Bayesian Methods for Trait Evolution

Paul Bastide[1]

[1] IMAG, Université de Montpellier, CNRS

December 2021

Bayesian Phylogeography

Outline

Maximum Likelihood

Likelihood: Distribution of the data given the parameters.

$$p\left(\mathbf{Y}, \mathbf{S} \mid \boldsymbol{\theta}, \mathcal{T}, \psi\right)$$

Maximum Likelihood

Likelihood: Distribution of the data given the parameters.

$$p\left(\mathbf{Y}, \mathbf{S} \mid \boldsymbol{\theta}, \mathcal{T}, \boldsymbol{\psi}\right)$$

Maximum likelihood: find the "best" possible parameters.

$$(\hat{\boldsymbol{\theta}}, \hat{\mathcal{T}}, \hat{\boldsymbol{\psi}}) = \underset{\boldsymbol{\theta}, \mathrm{tr}, \boldsymbol{\psi}}{\operatorname{argmax}}\, p\left(\mathbf{Y}, \mathbf{S} \mid \boldsymbol{\theta}, \mathcal{T}, \boldsymbol{\psi}\right)$$

Maximum Likelihood

Likelihood: Distribution of the data given the parameters.

$$p\,(\,\mathbf{Y}, \mathbf{S} \mid \boldsymbol{\theta}, \mathcal{T}, \boldsymbol{\psi}\,)$$

Maximum likelihood: find the "best" possible parameters.

$$(\hat{\boldsymbol{\theta}}, \hat{\mathcal{T}}, \hat{\boldsymbol{\psi}}) = \underset{\boldsymbol{\theta},\mathrm{tr},\boldsymbol{\psi}}{\operatorname{argmax}}\, p\,(\,\mathbf{Y}, \mathbf{S} \mid \boldsymbol{\theta}, \mathcal{T}, \boldsymbol{\psi}\,)$$

Limitations:

- No a priori knowledge.
  All parameters are equally probable.

Maximum Likelihood

Likelihood: Distribution of the data given the parameters.

$$p\,(\,\mathbf{Y}, \mathbf{S} \mid \boldsymbol{\theta}, \mathcal{T}, \boldsymbol{\psi}\,)$$

Maximum likelihood: find the "best" possible parameters.

$$(\hat{\boldsymbol{\theta}}, \hat{\mathcal{T}}, \hat{\boldsymbol{\psi}}) = \underset{\boldsymbol{\theta}, \mathrm{tr}, \boldsymbol{\psi}}{\operatorname{argmax}}\, p\,(\,\mathbf{Y}, \mathbf{S} \mid \boldsymbol{\theta}, \mathcal{T}, \boldsymbol{\psi}\,)$$

Limitations:

- No a priori knowledge.
  All parameters are equally probable.
- Point estimate.
  No notion of uncertainty.

Maximum Likelihood

Likelihood: Distribution of the data given the parameters.

$$p\,(\mathbf{Y}, \mathbf{S} \mid \boldsymbol{\theta}, \mathcal{T}, \boldsymbol{\psi}\,)$$

Maximum likelihood: find the "best" possible parameters.

$$(\hat{\boldsymbol{\theta}}, \hat{\mathcal{T}}, \hat{\boldsymbol{\psi}}) = \underset{\boldsymbol{\theta},\mathrm{tr},\boldsymbol{\psi}}{\operatorname{argmax}}\, p\,(\mathbf{Y}, \mathbf{S} \mid \boldsymbol{\theta}, \mathcal{T}, \boldsymbol{\psi}\,)$$

Limitations:

- No a priori knowledge.
  All parameters are equally probable.
- Point estimate.
  No notion of uncertainty.

  $\rightarrow$ Use Bayesian inference.

Bayesian Inference

Likelihood: Distribution of the data given the parameters.

$$p\left(\mathbf{Y}, \mathbf{S} \mid \boldsymbol{\theta}, \mathcal{T}, \psi\right)$$

Bayesian Inference

Likelihood: Distribution of the data given the parameters.

$$p\left(\mathbf{Y}, \mathbf{S} \mid \boldsymbol{\theta}, \mathcal{T}, \boldsymbol{\psi}\right)$$

Prior: Distribution of the parameters.

$$p\left(\boldsymbol{\theta}, \mathcal{T}, \boldsymbol{\psi}\right)$$

Bayesian Inference

Likelihood: Distribution of the data given the parameters.

$$p\left(\mathbf{Y}, \mathbf{S} \mid \boldsymbol{\theta}, \mathcal{T}, \psi\right)$$

Prior: Distribution of the parameters.

$$p\left(\boldsymbol{\theta}, \mathcal{T}, \psi\right)$$

Posterior: Distribution of the parameters given the data.

$$p\left(\boldsymbol{\theta}, \mathcal{T}, \psi \mid \mathbf{Y}, \mathbf{S}\right)$$

Bayesian Inference

Likelihood: Distribution of the data given the parameters.

$$p\,(\mathbf{Y}, \mathbf{S} \mid \boldsymbol{\theta}, \mathcal{T}, \boldsymbol{\psi}\,)$$

Prior: Distribution of the parameters.

$$p\,(\boldsymbol{\theta}, \mathcal{T}, \boldsymbol{\psi})$$

Posterior: Distribution of the parameters given the data.

$$p\,(\boldsymbol{\theta}, \mathcal{T}, \boldsymbol{\psi} \mid \mathbf{Y}, \mathbf{S}\,)$$

Bayesian Inference: Can we learn about the posterior ?

Bayes Theorem

Bayes Theorem:

$$p(\mathbf{A} \mid \mathbf{B}) = \frac{p(\mathbf{B} \mid \mathbf{A})\, p(\mathbf{A})}{p(\mathbf{B})}$$

Bayes Theorem

Bayes Theorem:

$$p(\mathbf{A} \mid \mathbf{B}) = \frac{p(\mathbf{B} \mid \mathbf{A})\, p(\mathbf{A})}{p(\mathbf{B})}$$

Bayesian Inference:

$$p(\boldsymbol{\theta}, \mathcal{T}, \psi \mid \mathbf{Y}, \mathbf{S}) = \frac{p(\mathbf{Y}, \mathbf{S} \mid \boldsymbol{\theta}, \mathcal{T}, \psi)\, p(\boldsymbol{\theta}, \mathcal{T}, \psi)}{p(\mathbf{Y}, \mathbf{S})}$$

Bayes Theorem

Bayes Theorem:

$$p\,(\mathbf{A} \mid \mathbf{B}) = \frac{p\,(\mathbf{B} \mid \mathbf{A})\,p\,(\mathbf{A})}{p\,(\mathbf{B})}$$

Bayesian Inference:

$$p\,(\boldsymbol{\theta}, \mathcal{T}, \psi \mid \mathbf{Y}, \mathbf{S}) = \frac{p\,(\mathbf{Y}, \mathbf{S} \mid \boldsymbol{\theta}, \mathcal{T}, \psi)\,p\,(\boldsymbol{\theta}, \mathcal{T}, \psi)}{p\,(\mathbf{Y}, \mathbf{S})}$$

- $p\,(\mathbf{Y}, \mathbf{S})$ does not depend on the parameters (constant).

Bayes Theorem

Bayes Theorem:

$$p(\mathbf{A} \mid \mathbf{B}) = \frac{p(\mathbf{B} \mid \mathbf{A})\, p(\mathbf{A})}{p(\mathbf{B})}$$

Bayesian Inference:

$$p(\boldsymbol{\theta}, \mathcal{T}, \psi \mid \mathbf{Y}, \mathbf{S}) = \frac{p(\mathbf{Y}, \mathbf{S} \mid \boldsymbol{\theta}, \mathcal{T}, \psi)\, p(\boldsymbol{\theta}, \mathcal{T}, \psi)}{p(\mathbf{Y}, \mathbf{S})}$$

- $p(\mathbf{Y}, \mathbf{S})$ does not depend on the parameters (constant).
- $p(\mathbf{Y}, \mathbf{S}) = \int_{\boldsymbol{\theta}, \mathcal{T}, \psi} p(\mathbf{Y}, \mathbf{S} \mid \boldsymbol{\theta}, \mathcal{T}, \psi)\, p(\boldsymbol{\theta}, \mathcal{T}, \psi)$
  marginal likelihood, very hard to compute.

Bayes Theorem

Bayes Theorem:

$$p(\mathbf{A} \mid \mathbf{B}) = \frac{p(\mathbf{B} \mid \mathbf{A})\, p(\mathbf{A})}{p(\mathbf{B})}$$

Bayesian Inference:

$$p(\boldsymbol{\theta}, \mathcal{T}, \psi \mid \mathbf{Y}, \mathbf{S}) = \frac{p(\mathbf{Y}, \mathbf{S} \mid \boldsymbol{\theta}, \mathcal{T}, \psi)\, p(\boldsymbol{\theta}, \mathcal{T}, \psi)}{p(\mathbf{Y}, \mathbf{S})}$$

- $p(\mathbf{Y}, \mathbf{S})$ does not depend on the parameters (constant).
- $p(\mathbf{Y}, \mathbf{S}) = \int_{\boldsymbol{\theta}, \mathcal{T}, \psi} p(\mathbf{Y}, \mathbf{S} \mid \boldsymbol{\theta}, \mathcal{T}, \psi)\, p(\boldsymbol{\theta}, \mathcal{T}, \psi)$ marginal likelihood, very hard to compute.

We cannot compute the posterior, but we can sample from it.

Bayes Inference by Sampling

Assume we can sample from the posterior:

$$(\boldsymbol{\theta}_i, \mathcal{T}_i, \boldsymbol{\psi}_i) \sim p\left(\boldsymbol{\theta}, \mathcal{T}, \boldsymbol{\psi} \mid \mathbf{Y}, \mathbf{S}\right) \quad \text{iid}$$

Bayes Inference by Sampling

Assume we can sample from the posterior:

$$(\boldsymbol{\theta}_i, \mathcal{T}_i, \boldsymbol{\psi}_i) \sim p\left(\boldsymbol{\theta}, \mathcal{T}, \boldsymbol{\psi} \mid \mathbf{Y}, \mathbf{S}\right) \quad \text{iid}$$

We can get some moments estimates.
E.g. for continuous parameters:

$$\mathbb{E}\left[\boldsymbol{\theta} \mid \mathbf{Y}, \mathbf{S}\right] \approx \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\theta}_i$$

$$\mathbb{V}\mathrm{ar}\left[\boldsymbol{\theta} \mid \mathbf{Y}, \mathbf{S}\right] \approx \frac{1}{n-1} \sum_{i=1}^{n} \left(\boldsymbol{\theta}_i - \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\theta}_i\right)^2$$

Bayes Inference by Sampling

Assume we can sample from the posterior:

$$(\boldsymbol{\theta}_i, \mathcal{T}_i, \boldsymbol{\psi}_i) \sim p\left(\boldsymbol{\theta}, \mathcal{T}, \boldsymbol{\psi} \mid \mathbf{Y}, \mathbf{S}\right) \quad \text{iid}$$

We can get some moments estimates.
E.g. for continuous parameters:

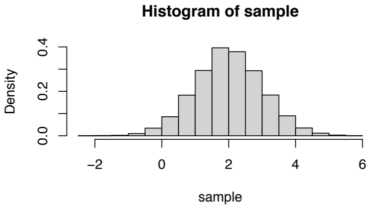
$$\mathbb{E}\left[\boldsymbol{\theta} \mid \mathbf{Y}, \mathbf{S}\right] \approx \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\theta}_i$$

$$\mathbb{V}\text{ar}\left[\boldsymbol{\theta} \mid \mathbf{Y}, \mathbf{S}\right] \approx \frac{1}{n-1} \sum_{i=1}^{n} \left(\boldsymbol{\theta}_i - \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\theta}_i\right)^2$$

We can estimate the parameters distribution (histogram).

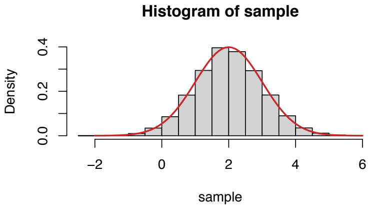
Bayes Inference by Sampling

```
set.seed(18300730)
# n=1000 sample from a mystery distribution
sample <- rmystery(10000)
# Histogram
hist(sample, probability = TRUE)
```



Bayes Inference by Sampling

```
set.seed(18300730)
# n=1000 sample from a mystery distribution
sample <- rmystery(10000)
# Histogram
hist(sample, probability = TRUE)
# distribution
x <- seq(-2, 6, 0.1)
lines(x, dnorm(x, 2, 1), lwd = 2, col = "firebrick3")
```



Bayesian Inference

Goal: Sample from the posterior

$$p\left(\boldsymbol{\theta}, \mathcal{T}, \boldsymbol{\psi} \mid \mathbf{Y}, \mathbf{S}\right) = \frac{p\left(\mathbf{Y}, \mathbf{S} \mid \boldsymbol{\theta}, \mathcal{T}, \boldsymbol{\psi}\right) p\left(\boldsymbol{\theta}, \mathcal{T}, \boldsymbol{\psi}\right)}{p\left(\mathbf{Y}, \mathbf{S}\right)}$$

That we know only up to a constant.

Bayesian Phylogenetics

Goal:

$$p\left(\boldsymbol{\theta}, \mathcal{T}, \boldsymbol{\psi} \mid \mathbf{Y}, \mathbf{S}\right)$$

Bayesian Phylogenetics

Goal:

$$p\left(\boldsymbol{\theta}, \mathcal{T}, \boldsymbol{\psi} \mid \mathbf{Y}, \mathbf{S}\right) \propto p\left(\mathbf{Y}, \mathbf{S} \mid \boldsymbol{\theta}, \mathcal{T}, \boldsymbol{\psi}\right) p\left(\boldsymbol{\theta}, \mathcal{T}, \boldsymbol{\psi}\right)$$

Bayesian Phylogenetics

Goal:

$$\begin{aligned} p(\boldsymbol{\theta}, \mathcal{T}, \boldsymbol{\psi} \mid \mathbf{Y}, \mathbf{S}) &\propto p(\mathbf{Y}, \mathbf{S} \mid \boldsymbol{\theta}, \mathcal{T}, \boldsymbol{\psi})\, p(\boldsymbol{\theta}, \mathcal{T}, \boldsymbol{\psi}) \\ &\propto p(\mathbf{Y} \mid \boldsymbol{\theta}, \mathcal{T})\, p(\mathbf{S} \mid \mathcal{T}, \boldsymbol{\psi})\, p(\boldsymbol{\theta}, \mathcal{T}, \boldsymbol{\psi}) \end{aligned}$$

Assumption: $\mathbf{Y}$ and $\mathbf{S}$ independent conditionally on $\mathcal{T}$.

Bayesian Phylogenetics

Goal:

$$
\begin{aligned}
p\left(\boldsymbol{\theta}, \mathcal{T}, \psi \mid \mathbf{Y}, \mathbf{S}\right) &\propto p\left(\mathbf{Y}, \mathbf{S} \mid \boldsymbol{\theta}, \mathcal{T}, \boldsymbol{\psi}\right) p\left(\boldsymbol{\theta}, \mathcal{T}, \psi\right) \\
&\propto p\left(\mathbf{Y} \mid \boldsymbol{\theta}, \mathcal{T}\right) p\left(\mathbf{S} \mid \mathcal{T}, \boldsymbol{\psi}\right) p\left(\boldsymbol{\theta}, \mathcal{T}, \boldsymbol{\psi}\right) \\
&\propto p\left(\mathbf{Y} \mid \boldsymbol{\theta}, \mathcal{T}\right) p\left(\boldsymbol{\theta}\right) p\left(\mathbf{S} \mid \mathcal{T}, \psi\right) p\left(\mathcal{T}, \psi\right)
\end{aligned}
$$

Assumption: **Y** and **S** independent conditionally on $\mathcal{T}$.

For now: $\mathcal{T}$ fixed.

Markov Chain Monte Carlo (Larget, 2005)

Goal: Sample from $p(\boldsymbol{\theta} \mid \mathbf{Y}) \propto p(\mathbf{Y} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})$

Markov Chain Monte Carlo (Larget, 2005)

Goal: Sample from $p(\boldsymbol{\theta} \mid \mathbf{Y}) \propto p(\mathbf{Y} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})$

Problem: We do not know the normalizing constant.

Markov Chain Monte Carlo (Larget, 2005)

Goal: Sample from $p(\boldsymbol{\theta} \mid \mathbf{Y}) \propto p(\mathbf{Y} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})$

Problem: We do not know the normalizing constant.

Metropolis - Hasting: Iterate:

- Draw $\boldsymbol{\theta}^*$ in $q\left(\boldsymbol{\theta} \,\middle|\, \boldsymbol{\theta}^t\right)$.
- Set $\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^*$ with probability:

$$r_t = \min\left\{1, \frac{p\left(\mathbf{Y} \mid \boldsymbol{\theta}^*\right)}{p\left(\mathbf{Y} \,\middle|\, \boldsymbol{\theta}^t\right)} \frac{p\left(\boldsymbol{\theta}^*\right)}{p\left(\boldsymbol{\theta}^t\right)} \frac{q(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^*)}{q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(t)})}\right\}.$$

Markov Chain Monte Carlo (Larget, 2005)

Goal: Sample from $p(\boldsymbol{\theta} \mid \mathbf{Y}) \propto p(\mathbf{Y} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})$

Problem: We do not know the normalizing constant.

Metropolis - Hasting: Iterate:

- Draw $\boldsymbol{\theta}^*$ in $q\left(\boldsymbol{\theta} \mid \boldsymbol{\theta}^t\right)$.
- Set $\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^*$ with probability:

$$r_t = \min\left\{1, \frac{p(\mathbf{Y} \mid \boldsymbol{\theta}^*)}{p\left(\mathbf{Y} \mid \boldsymbol{\theta}^t\right)} \frac{p(\boldsymbol{\theta}^*)}{p\left(\boldsymbol{\theta}^t\right)} \frac{q(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{\theta}^*)}{q(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^{(t)})}\right\}.$$

- The $\boldsymbol{\theta}^{(t)}$ is a Markov Chain.
- Its stationary distribution is $p(\boldsymbol{\theta} \mid \mathbf{Y})$.
  $\rightarrow$ For "t large enough", $\boldsymbol{\theta}^{(t)} \sim p(\boldsymbol{\theta} \mid \mathbf{Y})$.

Markov Chain Monte Carlo (Larget, 2005)

Goal: Sample from $p\left(\boldsymbol{\theta} \mid \mathbf{Y}\right) \propto p\left(\mathbf{Y} \mid \boldsymbol{\theta}\right) p\left(\boldsymbol{\theta}\right)$

Problem: We do not know the normalizing constant.

Metropolis - Hasting: Iterate:

- Draw $\boldsymbol{\theta}^*$ in $q\left(\boldsymbol{\theta} \;\middle|\; \boldsymbol{\theta}^t\right)$.
- Set $\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^*$ with probability:

$$r_t = \min\left\{1, \frac{p\left(\mathbf{Y} \mid \boldsymbol{\theta}^*\right)}{p\left(\mathbf{Y} \;\middle|\; \boldsymbol{\theta}^t\right)} \frac{p\left(\boldsymbol{\theta}^*\right)}{p\left(\boldsymbol{\theta}^t\right)} \frac{q(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^*)}{q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(t)})}\right\}.$$

Animation: Chi Feng's website

Markov Chain Monte Carlo (Larget, 2005)

Goal: Sample from $p\left(\boldsymbol{\theta} \mid \mathbf{Y}\right) \propto p\left(\mathbf{Y} \mid \boldsymbol{\theta}\right) p\left(\boldsymbol{\theta}\right)$

Problem: We do not know the normalizing constant.

Metropolis - Hasting: Iterate:

- Draw $\boldsymbol{\theta}^*$ in $q\left(\boldsymbol{\theta} \mid \boldsymbol{\theta}^t\right)$.
- Set $\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^*$ with probability:

$$r_t = \min\left\{1, \frac{p\left(\mathbf{Y} \mid \boldsymbol{\theta}^*\right)}{p\left(\mathbf{Y} \mid \boldsymbol{\theta}^t\right)} \frac{p\left(\boldsymbol{\theta}^*\right)}{p\left(\boldsymbol{\theta}^t\right)} \frac{q(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{\theta}^*)}{q(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^{(t)})}\right\}.$$

Gibbs: Rejection free.

- Split $\boldsymbol{\theta} = (\boldsymbol{\theta}_{[1]}, \dots, \boldsymbol{\theta}_{[K]})$.
- Draw $\boldsymbol{\theta}^*$ in $p(\boldsymbol{\theta}_{[k]} \mid \boldsymbol{\theta}^{(t)}_{[-k]}, \mathbf{Y})$ so that $r_t = 1$.

BM: Gibbs with Conjugate Priors

Likelihood:

$$\mathbf{Y}|\mathbf{R}, \boldsymbol{\mu} \sim \mathcal{MN}(\mathbf{1}_n \boldsymbol{\mu}^T, \mathbf{V}, \mathbf{R})$$

BM: Gibbs with Conjugate Priors

Likelihood:

$$\mathbf{Y}|\mathbf{R}, \boldsymbol{\mu} \sim \mathcal{MN}(\mathbf{1}_n\boldsymbol{\mu}^T, \mathbf{V}, \mathbf{R})$$

Conjugate Priors:

$$\mathbf{R} \sim \mathcal{IW}(\mathbf{R}_0, \nu)$$
$$\boldsymbol{\mu}|\mathbf{R} \sim \mathcal{N}(\boldsymbol{\mu}_0, \kappa_0^{-1}\mathbf{R})$$

BM: Gibbs with Conjugate Priors

Likelihood:

$$\mathbf{Y}|\mathbf{R}, \boldsymbol{\mu} \sim \mathcal{MN}(\mathbf{1}_n \boldsymbol{\mu}^T, \mathbf{V}, \mathbf{R})$$

Conjugate Priors:

$$\mathbf{R} \sim \mathcal{IW}(\mathbf{R}_0, \nu)$$
$$\boldsymbol{\mu}|\mathbf{R} \sim \mathcal{N}(\boldsymbol{\mu}_0, \kappa_0^{-1}\mathbf{R})$$

Gibbs:

$$\mathbf{R}|\mathbf{Y}, \boldsymbol{\mu} \sim \mathcal{IW}(\mathbf{R}_n, \nu_n) \quad \text{with} \quad \mathbf{R}_n = f(\mathbf{Y}, \boldsymbol{\mu}, \mathbf{V})$$

BM: Gibbs with Conjugate Priors

Likelihood:

$$\mathbf{Y}|\mathbf{R}, \boldsymbol{\mu} \sim \mathcal{MN}(\mathbf{1}_n \boldsymbol{\mu}^T, \mathbf{V}, \mathbf{R})$$

Conjugate Priors:

$$\mathbf{R} \sim \mathcal{IW}(\mathbf{R}_0, \nu)$$
$$\boldsymbol{\mu}|\mathbf{R} \sim \mathcal{N}(\boldsymbol{\mu}_0, \kappa_0^{-1}\mathbf{R})$$

Gibbs:

$$\mathbf{R}|\mathbf{Y}, \boldsymbol{\mu} \sim \mathcal{IW}(\mathbf{R}_n, \nu_n) \quad \text{with} \quad \mathbf{R}_n = f(\mathbf{Y}, \boldsymbol{\mu}, \mathbf{V})$$

$\hookrightarrow$ Automatic sampling in the space of variance matrices.

Wishart Distribution

Wishart: generalization of the gamma distribution to matrices.

$$\mathbf{P} \sim \mathcal{W}_p(\mathbf{P}_0, \nu), \quad \nu \geq p$$

Wishart Distribution

Wishart: generalization of the gamma distribution to matrices.

$$\mathbf{P} \sim \mathcal{W}_p(\mathbf{P}_0, \nu), \quad \nu \geq p$$

Moments:

$$\mathbb{E}\left[\mathbf{P}\right] = \nu \mathbf{P}_0 \qquad \mathbb{V}\mathrm{ar}\left[\mathbf{P}_{ij}\right] = \nu[(P_0)_{ij}^2 + (P_0)_{ii}(P_0)_{jj}]$$

Wishart Distribution

Wishart: generalization of the gamma distribution to matrices.

$$\mathbf{P} \sim \mathcal{W}_p(\mathbf{P}_0, \nu), \quad \nu \geq p$$

Moments:

$$\mathbb{E}\left[\mathbf{P}\right] = \nu \mathbf{P}_0 \qquad \mathbb{V}\mathrm{ar}\left[\mathbf{P}_{ij}\right] = \nu[(P_0)_{ij}^2 + (P_0)_{ii}(P_0)_{jj}]$$

Sample Variance: If $\mathbf{Y}_i \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ are n iid vectors of size p, then:

$$\hat{\boldsymbol{\Sigma}} = \sum_{i=1}^{n} (\mathbf{Y}_i - \boldsymbol{\mu})(\mathbf{Y}_i - \boldsymbol{\mu})^T \sim \mathcal{W}_p(\boldsymbol{\Sigma}, n)$$

Wishart Distribution

Wishart: generalization of the gamma distribution to matrices.

$$\mathbf{P} \sim \mathcal{W}_p(\mathbf{P}_0, \nu), \quad \nu \geq p$$

Moments:

$$\mathbb{E}\left[\mathbf{P}\right] = \nu \mathbf{P}_0 \qquad \mathbb{V}\mathrm{ar}\left[\mathbf{P}_{ij}\right] = \nu[(P_0)_{ij}^2 + (P_0)_{ii}(P_0)_{jj}]$$

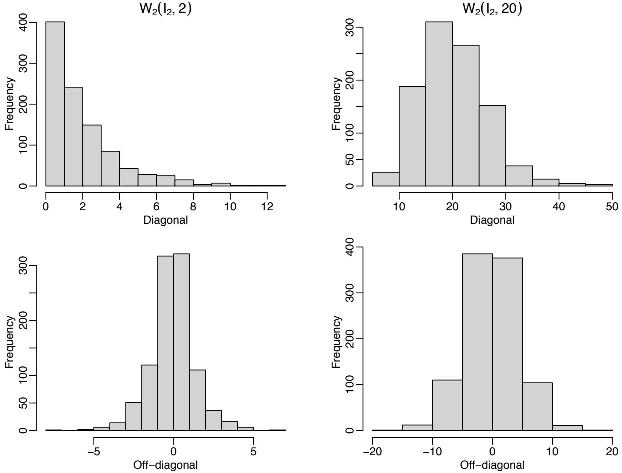
Sample Variance: If $\mathbf{Y}_i \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ are n iid vectors of size p, then:

$$\hat{\boldsymbol{\Sigma}} = \sum_{i=1}^{n} (\mathbf{Y}_i - \boldsymbol{\mu})(\mathbf{Y}_i - \boldsymbol{\mu})^T \sim \mathcal{W}_p(\boldsymbol{\Sigma}, n)$$

Inverse Wishart:

$$\mathbf{R} \sim \mathcal{IW}(\mathbf{R}_0, \nu) \iff \mathbf{R}^{-1} \sim \mathcal{W}_p(\mathbf{R}_0^{-1}, \nu), \quad \nu > p - 1$$

Wishart Distribution

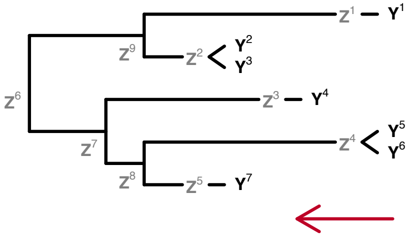


Diagonal and off-diagonal elements are related.

LKJ

Outline

General Gaussian Model: Likelihood



$$\mathbf{Z}^r \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Gamma}) \qquad \text{root}$$

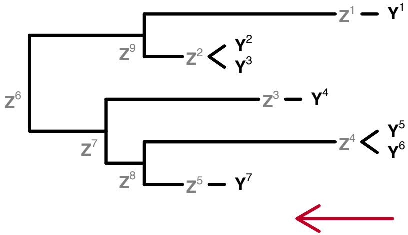
$$\mathbf{Z}^j \mid \mathbf{Z}^{\text{pa}(j)} \sim \mathcal{N}\left(\mathbf{q}_j \mathbf{Z}^{\text{pa}(j)} + \mathbf{r}_j,\ \boldsymbol{\Sigma}_j\right) \qquad \text{nodes}$$

$$\mathbf{Y}^i \mid \mathbf{Z}^{\text{pa}(i)} \sim \mathcal{N}\left(\mathbf{Z}^{\text{pa}(i)}, \mathbf{S}_i\right) \qquad \text{obs}$$

Likelihood: $\log p(\mathbf{Y})$ in one post-order traversal.

$\hookrightarrow O(N)$.

General Gaussian Model: Likelihood



$$\mathbf{Z}^r \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Gamma}) \quad \text{root}$$

$$\mathbf{Z}^j \mid \mathbf{Z}^{\mathrm{pa}(j)} \sim \mathcal{N}\left(\mathbf{q}_j \mathbf{Z}^{\mathrm{pa}(j)} + \mathbf{r}_j,\ \boldsymbol{\Sigma}_j\right) \quad \text{nodes}$$

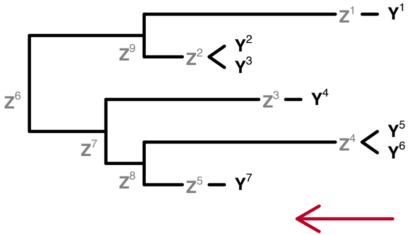
$$\mathbf{Y}^i \mid \mathbf{Z}^{\mathrm{pa}(i)} \sim \mathcal{N}\left(\mathbf{Z}^{\mathrm{pa}(i)}, \mathbf{S}_i\right) \quad \text{obs}$$

Likelihood: $\log p(\mathbf{Y})$ in one post-order traversal.

$\hookrightarrow$ $O(N)$.

$\hookrightarrow$ “Pruning”, “Gaussian elimination”, “Phylogenetic Kalman filter”, ...

General Gaussian Model: Likelihood



$$\mathbf{Z}^r \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Gamma}) \quad \text{root}$$

$$\mathbf{Z}^j \,\middle|\, \mathbf{Z}^{\mathrm{pa}(j)} \sim \mathcal{N}\left(\mathbf{q}_j \mathbf{Z}^{\mathrm{pa}(j)} + \mathbf{r}_j,\ \boldsymbol{\Sigma}_j\right) \quad \text{nodes}$$

$$\mathbf{Y}^i \,\middle|\, \mathbf{Z}^{\mathrm{pa}(i)} \sim \mathcal{N}\left(\mathbf{Z}^{\mathrm{pa}(i)}, \mathbf{S}_i\right) \quad \text{obs}$$

Likelihood: $\log p(\mathbf{Y})$ in one post-order traversal.

$\hookrightarrow$ $O(N)$.

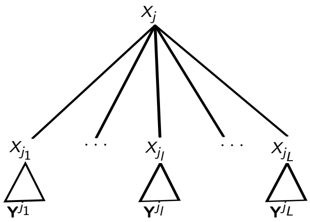
$\hookrightarrow$ "Pruning", "Gaussian elimination", "Phylogenetic Kalman filter", ...

Difficulty: Numerical robustness.

Likelihood Computation

Likelihood function:

$$f_{\mathbf{Y}^j|X_j}(\mathbf{Y}^j; a) = A_j(\mathbf{Y}^j)\Phi_{M_j(\mathbf{Y}^j), S_j^2(\mathbf{Y}^j)}(a)$$



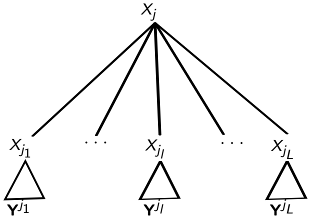
Likelihood Computation

Likelihood function:

$$f_{\mathbf{Y}^j|X_j}(\mathbf{Y}^j; a) = A_j(\mathbf{Y}^j)\Phi_{M_j(\mathbf{Y}^j), S_j^2(\mathbf{Y}^j)}(a)$$

Initialization: for tips

$$f_{Y_i|Y_i}(Y_i; a) = \Phi_{Y_i,0}(a)$$



Likelihood Computation

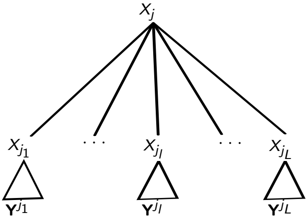
Likelihood function:

$$f_{\mathbf{Y}^j | X_j}(\mathbf{Y}^j; a) = A_j(\mathbf{Y}^j)\Phi_{M_j(\mathbf{Y}^j), S_j^2(\mathbf{Y}^j)}(a)$$

Initialization: for tips

$$f_{Y_i | Y_i}(Y_i; a) = \Phi_{Y_i, 0}(a)$$

Propagation:

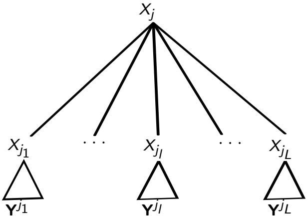


$$f_{\mathbf{Y}^j | X_j}(\mathbf{Y}^j; a) = \prod_{l=1}^{L} f_{\mathbf{Y}^{j_l} | X_j}(\mathbf{Y}^{j_l}; a) \quad \text{conditional independence}$$

Likelihood Computation

Likelihood function:

$$f_{\mathbf{Y}^j|X_j}(\mathbf{Y}^j; a) = A_j(\mathbf{Y}^j)\Phi_{M_j(\mathbf{Y}^j), S_j^2(\mathbf{Y}^j)}(a)$$



Initialization: for tips

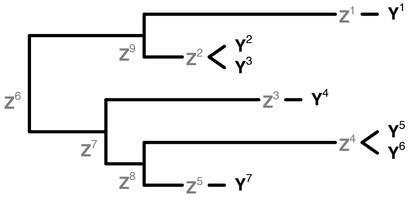
$$f_{Y_i|Y_i}(Y_i; a) = \Phi_{Y_i,0}(a)$$

Propagation:

$$f_{\mathbf{Y}^j|X_j}(\mathbf{Y}^j; a) = \prod_{l=1}^{L} f_{\mathbf{Y}^{j_l}|X_j}(\mathbf{Y}^{j_l}; a) \quad \text{conditional independence}$$

$$f_{\mathbf{Y}^{j_l}|X_j}(\mathbf{Y}^{j_l}; a) = \int_{\mathbb{R}} f_{\mathbf{Y}^{j_l}|X_{j_l}}(\mathbf{Y}^{j_l}; b) f_{X_{j_l}|X_j}(b; a)db \quad \text{explicit (Gaussian)}$$

Ancestral State Reconstruction



$$
\begin{aligned}
\mathbf{Z}^r &\sim \mathcal{N}\left(\boldsymbol{\mu}, \boldsymbol{\Gamma}\right) && \text{root} \\
\mathbf{Z}^j \mid \mathbf{Z}^{\mathrm{pa}(j)} &\sim \mathcal{N}\left(\mathbf{q}_j \mathbf{Z}^{\mathrm{pa}(j)} + \mathbf{r}_j,\ \boldsymbol{\Sigma}_j\right) && \text{nodes} \\
\mathbf{Y}^i \mid \mathbf{Z}^{\mathrm{pa}(i)} &\sim \mathcal{N}\left(\mathbf{Z}^{\mathrm{pa}(i)}, \mathbf{S}_i\right) && \text{obs}
\end{aligned}
$$

Ancestral State Reconstruction: compute $\mathbb{E}\left[\mathbf{Z} \mid \mathbf{Y}\right]$

Ancestral State Reconstruction: Naive Gaussian

$$\begin{pmatrix}\mathbf{Z}\\ \mathbf{Y}\end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix}\mu_{\mathbf{Z}}\\ \mu_{\mathbf{Y}}\end{pmatrix};\ \begin{pmatrix}\mathbf{V}_{\mathbf{ZZ}} & \mathbf{V}_{\mathbf{ZY}}\\ \mathbf{V}_{\mathbf{YZ}} & \mathbf{V}_{\mathbf{YY}}\end{pmatrix}\right)$$

Ancestral State Reconstruction: Naive Gaussian

$$\begin{pmatrix} \mathbf{Z} \\ \mathbf{Y} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_{\mathbf{Z}} \\ \mu_{\mathbf{Y}} \end{pmatrix};\ \begin{pmatrix} \mathbf{V}_{\mathbf{ZZ}} & \mathbf{V}_{\mathbf{ZY}} \\ \mathbf{V}_{\mathbf{YZ}} & \mathbf{V}_{\mathbf{YY}} \end{pmatrix}\right)$$

$$\mathbf{Z} \mid \mathbf{Y} \sim \mathcal{N}\left(\mu_{\mathbf{Z}} + \mathbf{V}_{\mathbf{ZY}}\mathbf{V}_{\mathbf{YY}}^{-1}\mu_{\mathbf{Y}};\ \mathbf{V}_{\mathbf{ZZ}} - \mathbf{V}_{\mathbf{ZY}}\mathbf{V}_{\mathbf{YY}}^{-1}\mathbf{V}_{\mathbf{YZ}}\right)$$

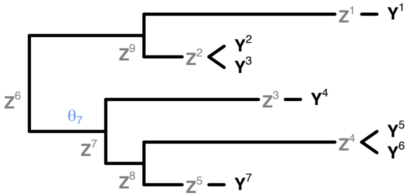
Ancestral State Reconstruction: Naive Gaussian

$$\begin{pmatrix} \mathbf{Z} \\ \mathbf{Y} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_{\mathbf{Z}} \\ \mu_{\mathbf{Y}} \end{pmatrix};\ \begin{pmatrix} \mathbf{V}_{\mathbf{ZZ}} & \mathbf{V}_{\mathbf{ZY}} \\ \mathbf{V}_{\mathbf{YZ}} & \mathbf{V}_{\mathbf{YY}} \end{pmatrix}\right)$$

$$\mathbf{Z} \mid \mathbf{Y} \sim \mathcal{N}\left(\mu_{\mathbf{Z}} + \mathbf{V}_{\mathbf{ZY}}\mathbf{V}_{\mathbf{YY}}^{-1}\mu_{\mathbf{Y}};\ \mathbf{V}_{\mathbf{ZZ}} - \mathbf{V}_{\mathbf{ZY}}\mathbf{V}_{\mathbf{YY}}^{-1}\mathbf{V}_{\mathbf{YZ}}\right)$$

$\rightarrow$ requires a matrix inversion $O(n^3)$.

Ancestral State Reconstruction: Efficient Computations



$$\mathbf{Z}^r \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Gamma}) \qquad \text{root}$$

$$\mathbf{Z}^j \mid \mathbf{Z}^{\mathrm{pa}(j)} \sim \mathcal{N}\left(\mathbf{q}_j \mathbf{Z}^{\mathrm{pa}(j)} + \mathbf{r}_j,\ \boldsymbol{\Sigma}_j\right) \qquad \text{nodes}$$

$$\mathbf{Y}^i \mid \mathbf{Z}^{\mathrm{pa}(i)} \sim \mathcal{N}\left(\mathbf{Z}^{\mathrm{pa}(i)}, \mathbf{S}_i\right) \qquad \text{obs}$$

Conditional Moments: in one pre-order traversal.

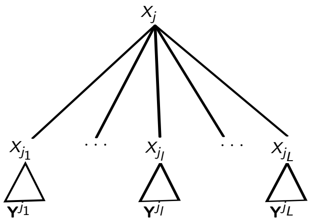
$\hookrightarrow$ $\mathbb{E}[\mathbf{Z}_j \mid \mathbf{Y}]$, $\mathbb{V}\mathrm{ar}[\mathbf{Z}_j \mid \mathbf{Y}]$.

$\hookrightarrow$ $O(N)$.

Conditional Moments Computation

Conditional Likelihood function:

$$f_{X_j \mid \mathbf{Y}}(a; \mathbf{Y}) \propto \Phi_{N_j, T_j^2}(a)$$



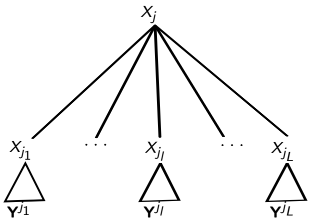
Conditional Moments Computation

Conditional Likelihood function:

$$f_{X_j|\mathbf{Y}}(a;\mathbf{Y}) \propto \Phi_{N_j, T_j^2}(a)$$

Initialization: for the root

$$f_{X_r|\mathbf{Y}}(a;\mathbf{Y}) \propto f_{\mathbf{Y}|X_r}(\mathbf{Y};a) f_{X_r}(a)$$



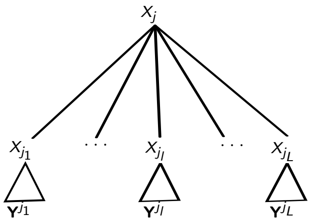
Conditional Moments Computation

Conditional Likelihood function:

$$f_{X_j|\mathbf{Y}}(a; \mathbf{Y}) \propto \Phi_{N_j, T_j^2}(a)$$

Initialization: for the root

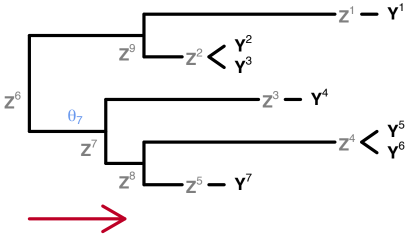
$$f_{X_r|\mathbf{Y}}(a; \mathbf{Y}) \propto f_{\mathbf{Y}|X_r}(\mathbf{Y}; a) f_{X_r}(a)$$



Propagation:

$$f_{X_j|\mathbf{Y}}(a; \mathbf{Y}) = \int_{\mathbb{R}} f_{X_j|X_{pa(j)}}(a; b) f_{X_{\mathrm{pa}(j)}|\mathbf{Y}}(b; \mathbf{Y}) db \quad \text{Gaussians}$$

Efficient Computations: Gradient



$$\mathbf{Z}^r \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Gamma}) \quad \text{root}$$

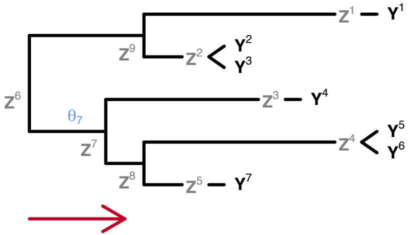
$$\mathbf{Z}^j \mid \mathbf{Z}^{\mathrm{pa}(j)} \sim \mathcal{N}\left(\mathbf{q}_j \mathbf{Z}^{\mathrm{pa}(j)} + \mathbf{r}_j,\ \boldsymbol{\Sigma}_j\right) \quad \text{nodes}$$

$$\mathbf{Y}^i \mid \mathbf{Z}^{\mathrm{pa}(i)} \sim \mathcal{N}\left(\mathbf{Z}^{\mathrm{pa}(i)}, \mathbf{S}_i\right) \quad \text{obs}$$

Branch-specific Gradient: in one pre-order traversal.

$$\begin{aligned} \nabla_{\boldsymbol{\theta}_k}\left[\log p\left(\mathbf{Y} \mid \boldsymbol{\theta}_k\right)\right] &= \mathbb{E}\left[\nabla_{\boldsymbol{\theta}_k}\left[\log p\left(\mathbf{Z}^k, \mathbf{Y} \mid \boldsymbol{\theta}_k\right)\right] \mid \mathbf{Y}\right] \\ &= \mathbb{E}\left[\nabla_{\boldsymbol{\theta}_k}\left[\log p\left(\mathbf{Z}^k \mid \mathbf{Y}_{\lceil k \rceil}, \boldsymbol{\theta}_k\right)\right] \mid \mathbf{Y}\right]. \end{aligned}$$

Efficient Computations: Gradient



$$\mathbf{Z}^r \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Gamma}) \qquad \text{root}$$

$$\mathbf{Z}^j \,\middle|\, \mathbf{Z}^{\mathrm{pa}(j)} \sim \mathcal{N}\left(\mathbf{q}_j \mathbf{Z}^{\mathrm{pa}(j)} + \mathbf{r}_j,\ \boldsymbol{\Sigma}_j\right) \qquad \text{nodes}$$

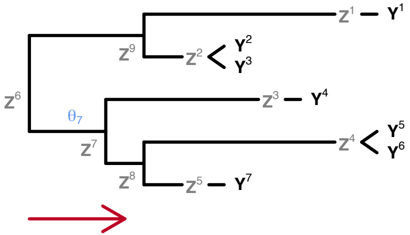
$$\mathbf{Y}^i \,\middle|\, \mathbf{Z}^{\mathrm{pa}(i)} \sim \mathcal{N}\left(\mathbf{Z}^{\mathrm{pa}(i)}, \mathbf{S}_i\right) \qquad \text{obs}$$

Branch-specific Gradient: in one pre-order traversal.

$$\begin{aligned}\nabla_{\boldsymbol{\theta}_k}\left[\log p\left(\mathbf{Y} \mid \boldsymbol{\theta}_k\right)\right] &= \mathbb{E}\left[\nabla_{\boldsymbol{\theta}_k}\left[\log p\left(\mathbf{Z}^k, \mathbf{Y} \mid \boldsymbol{\theta}_k\right)\right] \mid \mathbf{Y}\right] \\ &= \mathbb{E}\left[\nabla_{\boldsymbol{\theta}_k}\left[\log p\left(\mathbf{Z}^k \mid \mathbf{Y}_{\lceil k \rceil}, \boldsymbol{\theta}_k\right)\right] \mid \mathbf{Y}\right].\end{aligned}$$

Chain rule: Gradient w.r.t. any parameter in $O(N)$.

Efficient Computations: Gradient



$$
\begin{aligned}
\mathbf{Z}^r &\sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Gamma}) && \text{root}\\
\mathbf{Z}^j \mid \mathbf{Z}^{\mathrm{pa}(j)} &\sim \mathcal{N}\left(\mathbf{q}_j \mathbf{Z}^{\mathrm{pa}(j)} + \mathbf{r}_j,\ \boldsymbol{\Sigma}_j\right) && \text{nodes}\\
\mathbf{Y}^i \mid \mathbf{Z}^{\mathrm{pa}(i)} &\sim \mathcal{N}\left(\mathbf{Z}^{\mathrm{pa}(i)}, \mathbf{S}_i\right) && \text{obs}
\end{aligned}
$$

Branch-specific Gradient: in one pre-order traversal.

$$
\begin{aligned}
\nabla_{\boldsymbol{\theta}_k}\left[\log p\left(\mathbf{Y} \mid \boldsymbol{\theta}_k\right)\right] &= \mathbb{E}\left[\nabla_{\boldsymbol{\theta}_k}\left[\log p\left(\mathbf{Z}^k, \mathbf{Y} \mid \boldsymbol{\theta}_k\right)\right] \mid \mathbf{Y}\right]\\
&= \mathbb{E}\left[\nabla_{\boldsymbol{\theta}_k}\left[\log p\left(\mathbf{Z}^k \mid \mathbf{Y}_{\lceil k \rceil}, \boldsymbol{\theta}_k\right)\right] \mid \mathbf{Y}\right].
\end{aligned}
$$

Chain rule: Gradient w.r.t. any parameter in $O(N)$. $\rightarrow$ HMC

Outline

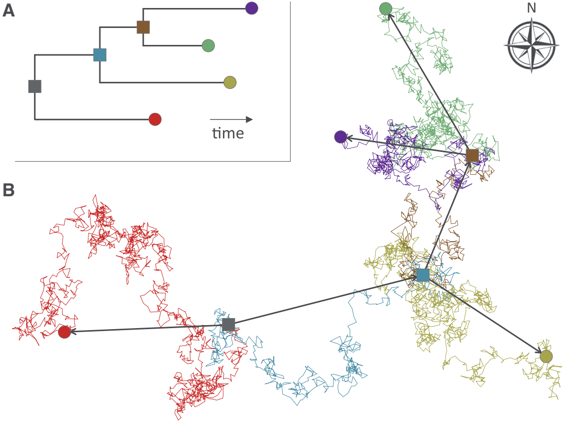
1. Bayesian for Continuous Traits
2. Pruning Algorithms
3. Phylogeography
    - Phylogeography
    - Relaxed Random Walk

Phylogeography

Trait: Geographical coordinates: latitude and longitude.

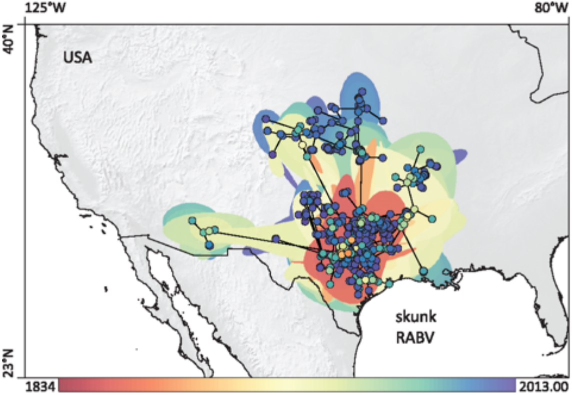
Phylogeography

Trait: Geographical coordinates: latitude and longitude.



Dellicour et al. (2021)

Phylogeography



Dellicour et al. (2017)

Spread on an heterogeneous geographical region.

Relaxed Random Walk (Lemey et al., 2010)

Model: BM with varying ("relaxed") variance.

$$\mathbf{Z}^j \,\Big|\, \mathbf{Z}^{\mathrm{pa}(j)} \sim \mathcal{N}\left(\mathbf{Z}^{\mathrm{pa}(j)}, \phi_j \mathbf{R}\right)$$

Idea: Each branch has its own propagation speed.

$\rightarrow$ Allow for varying speed of propagation.

Relaxed Random Walk (Lemey et al., 2010)

Model: BM with varying ("relaxed") variance.

$$\mathbf{Z}^j \,\Big|\, \mathbf{Z}^{\mathrm{pa}(j)} \sim \mathcal{N}\left(\mathbf{Z}^{\mathrm{pa}(j)}, \phi_j \mathbf{R}\right)$$

Idea: Each branch has its own propagation speed.

$\rightarrow$ Allow for varying speed of propagation.

Regularisation: Set a prior on the $\phi_j \sim \mathcal{L}(s)$

Examples:

$$\phi_j \sim \mathrm{Gamma}(\nu/2, \nu/2)$$
$$\phi_j \sim \mathrm{LogNormal}(1, \sigma)$$

Relaxed Random Walk (Lemey et al., 2010)

$$\mathbf{Z}^j \;\Big|\; \mathbf{Z}^{\mathrm{pa}(j)} \sim \mathcal{N}\left(\mathbf{Z}^{\mathrm{pa}(j)}, \phi_j \mathbf{R}\right) \qquad \phi_j \sim \mathcal{L}(s)$$

$$p\left(\mathbf{R}, \phi, s, \mathcal{T}, \psi \mid \mathbf{Y}, \mathbf{S}\right) \propto p\left(\mathbf{Y}, \mathbf{S} \mid \mathbf{R}, \phi, s, \mathcal{T}, \psi\right) p\left(\mathbf{R}, \phi, s, \mathcal{T}, \psi\right)$$

Relaxed Random Walk (Lemey et al., 2010)

$$\mathbf{Z}^j \,\Big|\, \mathbf{Z}^{\mathrm{pa}(j)} \sim \mathcal{N}\left(\mathbf{Z}^{\mathrm{pa}(j)}, \phi_j \mathbf{R}\right) \qquad \phi_j \sim \mathcal{L}(s)$$

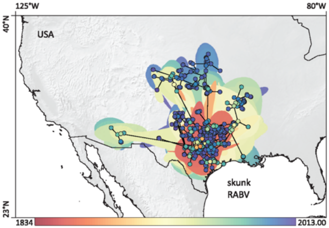
$$\begin{aligned} p\left(\mathbf{R}, \phi, s, \mathcal{T}, \psi \mid \mathbf{Y}, \mathbf{S}\right) &\propto p\left(\mathbf{Y}, \mathbf{S} \mid \mathbf{R}, \phi, s, \mathcal{T}, \psi\right) p\left(\mathbf{R}, \phi, s, \mathcal{T}, \psi\right) \\ &\propto p\left(\mathbf{Y} \mid \mathbf{R}, \phi, s, \mathcal{T}\right) p\left(\mathbf{S} \mid \mathcal{T}, \psi\right) p\left(\mathbf{R}, \phi, s, \mathcal{T}, \psi\right) \end{aligned}$$

Relaxed Random Walk (Lemey et al., 2010)

$$\mathbf{Z}^j \,\Big|\, \mathbf{Z}^{\mathrm{pa}(j)} \sim \mathcal{N}\left(\mathbf{Z}^{\mathrm{pa}(j)}, \phi_j \mathbf{R}\right) \qquad \phi_j \sim \mathcal{L}(s)$$

$$\begin{aligned}
p\left(\mathbf{R}, \phi, s, \mathcal{T}, \psi \mid \mathbf{Y}, \mathbf{S}\right) &\propto p\left(\mathbf{Y}, \mathbf{S} \mid \mathbf{R}, \phi, s, \mathcal{T}, \psi\right) p\left(\mathbf{R}, \phi, s, \mathcal{T}, \psi\right) \\
&\propto p\left(\mathbf{Y} \mid \mathbf{R}, \phi, s, \mathcal{T}\right) p\left(\mathbf{S} \mid \mathcal{T}, \psi\right) p\left(\mathbf{R}, \phi, s, \mathcal{T}, \psi\right) \\
&\propto p\left(\mathbf{Y} \mid \mathbf{R}, \phi, \mathcal{T}\right) p\left(\phi \mid s\right) p\left(s\right) p\left(\mathbf{R}\right) \\
&\qquad \times p\left(\mathbf{S} \mid \mathcal{T}, \psi\right) p\left(\mathcal{T}, \psi\right)
\end{aligned}$$

Phylogeography

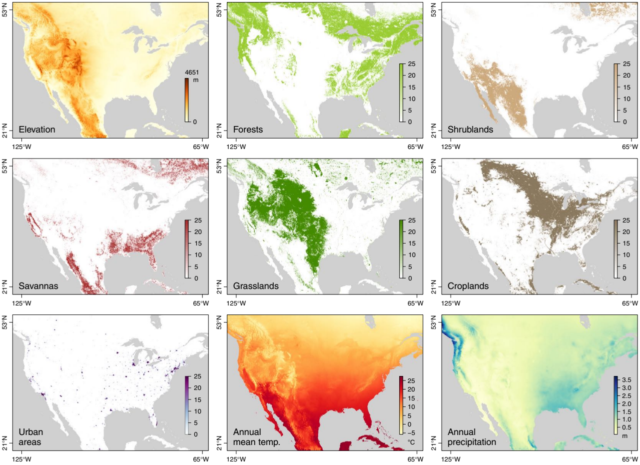


Dellicour et al. (2017)

Relaxed Random Walk

- Simple model.
- Flexible: accomodate for heterogeneous landscape.
- Ancestral state reconstruction: geopraphical spread of the epidemy.

Phylogeography: post-hoc analyses



Dellicour et al. (2019)

Bibliography

Betancourt. 2017. *arXiv e-print*. p. 170102434.

Dellicour, Gill, Faria, et al. 2021. *Molecular Biology and Evolution*. .

Dellicour, Rose, Faria, et al. 2017. *Molecular Biology and Evolution*. 34:2563–2571.

Dellicour, Troupin, Jahanbakhsh, et al. 2019. *Molecular Ecology*. 28:4335–4350.

Larget. 2005. In: Statistical Methods in Molecular Evolution, New York: Springer-Verlag, pp. 45–62.

Lemey, Rambaut, Welch, et al. 2010. *Molecular Biology and Evolution*. 27:1877–1885.

Lewandowski, Kurowicka, Joe. 2009. *Journal of Multivariate Analysis*. 100:1989–2001.

Thank you for listening

Institut Montpelliérain Alexander Grothendieck

Appendices

LKJ Distribution (Lewandowski, Kurowicka, and Joe, 2009)

Idea:
Use two different distributions for the variance and the correlation.

LKJ Distribution (Lewandowski, Kurowicka, and Joe, 2009)

Idea:
Use two different distributions for the variance and the correlation.

Decomposition: Use correlation matrix **C**

$$\mathbf{R} = \begin{pmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_p \end{pmatrix} \begin{pmatrix} 1 & & C_{kl} \\ & \ddots & \\ C_{kl} & & 1 \end{pmatrix} \begin{pmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_p \end{pmatrix}$$

LKJ Distribution (Lewandowski, Kurowicka, and Joe, 2009)

Idea:
Use two different distributions for the variance and the correlation.

Decomposition: Use correlation matrix $\mathbf{C}$

$$\mathbf{R} = \begin{pmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_p \end{pmatrix} \begin{pmatrix} 1 & & C_{kl} \\ & \ddots & \\ C_{kl} & & 1 \end{pmatrix} \begin{pmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_p \end{pmatrix}$$

$\boldsymbol{\sigma}$: Real positive $\rightarrow$ log-normal, gamma, half Gaussian, ...

LKJ Distribution (Lewandowski, Kurowicka, and Joe, 2009)

Idea:
Use two different distributions for the variance and the correlation.

Decomposition: Use correlation matrix **C**

$$\mathbf{R} = \begin{pmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_p \end{pmatrix} \begin{pmatrix} 1 & & C_{kl} \\ & \ddots & \\ C_{kl} & & 1 \end{pmatrix} \begin{pmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_p \end{pmatrix}$$

$\boldsymbol{\sigma}$: Real positive $\rightarrow$ log-normal, gamma, half Gaussian, ...

C: Correlation matrix $\rightarrow$ "LKJ" distribution.

LKJ Distribution (Lewandowski, Kurowicka, and Joe, 2009)

LKJ distribution:

$$\mathrm{LKJ}(\mathbf{C} \mid \eta) = c_p(\eta)\, |\mathbf{C}|^{\eta-1}$$

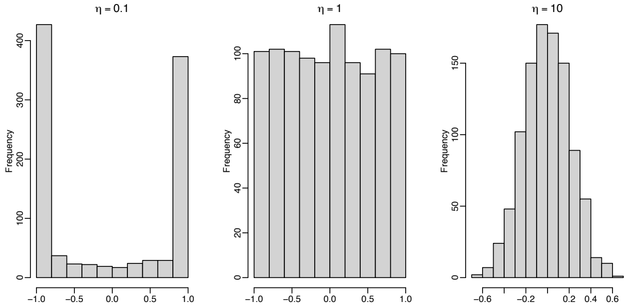
$\eta = 1$:	Uniform.
$\eta > 1$:	Peak around identity matrix.
$0 < \eta < 1$:	Trough around identity matrix.

LKJ Distribution (Lewandowski, Kurowicka, and Joe, 2009)

LKJ distribution:

$$\mathrm{LKJ}(\mathbf{C} \mid \eta) = c_p(\eta)\,|\mathbf{C}|^{\eta-1}$$

$\eta = 1$: Uniform.
$\eta > 1$: Peak around identity matrix.
$0 < \eta < 1$: Trough around identity matrix.



LKJ Distribution (Lewandowski, Kurowicka, and Joe, 2009)

Idea:
Use two different distributions for the variance and the correlation.

Decomposition: $\mathbf{R} = \boldsymbol{\sigma}\mathbf{C}\boldsymbol{\sigma}$

$\boldsymbol{\sigma}$: Real positive $\rightarrow$ log-normal, gamma, half Gaussian, ...

$\mathbf{C}$: Correlation matrix $\rightarrow$ "LKJ" distribution.

LKJ Distribution (Lewandowski, Kurowicka, and Joe, 2009)

Idea:
Use two different distributions for the variance and the correlation.

Decomposition: $\mathbf{R} = \boldsymbol{\sigma}\mathbf{C}\boldsymbol{\sigma}$

$\boldsymbol{\sigma}$: Real positive $\rightarrow$ log-normal, gamma, half Gaussian, ...

$\mathbf{C}$: Correlation matrix $\rightarrow$ "LKJ" distribution.

Problem: We cannot use the Gibbs sampler anymore.

LKJ Distribution (Lewandowski, Kurowicka, and Joe, 2009)

Idea:
Use two different distributions for the variance and the correlation.

Decomposition: $\mathbf{R} = \boldsymbol{\sigma}\mathbf{C}\boldsymbol{\sigma}$

$\boldsymbol{\sigma}$: Real positive $\rightarrow$ log-normal, gamma, half Gaussian, ...

$\mathbf{C}$: Correlation matrix $\rightarrow$ "LKJ" distribution.

Problem: We cannot use the Gibbs sampler anymore.

Problem: We cannot sample each coefficient of $\mathbf{C}$ independently.
$\hookrightarrow$ $\mathbf{C}$ has a complex structure.

LKJ Distribution (Lewandowski, Kurowicka, and Joe, 2009)

Idea:
Use two different distributions for the variance and the correlation.

Decomposition: $\mathbf{R} = \boldsymbol{\sigma}\mathbf{C}\boldsymbol{\sigma}$

$\boldsymbol{\sigma}$: Real positive $\rightarrow$ log-normal, gamma, half Gaussian, ...

$\mathbf{C}$: Correlation matrix $\rightarrow$ "LKJ" distribution.

Problem: We cannot use the Gibbs sampler anymore.

Problem: We cannot sample each coefficient of $\mathbf{C}$ independently.
$\hookrightarrow$ $\mathbf{C}$ has a complex structure.

Solution: Metropolis-Hasting in a constrained space.

MH in constrained space

Transformation:

$$f : \begin{cases} \mathcal{C}_q \to \mathbb{R}^q \\ \boldsymbol{\theta} \mapsto \boldsymbol{\nu} = f(\boldsymbol{\theta}) \end{cases}$$

MH in constrained space

Transformation:

$$f : \begin{cases} \mathcal{C}_q \to \mathbb{R}^q \\ \boldsymbol{\theta} \mapsto \boldsymbol{\nu} = f(\boldsymbol{\theta}) \end{cases}$$

Distribution: For a distribution π:

$$\pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = \pi_{\boldsymbol{\nu}}(f(\boldsymbol{\theta})) \times |J_f(\boldsymbol{\theta})|$$

MH in constrained space

Transformation:

$$f : \begin{cases} \mathcal{C}_q \to \mathbb{R}^q \\ \boldsymbol{\theta} \mapsto \boldsymbol{\nu} = f(\boldsymbol{\theta}) \end{cases}$$

Distribution: For a distribution π:

$$\pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = \pi_{\boldsymbol{\nu}}(f(\boldsymbol{\theta})) \times |J_f(\boldsymbol{\theta})|$$

Metropolis - Hasting: Iterate:

- Draw $\boldsymbol{\nu}^*$ in $q\left(\boldsymbol{\nu} \mid \boldsymbol{\nu}^t\right)$.
- Set $\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^* = f^{-1}(\boldsymbol{\nu}^*)$ with probability

$$r_t = \min\left\{1, \frac{p\left(\mathbf{Y} \mid \boldsymbol{\theta}^*\right)}{p\left(\mathbf{Y} \mid \boldsymbol{\theta}^t\right)} \frac{p\left(\boldsymbol{\theta}^*\right)}{p\left(\boldsymbol{\theta}^t\right)} \frac{q(\boldsymbol{\nu}^{(t)}|\boldsymbol{\nu}^*)}{q(\boldsymbol{\nu}^*|\boldsymbol{\nu}^{(t)})} \frac{|J_f(\boldsymbol{\theta}^{(t)})|}{|J_f(\boldsymbol{\theta}^*)|}\right\}.$$

Variance: LKJ transformation (Lewandowski, Kurowicka, and Joe, 2009)

Variance: LKJ transformation (Lewandowski, Kurowicka, and Joe, 2009)

Decomposition: Use correlation matrix $\mathbf{C}$

$$\mathbf{R} = \begin{pmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_p \end{pmatrix} \begin{pmatrix} 1 & & C_{kl} \\ & \ddots & \\ C_{kl} & & 1 \end{pmatrix} \begin{pmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_p \end{pmatrix}$$

Variance: LKJ transformation (Lewandowski, Kurowicka, and Joe, 2009)

Decomposition: Use correlation matrix $\mathbf{C}$

$$\mathbf{R} = \begin{pmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_p \end{pmatrix} \begin{pmatrix} 1 & & C_{kl} \\ & \ddots & \\ C_{kl} & & 1 \end{pmatrix} \begin{pmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_p \end{pmatrix}$$

$\boldsymbol{\sigma}$: Real positive

Variance: LKJ transformation (Lewandowski, Kurowicka, and Joe, 2009)

Decomposition: Use correlation matrix $\mathbf{C}$

$$\mathbf{R} = \begin{pmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_p \end{pmatrix} \begin{pmatrix} 1 & & C_{kl} \\ & \ddots & \\ C_{kl} & & 1 \end{pmatrix} \begin{pmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_p \end{pmatrix}$$

$\boldsymbol{\sigma}$: Real positive

$\mathbf{C}$: Correlation matrix

Variance: LKJ transformation (Lewandowski, Kurowicka, and Joe, 2009)

Decomposition: Use correlation matrix **C**

$$\mathbf{R} = \begin{pmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_p \end{pmatrix} \begin{pmatrix} 1 & & C_{kl} \\ & \ddots & \\ C_{kl} & & 1 \end{pmatrix} \begin{pmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_p \end{pmatrix}$$

$\boldsymbol{\sigma}$: Real positive

C: Correlation matrix

LKJ: Transformation on the space of correlation matrices.

Variance: LKJ transformation (Lewandowski, Kurowicka, and Joe, 2009)

Decomposition: Use correlation matrix $\mathbf{C}$

$$\mathbf{R} = \begin{pmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_p \end{pmatrix} \begin{pmatrix} 1 & & C_{kl} \\ & \ddots & \\ C_{kl} & & 1 \end{pmatrix} \begin{pmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_p \end{pmatrix}$$

$\boldsymbol{\sigma}$: Real positive

$\mathbf{C}$: Correlation matrix

LKJ: Transformation on the space of correlation matrices.
$\hookrightarrow$ Use "vine" theory.
$\hookrightarrow$ Easier and more efficient: Cholesky representation.

Variance: LKJ transformation (Lewandowski, Kurowicka, and Joe, 2009)

$$\mathbf{C} = \mathbf{W}^T\mathbf{W} = \begin{pmatrix} 1 & W_{12} & \cdots & W_{1p} \\ & W_{22} & & \vdots \\ & & \ddots & \vdots \\ 0 & & & W_{pp} \end{pmatrix}^T \begin{pmatrix} 1 & W_{12} & \cdots & W_{1p} \\ & W_{22} & & \vdots \\ & & \ddots & \vdots \\ 0 & & & W_{pp} \end{pmatrix}$$

With:

- $W_{1k}^2 + \cdots + W_{kk}^2 = 1$ (correlation)
- $W_{kk} > 0$ (identifiability)

Variance: LKJ transformation (Lewandowski, Kurowicka, and Joe, 2009)

$$\mathbf{C} = \mathbf{W}^T\mathbf{W} = \begin{pmatrix} 1 & W_{12} & \cdots & W_{1p} \\ & W_{22} & & \vdots \\ & & \ddots & \vdots \\ 0 & & & W_{pp} \end{pmatrix}^T \begin{pmatrix} 1 & W_{12} & \cdots & W_{1p} \\ & W_{22} & & \vdots \\ & & \ddots & \vdots \\ 0 & & & W_{pp} \end{pmatrix}$$

With:

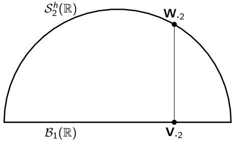
- $W_{1k}^2 + \cdots + W_{kk}^2 = 1$ (correlation)
- $W_{kk} > 0$ (identifiability)

$\hookrightarrow$ Each column k is in the half euclidian unit sphere $\mathcal{S}_k^{\mathrm{h}}(\mathbb{R})$.

Sampling in the Sphere

→ Sphere $\mathcal{S}_k^{\mathrm{h}}(\mathbb{R})$ to euclidean ball $\mathcal{B}_{k-1}(\mathbb{R})$:

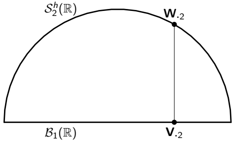
$$\mathbf{F} : \begin{cases} \mathcal{B}_{k-1}(\mathbb{R}) & \to \mathcal{S}_k^{\mathrm{h}}(\mathbb{R}) \\ \mathbf{V}_{\cdot k} & \mapsto \mathbf{W}_{\cdot k} = \left(\mathbf{V}_{\cdot k},\ \sqrt{1 - \|\mathbf{V}_{\cdot k}\|^2}\right) \end{cases}$$



Sampling in the Sphere

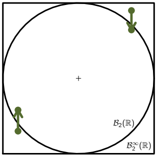
→ Sphere $\mathcal{S}_k^{\mathrm{h}}(\mathbb{R})$ to euclidean ball $\mathcal{B}_{k-1}(\mathbb{R})$:

$$\mathbf{F} : \begin{cases} \mathcal{B}_{k-1}(\mathbb{R}) & \to \mathcal{S}_k^{\mathrm{h}}(\mathbb{R}) \\ \mathbf{V}_{\cdot k} & \mapsto \mathbf{W}_{\cdot k} = \left(\mathbf{V}_{\cdot k},\ \sqrt{1 - \|\mathbf{V}_{\cdot k}\|^2}\right) \end{cases}$$



→ $\mathcal{B}_{k-1}(\mathbb{R})$ to infinite-norm ball $\mathcal{B}_{k-1}^{\infty}(\mathbb{R})$: "LKJ"

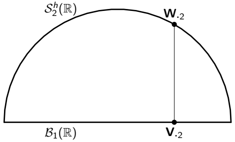
$$\mathbf{LKJ}_i(\mathbf{z}) = \begin{cases} z_i & \text{if } i = 1 \\ z_i \prod_{k=1}^{i-1}(1 - z_k^2)^{1/2} & \text{if } 1 < i \leq k-1. \end{cases}$$



Sampling in the Sphere

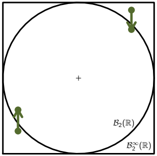
→ Sphere $\mathcal{S}_k^{\mathrm{h}}(\mathbb{R})$ to euclidean ball $\mathcal{B}_{k-1}(\mathbb{R})$:

$$\mathbf{F} : \begin{cases} \mathcal{B}_{k-1}(\mathbb{R}) & \to \mathcal{S}_k^{\mathrm{h}}(\mathbb{R}) \\ \mathbf{V}_{\cdot k} & \mapsto \mathbf{W}_{\cdot k} = \left(\mathbf{V}_{\cdot k},\ \sqrt{1 - \|\mathbf{V}_{\cdot k}\|^2}\right) \end{cases}$$



→ $\mathcal{B}_{k-1}(\mathbb{R})$ to infinite-norm ball $\mathcal{B}_{k-1}^{\infty}(\mathbb{R})$: "LKJ"

$$\mathbf{LKJ}_i(\mathbf{z}) = \begin{cases} z_i & \text{if } i = 1 \\ z_i \prod_{k=1}^{i-1}(1 - z_k^2)^{1/2} & \text{if } 1 < i \leq k-1. \end{cases}$$



→ $\mathcal{B}_{k-1}^{\infty}(\mathbb{R})$ to $\mathbb{R}^{k-1}$: "Fisher Z" transform

$$\tanh^{-1} : \begin{cases}]-1\,,1[\to \mathbb{R} \\ \qquad x \mapsto \dfrac{1}{2}\ln\left(\dfrac{1+x}{1-x}\right) \end{cases}$$

Variance: LKJ transformation (Lewandowski, Kurowicka, and Joe, 2009)

Decomposition: Use correlation matrix $\mathbf{C}$

$$\mathbf{R} = \begin{pmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_p \end{pmatrix} \begin{pmatrix} 1 & & C_{kl} \\ & \ddots & \\ C_{kl} & & 1 \end{pmatrix} \begin{pmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_p \end{pmatrix}$$

$\boldsymbol{\sigma}$: Real positive

$\mathbf{C}$: Correlation matrix

LKJ transformation: Smooth transformation from $\mathbb{R}^{p(p-1)/2}$ to the space of correlation matrices.

Variance: LKJ transformation (Lewandowski, Kurowicka, and Joe, 2009)

Decomposition: Use correlation matrix $\mathbf{C}$

$$\mathbf{R} = \begin{pmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_p \end{pmatrix} \begin{pmatrix} 1 & & C_{kl} \\ & \ddots & \\ C_{kl} & & 1 \end{pmatrix} \begin{pmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_p \end{pmatrix}$$

$\boldsymbol{\sigma}$: Real positive

$\mathbf{C}$: Correlation matrix

LKJ transformation: Smooth transformation from $\mathbb{R}^{p(p-1)/2}$ to the space of correlation matrices.

Jacobian: Can be computed.

OU: No Gibbs

Likelihood:

$$\mathbf{Y}|\mathbf{A}, \mathbf{R}, \boldsymbol{\mu} \not\sim \mathcal{MN}$$

OU: No Gibbs

Likelihood:

$$\mathbf{Y}|\mathbf{A}, \mathbf{R}, \boldsymbol{\mu} \not\sim \mathcal{MN}$$

Conjugate Priors: ???

OU: No Gibbs

Likelihood:

$$\mathbf{Y}|\mathbf{A}, \mathbf{R}, \boldsymbol{\mu} \not\sim \mathcal{MN}$$

Conjugate Priors: ???

Gibbs: Not possible.

OU: No Gibbs

Likelihood:

$$\mathbf{Y}|\mathbf{A}, \mathbf{R}, \boldsymbol{\mu} \not\sim \mathcal{MN}$$

Conjugate Priors: ???

Gibbs: Not possible.

Metropolis - Hasting:

Need to sample in constrained spaces (**A**, **R**).

back

Hamiltonian Monte Carlo (Betancourt, 2017)

Idea: Introduce the "momentum" $\mathbf{p}$ of the parameters $\mathbf{q}$.

Hamiltonian Monte Carlo (Betancourt, 2017)

Idea: Introduce the “momentum” $\mathbf{p}$ of the parameters $\mathbf{q}$.

$$\text{Hamiltonian } H(\mathbf{q}, \mathbf{p}) = \text{Potential energy} + \text{Kinetic energy}$$

Hamiltonian Monte Carlo (Betancourt, 2017)

Idea: Introduce the "momentum" $\mathbf{p}$ of the parameters $\mathbf{q}$.

$$\begin{aligned}\text{Hamiltonian } H(\mathbf{q}, \mathbf{p}) &= \text{Potential energy} + \text{Kinetic energy}\\ &= U(\mathbf{q}) + K(\mathbf{p})\\ &= -\log \underbrace{p(\mathbf{Y}|\mathbf{q})p(\mathbf{q})}_{\text{posterior}} - \log \underbrace{\phi(\mathbf{p})}_{\text{Gaussian}}\end{aligned}$$

Hamiltonian Monte Carlo (Betancourt, 2017)

Idea: Introduce the "momentum" $\mathbf{p}$ of the parameters $\mathbf{q}$.

$$\begin{aligned}\text{Hamiltonian } H(\mathbf{q}, \mathbf{p}) &= \text{Potential energy} + \text{Kinetic energy} \\ &= U(\mathbf{q}) + K(\mathbf{p}) \\ &= -\log \underbrace{p(\mathbf{Y}|\mathbf{q})p(\mathbf{q})}_{\text{posterior}} - \log \underbrace{\phi(\mathbf{p})}_{\text{Gaussian}}\end{aligned}$$

H **total energy** invariant by Hamiltonian dynamic: $\begin{cases} \dfrac{d\mathbf{p}}{dt} = \nabla_{\mathbf{q}} U(\mathbf{q}) \\ \dfrac{d\mathbf{q}}{dt} = -\nabla_{\mathbf{p}} K(\mathbf{p}) \end{cases}$

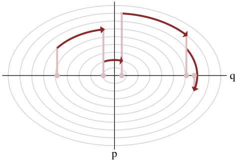
Hamiltonian Monte Carlo

(Betancourt, 2017)

Idea: Introduce the "momentum" $\mathbf{p}$ of the parameters $\mathbf{q}$.

$$\begin{aligned}\text{Hamiltonian } H(\mathbf{q},\mathbf{p}) &= \text{Potential energy} + \text{Kinetic energy}\\ &= U(\mathbf{q}) + K(\mathbf{p})\\ &= -\log \underbrace{p(\mathbf{Y}|\mathbf{q})p(\mathbf{q})}_{\text{posterior}} - \log \underbrace{\phi(\mathbf{p})}_{\text{Gaussian}}\end{aligned}$$

H **total energy** invariant by Hamiltonian dynamic: $\begin{cases} \dfrac{d\mathbf{p}}{dt} = \nabla_{\mathbf{q}} U(\mathbf{q}) \\ \dfrac{d\mathbf{q}}{dt} = -\nabla_{\mathbf{p}} K(\mathbf{p}) \end{cases}$



1. Draw random moments $\mathbf{p}$.
2. Propose a new $\mathbf{q}$ from the dynamic.