Supplemental Information

1 Post-order Traversal Likelihood Calculations

We seek to compute $p(\mathbf{Y} \,|\, \mathbf{L}, \boldsymbol{\Lambda}, \mathcal{F})$ in $\mathcal{O}(NPK^2 + NK^3)$ by adapting the methods developed by Bastide et al. (2018), Mitov et al. (2020) and Hassler et al. (2020). Let $\mathbf{Y}^{\text{obs}} = \left(\mathbf{y}_1^{\text{obs}}, \ldots, \mathbf{y}_N^{\text{obs}}\right)^t$ be the $N \times P$ matrix of observed data, where all missing measurements in $\mathbf{Y}$ have been replaced with 0's. This post-order algorithm requires that one can compute the partial mean $\mathbf{m}_i$, precision $\mathbf{P}_i$ and remainder r_i such that

$$\begin{aligned} p\left(\mathbf{y}_i^{\text{obs}} \,\middle|\, \mathbf{f}_i, \mathbf{L}, \boldsymbol{\Lambda}\right) &= r_i \hat{\theta}(\mathbf{f}_i; \mathbf{m}_i, \mathbf{P}_i)\,, \text{ where} \\ \hat{\theta}(x; \boldsymbol{\mu}, \mathbf{P}) &= (2\pi)^{-\text{rank}(\mathbf{P})/2}\, \hat{\det}(\mathbf{P})^{1/2} \exp\left(-\frac{1}{2}\,(x - \boldsymbol{\mu})^t\, \mathbf{P}\,(x - \boldsymbol{\mu})\right), \end{aligned} \tag{1}$$

rank($\mathbf{P}$) is the number of non-zero singular values of $\mathbf{P}$ and $\hat{\det}(\mathbf{P})$ is the product of the non-zero singular values of $\mathbf{P}$. We also define the indicator matrices $\boldsymbol{\delta}_i = \text{diag}[\delta_{i1}, \ldots, \delta_{iP}]$ where $\delta_{ij} = 1$ if y_{ij} is observed and $\delta_{ij} = 0$ if it is missing. Finally, we define $P_i^{\text{obs}} = \sum_{j=1}^{P} \delta_{ij}$ as the number of observed traits for taxon i.

In the context of PFA, we calculate

$$\begin{aligned} \log p\left(\mathbf{y}_i^{\text{obs}} \,\middle|\, \mathbf{f}_i, \mathbf{L}, \boldsymbol{\Lambda}\right) = &-\frac{\text{rank}(\boldsymbol{\delta}_i \boldsymbol{\Lambda} \boldsymbol{\delta}_i)}{2} \log 2\pi + \frac{1}{2} \log \hat{\det}(\boldsymbol{\delta}_i \boldsymbol{\Lambda} \boldsymbol{\delta}_i) \\ &- \frac{1}{2}\left(\mathbf{y}_i^{\text{obs}} - \mathbf{L}^t \mathbf{f}_i\right)^t \boldsymbol{\delta}_i \boldsymbol{\Lambda} \boldsymbol{\delta}_i \left(\mathbf{y}_i^{\text{obs}} - \mathbf{L}^t \mathbf{f}_i\right) \\ = &\log r_i + \log \hat{\theta}(\mathbf{f}_i; \mathbf{m}_i, \mathbf{P}_i)\,, \text{ where} \end{aligned} \tag{2}$$

the precision $\mathbf{P}_i = \mathbf{L} \boldsymbol{\delta}_i \boldsymbol{\Lambda} \boldsymbol{\delta}_i \mathbf{L}^t$, the mean $\mathbf{m}_i$ is a solution to $\mathbf{P}_i \mathbf{m}_i = \mathbf{L}^t \boldsymbol{\delta}_i \boldsymbol{\Lambda} \boldsymbol{\delta}_i \mathbf{y}_i^{\text{obs}}$ and

$$\begin{aligned} \log r_i = &-\frac{P_i^{\text{obs}} - \text{rank}(\mathbf{P}_i)}{2} \log 2\pi + \frac{1}{2}\left(\sum_{j=1}^{P} \delta_{ij} \log \lambda_j - \log \hat{\det}(\mathbf{P}_i)\right) \\ &- \frac{1}{2}\left[\mathbf{y}_i^{\text{obs}\,t} \boldsymbol{\delta}_i \boldsymbol{\Lambda} \boldsymbol{\delta}_i \mathbf{y}_i^{\text{obs}} - \mathbf{m}_i^t \mathbf{P}_i \mathbf{m}_i\right]. \end{aligned} \tag{3}$$

See SI Section 1.1 for detailed calculations. As $\boldsymbol{\Lambda}$ is diagonal, computing all $\mathbf{P}_i$ has complexity $\mathcal{O}(NPK^2)$, which dominates the computation time for these operations.

After computing $\mathbf{m}_i$, $\mathbf{P}_i$ and r_i, the Hassler et al. (2020) algorithm requires minor modification to compute the likelihood $p\left(\mathbf{Y}^{\text{obs}} \,\middle|\, \mathbf{L}, \boldsymbol{\Lambda}, \mathcal{F}\right)$ in $\mathcal{O}(NK^3)$ additional time. Specifically, $\mathbf{P}_i$ may not be invertible via the special inverse defined in Hassler et al. (2020). SI Section 1.2

offers an alternative approach that avoids this inversion via the continuously rediscovered identity $(\mathbf{A}+\mathbf{B})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\left(\mathbf{I}+\mathbf{B}\mathbf{A}^{-1}\right)^{-1}\mathbf{B}\mathbf{A}^{-1}$ for conformable square matrices $\mathbf{A}$ and $\mathbf{B}$ (Henderson et al., 1959; Henderson and Searle, 1981). We also utilize a more numerically stable modification of this post-order algorithm proposed by Bastide et al. (2021).

1.1 Partial Likelihood Calculations Under the Latent Factor Model

We present the detailed calculations from SI Equation 2.

$$
\begin{aligned}
\log p\big(\mathbf{y}_i^{\text{obs}} \,\big|\, \mathbf{f}_i, \mathbf{L}, \boldsymbol{\Lambda}\big) &= -\frac{\operatorname{rank}(\boldsymbol{\delta}_i\boldsymbol{\Lambda}\boldsymbol{\delta}_i)}{2}\log 2\pi + \frac{1}{2}\log \hat{\det}(\boldsymbol{\delta}_i\boldsymbol{\Lambda}\boldsymbol{\delta}_i) \\
&\qquad - \frac{1}{2}\left(\mathbf{y}_i^{\text{obs}} - \mathbf{L}^t\mathbf{f}_i\right)^t \boldsymbol{\delta}_i\boldsymbol{\Lambda}\boldsymbol{\delta}_i \left(\mathbf{y}_i^{\text{obs}} - \mathbf{L}^t\mathbf{f}_i\right) \\
&= -\frac{P_i^{\text{obs}}}{2}\log 2\pi + \frac{1}{2}\sum_{j=1}^{P}\delta_{ij}\log\lambda_j \\
&\qquad - \frac{1}{2}\left[\mathbf{f}_i^t\mathbf{L}^t\boldsymbol{\delta}_i\boldsymbol{\Lambda}\boldsymbol{\delta}_i\mathbf{L}^t\mathbf{f}_i - 2\mathbf{f}_i^t\mathbf{L}^t\boldsymbol{\delta}_i\boldsymbol{\Lambda}\boldsymbol{\delta}_i\mathbf{y}_i^{\text{obs}} + \mathbf{y}_i^{\text{obs}^t}\boldsymbol{\delta}_i\boldsymbol{\Lambda}\boldsymbol{\delta}_i\mathbf{y}_i^{\text{obs}}\right] \\
&= -\frac{P_i^{\text{obs}}}{2}\log 2\pi + \frac{1}{2}\sum_{j=1}^{P}\delta_{ij}\log\lambda_j \\
&\qquad - \frac{1}{2}\left[(\mathbf{f}_i - \mathbf{m}_i)^t\,\mathbf{P}_i\,(\mathbf{f}_i - \mathbf{m}_i)\right] - \frac{1}{2}\left[\mathbf{y}_i^{\text{obs}^t}\boldsymbol{\delta}_i\boldsymbol{\Lambda}\boldsymbol{\delta}_i\mathbf{y}_i^{\text{obs}} - \mathbf{m}_i^t\mathbf{P}_i\mathbf{m}_i\right] \\
&= \log r_i - \frac{\operatorname{rank}(\mathbf{P}_i)}{2}\log 2\pi + \frac{1}{2}\log\hat{\det}(\mathbf{P}_i) \\
&\qquad - \frac{1}{2}\left[(\mathbf{f}_i - \mathbf{m}_i)^t\,\mathbf{P}_i\,(\mathbf{f}_i - \mathbf{m}_i)\right] \\
&= \log r_i + \log\hat{\theta}(\mathbf{f}_i; \mathbf{m}_i, \mathbf{P}_i)\,,\ \text{ where}
\end{aligned}
\tag{4}
$$

the partial precision $\mathbf{P}_i = \mathbf{L}\boldsymbol{\delta}_i\boldsymbol{\Lambda}\boldsymbol{\delta}_i\mathbf{L}^t$, the partial mean $\mathbf{m}_i$ is a (not necessarily unique) solution to $\mathbf{P}_i\mathbf{m}_i = \mathbf{L}^t\boldsymbol{\delta}_i\boldsymbol{\Lambda}\boldsymbol{\delta}_i\mathbf{y}_i^{\text{obs}}$ and the remainder

$$
\begin{aligned}
\log r_i = -\frac{P_i^{\text{obs}} - \operatorname{rank}(\mathbf{P}_i)}{2}\log 2\pi + \frac{1}{2}\left(\sum_{j=1}^{P}\delta_{ij}\log\lambda_j - \log\hat{\det}(\mathbf{P}_i)\right) \\
- \frac{1}{2}\left[\mathbf{y}_i^{\text{obs}^t}\boldsymbol{\delta}_i\boldsymbol{\Lambda}\boldsymbol{\delta}_i\mathbf{y}_i^{\text{obs}} - \mathbf{m}_i^t\mathbf{P}_i\mathbf{m}_i\right].
\end{aligned}
\tag{5}
$$

1.2 Special Inverse Calculations

One challenge that the PFA model poses to this approach is that the partial precisions at the tips $\mathbf{P}_i$ for $i = 1, \ldots, N$ may not be invertible via the pseudoinverse used by Hassler et al. (2020). The post-order traversal algorithm requires that for each internal node ν_j for $j =$

$N+1,\dots,2N-1$ in $\mathcal{F}$, we must compute $\mathbf{P}_j^*$ such that $p\big(\mathbf{Y}_{\lfloor j \rfloor} \,\big|\, \mathbf{f}_{\text{pa}(j)}\big) = r_j \hat{\theta}\big(\mathbf{f}_{\text{pa}(j)}; \mathbf{m}_j, \mathbf{P}_j^*\big)$, where $\mathbf{Y}_{\lfloor j \rfloor}$ represents the trait values of all terminal descendants of node ν_j. In the PFA model, this results in $\mathbf{P}_j^* = \big(\mathbf{P}_j^{-1} + t_j \mathbf{I}_K\big)^{-1}$. However, it is possible that the initial partial precisions $\mathbf{P}_i$ at the tip nodes $\nu_1,\dots,\nu_N$ may be rank-deficient. This situation arises, for example, when the number of non-missing traits P_i^{obs} at taxon i is less than the number of factors K. To avoid this inversion, we use an algebraic slight-of-hand to compute $\mathbf{P}_j^*$ in terms of $\mathbf{P}_j$ directly (rather than its non-existing inverse). Specifically we use an identity for the inverse of the sum of two square matrices that has been discovered and forgotten several times (see, for example, Henderson et al., 1959; Henderson and Searle, 1981)

$$(\mathbf{A}+\mathbf{B})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\left(\mathbf{I} + \mathbf{B}\mathbf{A}^{-1}\right)^{-1}\mathbf{B}\mathbf{A}^{-1}. \tag{6}$$

Applying this to our equation for $\mathbf{P}_j^*$, we get

$$\mathbf{P}_j^* = \mathbf{P}_j - t_j \mathbf{P}_j \left(\mathbf{I}_k + t_j \mathbf{P}_j\right)^{-1} \mathbf{P}_j. \tag{7}$$

Note that the matrix $\mathbf{I}_K + t_j \mathbf{P}_j$ is the sum of the positive semi-definite matrix $t_j \mathbf{P}_j$ with the positive definite matrix $\mathbf{I}_K$ and is therefore invertible. As such, computing $\mathbf{P}_j^*$ is indeed possible and the Hassler et al. (2020) algorithm can proceed to compute the likelihood.

2 Sampling from the Loadings L via Data Augmentation

To employ the Gibbs sampler of Tolkoff et al. (2017) to sample from the loading $\mathbf{L}$, we follow the procedure below:

1. Sample from $\mathbf{F} \,\big|\, \mathbf{Y}^{\text{obs}}, \mathbf{L}, \mathbf{\Lambda}, \mathcal{F}$ via the pre-order algorithm of Hassler et al. (2020)

2. Sample from $\mathbf{L} \,\big|\, \mathbf{Y}^{\text{obs}}, \mathbf{F}, \mathbf{\Lambda}$ via the methods discussed in Lopes and West (2004)

2.1 Pre-Order Data Augmentation Algorithm

We seek to sample from $\mathbf{F} \,\big|\, \mathbf{Y}^{\text{obs}}, \mathbf{L}, \mathbf{\Lambda}, \mathcal{F}$ via the pre-order algorithm of Hassler et al. (2020). This procedure relies on first computing the statistics $\mathbf{m}_i$ and $\mathbf{P}_i$ such that

$$p\big(\mathbf{Y}_{\lfloor i \rfloor}^{\text{obs}} \,\big|\, \mathbf{f}_i, \mathbf{L}, \mathbf{\Lambda}, \mathcal{F}\big) \propto \hat{\theta}(\mathbf{f}_i; \mathbf{m}_i, \mathbf{P}_i) \tag{8}$$

for $i = 1, \ldots, 2N-1$ (i.e. all nodes in the tree), where $\mathbf{Y}^{\text{obs}}_{\lfloor i \rfloor}$ is the subset of $\mathbf{Y}^{\text{obs}}$ restricted to the descendants of node ν_i. We compute these statistics at the tips as described in Section 2.1.1 and at internal nodes as described in Section 2.1.2 of Hassler et al. (2020).

Once we have computed these statistics, we draw the factors at the root from their full conditional distribution $\mathbf{f}_{2N-1} \,\big|\, \mathbf{Y}^{\text{obs}}, \mathbf{L}, \boldsymbol{\Lambda}, \mathcal{F}, \boldsymbol{\mu}_0, \kappa_0$ as described by Equation 13 in Hassler et al. (2020). After sampling the factors $\mathbf{f}_{2N-1}$ at the root node ν_{2N-1} from their full conditional distribution, we perform a pre-order traversal of the tree sampling from $\mathbf{f}_i \,\Big|\, \mathbf{f}_{\text{pa}(i)}, \mathbf{Y}^{\text{obs}}_{\lfloor i \rfloor}, \mathbf{L}, \boldsymbol{\Lambda}, \mathcal{F}$ for $j = 1, \ldots, 2N-2$ as described in Section 2.2.1 of Hassler et al. (2020). After we have completed this pre-order traversal, we have sampled from the full conditional distribution of $\mathbf{F} = (\mathbf{f}_1, \ldots, \mathbf{f}_N)^t$.

2.2 Conjugate Gibbs Sampler on the Loadings L

Here we describe our procedure for sampling from $\mathbf{L} \,\big|\, \mathbf{Y}^{\text{obs}}, \mathbf{F}, \boldsymbol{\Lambda}$ via the conjugate Gibbs sampler developed by Lopes and West (2004) and Tolkoff et al. (2017). Let us first introduce notation related to both structured sparsity in the loadings and missing data. Let the K-dimensional vector $\boldsymbol{\ell}_j$ and N-dimensional vector $\mathbf{y}'_j$ be the j^{th} column of $\mathbf{L}$ and $\mathbf{Y}$ respectively for $j = 1, \ldots, P$. Let $\mathbf{x}_j \subseteq \{1, \ldots, K\}$ be the indices corresponding to the unconstrained elements of $\boldsymbol{\ell}_j$ (i.e. those that are not fixed at 0), and let $\mathbf{z}_j \subseteq \{1, \ldots, N\}$ be the indices of the observed (non-missing) elements of $\mathbf{y}'_j$. Finally let the sub-vectors $\boldsymbol{\ell}_{j,\mathbf{x}_j}$ and $\mathbf{f}_{i,\mathbf{x}_j}$ be the elements of $\boldsymbol{\ell}_j$ and $\mathbf{f}_i$, respectively, restricted to the indices in $\mathbf{x}_j$, and let $\mathbf{y}'_{j,\mathbf{z}_j}$ be the elements of $\mathbf{y}'_j$ restricted to the elements in $\mathbf{z}_j$ for $i = 1, \ldots, N$ and $j = 1, \ldots, P$. Note that conditional on the latent factors, the full conditional distributions of each column of the loadings are independent. Additionally, the full conditional of $\boldsymbol{\ell}_j$ depends only on $\mathbf{y}'_j$, and does not depend on the other columns of the data matrix $\mathbf{Y}$ (Lopes and West, 2004). As such, we draw from $\boldsymbol{\ell}_{j,\mathbf{x}_j} \,\Big|\, \mathbf{F}, \mathbf{y}'_{j,\mathbf{z}_j}, \boldsymbol{\Lambda}$ as follows:

$$\begin{aligned}
p\Big(\boldsymbol{\ell}_{j,\mathbf{x}_j} \,\Big|\, \mathbf{y}'_{j,\mathbf{z}_j}, \mathbf{F}, \lambda_j\Big) &\propto p\Big(\mathbf{y}'_{j,\mathbf{z}_j} \,\Big|\, \boldsymbol{\ell}_{j,\mathbf{x}_j}, \mathbf{L}, \lambda_j\Big) p\big(\boldsymbol{\ell}_{j,\mathbf{x}_j}\big) \\
&= \prod_{i \in \mathbf{z}_j} p\big(y_{ij} \,\big|\, \mathbf{f}_i, \boldsymbol{\ell}_{j,\mathbf{x}_j}, \lambda_j\big) p\big(\boldsymbol{\ell}_{j,\mathbf{x}_j}\big) \\
&= \prod_{i \in \mathbf{z}_j} \theta\Big(y_{ij}; \boldsymbol{\ell}^t_{j,\mathbf{x}_j} \mathbf{f}_{i,\mathbf{x}_j}, \lambda_j\Big)\, \theta\big(\boldsymbol{\ell}_{j,\mathbf{x}_j}; \mathbf{0}, \boldsymbol{\Lambda}_j\big) \\
&= \theta\big(\boldsymbol{\ell}_{j,\mathbf{x}_j}; \boldsymbol{\eta}_j, \boldsymbol{\Gamma}_j\big)
\end{aligned} \tag{9}$$

where

$$
\begin{aligned}
\mathbf{\Lambda}_j &= \frac{1}{\sigma^2}\mathbf{I}_{|\mathbf{x}_j|}, \\
\mathbf{\Gamma}_j &= \mathbf{\Lambda}_j + \lambda_j \sum_{i \in \mathbf{z}_j} \mathbf{f}_{i,\mathbf{x}_j}\mathbf{f}_{i,\mathbf{x}_j}^t, \\
\boldsymbol{\eta}_j &= \mathbf{\Gamma}_j^{-1}\left(\mathbf{\Lambda}_j\mathbf{0} + \lambda_j \sum_{i \in \mathbf{z}_j} y_{ij}\mathbf{f}_{i,\mathbf{x}_j}\right)
\end{aligned}
\tag{10}
$$

and $\theta(\mathbf{x}; \boldsymbol{\mu}, \mathbf{P})$ is the multivariate normal density function with argument $\mathbf{x}$, mean $\boldsymbol{\mu}$ and precision $\mathbf{P}$.

Computing $\mathbf{\Gamma}_j$ has computational complexity $\mathcal{O}(NK^2)$, so computing all P precisions has overall complexity $\mathcal{O}(NPK^2)$. Once the precisions have been computed, computing the means has complexity $\mathcal{O}(NPK + PK^3)$, which contributes relatively little to overall computation time as $N >> K$ for most problems. Note that if the data are completely observed and there is no structured sparsity in the loadings, then $\mathbf{\Gamma}_j = \mathbf{\Lambda}_j + \lambda_j\mathbf{F}^t\mathbf{F}$. In that case, we only need to compute $\mathbf{F}^t\mathbf{F}$ once (not P times), which brings the overall complexity down to $\mathcal{O}(NPK)$ (as we still need to compute the means for al P columns of $\mathbf{L}$). Drawing all $\boldsymbol{\ell}_j$ for $j = 1, \ldots, P$ results in a complete sample from the full conditional distribution of $\mathbf{L}$.

3 Loadings Gradient Calculation

We calculate the gradient of the likelihood with respect to each column of the loadings $\boldsymbol{\ell}_j$ individually to accommodate variation in the missing data structure across traits. Note that in the calculations below, we omit explicit dependence on the residual precision $\mathbf{\Lambda}$ and tree $\mathcal{F}$ in the interest of notational simplicity.

$$
\begin{aligned}
\nabla_{\boldsymbol{\ell}_j} \log p\big(\mathbf{Y}^{\text{obs}} \,\big|\, \mathbf{L}\big) &= \frac{1}{p\big(\mathbf{Y}^{\text{obs}} \,|\, \mathbf{L}\big)} \nabla_{\boldsymbol{\ell}_j} p\big(\mathbf{Y}^{\text{obs}} \,\big|\, \mathbf{L}\big) \\
&= \frac{1}{p\big(\mathbf{Y}^{\text{obs}} \,|\, \mathbf{L}\big)} \nabla_{\boldsymbol{\ell}_j} \left[\int p\big(\mathbf{Y}^{\text{obs}} \,\big|\, \mathbf{F}, \mathbf{L}\big) p(\mathbf{F}) \mathrm{d}\mathbf{F}\right] \\
&= \frac{1}{p\big(\mathbf{Y}^{\text{obs}} \,|\, \mathbf{L}\big)} \int p(\mathbf{F}) \nabla_{\boldsymbol{\ell}_j} p\big(\mathbf{Y}^{\text{obs}} \,\big|\, \mathbf{F}, \mathbf{L}\big) \mathrm{d}\mathbf{F}.
\end{aligned}
\tag{11}
$$

Based on the fact that the elements of $\mathbf{Y}^{\text{obs}}$ are independent conditional on the loadings and factors, we have:

$$\begin{aligned} p\big(\mathbf{Y}^{\text{obs}} \,\big|\, \mathbf{F}, \mathbf{L}\big) &= \prod_{i=1}^{N} \prod_{k=1}^{P} p\big(y_{ij} \,|\, \mathbf{f}_i, \mathbf{L}\big)^{\delta_{ik}} \\ &= \prod_{i=1}^{N} \prod_{k=1}^{P} (2\pi\lambda_k)^{-\delta_{ik}/2} \exp\left(-\frac{1}{2}\lambda_k \delta_{ik} \big(y_{ik} - \mathbf{f}_i^t \boldsymbol{\ell}_k\big)^2\right) \\ &= c \exp\left(-\frac{1}{2} \sum_{i=1}^{N} \sum_{k=1}^{P} \lambda_k \delta_{ik} \big(y_{ik} - \mathbf{f}_i^t \boldsymbol{\ell}_k\big)^2\right), \end{aligned} \tag{12}$$

where δ_{ij} is an indicator that equals 1 if y_{ij} is observed and 0 if it is missing, and c is a normalization constant that does not depend on the loadings $\mathbf{L}$. Therefore,

$$\begin{aligned} \nabla_{\boldsymbol{\ell}_j} p\big(\mathbf{Y}^{\text{obs}} \,\big|\, \mathbf{F}, \mathbf{L}\big) &= \nabla_{\boldsymbol{\ell}_j} \left[c \exp\left(-\frac{1}{2} \sum_{i=1}^{N} \sum_{k=1}^{P} \lambda_k \delta_{ik} \big(y_{ik} - \mathbf{f}_i^t \boldsymbol{\ell}_k\big)^2\right)\right] \\ &= c \exp\left(-\frac{1}{2} \sum_{i=1}^{N} \sum_{k=1}^{P} \lambda_k \delta_{ik} \big(y_{ik} - \mathbf{f}_i^t \boldsymbol{\ell}_k\big)^2\right) \\ &\qquad \times \nabla_{\boldsymbol{\ell}_j} \left[-\frac{1}{2} \sum_{i=1}^{N} \sum_{k=1}^{P} \lambda_k \delta_{ik} \big(y_{ik} - \mathbf{f}_i^t \boldsymbol{\ell}_k\big)^2\right] \\ &= p\big(\mathbf{Y}^{\text{obs}} \,\big|\, \mathbf{F}, \mathbf{L}\big) \times \nabla_{\boldsymbol{\ell}_j} \left[-\frac{1}{2} \sum_{i=1}^{N} \sum_{k=1}^{P} \lambda_k \delta_{ik} \big(y_{ik} - \mathbf{f}_i^t \boldsymbol{\ell}_k\big)^2\right] \\ &= p\big(\mathbf{Y}^{\text{obs}} \,\big|\, \mathbf{F}, \mathbf{L}\big) \times -\frac{1}{2}\lambda_j \sum_{i=1}^{N} \delta_{ij} \nabla_{\boldsymbol{\ell}_j} \left[\big(y_{ij} - \mathbf{f}_i^t \boldsymbol{\ell}_j\big)^2\right] \\ &= p\big(\mathbf{Y}^{\text{obs}} \,\big|\, \mathbf{F}, \mathbf{L}\big) \lambda_j \sum_{i=1}^{N} \delta_{ij} \mathbf{f}_i \big(y_{ij} - \mathbf{f}_i^t \boldsymbol{\ell}_j\big) \\ &= p\big(\mathbf{Y}^{\text{obs}} \,\big|\, \mathbf{F}, \mathbf{L}\big) \lambda_j \left(\mathbf{F}^t \boldsymbol{\delta}_j' \mathbf{y}_j^{\text{obs}'} - \mathbf{F}^t \boldsymbol{\delta}_j' \mathbf{F} \boldsymbol{\ell}_j\right) \end{aligned} \tag{13}$$

where $\mathbf{y}_j^{\text{obs}\prime}$ is the j^{th} column of $\mathbf{Y}^{\text{obs}}$ and $\boldsymbol{\delta}_j' = \text{diag}[\delta_{1j}, \ldots, \delta_{Nj}]$ is a diagonal matrix of observed-data indicators. Using this result in SI Equation 11, we calculate

$$\begin{aligned}
\nabla_{\boldsymbol{\ell}_j} \log p\big(\mathbf{Y}^{\text{obs}} \,\big|\, \mathbf{L}\big) &= \int \frac{p(\mathbf{F}) p\big(\mathbf{Y}^{\text{obs}} \,\big|\, \mathbf{F}, \mathbf{L}\big)}{p\big(\mathbf{Y}^{\text{obs}} \,|\, \mathbf{L}\big)} \lambda_j \left(\mathbf{F}^t \boldsymbol{\delta}_j' \mathbf{y}_j^{\text{obs}\prime} - \mathbf{F}^t \boldsymbol{\delta}_j' \mathbf{F} \boldsymbol{\ell}_j\right) \mathrm{d}\mathbf{F} \\
&= \int p\big(\mathbf{F} \,\big|\, \mathbf{Y}^{\text{obs}}, \mathbf{L}\big) \lambda_j \left(\mathbf{F}^t \boldsymbol{\delta}_j' \mathbf{y}_j^{\text{obs}\prime} - \mathbf{F}^t \boldsymbol{\delta}_j' \mathbf{F} \boldsymbol{\ell}_j\right) \mathrm{d}\mathbf{F} \\
&= \mathbb{E}\Big[\lambda_j \left(\mathbf{F}^t \boldsymbol{\delta}_j' \mathbf{y}_j^{\text{obs}\prime} - \mathbf{F}^t \boldsymbol{\delta}_j' \mathbf{F} \boldsymbol{\ell}_j\right) \,\Big|\, \mathbf{Y}^{\text{obs}}, \mathbf{L}\Big] \\
&= \lambda_j \mathbb{E}\big[\mathbf{F}^t \,\big|\, \mathbf{Y}^{\text{obs}}, \mathbf{L}\big] \boldsymbol{\delta}_j' \mathbf{y}_j^{\text{obs}\prime} - \lambda_j \mathbb{E}\big[\mathbf{F}^t \boldsymbol{\delta}_j' \mathbf{F} \,\big|\, \mathbf{Y}^{\text{obs}}, \mathbf{L}\big] \boldsymbol{\ell}_j.
\end{aligned} \tag{14}$$

Note that

$$\begin{aligned}
\mathbb{E}\big[\mathbf{F}^t \boldsymbol{\delta}_j' \mathbf{F} \,\big|\, \mathbf{Y}^{\text{obs}}, \mathbf{L}\big] &= \sum_{i=1}^{N} \delta_{ij} \mathbb{E}\big[\mathbf{f}_i \mathbf{f}_i^t \,\big|\, \mathbf{Y}^{\text{obs}}, \mathbf{L}\big] \\
&= \sum_{i=1}^{N} \delta_{ij} \mathbb{V}\big[\mathbf{f}_i \,\big|\, \mathbf{Y}^{\text{obs}}, \mathbf{L}\big] + \delta_{ij} \mathbb{E}\big[\mathbf{f}_i \,\big|\, \mathbf{Y}^{\text{obs}}, \mathbf{L}\big] \mathbb{E}\big[\mathbf{f}_i \,\big|\, \mathbf{Y}^{\text{obs}}, \mathbf{L}\big]^t.
\end{aligned} \tag{15}$$

We compute $\mathbb{E}\big[\mathbf{f}_i \,\big|\, \mathbf{Y}^{\text{obs}}, \mathbf{L}\big]$ and $\mathbb{V}\big[\mathbf{f}_i \,\big|\, \mathbf{Y}^{\text{obs}}, \mathbf{L}\big]$ for $i = 1, \ldots, N$ in $\mathcal{O}(NPK^2 + NK^3)$ via a post-order likelihood calculation algorithm (see SI Section 1) followed by the pre-order algorithms independently developed by Bastide et al. (2018) and Fisher et al. (2020).

For the case where there is no missing data, we can simplify SI Equation 13 to be

$$\nabla_{\mathbf{L}} p\big(\mathbf{Y} \,|\, \mathbf{F}, \mathbf{L}\big) = p\big(\mathbf{Y} \,|\, \mathbf{F}, \mathbf{L}\big) \left[\mathbf{F}^t \mathbf{Y} \boldsymbol{\Lambda} - \mathbf{F}^t \mathbf{F} \mathbf{L} \boldsymbol{\Lambda}\right]. \tag{16}$$

4 Post-Processing Procedure

We employ singular value decomposition (SVD) to enforce the orthogonality constraint on the loadings via post-processing. In practice, we sample from the orthogonally-constrained loadings as follows. Let $\mathbf{L}^{(n)}$ be a sample from the posterior distribution $\mathbf{L} \,|\, \mathbf{Y}$ at the n^{th} state in the MCMC chain. For each $\mathbf{L}^{(n)}$, we compute the SVD $\mathbf{L}^{(n)} = \mathbf{U}^{(n)} \boldsymbol{\Sigma}^{(n)} \mathbf{V}^{(n)}$ where $\mathbf{U}^{(n)}$ is a $K \times K$ orthonormal matrix and $\boldsymbol{\Sigma}^{(n)}$ and $\mathbf{V}^{(n)}$ retain their constraints from Section 2.2.2 (i.e. $\boldsymbol{\Sigma}^{(n)}$ is diagonal with descending positive entries and $\mathbf{V}^{(n)} {\mathbf{V}^{(n)}}^t = \mathbf{I}_K$). While the parameter $\mathbf{U}$ is not identifiable, $\boldsymbol{\Sigma}$ and $\mathbf{V}$ are (Holbrook et al., 2016). As such, we then treat ${\mathbf{L}^{\perp}}^{(n)} = \boldsymbol{\Sigma}^{(n)} \mathbf{V}^{(n)}$ as (now identifiable) samples from the posterior of the loadings. If we also sample the factors $\mathbf{F}$, we rotate the factors to sample from ${\mathbf{F}^{\perp}}^{(n)} = \mathbf{F}^{(n)} \mathbf{U}^{(n)}$ to ensure that ${\mathbf{F}^{\perp}}^{(n)} {\mathbf{L}^{\perp}}^{(n)} = \mathbf{F}^{(n)} \mathbf{U}^{(n)} \boldsymbol{\Sigma}^{(n)} \mathbf{V}^{(n)} = \mathbf{F}^{(n)} \mathbf{L}^{(n)}$.

5 Sampling from $\mathbf{\Sigma}$

We define the K-vector $\boldsymbol{\sigma}$ such that $\mathbf{\Sigma} = \text{diag}[\boldsymbol{\sigma}]$ and sample $\boldsymbol{\sigma}$ as follows (see SI Section 5.1 for derivation):

$$\begin{aligned} \boldsymbol{\sigma} \,\big|\, \mathbf{Y}^{\text{obs}}, \mathbf{F}, \mathbf{V}, \mathbf{\Lambda} &\sim \text{MVN}\big(\boldsymbol{\mu}_\sigma, \mathbf{P}_\sigma^{-1}\big), \text{ where} \\ \mathbf{P}_\sigma &= \text{diag}[\boldsymbol{\tau}] + \sum_{j=1}^{P} \lambda_j \,\text{diag}[\mathbf{v}_j]\mathbf{F}^t \boldsymbol{\delta}_j' \mathbf{F}\,\text{diag}[\mathbf{v}_j], \\ \boldsymbol{\mu}_\sigma &= \mathbf{P}_\sigma^{-1} \left(\sum_{j=1}^{P} \lambda_j \,\text{diag}[\mathbf{v}_j]\mathbf{F}^t \boldsymbol{\delta}_j' \mathbf{y}_j^{\text{obs}\prime} \right), \end{aligned} \tag{17}$$

$\boldsymbol{\tau} = (\tau_1, \ldots, \tau_K)$ and $\mathbf{v}_j$ is the j^{th} column of $\mathbf{V}$.

While the prior encourages the elements of $\boldsymbol{\sigma}$ to have descending absolute value, it does not enforce this constraint strictly. As discussed in Section 2.2.2, for some problems a strict ordering with forced spacing may be necessary in practice for full identifiability. In these cases we employ a rejection sampler where we draw from the full conditional distribution of $\boldsymbol{\sigma}$ using the unrestricted multivariate normal distribution and reject any samples that do not conform to the particular constraint. As the unconstrained prior already induces a soft ordering, we find that this rejection sampler typically has high acceptance probability.

5.1 Loadings Scale Full Conditional Distribution

We detail our derivation of SI Equation 17 below. Recall that we define the K-vector $\boldsymbol{\sigma}$ such that $\boldsymbol{\Sigma} = \text{diag}[\boldsymbol{\sigma}]$, and note that all proportional symbols imply log-proportional:

$$
\begin{aligned}
&\log p\big(\boldsymbol{\sigma} \,\big|\, \mathbf{Y}^{\text{obs}}, \mathbf{F}, \mathbf{V}, \boldsymbol{\Lambda}\big) \\
&\qquad \propto \log p\big(\mathbf{Y}^{\text{obs}} \,\big|\, \boldsymbol{\sigma}, \mathbf{F}, \mathbf{V}, \boldsymbol{\Lambda}\big) + \log p(\boldsymbol{\sigma}) \\
&\qquad = \sum_{j=1}^{P} \log p\Big(\mathbf{y}_j^{\text{obs}'} \,\Big|\, \boldsymbol{\sigma}, \mathbf{F}, \mathbf{v}_j, \lambda_j\Big) + \log p(\boldsymbol{\sigma}) \\
&\qquad \propto -\frac{1}{2}\sum_{j=1}^{P} \lambda_j \left(\mathbf{F}\boldsymbol{\Sigma}\mathbf{v}_j - \mathbf{y}_j^{\text{obs}'}\right)^t \boldsymbol{\delta}'_j \left(\mathbf{F}\boldsymbol{\Sigma}\mathbf{v}_j - \mathbf{y}_j^{\text{obs}'}\right) + \log p(\boldsymbol{\sigma}) \\
&\qquad \propto -\frac{1}{2}\sum_{j=1}^{P} \lambda_j \left(\mathbf{v}_j^t\boldsymbol{\Sigma}\mathbf{F}^t\boldsymbol{\delta}'_j\mathbf{F}\boldsymbol{\Sigma}\mathbf{v}_j - 2\mathbf{v}_j^t\boldsymbol{\Sigma}\mathbf{F}^t\boldsymbol{\delta}'_j\mathbf{y}_j^{\text{obs}'}\right) + \log p(\boldsymbol{\sigma}) \\
&\qquad \propto -\frac{1}{2}\sum_{j=1}^{P} \lambda_j \left(\boldsymbol{\sigma}^t\,\text{diag}[\mathbf{v}_j]\mathbf{F}^t\boldsymbol{\delta}'_j\mathbf{F}\,\text{diag}[\mathbf{v}_j]\boldsymbol{\sigma} - 2\boldsymbol{\sigma}^t\,\text{diag}[\mathbf{v}_j]\mathbf{F}^t\boldsymbol{\delta}'_j\mathbf{y}_j^{\text{obs}'}\right) + \log p(\boldsymbol{\sigma}) \\
&\qquad \propto -\frac{1}{2}\boldsymbol{\sigma}^t\left(\sum_{j=1}^{P}\lambda_j\,\text{diag}[\mathbf{v}_j]\mathbf{F}^t\boldsymbol{\delta}'_j\mathbf{F}\,\text{diag}[\mathbf{v}_j]\right)\boldsymbol{\sigma} \\
&\qquad\qquad - \boldsymbol{\sigma}^t\left(\sum_{j=1}^{P}\lambda_j\,\text{diag}[\mathbf{v}_j]\mathbf{F}^t\boldsymbol{\delta}'_j\mathbf{y}_j^{\text{obs}'}\right) + \log p(\boldsymbol{\sigma}) \\
&\qquad \propto -\frac{1}{2}\boldsymbol{\sigma}^t\left(\text{diag}[\boldsymbol{\tau}] + \sum_{j=1}^{P}\lambda_j\,\text{diag}[\mathbf{v}_j]\mathbf{F}^t\boldsymbol{\delta}'_j\mathbf{F}\,\text{diag}[\mathbf{v}_j]\right)\boldsymbol{\sigma} \\
&\qquad\qquad - \boldsymbol{\sigma}^t\left(\sum_{j=1}^{P}\lambda_j\,\text{diag}[\mathbf{v}_j]\mathbf{F}^t\boldsymbol{\delta}'_j\mathbf{y}_j^{\text{obs}'}\right) \\
&\qquad \propto -\frac{1}{2}\left(\boldsymbol{\sigma} - \boldsymbol{\mu}_\sigma\right)^t\mathbf{P}_\sigma\left(\boldsymbol{\sigma} - \boldsymbol{\mu}_\sigma\right),
\end{aligned}
\tag{18}
$$

where

$$
\begin{aligned}
\mathbf{P}_\sigma &= \text{diag}[\boldsymbol{\tau}] + \sum_{j=1}^{P}\lambda_j\,\text{diag}[\mathbf{v}_j]\mathbf{F}^t\boldsymbol{\delta}'_j\mathbf{F}\,\text{diag}[\mathbf{v}_j] \text{ and} \\
\boldsymbol{\mu}_\sigma &= \mathbf{P}_\sigma^{-1}\left(\sum_{j=1}^{P}\lambda_j\,\text{diag}[\mathbf{v}_j]\mathbf{F}^t\boldsymbol{\delta}'_j\mathbf{y}_j^{\text{obs}'}\right)
\end{aligned}
\tag{19}
$$

This implies

$$
\log p\big(\boldsymbol{\sigma} \,\big|\, \mathbf{Y}^{\text{obs}}, \mathbf{F}, \mathbf{V}, \boldsymbol{\Lambda}\big) = \theta\big(\boldsymbol{\sigma}; \boldsymbol{\mu}_\sigma, \mathbf{P}_\sigma\big)\,. \tag{20}
$$

6 Sign Constraint on the Loadings

Regardless of which prior (i.i.d. vs shrinkage) or constraint (sparsity vs orthogonality) we choose, we must enforce a sign constraint on a single element in each row of $\mathbf{L}$ for full identifiability. Let $\gamma_k \in \{1, \dots, P\}$ be the index of the K^{th} row of $\mathbf{L}$ with the sign constraint (i.e. require $\ell_{\gamma_k k} \geq 0$). If the sample $\ell_{k\gamma_k}^{(n)} < 0$, then we simply multiply row k of $\mathbf{L}^{(n)}$ by -1 to ensure $\ell_{k\gamma_k}^{(n)} \geq 0$. These K sign-constrained elements are not required to be in the same row of $\mathbf{L}$, and we choose these rows in a way that maximizes the posterior identifiability of $\mathbf{L}$. In practice, we apply a simple heuristic where for $k = 1, \dots, K$

$$\gamma_k = \underset{j \in 1, \dots, P}{\arg\max} \left(\frac{\bar{\ell}_{jk}^{\text{abs}}}{\sqrt{\sum_{n=1}^{M} \left(\left| \ell_{jk}^{(n)} \right| - \bar{\ell}_{jk}^{\text{abs}} \right)^2}} \right) \quad \text{and} \quad \bar{\ell}_{jk}^{\text{abs}} = \frac{1}{M} \sum_{n=1}^{M} \left| \ell_{jk}^{(n)} \right| . \tag{21}$$

In the absence of sign constraints, the marginal posteriors of many elements of $\mathbf{L}$ are bimodal and symmetric across zero. Our heuristic aims to find an index in each column of $\mathbf{L}$ with low mass near 0 and simply chose the positive mode.

7 Sampling from $\mathbf{\Lambda}$

Regardless of the prior on the loadings, we sample from $\mathbf{\Lambda} \,\big|\, \mathbf{F}, \mathbf{Y}^{\text{obs}}, \mathbf{L}$ using the same conjugate Gibbs sampler as Tolkoff et al. (2017) in conjunction with the data augmentation algorithm from Section 3.1.1. The Gamma$(a_{\mathbf{\Lambda}}, b_{\mathbf{\Lambda}})$ (shape, rate parameterization) prior on the diagonal elements of $\mathbf{\Lambda}$ results in a simple expression for the full conditional distribution of λ_j for $j = 1, \dots, P$ conditional on the factors $\mathbf{F}$. Specifically, each λ_j is distributed as

$$\lambda_j \,\big|\, \mathbf{Y}^{\text{obs}}, \mathbf{F}, \mathbf{L} \sim \text{Gamma}\left(a_{\mathbf{\Lambda}} + \frac{N_j^{\text{obs}}}{2}, b_{\mathbf{\Lambda}} + \frac{1}{2} \sum_{i=1}^{N} \delta_{ij} \left(y_{ij} - \boldsymbol{\ell}_j^t \mathbf{f}_i \right)^2 \right) . \tag{22}$$

This computation only requires run time $\mathcal{O}(NPK)$ and, in our experience, time spent estimating $\mathbf{\Lambda}$ does not contribute significantly to the overall run time of the MCMC chain.

Note that as with the loadings in Section 3.1.2, we also derive a strategy for sampling from these precisions without conditioning on $\mathbf{F}$ via HMC. As we are satisfied with the Tolkoff et al. (2017) procedure, we have not implemented this strategy, but the derivation can be found below. Naturally, this HMC sampler requires we compute the gradient of the likelihood with respect to the loadings as follows:

$$
\begin{aligned}
\frac{\partial \log p\big(\mathbf{Y}^{\text{obs}} \,\big|\, \boldsymbol{\Lambda}\big)}{\partial \lambda_j} &= \frac{1}{p\big(\mathbf{Y}^{\text{obs}} \,|\, \boldsymbol{\Lambda}\big)} \int p(\mathbf{F}) \frac{\partial p\big(\mathbf{Y}^{\text{obs}} \,\big|\, \mathbf{F}, \boldsymbol{\Lambda}\big)}{\partial \lambda_j} \mathrm{d}\mathbf{F} \\
&= \frac{1}{p\big(\mathbf{Y}^{\text{obs}} \,|\, \boldsymbol{\Lambda}\big)} \int p(\mathbf{F}) p\big(\mathbf{Y}^{\text{obs}} \,\big|\, \mathbf{F}, \boldsymbol{\Lambda}\big) \\
&\qquad \times \left(\frac{N_j^{\text{obs}}}{2} \lambda_j^{-1} - \frac{1}{2} \left(\mathbf{F}\boldsymbol{\ell}_j - \mathbf{y}_j^{\text{obs}\prime} \right)^t \boldsymbol{\delta}_j' \left(\mathbf{F}\boldsymbol{\ell}_j - \mathbf{y}_j^{\text{obs}\prime} \right) \right) \mathrm{d}\mathbf{F} \\
&= \mathbb{E}\left[\frac{N_j^{\text{obs}}}{2} \lambda_j^{-1} - \frac{1}{2} \left(\mathbf{F}\boldsymbol{\ell}_j - \mathbf{y}_j^{\text{obs}\prime} \right)^t \boldsymbol{\delta}_j' \left(\mathbf{F}\boldsymbol{\ell}_j - \mathbf{y}_j^{\text{obs}\prime} \right) \,\middle|\, \mathbf{Y}^{\text{obs}}, \boldsymbol{\Lambda} \right] \\
&= \frac{N_j^{\text{obs}}}{2} \lambda_j^{-1} - \frac{1}{2} \boldsymbol{\ell}_j^t \mathbb{E}\big[\mathbf{F}^t \boldsymbol{\delta}_j' \mathbf{F} \,\big|\, \mathbf{Y}^{\text{obs}}, \boldsymbol{\Lambda}\big] \boldsymbol{\ell}_j + \boldsymbol{\ell}_j^t \mathbb{E}\big[\mathbf{F}^t \,\big|\, \mathbf{Y}^{\text{obs}}, \boldsymbol{\Lambda}\big] \boldsymbol{\delta}_j' \mathbf{y}_j^{\text{obs}\prime} \\
&\quad - \frac{1}{2} \mathbf{y}_j^{\text{obs}\prime t} \boldsymbol{\delta}_j' \mathbf{y}_j^{\text{obs}\prime}
\end{aligned}
\tag{23}
$$

The conditional expectations of the latent factors are the same as in Section 3.1.2. Note that we restrict $\boldsymbol{\Lambda}$ to be diagonal, so we only consider the diagonal elements of the gradient. Once we have computed this gradient, we employ it in standard HMC to sample from the full conditional of $\boldsymbol{\Lambda}$.

8 Timing

8.1 Simulation Details

To simulate each data set for the timing comparison, we generate a random coalescent tree with N tips (Kingman, 1982). We then simulate the factors $\mathbf{F}$ according to K independent Brownian diffusion processes on the tree and subsequently re-scale the factors so that each column has unit variance. We draw $\mathbf{V}$ from a uniform distribution on the Stiefel manifold. To avoid identifiability challenges associated with values of $\boldsymbol{\Sigma}$ having similar magnitudes, we set $\sigma_k = 2^{-k}\sqrt{P}$ for $k = 1, \ldots, K$. Note that we multiply by $\sqrt{P}$ so that the expectations of $\ell_{kj}^2 = \sigma_k^2 v_{kj}^2$ remain the same regardless of P. We sample the residual variances λ_j^{-1} independently from Gamma$(2, 4)$ for $j = 1, \ldots, P$, which keeps the contribution of the residual variance to the total variance similar to that of the latent factors. Finally, we draw $\boldsymbol{\epsilon} \sim \text{MN}\big(\mathbf{0}, \mathbf{I}_N, \boldsymbol{\Lambda}^{-1}\big)$ and compute $\mathbf{Y} = \mathbf{F}\boldsymbol{\Sigma}\mathbf{V} + \boldsymbol{\epsilon}$. As all methods rely on the same principles for handling missing data, we do not remove any observations from the simulated data sets.

When performing inference, we assume the tree is fixed to its true value used to simulate the factors $\mathbf{F}$. We use the orthogonality constraint on the loadings and employ the post-processing regime discussed in Section 3.1.3 to rotate results from each sampler (except the one associated with the orthogonal shrinkage prior) to enforce this constraint. For the model with the orthogonal shrinkage prior, we assume both forced ordering and spacing ($\alpha = 0.9$).

8.2 Effective Sample Size Calculations

To understand the relative performance of each inference regime, we compare the effective sample size (ESS) per second of the loadings across all four samplers. Draws from an MCMC simulation are often auto-correlated, and the total number of steps in the chain is rarely a direct proxy for our confidence in the posterior estimates. ESS approximates the number of *independent* samples from the chain. As researchers typically set a minimum ESS threshold to determine the length of MCMC simulations, we compare the minimum ESS per unit time. Let $\text{ESS}_{kj}^{(m)}$ be the effective sample size for ℓ_{kj} in replicate m and $\text{ESS}_{\text{min}}^{(m)} = \min_{k,j} \text{ESS}_{kj}^{(m)}$ for $m = 1, \ldots, 3$. We compute $\overline{\text{ESS}}_{\text{min}} = \frac{1}{3} \sum_{m=1}^{3} \text{ESS}_{\text{min}}^{(m)} / t^{(m)}$ for all models, where $t^{(m)}$ is the time required for the m^{th} MCMC simulation. Actual ESS values were calculated using the Julia package MCMCDiagnostics.jl. We compare these values in Figure 1 and SI Table 1.

N	P	K	minimum ESS per minute				speed increase over sampled		
			Sampled	Gibbs	HMC	orthogonal	Gibbs	HMC	orthogonal
50	10	1	530	5100	5700	2000	9.8×	11.0×	3.8×
		2	500	3900	2500	810	7.8×	4.9×	1.6×
		4	680	2200	1400	450	3.3×	2.0×	0.7×
	100	1	190	1400	1700	170	7.6×	9.1×	0.89×
		2	150	1000	870	130	7.1×	5.9×	0.89×
		4	52	550	250	20	11×	4.7×	0.39×
	1000	1	34	460	250	5.2	14×	7.4×	0.15×
		2	27	390	85	0.87	14×	3.1×	0.032×
		4	23	320	23	0.51	14×	1.0×	0.022×
100	10	1	270	4100	3000	1100	15×	11×	4.0×
		2	160	2100	2000	400	13×	12×	2.5×
		4	51	680	500	110	13×	9.9×	2.1×
	100	1	33	360	480	94	11×	14×	2.9×
		2	18	240	290	35	13×	16×	1.9×
		4	17	200	83	38	12×	4.8×	2.2×
	1000	1	3.9	54	53	2.9	14×	14×	0.75×
		2	2.5	82	15	0.98	33×	5.8×	0.39×
		4	2.0	99	5.3	0.19	49×	2.6×	0.092×
500	10	1	5.0	740	460	170	150×	92×	33×
		2	3.4	260	280	59	77×	83×	17×
		4	1.7	160	170	30	93×	98×	18×
	100	1	0.77	95	110	25	120×	140×	32×
		2	0.37	20	28	5.4	56×	77×	15×
		4	0.46	18	12	3.7	40×	25×	8.1×
	1000	1	0.02	1.8	0.71	0.68	90×	35×	34×
		2	0.018	2.4	0.65	0.11	130×	36×	6.1×
		4	0.011	1.5	0.16	0.032	140×	15×	2.9×
1000	10	1	1.1	170	290	58	160×	270×	54×
		2	0.54	84	190	28	160×	350×	52×
		4	0.24	49	80	10	210×	340×	44×
	100	1	0.098	35	38	9.2	350×	390×	94×
		2	0.064	15	12	2.8	230×	180×	44×
		4	0.065	7.6	5.8	1.0	120×	90×	15×
	1000	1	0.0017	0.5	0.25	0.3	300×	150×	180×
		2	0.0015	0.67	0.15	0.085	450×	100×	57×
		4	0.0015	0.4	0.06	0.02	270×	40×	14×

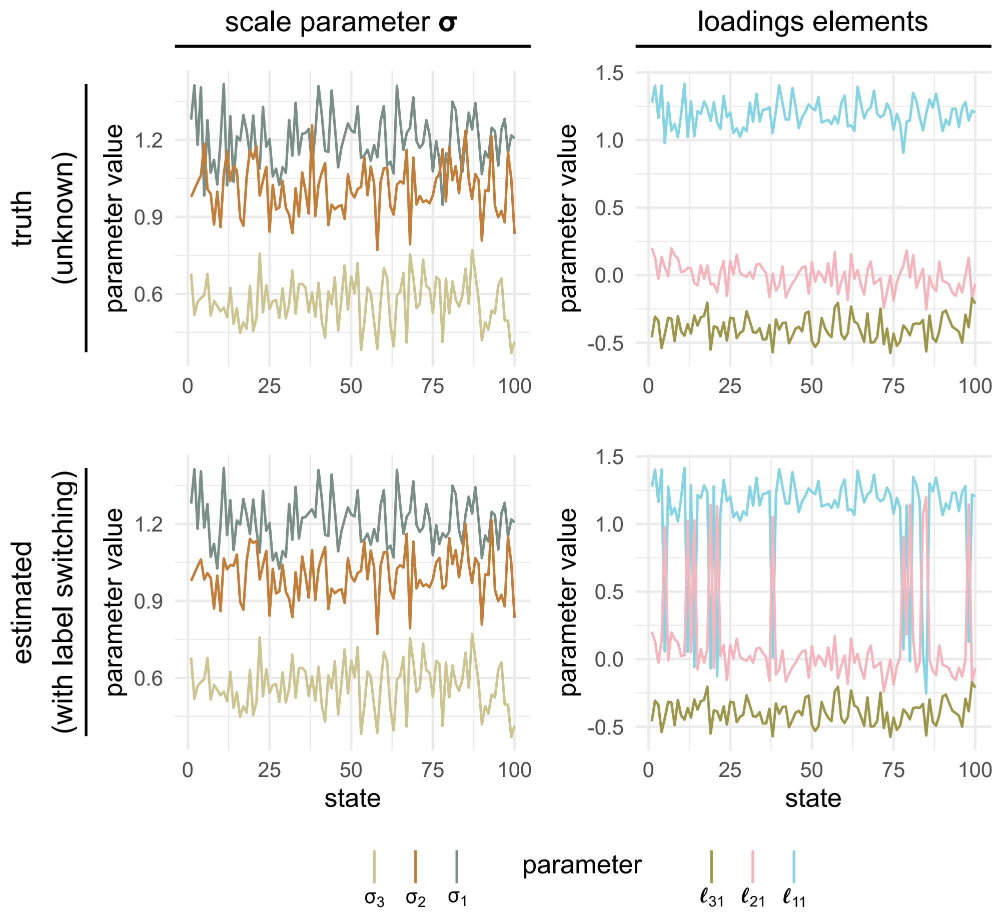
Table 1: Comparison of computational efficiency. Effective sample size computed using the Julia package MCMCDiagnostics.jl.

9 Identifying Label Switching

As discussed in Section 5.2, the post-processing algorithm used to induce orthogonality when sampling under the i.i.d. prior can result in label switching. This phenomenon occurs when elements of the scale parameter $\boldsymbol{\sigma}$ have significantly overlapping posterior distributions. As the post-processing algorithm orders the factors based on the magnitude of the scale parameters, it will swap two factors when their scales switch in order. Because of this, the estimated posterior distribution of the loadings and factor values associated with two factors undergoing label-switching will be a mixture of some (unknown) underlying distributions that we are trying to estimate. This mixing can obscure signals in our data.

Consider SI Figure 1 as a toy example where we know the true underlying distributions. In practice, we do not know these distributions (if we did we wouldn't need to infer them). We assume a 3-factor model where the posterior of scales σ_1 and σ_2 are slightly overlapping. If we then order the rows of the loadings according to the the scales $\boldsymbol{\sigma}$, the estimated rows of the loadings clearly switch at the places where the true $\sigma_1 < \sigma_2$. We see evidence of this occurring in the plot of the loadings where samples from the posterior of ℓ_{11}, which are normally greater than 1, occasionally have unusually low values near 0. At the same points in the chain samples from ℓ_{21}, which are normally near 0, have unusually high values near 1. It appears that the estimated samples from the posterior of ℓ_{11} and ℓ_{21} are occasionally switching between the two.

Label switching is not always as obvious as the simple example depicted here in SI Figure 1. In Figure 2, all elements of $\boldsymbol{\sigma}$ appear close to each other and there is likely a higher degree of overlap between pairs of factors. Rather than obvious switching, the posteriors of the loadings under the i.i.d. prior appear to blend into each other. While it is possible that the posteriors of the loadings really are overlapping, the apparently overlapping scale parameters and skewed tails of the each of loadings posterior densities toward the mean of the other distribution suggests label switching. Repeating the analysis with the orthogonal shrinkage prior reveals distinct posterior distributions in the relevant parameters of the loadings, confirming that label switching is occurring under first analysis (i.i.d. prior with post processing).



Supplementary Figure 1: Example of label switching. The top trace plots are samples from a known distribution. Note that in practice, we to not know the true underlying distribution. The bottom plot demonstrates how ordering the scale parameters can induce label switching between rows of the loadings. Here, there is label switching between the first two factors, but not the third. The switching in the estimated parameters occurs at the MCMC states where the "true" $\sigma_1 < \sigma_2$ (normally the reverse is true).

10 Cross Validation

Our model selection strategy seeks to identify the shrinkage strength (when using the shrinkage prior) or number of factors (when using the i.i.d. prior) that provides optimal predictive performance via cross-validation. To this end, we posit M sub-models characterized by the meta-parameters $\mathbf{\Omega}_1, \ldots, \mathbf{\Omega}_M$. Under the i.i.d. prior, $\mathbf{\Omega}_i = K^{[i]}$ is the number of factors in model i. For example, our default for the i.i.d. prior assumes $K_{\max} = 5$ and $M = 5$ models with $(K^{[1]}, \ldots, K^{[M]}) = (1, 2, 3, 4, 5)$. Under the shrinkage prior, let $\mathbf{\Omega}_i = \{\mathbf{a}^{[i]}, \mathbf{b}^{[i]}\}$ be the shapes and rates, respectively, of the gamma priors on the shrinkage multipliers $\nu_1, \ldots, \nu_K$ for model i. We typically retain $K_{\max} = 5$ and define the 5 sub-models as $\mathbf{a}^{[i]} = 10^{(i+1)/2}\mathbf{1}_{K_{\max}}$

and $\mathbf{b}^{[i]} = \mathbf{1}_{K_{\max}}$ for $i = 1, \ldots, 5$.

We evaluate the predictive performance of each model on R replicate data sets via R-fold cross-validation. For each replicate $j = 1, \ldots, R$, we randomly partition the observed data $\mathbf{Y}^{\text{obs}}$ into a training set $\mathbf{Y}_j^{\text{tr}}$ containing $(100 - \frac{100}{R})\%$ of the data and a validation set $\mathbf{Y}_j^{\text{val}}$ with the remaining $\frac{100}{R}\%$ such that each observation occurs in exactly one validation set.

Let $\mathbf{\Theta} = \{\mathbf{L}, \mathbf{\Lambda}\}$ be the model parameters relevant to the likelihood. We first approximate $p\big(\mathbf{\Theta} \,\big|\, \mathbf{Y}_j^{\text{tr}}, \mathbf{\Omega}_i\big)$ for $i = 1, \ldots, M$, $j = 1, \ldots, R$ via MCMC simulation as described in Section 3. We then compute the expected log predictive density (Gelman et al., 2013) $\pi_{ij} = \mathbb{E}\big[\log p\big(\mathbf{Y}_j^{\text{val}} \,\big|\, \mathbf{Y}_j^{\text{tr}}, \mathbf{\Theta}_{ij}\big)\big]$ for $i = 1, \ldots, M$, $j = 1, \ldots, R$, where $\mathbf{\Theta}_{ij}$ is a random variable with density $p\big(\mathbf{\Theta} \,\big|\, \mathbf{Y}_j^{\text{tr}}, \mathbf{\Omega}_i\big)$. We select $\mathbf{\Omega}_m$, where $m = \arg\max_i \frac{1}{R} \sum_j \pi_{ij}$, as the optimal model and approximate $p\big(\mathbf{L}, \mathbf{\Lambda} \,\big|\, \mathbf{Y}^{\text{obs}}, \mathbf{\Omega}_m\big)$ as the final step in the analysis plan.

11 Phylogenetic Latent Liability Model

In the case of binary traits, we assume the latent liability model of Cybis et al. (2015). Specifically, rather than assuming the observations $\mathbf{Y} = \mathbf{FL} + \boldsymbol{\epsilon}$, we introduce an additional latent variable $\mathbf{Z} = \{z_{ij}\}$ for $i = 1, \ldots, N$, $j = 1, \ldots, P$ and assume $\mathbf{Z} = \mathbf{FL} + \boldsymbol{\epsilon}$. These latent liabilities z_{ij} are connected to the observations y_{ij} via the link function $y_{ij} = g_j(z_{ij})$ where $g_j(x) = x$ if trait j is continuous, $g_j(x) = 1\{x \leq 0\}$ if j is binary.

Under this model, the full conditional distributions of the latent liabilities are independent truncated Gaussian distributions with densities

$$p\big(z_{ij} \,\big|\, y_{ij}, \mathbf{f}_i, \boldsymbol{\ell}_j, \lambda_j, \mathbf{t}_j\big) \sim \theta\big(z_{ij}; \mathbf{f}_i^t \boldsymbol{\ell}_j, \lambda_j\big)\, 1\{g_j(z_{ij}) = y_{ij}\}\,. \tag{24}$$

As these full conditional distributions are independent, we can sample from them efficiently via a simple rejection sampler. Specifically, we first draw from $\mathbf{F} \,|\, \mathbf{Z}, \mathbf{\Lambda}, \mathcal{F}$ as in Section 3.1.1. We then sample the proposal $z_{ij} \sim \mathcal{N}\big(\mathbf{f}_i^t \boldsymbol{\ell}_j, 1/\lambda_j\big)$ that we accept if $g_j(z_{ij}) = y_{ij}$ and reject otherwise. Note that for each discrete trait j, we must also fix $\lambda_j = 1$ to ensure the variance of the latent traits j are identifiable (see Tolkoff et al., 2017).

12 Phylogenetic Tree Inference

12.1 Yeast

For they yeast analysis, we first infer a phylogenetic tree for the 154 phenotyped strains using the 2.8 megabase DNA sequence alignment of Gallone et al. (2016) (see subsection

Phylogenetic Tree for the Sequenced Collection in *Methods* of Gallone et al. (2016) for details). Our phylogenetic tree model includes an uncorrelated relaxed clock model (Drummond et al., 2006), an HKY+G substitution model (Hasegawa et al., 1985; Yang, 1994) and a constant-population coalescent prior on the tree (Kingman, 1982).

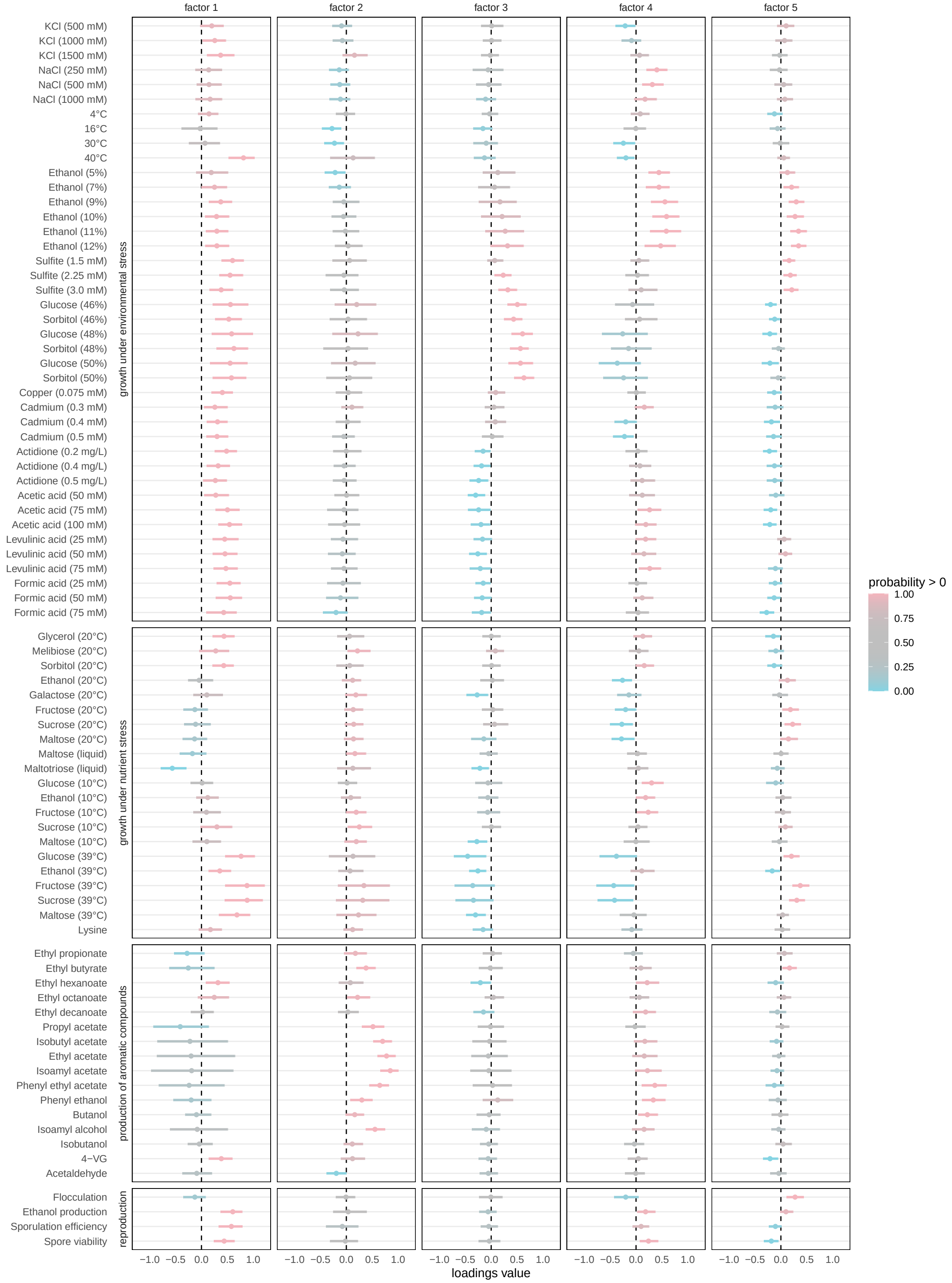
We perform MCMC simulation via BEAST (Suchard et al., 2018) to approximate the posterior distribution of the phylogenetic tree. We run the MCMC chain for 10 million states, sampling the tree and related parameters every thousand states and the factor related parameters every 10 thousand states. Inspection of relevant trace plots indicated the the MCMC chain had achieved stationarity by 1 million states, and we exclude the first million states as burn-in. We compute the maximum clade credibility (MCC) tree as a point estimate of the phylogenetic tree using TreeAnnotator (Rambaut and Drummond, 2015).

12.2 New World Monkeys

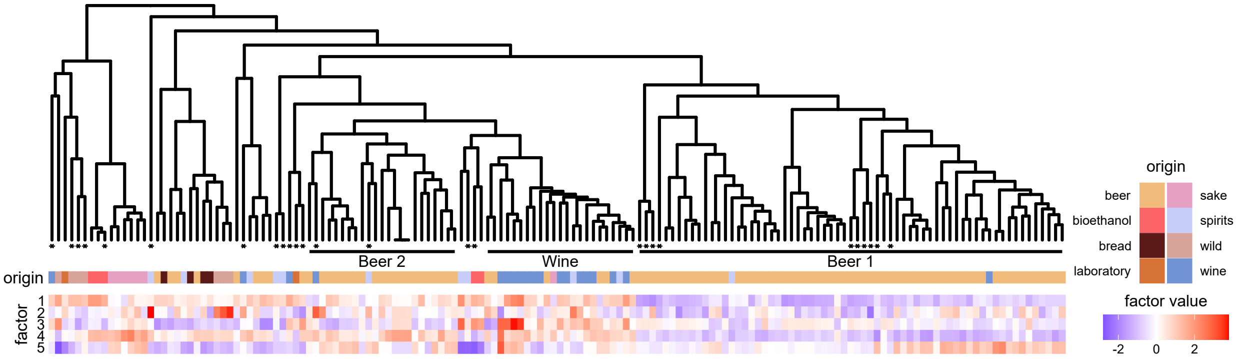
We simultaneously infer the NWM tree structure with the latent factor model using DNA sequence alignments of Aristide et al. (2015). To infer the tree structure, we partition the taxa into four monophyletic clades consisting of the 1) *Atelidae*, 2) *Aotidae* and *Callitrichidae*, 3) *Cebidae* and 4) *Pitheciidae* respectively and place zero prior probability on tree topologies that do not maintain these clades. Otherwise, we use the same phylogenetic tree model and inference procedure as described in SI Section 12.1.

13 Additional Results

We present the full results of our yeast analysis below.



Supplementary Figure 2: Posterior summary of loadings of 5-factor PFA on yeast data set. The first factor primarily captures differences associated with tolerance to environment and nutrient stress as well as reproductive ability. Dots represent posterior means while bars cover the 95% highest posterior density (HPD) interval. Colors represent the posterior probability that the parameter is greater than 0.



Supplementary Figure 3: All five factors plotted on yeast phylogeny with strain origin. Stars at the tips indicate mosaic strains as identified by Gallone et al. (2016). The first factor separates the Beer 1 clade from the remaining strains.

Supplemental References

Aristide, L., A. L. Rosenberger, M. F. Tejedor, and S. I. Perez (2015). Modeling lineage and phenotypic diversification in the New World monkey (Platyrrhini, Primates) radiation. *Molecular Phylogenetics and Evolution 82*, 375–385.

Bastide, P., C. Ané, S. Robin, and M. Mariadassou (2018). Inference of adaptive shifts for multivariate correlated traits. *Systematic Biology 67*(4), 662–680.

Bastide, P., L. S. T. Ho, G. Baele, P. Lemey, and M. A. Suchard (2021). Efficient Bayesian inference of general Gaussian models on large phylogenetic trees. *The Annals of Applied Statistics 15*(2), 971 – 997.

Cybis, G., J. Sinsheimer, T. Bedford, A. Mather, P. Lemey, and M. Suchard (2015). Assessing phenotypic correlation through the multivariate phylogenetic latent liability model. *Annals of Applied Statistics 9*, 969 – 991.

Drummond, A. J., S. Y. Ho, M. J. Phillips, and A. Rambaut (2006). Relaxed phylogenetics and dating with confidence. *PLoS Biology 4*(5), e88.

Fisher, A. A., X. Ji, Z. Zhang, P. Lemey, and M. A. Suchard (2020). Relaxed random walks at scale. *Systematic Biology 70*(2), 258–267.

Gallone, B., J. Steensels, T. Prahl, L. Soriaga, V. Saels, B. Herrera-Malaver, A. Merlevede, M. Roncoroni, K. Voordeckers, L. Miraglia, et al. (2016). Domestication and divergence of *Saccharomyces cerevisiae* beer yeasts. *Cell 166*(6), 1397–1410.

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin (2013). *Bayesian data analysis*. CRC press.

Hasegawa, M., H. Kishino, and T. Yano (1985). Dating of human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution 22*(2), 160–174.

Hassler, G., M. R. Tolkoff, W. L. Allen, L. S. T. Ho, P. Lemey, and M. A. Suchard (2020). Inferring phenotypic trait evolution on large trees with many incomplete measurements. *Journal of the American Statistical Association 0*(0), 1–15.

Henderson, C. R., O. Kempthorne, S. R. Searle, and C. M. von Krosigk (1959, June). The estimation of environmental and genetic trends from records subject to culling. *Biometrics 15*(2), 192–218.

Henderson, H. V. and S. R. Searle (1981). On deriving the inverse of a sum of matrices. *SIAM Reviews 23*(1), 53–60.

Holbrook, A., A. Vandenberg-Rodes, and B. Shahbaba (2016). Bayesian inference on matrix manifolds for linear dimensionality reduction. *arXiv preprint arXiv:1606.04478*.

Kingman, J. F. C. (1982). The coalescent. *Stochastic Processes and their Applications 13*(3), 235–248.

Lopes, H. F. and M. West (2004). Bayesian model assessment in factor analysis. *Statistica Sinica 14*, 41–67.

Mitov, V., K. Bartoszek, G. Asimomitis, and T. Stadler (2020). Fast likelihood calculation for multivariate Gaussian phylogenetic models with shifts. *Theoretical Population Biology 131*, 66–78.

Rambaut, A. and A. Drummond (2015). TreeAnnotator v1. 8.2. *MCMC Output Analysis*.

Suchard, M. A., P. Lemey, G. Baele, D. L. Ayres, A. J. Drummond, and A. Rambaut (2018). Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evolution 4*(1), vey016.

Tolkoff, M. R., M. E. Alfaro, G. Baele, P. Lemey, and M. A. Suchard (2017). Phylogenetic factor analysis. *Systematic Biology 67*(3), 384–399.

Yang, Z. (1994). Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *Journal of Molecular Evolution 39*(3), 306–314.