, e2001855.

Bolin, David and Finn Lindgren (2015). "Excursion and contour uncertainty regions for latent Gaussian models". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 77.1, pp. 85–106.

Boni, Maciej F et al. (2013). "Economic epidemiology of avian influenza on smallholder poultry farms". In: *Theoretical population biology* 90, pp. 135–144.

Botev, Zdravko and Leo Belzile (2021). *TruncatedNormal: Truncated Multivariate Normal and Student Distributions*. R package version 2.2.2. URL: https://CRAN.R-project.org/package=TruncatedNormal.

Botev, Zdravko I (2017). "The normal law under linear restrictions: simulation and estimation via minimax tilting". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 79.1, pp. 125–148.

Bouchard-Côté, Alexandre, Sebastian J Vollmer, and Arnaud Doucet (2018). "The bouncy particle sampler: A nonreversible rejection-free Markov chain Monte Carlo method". In: *Journal of the American Statistical Association* 113.522, pp. 855–867.

Brock, Kristian (2020). *trialr: Clinical Trial Designs in 'rstan'*. R package version 0.1.5. URL: https://CRAN.R-project.org/package=trialr.

Cartinhour, Jack (1990). "One-dimensional marginal density functions of a truncated multivariate normal density function". In: *Communications in Statistics-Theory and Methods* 19.1, pp. 197–203.

Cavalli-Sforza, Luigi L and Anthony WF Edwards (1967). "Phylogenetic analysis: models and estimation procedures". In: *Evolution* 21.3, pp. 550–570.

Chakraborty, Antik, Rihui Ou, and David B Dunson (2021). "Bayesian inference on high-dimensional multivariate binary data". In: *arXiv preprint arXiv:2106.02127*.

Chib, Siddhartha and Edward Greenberg (1998). "Analysis of multivariate probit models". In: *Biometrika* 85.2, pp. 347–361.

Claiborne, Daniel T et al. (2015). "Replicative fitness of transmitted HIV-1 drives acute immune activation, proviral load in memory CD4+ T cells, and disease progression". In: *Proceedings of the National Academy of Sciences* 112.12, E1480–E1489.

Clark, James S et al. (2017). "Generalized joint attribute modeling for biodiversity analysis: Median-zero, multivariate, multifarious data". In: *Ecological Monographs* 87.1, pp. 34–56.

Crawford, Hayley et al. (2007). "Compensatory mutation partially restores fitness and delays reversion of escape mutation within the immunodominant HLA-B*5703-restricted Gag epitope in chronic human immunodeficiency virus type 1 infection". In: *J Virol* 81.15, pp. 8346–51.

Cybis, Gabriela B et al. (2015). "Assessing phenotypic correlation through the multivariate phylogenetic latent liability model". In: *Annals of Applied Statistics* 9.2, pp. 969–991.

Daniels, Robert et al. (2003). "N-linked glycans direct the cotranslational folding pathway of influenza hemagglutinin". In: *Molecular cell* 11.1, pp. 79–90.

Datta, Abhirup et al. (2016). "Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets". In: *Journal of the American Statistical Association* 111.514, pp. 800–812.

Davis, Mark HA (1984). "Piecewise-deterministic Markov processes: A general class of non-diffusion stochastic models". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 46.3, pp. 353–376.

Deligiannidis, George, Alexandre Bouchard-Côté, and Arnaud Doucet (2019). "Exponential ergodicity of the bouncy particle sampler". In: *The Annals of Statistics* 47.3, pp. 1268–1287.

Deligiannidis, George et al. (2021). "Randomized Hamiltonian Monte Carlo as scaling limit of the bouncy particle sampler and dimension-free convergence rates". In: *The Annals of Applied Probability* 31.6, pp. 2612–2662.

Demmel, James W (1997). *Applied numerical linear algebra*. SIAM.

Dobra, Adrian et al. (2004). "Sparse graphical models for exploring gene expression data". In: *Journal of Multivariate Analysis* 90.1, pp. 196–212.

Draenert, Rika et al. (2004). "Immune selection for altered antigen processing leads to cytotoxic T lymphocyte escape in chronic HIV-1 infection". In: *J Exp Med* 199.7, pp. 905–15.

Dunson, David B (2000). "Bayesian latent variable models for clustered mixed outcomes". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 62.2, pp. 355–366.

Dunson, David B and JE Johndrow (2020). "The Hastings algorithm at fifty". In: *Biometrika* 107.1, pp. 1–23.

Edwards, Anthony WF and Luigi Luca Cavalli-Sforza (1965). "A method for cluster analysis". In: *Biometrics*, pp. 362–375.

Fearnhead, Paul et al. (2018). "Piecewise deterministic Markov processes for continuous-time Monte Carlo". In: *Statistical Science* 33.3, pp. 386–412.

Fedorov, Valerii, Yuehui Wu, and Rongmei Zhang (2012). "Optimal dose-finding designs with correlated continuous and discrete responses". In: *Statistics in medicine* 31.3, pp. 217–234.

Felsenstein, Joseph (1981). "Evolutionary trees from DNA sequences: a maximum likelihood approach". In: *Journal of Molecular Evolution* 17.6, pp. 368–376.

Felsenstein, Joseph (1985). "Phylogenies and the comparative method". In: *The American Naturalist* 125.1, pp. 1–15.

— (2005). "Using the quantitative genetic threshold model for inferences between and within species". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 360.1459, pp. 1427–1434.

— (2011). "A comparative method for both discrete and continuous characters using the threshold model". In: *The American Naturalist* 179.2, pp. 145–156.

Fisher, Alexander A et al. (2021a). "Relaxed random walks at scale". In: *Systematic Biology* 70.2, pp. 258–267.

Fisher, Alexander A et al. (2021b). "Shrinkage-based random local clocks with scalable inference". In: *arXiv preprint arXiv:2105.07119*.

Fitch, Walter M (1971). "Toward defining the course of evolution: minimum change for a specific tree topology". In: *Systematic Biology* 20.4, pp. 406–416.

Fitch, Walter M and Emanuel Margoliash (1967). "Construction of phylogenetic trees: a method based on mutation distances as estimated from cytochrome c sequences is of general applicability." In: *Science* 155.3760, pp. 279–284.

Fulton, Michelle and Scott A Hodges (1999). "Floral isolation between Aquilegia formosa and Aquilegia pubescens". In: *Proceedings of the Royal Society of London. Series B: Biological Sciences* 266.1435, pp. 2247–2252.

Gelman, Andrew (2006). "Multilevel (hierarchical) modeling: what it can and cannot do". In: *Technometrics* 48.3, pp. 432–435.

Gelman, Andrew, Donald B Rubin, et al. (1992). "Inference from iterative simulation using multiple sequences". In: *Statistical science* 7.4, pp. 457–472.

Gelman, Andrew et al. (2013). *Bayesian data analysis*. Chapman and Hall/CRC.

Geman, Stuart and Donald Geman (1984). "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images". In: *IEEE Transactions on pattern analysis and machine intelligence* 6, pp. 721–741.

Geweke, John (1991). "Efficient simulation from the multivariate normal and student-t distributions subject to linear constraints and the evaluation of constraint probabilities". In: *Computing science and statistics: Proceedings of the 23rd symposium on the interface*. Citeseer, pp. 571–578.

Gill, Mandev S et al. (2017). "A relaxed directional random walk model for phylogenetic trait evolution". In: *Systematic biology* 66.3, pp. 299–319.

Goulder, Philip JR and Bruce D Walker (2012). "HIV and HLA class I: an evolving relationship". In: *Immunity* 37.3, pp. 426–440.

Grafen, Alan (1989). "The phylogenetic regression". In: *Philosophical Transactions of the Royal Society of London. B, Biological Sciences* 326.1233, pp. 119–157.

Hassler, Gabriel et al. (2020). "Inferring phenotypic trait evolution on large trees with many incomplete measurements". In: *Journal of the American Statistical Association*, pp. 1–15.

Hassler, Gabriel W et al. (2021). "Principled, practical, flexible, fast: a new approach to phylogenetic factor analysis". In: *arXiv preprint arXiv:2107.01246*.

Hassler, Gabriel W et al. (2023). "Data Integration in Bayesian Phylogenetics". In: *Annual Review of Statistics and Its Application*.

Hastings, W Keith (1970). "Monte Carlo sampling methods using Markov chains and their applications". In.

Heaton, Matthew J et al. (2019). "A case study competition among methods for analyzing large spatial data". In: *Journal of Agricultural, Biological and Environmental Statistics* 24.3, pp. 398–425.

Hebert, Daniel N et al. (1997). "The number and location of glycans on influenza hemagglutinin determine folding and association with calnexin and calreticulin". In: *The Journal of cell biology* 139.3, pp. 613–623.

Heinonen, Markus et al. (2019). "Bayesian metabolic flux analysis reveals intracellular flux couplings". In: *Bioinformatics* 35.14, pp. i548–i557.

Hemelaar, Joris (2012). "The origin and diversity of the HIV-1 pandemic". In: *Trends Mol Med* 18.3, pp. 182–92.

Hodges, Scott A et al. (2002). "Genetics of floral traits influencing reproductive isolation between Aquilegia formosa and Aquilegia pubescens". In: *The American Naturalist* 159.S3, S51–S60.

Hoffman, Matthew D and Andrew Gelman (2014). "The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo." In: *Journal of Machine Learning Research* 15.1, pp. 1593–1623.

Höhna, Sebastian et al. (2016). "RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language". In: *Systematic biology* 65.4, pp. 726–736.

Huang, Alan, Matthew P Wand, et al. (2013). "Simple marginally noninformative prior distributions for covariance matrices". In: *Bayesian Analysis* 8.2, pp. 439–452.

Huang, Kuan-Hsiang Gary et al. (2011). "Progression to AIDS in South Africa is associated with both reverting and compensatory viral mutations". In: *PloS One* 6.4, e19018.

Irvine, Kathryn M, TJ Rodhouse, and Ilai N Keren (2016). "Extending ordinal regression with a latent zero-augmented beta distribution". In: *Journal of Agricultural, Biological and Environmental Statistics* 21.4, pp. 619–640.

Ives, Anthony R and Theodore Garland (2009). "Phylogenetic logistic regression for binary dependent variables". In: *Systematic Biology* 59.1, pp. 9–26.

Ji, Xiang et al. (2020). "Gradients do grow on trees: a linear-time O(N)-dimensional gradient for statistical phylogenetics". In: *Molecular biology and evolution* 37.10, pp. 3047–3060.

Ji, Xiang et al. (2021). "Scalable Bayesian divergence time estimation with ratio transformations". In: *arXiv preprint arXiv:2110.13298*.

Katzfuss, Matthias (2017). "A multi-resolution approximation for massive spatial datasets". In: *Journal of the American Statistical Association* 112.517, pp. 201–214.

Kingman, John Frank Charles (1982). "The coalescent". In: *Stochastic processes and their applications* 13.3, pp. 235–248.

Kotecha, Jayesh H and Petar M Djuric (1999). "Gibbs sampling approach for generation of truncated multivariate Gaussian random variables". In: *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No. 99CH36258)*. Vol. 3. IEEE, pp. 1757–1760.

Lachaab, Mohamed et al. (2006). "Modeling preference evolution in discrete choice models: A Bayesian state-space approach". In: *Quantitative Marketing and Economics* 4.1, pp. 57–81.

Lauritzen, Steffen L (1996). *Graphical models*. Vol. 17. Clarendon Press.

Lehnert, Judith et al. (2019). "Large-scale Bayesian spatial-temporal regression with application to cardiac MR-perfusion imaging". In: *SIAM Journal on Imaging Sciences* 12.4, pp. 2035–2062.

Leimkuhler, Benedict and Sebastian Reich (2004). *Simulating Hamiltonian dynamics*. 14. Cambridge university press.

Lewandowski, Daniel, Dorota Kurowicka, and Harry Joe (2009a). "Generating random correlation matrices based on vines and extended onion method". In: *Journal of Multivariate Analysis* 100.9, pp. 1989–2001.

— (2009b). "Generating random correlation matrices based on vines and extended onion method". In: *Journal of multivariate analysis* 100.9, pp. 1989–2001.

Lewis, Paul O (2001). "A likelihood approach to estimating phylogeny from discrete morphological character data". In: *Systematic Biology* 50.6, pp. 913–925.

Li, Shuying, Dennis K Pearl, and Hani Doss (2000). "Phylogenetic tree construction using Markov chain Monte Carlo". In: *Journal of the American statistical Association* 95.450, pp. 493–508.

Li, Zehang Richard, Tyler H McComick, and Samuel J Clark (2020). "Using Bayesian latent Gaussian graphical models to infer symptom associations in verbal autopsies". In: *Bayesian analysis* 15.3, p. 781.

Lin, Borong et al. (2020). "Role of protein glycosylation in host-pathogen interaction". In: *Cells* 9.4, p. 1022.

Liu, Jun S, Wing H Wong, and Augustine Kong (1995). "Covariance structure and convergence rate of the Gibbs sampler with various scans". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 57.1, pp. 157–169.

Lowry, David B et al. (2008). "The strength and genetic basis of reproductive isolating barriers in flowering plants". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 363.1506, pp. 3009–3021.

Lyubetsky, Vassily, William H Piel, and Dietmar Quandt (2014). "Current advances in molecular phylogenetics". In: *BioMed research international* 2014.

Ma, Wenjun (2020). "Swine influenza virus: Current status and challenge". In: *Virus research* 288, p. 198118.

Martinez-Picado, Javier et al. (2006). "Fitness cost of escape mutations in p24 *gag* in association with control of human immunodeficiency virus type 1". In: *Journal of Virology* 80.7, pp. 3617–3623.

Mellquist, JL et al. (1998). "The amino acid following an Asn-X-Ser/Thr sequon is an important determinant of N-linked core glycosylation efficiency". In: *Biochemistry* 37.19, pp. 6833–6837.

Metropolis, Nicholas et al. (1953). "Equation of State Calculations by Fast Computing Machines". In: *Journal of Chemical Physics* 21.6, pp. 1087–1092.

Molstad, Aaron J, Li Hsu, and Wei Sun (2021). "Gaussian process regression for survival time prediction with genome-wide gene expression". In: *Biostatistics* 22.1, pp. 164–180.

Mrode, Raphael A (2014). *Linear models for the prediction of animal breeding values*. Cabi.

Murray, Jared S et al. (2013). "Bayesian Gaussian copula factor models for mixed data". In: *Journal of the American Statistical Association* 108.502, pp. 656–665.

Neal, Radford M. (2011). "MCMC using Hamiltonian dynamics". In: *Handbook of Markov Chain Monte Carlo*. Ed. by Steve Brooks et al. Vol. 2. CRC Press New York, NY.

Neapolitan, Richard, Diyang Xue, and Xia Jiang (2014). "Modeling the altered expression levels of genes on signaling pathways in tumors as causal Bayesian networks". In: *Cancer Informatics* 13, CIN–S13578.

Nesterov, Yurii (2009). "Primal-dual subgradient methods for convex problems". In: *Mathematical programming* 120.1, pp. 221–259.

Nishimura, Akihiko, David B Dunson, and Jianfeng Lu (2020). "Discontinuous Hamiltonian Monte Carlo for discrete parameters and discontinuous likelihoods". In: *Biometrika* 107.2, pp. 365–380.

Nishimura, Akihiko, Zhenyu Zhang, and Marc A Suchard (2021). "Hamiltonian zigzag sampler got more momentum than its Markovian counterpart: Equivalence of two zigzags under a momentum refreshment limit". In: *arXiv preprint arXiv:2104.07694*.

Nomura, Shigeru et al. (2013). "Significant reductions in Gag-protease-mediated HIV-1 replication capacity during the course of the epidemic in Japan". In: *Journal of Virology* 87.3, pp. 1465–1476.

Olusola, Babatunde A, David O Olaleye, and Georgina N Odaibo (2020). "Non-synonymous Substitutions in HIV-1 GAG Are Frequent in Epitopes Outside the Functionally Conserved Regions and Associated With Subtype Differences". In: *Front Microbiol* 11, p. 615721.

O'Meara, Brian C (2012). "Evolutionary inferences from phylogenies: a review of methods". In: *Annual Review of Ecology, Evolution, and Systematics* 43, pp. 267–285.

Östbye, Henrik et al. (2020). "N-linked glycan sites on the influenza A virus neuraminidase head domain are required for efficient viral incorporation and replication". In: *Journal of Virology* 94.19, e00874–20.

Pagel, Mark (1994). "Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters". In: *Proceedings of the Royal Society of London. Series B: Biological Sciences* 255.1342, pp. 37–45.

Pakman, Ari and Liam Paninski (2014). "Exact Hamiltonian Monte Carlo for truncated multivariate Gaussians". In: *Journal of Computational and Graphical Statistics* 23.2, pp. 518–542.

Payne, Rebecca et al. (2014). "Impact of HLA-driven HIV adaptation on virulence in populations of high HIV seroprevalence". In: *Proceedings of the National Academy of Sciences* 111.50, E5393–E5400.

Peters, E.A. J. F. and G. de With (2012). "Rejection-free Monte Carlo sampling for general potentials". In: *Physical Review E* 85.2, p. 026703.

Pitt, Michael, David Chan, and Robert Kohn (2006). "Efficient Bayesian inference for Gaussian copula regression models". In: *Biometrika* 93.3, pp. 537–554.

Plummer, Martyn et al. (2006). "CODA: Convergence Diagnosis and Output Analysis for MCMC". In: *R News* 6.1, pp. 7–11. URL: https://journal.r-project.org/archive/.

Pourmohamad, Tony, Herbert KH Lee, et al. (2016). "Multivariate stochastic process models for correlated responses of mixed type". In: *Bayesian Analysis* 11.3, pp. 797–820.

Prince, Jessica L et al. (2012). "Role of transmitted *gag* CTL polymorphisms in defining replicative capacity and early HIV-1 pathogenesis". In: *PLoS Pathogens* 8.11, e1003041.

Pybus, Oliver G et al. (2012). "Unifying the spatial epidemiology and molecular evolution of emerging epidemics". In: *Proceedings of the National Academy of Sciences* 109.37, pp. 15066–15071.

Rannala, Bruce and Ziheng Yang (1996). "Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference". In: *Journal of molecular evolution* 43.3, pp. 304–311.

Ronquist, Fredrik et al. (2012). "MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space". In: *Systematic biology* 61.3, pp. 539–542.

Rosas-Guerrero, Víctor et al. (2014). "A quantitative review of pollination syndromes: do floral traits predict effective pollinators?" In: *Ecology letters* 17.3, pp. 388–400.

Rue, Håvard, Sara Martino, and Nicolas Chopin (2009). "Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations". In: *Journal of the royal statistical society: Series b (statistical methodology)* 71.2, pp. 319–392.

Schliep, Erin M and Jennifer A Hoeting (2013). "Multilevel latent Gaussian process model for mixed discrete and continuous multivariate response data". In: *Journal of Agricultural, Biological, and Environmental Statistics* 18.4, pp. 492–513.

Shahbaba, Babak et al. (2014). "Split Hamiltonian Monte Carlo". In: *Statistics and Computing* 24.3, pp. 339–349.

Sinsheimer, Janet S, James A Lake, and Roderick JA Little (1996). "Bayesian hypothesis testing of four-taxon topologies using molecular sequence data". In: *Biometrics*, pp. 193–210.

Skehel, JJ et al. (1984). "A carbohydrate side chain on hemagglutinins of Hong Kong influenza viruses inhibits recognition by a monoclonal antibody". In: *Proceedings of the National Academy of Sciences* 81.6, pp. 1779–1783.

Sokal, RR and Peter HA Sneath (1963). "Principles of numerical taxonomy". In: *WH Freedman and Company* 359.

Song, Daesub et al. (2008). "Transmission of avian influenza virus (H3N2) to dogs". In: *Emerging infectious diseases* 14.5, p. 741.

Song, Hongshuo et al. (2012). "Impact of immune escape mutations on HIV-1 fitness in the context of the cognate transmitted/founder genome". In: *Retrovirology* 9.1, p. 89.

Souris, Allyson, Anirban Bhattacharya, and Debdeep Pati (2018). "The Soft Multivariate Truncated Normal Distribution with Applications to Bayesian Constrained Estimation". In: *arXiv preprint arXiv:1807.09155*.

Stan Development Team (2018). *Stan Modeling Language Users Guide and Reference Manual, Version 2.18.0*. URL: http://mc-stan.org/.

Strang, Gilbert (1968). "On the construction and comparison of difference schemes". In: *SIAM journal on numerical analysis* 5.3, pp. 506–517.

Suchard, Marc A, Robert E Weiss, and Janet S Sinsheimer (2001). "Bayesian selection of continuous-time Markov chain evolutionary models". In: *Molecular biology and evolution* 18.6, pp. 1001–1013.

Suchard, Marc A et al. (2018). "Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10". In: *Virus Evolution* 4.1, vey016.

Tate, Michelle D et al. (2014). "Playing hide and seek: how glycosylation of the influenza virus hemagglutinin can modulate the immune response to infection". In: *Viruses* 6.3, pp. 1294–1316.

Tobin, James (1958). "Estimation of relationships for limited dependent variables". In: *Econometrica: journal of the Econometric Society*, pp. 24–36.

Tokuda, Tomoki et al. (2011). "Visualizing distributions of covariance matrices". In: *Columbia Univ., New York, USA, Tech. Rep*, pp. 18–18.

Tolkoff, Max R et al. (2018). "Phylogenetic factor analysis". In: *Systematic biology* 67.3, pp. 384–399.

Trovão, Nídia S et al. (2022). "Comparative evolution of the influenza virus A/H1 and A/H3 head and stalk domains across host species". In: *In Preparation*.

Trovão, Nídia S and Martha I Nelson (2020). "When Pigs Fly: Pandemic influenza enters the 21st century". In: *PLoS pathogens* 16.3, e1008259.

Troyer, Ryan M et al. (2009). "Variable fitness impact of HIV-1 escape mutations to cytotoxic T lymphocyte (CTL) response". In: *PLoS Pathog* 5.4, e1000365.

Tsionas, Efthymios G and Panayotis G Michaelides (2016). "A spatial stochastic frontier model with spillovers: Evidence for Italian regions". In: *Scottish Journal of Political Economy* 63.3, pp. 243–257.

Tung Ho, Lam si and Cécile Ané (2014). "A linear-time algorithm for Gaussian and non-Gaussian trait evolution models". In: *Systematic Biology* 63.3, pp. 397–408.

Vitezica, Zulma G, Luis Varona, and Andres Legarra (2013). "On the additive and dominant variance and covariance of individuals within the genomic selection scope". In: *Genetics* 195.4, pp. 1223–1230.

Wang, Hao et al. (2012). "Bayesian graphical lasso models and efficient posterior computation". In: *Bayesian Analysis* 7.4, pp. 867–886.

Webster, Robert G et al. (1992). "Evolution and ecology of influenza A viruses". In: *Microbiological reviews* 56.1, pp. 152–179.

Whittall, Justen B and Scott A Hodges (2007). "Pollinator shifts drive increasingly long nectar spurs in columbine flowers". In: *Nature* 447.7145, pp. 706–709.

Wood, S.N (2017). *Generalized Additive Models: An Introduction with R*. 2nd ed. Chapman and Hall/CRC.

Wright, Jaclyn K et al. (2010). "Gag-protease-mediated replication capacity in HIV-1 subtype C chronic infection: associations with HLA type and clinical parameters". In: *Journal of Virology* 84.20, pp. 10820–10831.

Wright, Jaclyn K et al. (2012). "Impact of HLA-B* 81-associated mutations in HIV-1 *gag* on viral replication capacity". In: *Journal of Virology* 86.6, pp. 3193–3199.

Wright, Sewall (1934). "An analysis of variability in number of digits in an inbred strain of Guinea pigs". In: *Genetics* 19.6, p. 506.

Yang, Ziheng (2006). *Computational molecular evolution*. Oxford University Press.

Zareifard, Hamid and Majid Jafari Khaledi (2021). "A heterogeneous Bayesian regression model for skewed spatial data". In: *Spatial Statistics* 46, p. 100545.

Zhang, Zhenyu et al. (2021). "Large-scale inference of correlation among mixed-type biological traits with phylogenetic multivariate probit models". In: *The Annals of Applied Statistics* 15.1, pp. 230–251.

Zhang, Zhenyu et al. (2022a). "Hamiltonian zigzag accelerates large-scale inference for conditional dependencies between complex biological traits". In: *arXiv preprint arXiv:2201.07291*.

Zhang, Zhenyu et al. (2022b). "hdtg: An R package for high-dimensional truncated normal simulation". In: *under review*.