Universidade Federal do Rio Grande do Sul
Instituto de Matemática e Estatística
Programa de Pós-Graduação em Estatística (PPG-Est)

Sparse precision matrix estimation in phylogenetic trait evolution models

Felipe Grillo Pinheiro

Porto Alegre, Março de 2022.

Dissertação submetida por Felipe Grillo Pinheiro como requisito parcial para a obtenção do título de Mestre em Estatística pelo Programa de Pós-Graduação em Estatística da Universidade Federal do Rio Grande do Sul (UFRGS).

Orientadora:

Dra. Gabriela Bettella Cybis (PPG-Est UFRGS)

Coorientadora:

Dra. Taiane Schaedler Prass (PPG-Est UFRGS)

Banca Examinadora:

Dr. Marc Adam Suchard (UCLA)

Dr. Nelson Jurandi Rosa Fagundes (PPGBM UFRGS)

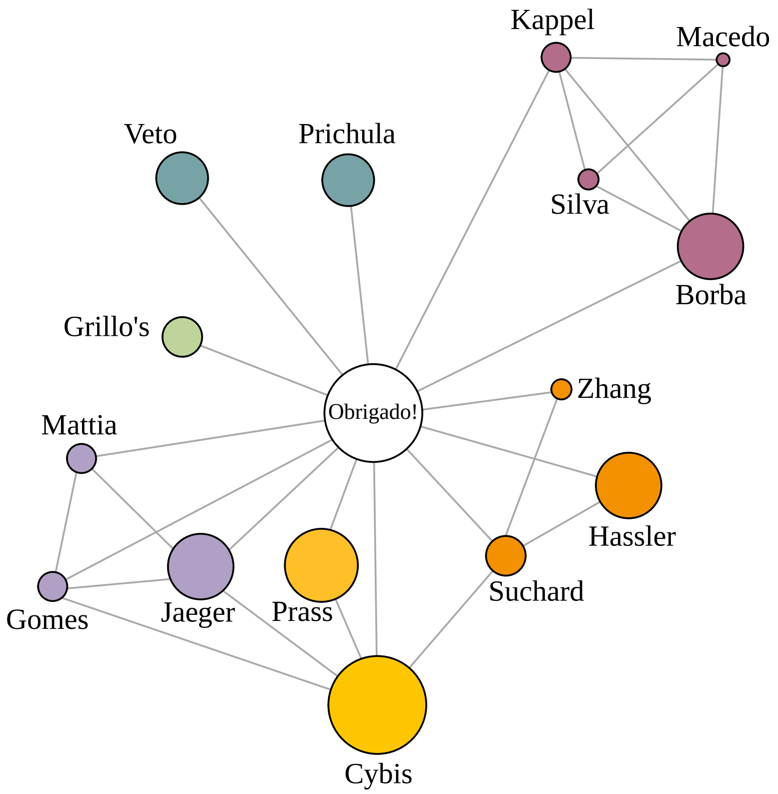
Dr. Guilherme Pumi (PPG-Est UFRGS)

Março, 2022

Dedicatória

Para Davi e Samuel

Agradecimentos



Jaqueline Jaeger, foste o meu principal apoio emocional em todos os (muitos) momentos difícieis desta trajeitória. Minha experiência foi única porque dividi ela contigo. Nossa amizade é extraterrestre.

Laíse Borba, és meu outro pilar e minha terceira irmã. Agradeço por todo cuidado e carinho que sempre recebo de ti. Obrigado Vanessa Kappel, Amanda da Silva e Sandra Macedo pela compreensão e suporte no nosso ambiente trabalho. Agradeço à minha família por compreender a minha ausência.

Arturito Mattia e Ratataela Gomes obrigado por dividirem essa experiência diariamente. Obrigado Nicole Veto por provar que a distância física não define nossa amizade e Janira Prichula por transformar os meus rabiscos nas duas lindas ilustrações científicas que valorizaram muito o manuscrito.

I would like to thank Zhenyu Zhang for helping me in the early stages of this project. I'm forever greatfull to Gabriel Hassler for his incredible support with BEAST. I also thank Marc Suchard for patiently helping me with git commits.

Taiane Prass, tê-la como coorientadora me deu muita segurança durante o desenvolvimento deste trabalho. Obrigado pela dedicação, pelas nossas incríveis reuniões com duração média de 5 horas e pelo auxílio imediato sempre que precisei.

Gabriela Cybis, obrigado por propor este desafio e apostar que eu daria conta. Agradeço por ter me dado a liberdade de propor minhas ideias e, ao mesmo tempo, direcioná-las, impô-las algum limite ou mesmo destrui-las. Sou imensamente grato também pelo tempo dedicado às correções do manuscrito e por todos os ajustes e melhorias propostos.

Resumo

Os modelos filogenéticos para evolução de traços (fenotípicos) permitem a estimativa de correlações evolutivas entre um conjunto de traços observados numa amostra de organismos relacionados. Ao modelar diretamente a evolução dos traços numa árvore filogenética num contexto Bayesiano, a estrutura do modelo nos permite controlar para a história evolutiva compartilhada entre os organismos da amostra e evitar as inferências espúrias originadas pelo parentesco. Nestes modelos, as correlações relevantes são definidas por meio do intervalo de credibilidade das correlações marginais. No entanto, as correlações selecionadas por si só podem não fornecer a melhor informação sobre as relações entre as características em estudo. A sua estrutura de associação, em contraste, fornece uma informação clara sobre associações diretas entre os traços em estudo. A fim de empregar um método baseado em modelo para identificar a estrutura de associação subjacente entre as variáveis, exploramos a utilização de modelos Gaussianos com grafos (GGM) para a seleção das covariâncias. Modelamos a matriz de precisão com a distribuição G-Wishart, uma priori conjugada que resulta em estimativas de precisão esparsa. Avaliamos a nossa abordagem através de simulações de Monte Carlo e comparamos os resultados com o método padrão, onde nenhuma estrutura de associação é explicitamente modelada. Também testamos a nossa abordagem para examinar a estrutura de associação e correlações evolutivas em dois conjuntods de dados: um envolvendo traços fenotípicos dos tentilhões de Darwin e outro envolvendo traços genômicos e fenotípicos de procariotos. A nossa abordagem fornece uma solução sistemática para a eliminação de correlações espúrias e melhor inferência para as matrizes de precisão e correlação, especialmente para as variáveis condicionalmente independentes, que são o alvo da esparsidade nos GGMs. Combinar a inferência das correlações evolutivas e da estrutura de associação permite uma seleção mais precisa das características que potencialmente interagiram ou interagem ao longo do processo evolutivo dos organismos estudados.

Palavras-chave: *inferência bayesiana*, *modelos de evolução de traços filogenéticos*, *modelos gaussianos com grafos*, *filogenética*

Abstract

Phylogenetic trait evolution models allow for the estimation of evolutionary correlations between a set of traits observed in a sample of related organisms. By directly modeling the evolution of the traits on a phylogenetic tree in a Bayesian framework, the model's structure allows us to control for shared evolutionary history between the organisms in the sample and avoid spurious inference that could have been originated from common ancestors. In these models, relevant correlations are obtained through the high posterior density interval of marginal correlations. However, the selected correlations alone may not provide the best information regarding trait relationships. Their association structure, in contrast, provide straightforward information about direct associations. In order to employ a model based method to identify the underlying association structure between the variables we explore the use of Gaussian graphical models (GGM) for covariance selection. We model the precision matrix with a G-Wishart conjugate prior which results in sparse precision estimates. We evaluate our approach through Monte Carlo simulations and compare the results to the standard method, where no association structure is explicitly modeled. We also test our approach to examine the association structure and evolutionary correlations of Darwin's finches phenotypic traits and prokaryotic genomic and phenotypic traits. Our approach provides a systematic solution for elimination of spurious correlations and better inference for the precision and correlation matrices, especially for conditionally independent variables, which are the target for sparsity in GGMs. Combining correlation and association structure inference allows for a more precise selection of candidate traits that may interact along the evolutionary process of related organisms.

Keywords: *Bayesian inference*, *Trait evolution model*, *Gaussian graphical models*, *phylogenetics*

Index

Chapter 1

Resumo Expandido

Entender como características genotípicas e fenotípicas interagem é um dos grandes desafios da biologia evolutiva. Eventos como duplicação, deleção, mutação e recombinação de sequencias genômicas são cruciais para a adaptação dos organismos vivos às mudanças ambientais. Neste cenário, as filogenias ajudam a compreender e correlacionar estes fenômenos com a evolução dos traços fenotípicos (McClintock, 1984; Vulić et al., 1999; Archibald and Roger, 2002; Logares et al., 2007; Shan and Li, 2008).

Em estudos comparativos, naturalmente surge o interesse em avaliar as relações entre um conjunto de variáveis de interesse. Estimar e identificar correlações relevantes entre traços fenotípicos colabora com o entendimento dos processos evolutivos potencialmente existentes entre as variáveis em estudo. Entretanto, para estimar correlações de forma adequada é necessário separá-las daquelas induzidas pelo parentesco genético, isto é, remover os efeitos espúrios gerados aleatoriamente ao longo da história evolutiva compartilhada pelos organismos em estudo — que podem gerar vieses nas estimativas destas correlações. Os modelos filogenéticos para evolução de traços fenotípicos são alternativas para medir correlações evolutivas livres do efeito do parentesco.

Para modelar conjuntamente as correlações evolutivas destas características ao longo de uma árvore filogenética desconhecida — mas estimável —, vários modelos filogenéticos para a evolução de características fenotípicas tem sido propostos nos últimos anos, com base no *Threshold model* de Felsenstein (2012), (Cybis et al., 2015; Hassler et al., 2020; Zhang et al., 2021). Estes modelos assumem variáveis latentes contínuas não observadas para cada organismo amostrado que surgem por meio de um processo de difusão Browniano multivariado (MBD) ao longo de uma árvore filogenética inferida a partir de sequências moleculares. Este modelo MBD é caracterizado por uma matriz de precisão $\mathbf{K} = \mathbf{\Sigma}^{-1}$, a inversa da matriz de covariâncias, da qual são obtidas as correlações evolutivas entre as características fenotípicas $\mathbf{R}$, o parâmetro principal nestes modelos

A correlação do processo de difusão informa a correlação entre parâmetros latentes, que é uma aproximação para a correlação entre os traços fenotípicos observados. A correlação do processo de difusão pode ser vista como o efeito combinado de fatores genéticos relevantes (por exemplo, deriva genética) que afetam os traços observados após o ajuste para a história evolutiva compartilhada entre os taxa.

O *Threshold model* proposto por (Felsenstein, 2012) adapta o processo de difusão browniano para permitir a estimativa de correlação entre traços fenotípicos binários e contínuos. O objetivo de Cybis et al. (2015) foi transpor este modelo para o contexto Bayesiano e expandi-lo ao dar origem ao modelo filogenético de variável latente — *Phylogenetic Multivariate Latent Liability Model* (PMLLM) — para a evolução de traços de tipo misto, contabilizando dados contínuos, binários, categóricos e ordinais. A principal contribuição de Zhang et al. (2021) foi desenvolver uma estrutura de inferência eficiente, chamado modelo de modelo probit filogenético multivariado — *Phylogenetic Multivariate Probit Model* (PMPM) —, baseado em um amostrador que importa ideias da física partículas (BPS), para amostrar os parâmetros latentes a partir de uma distribuição normal truncada com dimensionalidade, melhorando assim a eficiência do modelo em comparação com o esquema de MCMC em Cybis et al. (2015). Além disso, Hassler et al. (2020) ampliou o mecanismo de amostragem das variáveis latentes para computar a verossimilhança de forma eficiente sob um cenário de traços contínuos com dados incompletos, o que permite considerar estimativas de árvores muito maiores.

Apesar dos esforços para melhorar a eficiência e expandir a aplicabilidade do modelo, os coeficientes de correlação relevantes tem sido determinados pela avaliação da sua distribuição posteriori marginal utilizando um intervalo credibilidade (*High posterior density* (HPD) *interval*). Uma vez que, em muitos problemas, é esperado que apenas uma pequena porção dos traços observados seja realmente interligada, é desejável controlar para sinais falsos positivos e evitar estimativas espúrias — para além daquelas que já foram controladas por meio da árvore filogenética — especialmente em problemas de alta dimensionalidade onde o número de características a serem estudadas p é potencialmente grande.

Uma solução sistemática natural é estimar uma matriz de precisão esparsa no processo de difusão browniano. O padrão de esparsidade desta matriz reflete na matriz de correlação correspondente, potencialmente fixando em zero alguns dos elementos fora da sua diagonal principal, conforme desejado. Uma forma de obter a esparsidade é condicionando a matriz de precisão ao espaço de matrizes positivas definidas com entradas zero consistentes com um grafo que retrata a estrutura de dependência entre os traços (Dempster, 1972). Para variáveis gaussianas multivariadas com matriz de precisão $\mathbf{K} = \{k_{ij}\}$, tais como as variáveis latentes $\mathbf{X}$ oriundas do MBD, esta estrutura de associação pode ser traduzida pela (in)dependência condicional embutida na matriz de precisão. Uma entrada $k_{ij} = 0$ implica que as variáveis correspondentes são condicionalmente independentes dadas todas as outras variáveis do modelo (Li et al., 2020; Mitsakakis, 2010; Talhouk et al., 2012).

Os modelos gaussianos com grafo — *Gaussian graphical models* (GGM) — são ferramentas convenientes para modelar relações de (in)dependência condicional entre variáveis (Carvalho and Scott, 2009). Um GGM é um modelo probabilístico no qual a estrutura de (in)dependência condicional de $\mathbf{K}$ é também representada por um grafo $\mathbf{G}$ (Atay-Kayis and Massam, 2005; Letac and Massam, 2007; Mohammadi and Wit, 2015). Neste contexto, as entradas não-zero fora da diagonal principal de **K** correspondem também às arestas existentes no grafo não-direcionado **G**. Neste grafo, as variáveis são representadas pelos vértices (ou nós) e a presença ou ausência de arestas indica se existe ou não uma associação direta entre elas, ou, mais tecnicamente, representa a sua (in)dependência condicional dadas as outras variáveis.

Na inferência Bayesiana, a distribuição G-Wishart é a priori conjugada para matrizes de precisão estruturadas em variáveis com distribuição normal multivariada. Por esta razão, é uma escolha conveniente (Boom et al., 2021; Williams, 2021) e muitas abordagens diferentes foram propostas para calcular ou aproximar a sua constante de normalização, que é um grande desafio e o principal gargalo na eficiência computacional dos GGMs.

A fim de utilizar um método baseado em modelos para identificar esta estrutura de associação, exploramos a utilização de modelos Gaussianos com grafos (GGM) para a seleção de covariâncias. Neste estudo, propomos uma abordagem Bayesiana, num modelo chamado *Sparse Phylogenetic Trait Evolution Model* (SPTE), para inferência de uma matriz de precisão esparsa **K**, adaptando o modelo de difusão browniano para traços contínuos ao contexto da seleção de covariâncias. Ao fazê-lo, introduzimos outro parâmetro de interesse, o grafo do processo de difusão **G** que complementa as informações fornecidas pelas estimativas de correlação evolutivas tradicionalmente estimadas nos modelos filogenéticos de evolução de traços fenotípicos. Implementamos nosso modelo no módulo de desenvolvimento do software BEAST (Drummond et al., 2012).

Realizamos dois estudos de simulação para comparar o desempenho do nosso modelo esparso com o do modelo filogenético tradicional de evolução de traços fenotípicos (modelo completo), em que a (in)dependência condicional não é explicitamente modelada. Adicionalmente, aplicamos o modelo esparso e completo para examinar a estrutura de associação e a correlação evolutiva dos traços fenotípicos dos tentilhões de Darwin, bem como das características genômicas e fenotípicas de um conjunto massivo de espécies de procariotos.

Com base em estudos de simulação e aplicação, nossa abordagem melhora significativamente os modelos tradicionais de evolução de traços fenotípicos em termos de modelagem e inferência. Nosso modelo fornece melhores estimativas para a matriz de precisão e de correlação, especialmente para variáveis independentes — que são o alvo principal da esparsidade —, ao mesmo tempo que exibe o erros quadráticos médios (EQM) semelhantes para variáveis dependentes. Além disso, nosso modelo pode identificar com precisão a estrutura de associação entre os traços fenotípicos, o que realça as vantagens de uma abordagem baseada no modelo para a seleção de covariância.

Ao aplicar o modelo tradicional podemos apenas identificar traços significativamente correlacionados, discutir a força de suas correlações, e usá-las para orientar a procura de potenciais explicações — num correlograma possivelmente denso. Por outro lado, mais do que simplesmente inferir as correlações evolutivas, o modelo esparso informa também sobre a estrutura de associação entre os traços, que é codificada no grafo estimado. A estrutura de associação pode ajudar a refinar a procura de mecanismos potenciais para explicar as (in)dependências condicionais reveladas pelo grafo subjacente às dependências apresentadas pelos correlogramas.

Os resultados das nossas aplicações indicam que a combinação da informação das correlações com as independências condicionais do grafo, no entanto, permite uma seleção mais precisa dos traços candidatos a interagir ao longo do processo evolutivo dos organismos relacionados. Por este motivo, a aprendizagem da estrutura da associação é imperativa, particularmente quando a dimensão dos traços fenotípicos p aumenta.

Outra vantagem importante da nossa abordagem é que, com as adaptações adequadas, ela pode ser integrada com uma vasta gama de modelos filogenéticos gaussianos, devido à sua facilidade de utilização. Sob uma perspectiva computacional, incluindo a estimativa gráfica no modelo MBD apenas requer alterações no mecanismo de atualização da matriz de precisão para obter conjuntamente $p(\mathbf{K}, \mathbf{G})$. Portanto, como as alterações consistem essencialmente em modificações na escolha das prioris e hiperprioris, para além de todo o mecanismo de atualização do grafo, as abordagens empregadas para os cálculos de probabilidade permanecem intactas. Esta é uma característica desejável e conveniente porque permite que nossa abordagem potencialmente se beneficie de qualquer melhoria computacional futura em modelos de evolução de traços fenotípicos.

Por exemplo, no estudo com os procariotos fomos capazes de efetuar a seleção de covariância neste conjunto massivo de dados, baseando-nos na abordagem eficiente desenvolvida por Hassler et al. (2020) que integra os valores faltantes e permite análises previamente intratáveis em grandes árvores. Embora não exploremos isto em simulações ou aplicação, o modelo de evolução de traço filogenético esparso pode ser ainda adaptado para lidar com dados binários, categóricos e ordinais como em Cybis et al. (2015); Zhang et al. (2021), o que apenas acrescentará à ampla aplicabilidade do modelo.

Chapter 2

Phylogenetics

Bayesian phylogenetic methods for trait evolution were developed to infer the evolution between phenotypic traits while simultaneously controlling for shared evolutionary history of sampled organisms. In this dissertation, we propose the estimation of an additional parameter — the diffusion graph **G** —, in the context of Bayesian trait evolution models. Although there is still room for straightforward improvements on computational efficiency, the project presented here provides a rich contribution to modeling and inference in phylogenetic trait evolution models and can be easily coupled with new methodologies and benefit from future contributions in the field.

In the following sections we provide a short introduction on phylogenetic trees and how they are estimated in a Bayesian context. Additionally, we present the basic ideas underlying conditional independence in Gaussian variables and explain how they are connected to a Gaussian graphical model that aims to perform covariance selection through a graph.

2.1 Phylogenetic Trees

It is well-known that the genetic material of every living organism undergoes mutations over time and that part of that variation become fixed (Cordero and Janzen, 2013; Cvijović et al., 2015; Hössjer et al., 2021). Species that arise from a common ancestor accumulate distinct mutations over time (Lemey et al., 2009; Carlin, 2011), and the number of accumulated mutations tends to be proportional to the divergence time between species (Safran and Nosil, 2012)

Phylogeny is the study of the history of evolutionary relationships between genes, species or populations which is represented by a tree diagram that explains ancestral relationships (Nixon, 2001) through the connection of adjacent nodes (which represent studied organisms or sequences) along branches, such as lines that interconnect these nodes. Elucidating the evolutionary history of genes and species is one of the goals of molecular evolution (Gillespie, 1994; Nixon, 2001; Leliaert et al., 2012). For that purpose, molecular phylogeny methods make it possible to reconstruct, from a set of nucleotide or protein sequences, the history of successive divergences that cause interspecific genetic variations during evolution of related taxa from a common ancestor (Nixon, 2001). The reconstruction of phylogenetic trees is a statistical challenge, as the accuracy of tree estimates also depends on available statistical methods. In this context, molecular phylogenetics has focused on improving models for estimating phylogenetic trees based on sequence alignments. Thus, phylogenies help to understand and correlate different evolutionary phenomena and environment associations to the evolution of phenotypic traits (McClintock, 1984; Vulić et al., 1999; Archibald and Roger, 2002; Logares et al., 2007; Shan and Li, 2008).

A phylogenetic tree is an acyclic graph with N nodes of degree 1 — the node is only connected to another one —, representing the N organisms in the sample. These nodes are denoted by $\nu_1, \ldots, \nu_N$ and are usually termed as tips. The tree also has $N-2$ internal nodes of degree 3, denoted by $\nu_{N+1}, \ldots, \nu_{2N-2}$, that represent common ancestors to two or more organisms in the sample. Additionally, the tree may have one root node of degree 2 denoted by ν_{2N-1}, representing the most recent common ancestor of all N organisms, and we say that the tree is rooted. The branch lengths of the tree $\boldsymbol{t} = (t_1, \ldots, t_{2N-2})$ on the edges of a rooted tree represent elapsed evolutionary time between two nodes. Figure 2.1 presents an example rooted tree with $N = 3$ tips.

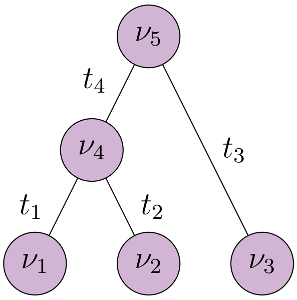


Figure 2.1: Example rooted tree with $N = 3$ tips

Phylogenetic methods use sequence data $\mathbf{S}$ to estimate a phylogenetic tree topology $\mathscr{T}$ that represents the evolutionary relationship between N related organisms. The $N \times L$ sequence matrix $\mathbf{S} = \{s_{ij}\}$ contains N aligned DNA or RNA sequences of length L from each of the organisms in the sample. In order to estimate the tree $\mathscr{T}$ from sequence data, we require a model for computing the probabilities of changes in the molecular sequences over evolutionary time. For each site of the molecular sequence, this process is usually modelled by a continuous time Markov chain (CTMC) defined by a infinitesimal rate matrix $\mathbf{Q}$ from which the transition probabilities between the DNA/RNA basis {A,G,C,T/U} can be obtained. The Markovian property of the base substitution process implies that, after two lineages split, their mutation processes are independent, given their most common recent ancestor. Propagation of this property throughout the tree leads to the tree likelihood for site j in the sequences $\mathbf{S}_j$, where j indicates the j-th column of genetic sequences. The tree likelihood $\mathcal{L}(\mathbf{Q}, \mathbf{t}, \mathscr{F}|\mathbf{S}_j) = p(\mathbf{S}_j|\mathbf{Q}, \mathbf{t}, \mathscr{F})$ computes the probability of the data for site j, given the molecular evolution process on the tree. Note that as we do not observe the internal nodes, we need to integrate over all possible base combinations for these nodes. This likelihood is computed using a pruning algorithm that traverses the tree in post order, keeping track of conditional probabilities, and evaluates the likelihood through $\mathcal{O}(N)$ operations (Felsenstein, 1981).

To obtain the likelihood for the whole matrix $\mathbf{S}$, one must assume a model for molecular evolution across sites. Assuming all sites to be independent and identically distributed, we compute the overall likelihood as

$$p(\mathbf{S}|\mathbf{Q}, \mathbf{t}, \mathscr{F}) \propto \prod_{j=1}^{L} p(\mathbf{S}_j|\mathbf{Q}, \mathbf{t}, \mathscr{F}). \tag{2.1}$$

This independent and identically distributed model is oversimplified, but it serves as a didactic illustration of how the likelihood is calculated for the sequence data (Felsenstein and Felenstein, 2004).

2.2 Bayesian inference of phylogenetic trees

In a Bayesian analysis, inference on the phylogeny is based upon the posterior probability $p(\boldsymbol{\theta}|\mathbf{S})$ of parameters $\boldsymbol{\theta}$, given the sequence data $\mathbf{S}$ (Huelsenbeck and Ronquist, 2001). Here $\boldsymbol{\theta}$ collects all the phylogenetic parameters, e.g. $\boldsymbol{\theta} = \{\mathbf{Q}, \mathbf{t}, \mathscr{F}\}$.

Through Bayes theorem, the posterior can be computed as

$$p(\boldsymbol{\theta}|\mathbf{S}) = \frac{p(\mathbf{S}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{S})}, \tag{2.2}$$

where $p(\boldsymbol{\theta})$ is the prior distribution representing our knowledge about $\boldsymbol{\theta}$ and the normalizing constant $p(\mathbf{S})$ is the marginal likelihood of the data $\mathbf{S}$. The likelihood $p(\mathbf{S}|\boldsymbol{\theta})$ of the molecular evolution process can be obtained through expression (2.1).

To compute the posterior in (2.2), we would also need an expression for the normalizing constant $p(\mathbf{S})$, which can be computed as the integral

$$p(\mathbf{S}) = \int p(\mathbf{S}|\boldsymbol{\theta})p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}. \tag{2.3}$$

However, since $\boldsymbol{\theta} = \{\mathbf{Q}, \mathbf{t}, \mathscr{F}\}$, evaluating (2.2) requires integrating over the space of all possible tree topologies, possible branch length combinations and base substitution parameters. Bayesian phylogenetic inference generally relies on Markov chain Monte Carlo (MCMC) due to the computational intractability of $p(\mathbf{S})$.

2.3 Markov chain Monte Carlo

Monte Carlo integration is a simulation method for estimating multidimensional integrals. Suppose we wish to estimate the expected value of $h(\boldsymbol{\theta})$, then

$$\mathbb{E}(h(\boldsymbol{\theta})|\mathbf{S}) = \int h(\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{S})d\boldsymbol{\theta}, \tag{2.4}$$

where $p(\boldsymbol{\theta}|\mathbf{S})$ is the posterior distribution of $\boldsymbol{\theta}$. If we cannot analytically evaluate the integral, random samples $\boldsymbol{\theta}^{(1)}, \dots, \boldsymbol{\theta}^{(m)}$ from the distribution $p(\boldsymbol{\theta}|\mathbf{S})$ can be used to estimate $h(\boldsymbol{\theta})$ as

$$\widehat{\mathbb{E}(h(\boldsymbol{\theta}))} = \frac{1}{m}\sum_{i=1}^{m} h(\boldsymbol{\theta}^{(i)}). \tag{2.5}$$

The samples can also be used to obtain the variance of the estimates and marginal distributions on individual components of $\boldsymbol{\theta}$.

However, for phylogenetic models it is generally not straightforward to generate samples from the posterior distribution $p(\boldsymbol{\theta}|\mathbf{S})$. MCMC methods use Markov chains to generate dependent samples of the target distribution. These chains are constructed to be ergodic and have equilibrium distribution $p(\boldsymbol{\theta}|\mathbf{S})$. Consequently, the process is asymptotically guaranteed to achieve the target distribution.

The construction of ergodic Markov chains with the correct stationary distribution is central to MCMC. The two most used methods for producing these chains are the Metropolis-Hastings method (Metropolis et al., 1953; Hastings, 1970) and the Gibbs sampler (Geman and Geman, 1984). Neither method requires the evaluation of the normalizing constant in expression (2.3) to generate samples from the posterior distribution.

Metropolis-Hastings algorithms rely on a two step procedure to generate consecutive posterior samples for $\boldsymbol{\theta}$. First a new state $\boldsymbol{\theta}^p$ is proposed according to a proposal distribution $q_{\theta^{(k)}}(\boldsymbol{\theta}^p)$, that usually depends on the current state $\boldsymbol{\theta}^{(k)}$. Then, the new state is accepted $\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^p$, with probability

$$A(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}^p) = \min\left\{1, \frac{q_{\theta^p}(\boldsymbol{\theta}^{(k)})p(\boldsymbol{\theta}^p|\mathbf{S})}{q_{\theta^{(k)}}(\boldsymbol{\theta}^p)p(\boldsymbol{\theta}^{(k)}|\mathbf{S})}\right\} \tag{2.6}$$

or rejected $\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)}$. Note that only the ratio of posterior probabilities is required for this evaluation because the normalizing constants $p(\mathbf{S})$ cancel out.

The Gibbs sampler is a special case of the Metropolis-Hastings algorithm (Brooks et al., 2011). In a Gibbs update the proposal is from a conditional distribution of the desired equilibrium distribution, therefore, it is always accepted. Gibbs samplers divide the parameter $\boldsymbol{\theta}$ into M components $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_M)$, and update each individual component $\boldsymbol{\theta}_m$ at a time. New samples for each $\boldsymbol{\theta}_m$ are drawn from their conditional distribution $p(\boldsymbol{\theta}_m|\boldsymbol{\theta}_{-m}, \mathbf{S})$, where $\boldsymbol{\theta}_m = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_{m-1}, \boldsymbol{\theta}_{m+1}, \dots, \boldsymbol{\theta}_M)$ represents all other component parameters in $\boldsymbol{\theta}$.

In these complex phylogenetic models, however, full conditional distributions are not always available for all the parameter components. Metropolis-Hastings algorithm can be used to generate samples for individual parameter components for which a Gibbs sampler is not available. This approach produces a "Metropolis-within-Gibbs" sampler, in which some parameter components are updated based on full conditional probabilities, and others are updated using Metropolis-Hastings algorithm. The phylogenetic methods presented in this dissertation exploit the flexibility of this combination of Metropolis and Gibbs approach.

Chapter 3

Covariance Selection and Gaussian Graphical Models

3.1 Covariance Selection

The principle of parsimony in parametric model fitting posits that parameters should be introduced sparingly and only when the data indicate they are required (Dempster, 1972). There should be a trade-off between costs and benefits to avoid suffering from underfitting — by ignoring factors or variables that should be included in the model (misspecification) —, and overfitting — by including too many redundant or unnecessary parameters. When facing high dimensional problems, it is often useful to impose a structure of association between analysed variables. Additionally, researchers may be interested in comparing different hypotheses of patterns of association, or the data might follow a natural grouping, where certain subsets of variables are likely to express higher degree of association with each other and less association with other groups of variables (Talhouk et al., 2012).

Consider a phenomenon modeled with a p-dimensional multivariate normal distribution $\mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma})$, then, imposing a structure on the covariance $\mathbf{\Sigma}$ (covariance selection) or correlation matrix $\mathbf{R}$ (correlation selection) may help to shrink dimensionality of the cost of estimating a covariance or correlation for every possible pair of variables when, in fact, a significant part of them is expected to be independent and uncorrelated, i.e. constrained to zero, — specially when the number of variables p is large. In sampling schemes, however, directly constraining $\mathbf{\Sigma}$ and $\mathbf{R}$ entries to zero may violate the positive-definiteness of these matrix estimates. Alternatively, an easier task would be to impose sparsity in their inverses, the precision matrix $\mathbf{K} = \mathbf{\Sigma}^{-1}$ and the partial correlation $\mathbf{R}^{-1}$. The zero pattern in $\mathbf{K}$ and $\mathbf{R}^{-1}$ can be potentially inherited by $\mathbf{\Sigma}$ and $\mathbf{R}$ depending on matrix structure, e.g. block diagonal matrices. On the other hand, sparsity in precision and partial correlations has a different interpretation as it represents the (in)conditional dependence structure — association structure — between the analysed variables in Gaussian models. Thereafter, although we refer to such a covariance/correlation selection problem, it is actually their inverses ($\mathbf{K}$ and $\mathbf{R}^{-1}$), that display zero constrained off-diagonal elements related to the association structure (Gaskins, 2019).

For Gaussian variables, Dempster (1972) proposed a parameter reduction scheme by setting off-diagonal elements of the precision matrix $\mathbf{K}$ to zero. This results in a pattern entries constraint to zero in $\mathbf{K}$, and is called covariance selection model or Gaussian graphical model as it represents a pairwise conditional independence structure (Lenkoski and Dobra, 2008). In this case, zero entries in $\mathbf{K}$ indicates that corresponded variables are conditionally independent (CI) given the remaining variables in the model, whereas conditionally dependent variables (CD) imply direct associations between them also given the rest.

Thus, imposing a structure of association may help achieve better modeling and inference — and computational efficiency, depending on model context —, particularly when some degree of sparsity is expected in the association structure between variables of scientific interest. Also, the amount of noise in a fitted model due to errors of estimation is substantially reduced when we favor parameter reduction. This is because the number of free parameters to be estimated is a smaller subset of the total number of parameters under a "full model", where all possible associations are estimated (Dempster, 1972).

3.2 *Conditional independence*

We recall an important property of the Gaussian distribution that connects conditional independence with the precision matrix. As in Mitsakakis (2010), consider a p-dimensional random vector $\mathbf{X} = (X_1, \ldots, X_p)^t$ and the set of variable indices $V = \{1, \ldots, p\}$ such that $\mathbf{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \mathbf{K}^{-1})$. Then $\mathbf{K}$, written as

$$\mathbf{K} = \begin{bmatrix} k_{11} & \ldots & k_{1p} \\ \vdots & \ddots & \vdots \\ k_{p1} & \ldots & k_{pp} \end{bmatrix},$$

is the precision matrix where, the components X_i and X_j are conditionally independent given the rest of the components $(X_h)_{h \in V \setminus \{i,j\}}$ if and only if $k_{ij} = 0$. Here the $\setminus$ symbol represent the set difference. This can easily be seen if we consider the joint conditional distribution of (X_i, X_j), given $(X_h)_{h \in V \setminus \{i,j\}}$, known to have a bivariate normal distribution with covariance matrix

$$\boldsymbol{\Sigma}_{ij.V \setminus \{i,j\}} = \begin{bmatrix} k_{ii} & k_{ij} \\ k_{ji} & k_{jj} \end{bmatrix}^{-1} = \frac{1}{k_{ii}k_{jj} - k_{ij}^2} \begin{bmatrix} k_{jj} & -k_{ji} \\ -k_{ij} & k_{ii} \end{bmatrix}.$$

Therefore X_i and X_j are conditionally independent given $(X_h)_{h \in V \setminus \{i,j\}}$ if and only if

$$\left[\boldsymbol{\Sigma}_{ij.V \setminus \{i,j\}}\right]_{12} = \left[\boldsymbol{\Sigma}_{ij.V \setminus \{i,j\}}\right]_{21} = 0,$$

i.e. if and only if $k_{ij} = k_{ji} = 0$. In summary

$$\forall\ i, j \in V,\ X_i \perp\!\!\!\perp X_j | X_{V \setminus \{i,j\}} \Leftrightarrow k_{ij} = 0.$$

After understanding the link between conditional independence and the zero pattern in the precision matrix $\mathbf{K}$, in the following section we introduce the Gaussian Graphical Models and show how the association structure embedded in $\mathbf{K}$ is modeled through a graph.

3.3 Gaussian Graphical Models

Gaussian graphical models provide a convenient framework for imposing a conditional independence structure of association between variables (Mohammadi and Wit, 2015; Mohammadi et al., 2021; Williams, 2021). Here we introduce some notation and the structure of undirected Gaussian graphical models and show how to perform covariance selection for multivariate Gaussian variables $\mathbf{X}$. We refer the interested reader to Lauritzen (1996) for detailed information.

Let $\mathbf{G} = (V, E)$ be an undirected graph, where $V = \{1, 2, \dots, p\}$ is a finite set of vertices (or nodes) and $E \subset V \times V$ is the set of existing edges. The vertices V represent the variables $\mathbf{X}$ measured for each observation. Also let

$$\mathcal{W} = \{(i, j) \mid i, j \in V,\ i < j\}$$

and $\overline{E} = W \setminus E$ denotes the set of non-existing edges. Without loss of generality, as in Mohammadi and Wit (2015) and Letac et al. (2017), we define a zero mean Gaussian graphical model (GGM) with respect to the graph $\mathbf{G}$ as

$$\mathcal{M}_G = \{\mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma}), \mathbf{K} = \mathbf{\Sigma}^{-1} \in \mathbb{P}_G\}, \tag{3.1}$$

where $\mathbb{P}_G$ is the space of $p \times p$ positive definite matrices with zero entries (i, j) consistent with $\mathbf{G}$. Let $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n)$ be an independent and identically distributed sample of size n from model $\mathcal{M}_G$. Then, the likelihood in GGM is given by

$$p(\mathbf{X}|\mathbf{K}, \mathbf{G}) = \frac{|\mathbf{K}|^{n/2}}{(2\pi)^{np/2}} \exp\left\{-\frac{1}{2}\mathrm{tr}(\mathbf{UK})\right\}, \tag{3.2}$$

where $\mathbf{U} = \mathbf{X}'\mathbf{X}$. The joint posterior distribution for $\mathbf{K}$ and $\mathbf{G}$, in a Bayesian context, can be then factored as

$$p(\mathbf{K}, \mathbf{G}|\mathbf{X}) \propto p(\mathbf{X}|\mathbf{K}, \mathbf{G})p(\mathbf{K}|\mathbf{G})p(\mathbf{G}). \tag{3.3}$$

For simplicity, we can set a discrete uniform distribution over the graph space $\mathcal{G}$ —

the space of all graphs with p edges — for the prior distribution of the graph,

$$p(\mathbf{G}) = \frac{1}{|\mathcal{G}|}, \tag{3.4}$$

for each $\mathbf{G} \in \mathcal{G}$, where $|\mathcal{G}|$ is the cardinality of the graph space given by $|\mathcal{G}| = 2^{p(p-1)/2}$. For the prior distribution on the structured precision matrix $p(\mathbf{K}|\mathbf{G})$, we can use the G-Wishart distribution (Roverato, 2002; Atay-Kayis and Massam, 2005). The G-Wishart distribution $\mathcal{W}_G(\delta, \mathbf{D})$ has density

$$f_G(\mathbf{K}; \delta, \mathbf{D}) = \frac{1}{I_G(\delta, \mathbf{D})} |\mathbf{K}|^{(\delta-2)/2} \exp\left\{ -\frac{1}{2} \mathrm{tr}(\mathbf{DK}) \right\} \mathbf{1}_{\mathbf{K} \in \mathbb{P}_G}, \tag{3.5}$$

with parameters δ and $\mathbf{D}$, where $\delta > 0$ represents the degrees of freedom (or shape parameter), $\mathbf{D}$ is a symmetric positive-definite rate matrix, $\mathbf{1}_{\mathbf{K} \in \mathbb{P}_G}$ is the indicator function that $\mathbf{K}$ is restricted to $\mathbb{P}_G$, and $I_G(\delta, \mathbf{D})$ is the normalizing constant,

$$I_G(\delta, \mathbf{D}) = \int_{\mathbf{K} \in \mathbb{P}_G} |\mathbf{K}|^{(\delta-2)/2} \exp\left\{ -\frac{1}{2} \mathrm{tr}(\mathbf{DK}) \right\} d\mathbf{K}. \tag{3.6}$$

When $\mathbf{G}$ is complete or decomposable, we have explicit formulas for the normalizing constant $I_G(\delta, \mathbf{D})$ (Roverato, 2002). For non-decomposable graphs we can approximate $I_G(\delta, \mathbf{D})$ using the Monte Carlo method of Atay-Kayis and Massam (2005) or the Laplace approximation of Lenkoski and Dobra (2011). Note that the G-Wishart prior (3.5) is conjugate to the likelihood in (3.2), conditional on graph $\mathbf{G}$. Therefore the posterior distribution of $\mathbf{K}|\mathbf{G}$ is also G-Wishart $\mathcal{W}_G(\delta^\star, \mathbf{D}^\star)$ where $\delta^\star = \delta + n$ and $\mathbf{D}^\star = \mathbf{D} + \mathbf{U}$.

Under the G-Wishart $\mathcal{W}_G(\delta, \mathbf{D})$ conjugate prior on $\mathbf{K}|\mathbf{G}$ for the Gaussian variables $\mathbf{X}$, the joint density of $(\mathbf{X}, \mathbf{K}, \mathbf{G})$ is

$$\begin{aligned} p(\mathbf{X}, \mathbf{K}, \mathbf{G}) &= p(\mathbf{X}|\mathbf{K}, \mathbf{G}) p(\mathbf{K}|\mathbf{G}) p(\mathbf{G}) \\ &= \frac{|\mathbf{K}|^{n/2}}{(2\pi)^{np/2}} \exp\left\{ -\frac{1}{2} \mathrm{tr}(\mathbf{UK}) \right\} \frac{1}{I(\delta, \mathbf{D})} |\mathbf{K}|^{(\delta-2)/2} \exp\left\{ -\frac{1}{2} \mathrm{tr}(\mathbf{DK}) \right\} \mathbf{1}_{\mathbf{K} \in \mathbb{P}_G} \frac{1}{|\mathcal{G}|} \\ &= \frac{1}{(2\pi)^{np/2}} \frac{1}{|\mathcal{G}|} \frac{1}{I_G(\delta, \mathbf{D})} |\mathbf{K}|^{(\delta^\star-2)/2} \exp\left\{ -\frac{1}{2} \mathrm{tr}(\mathbf{D}^\star \mathbf{K}) \right\} \mathbf{1}_{\mathbf{K} \in \mathbb{P}_G}. \end{aligned} \tag{3.7}$$

The marginal likelihood of data $\mathbf{X}$ given the graph $\mathbf{G}$ is given by

$$p(\mathbf{X}|\mathbf{G}) = \frac{p(\mathbf{X}, \mathbf{G})}{p(\mathbf{G})} = \frac{\int_{\mathbf{K} \in \mathbb{P}_G} p(\mathbf{X}, \mathbf{K}, \mathbf{G})\, d\mathbf{K}}{p(\mathbf{G})}. \tag{3.8}$$

By replacing the kernel of the integral in Equation (3.8) by the joint density (3.7) we have

$$
\begin{aligned}
p(\mathbf{X}|\mathbf{G}) &= |\mathcal{G}| \frac{1}{(2\pi)^{np/2}} \frac{1}{|\mathcal{G}|} \frac{1}{I_G(\delta, \mathbf{D})} \int_{\mathbf{K} \in \mathbb{P}_G} |\mathbf{K}|^{(\delta^\star - 2)/2} \exp\left\{ -\frac{1}{2} \mathrm{tr}(\mathbf{D}^\star \mathbf{K}) \right\} \mathbf{1}_{\mathbf{K} \in \mathbb{P}_G} \, d\mathbf{K} \\
&= \frac{1}{(2\pi)^{np/2}} \frac{I_G(\delta^\star, \mathbf{D}^\star)}{I_G(\delta, \mathbf{D})},
\end{aligned}
\tag{3.9}
$$

and the posterior density of $\mathbf{G}$ given $\mathbf{X}$ is

$$
p(\mathbf{G}|\mathbf{X}) \propto p(\mathbf{G}) p(\mathbf{X}|\mathbf{G}) = \frac{p(\mathbf{G})}{(2\pi)^{np/2}} \frac{I_G(\delta^\star, \mathbf{D}^\star)}{I_G(\delta, \mathbf{D})}. \tag{3.10}
$$

Computing the marginal likelihood (3.8) or the posterior distribution (3.10) is reduced to the problem of computing normalising constants of the type $I_G(\delta, \mathbf{D})$, with $\delta > 0$ and $\mathbf{D}$ positive definite, which are sufficient conditions for convergence of the normalizing constants, i.e. $I_G(\delta, \mathbf{D}) < \infty$ (Mitsakakis, 2010, Lemma 3.2.1).

MCMC methods for the posterior of the graph structure often aim to select graph proposals with higher posterior probability $p(\mathbf{G}|\mathbf{X})$ (Atay-Kayis and Massam, 2005; Mitsakakis et al., 2011). To select a new graph the posterior probability of the new candidate, $\mathbf{G}_p$ is compared with the posterior probability of the "current state" graph $\mathbf{G}_c$. To perform graph selection we start with a current graph $\mathbf{G}_c$, randomly select the index of two vertices and add or delete the correspondent edge in order to switch its current value and propose a new graph $\mathbf{G}_p$. Notice that $\mathbf{G}_p$ and $\mathbf{G}_c$ differ only by a single edge. The proposed graph $\mathbf{G}_p$, can be then accepted according to the following Metropolis-Hastings (MH) acceptance probability

$$
\alpha = \min\left\{1, \frac{p(\mathbf{G}_p|\mathbf{X})}{p(\mathbf{G}_c|\mathbf{X})}\right\} = \min\left\{1, \frac{I_{G_p}(\delta^\star, \mathbf{D}^\star)}{I_{G_c}(\delta, \mathbf{D})} \frac{I_{G_c}(\delta^\star, \mathbf{D}^\star)}{I_{G_p}(\delta, \mathbf{D})}\right\}. \tag{3.11}
$$

The new graph is then used to sample from $\mathcal{W}_G(\delta^\star, \mathbf{D}^\star)$ the posterior distribution of $p(\mathbf{K}|\mathbf{G}, \mathbf{X})$.

Chapter 4

Wishart Family Distributions

In GGM literature, a complete definition for the parametrization of Wishart and G-Wishart distribution is often not provided. Wishart and inverse Wishart distributions can be viewed as the generalization of gamma and inverse gamma distributions to multiple dimensions. In Bayesian analysis of multivariate Gaussian variables, inverse Wishart and Wishart are also the standard conjugate prior distributions for canonical covariance $\mathbf{\Sigma}$ and precision matrices $\mathbf{K} = \mathbf{\Sigma}^{-1}$, respectively, regarding a full model. Here we refer as full model a model that assumes no particular association structure for the variables or, equivalently, in a GGM context, a model where a full graph is assumed $\mathbf{G}_{Full} = \{g_{ij} = 1\}$ for $1 \leq i < j \leq p$.

Both distributions are parametrized in terms of degrees of freedom or shape, and scale matrix parameters. However, both distributions can be presented in two different parametrizations concerning their degrees of freedom. Additionally, the scale matrix can be expressed as inverse scale or rate matrix in the Wishart distribution as an analogy to the rate and scale parameters in gamma and inverse gamma distribution parametrizations. Here we single out both parametrizations for Wishart and inverse Wishart distributions and their relation in terms of degrees of freedom (shape), i.e. the induced distribution for $\mathbf{K}$ when $\mathbf{\Sigma}$ follows an inverse Wishart distribution and vice versa. We also compare Wishart and G-Wishart distributions in order to elucidate their equivalence when the graph $\mathbf{G}$ is complete.

4.1 Wishart and inverse Wishart

The usual parametrization for the Wishart and inverse Wishart (Muirhead, 1982; Gelman et al., 2013) is in terms of degrees of freedom ν. For the inverse Wishart distribution, we write $\mathbf{\Sigma} \sim IW(\nu, \mathbf{S})$, with support on the space of $p \times p$ positive definite matrices, and density

$$f_p(\mathbf{\Sigma}; \nu, \mathbf{S}) = \frac{|\mathbf{S}|^{\nu/2}}{2^{\nu p/2}\Gamma_p\left(\frac{\nu}{2}\right)} |\mathbf{\Sigma}|^{-(\nu+p+1)/2} \exp\left\{-\frac{1}{2}\mathrm{tr}(\mathbf{\Sigma}^{-1}\mathbf{S})\right\}, \tag{4.1}$$

where $\nu > p-1$ is the degrees of freedom (or shape parameter), $\mathbf{S}$ is a symmetric, positive definite $p \times p$ scale matrix, and $\Gamma(\alpha)$ is the multivariate gamma function that has the form $\Gamma_p(\alpha) = \pi^{p(p-1)/4} \prod_{i=1}^{p} \Gamma\left(\alpha - \frac{1}{2}(i-1)\right)$. On the other hand, the Wishart distribution, also with support on the space of $p \times p$ positive definite matrices, has density

$$f_p(\mathbf{K}; \nu, \mathbf{S}) = \frac{|\mathbf{S}|^{-\nu/2}}{2^{\nu p/2}\Gamma_p\left(\frac{\nu}{2}\right)} |\mathbf{K}|^{(\nu-p-1)/2} \exp\left\{-\frac{1}{2}\mathrm{tr}(\mathbf{K}\mathbf{S}^{-1})\right\}, \tag{4.2}$$

with degrees of freedom (or shape parameter) $\nu > p-1$ and symmetric positive definite $p \times p$ scale matrix $\mathbf{S}$.

From Dawid (1981) and Roverato (2000) we have another parametrization of Wishart and inverse Wishart distributions in terms of the shape parameter δ that relates to ν according to

$$\nu = \delta + p - 1. \tag{4.3}$$

By replacing $\nu = \delta + p - 1$ in density (4.1), we write $\mathbf{\Sigma} \sim IW(\delta, \mathbf{S})$ with density

$$f_p(\mathbf{\Sigma}; \delta, \mathbf{S}) = \frac{|\mathbf{S}|^{(\delta+p-1)/2}}{2^{(\delta+p-1)p/2}\Gamma_p\left(\frac{\delta+p-1}{2}\right)} |\mathbf{\Sigma}|^{-(\delta+2p)/2} \exp\left\{-\frac{1}{2}\mathrm{tr}(\mathbf{\Sigma}^{-1}\mathbf{S})\right\}, \tag{4.4}$$

with degrees of freedom (or shape parameter) $\delta > 0$ and symmetric positive-definite $p \times p$ scale matrix $\mathbf{S}$. Similarly, using density (4.2), we say $\mathbf{K} \sim W(\delta, \mathbf{S})$ and define the Wishart density in terms of δ as

$$f_p(\mathbf{K}; \delta, \mathbf{S}) = \frac{|\mathbf{S}|^{-(\delta+p-1)/2}}{2^{(\delta+p-1)p/2}\Gamma_p\left(\frac{\delta+p-1}{2}\right)} |\mathbf{K}|^{(\delta-2)/2} \exp\left\{-\frac{1}{2}\mathrm{tr}(\mathbf{K}\mathbf{S}^{-1})\right\}, \tag{4.5}$$

with degrees of freedom (or shape) $\delta > 0$ and symmetric, positive definite $p \times p$ scale matrix $\mathbf{S}$.

4.2 Wishart and inverse Wishart relationship

The parametrization using δ is useful because the distribution induced by $\mathbf{\Sigma} \sim IW(\delta = \delta, \mathbf{S})$ on $\mathbf{K} = \mathbf{\Sigma}^{-1}$ is Wishart, $\mathbf{K} \sim W(\nu = \delta + p - 1, \mathbf{S}^{-1})$ (Dawid, 1981; Roverato, 2000). Notice that $\mathbf{S}^{-1}$ is the inverse of the scale parameter $\mathbf{S}$ of Wishart and inverse Wishart distributions in parametrizations (4.2) and (4.1) and can also be referred to as inverse scale or rate matrix. If we take $\mathbf{D} = \mathbf{S}^{-1}$, we can express the Wishart density (4.5) in terms of shape parameter δ and rate matrix $\mathbf{D}$, i.e. $\mathbf{K} \sim W(\delta, \mathbf{D})$, with density

$$f_p(\mathbf{K};\delta,\mathbf{D}) = \frac{|\mathbf{D}|^{(\delta+p-1)/2}}{2^{(\delta+p-1)p/2}\Gamma_p\left(\frac{\delta+p-1}{2}\right)}|\mathbf{K}|^{(\delta-2)/2}\exp\left\{-\frac{1}{2}\mathrm{tr}(\mathbf{KD})\right\}, \tag{4.6}$$

with degrees of freedom $\delta > 0$ and symmetric positive-definite $p \times p$ rate (or inverse scale) matrix $\mathbf{D}$. The parametrization of the Wishart distribution in (4.6) is convenient because it emphasises the similarities between Wishart and G-Wishart distributions regarding a complete graph $\mathbf{G}_{Full}$.

4.3 G-Wishart distribution

The G-Wishart distribution was fist proposed by Roverato (2000) who derived its density from the hyper inverse Wishart distribution (Dawid and Lauritzen, 1993). We write $\mathbf{K} \sim \mathcal{W}_G(\delta,\mathbf{D})$, with support on the space of $p \times p$ positive definite matrices, and density

$$f_G(\mathbf{K};\delta,\mathbf{D}) = \frac{1}{I_G(\delta,\mathbf{D})}|\mathbf{K}|^{(\delta-2)/2}\exp\left\{-\frac{1}{2}\mathrm{tr}(\mathbf{DK})\right\}\mathbf{1}_{\mathbf{K}\in\mathbb{P}_G}, \tag{4.7}$$

with parameters δ and $\mathbf{D}$, where $\delta > 0$ (Mitsakakis, 2010, Lemma 3.2.1: $I_G(\delta,\mathbf{D}) < \infty$ for $\delta > 0$) is the degrees of freedom (or shape), $\mathbf{D}$ is a symmetric positive definite rate matrix, $\mathbb{P}_G$ is the space of $p \times p$ positive definite matrices with zero entries (i,j) whenever an edge is missing in the graph $\mathbf{G}$ and $I_G(\delta,\mathbf{D})$ is the normalizing constant as in Equation (3.6). Note that when $\mathbf{G}$ is a complete graph, the G-Wishart $\mathcal{W}_G(\delta,\mathbf{D})$ reduces to a Wishart distribution with rate matrix parametrization $\mathcal{W}_p(\delta,\mathbf{D})$ as in equation (4.6) and the normalizing constant becomes

$$I_G(\delta,\mathbf{D}) = 2^{(\delta+p-1)p/2}|\mathbf{D}|^{-(\delta+p-1)/2}\pi^{p(p-1)/4}\prod_{i=1}^{p}\Gamma\left(\frac{\delta+p-i}{2}\right), \tag{4.8}$$

which is equivalent to the normalizing constant in equation (4.6). Importantly, the rate parameter $\mathbf{D}$ in the G-Wishart distribution is also equivalent to the inverse of the scale matrix $\mathbf{S}$ in Wishart parametrization (4.5), i.e. $\mathbf{D} = \mathbf{S}^{-1}$. Table 4.1 summarizes the parametrizations of the aforementioned distributions.

Table 4.1: Parametrizations of Wishart, inverse Wishart and G-Wishart distributions in terms of degrees of freedom or shape parameter (ν, δ) and scale matrix $\mathbf{S}$ or rate matrix $\mathbf{D}$ parameters. The equivalence between the degrees of freedom (shape) is $\nu = \delta + p - 1$, and $\mathbf{D} = \mathbf{S}^{-1}$.

Distribution	Notation Parameters	Density function
Inverse Wishart	$\mathbf{\Sigma} \sim W(\nu, \mathbf{S})$ ν degrees of freedom $\mathbf{S}$ $p \times p$ Scale matrix	$f_p(\mathbf{\Sigma}; \nu, \mathbf{S}) = \frac{\lvert\mathbf{S}\rvert^{\nu/2}}{2^{\nu p/2}\Gamma_p\left(\frac{\nu}{2}\right)} \lvert\mathbf{\Sigma}\rvert^{-(\nu+p+1)/2} \exp\left\{-\frac{1}{2}\mathrm{tr}(\mathbf{\Sigma}^{-1}\mathbf{S})\right\}$
Inverse Wishart	$\mathbf{\Sigma} \sim W(\delta, \mathbf{S})$ δ degrees of freedom $\mathbf{S}$ $p \times p$ scale matrix	$f_p(\mathbf{\Sigma}; \delta, \mathbf{S}) = \frac{\lvert\mathbf{S}\rvert^{(\delta+p-1)/2}}{2^{(\delta+p-1)p/2}\Gamma_p\left(\frac{\delta+p-1}{2}\right)} \lvert\mathbf{\Sigma}\rvert^{-(\delta+2p)/2} \exp\left\{-\frac{1}{2}\mathrm{tr}(\mathbf{\Sigma}^{-1}\mathbf{S})\right\}$
Wishart	$\mathbf{K} \sim W(\nu, \mathbf{S})$ ν degrees of freedom $\mathbf{S}$ $p \times p$ scale matrix	$f_p(\mathbf{K}; \nu, \mathbf{S}) = \frac{\lvert\mathbf{S}\rvert^{-\nu/2}}{2^{\nu p/2}\Gamma_p(\nu/2)} \lvert\mathbf{K}\rvert^{(\nu-p-1)/2} \exp\left\{-\frac{1}{2}\mathrm{tr}(\mathbf{K}\mathbf{S}^{-1})\right\}$
Wishart	$\mathbf{K} \sim W(\delta, \mathbf{S})$ δ degrees of freedom $\mathbf{S}$ $p \times p$ scale matrix	$f_p(\mathbf{K}; \delta, \mathbf{S}) = \frac{\lvert\mathbf{S}\rvert^{-(\delta+p-1)/2}}{2^{(\delta+p-1)p/2}\Gamma_p\left(\frac{\delta+p-1}{2}\right)} \lvert\mathbf{K}\rvert^{(\delta-2)/2} \exp\left\{-\frac{1}{2}\mathrm{tr}(\mathbf{K}\mathbf{S}^{-1})\right\}$
Wishart	$\mathbf{K} \sim W(\delta, \mathbf{D})$ δ degrees of freedom $\mathbf{D}$ $p \times p$ rate matrix	$f_p(\mathbf{K}; \delta, \mathbf{D}) = \frac{\lvert\mathbf{D}\rvert^{(\delta+p-1)/2}}{2^{(\delta+p-1)p/2}\Gamma_p\left(\frac{\delta+p-1}{2}\right)} \lvert\mathbf{K}\rvert^{(\delta-2)/2} \exp\left\{-\frac{1}{2}\mathrm{tr}(\mathbf{K}\mathbf{D})\right\}$
G-Wishart	$\mathbf{K} \sim \mathcal{W}_G(\delta, \mathbf{D})$ δ degrees of freedom $\mathbf{D}$ $p \times p$ rate matrix $I_G(\delta, \mathbf{D})$ is the normalizing constant $\mathbb{P}_G$ is the space of $p \times p$ positive definite matrices with zero entries (i, j) whenever an edge is missing in the graph G	$f_p(\mathbf{K}; \delta, \mathbf{D}) = \frac{1}{I_G(\delta,\mathbf{D})} \lvert\mathbf{K}\rvert^{(\delta-2)/2} \exp\left\{-\frac{1}{2}\mathrm{tr}(\mathbf{K}\mathbf{D})\right\} \mathbf{1}_{\mathbf{K}\in\mathbb{P}_G}$

Chapter 5

Paper Pinheiro, Prass and Cybis

Sparse precision matrix estimation in phenotypic trait evolution models

Felipe G. Pinheiro[1], Taiane S. Prass[1], and Gabriela B. Cybis[1]

[1]*Department of Statistics, Federal University of Rio Grande do Sul, RS, Brazil*

April 29, 2022

Abstract

Phylogenetic trait evolution models allow for the estimation of evolutionary correlations between a set of traits observed in a sample of related organisms. By directly modeling the evolution of the traits on a phylogenetic tree in a Bayesian framework, the model's structure allows us to control for shared evolutionary history. In these models, relevant correlations are assessed through a post-process procedure based on the high posterior density interval of marginal correlations. However, the selected correlations alone may not provide all available information regarding trait relationships. Their association structure, in contrast, is likely to express some sparsity pattern and provide straightforward information about direct associations between traits. In order to employ a model-based method to identify this association structure we explore the use of Gaussian graphical models (GGM) for covariance selection. We model the precision matrix with a G-Wishart conjugate prior, which results in sparse precision estimates. We evaluate our approach through Monte Carlo simulations and applications that examine the association structure and evolutionary correlations of phenotypic traits in Darwin's finches and genomic and phenotypic traits in prokaryotes. Our approach provides accurate graph estimates and lower errors for the precision and correlation parameter estimates, especially for conditionally independent traits, which are the target for sparsity in GGMs.

1 Introduction

Estimating correlations between a set of traits evolving along a phylogenetic tree has been the focus of recent developments aiming to elucidate the association structure of the data. Here we explicitly include the association structure, through a graph parameter, in a Bayesian phylogenetic trait evolution model in order to improve modeling, inference, and our ability to unravel complex trait relationships.

Phylogenetic trait evolution models are important tools for investigating the evolutionary associations between a set of phenotypic and genotypic traits controlling for the shared evolutionary history of related organisms. Failing to control for this evolutionary history can mislead the inference of correlation between these traits as some of the associations could be simply a result from common ancestry rather than having an specific adaptive meaning. To jointly model the trait correlation evolution along an unknown tree, several versions of phylogenetic trait evolution models have been proposed over the last few years building upon the Brownian diffusion process and the threshold model proposed by Felsenstein (2012) (Cybis et al., 2015; Hassler et al., 2020; Zhang et al., 2021).

These models assume unobserved continuous latent variables for each tip taxon that arise from a multivariate Brownian diffusion (MBD) process along a phylogenetic tree inferred from molecular sequences. This MBD model is characterized by a precision matrix $\mathbf{K} = \mathbf{\Sigma}^{-1}$, the inverse covariance, from which the evolutionary trait correlations are obtained. The diffusion correlation is the parameter of scientific interest and informs the correlation between latent parameters, which is a proxy for the desired correlation between observed phenotype traits. The diffusion correlation can be viewed as the combined effect of relevant genetic factors (e.g. selective pressures, genetic linkage) that couples the evolution of observed traits after adjusting for the taxa shared evolutionary history.

The threshold model of Felsenstein (2012) adapted the MBD process to allow for evolutionary correlation estimation among binary and continuous phenotypic traits. The aim of Cybis et al. (2015) was to bring the threshold model to a Bayesian perspective and extend it to create the phylogenetic multivariate latent liability model (PMLLM) for the evolution of mixed-type traits, accounting for continuous, binary, categorical, and ordinal data.

The primary contribution of Zhang et al. (2021) was to develop an efficient inference framework, called phylogenetic multivariate probit model (PMPM), based on the bouncy particle sampler (BPS), to sample the latent parameters from a high-dimensional truncated normal distribution, thus improving mixing and efficiency compared to the MCMC scheme in Cybis et al. (2015). Although limited to continuous and binary outcomes, PMPM also takes advantage of a separation strategy over the covariance matrix that avoids the previous identifiability issue of using the Wishart distribution as the conjugate prior on the diffusion precision in a mixed-type trait model.

Additionally, Hassler et al. (2020) extended the sampling mechanism for the latent variables to provide a highly efficient likelihood computation under incomplete continuous trait data scenarios. This approach tremendously increases the available information used for phylogenetic trait correlation estimation by allowing the inclusion of both taxa with incomplete trait measurements and trait information not yet available for all taxa.

Despite the efforts to improve efficiency and expand the model's applicability, the significance of correlation coefficients is still determined by the evaluation of their marginal posterior distribution using a high posterior density (HPD) interval. Because, in many problems, only a small portion of the observed traits are actually expected to be interconnected in the underlying biology, it is desirable to control for false positive signals and avoid estimation of spurious correlation coefficients, specially for high-dimensional problems where the number of traits p is large. Moreover, in current practice, the selected correlations alone may not provide all available information regarding trait relationships. Their association structure — which explores the conditional independence —, in contrast, is likely to express some sparsity pattern and provide straightforward information about direct associations between traits. Additionally, one may want to favor a sparse representation of target parameters purely based on modeling motivations.

A natural systematic solution for this is to estimate a sparse diffusion precision matrix. The sparsity pattern on the diffusion precision reflects on the corresponding correlation matrix potentially shrinking some of its off-diagonals towards zero, as desired. One way to obtain sparsity is by conditioning the precision matrix to the space of positive definite matrices with zero entries consistent with a graph that depicts the dependency structure between traits (Dempster, 1972). In addition to impacting corresponding trait correlations, sparse precision estimates, in combination with a graph parameter, would also allow us to explore the underlying association structure between traits, i.e. their conditional (in)dependence structure.

For multivariate Gaussian variables with precision matrix $\mathbf{K} = \{k_{ij}\}$, such as the latent parameters, this structure of association can be translated by the conditional (in)dependence embedded in the precision matrix. Then entry $k_{ij} = 0$ implies that the corresponding variables are conditionally independent given all the other variables in the model (Li et al., 2020; Mitsakakis, 2010; Talhouk et al., 2012). We refer the interested reader to Maathuis et al. (2018, Ch.9) for detailed proofs on the connection between conditional independence and the zero pattern in the precision matrix.

Gaussian graphical models (GGM) are convenient tools for modeling conditional (in)dependence relationships among variables (Carvalho and Scott, 2009). A GGM is a probabilistic model in which the conditional (in)dependence structure on $\mathbf{K}$ is also represented by a graph $\mathbf{G}$ (Atay-Kayis and Massam, 2005; Letac and Massam, 2007; Mohammadi and Wit, 2015). In this context, non-zero entries in the off-diagonal of $\mathbf{K}$ also correspond to the existing edges in the undirected graph $\mathbf{G}$. In this graph, variables are represented by the vertices (or nodes) and the presence or absence of edges indicates whether exists a direct association between them, or, more technically, represents their conditional (in)dependence given the other variables. Because of the intricate structure between elements in $\mathbf{K}$, both $\mathbf{\Sigma}$ and the corresponding correlation matrix $\mathbf{R}$ do not always inherit the exact same zero pattern from the diffusion precision. Therefore, although we refer to such as covariance selection, it is actually the precision matrix that has the zero elements related to conditional independence which are being directly modeled (Gaskins, 2019).

In Bayesian inference, the G-Wishart distribution is the conjugate prior for structured precision matrices of a multivariate normal distribution. Because of its conjugacy the G-Wishart is a convenient choice (Boom et al., 2021; Williams, 2021) despite the fact that its normalizing constant does not have an analytical form for a general non-decomposable graph $\mathbf{G}$ (Boom et al., 2021). The normalizing constant is required for G-Wishart density computations and plays an essential role in model selection — the search for graphs (models) with high posterior density. For this reason, many different approaches have been proposed to compute or approximate the challenging $I_G(\delta, \mathbf{D})$. Some approaches focused on the advantages of estimating the normalizing constant for decomposable graphs (Letac and Massam, 2007) — the constant has closed-form in this context —, whereas others provide the pathway to its generalization to non-decomposable graphs (Roverato, 2000, 2002; Atay-Kayis and Massam, 2005), further improving or avoiding its calculation (Lenkoski and Dobra, 2011; Letac et al., 2017; Mohammadi and Wit, 2015), or developing closed-form expressions for specific graph configurations (Uhler et al., 2018).

In order to employ a model-based method to identify this association structure we explore the use of Gaussian graphical models (GGM) for covariance selection. In this study, we propose a Bayesian approach, called Sparse Phylogenetic Trait Evolution Model (SPTE), for inference of a sparse precision matrix $\mathbf{K}$ by adapting the MBD model for continuous traits to the context of covariance selection. By estimating the association structure through the graph we aim to benefit from: i) a systematic solution for elimination of spurious correlations between phenotypic traits; ii) parameter reduction, which is a major gain since the number of pairwise correlations scales quadratically in trait dimension; iii) subsequent error reduction in diffusion precision and correlation estimation; and iv) improving our ability to explore complex relationships between continuous traits.

The reminder of this paper proceeds as follows: In Section 2, we first describe the MBD process on phylogenetic trees, further expand it to account for covariance selection using GGM, and finally define the our SPTE model. In Section 3, we describe the inference framework for the MCMC and present a joint sampling scheme for the graph and sparse precision matrix. Section 4 presents the results from two simulation studies conducted to compare the performances of SPTE (also referred to as sparse model) to the ones of the traditional phylogenetic trait evolution model (full model). Then, we apply both sparse and full model to examine the association structure and evolutionary correlation of Darwin's finches phenotypic traits and prokaryote genomic and phenotypic growth properties in Section 5. Lastly, we evaluate the computational cost of our model in Section 6, and discuss the advantages, limitations and future directions of SPTE along with concluding remarks in Section 7.

2 Modeling

In this section we show how we connect the continuous p-dimensional observed traits $\mathbf{Y}$ and correspondent latent variables $\mathbf{X}$ to the phylogenetic tree $\mathscr{F}$ using the multivariate Brownian diffusion model to control for shared evolutionary history of N analysed taxa in the estimation of trait correlations $\mathbf{R}$. We explain how we adapt the MBD model to the context of covariance selection to finally present the Sparse Phylogenetic Trait Evolution Model (SPTE). We also provide details on graph estimation (model selection) and sampling strategy for our Bayesian inference framework.

2.1 ***Multivariate Brownian diffusion on Trees.*** Consider a data set of N aligned molecular sequences $\mathbf{S}$ from related organisms and N continuous p-dimensional trait observations $\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_N)^t$, where $\mathbf{Y}_k = (Y_{k1}, \ldots, Y_{kp})^t$ for $k = 1, \ldots, N$. 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 node set $\mathbb{V}$ and branch lengths $\mathbf{t}$. The node set $\mathbb{V} = (v_1, \ldots, v_{2N-1})$ contains N tip nodes of degree-1 $(v_1, \ldots, v_N)$, $N-2$ internal nodes of degree 3 $(v_{N+1}, \ldots, v_{2N-2})$ and one root node v_{2N-1} 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.

We associate each node h in $\mathscr{F}$ with a latent variable $\mathbf{X}_h \in \mathbb{R}^p$ for $h = 1, \ldots, 2N-1$. A multivariate Brownian diffusion process on $\mathscr{F}$ characterizes the evolutionary relationship between latent variables $\mathbf{X}$ and acts conditionally independently along each branch, such that $\mathbf{X}_h$ is multivariate normally distributed,

$$\mathbf{X}_h \sim \mathcal{N}_p(\mathbf{X}_{\text{pa}(h)}, t_h \mathbf{K}^{-1}), \tag{1}$$

centered at realized value $\mathbf{X}_{\text{pa}(h)}$, where $\text{pa}(h)$ denotes the parent node of h, with variance proportional to a $p \times p$ positive definite covariance matrix $\mathbf{K}^{-1}$ that is shared by all branches of $\mathscr{F}$. At the tips of $\mathscr{F}$, we collect the $N \times p$ matrix $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_N)^t$, where $\mathbf{X}_k = (X_{k1}, \ldots, X_{kp})^t$ for $k = 1, \ldots, N$, and map it to the observed traits $\mathbf{Y}$ through a stochastic link $p(\mathbf{Y}_k|\mathbf{X}_k)$. The form of the stochastic link varies depending on the nature of $\mathbf{Y}$. In the case of continuous traits, $p(\mathbf{Y}_k|\mathbf{X}_k)$ has a degenerate density at $\mathbf{X}_k$ (i.e. $\mathbf{Y}_k = \mathbf{X}_k$ with probability 1), when there are no missing data. When trait measurements are missing, we employ the pre-order missing data augmentation algorithm developed by Hassler et al. (2020, Section 2.2.1) in order to compute likelihood presented in (3). For models with mixed-type traits, such as PMLLM of Cybis et al. (2015) or PMPM of Zhang et al. (2021), $p(\mathbf{Y}_k|\mathbf{X}_k)$ stems from a deterministic mapping function $\mathbf{Y}_k = g(\mathbf{X}_k)$, that maps the elements of the continuous latent variables $\mathbf{X}_k$ at the tips to its corresponding categories for non-continuous traits. Nevertheless, here we restrict our model to continuous traits only. This is because allowing for mixed-type traits introduces an identifiability issue for the diffusion precision — when using conjugate priors directly on $\mathbf{K}$. We discuss this limitation, that can be overcome with different modeling strategies, in Section 7.

Because we only need the latent parameters $\mathbf{X}$ at the tips of $\mathscr{F}$, i.e. $(\mathbf{X}_1, \ldots, \mathbf{X}_N)^t$, to map $\mathbf{Y}$, we can compute the likelihood of $\mathbf{X}$ by integrating out $\mathbf{X}_{N+1}, \ldots, \mathbf{X}_{2N-1}$. In order to do so we adopt a conjugate prior on the root of the tree, $\mathbf{X}_{2N-1} \sim \mathcal{N}_p(\boldsymbol{\mu}_0, \tau_0^{-1}\mathbf{K}^{-1})$ with prior mean $\boldsymbol{\mu}_0$ and prior sample size τ_0 (Pybus et al., 2012). Hence, $\mathbf{X}$ follows a matrix-normal distribution,

$$\mathbf{X} \sim \mathcal{MN}_{N \times p}(\mathbf{M}, \boldsymbol{\Upsilon}, \mathbf{K}^{-1}), \tag{2}$$

where $\mathbf{M} = \mathbf{1}_N \boldsymbol{\mu}_0^t$ is an $N \times p$ mean matrix, $\mathbf{1}_N$ is a vector of length N populated by ones, $\mathbf{K}^{-1}$ is the $p \times p$ across-trait covariance matrix, and $\boldsymbol{\Upsilon} = \mathbf{V}(\mathscr{F}) + \tau_0^{-1}\mathbf{J}_N$ is an $N \times N$ across-taxa tree covariance matrix. The tree diffusion matrix $\mathbf{V}(\mathscr{F})$ is a deterministic function of $\mathscr{F}$, and $\mathbf{J}_N = \mathbf{1}_N \mathbf{1}_N^t$ is an $N \times N$ matrix of all ones and the term $\tau_0^{-1}\mathbf{J}_N$ comes from the integrated-out tree root prior (for further details on $\mathbf{V}(\mathscr{F})$, see Zhang et al., 2021, Figure 1). Combining the stochastic link $p(\mathbf{Y}|\mathbf{X})$ and Equation (2) we can consider an augmented likelihood of $\mathbf{Y}$ and $\mathbf{X}$ through the factorization

$$p(\mathbf{Y}, \mathbf{X}|\mathbf{K}, \mathscr{F}, \boldsymbol{\mu}_0, \tau_0) = p(\mathbf{Y}|\mathbf{X})p(\mathbf{X}|\mathbf{K}, \mathscr{F}, \boldsymbol{\mu}_0, \tau_0). \tag{3}$$

Building upon the tree and matrix-normal structures, several algorithms were developed to evaluate the trait likelihood keeping the linear scale with the number of taxa in complete data set scenarios (Tung Ho and Ané, 2014; Tolkoff et al., 2018; Bastide et al., 2018). When dealing with missing data, however, scalability becomes a bottleneck as the model requires data imputation or integration. To bypass this limitation, Hassler et al. (2020) propose an inference technique that integrates out missing values analytically that scales linearly with the number of taxa by using a post-order traversal algorithm under a MBD model to characterize trait evolution.

In the following section we describe how to incorporate sparsity on the diffusion precision to perform covariance selection using Gaussian graphical models for this trait evolution model. We further close this section declaring the prior specifications for the parameters in the model.

2.2 ***Model extension: Covariance Selection with Gaussian Graphical Model.*** Gaussian graphical models provide a simple and convenient framework for imposing a conditional independence structure of association between variables (Mohammadi and Wit, 2015; Williams, 2021). Here we introduce some basic notation and the structure of undirected Gaussian graphical models. We refer the interested reader to Lauritzen (1996) for detailed information.

Let $\mathbf{G} = (V, E)$ be an undirected graph defined by a finite set of vertices (or nodes) $V = \{1, 2, \ldots, p\}$ that represent the Gaussian variables, and a set of existing edges $E \subset \{(i, j)|1 \leq i < j \leq p\}$ that represent links among the nodes $i, j \in V$. We define a $\boldsymbol{\mu}_0$ mean Gaussian graphical model with respect to the graph $\mathbf{G}$ as the set of all Gaussian models such that

$$\mathcal{M}_G = \{\mathcal{N}_p(\boldsymbol{\mu}_0, \mathbf{K}^{-1}), \mathbf{K} \in \mathbb{P}_G\}, \tag{4}$$

where $\mathbb{P}_G$ is the space of $p \times p$ positive definite matrices with zero entries (i, j) consistent with $\mathbf{G}$, i.e. $k_{ij} = 0$ whenever $(i, j) \notin E$. Hence, in GGM, we assume that the precision matrix $\mathbf{K}$ depends on the graph $\mathbf{G}$.

2.3 ***Sparse Phylogenetic Trait Evolution Model (SPTE).*** Importing the GGM approach to the phylogenetic context we then restrict the precision matrix of the Brownian diffusion process to $\mathbb{P}_G$. This gives rise to the diffusion graph $\mathbf{G}$, a relevant parameter that represents the association structure of the partial correlations between the p components of the latent parameter $\mathbf{X}$.

We complete our model specification by choosing the prior distributions for the graph $\mathbf{G}$ and for the structured precision matrix $\mathbf{K}|\mathbf{G}$. For simplicity, we place a discrete uniform prior distribution over $\mathcal{G}$, the space of all graphs with fixed p edges (Mohammadi and Wit, 2015; Mohammadi et al., 2021), such that

$$p(\mathbf{G}) = \frac{1}{|\mathcal{G}|}, \tag{5}$$

for each $\mathbf{G} \in \mathcal{G}$, where $|\mathcal{G}| = 2^{p(p-1)/2}$ is the cardinality of the graph space. One can rather choose a different prior favoring denser or sparser graphs in the light of better knowledge about the association structure (Mohammadi and Wit, 2015).

For the prior distribution of the precision matrix $p(\mathbf{K}|\mathbf{G})$, we use the G-Wishart distribution (Roverato, 2002; Atay-Kayis and Massam, 2005), which is the conjugate prior, under Gaussian models, for structured precision matrices. The G-Wishart distribution $\mathcal{W}_G(\delta, \mathbf{D})$, with support on the space of $p \times p$ positive definite matrices, has density

$$f_G(\mathbf{K}; \delta, \mathbf{D}) = \frac{1}{I_G(\delta, \mathbf{D})} |\mathbf{K}|^{(\delta-2)/2} \exp\left\{-\frac{1}{2}\mathrm{tr}(\mathbf{DK})\right\} \mathbf{1}_{\mathbf{K} \in \mathbb{P}_G}, \tag{6}$$

with parameters δ and $\mathbf{D}$, where $\delta > 0$ (Mitsakakis, 2010, Lemma 3.2.1: $I_G(\delta, \mathbf{D}) < \infty$ for $\delta > 0$) represents the degrees of freedom (or shape parameter), $\mathbf{D}$ is a symmetric positive-definite rate matrix, and $I_G(\delta, \mathbf{D})$ is the normalizing constant,

$$I_G(\delta, \mathbf{D}) = \int_{\mathbf{K} \in \mathbb{P}_G} |\mathbf{K}|^{(\delta-2)/2} \exp\left\{-\frac{1}{2}\mathrm{tr}(\mathbf{DK})\right\} d\mathbf{K}. \tag{7}$$

We use the Monte Carlo method of Atay-Kayis and Massam (2005) to numerically approximate the prior and posterior normalizing constants required for graph updates during MCMC.

3 Inference

We single out the diffusion correlation $\mathbf{R}$ and the diffusion graph $\mathbf{G}$ as the primary parameters of scientific interest. The model is parametrized, however through the diffusion precision $\mathbf{K}$ which indirectly shapes the correlations according to the conditional (in)dependencies in the graph. We drop the posterior's dependence on the hyperparameters $(\boldsymbol{\Upsilon}, \boldsymbol{\mu}_0, \tau_0, \delta, \mathbf{D})$ to ease notation. Connecting the likelihood (3) to the priors (5) and

(6), we finally arrive at the posterior factorization

$$\begin{aligned}p(\mathbf{K},\mathbf{G},\mathscr{F}|\mathbf{Y},\mathbf{S}) &\propto p(\mathbf{Y}|\mathbf{K},\mathscr{F})p(\mathbf{K}|\mathbf{G})p(\mathbf{G})p(\mathbf{S},\mathscr{F})\\ &= \left(\int p(\mathbf{Y}|\mathbf{X})p(\mathbf{X}|\mathbf{K},\mathscr{F})\,\mathrm{d}\mathbf{X}\right)p(\mathbf{K}|\mathbf{G})p(\mathbf{G})p(\mathbf{S},\mathscr{F}). \qquad (8)\end{aligned}$$

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 all other parameters given $\mathscr{F}$ (Zhang et al., 2021). To obtain the posterior for this model we must integrate over the possible values for the unobserved latent variables at the tips of the tree. For the simple MBD with continuous traits and no missing data, nevertheless, there is no need to perform integration over $\mathbf{X}$ because of the nature of the link function. To approximate the posterior distributions via MCMC simulation, we apply a random scan Metropolis-within-Gibbs (Liu et al., 1995) approach by which we sample parameter blocks one at a time at random, using Gibbs steps whenever a known full conditional distribution is available.

3.1 *Sampling scheme.* We employ standard Bayesian phylogenetic algorithms to obtain $p(\mathscr{F},\mathbf{S})$ when the tree is unknown (Suchard et al., 2018). Alternatively, $\mathscr{F}$ can be fixed, in which case there is no need for sequence data $\mathbf{S}$. If trait data $\mathbf{Y}$ is incomplete, we draw from the full conditional distribution of $\mathbf{X}$ using the pre-order missing data augmentation algorithm developed by Hassler et al. (2020) with overall computational complexity $\mathcal{O}(Np^3)$.

A joint updating scheme, inspired by the factorization $p(\mathbf{G},\mathbf{K}) = p(\mathbf{G})p(\mathbf{K}|\mathbf{G})$, is considered for $\mathbf{K}$ and $\mathbf{G}$. First, the graph $\mathbf{G}$ is updated through a Metropolis-Hastings step whose target distribution is the marginal distribution of the graph. Then, $\mathbf{K}$ is updated conditional on the new $\mathbf{G}$ through a Gibbs step.

To update the graph $\mathbf{G}$ we need to compute the marginal distribution of $\mathbf{G}$, given all other parameters except $\mathbf{K}$ (see Supplementary Information (SI 2)). This marginal distribution is given by Equation (SI.8) as

$$p(\mathbf{G}|\mathbf{X},\delta,\mathbf{D},\mathscr{F}) \propto p(\mathbf{G}|\delta,\mathbf{D})p(\mathbf{X}|\mathbf{G},\delta,\mathbf{D},\mathscr{F}) = \frac{p(\mathbf{G}|\delta,\mathbf{D})}{(2\pi)^{Np/2}}\frac{I_G(\delta+N,\mathbf{D}+\boldsymbol{\Delta})}{I_G(\delta,\mathbf{D})}, \qquad (9)$$

where $\boldsymbol{\Delta} = (\mathbf{X}-\mathbf{1}_N\boldsymbol{\mu}_0^t)^t\left(\boldsymbol{\Upsilon}+\tau_0^{-1}\mathbf{J}_N\right)^{-1}(\mathbf{X}-\mathbf{1}_N\boldsymbol{\mu}_0^t)$, comes from integrating $\mathscr{F}$ with GGM. As in Hassler et al. (2020), we follow the methods in Tung Ho and Ané (2014) to compute $\boldsymbol{\Delta}$ via post-order traversal of the tree, which has computational complexity $\mathcal{O}(Np^2)$.

Based on this target distribution, the current graph $\mathbf{G}_c$ is updated using a Metropolis-Hastings step with new graph proposals $\mathbf{G}_p$ generated by switching the value of a randomly selected edge in $\mathbf{G}_c$. We set $\mathbf{G}_c = \mathbf{G}_p$ by accepting the proposed graph with probability

$$\alpha = \min\left\{1, \frac{I_{G_p}(\delta + N, \mathbf{D} + \mathbf{\Delta})}{I_{G_c}(\delta + N, \mathbf{D} + \mathbf{\Delta})} \frac{I_{G_c}(\delta, \mathbf{D})}{I_{G_p}(\delta, \mathbf{D})}\right\}. \tag{10}$$

After the update, $\mathbf{G}_c$ is then used to sample from the posterior distribution of $(\mathbf{K}|\mathbf{X}, \mathbf{G}_c, \delta, \mathbf{D}, \mathscr{F})$ in Equation (11).

Since we place a G-Wishart conjugate prior $\mathcal{W}_G(\delta, \mathbf{D})$ on the structured diffusion precision matrix $\mathbf{K}|\mathbf{X}, \mathbf{G}_c, \delta, \mathbf{D}, \mathscr{F}$, the full conditional distribution of diffusion precision is also G-Wishart

$$\mathbf{K}|\mathbf{X}, \mathbf{G}_c, \delta, \mathbf{D}, \mathscr{F} \sim \mathcal{W}_G\left(\delta + N, \mathbf{D} + \mathbf{\Delta}\right). \tag{11}$$

3.2 ***Parameter Estimates.*** After running the MCMC we process the posterior distributions of the parameters of scientific interest, namely $\mathbf{G}$, $\mathbf{K}$ and $\mathbf{R}$. We estimate the entries of the posterior graph $\hat{\mathbf{G}} = \{\hat{g}_{ij}\}$ from its MCMC iterations, after warm-up, through

$$\hat{g}_{ij} = \begin{cases} 1, & \text{if } BF^{(ij)}_{1:0} \geq 10^{1/2} \\ 0, & \text{otherwise} \end{cases}, \tag{12}$$

where $BF^{(ij)}_{1:0}$ is the Bayes factor calculated to evaluate the evidence in favor of including edge (i, j) in the posterior graph $\hat{\mathbf{G}}$, i.e. evidence for conditional dependence between i-th and j-th traits. A Bayes factor above the threshold $10^{1/2}$ indicates substantial evidence for the hypothesis according to the criteria in Jeffreys (1998). We compute the Bayes factor as

$$BF^{(ij)}_{1:0} = \frac{p(\hat{g}_{ij} = 1|\mathbf{Y}, \mathbf{X})}{p(\hat{g}_{ij} = 0|\mathbf{Y}, \mathbf{X})} \frac{p(\hat{g}_{ij} = 0)}{p(\hat{g}_{ij} = 1)} = \frac{\hat{pe}_{ij}}{(1 - \hat{pe}_{ij})}, \tag{13}$$

where the estimated posterior edge inclusion probability $\hat{pe}_{ij}$ is the proportion of posterior samples with graph entry $g_{ij} = 1$, for $1 \leq i < j \leq p$. Equivalently, the threshold $10^{1/2}$ corresponds to $\hat{pe}_{ij} \approx 0.76$. We assume equally likely models $p(\hat{g}_{ij} = 1) = p(\hat{g}_{ij} = 0) = 1/2$. Alternatively, one may want to favor different criteria for the threshold in the Bayes factor.

Furthermore, the entries of the posterior precision estimate $\hat{\mathbf{K}} = \{\hat{k}_{ij}\}$ are computed as the mean of the precision samples k_{ij} for all MCMC iterations whose corresponding graph edge g_{ij} is consistent with the marginal posterior graph $\hat{g}_{ij}$. We keep all MCMC samples, after warm-up, when computing the posterior mean correlation estimate $\hat{r}_{ij}$.

3.3 ***Implementation.*** We have implemented the proposed model in the development version (v1.10.5) of BEAST (Suchard et al., 2018).

4 Simulation Study

In order to understand the behavior of the proposed model (SPTE), alternatively referred to as sparse model, and its ability to recover the true graph structure, trait precision and correlation underlying the observed data, we conduct a simulation study consisting of two scenarios (*Sim* 1 and *Sim* 2) that combine different graph structures $\mathbf{G}_0$ and precision matrices $\mathbf{K}_0$. We also compare these results to the ones obtained from the full model, in which the association structure is not explicitly modeled through a graph. We use $N = 50$ and $p = 5$ for *Sim* 1, and $N = 100$ and $p = 10$ for *Sim* 2. The diffusion precision $\mathbf{K}_0$ used to generate the continuous trait observations in each simulation scenario and the corresponding diffusion correlation $\mathbf{R}_0$ are available in the Supplementary Information (SI 1). The true graph structure $\mathbf{G}_0$ can be directly recovered from the respective $\mathbf{K}_0$.

For each scenario, we simulate $RE = 1000$ Monte Carlo data sets and run MCMC chains to fit both sparse and full models. As non-informative priors, we take a uniform distribution over the graph space on $\mathbf{G}$, and a G-Wishart $\mathcal{W}_G(3, \mathbf{I_p})$ on $\mathbf{K}|\mathbf{G}$ for the sparse model, whereas we assume a Wishart $\mathcal{W}_p(2+p, \mathbf{I_p})$, with rate parametrization, on $\mathbf{K}$ for the full model, where $\mathbf{I_p}$ denotes the identity matrix of dimension p. The Wishart and G-Wishart hyperparameters are equivalent since $\nu = \delta + p - 1$, where ν and δ are the degrees of freendom of Wishart and G-Wishart distributions, respectively. For each Monte Carlo replication, we generate a random tree, $\mathscr{F}$, of size N and simulate the latent variables $\mathbf{X}_h$, $h = 1, \ldots, 2N-1$, traversing the tree from root $\mathbf{X}_{2N-1}$ to tips, using Equation (1) with diffusion precision $\mathbf{K}_0$. We set $\mathbf{Y}_k = \mathbf{X}_k$ for the tips $k = 1, \ldots, N$.

We approximate the posterior distribution of the graph structure $\mathbf{G}$ for the sparse model and the posterior distribution of the diffusion precision $\mathbf{K}$ and diffusion correlation $\mathbf{R}$ for both models. Simulations are tailored to reach an effective sample size ESS $\approx$ 500 after warm-up. For each Monte Carlo replication $re = 1, \ldots, RE$, we estimate the posterior graph $\hat{\mathbf{G}}^{(re)}$ using the Bayes factor criteria and threshold in Eq. (12).

Figure 1 presents the Monte Carlo posterior graph $\hat{\mathbf{G}}^{MC} = \{\hat{g}_{ij}^{MC}\}$, calculated as $\hat{\mathbf{G}}^{MC} = (RE)^{-1} \sum_{re=1}^{RE} \hat{\mathbf{G}}^{(re)}$ and the respective true graph structure for both simulation scenarios. Note that the Monte Carlo graph is a summary measure for the graph estimates in each Monte Carlo replication, not a direct estimate itself. Therefore, in Figure 1a, edge thickness and transparency represent $\hat{g}_{ij}^{MC}$, i.e. the proportion of Monte Carlo replicates that include edge (i, j) in the posterior graph.

In *Sim* 1, $\hat{\mathbf{G}}^{MC}$ perfectly matches the true graph structure, which means that, on average, the Monte Carlo replicates correctly estimate the association structure. In *Sim* 2, however, edges $(8, 9)$ and $(8, 10)$ are not always included in the posterior graphs replicates (Figure 1a).

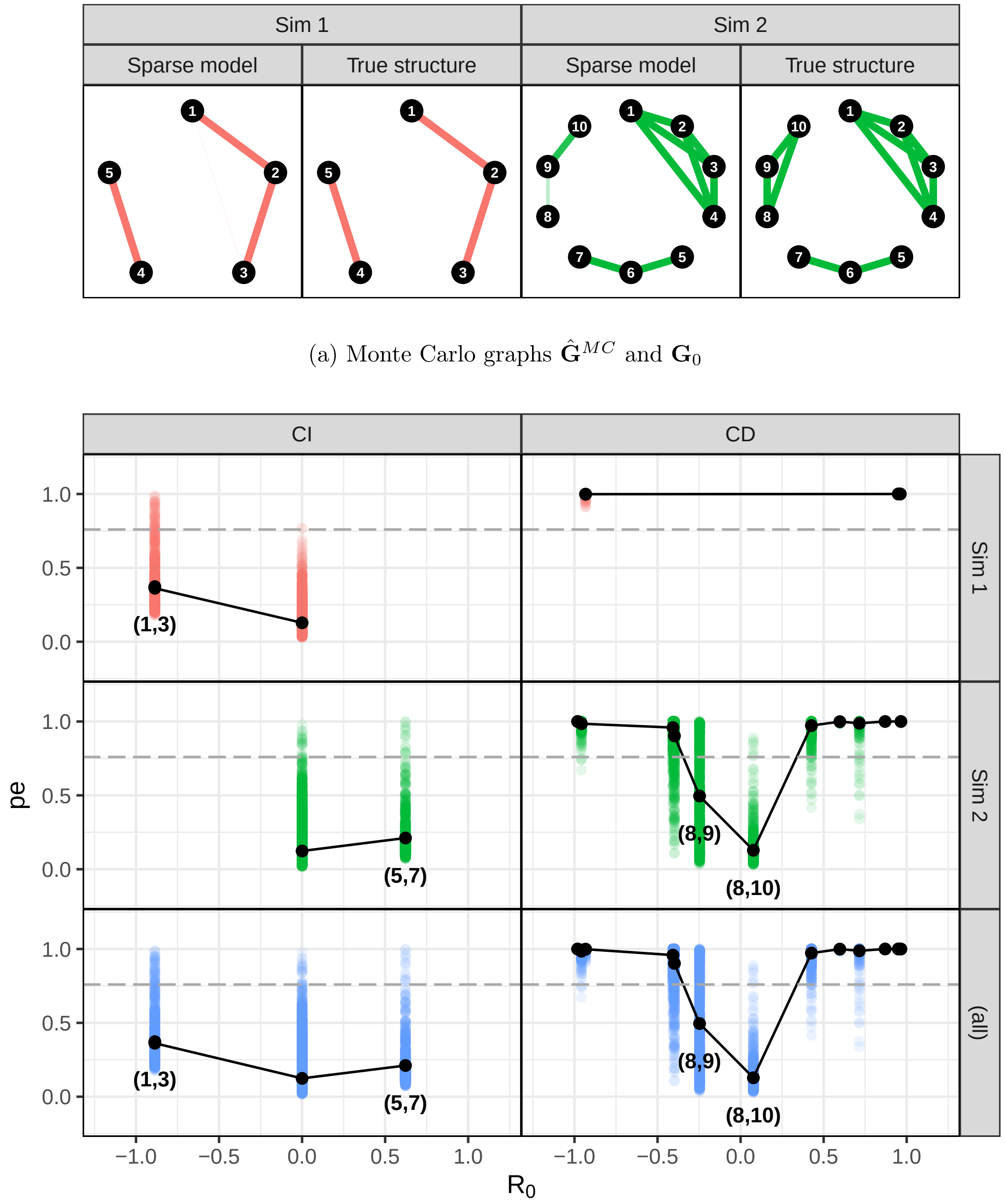
To help elucidate the inconsistencies in graph estimation, we also show, in Figure 1b, the posterior edge inclusion probabilities for each Monte Carlo replication $\hat{pe}_{ij}^{(re)}$ (colored dots), as well as their means $\hat{pe}_{ij}^{MC}$ (black dots). We stratify the edges by conditional (in)dependence type — as conditionally independent (CI), when $g_{0(ij)} = 0$, or conditionally dependent (CD), when $g_{0(ij)} = 1$ —, and simulation (*Sim* 1 or *Sim* 2), and plot them against the true correlation strength $\mathbf{R}_0$. The last grid in Figure 1b combines $\hat{\boldsymbol{pe}}$ from both simulations to facilitate the visualization, but should be interpreted with care, since sample sizes and trait dimensions are different between simulations.

We single out a few edges in Figure 1b in order to highlight a couple of model features. We point out that edge $(1, 3)$ in *Sim* 1 has mean posterior edge inclusion probability $\hat{pe}_{13}^{MC} = 0.372$ (black dot for $(1, 3)$ in Figure 1b), and mean posterior graph entry $\hat{g}_{13}^{MC} = 0.051$, which indicates that only 5.1% of the Monte Carlo replications display posterior edge inclusion probability greater than or equal to 0.76 (see the proportion of colored dots above the dashed lines in Figure 1b). This is relevant because it suggests that the absence of edges for CI variables can be correctly estimated even when the respective true correlation is very strong ($r_{0(13)} = -0.89$).

For CD variables, one can see that posterior edge inclusion probabilities are high for strong true correlations, but decrease as the true correlations approach zero (Figure 1b). This result suggests that even CD variables tend to have their posterior graph edge inclusion probabilities shrunk towards zero in the sparse model if the respective true correlations are weak. This is the case of the two edges in *Sim* 2 that display weak correlations, namely, -0.25 and 0.08 for edges $(8, 9)$ and $(8, 10)$, respectively.

Figure 2 presents the log mean squared-error (*log*MSE) for $\hat{\mathbf{K}}^{(re)}$ and $\hat{\mathbf{R}}^{(re)}$ estimates in both models. For this analysis we classify each parameter entry accordingly to its conditional (in)dependence type (CI or CD) and, additionally, by its dependence type (as independent (I), if $r_{0(ij)} = 0$ or dependent (D), if $r_{0(ij)} \neq 0$). Then, the categories are conditionally independent and independent (CI-I), conditionally independent and dependent (CI-D), and conditionally dependent and dependent (CD-D).

As expected, in both simulation scenarios, the mean precision *log*MSE for CI variables (CI-D and CI-I) is significantly lower in the sparse model, specially for the independent ones (CI-I), when compared to the full model estimates. This is an important result since CI variables are the main target for sparsity in GGM. Note that, for CD variables,



(a) Monte Carlo graphs $\hat{\mathbf{G}}^{MC}$ and $\mathbf{G}_0$

(b) Monte Carlo posterior edge inclusion probabilities $\boldsymbol{pe}$

Figure 1: Monte Carlo graphs and posterior edge inclusion probabilities for *Sim* 1 and *Sim* 2. (1a) Monte Carlo posterior graph $\hat{\mathbf{G}}^{MC}$ and true graph $\mathbf{G}_0$ in both simulations scenarios. Edge thickness and transparency represent the proportion of Monte Carlo replicates that include edge (i, j) in the posterior graph. (1b) Monte Carlo posterior edge inclusion probabilities $\hat{\boldsymbol{pe}}$ across $\mathbf{R}_0$. Colored points depict the proportion of edges $\hat{g}_{ij} = 1$ in the 1000 Monte Carlo replicates. The black points represent the $\hat{pe}_{ij}^{MC}$ of each pairwise correlation. Edges $(1, 3)$ from *Sim* 1 and $(5, 7)$, $(8, 9)$ and $(8, 10)$ from *Sim* 2 are indicated to illustrate important features of graph estimation.

the *log*MSE are equivalent between the models, which is consistent with the fact that Wishart and G-Wishart sampling processes are the same for CD variables (for details, see Atay-Kayis and Massam, 2005). The mean correlation *log*MSE for CI-I variables are also lower in sparse models. This result suggests that the better the estimates for precision matrices in the sparse model, the lower the correlation's *log*MSEs for this class of variables. Additionally, note that correlations between independent variables display similar log mean squared-errors regardless their conditional (in)dependence status, in both simulations.

Additionally, since full models rely on a post-process procedure over the posterior distribution of $\mathbf{R}$ to perform correlation selection, and sparse models directly estimate $\hat{\mathbf{G}}$, we calculate the accuracy, sensitivity, specificity, precision and F1-Score (see the confusion matrix and Table 1 in SI 1 for detailed definitions) for the respective model target parameter in order to establish a fair comparison between models. These metrics depict the overall performance of joint estimation of entries for each parameter. However, since they derived from confusion matrices, which are proper to evaluate binary classification systems, for full models, the target parameter in each Monte Carlo iteration is actually $\hat{\mathbf{R}}^H$, that is obtained by the classification of each marginal correlation into zero or non-zero using a specific HPD criteria. We define

$$\hat{r}_{ij}^{H(re)} = \begin{cases} 1, & \text{if } 0 \notin \text{HPD}_{\gamma\%}(r_{ij}^{(re)}) \\ 0, & \text{otherwise,} \end{cases}$$

where γ indicates the chosen percentage for HPD criteria.

Table 1 reports the comparisons of our sparse model with the full model applying $\text{HPD}_{90\%}$ and $\text{HPD}_{95\%}$ criteria. Our method performs well overall as its specificity, F1-score, precision and accuracy are higher than the full model in both simulations. Sensitivity is one for all models and criteria in *Sim* 1, which means that all the true edges are correctly identified in the sparse model as well as all the true non-zero correlations in the full model for both HPD criteria. In *Sim* 2, sensitivity is higher for the full model with $\text{HPD}_{90\%}$ (0.91 ± 0.054) mainly because of the lower percentage of posterior density required to define correlations as non-zero. Since $\mathbf{R}_0$ in Sim 2 (see SI.4) displays some weak correlations that tend to be shrunk by HPD correlation selection in the full model — or equivalently, whose corresponding graph entries is pushed towards zero in the sparse model —, the apparent best performance in sensitivity of $\text{HPD}_{90\%}$ criteria is actually a consequence of its less strict interval that favors non-zero correlation classifications, even for weak true correlations, at the cost of less specificity (0.85 ± 0.093).

Overall, the sparse model is better at identifying CI and CD variables than the HPD

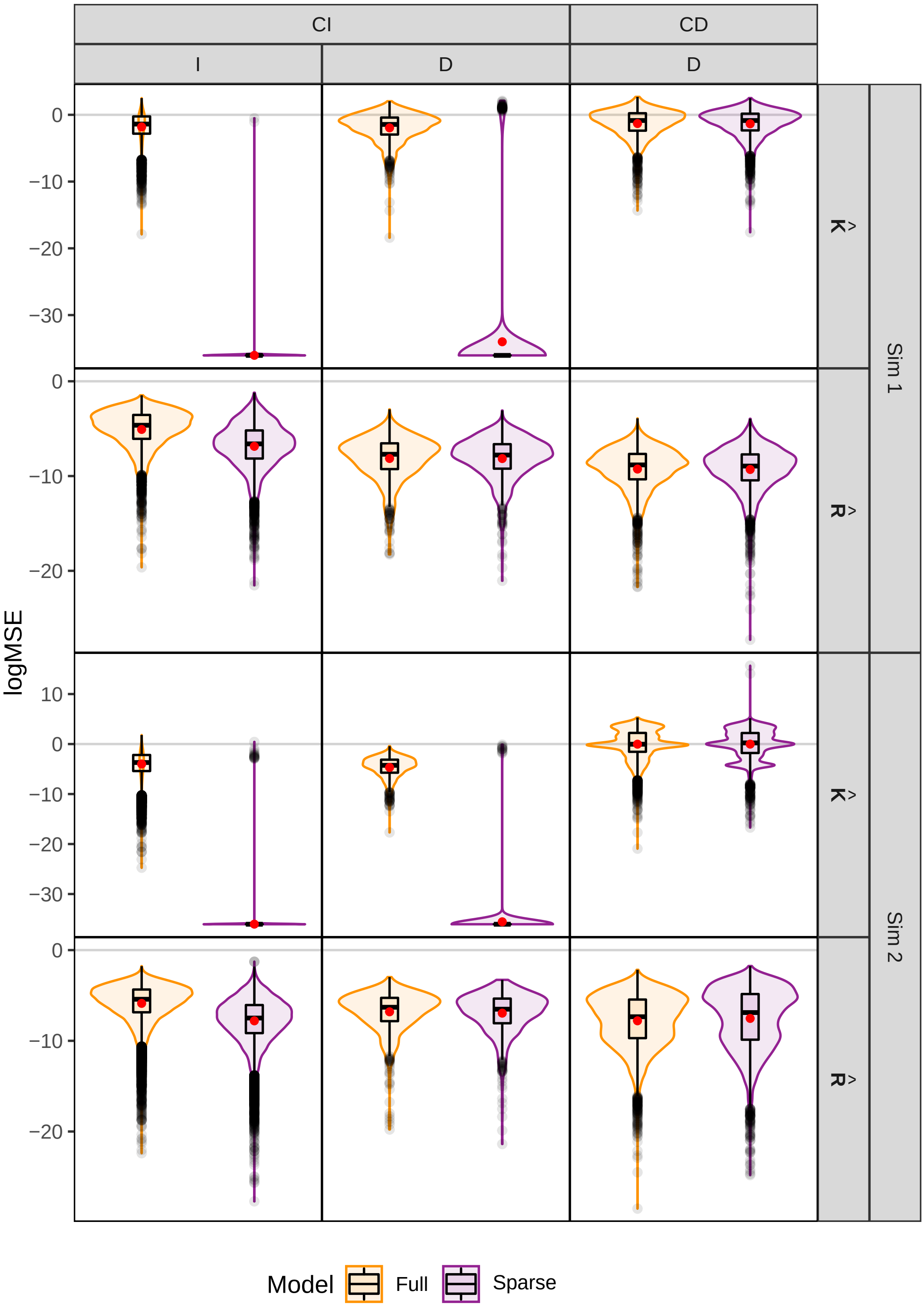


Figure 2: Log mean squared-error of posterior diffusion precision estimates $\hat{\mathbf{K}}^{(re)}$, and posterior diffusion correlation estimates $\hat{\mathbf{R}}^{(re)}$, over 1000 simulated replicates based on *Sim* 1 and *Sim* 2. Red dots represent the Monte Carlo mean values for *log*MSE. CI = conditionally independent; CD = conditionally dependent; I = independent; D = dependent.

criteria are at discriminating between correlated and uncorrelated variables in the full model. This is particularly the case for the precision metric, that depicts the proportion of true positives in the confusion matrix among all predicted as true (including the false positives), and F1-Score, which summarizes the performances by balancing sensitivity and precision. One can see that, across simulations, the precision metric for $\hat{\mathbf{G}}^{(re)}$ is at least 0.99 in the sparse model, but lies under 0.78 for $\hat{\mathbf{R}}^{H(re)}$ in the full model, indicating that the number of false positives is relatively high for post-process HPD procedures.

Table 1: Summary of performance measures for decision criteria in sparse and full models. In the full model, the decision is to classify the correlations as zero or non-zero according to a chosen HPD; in the sparse model, it is whether the edges should be included in the posterior graph or not. The table presents the mean (and standard deviation) for sensitivity, specificity, precision, F1-score, and accuracy of posterior graph estimates $\hat{\mathbf{G}}^{(re)}$ (sparse model) and post-processed diffusion correlation estimates $\hat{\mathbf{R}}^{H(re)}$ (full model). The F1-score reaches its best score at 1 and its worst at 0. The best model for each statistic is boldfaced.

Sim	Parameter	Criteria	Sensitivity		Specificity		Precision		F1-Score		Accuracy	
Sim 1	$\hat{\mathbf{G}}$	GGM	**1.00**	(0.000)	**0.99**	(0.032)	**0.99**	(0.056)	**0.97**	(0.059)	**0.98**	(0.042)
	$\hat{\mathbf{R}}^H$	HPD$_{90\%}$	**1.00**	(0.000)	0.72	(0.266)	0.66	(0.155)	0.79	(0.132)	0.81	(0.186)
	$\hat{\mathbf{R}}^H$	HPD$_{95\%}$	**1.00**	(0.000)	0.78	(0.210)	0.70	(0.127)	0.81	(0.107)	0.84	(0.147)
Sim 2	$\hat{\mathbf{G}}$	GGM	0.82	(0.055)	**1.00**	(0.005)	**1.00**	(0.018)	**0.90**	(0.034)	**0.96**	(0.014)
	$\hat{\mathbf{R}}^H$	HPD$_{90\%}$	**0.91**	(0.054)	0.85	(0.093)	0.70	(0.132)	0.78	(0.092)	0.87	(0.072)
	$\hat{\mathbf{R}}^H$	HPD$_{95\%}$	0.90	(0.055)	0.91	(0.072)	0.78	(0.120)	0.83	(0.079)	0.91	(0.056)

We also compute the accuracy of $\hat{\mathbf{G}}$ and $\hat{\mathbf{R}}^H$ directly on each parameter entry, stratifying the results by the combination of conditional (in)dependence type (CI or CD) and dependence type (I or D). This statistic provides better insight about how the conditional (in)dependence structure and correlation type can affect the model ability to estimate the association structure (sparse model) or perform correlation selection (full model with post-process HPD procedure).

In both simulation scenarios, the overall accuracy for each entry is higher for $\hat{\mathbf{G}}$ in the sparse model than it is for the analogous $\hat{\mathbf{R}}^H$ in the full model (Table 2). The sparse model also displays higher average accuracy for CI-I variables, which is one of the substantial gains of including the graph estimation in the MBD model. Note that the difference in accuracy between the models is highlighted by the fact that standard deviations in CI-I variables are relatively small in both simulation scenarios. On the other hand, for CI-D variables, the average accuracy is larger for the full model. The reason for that is the underlying strong correlation of CI-D variables in our simulation settings (see edges $(1, 3)$ in *Sim* 1, and $(5, 7)$ in *Sim* 2, Figure 1b), and how they negatively affect only sparse model performance. Although the final Monte Carlo graph matches the true graph structure, in *Sim* 1 (Figure 1b), the bias in edge $(1, 3)$ affects the average accuracy statistic for CI-D, mainly because $(1,3)$ is the only edge in this classification. The same conclusion holds for edge $(5,7)$, in *Sim* 2, which is also the only edge in this category in its simulation. This is the reason we are not able to display standard deviations for this category in Table 2.

In addition, both models and criteria performed well in terms of pairwise accuracy for CD-D variables. Despite the better accuracy for $\hat{\mathbf{R}}^H$ with the HPD$_{90\%}$ criteria in the full model, all three statistics are similar, specially when considering their high standard deviations. Finally, in a general perspective, these accuracy results corroborate the ones obtained for the *log*MSE of $\hat{\mathbf{K}}$ where better estimates for the sparse model are associated with CI-I and similar results between models are obtained for dependent variables (CI-D and CD-D).

Table 2: Summary of Monte Carlo accuracy measures on each parameter entries in two simulations scenarios for sparse and full models. The table presents the average (and standard deviation) accuracy for posterior graph $\hat{\mathbf{G}}^{(re)}$ edges in the sparse model, and for post-inference posterior correlation $\hat{\mathbf{R}}^{H(re)}$ entries in the full model. The number of pairwise entries classified in each category according to the conditional (in)dependency (CD or CI) and dependency types (D or I) is indicated by n. The best models are boldfaced.

Criteria	$\mathbf{G}_0$	$\mathbf{R}_0$	*Sim* 1			*Sim* 2		
			n	**Accuracy**	**(sd)**	n	**Accuracy**	**(sd)**
GGM	CD	D	3	**1.00**	(0.000)	11	0.82	(0.348)
HPD$_{90\%}$	CD	D	3	**1.00**	(0.000)	11	**0.91**	(0.236)
HPD$_{95\%}$	CD	D	3	**1.00**	(0.000)	11	0.90	(0.264)
GGM	CI	D	1	0.95	-	1	0.99	-
HPD$_{90\%}$	CI	D	1	**1.00**	-	1	**1.00**	-
HPD$_{95\%}$	CI	D	1	**1.00**	-	1	**1.00**	-
GGM	CI	I	6	**1.00**	(0.001)	33	**1.00**	(0.001)
HPD$_{90\%}$	CI	I	6	0.84	(0.007)	33	0.88	(0.010)
HPD$_{95\%}$	CI	I	6	0.90	(0.007)	33	0.94	(0.007)
GGM	Overall	Overall	10	**1.00**	(0.016)	45	**0.96**	(0.183)
HPD$_{90\%}$	Overall	Overall	10	0.91	(0.082)	45	0.89	(0.115)
HPD$_{95\%}$	Overall	Overall	10	0.94	(0.050)	45	0.93	(0.128)

5 Applications

We apply our method to two data sets to showcase the benefits of including graph estimation in the MBD model and to demonstrate how the association structure represented by the graph can lead to richer discussions when compared to traditional trait evolution models. For each application, we run both sparse and full models in equivalent setups. We choose $\text{HPD}_{95\%}$ as the criteria for post-process correlation selection.

5.1 ***Darwin's Finches.*** Evolution of Darwin's finches (Fringillidae, Passeriformes) is a classical example of adaptive radiation under natural selection. There are thirteen species in the Galápagos archipelago and another one on Cocos island. The wide variation in beak morphology is associated with the exploitation of a variety of ecological niches, because it allows finches to access particular types of food — including seeds, insects, and cactus flowers (Abzhanov et al., 2004, 2006) —, and likely played a role in the diversification of avian species (Mallarino et al., 2011). To assess the phenotypic correlations and the association structure between morphometric variables, we use the data set of 13 species of Darwin's finches in the study of Drummond et al. (2012). The data consist of a 2,065-bp partial nucleotide alignment of the mitochondrial control region and cytochrome b genes and five continuously measured phenotypic traits: culmen length (CulmenL), beak depth (BeakD), gonys width (GonysW), wing length (WingL), and tarsus length (TarsusL). We estimate the posterior graph $\hat{\mathbf{G}}$ for the sparse model and the diffusion correlation $\hat{\mathbf{R}}$ in both sparse and full models in order to illustrate the gain of explicitly accounting for the association structure in our inference framework.

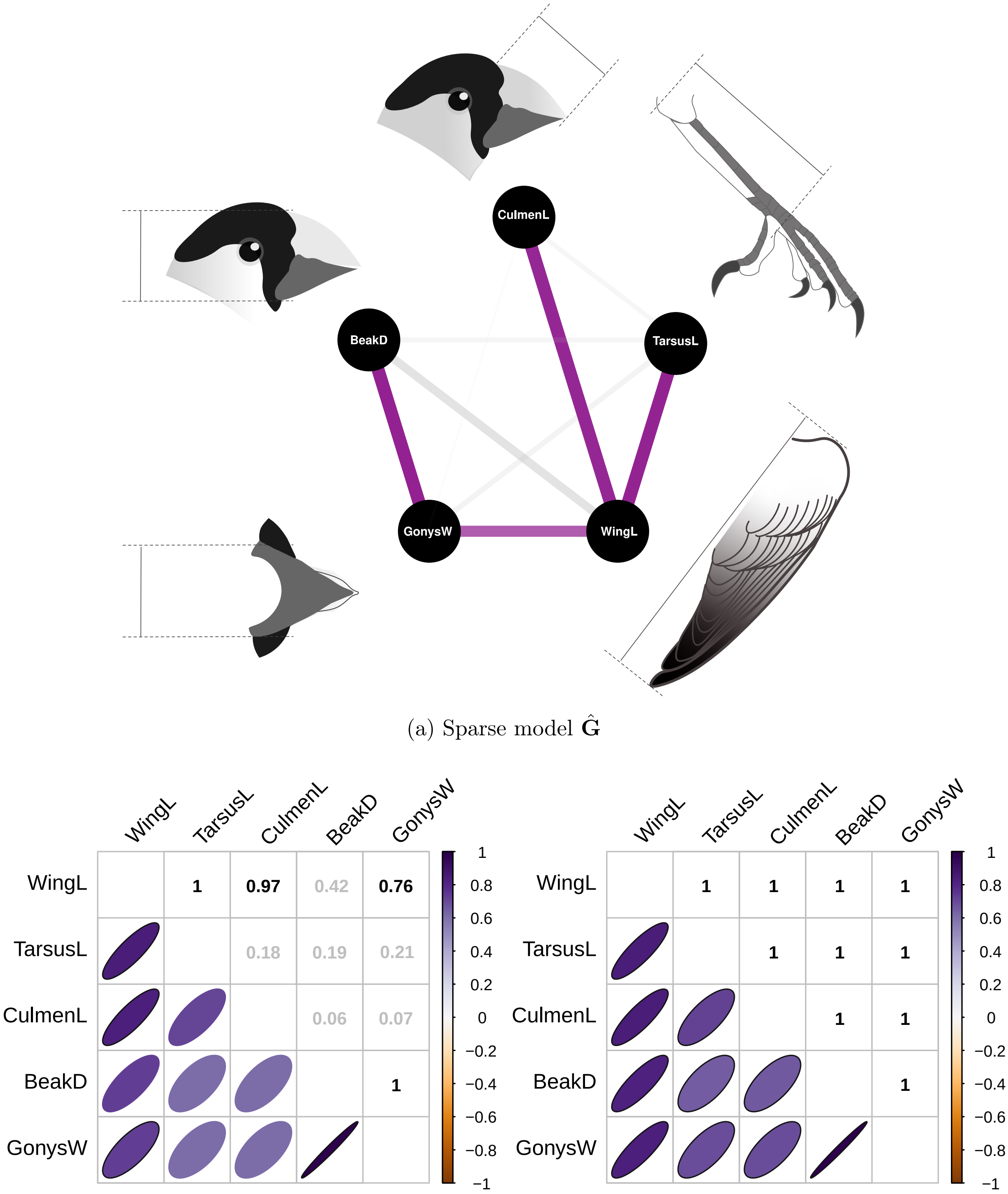
Figure 3a displays the estimated posterior graph $\hat{\mathbf{G}}$ for the association structure between Darwin's finches trait measurements, whereas Figures 3b, and 3c present the evolutionary correlation between those traits in the sparse and full models, respectively. The numbers above diagonal in each correlogram represent the posterior edge inclusion probabilities $\boldsymbol{pe}$ in the sparse model, and the posterior probability that correlations are of the same sign of its mean $\boldsymbol{ps}$, which is a proxy for the percentage whose HPD interval would not contain zero.

The correlograms of both sparse and full models show similar estimates, although pairwise correlations are slightly stronger in the full model (Figure 3b and 3c). From the full model perspective, all ten pairwise correlations are significant, since $\boldsymbol{ps} = 1$ for each trait combination (Figure 3c). Correlations are all positive and relatively strong, varying between 0.65 and 0.99. Interestingly, note that stronger correlations correspond to the pairwise variables that share an edge in the sparse model graph. Notice also that the association structure represented by the estimated graph strongly enhances our ability to interpret and translate the intricate correlations presented in the correlograms.

From Figure 3a one can see that the wing length is directly associated with tarsus length ($\hat{r}_{WingL,TarsusL} = 0.84$, $pe = 0.99$), culmen length ($\hat{r}_{WingL,CulmenL} = 0.83$, $pe = 0.97$), and gonys width ($\hat{r}_{WingL,GonysW} = 0.74$, $pe = 0.76$). However, conditioning on the wing length and the remaining variables, culmen length and tarsus length and gonys width are pairwise independent, which suggest that there is no evidence for direct interaction between these traits during their evolution among analysed finch species. Additionally, the beak depth is directly and exclusively associated with gonys width ($\hat{r}_{BeakD,GonysW} = 0.98$, $pe = 1$), highlighting their decisive conditional dependency given the other variables in the sparse model. The conditional independence found between culmen length and beak depth ($pe = 0.06$) and culmen length and gonys width ($pe = 0.07$) couple with the findings of a sequence of studies on the identification of a regulatory network governing the morphology of the prenasal cartilage (*pnc*) (Abzhanov et al., 2004, 2006), and another one controlling the premaxillary bone (*pmx*) (Mallarino et al., 2011). The modularity embedded in these conditional independencies might help explain how the Finches evolutionary system lead to so much phenotypic variability in beak morphology.

First, Abzhanov et al. (2004) found that the expression of the bone morphogenetic protein 4 (*Bmp4*) in the mesenchyme of the upper beaks strongly correlated with deep and broad beak morphology, explaining the linkage in their variation. This *Bmp4* regulatory pathway could explain why beak depth and gonys width share an edge in the posterior graph. However, it is important to consider that those results were obtained for the upper beak only, and our finches data set provide only: i) gonys width measures, which correspond to the lower beak; and ii) beak depth without discriminating between upper and lower parts. Therefore, a more detailed data set is required to build a proper causal connection between the conditional dependency found in the posterior graph and the *Bmp4* regulatory pathway in lower beaks measurements. Additionally, Abzhanov et al. (2006) found that local upregulation of the calmodulin-dependent pathway is likely to have been a component of the evolution of Darwin's finch species with elongated beak morphology and provide a mechanistic explanation for the conditional independence of beak evolution between length and width/depth axes. Both *Bmp4* and *CaM* regulate morphogenesis of the prenasal cartilage (*pnc*) in early development, which forms the initial beak skeleton. However, much of the beak diversity in birds depends on variation in the premaxillary bone (*pmx*), that forms later in development and becomes the most prominent functional and structural component of the adult upper beak/jaw.

Second, Mallarino et al. (2011) found that *TGFβIIr*, *β-catenin*, and *Dickkopf-3* are differentially expressed in the developing premaxillary bone of embryos of species with different beak shapes affecting beak length and depth, which might explain how the tightly coupled depth and width dimensions can evolve to some extent conditionally independently. Altogether, the two-module program of development involving independent regulating molecules offers unique insights into how different developmental pathways may be modified and combined to induce multidimensional shifts in beak morphology.



(a) Sparse model $\hat{\mathbf{G}}$

(b) Sparse model $\hat{\mathbf{R}}$ and $\boldsymbol{pe}$

(c) Full model $\hat{\mathbf{R}}$ and $\boldsymbol{ps}$

Figure 3: Association structure and correlation between Darwin's Finches morphometric traits in sparse and full models. Graph edge thickness and transparency represent the posterior edge inclusion probability $\boldsymbol{pe}$ in the sparse model. The ellipses below correlogram diagonal summarize the posterior mean correlations $\hat{r}_{ij}$ between each pair of traits. In the sparse model, the numbers above the diagonal report the posterior edge inclusion probability $\boldsymbol{pe}$ (Figure 3b), while in the full model they report the posterior probability that the correlation is of the same sign as its mean $\boldsymbol{ps}$, as an alternative visualization of HPD (Figure 3c). Highlighted numbers above diagonal indicate included edges in the posterior graph (sparse model) or significant correlations using $\text{HPD}_{95\%}$ (full model).

This hypothesis, however, might explain just a part of the complex beak measurement variation, since in the graph the edge between BeakD and GonysW display the highest posterior inclusion probability $pe_{BeakD,GonysW} = 1$, and the strongest correlation $\hat{r}_{BeakD,GonysW} = 0.98$, revealing a decisive support in favor of the conditionally dependence structure between these traits.

5.2 ***Prokaryotes.*** We revisit the application of Hassler et al. (2020) concerning correlation estimates among a set of genotypic and phenotypic prokaryote traits. Our approach extends the MBD model for phenotypic trait evolution in Hassler et al. (2020) to include the estimation of the diffusion graph representing the underlying association structure between traits.

Hassler et al. (2020) collect data for $N = 705$ prokaryotes, combining cell diameter (CellD), cell length (CellL), optimum temperature (Topt), and pH measurements from Goberna and Verdú (2016), as well as data on genome length (GenL), coding sequences length (CDS), and GC content (GC) from the prokaryotes table in NCBI Genome. As in Hassler et al. (2020), we use 16S sequences to infer the phylogeny. After fitting sparse and full models, according to model setup described in Hassler et al. (2020, Section 7.2), we obtain posterior samples for parameters of scientific interest. We discard the first 20% of the samples as warm-up. From the posterior distribution of $\mathbf{R}$ and $\mathbf{G}$ we estimate the posterior graph $\hat{\mathbf{G}}$, and the posterior correlation $\hat{\mathbf{R}}$.

Figure 4a displays the estimated posterior graph structure $\hat{\mathbf{G}}$ with the associations between trait measurements, and Figure 4b and 4c present the posterior evolutionary correlation $\hat{\mathbf{R}}$ between those traits in both sparse and full models. In the full model, marked ellipses and upper diagonal numbers single out significant correlations using $\text{HPD}_{95\%}$ criteria to perform correlation selection, whereas, in the sparse model, they indicate that corresponding edges are included in the posterior graph.

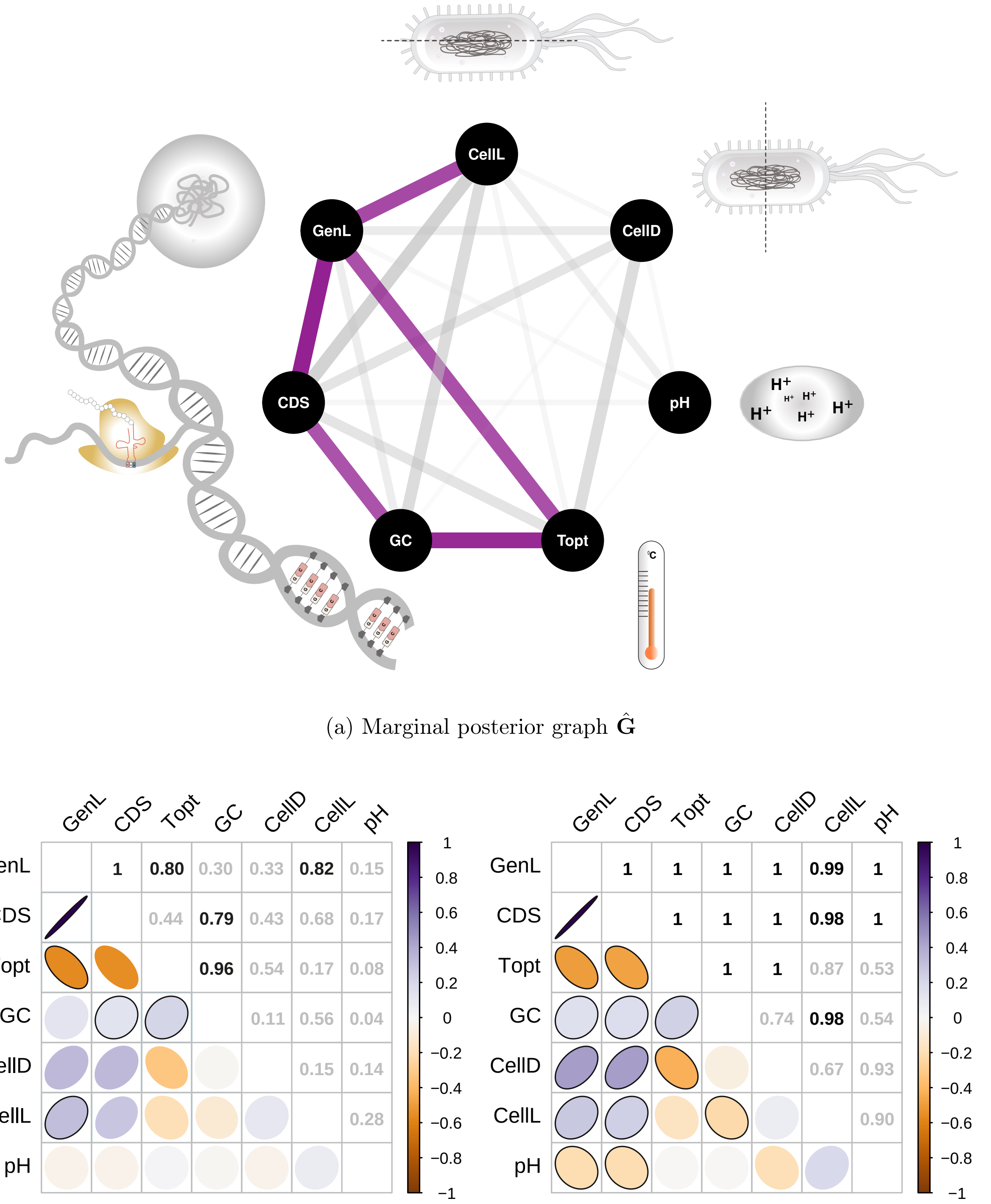
By comparing both correlograms, one can see that correlations are similar for most pairs of traits. However, we see an important divergence between the models in the correlation of pH with both genome length and coding sequence length. While the full model identifies those correlations as significant ($\hat{r}_{pH,\ genL} = -0.20$, $\text{HPD}_{95\%} = [-0.34, -0.05]$ and $\hat{r}_{pH,\ CDS} = -0.20$, $\text{HPD}_{95\%} = [-0.35, -0.06]$), the sparse model shrinks them towards zero ($\hat{r}_{pH,\ genL} \approx \hat{r}_{pH,\ CDS} = -0.04$). The posterior graph shows that pH is conditionally independent to all other traits in the sparse model (Figure 4a), suggesting that pH may have not coupled with any of them during prokaryote evolution.

We focus on the results obtained from the sparse model and take advantage of the information in the graph structure to drive our interpretations of the diffusion correlations. The graph structure suggests that the temperature might play an important role as a selective pressure in prokaryote evolution, since prokaryote optimum temperature are directly associate with two well discussed hypothesis, namely genome streamlining (Sela et al., 2016) and thermal adaptation (Bernardi and Bernardi, 1986).

One can see that higher optimal temperatures are directly associated with smaller genome length ($\hat{r}_{Topt,genL} = -0.56$, $pe = 0.80$) and smaller genomes are also directly associated with reduced coding sequences (CDS), as informed by their conditional dependence with high edge inclusion probability in the graph, and the extreme positive correlation between those traits ($\hat{r}_{genL,CDS} = 0.99$, $pe = 1$). Although CDS and optimal temperature display a relatively strong negative correlation ($\hat{r}_{Topt,CDS} = -0.56$), the estimated graph indicates that they might actually be conditionally independent given the genome length and the remaining traits in the model ($pe_{Topt,CDS} = 0.44$). In addition, genL is also directly associated and positively correlated with the cell length ($\hat{r}_{genL,cellL} = 0.32$, $pe = 0.82$). This result suggest that optimal temperature indirectly affects CellL and CDS through its direct effect on genome length. Those results corroborate with the genome streamlining hypothesis which states that certain prokaryotic genomes tend to be small in size, due to selection against the retention of non-coding DNA, and in favor of faster replication rates (Sela et al., 2016).

Moreover, we find optimal temperature to be also directly associated with the GC content in prokaryotic genomes displaying a positive correlation ($\hat{r}_{Topt,GC} = 0.21$, $pe = 0.96$). This result points to the polemic thermal adaptation hypothesis which posits that higher GC content is involved in adaptation to high temperatures because it may offer thermostability to genetic material (Bernardi and Bernardi, 1986).

Hurst and Merchant (2001) refuted the thermal adaptation hypothesis by founding no correlation between optimal temperatures and GC3 content which would provide a strong evidence for the hypothesis. Nevertheless, data from only 100 prokaryotes were used to achieve this result. Additionally, in a recent study, Hu et al. (2022) showed positive correlations between optimal growth temperature (Topt) and GC content both in bacterial and archaeal structural RNA genes and in bacterial whole genome sequences, chromosomal sequences, plasmid sequences, core genes, and accessory genes, providing additional support for the thermal adaptation hypothesis.



(a) Marginal posterior graph $\hat{\mathbf{G}}$

(b) Sparse model $\hat{\mathbf{R}}$ and $\boldsymbol{pe}$

(c) Full model $\hat{\mathbf{R}}$ and $\boldsymbol{ps}$

Figure 4: Association structure and correlation among prokaryotic growth properties. See Figure 3 for caption. In the sparse model, marked ellipses and numbers correspond to the edges included in the marginal posterior graph $\mathbf{G}$, while in the full model they highlight the significant correlations according to the $\text{HPD}_{95\%}$ criteria.

6 Computational Efficiency

Here we provide a summary of the computational cost of including graph estimation in SPTE model. We formalize our comparison by computing parameter update cost the proportion of time spent updating the parameter of interest ($\mathbf{K}$ in full model and $(\mathbf{K}, \mathbf{G})$ in sparse model) and the minimum and median effective sample size (ESS) per minute for the diffusion correlation under both sparse and full models (Table 3).

We fixed the phylogenetic tree in the simulation study, whereas estimate $\mathscr{F}$ in the application data sets. Notice that, in the simulation data, we can directly compare the computational cost of equipping the MBD model with GGM because we only update $(\mathbf{K}, \mathbf{G})$ in the sparse model and $\mathbf{K}$ in the full model. This is the reason why parameter update cost is 100% in both *Sim* 1 and *Sim* 2. When the tree $\mathscr{F}$ is fixed, we see that ESS per second decreases between 76 and 82%. Indeed, sparse models are computationally more expensive compared to full models due to the graph update burden.

On the other hand, when the tree is simultaneously estimated, most of the computational effort involve the phylogenetic tree estimation such that the computational cost of structure learning decreases compared to the global MCMC cost. One can see that the computational cost of updating the diffusion graph is less than 5% in the worst scenario. The Darwin's Finches and *Sim* 1 are both $p = 5$ data sets, but the speed-down for the median ESS/minute drops from 76.5% in *Sim* 1 to 20.3% in finches data. Note also that in the prokaryotes application, although updating $(\mathbf{K}, \mathbf{G})$ requires more computational effort than updating $\mathbf{K}$ only (full model), both update costs are negligible compared to the global update cost of MCMC (1.26% and 1.30%, respectively).

Table 3: Computational cost of graph estimation. We report MCMC sampling efficiency through the update parameter cost, which is the proportion of time spent to update the parameter of interest in comparison to global MCMC cost, and by computing the minimum and median effective sample size (ESS) per minute.

Data set	$\mathscr{F}$	N	p	Model	Update		$\mathbf{R}$ ESS/minute	
					parameter	cost	minimum	median
Sim 1	Fixed	50	5	Full	$\mathbf{K}$	100%	4445.59	5069.41
				Sparse	$(\mathbf{K}, \mathbf{G})$	100%	1053.96	1189.19
				Speed-down	-	-	**76.3%**	**76.5%**
Sim 2	Fixed	100	10	Full	$\mathbf{K}$	100%	1606.49	2273.61
				Sparse	$(\mathbf{K}, \mathbf{G})$	100%	285.94	426.14
				Speed-down	-	-	**82.2%**	**81.3%**
Darwin's finches	Estimated	13	5	Full	$\mathbf{K}$	0.66%	1244.43	1244.43
				Sparse	$(\mathbf{K}, \mathbf{G})$	4.83%	887.56	992.03
				Speed-down	-	-	**28.7%**	**20.3%**
Prokaryotes	Estimated	705	7	Full	$\mathbf{K}$	1.26%	1.02	2.45
				Sparse	$(\mathbf{K}, \mathbf{G})$	1.30%	0.35	1.84
				Speed-down	-	-	**66.1%**	**24.8%**

7 Discussion

We present a Bayesian inference framework to perform a model-based covariance selection, using Gaussian graphical models, in order to learn the association structure between continuous traits while jointly inferring the trait evolutionary correlations and the phylogenetic tree of related taxa through sequence data. By doing so, we introduce another parameter of scientific interest, the diffusion graph $\mathbf{G}$ that complements the information provided by the trait correlations estimated in these phylogenetic trait evolution models.

Our approach significantly improves upon traditional trait evolution models in terms of modeling and inference. As shown in the simulation study, our model also provides better estimates for the diffusion precision and diffusion correlation, specially for independent variables — which are the main target for sparsity —, while displaying similar logMSE for dependent variables (CI-D and CD-D). Additionally, SPTE model can accurately identify the association structure between traits. The statistical performance for graph estimation in the sparse model is better than the one for post-process correlation selection in the full model, which highlights the advantages of a model-based approach for covariance selection.

When applying the full model we can merely identify significantly correlated traits, discuss the strength of these correlations, and use it to guide the search for potential explanations — in a possibly dense correlogram. On the other hand, more than simply inferring the evolutionary correlations, the sparse model additionally informs about the association structure between traits, which is encoded in the estimated diffusion graph. The association structure can help us refine the search for potential mechanisms to explain the conditional (in)dependencies revealed by the diffusion graph underlying the dependencies presented by the correlograms.

By analyzing the correlograms alone we are not able to understand the nuances in trait relationships. In many cases, the correlations might capture indirect effects of the association structure itself rather than translate biological phenomenon. For example, if we have two variables A and B each independently associated with a third variable C, the changes in C might affect the variation in both A and B. These changes in A and B, thereafter, are likely to express some indirect pattern due to their individual connection to C. This indirect pattern can, ultimately, be captured by a correlation coefficient. This could explain why it is difficult to distinguish the correlations directly originated from causal relationships between traits in the study to the ones representing indirect effects of the underlying association structure — or the ones mediated by phenomena not considered in the study —, by just relying on the correlations.

The results from our applications indicate that combining information from correlations with the conditional independencies in the diffusion graph, however, allows for a more precise selection of candidate traits that may interact along the evolutionary process of related organisms. For this reason, learning the association structure is imperative, particularly when trait dimension p increases.

Another important advantage of our approach is that, with the appropriate adaptations, it can be integrated to a broad range of Gaussian models due to its readily use feature. Under a computational perspective, including graph estimation in the MBD model only requires changes in the precision matrix update mechanism to jointly obtain $p(\mathbf{K}, \mathbf{G})$. The proposed novelty does not make any impact on model likelihood, since model dependence on $\mathbf{G}$ is completely mediated by $\mathbf{K}$, i.e. $p(\mathbf{X}|\mathbf{K}, \mathbf{G}, \mathscr{F}) = p(\mathbf{X}|\mathbf{K}, \mathscr{F})$. Therefore, as the changes consists on prior and hyperprior structure choices, the approaches employed for likelihood computations remain intact. This is a desirable and convenient feature because it enables our GGM approach to potentially profit from any future computational improvement in trait evolution models. For example, in the prokaryotes application we were able to perform covariance selection on this massive data set by building upon the efficient approach developed by Hassler et al. (2020) which integrates out missing values and allows for previously intractable analyses on large trees.

The sparse model is computationally more expensive than the full model due to the additional steps required for graph estimation such as the computation of G-Wishart prior $I_G(\delta, \mathbf{D})$ and posterior $I_G(\delta + N, \mathbf{D} + \boldsymbol{\Delta})$ normalizing constants to perform graph updates. Additionally, chains must be run longer to account for increased complexity in the parametric space. This restriction, however, is not overly limiting because when we simultaneously estimate the phylogenetic tree, the computational cost of graph estimation pales in comparison to the global cost of MCMC.

One possible future improvement for our approach lies in the mechanism choice to perform graph updates. Graph estimation is incredibly challenging given the dimensionality of the graph space. The employed graph updates using the ratio of normalizing constants are convenient because they do not require any additional implementations, since G-Wishart normalizing constant approximations should be mandatorily implemented for G-Wishart likelihood computations. In spite of convenience, this is one of the early approaches to tackle this expensive step in GGM. Algorithms such as birth-death MCMC (BDMCMC) (Mohammadi and Wit, 2015) or G-Wishart weighted proposal algorithm (WWA) (Boom et al., 2021) are potential directions to explore.

Additionally, we did not perform extensive simulations to characterize the performance of the presented methodology. Examining a broader range of simulation conditions such as different graph structures $\mathbf{G}_0$ and trait dimension p, is an important future direction to improve SPTE.

Finally, while we do not explore this in simulations or application, as presented in Section 4 and 5, the sparse phylogenetic trait evolution model can be further adapted to deal with binary, categorical, and ordinal data as in Cybis et al. (2015); Zhang et al. (2021), which will only add to the model's broad applicability. The biggest challenge for this extension is how to bypass the identifiability issue on the diffusion precision. One way to achieve that is using a parameter expansion for data augmentation (PXDA) approach (Chib and Greenberg, 1998; Talhouk et al., 2012).

References

Abzhanov, A., Kuo, W. P., Hartmann, C., Grant, B. R., Grant, P. R., and Tabin, C. J. (2006). The calmodulin pathway and evolution of elongated beak morphology in darwin's finches. *Nature*, 442(7102):563–567.

Abzhanov, A., Protas, M., Grant, B. R., Grant, P. R., and Tabin, C. J. (2004). Bmp4 and morphological variation of beaks in darwin's finches. *Science*, 305(5689):1462–1465.

Atay-Kayis, A. and Massam, H. (2005). A monte carlo method for computing the marginal likelihood in nondecomposable gaussian graphical models. *Biometrika*, 92(2):317–335.

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

Bernardi, G. and Bernardi, G. (1986). Compositional constraints and genome evolution. *Journal of molecular evolution*, 24(1):1–11.

Boom, W. v. d., Beskos, A., and De Iorio, M. (2021). The g-wishart weighted proposal algorithm: Efficient posterior computation for gaussian graphical models. *arXiv preprint arXiv:2108.01308*.

Carvalho, C. M. and Scott, J. G. (2009). Objective bayesian model selection in gaussian graphical models. *Biometrika*, 96(3):497–512.

Chib, S. and Greenberg, E. (1998). Analysis of multivariate probit models. *Biometrika*, 85(2):347–361.

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. *The Annals of Applied Statistics*, 9(2):969.

Dempster, A. P. (1972). Covariance selection. *Biometrics*, 28(1):157–175.

Drummond, A. J., Suchard, M. A., Xie, D., and Rambaut, A. (2012). Bayesian phylogenetics with beauti and the beast 1.7. *Molecular Biology and Evolution*, 29(8):1969–1973.

Felsenstein, J. (2012). A comparative method for both discrete and continuous characters using the threshold model. *The American Naturalist*, 179(2):145–156.

Gaskins, J. (2019). Hyper markov laws for correlation matrices. *Statistica Sinica*, 29(1):165–184.

Goberna, M. and Verdú, M. (2016). Predicting microbial traits with phylogenies. *The ISME Journal*, 10(4):959–967.

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

Hu, E.-Z., Lan, X.-R., Liu, Z.-L., Gao, J., and Niu, D.-K. (2022). A positive correlation between gc content and growth temperature in prokaryotes. *BMC genomics*, 23(1):1–17.

Hurst, L. D. and Merchant, A. R. (2001). High guanine–cytosine content is not an adaptation to high temperature: a comparative analysis amongst prokaryotes. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 268(1466):493–497.

Jeffreys, H. (1998). *The theory of probability*. OUP Oxford.

Lauritzen, S. L. (1996). *Graphical Models*. Clarendon Press.

Lenkoski, A. and Dobra, A. (2011). Computational aspects related to inference in gaussian graphical models with the g-wishart prior. *Journal of Computational and Graphical Statistics*, 20(1):140–157.

Letac, G. and Massam, H. (2007). Wishart distributions for decomposable graphs. *The Annals of Statistics*, 35(3):1278 – 1323.

Letac, G., Massam, H., and Mohammadi, R. (2017). The ratio of normalizing constants for bayesian graphical gaussian model selection. *arXiv preprint arXiv:1706.04416*.

Li, Z. R., McComick, T. H., and Clark, S. J. (2020). Using bayesian latent gaussian graphical models to infer symptom associations in verbal autopsies. *Bayesian Analysis*, 15(3):781.

Liu, J. S., Wong, W. H., and Kong, A. (1995). Covariance structure and convergence rate of the gibbs sampler with various scans. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):157–169.

Maathuis, M., Drton, M., Lauritzen, S., and Wainwright, M. (2018). *Handbook of graphical models*. CRC Press.

Mallarino, R., Grant, P. R., Grant, B. R., Herrel, A., Kuo, W. P., and Abzhanov, A. (2011). Two developmental modules establish 3d beak-shape variation in darwin's finches. *Proceedings of the National Academy of Sciences*, 108(10):4057–4062.

Mitsakakis, N. (2010). *Bayesian Methods in Gaussian Graphical Models*. PhD thesis, University of Toronto.

Mohammadi, A. and Wit, E. C. (2015). Bayesian Structure Learning in Sparse Gaussian Graphical Models. *Bayesian Analysis*, 10(1):109 – 138.

Mohammadi, R., Massam, H., and Letac, G. (2021). Accelerating bayesian structure learning in sparse gaussian graphical models. *Journal of the American Statistical Association*, pages 1–14.

Pybus, O. G., Suchard, M. A., Lemey, P., Bernardin, F. J., Rambaut, A., Crawford, F. W., Gray, R. R., Arinaminpathy, N., Stramer, S. L., Busch, M. P., et al. (2012). Unifying the spatial epidemiology and molecular evolution of emerging epidemics. *Proceedings of the National Academy of Sciences*, 109(37):15066–15071.

Roverato, A. (2000). Cholesky decomposition of a hyper inverse wishart matrix. *Biometrika*, 87(1):99–112.

Roverato, A. (2002). Hyper inverse wishart distribution for non-decomposable graphs and its application to bayesian inference for gaussian graphical models. *Scandinavian Journal of Statistics*, 29(3):391–411.

Sela, I., Wolf, Y. I., and Koonin, E. V. (2016). Theory of prokaryotic genome evolution. *Proceedings of the National Academy of Sciences*, 113(41):11399–11407.

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 evolution*, 4(1):vey016.

Talhouk, A., Doucet, A., and Murphy, K. (2012). Efficient bayesian inference for multivariate probit models with sparse inverse correlation matrices. *Journal of Computational and Graphical Statistics*, 21(3):739–757.

Tolkoff, M. R., Alfaro, M. E., Baele, G., Lemey, P., and Suchard, M. A. (2018). Phylogenetic factor analysis. *Systematic Biology*, 67(3):384–399.

Tung Ho, L. s. and Ané, C. (2014). A linear-time algorithm for gaussian and non-gaussian trait evolution models. *Systematic Biology*, 63(3):397–408.

Uhler, C., Lenkoski, A., Richards, D., et al. (2018). Exact formulas for the normalizing constants of wishart distributions for graphical models. *The Annals of Statistics*, 46(1):90–118.

Williams, D. R. (2021). Bayesian estimation for gaussian graphical models: Structure learning, predictability, and network comparisons. *Multivariate Behavioral Research*, 56(2):336–352.

Zhang, Z., Nishimura, A., Bastide, P., Ji, X., Payne, R. P., Goulder, P., Lemey, P., and Suchard, M. A. (2021). Large-scale inference of correlation among mixed-type biological traits with phylogenetic multivariate probit models. *The Annals of Applied Statistics*, 15(1):230–251.

Supplementary Information

SI 1 Simulation Study

Here we provide more the details for the simulations conducted to compare the performances of both sparse and full models. First we present the true diffusion precision matrices and the corresponding diffusion correlation matrices for simulation *Sim* 1 and *Sim* 2. For *Sim* 1 we define $\mathbf{K}_0$ and $\mathbf{R}_0$ as

$$\mathbf{K}_0 = \begin{pmatrix} 3.87 & 3.62 & 0 & 0 & 0 \\ 3.62 & 8.50 & -4.87 & 0 & 0 \\ 0 & -4.87 & 5.13 & 0 & 0 \\ 0 & 0 & 0 & 6.55 & -6.29 \\ 0 & 0 & 0 & 6.29 & 6.55 \end{pmatrix}, \tag{SI.1}$$

and

$$\mathbf{R}_0 = \begin{pmatrix} 1.00 & -0.93 & -0.89 & 0 & 0 \\ -0.93 & 1.00 & 0.95 & 0 & 0 \\ -0.89 & 0.95 & 1.00 & 0 & 0 \\ 0 & 0 & 0 & 1.00 & 0.96 \\ 0 & 0 & 0 & 0.96 & 1.00 \end{pmatrix} \tag{SI.2}$$

For *Sim* 2 we define

$$\mathbf{K}_0 = \begin{pmatrix} 9.52 & -7.26 & 4.25 & -3.64 & 0 & 0 & 0 & 0 & 0 & 0 \\ -7.26 & 8.17 & -3.55 & 2.34 & 0 & 0 & 0 & 0 & 0 & 0 \\ 4.25 & -3.55 & 10.3 & 3.34 & 0 & 0 & 0 & 0 & 0 & 0 \\ -3.64 & 2.34 & 3.34 & 4.46 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.59 & -0.59 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -0.59 & 2.91 & -1.54 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1.54 & 1.35 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 7.84 & 1.12 & 0.12 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1.12 & 2.86 & 1.18 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.12 & 1.18 & 3.16 \end{pmatrix}, \tag{SI.3}$$

and

$$\mathbf{R}_0 = \begin{pmatrix} 1 & 0.6 & -0.96 & 0.97 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.6 & 1 & -0.41 & 0.43 & 0 & 0 & 0 & 0 & 0 & 0 \\ -0.96 & -0.41 & 1 & -0.98 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.97 & 0.43 & -0.98 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0.72 & 0.62 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.72 & 1 & 0.87 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.62 & 0.87 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -0.25 & 0.08 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -0.25 & 1 & -0.4 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.08 & -0.4 & 1 \end{pmatrix}. \tag{SI.4}$$

Table 1 show the definifion of the statistics used to evaluate the performances of both full and sparse models presented in Section 4, based on the following confusion matrix

	True	
Predicted	1	0
1	TP	FP
0	FN	TN

where TP is the true positives, FP is the false positives, FN is the false negatives and TN represent the true negatives. In our model, the variables are correspond to presence or absence of edges in sparse models or non-zero or zero correlations.

Table 1: Definition of the statistics used to assess the performances of graph estimation in the sparse model and post-process correlation selection with HPD criteria in the full model. TP = True Positive; TN = True Negative; FP = False Positive; FN = False Negative.

Statistic	**Definition**
Accuracy	$\frac{TP+TN}{TP+FP+FN+TN}$
Sensitivity	$\frac{TP}{TP+FN}$
Specificity	$\frac{TN}{TN+FP}$
Precision	$\frac{TP}{TP+FP}$
F1-Score	$\frac{2TP}{2TP+FP+FN}$

SI 2 Graph updates

Here we present further details to obtain the marginal distribution of $\mathbf{G}$ given all the other parameters in the model, except $\mathbf{K}$. Under the non-informative prior choices for diffusion graph in (5) and diffusion precision (6), the joint density $p(\mathbf{X}, \mathbf{K}, \mathbf{G}|\mathscr{F}, \delta, \mathbf{D})$ is given by

$$\begin{aligned}
p(\mathbf{X}, \mathbf{K}, \mathbf{G}|\delta, \mathbf{D}, \mathscr{F}) &= p(\mathbf{X}|\mathbf{K}, \mathscr{F})p(\mathbf{K}|\mathbf{G}, \delta, \mathbf{D}, \mathscr{F})p(\mathbf{G}|\delta, \mathbf{D}) \\
&= \frac{|\mathbf{K}|^{n/2}}{(2\pi)^{np/2}} \exp\left\{-\frac{1}{2}\text{tr}(\mathbf{UK})\right\} \\
&\quad \times \frac{1}{I(\delta, \mathbf{D})}|\mathbf{K}|^{(\delta-2)/2} \exp\left\{-\frac{1}{2}\text{tr}(\mathbf{DK})\right\}\mathbf{1}_{\mathbf{K}\in\mathbb{P}_G}\frac{1}{|\mathcal{G}|} \\
&= \frac{1}{(2\pi)^{np/2}}\frac{1}{|\mathcal{G}|}\frac{1}{I_G(\delta, \mathbf{D})}|\mathbf{K}|^{(\delta^\star-2)/2} \exp\left\{-\frac{1}{2}\text{tr}(\mathbf{D}^\star\mathbf{K})\right\}\mathbf{1}_{\mathbf{K}\in\mathbb{P}_G}. \qquad \text{(SI.5)}
\end{aligned}$$

where $\delta^\star = \delta + N$, $\mathbf{D}^\star = \mathbf{D} + \mathbf{\Delta}$, and $\mathbf{\Delta} = (\mathbf{X} - \mathbf{1}_N\boldsymbol{\mu}_0^t)^t\left(\mathbf{\Upsilon} + \tau_0^{-1}\mathbf{J}_N\right)^{-1}(\mathbf{X} - \mathbf{1}_N\boldsymbol{\mu}_0^t)$. In order to obtain the marginal distribution of $\mathbf{X}$, given all other parameters except $\mathbf{K}$, we integrate the joint distribution (SI.5) over the possible values for $\mathbf{K}$,

$$p(\mathbf{X}|\mathbf{G}, \delta, \mathbf{D}, \mathscr{F}) = \frac{p(\mathbf{X}, \mathbf{G}|\delta, \mathbf{D}, \mathscr{F})}{p(\mathbf{G}|\delta, \mathbf{D}, \mathscr{F})} = \frac{\int_{\mathbf{K}\in\mathbb{P}_G} p(\mathbf{X}, \mathbf{K}, \mathbf{G}|\delta, \mathbf{D}, \mathscr{F})\, d\mathbf{K}}{p(\mathbf{G}|\delta, \mathbf{D}, \mathscr{F})}. \qquad \text{(SI.6)}$$

By replacing the joint density (SI.5) in the kernel of the integral in Equation (SI.6), we have

$$\begin{aligned}
p(\mathbf{X}|\mathbf{G}, \delta, \mathbf{D}, \mathscr{F}) &= |\mathcal{G}|\frac{1}{(2\pi)^{Np/2}}\frac{1}{|\mathcal{G}|}\frac{1}{I_G(\delta, \mathbf{D})}\int_{\mathbf{K}\in\mathbb{P}_G}|\mathbf{K}|^{(\delta^\star-2)/2} \exp\left\{-\frac{1}{2}\text{tr}(\mathbf{D}^\star\mathbf{K})\right\}\mathbf{1}_{\mathbf{K}\in\mathbb{P}_G}\, d\mathbf{K} \\
&= \frac{1}{(2\pi)^{Np/2}}\frac{I_G(\delta^\star, \mathbf{D}^\star)}{I_G(\delta, \mathbf{D})}. \qquad \text{(SI.7)}
\end{aligned}$$

Note that the kernel of the integral in Equation (SI.7) corresponds to the posterior of the diffusion precision, whose distribution is $\mathcal{W}_G(\delta^\star = \delta + N, \mathbf{D}^\star = \mathbf{D} + \mathbf{\Delta})$. Therefore the marginal distribution of $\mathbf{G}$, given all other parameters except $\mathbf{K}$,

$$p(\mathbf{G}|\mathbf{X}, \delta, \mathbf{D}, \mathscr{F}) \propto p(\mathbf{G}|\delta, \mathbf{D})p(\mathbf{X}|\mathbf{G}, \delta, \mathbf{D}, \mathscr{F}) = \frac{p(\mathbf{G}|\delta, \mathbf{D})}{(2\pi)^{Np/2}}\frac{I_G(\delta + N, \mathbf{D} + \mathbf{\Delta})}{I_G(\delta, \mathbf{D})}. \qquad \text{(SI.8)}$$

Hence, computing the marginal likelihood (SI.6) or the posterior distribution (SI.8) is reduced to the problem of computing normalising constants of the type $I_G(\delta, \mathbf{D})$, with $\delta > 0$ and $\mathbf{D}$ positive definite, which are sufficient conditions for convergence of the normalizing constants (Mitsakakis, 2010, Lemma 3.2.1: $I_G(\delta, \mathbf{D}) < \infty$ for $\delta > 0$).

Bibliography

Archibald, J. M. and Roger, A. J. (2002). Gene duplication and gene conversion shape the evolution of archaeal chaperonins. *Journal of molecular biology*, 316(5):1041–1050.

Atay-Kayis, A. and Massam, H. (2005). A monte carlo method for computing the marginal likelihood in nondecomposable gaussian graphical models. *Biometrika*, 92(2):317–335.

Boom, W. v. d., Beskos, A., and De Iorio, M. (2021). The g-wishart weighted proposal algorithm: Efficient posterior computation for gaussian graphical models. *arXiv preprint arXiv:2108.01308*.

Brooks, S., Gelman, A., Jones, G., and Meng, X.-L. (2011). *Handbook of markov chain monte carlo*. CRC press.

Carlin, J. (2011). Mutations are the raw materials of evolution. *Nature Education Knowledge*, 3(10):10.

Carvalho, C. M. and Scott, J. G. (2009). Objective bayesian model selection in gaussian graphical models. *Biometrika*, 96(3):497–512.

Cordero, G. A. and Janzen, F. (2013). Does life history affect molecular evolutionary rates? *Nature Education Knowledge*, 4(4):1.

Cvijović, I., Good, B. H., Jerison, E. R., and Desai, M. M. (2015). Fate of a mutation in a fluctuating environment. *Proceedings of the National Academy of Sciences*, 112(36):E5021–E5028.

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. *The Annals of Applied Statistics*, 9(2):969.

Dawid, A. P. (1981). Some matrix-variate distribution theory: notational considerations and a bayesian application. *Biometrika*, 68(1):265–274.

Dawid, A. P. and Lauritzen, S. L. (1993). Hyper markov laws in the statistical analysis of decomposable graphical models. *The Annals of Statistics*, pages 1272–1317.

Dempster, A. P. (1972). Covariance selection. *Biometrics*, 28(1):157–175.

Drummond, A. J., Suchard, M. A., Xie, D., and Rambaut, A. (2012). Bayesian phylogenetics with beauti and the beast 1.7. *Molecular Biology and Evolution*, 29(8):1969–1973.

Felsenstein, J. (1981). Evolutionary trees from dna sequences: a maximum likelihood approach. *Journal of molecular evolution*, 17(6):368–376.

Felsenstein, J. (2012). A comparative method for both discrete and continuous characters using the threshold model. *The American Naturalist*, 179(2):145–156.

Felsenstein, J. and Felenstein, J. (2004). *Inferring phylogenies*, volume 2. Sinauer associates Sunderland, MA.

Gaskins, J. (2019). Hyper markov laws for correlation matrices. *Statistica Sinica*, 29(1):165–184.

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

Geman, S. and Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, (6):721–741.

Gillespie, J. H. (1994). *The causes of molecular evolution*, volume 2. Oxford University Press On Demand.

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

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications.

Hössjer, O., Bechly, G., and Gauger, A. (2021). On the waiting time until coordinated mutations get fixed in regulatory sequences. *Journal of Theoretical Biology*, 524:110657.

Huelsenbeck, J. P. and Ronquist, F. (2001). Mrbayes: Bayesian inference of phylogenetic trees. *Bioinformatics*, 17(8):754–755.

Lauritzen, S. L. (1996). *Graphical Models*. Clarendon Press.

Leliaert, F., Smith, D. R., Moreau, H., Herron, M. D., Verbruggen, H., Delwiche, C. F., and De Clerck, O. (2012). Phylogeny and molecular evolution of the green algae. *Critical reviews in plant sciences*, 31(1):1–46.

Lemey, P., Salemi, M., and Vandamme, A.-M. (2009). *The phylogenetic handbook: a practical approach to phylogenetic analysis and hypothesis testing*. Cambridge University Press.

Lenkoski, A. and Dobra, A. (2008). Bayesian structural learning and estimation in gaussian graphical models. Technical Report 545, Department of Statistics, University of Washington.

Lenkoski, A. and Dobra, A. (2011). Computational aspects related to inference in gaussian graphical models with the g-wishart prior. *Journal of Computational and Graphical Statistics*, 20(1):140–157.

Letac, G. and Massam, H. (2007). Wishart distributions for decomposable graphs. *The Annals of Statistics*, 35(3):1278 – 1323.

Letac, G., Massam, H., and Mohammadi, R. (2017). The ratio of normalizing constants for bayesian graphical gaussian model selection. *arXiv preprint arXiv:1706.04416*.

Li, Z. R., McComick, T. H., and Clark, S. J. (2020). Using bayesian latent gaussian graphical models to infer symptom associations in verbal autopsies. *Bayesian Analysis*, 15(3):781.

Logares, R., Rengefors, K., Kremp, A., Shalchian-Tabrizi, K., Boltovskoy, A., Tengs, T., Shurtleff, A., and Klaveness, D. (2007). Phenotypically different microalgal morphospecies with identical ribosomal dna: a case of rapid adaptive evolution? *Microbial Ecology*, 53(4):549–561.

McClintock, B. (1984). The significance of responses of the genome to challenge. *Science*, 226(4676):792–801.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092.

Mitsakakis, N. (2010). *Bayesian Methods in Gaussian Graphical Models*. PhD thesis, University of Toronto.

Mitsakakis, N., Massam, H., and Escobar, M. D. (2011). A metropolis-hastings based method for sampling from the g-wishart distribution in gaussian graphical models. *Electronic Journal of Statistics*, 5:18–30.

Mohammadi, A. and Wit, E. C. (2015). Bayesian Structure Learning in Sparse Gaussian Graphical Models. *Bayesian Analysis*, 10(1):109 – 138.

Mohammadi, R., Massam, H., and Letac, G. (2021). Accelerating bayesian structure learning in sparse gaussian graphical models. *Journal of the American Statistical Association*, pages 1–14.

Muirhead, R. J. (1982). Aspects of multivariate statistical theory.

Nixon, K. C. (2001). Phylogeny. In Levin, S. A., editor, *Encyclopedia of Biodiversity (Second Edition)*, pages 16–23. Academic Press, Waltham, second edition edition.

Roverato, A. (2000). Cholesky decomposition of a hyper inverse wishart matrix. *Biometrika*, 87(1):99–112.

Roverato, A. (2002). Hyper inverse wishart distribution for non-decomposable graphs and its application to bayesian inference for gaussian graphical models. *Scandinavian Journal of Statistics*, 29(3):391–411.

Safran, R. J. and Nosil, P. (2012). Speciation: the origin of new species. *Nature Education Knowledge*, 3(10):17.

Shan, Y. and Li, X.-Q. (2008). Maximum gene-support tree. *Evolutionary Bioinformatics*, 4:EBO–S652.

Talhouk, A., Doucet, A., and Murphy, K. (2012). Efficient bayesian inference for multivariate probit models with sparse inverse correlation matrices. *Journal of Computational and Graphical Statistics*, 21(3):739–757.

Vulić, M., Lenski, R. E., and Radman, M. (1999). Mutation, recombination, and incipient speciation of bacteria in the laboratory. *Proceedings of the National Academy of Sciences*, 96(13):7348–7351.

Williams, D. R. (2021). Bayesian estimation for gaussian graphical models: Structure learning, predictability, and network comparisons. *Multivariate Behavioral Research*, 56(2):336–352.

Zhang, Z., Nishimura, A., Bastide, P., Ji, X., Payne, R. P., Goulder, P., Lemey, P., and Suchard, M. A. (2021). Large-scale inference of correlation among mixed-type biological traits with phylogenetic multivariate probit models. *The Annals of Applied Statistics*, 15(1):230–251.