The Annals of Applied Statistics
2021, Vol. 15, No. 1, 230–251
https://doi.org/10.1214/20-AOAS1394

LARGE-SCALE INFERENCE OF CORRELATION AMONG MIXED-TYPE BIOLOGICAL TRAITS WITH PHYLOGENETIC MULTIVARIATE PROBIT MODELS

BY ZHENYU ZHANG[1], AKIHIKO NISHIMURA[2], PAUL BASTIDE[3], XIANG JI[4], REBECCA P. PAYNE[5], PHILIP GOULDER[6], PHILIPPE LEMEY[7] AND MARC A. SUCHARD[8]

[1]*Department of Biostatistics, Fielding School of Public Health, University of California, Los Angeles, zyz606@ucla.edu*
[2]*Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, akihiko4@gmail.com*
[3]*IMAG, Université de Montpellier, CNRS, paul.bastide@umontpellier.fr*
[4]*Department of Mathematics, School of Science & Engineering, Tulane University, xji4@tulane.edu*
[5]*Translational and Clinical Research Institute, Newcastle University, rebecca.payne2@ncl.ac.uk*
[6]*Department of Paediatrics, University of Oxford, HIV Pathogenesis Programme, Doris Duke Medical Research Institute, University of KwaZulu-Natal, Ragon Institute of MGH, MIT and Harvard University, philip.goulder@paediatrics.ox.ac.uk*
[7]*Department of Microbiology, Immunology and Transplantation, Rega Institute, KU Leuven, philippe.lemey@kuleuven.be*
[8]*Departments of Biomathematics, Biostatistics and Human Genetics, University of California, Los Angeles, msuchard@ucla.edu*

Inferring concerted changes among biological traits along an evolutionary history remains an important yet challenging problem. Besides adjusting for spurious correlation induced from the shared history, the task also requires sufficient flexibility and computational efficiency to incorporate multiple continuous and discrete traits as data size increases. To accomplish this, we jointly model mixed-type traits by assuming latent parameters for binary outcome dimensions at the tips of an unknown tree informed by molecular sequences. This gives rise to a phylogenetic multivariate probit model. With large sample sizes, posterior computation under this model is problematic, as it requires repeated sampling from a high-dimensional truncated normal distribution. Current best practices employ multiple-try rejection sampling that suffers from slow-mixing and a computational cost that scales quadratically in sample size. We develop a new inference approach that exploits: (1) the bouncy particle sampler (BPS) based on piecewise deterministic Markov processes 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. In an application with 535 HIV viruses and 24 traits that necessitates sampling from a 12,840-dimensional truncated normal, our method makes it possible to estimate the across-trait correlation and detect factors that affect the pathogen's capacity to cause disease. This inference framework is also applicable to a broader class of covariance structures beyond comparative biology.

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

Received December 2019; revised July 2020.
Key words and phrases. Bayesian phylogenetics, probit models, bouncy particle sampler, dynamic programming, HIV evolution.

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 (Schliep and Hoeting (2013), Irvine, Rodhouse and Keren (2016), Clark et al. (2017)), optimal design (Fedorov, Wu and Zhang (2012)) and computer experiments (Pourmohamad and Lee (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)).

Alternative approaches for mixed-type traits on unknown trees are limited. Phylogenetic regression models (Grafen (1989)) assume a known fixed tree, and their logistic extensions (Ives and Garland (2010)) 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 methods (Pagel (1994), Lewis (2001)) are popular for multiple binary traits but, restrictively, assume independence between traits given the tree.

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 autocorrelated, 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 (BPS) (Bouchard-Côté, Vollmer and Doucet (2018)) 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.

2. Modeling.

2.1. *Phylogenetic multivariate probit model for mixed-type traits.* Consider N biological taxa, each with P trait measurements. These measurements partition as $\mathbf{Y} = \{y_{ij}\} = [\mathbf{Y^b}, \mathbf{Y^c}]$ 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 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,

$$\mathbf{X}_i \sim \mathcal{N}(\mathbf{X}_{\mathrm{pa}(i)}, t_i \boldsymbol{\Omega}), \tag{2.1}$$

centered at its parent node value $\mathbf{X}_{\mathrm{pa}(i)}$ with across-trait, per-unit-time, $P \times P$ variance matrix $\boldsymbol{\Omega}$ 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

$$y_{ij} = g(x_{ij}) = \begin{cases} \mathrm{sign}(x_{ij}), & j = 1, \ldots P_b, \\ x_{ij}, & j = P_b + 1, \ldots, P, \end{cases} \tag{2.2}$$

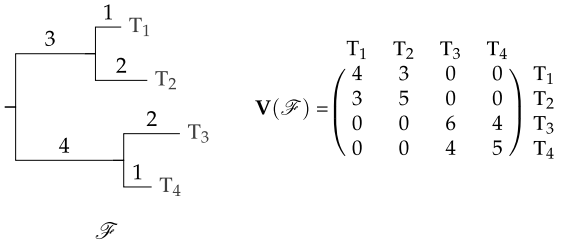

FIG. 1. *A four-taxon phylogenetic tree* $\mathscr{F}$ *with tips* (T_1, T_2, T_3, T_4) *and their corresponding tree diffusion matrix* $\mathbf{V}(\mathscr{F})$.

where $\text{sign}(x_{ij})$ 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, $\mathbf{X}_{2N-1} \sim \mathcal{N}(\boldsymbol{\mu}_0, \tau_0^{-1}\boldsymbol{\Omega})$ with prior mean $\boldsymbol{\mu}_0$ and prior sample size τ_0. Then, $\mathbf{X}$ follows a matrix normal distribution,

$$\mathbf{X} \sim \text{MTN}_{NP}(\mathbf{M}, \boldsymbol{\Upsilon}, \boldsymbol{\Omega}), \tag{2.3}$$

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 $\boldsymbol{\Upsilon} = \mathbf{V}(\mathscr{F}) + \tau_0^{-1}\mathbf{J}$ (Pybus et al. (2012)). The tree diffusion matrix $\mathbf{V}(\mathscr{F})$ is a deterministic function of $\mathscr{F}$, 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 1 illustrates how the tree structure determines $\mathbf{V}(\mathscr{F})$: 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 (2.2) and (2.3) enables us to write down the augmented likelihood of $\mathbf{X}$ and $\mathbf{Y}$ through the factorization

$$p(\mathbf{Y}, \mathbf{X} \mid \boldsymbol{\Upsilon}, \boldsymbol{\Omega}, \boldsymbol{\mu}_0, \tau_0, g) = p(\mathbf{Y} \mid \mathbf{X})\, p(\mathbf{X} \mid \boldsymbol{\Upsilon}, \boldsymbol{\Omega}, \boldsymbol{\mu}_0, \tau_0), \tag{2.4}$$

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.

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 $\boldsymbol{\Omega}^{-1}$ 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 $\boldsymbol{\Omega}_{jj}$, the jth diagonal element of $\boldsymbol{\Omega}$. If we scale them to kx_{ij} and $\boldsymbol{\Omega}_{jj}/k$ by any positive number k, then, according to (2.3), 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 $\boldsymbol{\Omega}^{-1}$ 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 $\boldsymbol{\Omega}$ into an across-trait correlation matrix and standard deviations, with a jointly uniform prior on the correlation matrix. Specifically, we decompose $\boldsymbol{\Omega} = \mathbf{DRD}$, 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 (2009)), with density

$$\text{LKJ}(\mathbf{R}|\eta) = c(\eta)\det(\mathbf{R})^{\eta-1}, \tag{2.5}$$

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 $i = P_b + 1, \ldots, P$ with mean 0 and variance 1 on the log scale. We describe how to carry out the posterior inference under this prior in Section 3.2. There exists other methods for specifying a prior distribution on **DRD**. For example, Huang and Wand (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 nonzero 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 and Wand (2013).

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

$$p(\mathbf{R}, \mathbf{D}, \mathbf{X}, \mathscr{F} \mid \mathbf{Y}, \mathbf{S}) \propto p(\mathbf{Y} \mid \mathbf{X}) \times p(\mathbf{X} \mid \mathbf{R}, \mathbf{D}, \mathscr{F}) \times p(\mathbf{R}, \mathbf{D}) \times p(\mathbf{S} \mid \mathscr{F}) \times p(\mathscr{F}) \tag{3.1}$$

via a random-scan Gibbs scheme (Liu, Wong and Kong (1995)) and drop the posterior's dependence on the hyperparameters $(\boldsymbol{\Upsilon}, \boldsymbol{\mu}_0, \tau_0, g)$ 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 $p(\mathbf{S} \mid \mathscr{F})$ 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 nonconjugate prior on $\boldsymbol{\Omega} = \mathbf{DRD}$.

3.1. *BPS for updating high-dimensional latent parameters.* BPS is a nonreversible "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 toward areas of low probability or, equivalently, of high energy is more likely to be rejected than one toward 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.

3.1.1. *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 interevent times $s_{k+1} = t_{k+1} - t_k$ with rate

$$\lambda\big(\mathbf{x}(t), \mathbf{v}_k\big) = \max\big\{0, \langle \mathbf{v}_k, \nabla U\big(\mathbf{x}(t)\big)\rangle\big\}, \tag{3.2}$$

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 prespecified 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

$$s_{\min} = \operatorname*{argmin}_{s \geq 0} U(\mathbf{x}_{k-1} + s\mathbf{v}_{k-1}) \quad \text{and} \quad U_{\min} = U(\mathbf{x}_{k-1} + s_{\min}\mathbf{v}_{k-1}). \tag{3.3}$$

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

$$U(\mathbf{x}_{k-1} + s_k\mathbf{v}_{k-1}) - U_{\min} = T \quad \text{and} \quad s_k > s_{\min}. \tag{3.4}$$

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

$$\mathbf{x}_k \leftarrow \mathbf{x}_{k-1} + s_k\mathbf{v}_{k-1}, \qquad \mathbf{v}_k \leftarrow \mathbf{v}_{k-1} - 2\frac{\langle \mathbf{v}_{k-1}, \nabla U(\mathbf{x}_k)\rangle}{\|\nabla U(\mathbf{x}_k)\|^2}\nabla U(\mathbf{x}_k). \tag{3.5}$$

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 nonreversible MCMC suggests that adding further events only reduces the efficiency (Bierkens and Duncan (2017), Andrieu and Livingstone (2019)).

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

$$s_{\text{bd},k} = \inf_{s>0}\{\mathbf{x}_{k-1} + s\mathbf{v}_{k-1} \notin D\}. \tag{3.6}$$

Then, the bounce time is given by $s_k = \min\{s_{\text{bd},k}, s_{\text{gr},k}\}$, where $s_{\text{gr},k}$ denotes the gradient event time of (3.4). If $s_{\text{bd},k} < s_{\text{gr},k}$, we have a boundary bounce, and the position is updated as in (3.5) while the velocity is updated as

$$\mathbf{v}_k \leftarrow \mathbf{v}_{k-1} - 2\langle \mathbf{v}_{k-1}, \boldsymbol{\nu}\rangle \boldsymbol{\nu}, \tag{3.7}$$

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

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

$$\mathbf{x} \sim \mathcal{N}(\mathbf{m}, \boldsymbol{\Sigma}) \quad \text{subject to} \quad \mathbf{x} \in D = \{\text{sign}(\mathbf{x}) = \mathbf{y}\} \quad \text{for } \mathbf{y} \in \{\pm 1\}^d. \tag{3.8}$$

Importantly, we can implement BPS so that, aside from 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 $\{\text{sign}(\mathbf{x}) = \mathbf{y}\}$, 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,

$$\begin{gathered} U(\mathbf{x} + s\mathbf{v}) = \frac{1}{2}\langle \mathbf{v}, \boldsymbol{\varphi}_{\mathbf{v}}\rangle s^2 + \langle \mathbf{v}, \boldsymbol{\varphi}_{\mathbf{x}}\rangle s + \frac{1}{2}\langle \mathbf{x} - \mathbf{m}, \boldsymbol{\varphi}_{\mathbf{x}}\rangle + C \\ \text{where } \boldsymbol{\varphi}_{\mathbf{v}} = \boldsymbol{\Phi}\mathbf{v} \text{ and } \boldsymbol{\varphi}_{\mathbf{x}} = \boldsymbol{\Phi}(\mathbf{x} - \mathbf{m}) = \nabla U(\mathbf{x}). \end{gathered} \tag{3.9}$$

The solution to the optimization problem (3.3) is given by

$$\begin{aligned} s_{\min} &= \max\{0, -\langle \mathbf{v}, \boldsymbol{\varphi}_{\mathbf{x}}\rangle / \langle \mathbf{v}, \boldsymbol{\varphi}_{\mathbf{v}}\rangle\}, \\ U_{\min} &= \frac{1}{2}\langle \mathbf{v}, \boldsymbol{\varphi}_{\mathbf{v}}\rangle s_{\min}^2 + \langle \mathbf{v}, \boldsymbol{\varphi}_{\mathbf{x}}\rangle s_{\min} + \frac{1}{2}\langle \mathbf{x} - \mathbf{m}, \boldsymbol{\varphi}_{\mathbf{x}}\rangle + C. \end{aligned} \tag{3.10}$$

It follows from (3.9) that the gradient event time in (3.4) coincides with the larger root of the quadratic equation $as^2 + bs + c = 0$ with

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

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 (3.5) with

$$\nabla U(\mathbf{x} + s\mathbf{v}) = \boldsymbol{\varphi}_{\mathbf{x}+s\mathbf{v}} = \boldsymbol{\Phi}(\mathbf{x} - \mathbf{m}) + s\boldsymbol{\Phi}\mathbf{v} = \boldsymbol{\varphi}_{\mathbf{x}} + s\boldsymbol{\varphi}_{\mathbf{v}}. \tag{3.11}$$

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 $\bigcup_i \{x_i = 0\}$. When $\text{sign}(x_i) = \text{sign}(v_i)$, where x_i and v_i denotes the ith coordinate of particle position and velocity, the particle is moving away from the ith 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

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

When a boundary event takes place, the particle bounces against the plane orthogonal to the standard basis vector $\boldsymbol{v} = \boldsymbol{e}_{i_{\text{bd}}}$. 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

$$\boldsymbol{\varphi}_{\mathbf{v}^*} = \boldsymbol{\Phi}\mathbf{v}^* = \boldsymbol{\varphi}_{\mathbf{v}} + 2v^*_{i_{\text{bd}}}\boldsymbol{\Phi}\boldsymbol{e}_{i_{\text{bd}}} \quad \text{where } v^*_{i_{\text{bd}}} = -v_{i_{\text{bd}}}. \tag{3.12}$$

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 3.1.3) can be reused 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 $p(\mathbf{x}_b \mid \mathbf{x}_c)$ (details in Appendix A.1). We choose the tuning parameter t_{total} based on a heuristic that works well in practice (Section A.2).

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
  $\varphi_{\mathbf{v},\mathbf{x}} \leftarrow \mathbf{v}^\top\boldsymbol{\varphi}_{\mathbf{x}}, \varphi_{\mathbf{v},\mathbf{v}} \leftarrow \mathbf{v}^\top\boldsymbol{\varphi}_{\mathbf{v}}$
  ▷ find gradient event time
  $s_{\min} \leftarrow \max\{0, -\varphi_{\mathbf{v},\mathbf{x}}/\varphi_{\mathbf{v},\mathbf{v}}\}$
  $T \sim \text{Exp}(1)$
  $a \leftarrow \frac{1}{2}\varphi_{\mathbf{v},\mathbf{v}},\ b \leftarrow \varphi_{\mathbf{v},\mathbf{x}},\ c \leftarrow -\frac{1}{2}s^2_{\min}\varphi_{\mathbf{v},\mathbf{v}} - s_{\min}\varphi_{\mathbf{v},\mathbf{x}} - T$
  $s_{\text{gr}} \leftarrow (-b + \sqrt{b^2 - 4ac})/(2a)$
  ▷ find truncation event time at coordinate i
  $s_{\text{bd}} \leftarrow \text{argmin}_i\, x_i/v_i,\ \text{for } i \text{ with } x_i v_i < 0$
  ▷ 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

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 (2.3) and the equivalence between matrix normal and multivariate normal distributions, to sample latent parameters $\mathbf{X}$ from their conditional posterior the target distribution (3.8) specifies as $\mathbf{x} = \mathrm{vec}(\mathbf{X})$, $\mathbf{m} = \mathrm{vec}(\mathbf{M})$, $\boldsymbol{\Sigma} = \boldsymbol{\Omega} \otimes \boldsymbol{\Upsilon}$ and $\mathbf{y} = \mathrm{vec}(\mathbf{Y})$, where $\mathrm{vec}(\cdot)$ 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 such that $\boldsymbol{\Upsilon}^{-1}$ is known, the typical computation proceeds via

$$\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \mathbf{m}) = (\boldsymbol{\Omega}^{-1} \otimes \boldsymbol{\Upsilon}^{-1})(\mathbf{x} - \mathbf{m}) = \mathrm{vec}\big(\boldsymbol{\Upsilon}^{-1}(\mathbf{X} - \mathbf{M})\boldsymbol{\Omega}^{-1}\big), \tag{3.13}$$

with a cost $\mathcal{O}(N^2 P + N P^2)$. 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 $\mathcal{O}(N P^2)$. 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*

$$\boldsymbol{\varphi}_{\mathbf{x}} = \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \mathbf{m}) = \begin{pmatrix} \mathbf{Q}_1(\mathbf{X}_1 - \boldsymbol{\mu}_1) \\ \vdots \\ \mathbf{Q}_N(\mathbf{X}_N - \boldsymbol{\mu}_N) \end{pmatrix}, \tag{3.14}$$

where $\boldsymbol{\mu}_i$ and $\mathbf{Q}_i$ are the mean and the precision matrix of the distributions $p(\mathbf{X}_i \mid \mathbf{X}_{(i)})$ for $i = 1, \ldots, N$, and $p(\mathbf{X}_i \mid \mathbf{X}_{(i)})$ 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 $p(\mathbf{X}_i \mid \mathbf{X}_{(i)})$ are also multivariate normal. Note that

$$\frac{\partial}{\partial \mathbf{x}}\big[\log p(\mathbf{x})\big] = -\frac{1}{2}\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \mathbf{m}). \tag{3.15}$$

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

$$\begin{aligned} \frac{\partial}{\partial \mathbf{X}_i}\big[\log p(\mathbf{x})\big] &= \frac{\partial}{\partial \mathbf{X}_i}\big[\log p(\mathbf{X}_i \mid \mathbf{X}_{(i)}) + \log p(\mathbf{X}_{(i)})\big] \\ &= \frac{\partial}{\partial \mathbf{X}_i}\big[\log p(\mathbf{X}_i \mid \mathbf{X}_{(i)})\big] \\ &= -\frac{1}{2}\mathbf{Q}_i(\mathbf{X}_i - \boldsymbol{\mu}_i). \end{aligned} \tag{3.16}$$

Equating (3.15) and (3.16) 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 $\mathcal{O}(N P^2)$ 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 (3.12) and update $\boldsymbol{\varphi}_{\mathbf{v}^*}$ in $\mathcal{O}(NP)$. For the conditional posterior distribution in our HIV application (Section 4), boundary bounces occur far more frequently than gradient ones, so the efficient update via (3.12) leads to further significant speed-up.

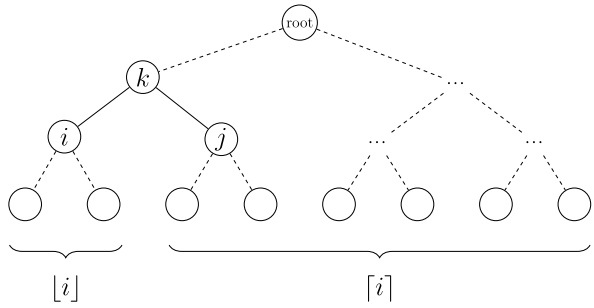

FIG. 2. *A sample tree to illustrate pre- and postorder traversals for efficiently computing* $p(\mathbf{X}_i \mid \mathbf{X}_{(i)})$. *In the triplet* (i, j, k), *parent node* k *has two children* i *and* j. *We group the tip nodes into two disjoint and exhaustive classes*: $\lfloor i \rfloor$ = *tree tips that are descendants to or include node* i, *and* $\lceil i \rceil$ = *tree tips that are not descendants to* i.

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 postorder (child $\rightarrow$ parent) and then again in preorder (parent $\rightarrow$ child) to calculate partial data likelihoods that lead to $\boldsymbol{\mu}_i$ and $\mathbf{Q}_i$. The postorder traversal begins at a tip and ends at the root, while preorder 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 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 postorder traversal the partial likelihoods of the data $\mathbf{X}_{\lfloor i \rfloor}$, given latent $\mathbf{X}_i$, is proportional to a MVN density of $\mathbf{X}_i$ in terms of a postorder mean $\mathbf{m}_i$ and variance $v_i\boldsymbol{\Omega}$ (Pybus et al. (2012)), that is,

$$p(\mathbf{X}_{\lfloor i \rfloor} \mid \mathbf{X}_i) \propto \text{MVN}(\mathbf{X}_i; \mathbf{m}_i, v_i\boldsymbol{\Omega}). \tag{3.17}$$

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

$$\begin{aligned} \mathbf{m}_k &= v_k\big[(v_i + t_i)^{-1}\mathbf{m}_i + (v_j + t_j)^{-1}\mathbf{m}_j\big] \quad \text{with} \\ v_k &= \big[(v_i + t_i)^{-1} + (v_j + t_j)^{-1}\big]^{-1}. \end{aligned} \tag{3.18}$$

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

$$p(\mathbf{X}_i \mid \mathbf{X}_{\lceil i \rceil}) \propto \text{MVN}(\mathbf{X}_i; \boldsymbol{\mu}_i, w_i\boldsymbol{\Omega}), \tag{3.19}$$

in terms of a preorder 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

$$\begin{aligned} \boldsymbol{\mu}_i &= w_i^*\big[(v_j + t_j)^{-1}\mathbf{m}_j + w_k^{-1}\boldsymbol{\mu}_k\big] \quad \text{with} \\ w_i^* &= \big[(v_j + t_j)^{-1} + w_k^{-1}\big]^{-1} \quad \text{and} \\ w_i &= w_i^* + t_i. \end{aligned} \tag{3.20}$$

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 postorder 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 (3.14) remains $\mathcal{O}(NP^2)$.

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)), 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. (2014)).

To introduce the main ideas behind HMC, we denote the distribution of interest by $p(\boldsymbol{\theta}) = p(\mathbf{R}, \mathbf{D} \mid \mathbf{X}, \mathscr{F})$. In order to sample from $\boldsymbol{\theta} = (\mathbf{R}, \mathbf{D})$, HMC introduces an auxiliary *momentum* variable $\boldsymbol{\phi} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and samples from the product density $p(\boldsymbol{\theta}, \boldsymbol{\phi}) = p(\boldsymbol{\theta})p(\boldsymbol{\phi})$. HMC explores the joint space $(\boldsymbol{\theta}, \boldsymbol{\phi})$ by approximating Hamiltonian dynamics that evolve according to the differential equation,

$$\frac{\mathrm{d}\boldsymbol{\theta}}{\mathrm{d}t} = \boldsymbol{\phi}, \qquad \frac{\mathrm{d}\boldsymbol{\phi}}{\mathrm{d}t} = \nabla \log p(\boldsymbol{\theta}). \tag{3.21}$$

More precisely, each HMC iteration proceeds as follows. We first draw a new value of $\boldsymbol{\phi}$ from its marginal distribution, then we approximate the evolution in (3.21) from time $t = 0$ to $t = \tau$ by applying $L = \lfloor \tau/\epsilon \rfloor$ steps of the *leapfrog* update with stepsize ϵ,

$$\boldsymbol{\phi} \leftarrow \boldsymbol{\phi} + \frac{\epsilon}{2}\nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta}), \qquad \boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \epsilon\boldsymbol{\phi}, \qquad \boldsymbol{\phi} \leftarrow \boldsymbol{\phi} + \frac{\epsilon}{2}\nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta}). \tag{3.22}$$

The end point of the approximated dynamics constitutes a valid *Metropolis* proposal (Metropolis et al. (1953)) that is accepted or rejected according to the standard acceptance probability formula.

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. We automate these choices via the stochastic optimization approach of Andrieu and Thoms (2008) and the *No-U-Turn* algorithm of Hoffman and Gelman (2014) that have been shown to achieve performance competitive with manually optimized HMC. 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)).

4. Application on HIV immune escape.

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 subtypes (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.

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's and South Africa's 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 Supplementary Material (Zhang et al. (2021)).

The heat map in Figure 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 four rows that relate to questions addressed by Payne et al. (2014), for example, difference in HLA escape mutations between the two countries and correlation between escape mutations and infection traits (VL and CD4 count) as well as replicative capacity. We identify 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

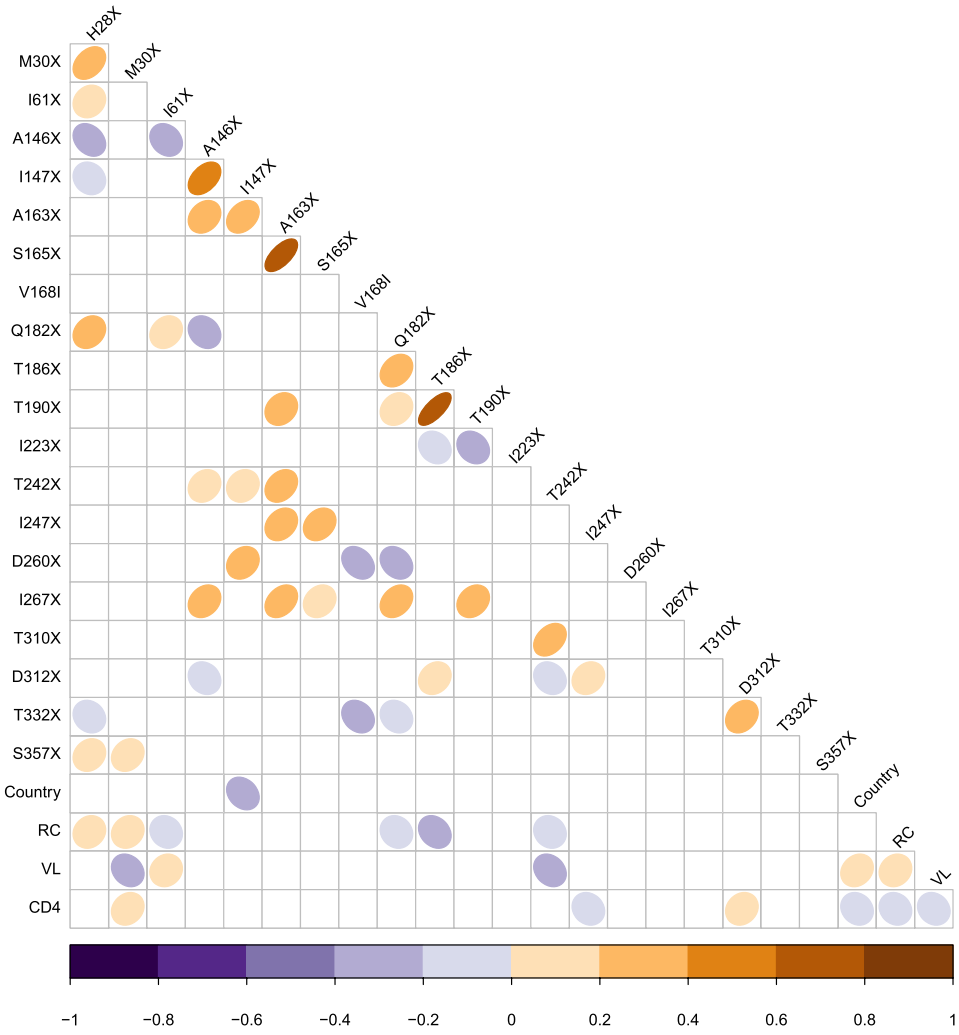

FIG. 3. *Significant across-trait correlation with* $< 10\%$ *posterior tail probability and their posterior mean estimates* (*in color*). *HIV gag mutations are named by the wild type amino acid state, the amino acid site number according to the standard reference genome* (*HXB2*), *and the amino acid "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*.

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) 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) to partly compensate for this impaired RC. The other strong positive correlation

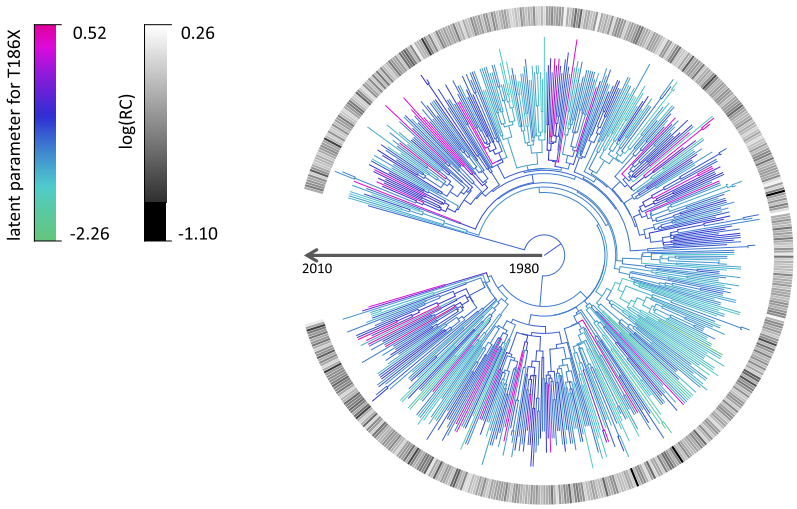

FIG. 4. *The maximum clade credibility tree with branches colored by the posterior mean of the latent parameter corresponding to mutation T*186*X. Outer circle shows log*(*RC*) *in gray scale.*

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)).

4.3. *Tree inference.* Figure 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.

5. Efficiency comparison and goodness-of-fit test.

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 1). This order-of-magnitude improvement

TABLE 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)	20,596 (271)	282 (20)	1468 (11)
Rejection	237 (20)	4707 (25)	3.8 (0.1)	137 (0.7)
Speed-up	23×	4.4×	74×	11×

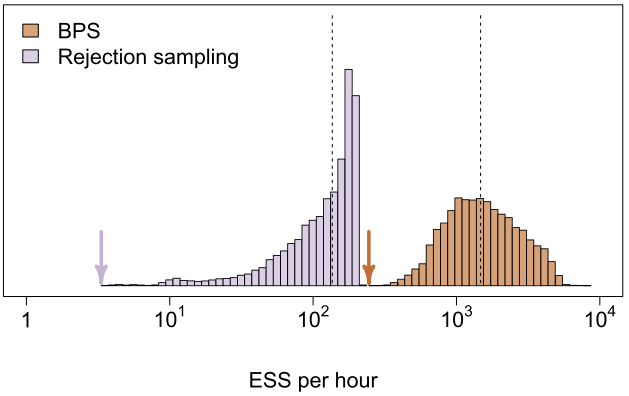

FIG. 5. *A representative histogram of ESS across latent parameters, sampled by BPS or rejection sampling in one hour runtime. Arrows and dashed lines denote the minimum and median ESS* ($N = 535$, $P = 24$).

is more clear in Figure 5. Because rejection sampling only updates one taxon per iteration, some latent parameters rarely change their values (Figure 6). As a result, the minimum ESS of multiple-try rejection sampling is much lower than BPS which simultaneously updates all latent dimensions.

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 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

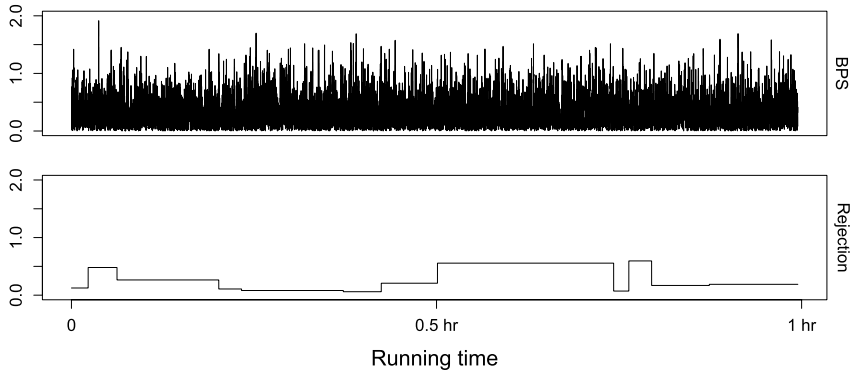

FIG. 6. *Trace plot of the latent parameter with the least ESS by rejection sampling (bottom) and trace plot of the same latent parameter sampled by BPS (top) for an one hour runtime. BPS and rejection sampling run* 1.1×10^4 *and* 2.6×10^5 *iterations, respectively* ($N = 535$, $P = 24$).

TABLE 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)

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 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.

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.1). 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 autocorrelation 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 toward zero. Ideas like the graphical lasso prior (Wang (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 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 (Vitezica, Varona and Legarra (2013), Mrode (2014)) 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.

APPENDIX A: BPS DETAILS

A.1. 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 $p(\mathbf{x}_b \mid \mathbf{x}_c)$. To make progress, we parameterize $p(\mathbf{x}_b \mid \mathbf{x}_c)$ in terms of $p(\mathbf{x})$ with partitioned mean $\mathbf{m} = (\mathbf{m}_b, \mathbf{m}_c)$ and precision matrix

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

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

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

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

$$U_{b|c}(\mathbf{x}_b + t\mathbf{v}_b) = \frac{t^2}{2}\mathbf{v}_b^\intercal \mathbf{\Phi}_{bb}\mathbf{v}_b + t\mathbf{v}_b^\intercal\big[\mathbf{\Phi}_{bb}(\mathbf{x}_b - \mathbf{m}_b) + \mathbf{\Phi}_{bc}(\mathbf{x}_c - \mathbf{m}_c)\big] + C. \tag{A.3}$$

This expression is equivalent to masking out the dimensions of $\mathbf{v}$ in (3.9) that corresponds to $\mathbf{x}_c$ via the vector $\tilde{\mathbf{v}} = (\mathbf{v}_b, \mathbf{0})$. To be explicit, we rewrite (A.3) as

$$U_{b|c}(\mathbf{x}_b + t\mathbf{v}_b) = \frac{t^2}{2}\tilde{\mathbf{v}}^\intercal \mathbf{\Phi}\tilde{\mathbf{v}} + t\tilde{\mathbf{v}}^\intercal \mathbf{\Phi}(\mathbf{x} - \mathbf{m}) + C. \tag{A.4}$$

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 $p(\mathbf{x}_b \mid \mathbf{x}_c)$ without any additional cost.

A.2. 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 autocorrelation 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, suggests that, roughly, the particle moves an average distance of $\sqrt{2/\pi}$ during one unit of time. We so conjecture that there

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

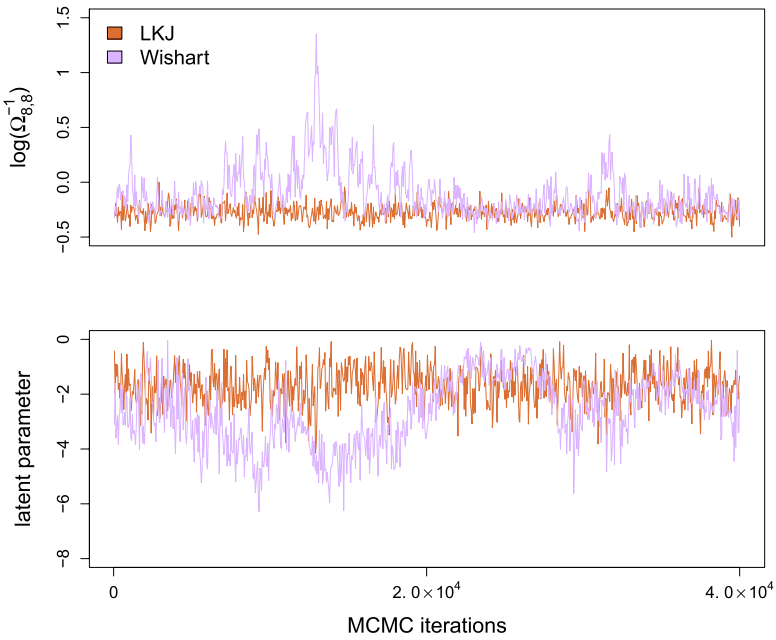

FIG. A.1. *Trace plot of a representative* $\mathbf{\Omega}^{-1}$ *element* (*top*) *in log scale and the latent parameter with the least ESS when assuming a Wishart prior on* $\mathbf{\Omega}^{-1}$ (*bottom*).

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 runtime.

APPENDIX B: 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. (2014)), 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.1).

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).

SUPPLEMENTARY MATERIAL

Data set and source code (DOI: 10.1214/20-AOAS1394SUPP; .zip). We provide the HIV data set and source code to reproduce results in the article.

REFERENCES

ANDRIEU, C. and LIVINGSTONE, S. (2019). Peskun–Tierney ordering for Markov chain and process Monte Carlo: Beyond the reversible scenario. Preprint. Available at arXiv:1906.06197.

ANDRIEU, C. and THOMS, J. (2008). A tutorial on adaptive MCMC. *Stat. Comput.* **18** 343–373. MR2461882 https://doi.org/10.1007/s11222-008-9110-y

BARBU, C. M., HONG, A., MANNE, J. M., SMALL, D. S., CALDERÓN, J. E. Q., SETHURAMAN, K., QUISPE-MACHACA, V., ANCCA-JUÁREZ, J., DEL CARPIO, J. G. C. et al. (2013). The effects of city streets on an urban disease vector. *PLoS Comput. Biol.* **9** e1002801.

BHATTACHARYA, A. and DUNSON, D. B. (2011). Sparse Bayesian infinite factor models. *Biometrika* **98** 291–306. MR2806429 https://doi.org/10.1093/biomet/asr013

BIERKENS, J. and DUNCAN, A. (2017). Limit theorems for the zig-zag process. *Adv. in Appl. Probab.* **49** 791–825. MR3694318 https://doi.org/10.1017/apr.2017.22

BIERKENS, J., BOUCHARD-CÔTÉ, A., DOUCET, A., DUNCAN, A. B., FEARNHEAD, P., LIENART, T., ROBERTS, G. and VOLLMER, S. J. (2018). Piecewise deterministic Markov processes for scalable Monte Carlo on restricted domains. *Statist. Probab. Lett.* **136** 148–154. MR3806858 https://doi.org/10.1016/j.spl.2018.02.021

BOUCHARD-CÔTÉ, A., VOLLMER, S. J. and DOUCET, A. (2018). The bouncy particle sampler: A nonreversible rejection-free Markov chain Monte Carlo method. *J. Amer. Statist. Assoc.* **113** 855–867. MR3832232 https://doi.org/10.1080/01621459.2017.1294075

CHIB, S. and GREENBERG, E. (1998). Analysis of multivariate probit models. *Biometrika* **85** 347–361.

CLARK, J. S., NEMERGUT, D., SEYEDNASROLLAH, B., TURNER, P. J. and ZHANG, S. (2017). Generalized joint attribute modeling for biodiversity analysis: Median-zero, multivariate, multifarious data. *Ecol. Monogr.* **87** 34–56.

CRAWFORD, H., PRADO, J. G., LESLIE, A., HUÉ, S., HONEYBORNE, I., REDDY, S., VAN DER STOK, M., MNCUBE, Z., BRANDER, C. 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. *J. Virol.* **81** 8346–51. https://doi.org/10.1128/JVI.00465-07

CYBIS, G. B., SINSHEIMER, J. S., BEDFORD, T., MATHER, A. E., LEMEY, P. and SUCHARD, M. A. (2015). Assessing phenotypic correlation through the multivariate phylogenetic latent liability model. *Ann. Appl. Stat.* **9** 969–991. MR3371344 https://doi.org/10.1214/15-AOAS821

DATTA, A., BANERJEE, S., FINLEY, A. O. and GELFAND, A. E. (2016). Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets. *J. Amer. Statist. Assoc.* **111** 800–812. MR3538706 https://doi.org/10.1080/01621459.2015.1044091

DRAENERT, R., LE GALL, S., PFAFFEROTT, K. J., LESLIE, A. J., CHETTY, P., BRANDER, C., HOLMES, E. C., CHANG, S.-C., FEENEY, M. E. et al. (2004). Immune selection for altered antigen processing leads to cytotoxic T lymphocyte escape in chronic HIV-1 infection. *J. Exp. Med.* **199** 905–915. https://doi.org/10.1084/jem.20031982

DUNSON, D. B. (2000). Bayesian latent variable models for clustered mixed outcomes. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **62** 355–366. MR1749544 https://doi.org/10.1111/1467-9868.00236

FEDOROV, V., WU, Y. and ZHANG, R. (2012). Optimal dose-finding designs with correlated continuous and discrete responses. *Stat. Med.* **31** 217–234. MR2878850 https://doi.org/10.1002/sim.4388

FELSENSTEIN, J. (1985). Phylogenies and the comparative method. *Amer. Nat.* **125** 1–15.

FELSENSTEIN, J. (2005). Using the quantitative genetic threshold model for inferences between and within species. *Philos. Trans. R. Soc. Lond. B, Biol. Sci.* **360** 1427–1434.

FELSENSTEIN, J. (2011). A comparative method for both discrete and continuous characters using the threshold model. *Amer. Nat.* **179** 145–156.

GELMAN, A. (2006). Multilevel (hierarchical) modeling: What it can and cannot do. *Technometrics* **48** 432–435. MR2252307 https://doi.org/10.1198/004017005000000661

GELMAN, A., RUBIN, D. B. et al. (1992). Inference from iterative simulation using multiple sequences. *Statist. Sci.* **7** 457–472.

GELMAN, A., CARLIN, J. B., STERN, H. S., DUNSON, D. B., VEHTARI, A. and RUBIN, D. B. (2014). *Bayesian Data Analysis*, 3rd ed. *Texts in Statistical Science Series*. CRC Press, Boca Raton, FL. MR3235677

GOULDER, P. J. and WALKER, B. D. (2012). HIV and HLA class I: An evolving relationship. *Immunity* **37** 426–440.

GRAFEN, A. (1989). The phylogenetic regression. *Philos. Trans. R. Soc. Lond. B, Biol. Sci.* **326** 119–157.

HEATON, M. J., DATTA, A., FINLEY, A. O. et al. (2019). A case study competition among methods for analyzing large spatial data. *J. Agric. Biol. Environ. Stat.* **24** 398–425. MR3996451 https://doi.org/10.1007/s13253-018-00348-w

HEMELAAR, J. (2012). The origin and diversity of the HIV-1 pandemic. *Trends Mol. Med.* **18** 182–192. https://doi.org/10.1016/j.molmed.2011.12.001

HOFFMAN, M. D. and GELMAN, A. (2014). The no-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.* **15** 1593–1623. MR3214779

HUANG, A. and WAND, M. P. (2013). Simple marginally noninformative prior distributions for covariance matrices. *Bayesian Anal.* **8** 439–451. MR3066948 https://doi.org/10.1214/13-BA815

HUANG, K.-H. G., GOEDHALS, D., CARLSON, J. M., BROCKMAN, M. A., MISHRA, S., BRUMME, Z. L., HICKLING, S., TANG, C. S., MIURA, T. et al. (2011). Progression to AIDS in South Africa is associated with both reverting and compensatory viral mutations. *PLoS ONE* **6** e19018.

IRVINE, K. M., RODHOUSE, T. J. and KEREN, I. N. (2016). Extending ordinal regression with a latent zero-augmented beta distribution. *J. Agric. Biol. Environ. Stat.* **21** 619–640. MR3576641 https://doi.org/10.1007/s13253-016-0265-2

IVES, A. R. and GARLAND, T. (2010). Phylogenetic logistic regression for binary dependent variables. *Syst. Biol.* **59** 9–26. https://doi.org/10.1093/sysbio/syp074

KATZFUSS, M. (2017). A multi-resolution approximation for massive spatial datasets. *J. Amer. Statist. Assoc.* **112** 201–214. MR3646566 https://doi.org/10.1080/01621459.2015.1123632

KINGMAN, J. F. C. (1982). The coalescent. *Stochastic Process. Appl.* **13** 235–248. MR0671034 https://doi.org/10.1016/0304-4149(82)90011-4

LEWANDOWSKI, D., KUROWICKA, D. and JOE, H. (2009). Generating random correlation matrices based on vines and extended onion method. *J. Multivariate Anal.* **100** 1989–2001. MR2543081 https://doi.org/10.1016/j.jmva.2009.04.008

LEWIS, P. O. (2001). A likelihood approach to estimating phylogeny from discrete morphological character data. *Syst. Biol.* **50** 913–925.

LIU, J. S., WONG, W. H. and KONG, A. (1995). Covariance structure and convergence rate of the Gibbs sampler with various scans. *J. Roy. Statist. Soc. Ser. B* **57** 157–169. MR1325382

MARTINEZ-PICADO, J., PRADO, J. G., FRY, E. E., PFAFFEROTT, K., LESLIE, A., CHETTY, S., THOBAKGALE, C., HONEYBORNE, I., CRAWFORD, H. et al. (2006). Fitness cost of escape mutations in p24 *gag* in association with control of human immunodeficiency virus type 1. *J. Virol.* **80** 3617–3623.

METROPOLIS, N., ROSENBLUTH, A. W., ROSENBLUTH, M. N., TELLER, A. H. and TELLER, E. (1953). Equation of state calculations by fast computing machines. *J. Chem. Phys.* **21** 1087–1092.

MRODE, R. A. (2014). *Linear Models for the Prediction of Animal Breeding Values*. Cabi.

MURRAY, J. S., DUNSON, D. B., CARIN, L. and LUCAS, J. E. (2013). Bayesian Gaussian copula factor models for mixed data. *J. Amer. Statist. Assoc.* **108** 656–665. MR3174649 https://doi.org/10.1080/01621459.2012.762328

NEAL, R. M. (2011). MCMC using Hamiltonian dynamics. In *Handbook of Markov Chain Monte Carlo. Chapman & Hall/CRC Handb. Mod. Stat. Methods* 113–162. CRC Press, Boca Raton, FL. MR2858447

NOMURA, S., HOSOYA, N., BRUMME, Z. L., BROCKMAN, M. A., KIKUCHI, T., KOGA, M., NAKAMURA, H., KOIBUCHI, T., FUJII, T. et al. (2013). Significant reductions in Gag-protease-mediated HIV-1 replication capacity during the course of the epidemic in Japan. *J. Virol.* **87** 1465–1476.

PAGEL, M. (1994). Detecting correlated evolution on phylogenies: A general method for the comparative analysis of discrete characters. *Proc. R. Soc. Lond., B Biol. Sci.* **255** 37–45.

PAYNE, R., MUENCHHOFF, M., MANN, J., ROBERTS, H. E., MATTHEWS, P., ADLAND, E., HEMPENSTALL, A., HUANG, K.-H., BROCKMAN, M. et al. (2014). Impact of HLA-driven HIV adaptation on virulence in populations of high HIV seroprevalence. *Proc. Natl. Acad. Sci. USA* **111** E5393–E5400.

PETERS, E. A. J. F. and DE WITH, G. (2012). Rejection-free Monte Carlo sampling for general potentials. *Phys. Rev. E* **85** 026703.

POURMOHAMAD, T. and LEE, H. K. H. (2016). Multivariate stochastic process models for correlated responses of mixed type. *Bayesian Anal.* **11** 797–820. MR3498046 https://doi.org/10.1214/15-BA976

PRINCE, J. L., CLAIBORNE, D. T., CARLSON, J. M., SCHAEFER, M., YU, T., LAHKI, S., PRENTICE, H. A., YUE, L., VISHWANATHAN, S. A. et al. (2012). Role of transmitted *gag* CTL polymorphisms in defining replicative capacity and early HIV-1 pathogenesis. *PLoS Pathog.* **8** e1003041.

PYBUS, O. G., SUCHARD, M. A., LEMEY, P., BERNARDIN, F. J., RAMBAUT, A., CRAWFORD, F. W., GRAY, R. R., ARINAMINPATHY, N., STRAMER, S. L. et al. (2012). Unifying the spatial epidemiology and molecular evolution of emerging epidemics. *Proc. Natl. Acad. Sci. USA* **109** 15066–15071.

RUE, H., MARTINO, S. and CHOPIN, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **71** 319–392. MR2649602 https://doi.org/10.1111/j.1467-9868.2008.00700.x

SCHLIEP, E. M. and HOETING, J. A. (2013). Multilevel latent Gaussian process model for mixed discrete and continuous multivariate response data. *J. Agric. Biol. Environ. Stat.* **18** 492–513. MR3142597 https://doi.org/10.1007/s13253-013-0136-z

SONG, H., PAVLICEK, J. W., CAI, F., BHATTACHARYA, T., LI, H., IYER, S. S., BAR, K. J., DECKER, J. M., GOONETILLEKE, N. et al. (2012). Impact of immune escape mutations on HIV-1 fitness in the context of the cognate transmitted/founder genome. *Retrovirology* **9** 89.

STAN DEVELOPMENT TEAM (2018). Stan Modeling Language Users Guide and Reference Manual, Version 2.18.0.

SUCHARD, M. A., WEISS, R. E. and SINSHEIMER, J. S. (2001). Bayesian selection of continuous-time Markov chain evolutionary models. *Mol. Biol. Evol.* **18** 1001–1013.

SUCHARD, M. A., LEMEY, P., BAELE, G., AYRES, D. L., DRUMMOND, A. J. and RAMBAUT, A. (2018). Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol.* **4** vey016. https://doi.org/10.1093/ve/vey016

TOKUDA, T., GOODRICH, B., VAN MECHELEN, I., GELMAN, A. and TUERLINCKX, F. (2011). Visualizing distributions of covariance matrices. Tech. Rep., 18–18, Columbia Univ., New York, USA.

TROYER, R. M., MCNEVIN, J., LIU, Y., ZHANG, S. C., KRIZAN, R. W., ABRAHA, A., TEBIT, D. M., ZHAO, H., AVILA, S. et al. (2009). Variable fitness impact of HIV-1 escape mutations to cytotoxic T lymphocyte (CTL) response. *PLoS Pathog.* **5** e1000365. https://doi.org/10.1371/journal.ppat.1000365

TUNG HO, L. S. and ANÉ, C. (2014). A linear-time algorithm for Gaussian and non-Gaussian trait evolution models. *Syst. Biol.* **63** 397–408.

VITEZICA, Z. G., VARONA, L. and LEGARRA, A. (2013). On the additive and dominant variance and covariance of individuals within the genomic selection scope. *Genetics* **195** 1223–1230.

WANG, H. (2012). Bayesian graphical lasso models and efficient posterior computation. *Bayesian Anal.* **7** 867–886. MR3000017 https://doi.org/10.1214/12-BA729

WRIGHT, S. (1934). An analysis of variability in number of digits in an inbred strain of Guinea pigs. *Genetics* **19** 506.

WRIGHT, J. K., BRUMME, Z. L., CARLSON, J. M., HECKERMAN, D., KADIE, C. M., BRUMME, C. J., WANG, B., LOSINA, E., MIURA, T. et al. (2010). Gag-protease-mediated replication capacity in HIV-1 subtype C chronic infection: Associations with HLA type and clinical parameters. *J. Virol.* **84** 10820–10831.

WRIGHT, J. K., NAIDOO, V. L., BRUMME, Z. L., PRINCE, J. L., CLAIBORNE, D. T., GOULDER, P. J., BROCKMAN, M. A., HUNTER, E. and NDUNG'U, T. (2012). Impact of HLA-B* 81-associated mutations in HIV-1 *gag* on viral replication capacity. *J. Virol.* **86** 3193–3199.

ZHANG, Z., NISHIMURA, A., BASTIDE, P., JI, X., PAYNE, R. P., GOULDER, P., LEMEY, P. and SUCHARD, M. A. (2021). Supplement to "Large-scale inference of correlation among mixed-type biological traits with phylogenetic multivariate probit models." https://doi.org/10.1214/20-AOAS1394SUPP