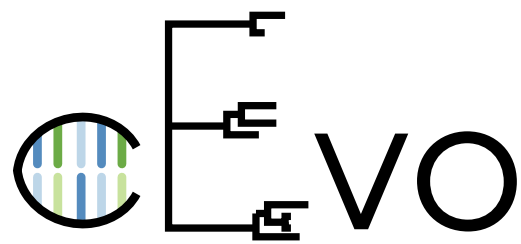
Exponential families for phylogeneticists

cEvo group meeting – Basel

Marc Manceau

June 23, 2020



4 reasons to speak about this concept

1. I didn't want to speak about ancestral population size inference,
2. I wanted to speak/teach something more on the methodological side,
3. Very nice concept of statistics that we could try to use more in phylogenetics,
4. Might give ideas for future projects aiming at speeding up Bayesian inference or provide you with nicely behaving building blocks for your next modeling work.

4 reasons to speak about this concept

1. I didn't want to speak about ancestral population size inference,
2. I wanted to speak/teach something more on the methodological side,
3. Very nice concept of statistics that we could try to use more in phylogenetics,
4. Might give ideas for future projects aiming at speeding up Bayesian inference or provide you with nicely behaving building blocks for your next modeling work.

4 reasons to speak about this concept

1. I didn't want to speak about ancestral population size inference,
2. I wanted to speak/teach something more on the methodological side,
3. Very nice concept of statistics that we could try to use more in phylogenetics,
4. Might give ideas for future projects aiming at speeding up Bayesian inference or provide you with nicely behaving building blocks for your next modeling work.

4 reasons to speak about this concept

1. I didn't want to speak about ancestral population size inference,
2. I wanted to speak/teach something more on the methodological side,
3. Very nice concept of statistics that we could try to use more in phylogenetics,
4. Might give ideas for future projects aiming at speeding up Bayesian inference or provide you with nicely behaving building blocks for your next modeling work.

4 reasons to speak about this concept

1. I didn't want to speak about ancestral population size inference,

2. I wanted to speak/teach something more on the methodological side,

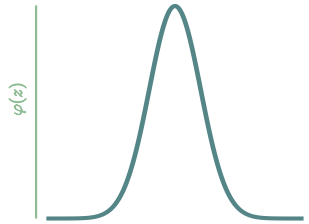
3. Very nice concept of statistics that we could try to use more in phylogenetics,

4. Might give ideas for future projects aiming at speeding up Bayesian inference or provide you with nicely behaving building blocks for your next modeling work.

Sketch of the presentation

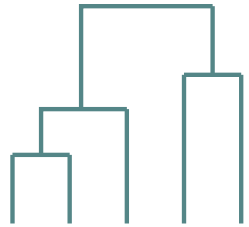
Some basics

- Definition
- Nice properties
- A few discrete examples
- A few continuous examples



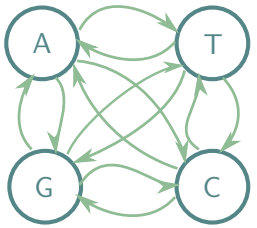
Phylodynamics modeling

- Kingman's coalescent
- Yule (pure birth) tree
- What about birth-death reconstructed trees ?



Trait evolution modeling

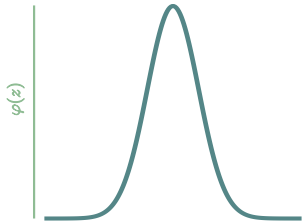
- Continuous trait evolution with BM
- Molecular evolution



Conclusion

Sketch of the presentation

Some basics
- Definition
- Nice properties
- A few discrete examples
- A few continuous examples



Phylodynamics modeling
- Kingman's coalescent
- Yule (pure birth) tree
- What about birth-death reconstructed trees ?

Trait evolution modeling
- Continuous trait evolution with BM
- Molecular evolution

Conclusion

Definition

A bit of lexicon related to exponential families

Definition 1

A family of probability distributions parametrized by a parameter θ is called an exponential family if its probability mass function, or density, can be expressed as

$$f(x|\theta) = h(x)e^{\eta(\theta)^t T(x) - A(\eta(\theta))}$$

natural parameter $\eta(\theta)$,
the distribution is said to be in its canonical form if $\eta(\theta) = \theta$,

sufficient statistic $T(x)$,
all information in the data that is related to the parameters θ.

log-partition function $A(\eta) = \ln\left(\int_x h(x)e^{\eta(\theta)^t T(x)}dx\right)$, which is the logarithm of the normalization factor, ensuring that f is a density.

Example 1

The family of exponential distributions $(\mathcal{E}(\lambda))_{\lambda\in\mathbb{R}^+}$ is an exponential family.
Indeed, we can express the density as,

$$f(x|\lambda) = \lambda e^{-\lambda x} = e^{-\lambda x + \ln\lambda}$$

And get the natural parameter $\eta := -\lambda$ and the sufficient statistic $T(x) = x$.

Definition

A bit of lexicon related to exponential families

Definition 1

A family of probability distributions parametrized by a parameter θ is called an exponential family if its probability mass function, or density, can be expressed as

$$f(x|\theta) = h(x)e^{\eta(\theta)^t T(x) - A(\eta(\theta))}$$

natural parameter $\eta(\theta)$,
the distribution is said to be in its canonical form if $\eta(\theta) = \theta$,

sufficient statistic $T(x)$,
all information in the data that is related to the parameters θ.

log-partition function $A(\eta) = \ln\left(\int_x h(x)e^{\eta(\theta)^t T(x)}dx\right)$, which is the logarithm of the normalization factor, ensuring that f is a density.

Example 1

The family of exponential distributions $(\mathcal{E}(\lambda))_{\lambda\in\mathbb{R}^+}$ is an exponential family.
Indeed, we can express the density as,

$$f(x|\lambda) = \lambda e^{-\lambda x} = e^{-\lambda x + \ln\lambda}$$

And get the natural parameter $\eta := -\lambda$ and the sufficient statistic $T(x) = x$.

Definition

A bit of lexicon related to exponential families

Definition 1

A family of probability distributions parametrized by a parameter θ is called an exponential family if its probability mass function, or density, can be expressed as

$$f(x|\theta) = h(x)e^{\eta(\theta)^t T(x) - A(\eta(\theta))}$$

natural parameter $\eta(\theta)$,
the distribution is said to be in its canonical form if $\eta(\theta) = \theta$,

sufficient statistic $T(x)$,
all information in the data that is related to the parameters θ.

log-partition function $A(\eta) = \ln\left(\int_x h(x)e^{\eta(\theta)^t T(x)}dx\right)$, which is the logarithm of the normalization factor, ensuring that f is a density.

Example 1

The family of exponential distributions $(\mathcal{E}(\lambda))_{\lambda\in\mathbb{R}^+}$ is an exponential family.
Indeed, we can express the density as,

$$f(x|\lambda) = \lambda e^{-\lambda x} = e^{-\lambda x + \ln\lambda}$$

And get the natural parameter $\eta := -\lambda$ and the sufficient statistic $T(x) = x$.

Definition

A bit of lexicon related to exponential families

Definition 1

A family of probability distributions parametrized by a parameter θ is called an exponential family if its probability mass function, or density, can be expressed as

$$f(x|\theta) = h(x)e^{\eta(\theta)^t T(x) - A(\eta(\theta))}$$

natural parameter $\eta(\theta)$,
the distribution is said to be in its canonical form if $\eta(\theta) = \theta$,

sufficient statistic $T(x)$,
all information in the data that is related to the parameters θ.

log-partition function $A(\eta) = \ln\left(\int_x h(x)e^{\eta(\theta)^t T(x)}dx\right)$, which is the logarithm of the normalization factor, ensuring that f is a density.

Example 1

The family of exponential distributions $(\mathcal{E}(\lambda))_{\lambda\in\mathbb{R}^+}$ is an exponential family.
Indeed, we can express the density as,

$$f(x|\lambda) = \lambda e^{-\lambda x} = e^{-\lambda x + \ln\lambda}$$

And get the natural parameter $\eta := -\lambda$ and the sufficient statistic $T(x) = x$.

Definition

A bit of lexicon related to exponential families

Definition 1

A family of probability distributions parametrized by a parameter θ is called an exponential family if its probability mass function, or density, can be expressed as

$$f(x|\theta) = h(x)e^{\eta(\theta)^t T(x) - A(\eta(\theta))}$$

natural parameter $\eta(\theta)$,
the distribution is said to be in its canonical form if $\eta(\theta) = \theta$,

sufficient statistic $T(x)$,
all information in the data that is related to the parameters θ.

log-partition function $A(\eta) = \ln\left(\int_x h(x)e^{\eta(\theta)^t T(x)}dx\right)$, which is the logarithm of the normalization factor, ensuring that f is a density.

Example 1

The family of exponential distributions $(\mathcal{E}(\lambda))_{\lambda\in\mathbb{R}^+}$ is an exponential family.
Indeed, we can express the density as,

$$f(x|\lambda) = \lambda e^{-\lambda x} = e^{-\lambda x + \ln\lambda}$$

And get the natural parameter $\eta := -\lambda$ and the sufficient statistic $T(x) = x$.

Definition

A bit of lexicon related to exponential families

Definition 1

A family of probability distributions parametrized by a parameter θ is called an exponential family if its probability mass function, or density, can be expressed as

$$f(x|\theta) = h(x)e^{\eta(\theta)^t T(x) - A(\eta(\theta))}$$

natural parameter $\eta(\theta)$,
the distribution is said to be in its canonical form if $\eta(\theta) = \theta$,

sufficient statistic $T(x)$,
all information in the data that is related to the parameters θ.

log-partition function $A(\eta) = \ln\left(\int_x h(x)e^{\eta(\theta)^t T(x)}dx\right)$, which is the logarithm of the normalization factor, ensuring that f is a density.

Example 1

The family of exponential distributions $(\mathcal{E}(\lambda))_{\lambda\in\mathbb{R}^+}$ is an exponential family.
Indeed, we can express the density as,

$$f(x|\lambda) = \lambda e^{-\lambda x} = e^{-\lambda x + \ln\lambda}$$

And get the natural parameter $\eta := -\lambda$ and the sufficient statistic $T(x) = x$.

Nice properties

The sufficient statistic summarizes nicely all iid observations

Property 1

One can summarize all the information coming from an arbitrary amount of iid random variables $X = (X_i)_{i=1}^n$ with only a fixed number of values, through the sufficient statistic

$$T(X) = \sum_{i=1}^{n} T(X_i) \quad .$$

Indeed, if $\mathcal{L}(\theta)$ is a member of an exponential family, and if $X = (X_i)_{i=1}^n$ is a sequence of n iid random variables distributed according to $\mathcal{L}(\theta)$, then the density of X is

$$f(x|\theta) = h(x_1)h(x_2)\ldots h(x_n)\exp\left(\eta(\theta)\sum_{i=1}^{n} T(x_i) - nA(\eta)\right)$$

Example 2

If $X = (X_i)_{i=1}^n$ are n iid exponentially distributed variables with rate λ, then all the information available to estimate λ is simply contained in

$$T(X) = \sum_{i=1}^{n} X_i$$

Nice properties

The sufficient statistic summarizes nicely all iid observations

Property 1

One can summarize all the information coming from an arbitrary amount of iid random variables $X = (X_i)_{i=1}^n$ with only a fixed number of values, through the sufficient statistic

$$T(X) = \sum_{i=1}^{n} T(X_i) \quad .$$

Indeed, if $\mathcal{L}(\theta)$ is a member of an exponential family, and if $X = (X_i)_{i=1}^n$ is a sequence of n iid random variables distributed according to $\mathcal{L}(\theta)$, then the density of X is

$$f(x|\theta) = h(x_1)h(x_2)\ldots h(x_n)\exp\left(\eta(\theta)\sum_{i=1}^{n} T(x_i) - nA(\eta)\right)$$

Example 2

If $X = (X_i)_{i=1}^n$ are n iid exponentially distributed variables with rate λ, then all the information available to estimate λ is simply contained in

$$T(X) = \sum_{i=1}^{n} X_i$$

Nice properties

The sufficient statistic summarizes nicely all iid observations

Property 1

One can summarize all the information coming from an arbitrary amount of iid random variables $X = (X_i)_{i=1}^n$ with only a fixed number of values, through the sufficient statistic

$$T(X) = \sum_{i=1}^{n} T(X_i) \quad .$$

Indeed, if $\mathcal{L}(\theta)$ is a member of an exponential family, and if $X = (X_i)_{i=1}^n$ is a sequence of n iid random variables distributed according to $\mathcal{L}(\theta)$, then the density of X is

$$f(x|\theta) = h(x_1)h(x_2)\ldots h(x_n)\exp\left(\eta(\theta)\sum_{i=1}^{n} T(x_i) - nA(\eta)\right)$$

Example 2

If $X = (X_i)_{i=1}^n$ are n iid exponentially distributed variables with rate λ, then all the information available to estimate λ is simply contained in

$$T(X) = \sum_{i=1}^{n} X_i$$

Nice properties

The sufficient statistic summarizes nicely all iid observations

Property 1

One can summarize all the information coming from an arbitrary amount of iid random variables $X = (X_i)_{i=1}^n$ with only a fixed number of values, through the sufficient statistic

$$T(X) = \sum_{i=1}^{n} T(X_i) \quad .$$

Indeed, if $\mathcal{L}(\theta)$ is a member of an exponential family, and if $X = (X_i)_{i=1}^n$ is a sequence of n iid random variables distributed according to $\mathcal{L}(\theta)$, then the density of X is

$$f(x|\theta) = h(x_1)h(x_2)\ldots h(x_n)\exp\left(\eta(\theta)\sum_{i=1}^{n} T(x_i) - nA(\eta)\right)$$

Example 2

If $X = (X_i)_{i=1}^n$ are n iid exponentially distributed variables with rate λ, then all the information available to estimate λ is simply contained in

$$T(X) = \sum_{i=1}^{n} X_i$$

Nice properties

The magic of the conjugate prior

Property 2

These exponential families admit conjugate priors that belong to another exponential family. I.e. if

$$X|\eta \sim f_\eta, \quad \text{where } f_\eta \text{ belongs to an exponential family } \mathcal{F}(\eta)$$

then there exists another exponential family $\mathcal{H}$ such that if $g \in \mathcal{H}$ and if

$$\eta \sim g, \quad \text{the posterior is given by}$$

$$\eta|X \sim h, \quad \text{where } h \text{ belongs to the same family } \mathcal{H}.$$

Nice properties

The magic of the conjugate prior

Property 2

These exponential families admit conjugate priors that belong to another exponential family. I.e. if

$$X|\eta \sim f_\eta, \quad \text{where } f_\eta \text{ belongs to an exponential family } \mathcal{F}(\eta)$$

then there exists another exponential family $\mathcal{H}$ such that if $g \in \mathcal{H}$ and if

$$\eta \sim g, \quad \text{the posterior is given by}$$

$$\eta|X \sim h, \quad \text{where } h \text{ belongs to the same family } \mathcal{H}.$$

Nice properties

The magic of the conjugate prior

Property 2

These exponential families admit conjugate priors that belong to another exponential family. I.e. if

$$X|\eta \sim f_\eta, \quad \text{where } f_\eta \text{ belongs to an exponential family } \mathcal{F}(\eta)$$

then there exists another exponential family $\mathcal{H}$ such that if $g \in \mathcal{H}$ and if

$$\eta \sim g, \quad \text{the posterior is given by}$$

$$\eta|X \sim h, \quad \text{where } h \text{ belongs to the same family } \mathcal{H}.$$

Indeed, we can even derive the density of this conjugate prior:

$$f(x|\eta) = h(x)e^{\eta^t T(x) - A(\eta)}$$

$$f(\eta|\chi, \nu) = p(\chi, \nu)e^{\eta\chi - \nu A(\eta)}$$

where χ, ν are hyperparameters.
One can directly check that the posterior is in the same family:

$$\begin{aligned} f(\eta|x, \chi, \nu) &\propto h(x)e^{\eta T(x) - A(\eta)} p(\chi, \nu)e^{\eta\chi - \nu A(\eta)} \\ &\propto e^{\eta(\chi + T(x)) - (\nu+1)A(\eta)} \end{aligned}$$

Nice properties

The magic of the conjugate prior

Property 2

These exponential families admit conjugate priors that belong to another exponential family. I.e. if

$$X|\eta \sim f_\eta, \quad \text{where } f_\eta \text{ belongs to an exponential family } \mathcal{F}(\eta)$$

then there exists another exponential family $\mathcal{H}$ such that if $g \in \mathcal{H}$ and if

$$\eta \sim g, \quad \text{the posterior is given by}$$

$$\eta|X \sim h, \quad \text{where } h \text{ belongs to the same family } \mathcal{H}.$$

Example 3

The exponential distribution is conjugate to a Gamma distribution.

$$\begin{aligned} \textit{Assume that} \quad & X|\lambda \sim \mathcal{E}(\lambda) \\ \textit{and} \quad & \lambda|\alpha, \beta \sim \Gamma(\alpha, \beta) \\ \Longrightarrow \quad & \lambda|X, \alpha, \beta \sim \Gamma(\alpha + 1, \beta + X) \quad . \end{aligned}$$

Nice properties

The magic of the conjugate prior

Property 2

These exponential families admit conjugate priors that belong to another exponential family. I.e. if

$$X|\eta \sim f_\eta, \quad \text{where } f_\eta \text{ belongs to an exponential family } \mathcal{F}(\eta)$$

then there exists another exponential family $\mathcal{H}$ such that if $g \in \mathcal{H}$ and if

$$\eta \sim g, \quad \text{the posterior is given by}$$

$$\eta|X \sim h, \quad \text{where } h \text{ belongs to the same family } \mathcal{H}.$$

Example 3

The exponential distribution is conjugate to a Gamma distribution.

$$\begin{aligned} \textit{Assume that} \quad & X|\lambda \sim \mathcal{E}(\lambda) \\ \textit{and} \quad & \lambda|\alpha,\beta \sim \Gamma(\alpha,\beta) \\ \Longrightarrow \quad & \lambda|X,\alpha,\beta \sim \Gamma(\alpha+1,\beta+X) \quad . \end{aligned}$$

Indeed, one can check the posterior density,

$$\begin{aligned} & f(X|\lambda) = \lambda e^{-\lambda X}, \\ & f(\lambda|\alpha,\beta) \propto \lambda^{\alpha-1} e^{-\beta\lambda} \\ \Longrightarrow \quad & f(\lambda|x,\alpha,\beta) \propto \lambda^{\alpha-1+1} e^{-(\beta+X)\lambda} \quad . \end{aligned}$$

A few discrete examples

The Poisson distribution

Example 4

The family of Poisson distributions $(\mathcal{P}(\lambda))_{\lambda \in \mathbb{R}^+}$ is an exponential family.
The conjugate prior is a Gamma distribution, with

$$\begin{aligned} \lambda|\alpha,\beta &\sim \Gamma(\alpha,\beta) \\ \Longrightarrow \quad \lambda|\alpha,\beta,x &\sim \Gamma(\alpha+x,\beta+1) \quad . \end{aligned}$$

Canonical form the probability mass function can be expressed as

$$f(x|\lambda) = \frac{1}{x!}\lambda^x e^{-\lambda} = \frac{1}{x!} e^{x \ln \lambda - \lambda}$$

The natural parameter is $\eta(\lambda) = \ln \lambda$ and the sufficient statistic is $T(x) = x$.

Conjugate prior If $X|\lambda \sim \mathcal{P}(\lambda)$ and $\lambda|\alpha,\beta \sim \Gamma(\alpha,\beta)$, we have

$$\begin{aligned} f(\lambda|\alpha,\beta) &\propto \lambda^{\alpha-1} e^{-\beta\lambda} \\ f(\lambda|x) &\propto \lambda^{\alpha-1+x} e^{-(\beta+1)\lambda} \\ \Longrightarrow \quad \lambda|x &\sim \Gamma(\alpha+x,\beta+1) \end{aligned}$$

Interpretation α is the total number of points observed, β is the total number of intervals.

A few discrete examples

The Poisson distribution

Example 4

The family of Poisson distributions $(\mathcal{P}(\lambda))_{\lambda\in\mathbb{R}^+}$ is an exponential family.
The conjugate prior is a Gamma distribution, with

$$\begin{aligned} & \lambda|\alpha,\beta \ \sim \Gamma(\alpha,\beta) \\ \Longrightarrow \ & \lambda|\alpha,\beta,x \ \sim \Gamma(\alpha+x,\beta+1) \ . \end{aligned}$$

Canonical form the probability mass function can be expressed as

$$f(x|\lambda) = \frac{1}{x!}\lambda^x e^{-\lambda} = \frac{1}{x!}e^{x\ln\lambda-\lambda}$$

The natural parameter is $\eta(\lambda) = \ln\lambda$ and the sufficient statistic is $T(x) = x$.

Conjugate prior If $X|\lambda \sim \mathcal{P}(\lambda)$ and $\lambda|\alpha,\beta \sim \Gamma(\alpha,\beta)$, we have

$$\begin{aligned} f(\lambda|\alpha,\beta) &\propto \lambda^{\alpha-1}e^{-\beta\lambda} \\ f(\lambda|x) &\propto \lambda^{\alpha-1+x}e^{-(\beta+1)\lambda} \\ \Longrightarrow \ \lambda|x &\sim \Gamma(\alpha+x,\beta+1) \end{aligned}$$

Interpretation α is the total number of points observed, β is the total number of intervals.

A few discrete examples

The Poisson distribution

Example 4

The family of Poisson distributions $(\mathcal{P}(\lambda))_{\lambda \in \mathbb{R}^+}$ *is an exponential family.*
The conjugate prior is a Gamma distribution, with

$$\begin{aligned} & \lambda|\alpha,\beta \ \sim \Gamma(\alpha,\beta) \\ \Longrightarrow \ & \lambda|\alpha,\beta,x \ \sim \Gamma(\alpha+x,\beta+1) \ . \end{aligned}$$

Canonical form the probability mass function can be expressed as

$$f(x|\lambda) = \frac{1}{x!}\lambda^x e^{-\lambda} = \frac{1}{x!} e^{x \ln \lambda - \lambda}$$

The natural parameter is $\eta(\lambda) = \ln \lambda$ and the sufficient statistic is $T(x) = x$.

Conjugate prior If $X|\lambda \sim \mathcal{P}(\lambda)$ and $\lambda|\alpha,\beta \sim \Gamma(\alpha,\beta)$, we have

$$\begin{aligned} & f(\lambda|\alpha,\beta) \propto \lambda^{\alpha-1} e^{-\beta\lambda} \\ & f(\lambda|x) \propto \lambda^{\alpha-1+x} e^{-(\beta+1)\lambda} \\ \Longrightarrow \ & \lambda|x \sim \Gamma(\alpha+x,\beta+1) \end{aligned}$$

Interpretation α is the total number of points observed, β is the total number of intervals.

A few discrete examples

The Poisson distribution

Example 4

The family of Poisson distributions $(\mathcal{P}(\lambda))_{\lambda \in \mathbb{R}^+}$ is an exponential family.
The conjugate prior is a Gamma distribution, with

$$\begin{aligned} & \lambda|\alpha,\beta \ \sim \Gamma(\alpha,\beta) \\ \Longrightarrow \ & \lambda|\alpha,\beta,x \ \sim \Gamma(\alpha+x,\beta+1) \ . \end{aligned}$$

Canonical form the probability mass function can be expressed as

$$f(x|\lambda) = \frac{1}{x!}\lambda^x e^{-\lambda} = \frac{1}{x!} e^{x \ln \lambda - \lambda}$$

The natural parameter is $\eta(\lambda) = \ln \lambda$ and the sufficient statistic is $T(x) = x$.

Conjugate prior If $X|\lambda \sim \mathcal{P}(\lambda)$ and $\lambda|\alpha,\beta \sim \Gamma(\alpha,\beta)$, we have

$$\begin{aligned} f(\lambda|\alpha,\beta) &\propto \lambda^{\alpha-1} e^{-\beta\lambda} \\ f(\lambda|x) &\propto \lambda^{\alpha-1+x} e^{-(\beta+1)\lambda} \\ \Longrightarrow \ \lambda|x &\sim \Gamma(\alpha+x,\beta+1) \end{aligned}$$

Interpretation α is the total number of points observed, β is the total number of intervals.

A few discrete examples

The Poisson distribution

Example 4

The family of Poisson distributions $(\mathcal{P}(\lambda))_{\lambda \in \mathbb{R}^+}$ *is an exponential family.*
The conjugate prior is a Gamma distribution, with

$$\begin{aligned} & \lambda|\alpha,\beta \ \sim \Gamma(\alpha,\beta) \\ \Longrightarrow \ & \lambda|\alpha,\beta,x \ \sim \Gamma(\alpha+x,\beta+1) \ . \end{aligned}$$

Canonical form the probability mass function can be expressed as

$$f(x|\lambda) = \frac{1}{x!}\lambda^x e^{-\lambda} = \frac{1}{x!}e^{x \ln \lambda - \lambda}$$

The natural parameter is $\eta(\lambda) = \ln \lambda$ and the sufficient statistic is $T(x) = x$.

Conjugate prior If $X|\lambda \sim \mathcal{P}(\lambda)$ and $\lambda|\alpha,\beta \sim \Gamma(\alpha,\beta)$, we have

$$\begin{aligned} & f(\lambda|\alpha,\beta) \propto \lambda^{\alpha-1}e^{-\beta\lambda} \\ & f(\lambda|x) \propto \lambda^{\alpha-1+x}e^{-(\beta+1)\lambda} \\ \Longrightarrow \ & \lambda|x \sim \Gamma(\alpha+x,\beta+1) \end{aligned}$$

Interpretation α is the total number of points observed, β is the total number of intervals.

A few discrete examples

Distributions related to Bernoulli experiments

Example 5

The following are all exponential families, conjugate to a Beta prior:
Geometric distributions $(\mathcal{G}(p))_{p\in(0,1)}$, *i.e. number of failures before the first success.*
Binomial distributions $(\mathcal{B}(n,p))_{p\in(0,1)}$, *i.e. number of successes in n successive experiments.*
Negative binomial distributions $(\mathcal{NB}(r,p))_{p\in(0,1)}$, *i.e. number of successes before finding r failures.*

Canonical form we can express these pmfs for the number of successes n_s and number of failures n_f,

$$f_1(n_f|p) = (1-p)^{n_f}\, p = e^{n_f \ln(1-p)+\ln p},$$

$$f_2(n_s|p) = \binom{n}{n_s} p^{n_s}(1-p)^{n-n_s} = \binom{n}{n_s} e^{n_s \ln \frac{p}{1-p} + n\ln(1-p)},$$

$$f_3(n_s|p) = \binom{n_s+r-1}{n_s}(1-p)^r p^{n_s} = \binom{n_s+r-1}{n_s} e^{n_s \ln p + r\ln(1-p)}$$

Conjugate prior Assume that $p|\alpha,\beta \sim$ Beta(α,β), then,

$$\text{if } n_f|p \sim \mathcal{G}(p) \implies p|n_f,\alpha,\beta \sim \text{Beta}(\alpha+1,\beta+n_f)$$
$$\text{if } n_s|p \sim \mathcal{B}(n,p) \implies p|n_s,\alpha,\beta \sim \text{Beta}(\alpha+n_s,\beta+(n-n_s))$$
$$\text{if } n_s|p \sim \mathcal{NB}(r,p) \implies p|n_s,\alpha,\beta \sim \text{Beta}(\alpha+n_s,\beta+r)$$

Interpretation $\alpha-1$ is the total number of successes, $\beta-1$ is the total number of failures.

With more than two outcomes the multinomial distribution is conjugate to the Dirichlet distribution.

A few discrete examples

Distributions related to Bernoulli experiments

Example 5

The following are all exponential families, conjugate to a Beta prior:
Geometric distributions $(\mathcal{G}(p))_{p\in(0,1)}$*, i.e. number of failures before the first success.*
Binomial distributions $(\mathcal{B}(n,p))_{p\in(0,1)}$*, i.e. number of successes in n successive experiments.*
Negative binomial distributions $(\mathcal{NB}(r,p))_{p\in(0,1)}$*, i.e. number of successes before finding r failures.*

Canonical form we can express these pmfs for the number of successes n_s and number of failures n_f,

$$f_1(n_f|p) = (1-p)^{n_f}\, p = e^{n_f \ln(1-p)+\ln p},$$

$$f_2(n_s|p) = \binom{n}{n_s} p^{n_s}(1-p)^{n-n_s} = \binom{n}{n_s} e^{n_s \ln \frac{p}{1-p}+n\ln(1-p)},$$

$$f_3(n_s|p) = \binom{n_s+r-1}{n_s}(1-p)^r p^{n_s} = \binom{n_s+r-1}{n_s} e^{n_s \ln p + r\ln(1-p)}$$

Conjugate prior Assume that $p|\alpha,\beta \sim$ Beta(α,β), then,

$$\text{if } n_f|p \sim \mathcal{G}(p) \implies p|n_f,\alpha,\beta \sim \text{Beta}(\alpha+1,\beta+n_f)$$
$$\text{if } n_s|p \sim \mathcal{B}(n,p) \implies p|n_s,\alpha,\beta \sim \text{Beta}(\alpha+n_s,\beta+(n-n_s))$$
$$\text{if } n_s|p \sim \mathcal{NB}(r,p) \implies p|n_s,\alpha,\beta \sim \text{Beta}(\alpha+n_s,\beta+r)$$

Interpretation $\alpha-1$ is the total number of successes, $\beta-1$ is the total number of failures.

With more than two outcomes the multinomial distribution is conjugate to the Dirichlet distribution.

A few discrete examples

Distributions related to Bernoulli experiments

Example 5

The following are all exponential families, conjugate to a Beta prior:
Geometric distributions $(\mathcal{G}(p))_{p\in(0,1)}$*, i.e. number of failures before the first success.*
Binomial distributions $(\mathcal{B}(n,p))_{p\in(0,1)}$*, i.e. number of successes in n successive experiments.*
Negative binomial distributions $(\mathcal{NB}(r,p))_{p\in(0,1)}$*, i.e. number of successes before finding r failures.*

Canonical form we can express these pmfs for the number of successes n_s and number of failures n_f,

$$f_1(n_f|p) = (1-p)^{n_f}\, p = e^{n_f \ln(1-p)+\ln p},$$

$$f_2(n_s|p) = \binom{n}{n_s} p^{n_s}(1-p)^{n-n_s} = \binom{n}{n_s} e^{n_s \ln \frac{p}{1-p} + n\ln(1-p)},$$

$$f_3(n_s|p) = \binom{n_s+r-1}{n_s}(1-p)^r p^{n_s} = \binom{n_s+r-1}{n_s} e^{n_s \ln p + r\ln(1-p)}$$

Conjugate prior Assume that $p|\alpha,\beta \sim \text{Beta}(\alpha,\beta)$, then,

$$\text{if } n_f|p \sim \mathcal{G}(p) \implies p|n_f,\alpha,\beta \sim \text{Beta}(\alpha+1,\beta+n_f)$$

$$\text{if } n_s|p \sim \mathcal{B}(n,p) \implies p|n_s,\alpha,\beta \sim \text{Beta}(\alpha+n_s,\beta+(n-n_s))$$

$$\text{if } n_s|p \sim \mathcal{NB}(r,p) \implies p|n_s,\alpha,\beta \sim \text{Beta}(\alpha+n_s,\beta+r)$$

Interpretation $\alpha-1$ is the total number of successes, $\beta-1$ is the total number of failures.

With more than two outcomes the multinomial distribution is conjugate to the Dirichlet distribution.

A few discrete examples

Distributions related to Bernoulli experiments

Example 5

The following are all exponential families, conjugate to a Beta prior:
Geometric distributions $(\mathcal{G}(p))_{p\in(0,1)}$, *i.e. number of failures before the first success.*
Binomial distributions $(\mathcal{B}(n,p))_{p\in(0,1)}$, *i.e. number of successes in n successive experiments.*
Negative binomial distributions $(\mathcal{NB}(r,p))_{p\in(0,1)}$, *i.e. number of successes before finding r failures.*

Canonical form we can express these pmfs for the number of successes n_s and number of failures n_f,

$$f_1(n_f|p) = (1-p)^{n_f}\, p = e^{n_f \ln(1-p)+\ln p},$$

$$f_2(n_s|p) = \binom{n}{n_s} p^{n_s}(1-p)^{n-n_s} = \binom{n}{n_s} e^{n_s \ln \frac{p}{1-p} + n\ln(1-p)},$$

$$f_3(n_s|p) = \binom{n_s+r-1}{n_s}(1-p)^r p^{n_s} = \binom{n_s+r-1}{n_s} e^{n_s \ln p + r\ln(1-p)}$$

Conjugate prior Assume that $p|\alpha,\beta \sim$ Beta(α,β), then,

if $n_f|p \sim \mathcal{G}(p) \implies p|n_f,\alpha,\beta \sim$ Beta$(\alpha+1, \beta+n_f)$

if $n_s|p \sim \mathcal{B}(n,p) \implies p|n_s,\alpha,\beta \sim$ Beta$(\alpha+n_s, \beta+(n-n_s))$

if $n_s|p \sim \mathcal{NB}(r,p) \implies p|n_s,\alpha,\beta \sim$ Beta$(\alpha+n_s, \beta+r)$

Interpretation $\alpha-1$ is the total number of successes, $\beta-1$ is the total number of failures.

With more than two outcomes the multinomial distribution is conjugate to the Dirichlet distribution.

A few discrete examples

Distributions related to Bernoulli experiments

Example 5

The following are all exponential families, conjugate to a Beta prior:
Geometric distributions $(\mathcal{G}(p))_{p\in(0,1)}$, *i.e. number of failures before the first success.*
Binomial distributions $(\mathcal{B}(n,p))_{p\in(0,1)}$, *i.e. number of successes in n successive experiments.*
Negative binomial distributions $(\mathcal{NB}(r,p))_{p\in(0,1)}$, *i.e. number of successes before finding r failures.*

Canonical form we can express these pmfs for the number of successes n_s and number of failures n_f,

$$f_1(n_f|p) = (1-p)^{n_f}\, p = e^{n_f \ln(1-p)+\ln p},$$

$$f_2(n_s|p) = \binom{n}{n_s} p^{n_s}(1-p)^{n-n_s} = \binom{n}{n_s} e^{n_s \ln \frac{p}{1-p} + n\ln(1-p)},$$

$$f_3(n_s|p) = \binom{n_s+r-1}{n_s}(1-p)^r p^{n_s} = \binom{n_s+r-1}{n_s} e^{n_s \ln p + r\ln(1-p)}$$

Conjugate prior Assume that $p|\alpha,\beta \sim$ Beta(α,β), then,

if $n_f|p \sim \mathcal{G}(p) \implies p|n_f,\alpha,\beta \sim$ Beta$(\alpha+1, \beta+n_f)$

if $n_s|p \sim \mathcal{B}(n,p) \implies p|n_s,\alpha,\beta \sim$ Beta$(\alpha+n_s, \beta+(n-n_s))$

if $n_s|p \sim \mathcal{NB}(r,p) \implies p|n_s,\alpha,\beta \sim$ Beta$(\alpha+n_s, \beta+r)$

Interpretation $\alpha-1$ is the total number of successes, $\beta-1$ is the total number of failures.

With more than two outcomes the multinomial distribution is conjugate to the Dirichlet distribution.

A few discrete examples

Distributions related to Bernoulli experiments

Example 5

The following are all exponential families, conjugate to a Beta prior:
Geometric distributions $(\mathcal{G}(p))_{p\in(0,1)}$*, i.e. number of failures before the first success.*
Binomial distributions $(\mathcal{B}(n,p))_{p\in(0,1)}$*, i.e. number of successes in n successive experiments.*
Negative binomial distributions $(\mathcal{NB}(r,p))_{p\in(0,1)}$*, i.e. number of successes before finding r failures.*

Canonical form we can express these pmfs for the number of successes n_s and number of failures n_f,

$$f_1(n_f|p) = (1-p)^{n_f}\, p = e^{n_f \ln(1-p)+\ln p},$$

$$f_2(n_s|p) = \binom{n}{n_s} p^{n_s}(1-p)^{n-n_s} = \binom{n}{n_s} e^{n_s \ln \frac{p}{1-p} + n\ln(1-p)},$$

$$f_3(n_s|p) = \binom{n_s+r-1}{n_s}(1-p)^r p^{n_s} = \binom{n_s+r-1}{n_s} e^{n_s \ln p + r\ln(1-p)}$$

Conjugate prior Assume that $p|\alpha,\beta \sim \text{Beta}(\alpha,\beta)$, then,

$$\text{if } n_f|p \sim \mathcal{G}(p) \implies p|n_f,\alpha,\beta \sim \text{Beta}(\alpha+1,\beta+n_f)$$

$$\text{if } n_s|p \sim \mathcal{B}(n,p) \implies p|n_s,\alpha,\beta \sim \text{Beta}(\alpha+n_s,\beta+(n-n_s))$$

$$\text{if } n_s|p \sim \mathcal{NB}(r,p) \implies p|n_s,\alpha,\beta \sim \text{Beta}(\alpha+n_s,\beta+r)$$

Interpretation $\alpha-1$ is the total number of successes, $\beta-1$ is the total number of failures.

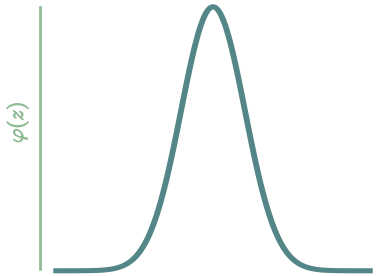
With more than two outcomes the multinomial distribution is conjugate to the Dirichlet distribution.

A few continuous examples

The Gaussian distribution

Example 6

The 3 family of Gaussian distributions with one parameter fixed or not, i.e. $(\mathcal{N}(\mu, \sigma^2))_{\mu \in \mathbb{R}}$ $(\mathcal{N}(\mu, \sigma^2))_{\sigma^2 \in \mathbb{R}^+}$ $(\mathcal{N}(\mu, \sigma^2))_{\mu \in \mathbb{R}, \sigma^2 \in R^+}$*, are all exponential families.*



Canonical forms are a bit messier on this example, but they respectively lead to,

$$\eta(\mu) = (\mu/\sigma^2) \text{ and } T(x) = x,$$
$$\eta(\sigma^2) = -1/(2\sigma^2) \text{ and } T(x) = x^2 - 2\mu x,$$
$$\eta(\mu, \sigma^2) = \left(\mu/\sigma^2, -1/(2\sigma^2)\right) \text{ and } T(x) = (x, x^2) \quad .$$

Conjugate prior they are respectively conjugate to the following priors:

$$\mu \sim \mathcal{N}(\mu_0, \sigma_0^2),$$
$$\sigma^2 \sim \Gamma^{-1}(\alpha, \beta),$$
$$(\mu, \sigma^2) \sim \mathcal{N}\Gamma^{-1}(\mu_0, \lambda, \alpha, \beta).$$

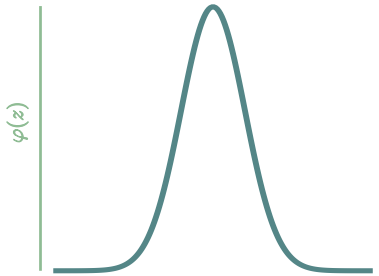
Multivariate normal Similar results still hold in higher dimension.

A few continuous examples

The Gaussian distribution

Example 6

The 3 family of Gaussian distributions with one parameter fixed or not, i.e. $(\mathcal{N}(\mu, \sigma^2))_{\mu \in \mathbb{R}}$ $(\mathcal{N}(\mu, \sigma^2))_{\sigma^2 \in \mathbb{R}^+}$ $(\mathcal{N}(\mu, \sigma^2))_{\mu \in \mathbb{R}, \sigma^2 \in R^+}$*, are all exponential families.*



Canonical forms are a bit messier on this example, but they respectively lead to,

$$\eta(\mu) = (\mu/\sigma^2) \text{ and } T(x) = x,$$
$$\eta(\sigma^2) = -1/(2\sigma^2) \text{ and } T(x) = x^2 - 2\mu x,$$
$$\eta(\mu, \sigma^2) = \left(\mu/\sigma^2, -1/(2\sigma^2)\right) \text{ and } T(x) = (x, x^2) \quad .$$

Conjugate prior they are respectively conjugate to the following priors:

$$\mu \sim \mathcal{N}(\mu_0, \sigma_0^2),$$
$$\sigma^2 \sim \Gamma^{-1}(\alpha, \beta),$$
$$(\mu, \sigma^2) \sim \mathcal{N}\Gamma^{-1}(\mu_0, \lambda, \alpha, \beta).$$

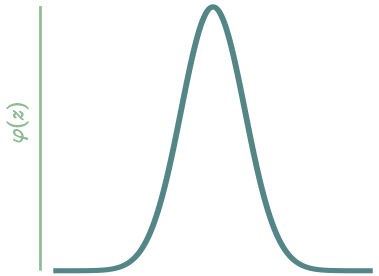
Multivariate normal Similar results still hold in higher dimension.

A few continuous examples

The Gaussian distribution

Example 6

The 3 family of Gaussian distributions with one parameter fixed or not, i.e. $(\mathcal{N}(\mu, \sigma^2))_{\mu \in \mathbb{R}}$ $(\mathcal{N}(\mu, \sigma^2))_{\sigma^2 \in \mathbb{R}^+}$ $(\mathcal{N}(\mu, \sigma^2))_{\mu \in \mathbb{R}, \sigma^2 \in R^+}$*, are all exponential families.*



Canonical forms are a bit messier on this example, but they respectively lead to,

$$\eta(\mu) = (\mu/\sigma^2) \text{ and } T(x) = x,$$
$$\eta(\sigma^2) = -1/(2\sigma^2) \text{ and } T(x) = x^2 - 2\mu x,$$
$$\eta(\mu, \sigma^2) = \left(\mu/\sigma^2, -1/(2\sigma^2)\right) \text{ and } T(x) = (x, x^2) \quad .$$

Conjugate prior they are respectively conjugate to the following priors:

$$\mu \sim \mathcal{N}(\mu_0, \sigma_0^2),$$
$$\sigma^2 \sim \Gamma^{-1}(\alpha, \beta),$$
$$(\mu, \sigma^2) \sim \mathcal{N}\Gamma^{-1}(\mu_0, \lambda, \alpha, \beta).$$

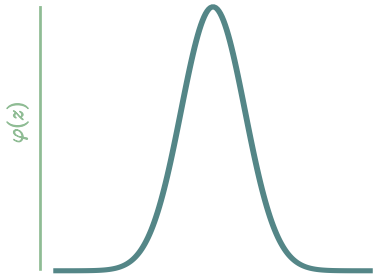
Multivariate normal Similar results still hold in higher dimension.

A few continuous examples

The Gaussian distribution

Example 6

The 3 family of Gaussian distributions with one parameter fixed or not, i.e. $(\mathcal{N}(\mu, \sigma^2))_{\mu \in \mathbb{R}}$ $(\mathcal{N}(\mu, \sigma^2))_{\sigma^2 \in \mathbb{R}^+}$ $(\mathcal{N}(\mu, \sigma^2))_{\mu \in \mathbb{R}, \sigma^2 \in R^+}$, *are all exponential families.*



Canonical forms are a bit messier on this example, but they respectively lead to,

$$\eta(\mu) = (\mu/\sigma^2) \text{ and } T(x) = x,$$
$$\eta(\sigma^2) = -1/(2\sigma^2) \text{ and } T(x) = x^2 - 2\mu x,$$
$$\eta(\mu, \sigma^2) = \left(\mu/\sigma^2, -1/(2\sigma^2)\right) \text{ and } T(x) = (x, x^2) \quad .$$

Conjugate prior they are respectively conjugate to the following priors:

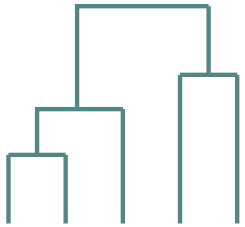
$$\mu \sim \mathcal{N}(\mu_0, \sigma_0^2),$$
$$\sigma^2 \sim \Gamma^{-1}(\alpha, \beta),$$
$$(\mu, \sigma^2) \sim \mathcal{N}\Gamma^{-1}(\mu_0, \lambda, \alpha, \beta).$$

Multivariate normal Similar results still hold in higher dimension.

Sketch of the presentation

Some basics
- Definition
- Nice properties
- A few discrete examples
- A few continuous examples

Phylodynamics modeling
- Kingman's coalescent
- Yule (pure birth) tree
- What about birth-death reconstructed trees ?

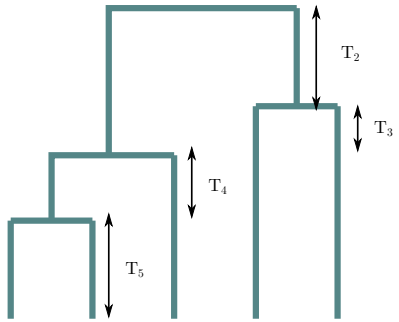


Trait evolution modeling
- Continuous trait evolution with BM
- Molecular evolution

Conclusion

Kingman's coalescent

- Let $\mathcal{T}$ be a tree with n leaves,
- Let $T_i :=$ time elapsed to go from i to $i-1$ lineages,



then $\mathcal{T}$ is said to follow a Kingman coalescent with parameter θ if its density is,

$$f(\mathcal{T}|\theta) = \prod_{i=2}^{n} \theta e^{-\theta \binom{i}{2} T_i} = \exp\left(-\theta \sum_{i=2}^{n} \binom{i}{2} T_i + (n-1) \ln \theta \right)$$

Exponential family with natural parameter $-\theta$ and sufficient statistic $\sum_{i=2}^{n} \binom{i}{2} T_i$.

Conjugate prior Assume $\theta \sim \Gamma(\alpha, \beta)$, then,

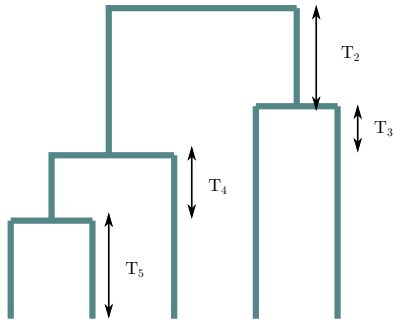
$$\begin{aligned} f(\theta|\mathcal{T}, \alpha, \beta) &\propto \theta^{\alpha-1} e^{-\beta\theta} \theta^{n-1} e^{-\theta \sum_{i=2}^{n} T_i} \\ &\propto \theta^{\alpha-1+(n-1)} e^{-\theta \left(\beta + \sum_{i=2}^{n} \binom{i}{2} T_i\right)} \end{aligned}$$

which means that $\theta|\mathcal{T}, \alpha, \beta \sim \Gamma\left(\alpha + n - 1,\ \beta + \sum_{i=2}^{n} \binom{i}{2} T_i\right)$.

In the literature I found a few papers using this, see Parag et al. (2020).

Kingman's coalescent

- Let $\mathcal{T}$ be a tree with n leaves,
- Let $T_i :=$ time elapsed to go from i to $i-1$ lineages,



then $\mathcal{T}$ is said to follow a Kingman coalescent with parameter θ if its density is,

$$f(\mathcal{T}|\theta) = \prod_{i=2}^{n} \theta e^{-\theta \binom{i}{2} T_i} = \exp\left(-\theta \sum_{i=2}^{n} \binom{i}{2} T_i + (n-1)\ln\theta\right)$$

Exponential family with natural parameter $-\theta$ and sufficient statistic $\sum_{i=2}^{n} \binom{i}{2} T_i$.

Conjugate prior Assume $\theta \sim \Gamma(\alpha, \beta)$, then,

$$f(\theta|\mathcal{T}, \alpha, \beta) \propto \theta^{\alpha-1} e^{-\beta\theta} \theta^{n-1} e^{-\theta \sum_{i=2}^{n} T_i}$$

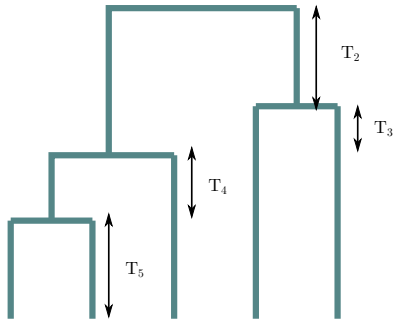
$$\propto \theta^{\alpha-1+(n-1)} e^{-\theta\left(\beta + \sum_{i=2}^{n} \binom{i}{2} T_i\right)}$$

which means that $\theta|\mathcal{T}, \alpha, \beta \sim \Gamma\left(\alpha + n - 1,\ \beta + \sum_{i=2}^{n} \binom{i}{2} T_i\right)$.

In the literature I found a few papers using this, see Parag et al. (2020).

Kingman's coalescent

- Let $\mathcal{T}$ be a tree with n leaves,
- Let $T_i :=$ time elapsed to go from i to $i-1$ lineages,



then $\mathcal{T}$ is said to follow a Kingman coalescent with parameter θ if its density is,

$$f(\mathcal{T}|\theta) = \prod_{i=2}^{n} \theta e^{-\theta \binom{i}{2} T_i} = \exp\left(-\theta \sum_{i=2}^{n} \binom{i}{2} T_i + (n-1) \ln \theta \right)$$

Exponential family with natural parameter $-\theta$ and sufficient statistic $\sum_{i=2}^{n} \binom{i}{2} T_i$.

Conjugate prior Assume $\theta \sim \Gamma(\alpha, \beta)$, then,

$$f(\theta|\mathcal{T}, \alpha, \beta) \propto \theta^{\alpha-1} e^{-\beta\theta} \theta^{n-1} e^{-\theta \sum_{i=2}^{n} T_i}$$

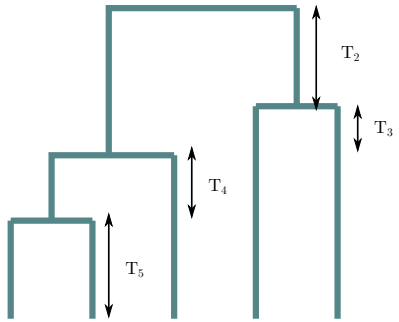
$$\propto \theta^{\alpha-1+(n-1)} e^{-\theta \left(\beta + \sum_{i=2}^{n} \binom{i}{2} T_i\right)}$$

which means that $\theta|\mathcal{T}, \alpha, \beta \sim \Gamma\left(\alpha + n - 1,\ \beta + \sum_{i=2}^{n} \binom{i}{2} T_i\right)$.

In the literature I found a few papers using this, see Parag et al. (2020).

Kingman's coalescent

- Let $\mathcal{T}$ be a tree with n leaves,
- Let $T_i :=$ time elapsed to go from i to $i-1$ lineages,



then $\mathcal{T}$ is said to follow a Kingman coalescent with parameter θ if its density is,

$$f(\mathcal{T}|\theta) = \prod_{i=2}^{n} \theta e^{-\theta \binom{i}{2} T_i} = \exp\left(-\theta \sum_{i=2}^{n} \binom{i}{2} T_i + (n-1) \ln \theta \right)$$

Exponential family with natural parameter $-\theta$ and sufficient statistic $\sum_{i=2}^{n} \binom{i}{2} T_i$.

Conjugate prior Assume $\theta \sim \Gamma(\alpha, \beta)$, then,

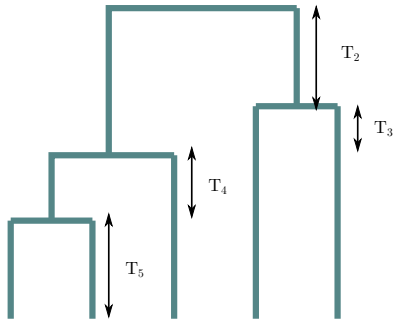
$$\begin{aligned} f(\theta|\mathcal{T}, \alpha, \beta) &\propto \theta^{\alpha-1} e^{-\beta\theta} \theta^{n-1} e^{-\theta \sum_{i=2}^{n} T_i} \\ &\propto \theta^{\alpha-1+(n-1)} e^{-\theta \left(\beta + \sum_{i=2}^{n} \binom{i}{2} T_i\right)} \end{aligned}$$

which means that $\theta|\mathcal{T}, \alpha, \beta \sim \Gamma\left(\alpha + n - 1,\ \beta + \sum_{i=2}^{n} \binom{i}{2} T_i\right)$.

In the literature I found a few papers using this, see Parag et al. (2020).

Kingman's coalescent

- Let $\mathcal{T}$ be a tree with n leaves,
- Let $T_i :=$ time elapsed to go from i to $i-1$ lineages,



then $\mathcal{T}$ is said to follow a Kingman coalescent with parameter θ if its density is,

$$f(\mathcal{T}|\theta) = \prod_{i=2}^{n} \theta e^{-\theta \binom{i}{2} T_i} = \exp\left(-\theta \sum_{i=2}^{n} \binom{i}{2} T_i + (n-1) \ln \theta \right)$$

Exponential family with natural parameter $-\theta$ and sufficient statistic $\sum_{i=2}^{n} \binom{i}{2} T_i$.

Conjugate prior Assume $\theta \sim \Gamma(\alpha, \beta)$, then,

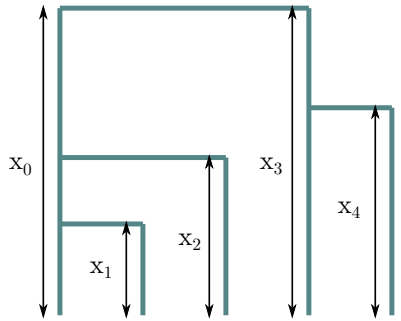
$$\begin{aligned} f(\theta|\mathcal{T}, \alpha, \beta) &\propto \theta^{\alpha-1} e^{-\beta\theta} \theta^{n-1} e^{-\theta \sum_{i=2}^{n} T_i} \\ &\propto \theta^{\alpha-1+(n-1)} e^{-\theta \left(\beta + \sum_{i=2}^{n} \binom{i}{2} T_i\right)} \end{aligned}$$

which means that $\theta|\mathcal{T}, \alpha, \beta \sim \Gamma\left(\alpha + n - 1,\ \beta + \sum_{i=2}^{n} \binom{i}{2} T_i\right)$.

In the literature I found a few papers using this, see Parag et al. (2020).

Yule (pure birth) tree

- Let $\mathcal{T}$ be a tree with n leaves,
- Let x_i be the depth of leaf i,



then $\mathcal{T}$ is said to be a Yule tree if it has the following density,

$$f(\mathcal{T}|\lambda) \propto \lambda^{n-1} \prod_{i=0}^{n-1} e^{-\lambda x_i} \propto \exp\left(-\lambda \sum_{i=0}^{n-1} x_i + (n-1)\ln\lambda\right)$$

Exponential family with natural parameter $-\lambda$ and sufficient statistic $\sum_{i=0}^{n-1} x_i$.

Conjugate prior It is thus again conjugate to a Gamma distribution.

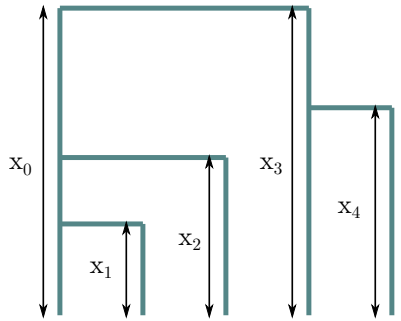
$$\lambda|\alpha,\beta \sim \Gamma(\alpha,\beta) \implies \lambda|x,\alpha,\beta \sim \Gamma\left(\alpha+n-1, \beta+\sum_{i=0}^{n-1} x_i\right)$$

In the literature I don't see much statistical work in phylogenetics based on pure birth anymore.

What about birth-death reconstructed trees then ?

Yule (pure birth) tree

- Let $\mathcal{T}$ be a tree with n leaves,
- Let x_i be the depth of leaf i,



then $\mathcal{T}$ is said to be a Yule tree if it has the following density,

$$f(\mathcal{T}|\lambda) \propto \lambda^{n-1} \prod_{i=0}^{n-1} e^{-\lambda x_i} \propto \exp\left(-\lambda \sum_{i=0}^{n-1} x_i + (n-1)\ln\lambda\right)$$

Exponential family with natural parameter $-\lambda$ and sufficient statistic $\sum_{i=0}^{n-1} x_i$.

Conjugate prior It is thus again conjugate to a Gamma distribution.

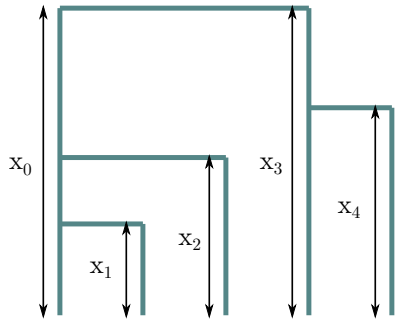
$$\lambda|\alpha,\beta \sim \Gamma(\alpha,\beta) \implies \lambda|x,\alpha,\beta \sim \Gamma\left(\alpha+n-1, \beta+\sum_{i=0}^{n-1} x_i\right)$$

In the literature I don't see much statistical work in phylogenetics based on pure birth anymore.

What about birth-death reconstructed trees then ?

Yule (pure birth) tree

- Let $\mathcal{T}$ be a tree with n leaves,
- Let x_i be the depth of leaf i,



then $\mathcal{T}$ is said to be a Yule tree if it has the following density,

$$f(\mathcal{T}|\lambda) \propto \lambda^{n-1} \prod_{i=0}^{n-1} e^{-\lambda x_i} \propto \exp\left(-\lambda \sum_{i=0}^{n-1} x_i + (n-1) \ln \lambda \right)$$

Exponential family with natural parameter $-\lambda$ and sufficient statistic $\sum_{i=0}^{n-1} x_i$.

Conjugate prior It is thus again conjugate to a Gamma distribution.

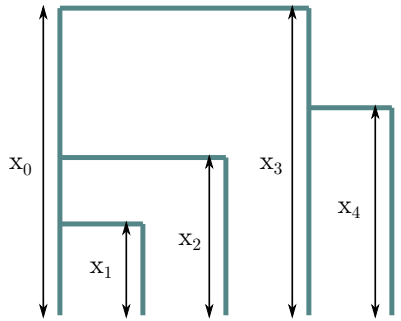
$$\lambda|\alpha,\beta \sim \Gamma(\alpha,\beta) \implies \lambda|x,\alpha,\beta \sim \Gamma\left(\alpha + n - 1, \beta + \sum_{i=0}^{n-1} x_i\right)$$

In the literature I don't see much statistical work in phylogenetics based on pure birth anymore.

What about birth-death reconstructed trees then ?

Yule (pure birth) tree

- Let $\mathcal{T}$ be a tree with n leaves,
- Let x_i be the depth of leaf i,



then $\mathcal{T}$ is said to be a Yule tree if it has the following density,

$$f(\mathcal{T}|\lambda) \propto \lambda^{n-1} \prod_{i=0}^{n-1} e^{-\lambda x_i} \propto \exp\left(-\lambda \sum_{i=0}^{n-1} x_i + (n-1) \ln \lambda \right)$$

Exponential family with natural parameter $-\lambda$ and sufficient statistic $\sum_{i=0}^{n-1} x_i$.

Conjugate prior It is thus again conjugate to a Gamma distribution.

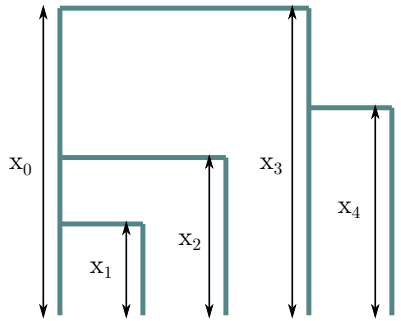
$$\lambda|\alpha,\beta \sim \Gamma(\alpha,\beta) \implies \lambda|x,\alpha,\beta \sim \Gamma\left(\alpha + n - 1, \beta + \sum_{i=0}^{n-1} x_i\right)$$

In the literature I don't see much statistical work in phylogenetics based on pure birth anymore.

What about birth-death reconstructed trees then ?

Yule (pure birth) tree

- Let $\mathcal{T}$ be a tree with n leaves,
- Let x_i be the depth of leaf i,



then $\mathcal{T}$ is said to be a Yule tree if it has the following density,

$$f(\mathcal{T}|\lambda) \propto \lambda^{n-1} \prod_{i=0}^{n-1} e^{-\lambda x_i} \propto \exp\left(-\lambda \sum_{i=0}^{n-1} x_i + (n-1) \ln \lambda \right)$$

Exponential family with natural parameter $-\lambda$ and sufficient statistic $\sum_{i=0}^{n-1} x_i$.

Conjugate prior It is thus again conjugate to a Gamma distribution.

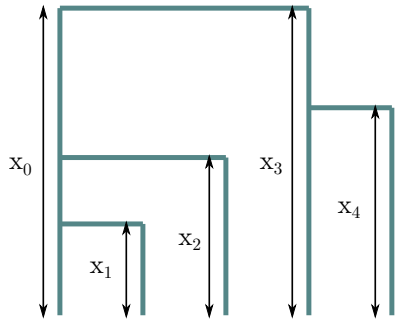
$$\lambda|\alpha,\beta \sim \Gamma(\alpha,\beta) \implies \lambda|x,\alpha,\beta \sim \Gamma\left(\alpha + n - 1, \beta + \sum_{i=0}^{n-1} x_i\right)$$

In the literature I don't see much statistical work in phylogenetics based on pure birth anymore.

What about birth-death reconstructed trees then ?

Yule (pure birth) tree

- Let $\mathcal{T}$ be a tree with n leaves,
- Let x_i be the depth of leaf i,



then $\mathcal{T}$ is said to be a Yule tree if it has the following density,

$$f(\mathcal{T}|\lambda) \propto \lambda^{n-1} \prod_{i=0}^{n-1} e^{-\lambda x_i} \propto \exp\left(-\lambda \sum_{i=0}^{n-1} x_i + (n-1) \ln \lambda \right)$$

Exponential family with natural parameter $-\lambda$ and sufficient statistic $\sum_{i=0}^{n-1} x_i$.

Conjugate prior It is thus again conjugate to a Gamma distribution.

$$\lambda|\alpha,\beta \sim \Gamma(\alpha,\beta) \implies \lambda|x,\alpha,\beta \sim \Gamma\left(\alpha + n - 1, \beta + \sum_{i=0}^{n-1} x_i\right)$$

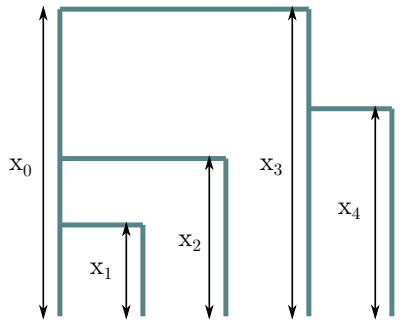
In the literature I don't see much statistical work in phylogenetics based on pure birth anymore.

What about birth-death reconstructed trees then ?

What about birth-death reconstructed trees ?

The density of the reconstructed tree is given by

$$f(\mathcal{T}|\lambda,\mu) = \lambda^{n-1}\prod_{i=0}^{n-1} p(x_i|\lambda,\mu)$$



where,

$$\begin{aligned} p(x_i|\lambda,\mu) &= \left(\frac{\lambda-\mu}{\lambda-\mu e^{-(\lambda-\mu)x_i}}\right)^2 e^{-(\lambda-\mu)x_i} \\ &= \exp\left(-(\lambda-\mu)x_i + 2\ln(\lambda-\mu) - 2\ln(\lambda-\mu e^{-(\lambda-\mu)x_i})\right) \end{aligned}$$

Not an exponential family since we cannot factorize $\eta(\theta)T(x)$ within the exponential in function p.

Conjugate prior it thus seems very optimistic to find an interesting conjugate prior.

Special cases for the critical process with $\lambda = \mu$, we have a different expression.

$$p(x_i|\lambda) = (\lambda x_i + 1)^{-2} = e^{-2\ln(\lambda x_i + 1)}$$

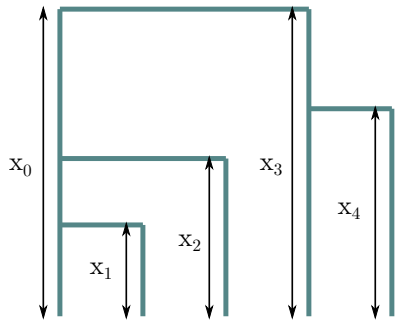
But this still does not define an exponential family.

Continuously observed process What happens if we look at the full birth-death trajectory ?

What about birth-death reconstructed trees ?

The density of the reconstructed tree is given by

$$f(\mathcal{T}|\lambda,\mu) = \lambda^{n-1}\prod_{i=0}^{n-1} p(x_i|\lambda,\mu)$$



where,

$$\begin{aligned} p(x_i|\lambda,\mu) &= \left(\frac{\lambda-\mu}{\lambda-\mu e^{-(\lambda-\mu)x_i}}\right)^2 e^{-(\lambda-\mu)x_i} \\ &= \exp\left(-(\lambda-\mu)x_i + 2\ln(\lambda-\mu) - 2\ln(\lambda-\mu e^{-(\lambda-\mu)x_i})\right) \end{aligned}$$

Not an exponential family since we cannot factorize $\eta(\theta)T(x)$ within the exponential in function p.

Conjugate prior it thus seems very optimistic to find an interesting conjugate prior.

Special cases for the critical process with $\lambda = \mu$, we have a different expression.

$$p(x_i|\lambda) = (\lambda x_i + 1)^{-2} = e^{-2\ln(\lambda x_i + 1)}$$

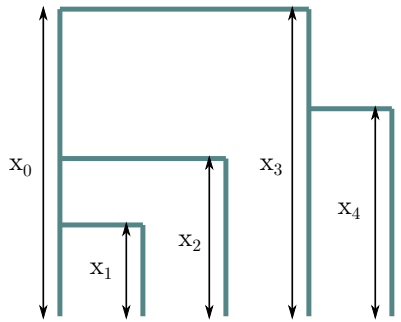
But this still does not define an exponential family.

Continuously observed process What happens if we look at the full birth-death trajectory ?

What about birth-death reconstructed trees ?

The density of the reconstructed tree is given by

$$f(\mathcal{T}|\lambda,\mu) = \lambda^{n-1}\prod_{i=0}^{n-1} p(x_i|\lambda,\mu)$$



where,

$$\begin{aligned} p(x_i|\lambda,\mu) &= \left(\frac{\lambda-\mu}{\lambda-\mu e^{-(\lambda-\mu)x_i}}\right)^2 e^{-(\lambda-\mu)x_i} \\ &= \exp\left(-(\lambda-\mu)x_i + 2\ln(\lambda-\mu) - 2\ln(\lambda-\mu e^{-(\lambda-\mu)x_i})\right) \end{aligned}$$

Not an exponential family since we cannot factorize $\eta(\theta)T(x)$ within the exponential in function p.

Conjugate prior it thus seems very optimistic to find an interesting conjugate prior.

Special cases for the critical process with $\lambda = \mu$, we have a different expression.

$$p(x_i|\lambda) = (\lambda x_i + 1)^{-2} = e^{-2\ln(\lambda x_i+1)}$$

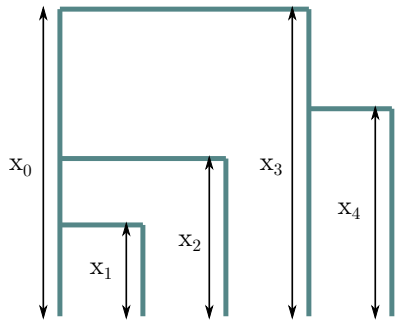
But this still does not define an exponential family.

Continuously observed process What happens if we look at the full birth-death trajectory ?

What about birth-death reconstructed trees ?

The density of the reconstructed tree is given by

$$f(\mathcal{T}|\lambda,\mu) = \lambda^{n-1}\prod_{i=0}^{n-1} p(x_i|\lambda,\mu)$$



where,

$$\begin{aligned} p(x_i|\lambda,\mu) &= \left(\frac{\lambda-\mu}{\lambda-\mu e^{-(\lambda-\mu)x_i}}\right)^2 e^{-(\lambda-\mu)x_i} \\ &= \exp\left(-(\lambda-\mu)x_i + 2\ln(\lambda-\mu) - 2\ln(\lambda-\mu e^{-(\lambda-\mu)x_i})\right) \end{aligned}$$

Not an exponential family since we cannot factorize $\eta(\theta)T(x)$ within the exponential in function p.

Conjugate prior it thus seems very optimistic to find an interesting conjugate prior.

Special cases for the critical process with $\lambda = \mu$, we have a different expression.

$$p(x_i|\lambda) = (\lambda x_i + 1)^{-2} = e^{-2\ln(\lambda x_i+1)}$$

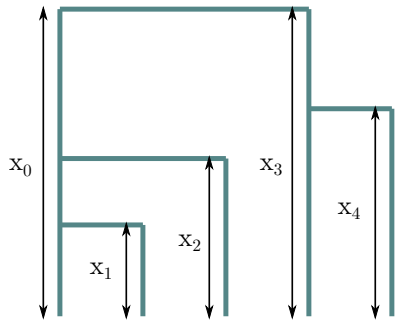
But this still does not define an exponential family.

Continuously observed process What happens if we look at the full birth-death trajectory ?

What about birth-death reconstructed trees ?

The density of the reconstructed tree is given by

$$f(\mathcal{T}|\lambda,\mu) = \lambda^{n-1}\prod_{i=0}^{n-1} p(x_i|\lambda,\mu)$$



where,

$$\begin{aligned} p(x_i|\lambda,\mu) &= \left(\frac{\lambda-\mu}{\lambda-\mu e^{-(\lambda-\mu)x_i}}\right)^2 e^{-(\lambda-\mu)x_i} \\ &= \exp\left(-(\lambda-\mu)x_i + 2\ln(\lambda-\mu) - 2\ln(\lambda-\mu e^{-(\lambda-\mu)x_i})\right) \end{aligned}$$

Not an exponential family since we cannot factorize $\eta(\theta)T(x)$ within the exponential in function p.

Conjugate prior it thus seems very optimistic to find an interesting conjugate prior.

Special cases for the critical process with $\lambda = \mu$, we have a different expression.

$$p(x_i|\lambda) = (\lambda x_i + 1)^{-2} = e^{-2\ln(\lambda x_i+1)}$$

But this still does not define an exponential family.

Continuously observed process What happens if we look at the full birth-death trajectory ?

What about birth-death reconstructed trees ?

The continuously observed process, see Crawford et al. (2018)

For any birth-death process parametrized with birth/death rates (λ_k, μ_k) in state k,

- let U_k and D_k be the total number of birth/death events from state k,
- let T_k be the total time spent in state k,

then the density of the continuously observed process is,

$$\mathbb{P}(X|(\lambda_k, \mu_k)) = \prod_{k=0}^{\infty} \lambda_k^{U_k} \mu_k^{D_k} e^{-(\lambda_k+\mu_k)T_k}$$

Exponential family with sufficient statistics (U_k, D_k, T_k).

What about birth-death reconstructed trees ?

The continuously observed process, see Crawford et al. (2018)

For any birth-death process parametrized with birth/death rates (λ_k, μ_k) in state k,

- let U_k and D_k be the total number of birth/death events from state k,
- let T_k be the total time spent in state k,

then the density of the continuously observed process is,

$$\mathbb{P}(X|(\lambda_k, \mu_k)) = \prod_{k=0}^{\infty} \lambda_k^{U_k} \mu_k^{D_k} e^{-(\lambda_k+\mu_k)T_k}$$

Exponential family with sufficient statistics (U_k, D_k, T_k).

What about birth-death reconstructed trees ?

The continuously observed process, see Crawford et al. (2018)

For the linear case it simplifies to,

$$\begin{aligned}\mathbb{P}(X|\lambda,\mu) &= \prod_{k=0}^{\infty}(k\lambda)^{U_k}(k\mu)^{D_k}e^{-k(\lambda+\mu)T_k} \\ &= \exp\left(\sum_k (U_k+D_k)lnk + \sum_k U_k \ln\lambda + \sum_k D_k \ln\mu - (\lambda+\mu)\sum_k kT_k\right) \\ &\propto \exp\left((\ln\lambda, \ln\mu, \lambda+\mu)^t(U, D, T_{\text{part}})\right)\end{aligned}$$

Exponential family with sufficient statistic (U, D, T_{part}),
where U, D are the number of birth and death events, and T_{part} is the total particle time, i.e. $\sum_k kT_k$.

Conjugate prior Assume that $\lambda, \mu \sim \Gamma(\alpha,\gamma) \otimes \Gamma(\beta,\gamma)$, i.e.

$$f(\lambda,\mu) \propto \lambda^{\alpha-1}e^{-\gamma\lambda}\mu^{\beta-1}e^{-\gamma\mu}$$

Then we get,

$$\begin{aligned}p(\lambda,\mu|X) &\propto \lambda^{\alpha-1}\mu^{\beta-1}e^{-\gamma(\lambda+\mu)}\lambda^{U}\mu^{D}e^{-(\lambda+\mu)T_{part}} \\ &\propto \lambda^{\alpha-1+U}\mu^{\beta-1+D}e^{-(\gamma+T_{part})(\lambda+\mu)}\end{aligned}$$

meaning that $\lambda, \mu|X \sim \Gamma(\alpha+U, \gamma+T_{\text{part}}) \otimes \Gamma(\beta+D, \gamma+T_{\text{part}})$.

What about birth-death reconstructed trees ?

The continuously observed process, see Crawford et al. (2018)

For the linear case it simplifies to,

$$\begin{aligned}\mathbb{P}(X|\lambda,\mu) &= \prod_{k=0}^{\infty}(k\lambda)^{U_k}(k\mu)^{D_k}e^{-k(\lambda+\mu)T_k} \\ &= \exp\left(\sum_k (U_k+D_k)lnk + \sum_k U_k \ln\lambda + \sum_k D_k \ln\mu - (\lambda+\mu)\sum_k kT_k\right) \\ &\propto \exp\left((\ln\lambda, \ln\mu, \lambda+\mu)^t(U, D, T_{\text{part}})\right)\end{aligned}$$

Exponential family with sufficient statistic (U, D, T_{part}),
where U, D are the number of birth and death events, and T_{part} is the total particle time, i.e. $\sum_k kT_k$.

Conjugate prior Assume that $\lambda, \mu \sim \Gamma(\alpha,\gamma) \otimes \Gamma(\beta,\gamma)$, i.e.

$$f(\lambda,\mu) \propto \lambda^{\alpha-1}e^{-\gamma\lambda}\mu^{\beta-1}e^{-\gamma\mu}$$

Then we get,

$$\begin{aligned}p(\lambda,\mu|X) &\propto \lambda^{\alpha-1}\mu^{\beta-1}e^{-\gamma(\lambda+\mu)}\lambda^{U}\mu^{D}e^{-(\lambda+\mu)T_{part}} \\ &\propto \lambda^{\alpha-1+U}\mu^{\beta-1+D}e^{-(\gamma+T_{part})(\lambda+\mu)}\end{aligned}$$

meaning that $\lambda, \mu|X \sim \Gamma(\alpha+U, \gamma+T_{\text{part}}) \otimes \Gamma(\beta+D, \gamma+T_{\text{part}})$.

What about birth-death reconstructed trees ?

The continuously observed process, see Crawford et al. (2018)

For the linear case it simplifies to,

$$\begin{aligned}\mathbb{P}(X|\lambda,\mu) &= \prod_{k=0}^{\infty}(k\lambda)^{U_k}(k\mu)^{D_k}e^{-k(\lambda+\mu)T_k} \\ &= \exp\left(\sum_k (U_k+D_k)lnk + \sum_k U_k \ln\lambda + \sum_k D_k \ln\mu - (\lambda+\mu)\sum_k kT_k\right) \\ &\propto \exp\left((\ln\lambda, \ln\mu, \lambda+\mu)^t(U, D, T_{\text{part}})\right)\end{aligned}$$

Exponential family with sufficient statistic (U, D, T_{part}),
where U, D are the number of birth and death events, and T_{part} is the total particle time, i.e. $\sum_k kT_k$.

Conjugate prior Assume that $\lambda, \mu \sim \Gamma(\alpha,\gamma) \otimes \Gamma(\beta,\gamma)$, i.e.

$$f(\lambda,\mu) \propto \lambda^{\alpha-1}e^{-\gamma\lambda}\mu^{\beta-1}e^{-\gamma\mu}$$

Then we get,

$$\begin{aligned}p(\lambda,\mu|X) &\propto \lambda^{\alpha-1}\mu^{\beta-1}e^{-\gamma(\lambda+\mu)}\lambda^{U}\mu^{D}e^{-(\lambda+\mu)T_{part}} \\ &\propto \lambda^{\alpha-1+U}\mu^{\beta-1+D}e^{-(\gamma+T_{part})(\lambda+\mu)}\end{aligned}$$

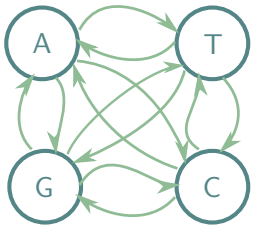
meaning that $\lambda, \mu|X \sim \Gamma(\alpha+U, \gamma+T_{\text{part}}) \otimes \Gamma(\beta+D, \gamma+T_{\text{part}})$.

Sketch of the presentation

Some basics
- Definition
- Nice properties
- A few discrete examples
- A few continuous examples

Phylodynamics modeling
- Kingman's coalescent
- Yule (pure birth) tree
- What about birth-death reconstructed trees ?

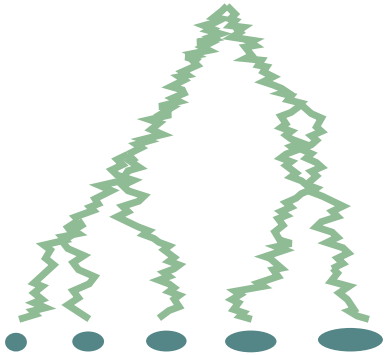
Trait evolution modeling
- Continuous trait evolution with BM
- Molecular evolution



Conclusion

Continuous trait evolution with BM

- Let $\mathcal{T}$ be a tree with n leaves,
- let $t_{k,l}$ denote the coalescence times between two leaves k, l,



then a Brownian Motion with initial (root) value μ and infinitesimal variance σ^2 has the following tip distribution,

$$(X_f) \sim \mathcal{N}_n\left(\mu V, \sigma^2 \Sigma_T\right) \quad \text{where} \quad (\Sigma_T)_{k,l} = t_{k,l} \quad \text{and} \quad V = (1, 1, ..., 1).$$

Conjugate prior We have here a Normal-Inverse Gamma conjugate prior:

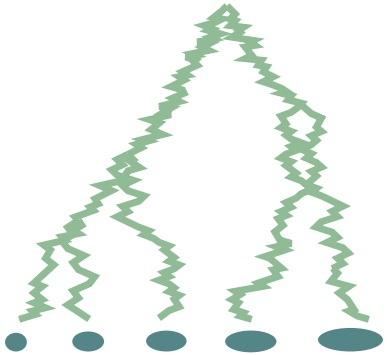
$$(\mu, \sigma^2) \sim \mathcal{N}\Gamma^{-1}(\mu_0, \lambda, \alpha, \beta)$$
$$(\mu, \sigma^2) | (X_f) \sim \mathcal{N}\Gamma^{-1}(.)$$

Ornstein-Uhlenbeck The distribution of tip values is not part of an exponential family.

In higher dimension It still holds, see for example Tolkoff et al. (2017).
With multiple traits evolving jointly along the tree, the conjugate prior to a multivariate normal distribution is called *Inverse-Wishart*.

Continuous trait evolution with BM

- Let $\mathcal{T}$ be a tree with n leaves,
- let $t_{k,l}$ denote the coalescence times between two leaves k, l,



then a Brownian Motion with initial (root) value μ and infinitesimal variance σ^2 has the following tip distribution,

$$(X_f) \sim \mathcal{N}_n\left(\mu V, \sigma^2 \Sigma_T\right) \quad \text{where} \quad (\Sigma_T)_{k,l} = t_{k,l} \quad \text{and} \quad V = (1, 1, \ldots, 1).$$

Conjugate prior We have here a Normal-Inverse Gamma conjugate prior:

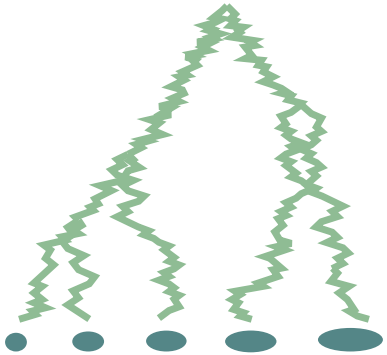
$$(\mu, \sigma^2) \sim \mathcal{N}\Gamma^{-1}(\mu_0, \lambda, \alpha, \beta)$$
$$(\mu, \sigma^2) | (X_f) \sim \mathcal{N}\Gamma^{-1}(.)$$

Ornstein-Uhlenbeck The distribution of tip values is not part of an exponential family.

In higher dimension It still holds, see for example Tolkoff et al. (2017).
With multiple traits evolving jointly along the tree, the conjugate prior to a multivariate normal distribution is called *Inverse-Wishart*.

Continuous trait evolution with BM

- Let $\mathcal{T}$ be a tree with n leaves,
- let $t_{k,l}$ denote the coalescence times between two leaves k, l,



then a Brownian Motion with initial (root) value μ and infinitesimal variance σ^2 has the following tip distribution,

$$(X_f) \sim \mathcal{N}_n\left(\mu V, \sigma^2 \Sigma_T\right) \quad \text{where} \quad (\Sigma_T)_{k,l} = t_{k,l} \quad \text{and} \quad V = (1, 1, \ldots, 1).$$

Conjugate prior We have here a Normal-Inverse Gamma conjugate prior:

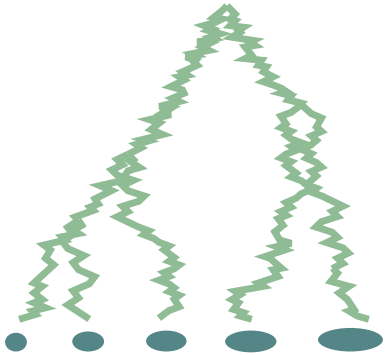
$$(\mu, \sigma^2) \sim \mathcal{N}\Gamma^{-1}(\mu_0, \lambda, \alpha, \beta)$$
$$(\mu, \sigma^2) | (X_f) \sim \mathcal{N}\Gamma^{-1}(.)$$

Ornstein-Uhlenbeck The distribution of tip values is not part of an exponential family.

In higher dimension It still holds, see for example Tolkoff et al. (2017).
With multiple traits evolving jointly along the tree, the conjugate prior to a multivariate normal distribution is called *Inverse-Wishart*.

Continuous trait evolution with BM

- Let $\mathcal{T}$ be a tree with n leaves,
- let $t_{k,l}$ denote the coalescence times between two leaves k, l,



then a Brownian Motion with initial (root) value μ and infinitesimal variance σ^2 has the following tip distribution,

$$(X_f) \sim \mathcal{N}_n\left(\mu V, \sigma^2 \Sigma_T\right) \quad \text{where} \quad (\Sigma_T)_{k,l} = t_{k,l} \quad \text{and} \quad V = (1, 1, \ldots, 1).$$

Conjugate prior We have here a Normal-Inverse Gamma conjugate prior:

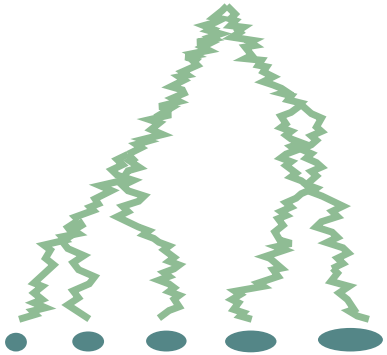
$$(\mu, \sigma^2) \sim \mathcal{N}\Gamma^{-1}(\mu_0, \lambda, \alpha, \beta)$$
$$(\mu, \sigma^2) | (X_f) \sim \mathcal{N}\Gamma^{-1}(.)$$

Ornstein-Uhlenbeck The distribution of tip values is not part of an exponential family.

In higher dimension It still holds, see for example Tolkoff et al. (2017).
With multiple traits evolving jointly along the tree, the conjugate prior to a multivariate normal distribution is called *Inverse-Wishart*.

Continuous trait evolution with BM

- Let $\mathcal{T}$ be a tree with n leaves,
- let $t_{k,l}$ denote the coalescence times between two leaves k, l,



then a Brownian Motion with initial (root) value μ and infinitesimal variance σ^2 has the following tip distribution,

$$(X_f) \sim \mathcal{N}_n\left(\mu V, \sigma^2 \Sigma_T\right) \quad \text{where} \quad (\Sigma_T)_{k,l} = t_{k,l} \quad \text{and} \quad V = (1, 1, \ldots, 1).$$

Conjugate prior We have here a Normal-Inverse Gamma conjugate prior:

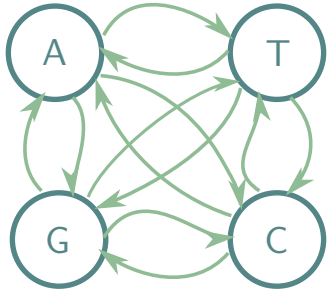
$$(\mu, \sigma^2) \sim \mathcal{N}\Gamma^{-1}(\mu_0, \lambda, \alpha, \beta)$$
$$(\mu, \sigma^2) | (X_f) \sim \mathcal{N}\Gamma^{-1}(.)$$

Ornstein-Uhlenbeck The distribution of tip values is not part of an exponential family.

In higher dimension It still holds, see for example Tolkoff et al. (2017).
With multiple traits evolving jointly along the tree, the conjugate prior to a multivariate normal distribution is called *Inverse-Wishart*.

Molecular evolution

- Consider a JC69 model with fixed transition rate α,
- and observe the state X_1 at time t_1 and X_2 at time t_2,



then the probability of the observation is,

$$\mathbb{P}(X_1, X_2 | t, \alpha) = \frac{3}{4}\left(1 - e^{-4\alpha t}\right) \mathbb{1}_{X_1 \neq X_2} + \frac{1}{4}\left(1 + 3e^{-4\alpha t}\right) \mathbb{1}_{X_1 = X_2}$$

Not an exponential family

What about the continuously observed process ? If we observe the whole trajectory $(X_t)_{t \in (t_1, t_2)}$, under any model of molecular evolution with transition rate matrix $Q = (q_{ij})$,

$$\mathbb{P}\left((X_t) | (q_{ij})\right) = \prod_{i=1}^{4} e^{-q_{ii} T_i} \prod_{j \neq i} q_{ij}^{U_{ij}}$$

which is an exponential family with sufficient statistics,

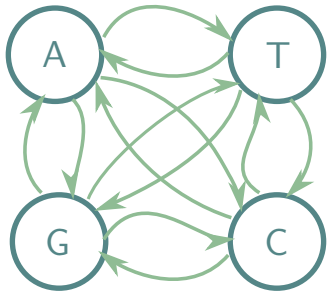
$$T_i := \int_{t_1}^{t_2} \mathbb{1}_{X_t = i} dt \quad \text{(total time spent in state } i\text{)}$$

$$U_{ij} := \text{ number of steps from } i \text{ to } j$$

With JC69, this simplifies to $e^{-3\alpha t} \alpha^U$ where U is simply the number of steps.

Molecular evolution

- Consider a JC69 model with fixed transition rate α,
- and observe the state X_1 at time t_1 and X_2 at time t_2,



then the probability of the observation is,

$$\mathbb{P}(X_1, X_2 | t, \alpha) = \frac{3}{4}\left(1 - e^{-4\alpha t}\right)\mathbb{1}_{X_1 \neq X_2} + \frac{1}{4}\left(1 + 3e^{-4\alpha t}\right)\mathbb{1}_{X_1 = X_2}$$

Not an exponential family

What about the continuously observed process ? If we observe the whole trajectory $(X_t)_{t\in(t_1,t_2)}$, under any model of molecular evolution with transition rate matrix $Q = (q_{ij})$,

$$\mathbb{P}\left((X_t)|(q_{ij})\right) = \prod_{i=1}^{4} e^{-q_{ii}T_i} \prod_{j\neq i} q_{ij}^{U_{ij}}$$

which is an exponential family with sufficient statistics,

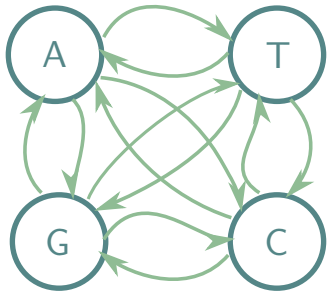
$$T_i := \int_{t_1}^{t_2} \mathbb{1}_{X_t = i} dt \quad \text{(total time spent in state } i\text{)}$$

$$U_{ij} := \text{ number of steps from } i \text{ to } j$$

With JC69, this simplifies to $e^{-3\alpha t}\alpha^U$ where U is simply the number of steps.

Molecular evolution

- Consider a JC69 model with fixed transition rate α,
- and observe the state X_1 at time t_1 and X_2 at time t_2,



then the probability of the observation is,

$$\mathbb{P}(X_1, X_2 | t, \alpha) = \frac{3}{4}\left(1 - e^{-4\alpha t}\right) \mathbb{1}_{X_1 \neq X_2} + \frac{1}{4}\left(1 + 3e^{-4\alpha t}\right) \mathbb{1}_{X_1 = X_2}$$

Not an exponential family

What about the continuously observed process ? If we observe the whole trajectory $(X_t)_{t \in (t_1, t_2)}$, under any model of molecular evolution with transition rate matrix $Q = (q_{ij})$,

$$\mathbb{P}\left((X_t) | (q_{ij})\right) = \prod_{i=1}^{4} e^{-q_{ii} T_i} \prod_{j \neq i} q_{ij}^{U_{ij}}$$

which is an exponential family with sufficient statistics,

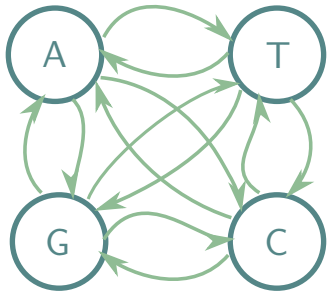
$$T_i := \int_{t_1}^{t_2} \mathbb{1}_{X_t = i} dt \quad \text{(total time spent in state } i\text{)}$$

$$U_{ij} := \text{ number of steps from } i \text{ to } j$$

With JC69, this simplifies to $e^{-3\alpha t} \alpha^U$ where U is simply the number of steps.

Molecular evolution

- Consider a JC69 model with fixed transition rate α,
- and observe the state X_1 at time t_1 and X_2 at time t_2,



then the probability of the observation is,

$$\mathbb{P}(X_1, X_2 | t, \alpha) = \frac{3}{4}\left(1 - e^{-4\alpha t}\right) \mathbb{1}_{X_1 \neq X_2} + \frac{1}{4}\left(1 + 3e^{-4\alpha t}\right) \mathbb{1}_{X_1 = X_2}$$

Not an exponential family

What about the continuously observed process ? If we observe the whole trajectory $(X_t)_{t \in (t_1, t_2)}$, under any model of molecular evolution with transition rate matrix $Q = (q_{ij})$,

$$\mathbb{P}\left((X_t) | (q_{ij})\right) = \prod_{i=1}^{4} e^{-q_{ii} T_i} \prod_{j \neq i} q_{ij}^{U_{ij}}$$

which is an exponential family with sufficient statistics,

$$T_i := \int_{t_1}^{t_2} \mathbb{1}_{X_t = i} dt \quad \text{(total time spent in state } i\text{)}$$

$$U_{ij} := \text{ number of steps from } i \text{ to } j$$

With JC69, this simplifies to $e^{-3\alpha t} \alpha^U$ where U is simply the number of steps.

Molecular evolution

The two most popular ways of building a MCMC to sample a target distribution ν

Metropolis-Hastings

Algorithm Initialize a first state x_0.
At step i, the chain being in state x_i,

1. Propose a next state y_{i+1} by drawing a realisation in distribution $q(x_i, \cdot)$.
2. Compute the ratio:
$$r(x_i, y_{i+1}) := \frac{\nu(y_{i+1})q(y_{i+1}, x_i)}{\nu(x_i)q(x_i, y_{i+1})}$$
3. Draw $u \sim \mathcal{U}(0, 1)$.
If $u \leq r$, set $x_{i+1} := y_{i+1}$.
otherwise, keep $x_{i+1} := x_i$.

Reversibility One can check that $\nu_x q_{xy} \min(1, r(x, y)) = \nu_y q_{yx} \min(1, r(y, x))$.
Hence, it converges to the stationary distribution ν.

Advantage One can use (almost) any proposal distribution q.

Drawback One needs to carefully tune q to ensure fast convergence.

Gibbs sampler

Algorithm First, initialize the chain in state x_0.
At step n, $x_n = (x_n^{(1)}, x_n^{(2)}, \ldots, x_n^{(k)})$,

1. Draw $i \sim \mathcal{U}\{1, 2, \ldots, k\}$.
2. Draw $x_{n+1}^{(i)}$ in the conditional law
$$p\left(X^{(i)} \mid \left(X^{(j)}\right)_{j \neq i} = \left(x_n^{(j)}\right)_{j \neq i}\right).$$
3. Fix $x_{n+1} := \left(x_n^{(1)}, x_n^{(2)}, \ldots, x_{n+1}^{(i)}, \ldots, x_n^{(k)}\right)$.

Reversibility One can check that $\nu_x p_{xy} = \nu_y p_{yx}$.
Hence, it converges to the stationary distribution ν.

Advantage It is generally assumed that it converges faster.

Drawback One needs to know how to sample step 2.

Molecular evolution

The two most popular ways of building a MCMC to sample a target distribution ν

Metropolis-Hastings

Algorithm Initialize a first state x_0.
At step i, the chain being in state x_i,

1. Propose a next state y_{i+1} by drawing a realisation in distribution $q(x_i, \cdot)$.
2. Compute the ratio:
$$r(x_i, y_{i+1}) := \frac{\nu(y_{i+1})q(y_{i+1}, x_i)}{\nu(x_i)q(x_i, y_{i+1})}$$
3. Draw $u \sim \mathcal{U}(0,1)$.
If $u \leq r$, set $x_{i+1} := y_{i+1}$.
otherwise, keep $x_{i+1} := x_i$.

Reversibility One can check that $\nu_x q_{xy} \min(1, r(x,y)) = \nu_y q_{yx} \min(1, r(y,x))$.
Hence, it converges to the stationary distribution ν.

Advantage One can use (almost) any proposal distribution q.

Drawback One needs to carefully tune q to ensure fast convergence.

Gibbs sampler

Algorithm First, initialize the chain in state x_0.
At step n, $x_n = (x_n^{(1)}, x_n^{(2)}, \ldots, x_n^{(k)})$,

1. Draw $i \sim \mathcal{U}\{1, 2, \ldots, k\}$.
2. Draw $x_{n+1}^{(i)}$ in the conditional law
$$p\left(X^{(i)} \mid \left(X^{(j)}\right)_{j \neq i} = \left(x_n^{(j)}\right)_{j \neq i}\right).$$
3. Fix $x_{n+1} := \left(x_n^{(1)}, x_n^{(2)}, \ldots, x_{n+1}^{(i)}, \ldots, x_n^{(k)}\right)$.

Reversibility One can check that $\nu_x p_{xy} = \nu_y p_{yx}$.
Hence, it converges to the stationary distribution ν.

Advantage It is generally assumed that it converges faster.

Drawback One needs to know how to sample step 2.

Molecular evolution

The two most popular ways of building a MCMC to sample a target distribution ν

Metropolis-Hastings

Algorithm Initialize a first state x_0.
At step i, the chain being in state x_i,

1. Propose a next state y_{i+1} by drawing a realisation in distribution $q(x_i, \cdot)$.
2. Compute the ratio:
$$r(x_i, y_{i+1}) := \frac{\nu(y_{i+1})q(y_{i+1}, x_i)}{\nu(x_i)q(x_i, y_{i+1})}$$
3. Draw $u \sim \mathcal{U}(0,1)$.
If $u \leq r$, set $x_{i+1} := y_{i+1}$.
otherwise, keep $x_{i+1} := x_i$.

Reversibility One can check that $\nu_x q_{xy} \min(1, r(x,y)) = \nu_y q_{yx} \min(1, r(y,x))$.
Hence, it converges to the stationary distribution ν.

Advantage One can use (almost) any proposal distribution q.

Drawback One needs to carefully tune q to ensure fast convergence.

Gibbs sampler

Algorithm First, initialize the chain in state x_0.
At step n, $x_n = (x_n^{(1)}, x_n^{(2)}, \ldots, x_n^{(k)})$,

1. Draw $i \sim \mathcal{U}\{1, 2, \ldots, k\}$.
2. Draw $x_{n+1}^{(i)}$ in the conditional law
$$p\left(X^{(i)} \mid \left(X^{(j)}\right)_{j \neq i} = \left(x_n^{(j)}\right)_{j \neq i}\right).$$
3. Fix $x_{n+1} := \left(x_n^{(1)}, x_n^{(2)}, \ldots, x_{n+1}^{(i)}, \ldots, x_n^{(k)}\right)$.

Reversibility One can check that $\nu_x p_{xy} = \nu_y p_{yx}$.
Hence, it converges to the stationary distribution ν.

Advantage It is generally assumed that it converges faster.

Drawback One needs to know how to sample step 2.

Molecular evolution

The two most popular ways of building a MCMC to sample a target distribution ν

Metropolis-Hastings

Algorithm Initialize a first state x_0.
At step i, the chain being in state x_i,

1. Propose a next state y_{i+1} by drawing a realisation in distribution $q(x_i, \cdot)$.
2. Compute the ratio:
$$r(x_i, y_{i+1}) := \frac{\nu(y_{i+1})q(y_{i+1}, x_i)}{\nu(x_i)q(x_i, y_{i+1})}$$
3. Draw $u \sim \mathcal{U}(0,1)$.
If $u \leq r$, set $x_{i+1} := y_{i+1}$.
otherwise, keep $x_{i+1} := x_i$.

Reversibility One can check that $\nu_x q_{xy} \min(1, r(x,y)) = \nu_y q_{yx} \min(1, r(y,x))$.
Hence, it converges to the stationary distribution ν.

Advantage One can use (almost) any proposal distribution q.

Drawback One needs to carefully tune q to ensure fast convergence.

Gibbs sampler

Algorithm First, initialize the chain in state x_0.
At step n, $x_n = (x_n^{(1)}, x_n^{(2)}, \ldots, x_n^{(k)})$,

1. Draw $i \sim \mathcal{U}\{1, 2, \ldots, k\}$.
2. Draw $x_{n+1}^{(i)}$ in the conditional law
$$p\left(X^{(i)} \mid \left(X^{(j)}\right)_{j \neq i} = \left(x_n^{(j)}\right)_{j \neq i}\right).$$
3. Fix $x_{n+1} := \left(x_n^{(1)}, x_n^{(2)}, \ldots, x_{n+1}^{(i)}, \ldots, x_n^{(k)}\right)$.

Reversibility One can check that $\nu_x p_{xy} = \nu_y p_{yx}$.
Hence, it converges to the stationary distribution ν.

Advantage It is generally assumed that it converges faster.

Drawback One needs to know how to sample step 2.

Molecular evolution

The two most popular ways of building a MCMC to sample a target distribution ν

Metropolis-Hastings

Algorithm Initialize a first state x_0.
At step i, the chain being in state x_i,

1. Propose a next state y_{i+1} by drawing a realisation in distribution $q(x_i, \cdot)$.
2. Compute the ratio:
$$r(x_i, y_{i+1}) := \frac{\nu(y_{i+1})q(y_{i+1}, x_i)}{\nu(x_i)q(x_i, y_{i+1})}$$
3. Draw $u \sim \mathcal{U}(0,1)$.
If $u \leq r$, set $x_{i+1} := y_{i+1}$.
otherwise, keep $x_{i+1} := x_i$.

Reversibility One can check that $\nu_x q_{xy} \min(1, r(x,y)) = \nu_y q_{yx} \min(1, r(y,x))$.
Hence, it converges to the stationary distribution ν.

Advantage One can use (almost) any proposal distribution q.

Drawback One needs to carefully tune q to ensure fast convergence.

Gibbs sampler

Algorithm First, initialize the chain in state x_0.
At step n, $x_n = (x_n^{(1)}, x_n^{(2)}, \ldots, x_n^{(k)})$,

1. Draw $i \sim \mathcal{U}\{1, 2, \ldots, k\}$.
2. Draw $x_{n+1}^{(i)}$ in the conditional law
$$p\left(X^{(i)} \mid \left(X^{(j)}\right)_{j \neq i} = \left(x_n^{(j)}\right)_{j \neq i}\right).$$
3. Fix $x_{n+1} := \left(x_n^{(1)}, x_n^{(2)}, \ldots, x_{n+1}^{(i)}, \ldots, x_n^{(k)}\right)$.

Reversibility One can check that $\nu_x p_{xy} = \nu_y p_{yx}$.
Hence, it converges to the stationary distribution ν.

Advantage It is generally assumed that it converges faster.

Drawback One needs to know how to sample step 2.

Molecular evolution

The two most popular ways of building a MCMC to sample a target distribution ν

Metropolis-Hastings

Algorithm Initialize a first state x_0.
At step i, the chain being in state x_i,

1. Propose a next state y_{i+1} by drawing a realisation in distribution $q(x_i, \cdot)$.
2. Compute the ratio:
$$r(x_i, y_{i+1}) := \frac{\nu(y_{i+1})q(y_{i+1}, x_i)}{\nu(x_i)q(x_i, y_{i+1})}$$
3. Draw $u \sim \mathcal{U}(0, 1)$.
If $u \leq r$, set $x_{i+1} := y_{i+1}$.
otherwise, keep $x_{i+1} := x_i$.

Reversibility One can check that $\nu_x q_{xy} \min(1, r(x, y)) = \nu_y q_{yx} \min(1, r(y, x))$.
Hence, it converges to the stationary distribution ν.

Advantage One can use (almost) any proposal distribution q.

Drawback One needs to carefully tune q to ensure fast convergence.

Gibbs sampler

Algorithm First, initialize the chain in state x_0.
At step n, $x_n = (x_n^{(1)}, x_n^{(2)}, \ldots, x_n^{(k)})$,

1. Draw $i \sim \mathcal{U}\{1, 2, \ldots, k\}$.
2. Draw $x_{n+1}^{(i)}$ in the conditional law
$$p\left(X^{(i)} \mid \left(X^{(j)}\right)_{j \neq i} = \left(x_n^{(j)}\right)_{j \neq i}\right).$$
3. Fix $x_{n+1} := \left(x_n^{(1)}, x_n^{(2)}, \ldots, x_{n+1}^{(i)}, \ldots, x_n^{(k)}\right)$.

Reversibility One can check that $\nu_x p_{xy} = \nu_y p_{yx}$.
Hence, it converges to the stationary distribution ν.

Advantage It is generally assumed that it converges faster.

Drawback One needs to know how to sample step 2.

Molecular evolution

The two most popular ways of building a MCMC to sample a target distribution ν

Metropolis-Hastings

Algorithm Initialize a first state x_0.
At step i, the chain being in state x_i,

1. Propose a next state y_{i+1} by drawing a realisation in distribution $q(x_i, \cdot)$.
2. Compute the ratio:
$$r(x_i, y_{i+1}) := \frac{\nu(y_{i+1})q(y_{i+1}, x_i)}{\nu(x_i)q(x_i, y_{i+1})}$$
3. Draw $u \sim \mathcal{U}(0, 1)$.
If $u \leq r$, set $x_{i+1} := y_{i+1}$.
otherwise, keep $x_{i+1} := x_i$.

Reversibility One can check that $\nu_x q_{xy} \min(1, r(x, y)) = \nu_y q_{yx} \min(1, r(y, x))$.
Hence, it converges to the stationary distribution ν.

Advantage One can use (almost) any proposal distribution q.

Drawback One needs to carefully tune q to ensure fast convergence.

Gibbs sampler

Algorithm First, initialize the chain in state x_0.
At step n, $x_n = (x_n^{(1)}, x_n^{(2)}, \ldots, x_n^{(k)})$,

1. Draw $i \sim \mathcal{U}\{1, 2, \ldots, k\}$.
2. Draw $x_{n+1}^{(i)}$ in the conditional law
$$p\left(X^{(i)} \mid \left(X^{(j)}\right)_{j \neq i} = \left(x_n^{(j)}\right)_{j \neq i}\right).$$
3. Fix $x_{n+1} := \left(x_n^{(1)}, x_n^{(2)}, \ldots, x_{n+1}^{(i)}, \ldots, x_n^{(k)}\right)$.

Reversibility One can check that $\nu_x p_{xy} = \nu_y p_{yx}$.
Hence, it converges to the stationary distribution ν.

Advantage It is generally assumed that it converges faster.

Drawback One needs to know how to sample step 2.

Molecular evolution

The two most popular ways of building a MCMC to sample a target distribution ν

Metropolis-Hastings

Algorithm Initialize a first state x_0.
At step i, the chain being in state x_i,

1. Propose a next state y_{i+1} by drawing a realisation in distribution $q(x_i, \cdot)$.
2. Compute the ratio:
$$r(x_i, y_{i+1}) := \frac{\nu(y_{i+1})q(y_{i+1}, x_i)}{\nu(x_i)q(x_i, y_{i+1})}$$
3. Draw $u \sim \mathcal{U}(0,1)$.
If $u \leq r$, set $x_{i+1} := y_{i+1}$.
otherwise, keep $x_{i+1} := x_i$.

Reversibility One can check that $\nu_x q_{xy} \min(1, r(x,y)) = \nu_y q_{yx} \min(1, r(y,x))$.
Hence, it converges to the stationary distribution ν.

Advantage One can use (almost) any proposal distribution q.

Drawback One needs to carefully tune q to ensure fast convergence.

Gibbs sampler

Algorithm First, initialize the chain in state x_0.
At step n, $x_n = (x_n^{(1)}, x_n^{(2)}, \ldots, x_n^{(k)})$,

1. Draw $i \sim \mathcal{U}\{1, 2, \ldots, k\}$.
2. Draw $x_{n+1}^{(i)}$ in the conditional law
$$p\left(X^{(i)} \mid \left(X^{(j)}\right)_{j \neq i} = \left(x_n^{(j)}\right)_{j \neq i}\right).$$
3. Fix $x_{n+1} := \left(x_n^{(1)}, x_n^{(2)}, \ldots, x_{n+1}^{(i)}, \ldots, x_n^{(k)}\right)$.

Reversibility One can check that $\nu_x p_{xy} = \nu_y p_{yx}$.
Hence, it converges to the stationary distribution ν.

Advantage It is generally assumed that it converges faster.

Drawback One needs to know how to sample step 2.

Molecular evolution

The two most popular ways of building a MCMC to sample a target distribution ν

Metropolis-Hastings

Algorithm Initialize a first state x_0.
At step i, the chain being in state x_i,

1. Propose a next state y_{i+1} by drawing a realisation in distribution $q(x_i, \cdot)$.
2. Compute the ratio:
$$r(x_i, y_{i+1}) := \frac{\nu(y_{i+1})q(y_{i+1}, x_i)}{\nu(x_i)q(x_i, y_{i+1})}$$
3. Draw $u \sim \mathcal{U}(0,1)$.
If $u \leq r$, set $x_{i+1} := y_{i+1}$.
otherwise, keep $x_{i+1} := x_i$.

Reversibility One can check that $\nu_x q_{xy} \min(1, r(x,y)) = \nu_y q_{yx} \min(1, r(y,x))$.
Hence, it converges to the stationary distribution ν.

Advantage One can use (almost) any proposal distribution q.

Drawback One needs to carefully tune q to ensure fast convergence.

Gibbs sampler

Algorithm First, initialize the chain in state x_0.
At step n, $x_n = (x_n^{(1)}, x_n^{(2)}, \ldots, x_n^{(k)})$,

1. Draw $i \sim \mathcal{U}\{1, 2, \ldots, k\}$.
2. Draw $x_{n+1}^{(i)}$ in the conditional law
$$p\left(X^{(i)} \mid \left(X^{(j)}\right)_{j \neq i} = \left(x_n^{(j)}\right)_{j \neq i}\right).$$
3. Fix $x_{n+1} := \left(x_n^{(1)}, x_n^{(2)}, \ldots, x_{n+1}^{(i)}, \ldots, x_n^{(k)}\right)$.

Reversibility One can check that $\nu_x p_{xy} = \nu_y p_{yx}$.
Hence, it converges to the stationary distribution ν.

Advantage It is generally assumed that it converges faster.

Drawback One needs to know how to sample step 2.

Molecular evolution

Example from the literature: Lartillot (2006)

- A nice illustration of Gibbs sampling using conjugacy properties and data augmentation,
- Comparison to an alternative MH-MCMC sampler.

1. along each branch j and at any site i, sample n_{ij} the total number of substitutions, (t_{ij}^k) the times at which substitutions occur, and (σ_{ij}^k) the successive states.
2. sampling the branch-length l given everything else.
   The prior is Gamma, conjugate to a Poisson variable.
   Posterior is a Gamma again with known parameters.
3. sampling the site-specific rates r given everything else.
   The prior is Gamma, conjugate to a Poisson variable.
   Posterior is Gamma again with known parameters.
4. sampling the stationary profile π given everything else.
   The prior is a Dirichlet, conjugate to a multinomial distribution.
   Posterior is Dirichlet again with known parameters.
5. update the hyperparameters with a MH step.

- each step requires a costly data-augmentation step,
- but decorrelation time with Gibbs-MCMC is one order of magnitude smaller than with MH-MCMC.

Molecular evolution

Example from the literature: Lartillot (2006)

- A nice illustration of Gibbs sampling using conjugacy properties and data augmentation,
- Comparison to an alternative MH-MCMC sampler.

1. along each branch j and at any site i, sample n_{ij} the total number of substitutions, (t_{ij}^k) the times at which substitutions occur, and (σ_{ij}^k) the successive states.
2. sampling the branch-length l given everything else.
   The prior is Gamma, conjugate to a Poisson variable.
   Posterior is a Gamma again with known parameters.
3. sampling the site-specific rates r given everything else.
   The prior is Gamma, conjugate to a Poisson variable.
   Posterior is Gamma again with known parameters.
4. sampling the stationary profile π given everything else.
   The prior is a Dirichlet, conjugate to a multinomial distribution.
   Posterior is Dirichlet again with known parameters.
5. update the hyperparameters with a MH step.

- each step requires a costly data-augmentation step,
- but decorrelation time with Gibbs-MCMC is one order of magnitude smaller than with MH-MCMC.

Molecular evolution

Example from the literature: Lartillot (2006)

- A nice illustration of Gibbs sampling using conjugacy properties and data augmentation,
- Comparison to an alternative MH-MCMC sampler.

1. along each branch j and at any site i, sample n_{ij} the total number of substitutions, (t_{ij}^k) the times at which substitutions occur, and (σ_{ij}^k) the successive states.
2. sampling the branch-length l given everything else.
   The prior is Gamma, conjugate to a Poisson variable.
   Posterior is a Gamma again with known parameters.
3. sampling the site-specific rates r given everything else.
   The prior is Gamma, conjugate to a Poisson variable.
   Posterior is Gamma again with known parameters.
4. sampling the stationary profile π given everything else.
   The prior is a Dirichlet, conjugate to a multinomial distribution.
   Posterior is Dirichlet again with known parameters.
5. update the hyperparameters with a MH step.

- each step requires a costly data-augmentation step,
- but decorrelation time with Gibbs-MCMC is one order of magnitude smaller than with MH-MCMC.

Molecular evolution

Example from the literature: Lartillot (2006)

- A nice illustration of Gibbs sampling using conjugacy properties and data augmentation,
- Comparison to an alternative MH-MCMC sampler.

1. along each branch j and at any site i, sample n_{ij} the total number of substitutions, (t_{ij}^k) the times at which substitutions occur, and (σ_{ij}^k) the successive states.
2. sampling the branch-length l given everything else.
   The prior is Gamma, conjugate to a Poisson variable.
   Posterior is a Gamma again with known parameters.
3. sampling the site-specific rates r given everything else.
   The prior is Gamma, conjugate to a Poisson variable.
   Posterior is Gamma again with known parameters.
4. sampling the stationary profile π given everything else.
   The prior is a Dirichlet, conjugate to a multinomial distribution.
   Posterior is Dirichlet again with known parameters.
5. update the hyperparameters with a MH step.

- each step requires a costly data-augmentation step,
- but decorrelation time with Gibbs-MCMC is one order of magnitude smaller than with MH-MCMC.

Molecular evolution

Example from the literature: Lartillot (2006)

- A nice illustration of Gibbs sampling using conjugacy properties and data augmentation,
- Comparison to an alternative MH-MCMC sampler.

1. along each branch j and at any site i, sample n_{ij} the total number of substitutions, (t_{ij}^k) the times at which substitutions occur, and (σ_{ij}^k) the successive states.
2. sampling the branch-length l given everything else.
   The prior is Gamma, conjugate to a Poisson variable.
   Posterior is a Gamma again with known parameters.
3. sampling the site-specific rates r given everything else.
   The prior is Gamma, conjugate to a Poisson variable.
   Posterior is Gamma again with known parameters.
4. sampling the stationary profile π given everything else.
   The prior is a Dirichlet, conjugate to a multinomial distribution.
   Posterior is Dirichlet again with known parameters.
5. update the hyperparameters with a MH step.

- each step requires a costly data-augmentation step,
- but decorrelation time with Gibbs-MCMC is one order of magnitude smaller than with MH-MCMC.

Molecular evolution

Example from the literature: Lartillot (2006)

- A nice illustration of Gibbs sampling using conjugacy properties and data augmentation,
- Comparison to an alternative MH-MCMC sampler.

1. along each branch j and at any site i, sample n_{ij} the total number of substitutions, (t_{ij}^k) the times at which substitutions occur, and (σ_{ij}^k) the successive states.
2. sampling the branch-length l given everything else.
   The prior is Gamma, conjugate to a Poisson variable.
   Posterior is a Gamma again with known parameters.
3. sampling the site-specific rates r given everything else.
   The prior is Gamma, conjugate to a Poisson variable.
   Posterior is Gamma again with known parameters.
4. sampling the stationary profile π given everything else.
   The prior is a Dirichlet, conjugate to a multinomial distribution.
   Posterior is Dirichlet again with known parameters.
5. update the hyperparameters with a MH step.

- each step requires a costly data-augmentation step,
- but decorrelation time with Gibbs-MCMC is one order of magnitude smaller than with MH-MCMC.

Molecular evolution

Example from the literature: Lartillot (2006)

- A nice illustration of Gibbs sampling using conjugacy properties and data augmentation,
- Comparison to an alternative MH-MCMC sampler.

1. along each branch j and at any site i, sample n_{ij} the total number of substitutions, (t_{ij}^k) the times at which substitutions occur, and (σ_{ij}^k) the successive states.
2. sampling the branch-length l given everything else.
   The prior is Gamma, conjugate to a Poisson variable.
   Posterior is a Gamma again with known parameters.
3. sampling the site-specific rates r given everything else.
   The prior is Gamma, conjugate to a Poisson variable.
   Posterior is Gamma again with known parameters.
4. sampling the stationary profile π given everything else.
   The prior is a Dirichlet, conjugate to a multinomial distribution.
   Posterior is Dirichlet again with known parameters.
5. update the hyperparameters with a MH step.

- each step requires a costly data-augmentation step,
- but decorrelation time with Gibbs-MCMC is one order of magnitude smaller than with MH-MCMC.

Molecular evolution

Example from the literature: Lartillot (2006)

- A nice illustration of Gibbs sampling using conjugacy properties and data augmentation,
- Comparison to an alternative MH-MCMC sampler.

1. along each branch j and at any site i, sample n_{ij} the total number of substitutions, (t_{ij}^k) the times at which substitutions occur, and (σ_{ij}^k) the successive states.
2. sampling the branch-length l given everything else.
   The prior is Gamma, conjugate to a Poisson variable.
   Posterior is a Gamma again with known parameters.
3. sampling the site-specific rates r given everything else.
   The prior is Gamma, conjugate to a Poisson variable.
   Posterior is Gamma again with known parameters.
4. sampling the stationary profile π given everything else.
   The prior is a Dirichlet, conjugate to a multinomial distribution.
   Posterior is Dirichlet again with known parameters.
5. update the hyperparameters with a MH step.

- each step requires a costly data-augmentation step,
- but decorrelation time with Gibbs-MCMC is one order of magnitude smaller than with MH-MCMC.

Molecular evolution

Example from the literature: Lartillot (2006)

- A nice illustration of Gibbs sampling using conjugacy properties and data augmentation,
- Comparison to an alternative MH-MCMC sampler.

1. along each branch j and at any site i, sample n_{ij} the total number of substitutions, (t_{ij}^k) the times at which substitutions occur, and (σ_{ij}^k) the successive states.
2. sampling the branch-length l given everything else.
   The prior is Gamma, conjugate to a Poisson variable.
   Posterior is a Gamma again with known parameters.
3. sampling the site-specific rates r given everything else.
   The prior is Gamma, conjugate to a Poisson variable.
   Posterior is Gamma again with known parameters.
4. sampling the stationary profile π given everything else.
   The prior is a Dirichlet, conjugate to a multinomial distribution.
   Posterior is Dirichlet again with known parameters.
5. update the hyperparameters with a MH step.

- each step requires a costly data-augmentation step,
- but decorrelation time with Gibbs-MCMC is one order of magnitude smaller than with MH-MCMC.

Conclusion

Take-home messages

Exponential families are beloved by statisticians, for good reasons.

In phylogenetics we might want to think more about them.

Among models of phylodynamics Kingman's coalescents, Yule trees, the continuously observed birth-death process, are exponential families.

Among models of trait evolution any continuously observed discrete space Markov process, BM, are exponential families.

Data augmentation with Gibbs sampling could represent a promising alternative to MH-MCMC.

A few relevant papers can be found here:

Crawford, F. W., Ho, L. S. T., and Suchard, M. A. (2018). Computational methods for birth-death processes. *Wiley Interdisciplinary Reviews: Computational Statistics*, 10(2):–1423.

Lartillot, N. (2006). Conjugate Gibbs sampling for Bayesian phylogenetic models. *Journal of computational biology*.

Parag, K. V., Pybus, O. G., and Wu, C.-H. (2020). Are skyline plot-based demographic estimates overly dependent on smoothing prior assumptions? *bioRxiv*.

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

Thank you for your attention !

Conclusion

Take-home messages

Exponential families are beloved by statisticians, for good reasons.

In phylogenetics we might want to think more about them.

Among models of phylodynamics Kingman's coalescents, Yule trees, the continuously observed birth-death process, are exponential families.

Among models of trait evolution any continuously observed discrete space Markov process, BM, are exponential families.

Data augmentation with Gibbs sampling could represent a promising alternative to MH-MCMC.

A few relevant papers can be found here:

Crawford, F. W., Ho, L. S. T., and Suchard, M. A. (2018). Computational methods for birth-death processes. *Wiley Interdisciplinary Reviews: Computational Statistics*, 10(2):–1423.

Lartillot, N. (2006). Conjugate Gibbs sampling for Bayesian phylogenetic models. *Journal of computational biology*.

Parag, K. V., Pybus, O. G., and Wu, C.-H. (2020). Are skyline plot-based demographic estimates overly dependent on smoothing prior assumptions? *bioRxiv*.

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

Thank you for your attention !

Conclusion

Take-home messages

Exponential families are beloved by statisticians, for good reasons.

In phylogenetics we might want to think more about them.

Among models of phylodynamics Kingman's coalescents, Yule trees, the continuously observed birth-death process, are exponential families.

Among models of trait evolution any continuously observed discrete space Markov process, BM, are exponential families.

Data augmentation with Gibbs sampling could represent a promising alternative to MH-MCMC.

A few relevant papers can be found here:

Crawford, F. W., Ho, L. S. T., and Suchard, M. A. (2018). Computational methods for birth-death processes. *Wiley Interdisciplinary Reviews: Computational Statistics*, 10(2):–1423.

Lartillot, N. (2006). Conjugate Gibbs sampling for Bayesian phylogenetic models. *Journal of computational biology*.

Parag, K. V., Pybus, O. G., and Wu, C.-H. (2020). Are skyline plot-based demographic estimates overly dependent on smoothing prior assumptions? *bioRxiv*.

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

Thank you for your attention !

Conclusion

Take-home messages

Exponential families are beloved by statisticians, for good reasons.

In phylogenetics we might want to think more about them.

Among models of phylodynamics Kingman's coalescents, Yule trees, the continuously observed birth-death process, are exponential families.

Among models of trait evolution any continuously observed discrete space Markov process, BM, are exponential families.

Data augmentation with Gibbs sampling could represent a promising alternative to MH-MCMC.

A few relevant papers can be found here:

Crawford, F. W., Ho, L. S. T., and Suchard, M. A. (2018). Computational methods for birth-death processes. *Wiley Interdisciplinary Reviews: Computational Statistics*, 10(2):–1423.

Lartillot, N. (2006). Conjugate Gibbs sampling for Bayesian phylogenetic models. *Journal of computational biology*.

Parag, K. V., Pybus, O. G., and Wu, C.-H. (2020). Are skyline plot-based demographic estimates overly dependent on smoothing prior assumptions? *bioRxiv*.

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

Thank you for your attention !

Conclusion

Take-home messages

Exponential families are beloved by statisticians, for good reasons.

In phylogenetics we might want to think more about them.

Among models of phylodynamics Kingman's coalescents, Yule trees, the continuously observed birth-death process, are exponential families.

Among models of trait evolution any continuously observed discrete space Markov process, BM, are exponential families.

Data augmentation with Gibbs sampling could represent a promising alternative to MH-MCMC.

A few relevant papers can be found here:

Crawford, F. W., Ho, L. S. T., and Suchard, M. A. (2018). Computational methods for birth-death processes. *Wiley Interdisciplinary Reviews: Computational Statistics*, 10(2):–1423.

Lartillot, N. (2006). Conjugate Gibbs sampling for Bayesian phylogenetic models. *Journal of computational biology*.

Parag, K. V., Pybus, O. G., and Wu, C.-H. (2020). Are skyline plot-based demographic estimates overly dependent on smoothing prior assumptions? *bioRxiv*.

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

Thank you for your attention !

Conclusion

Take-home messages

Exponential families are beloved by statisticians, for good reasons.

In phylogenetics we might want to think more about them.

Among models of phylodynamics Kingman's coalescents, Yule trees, the continuously observed birth-death process, are exponential families.

Among models of trait evolution any continuously observed discrete space Markov process, BM, are exponential families.

Data augmentation with Gibbs sampling could represent a promising alternative to MH-MCMC.

A few relevant papers can be found here:

Crawford, F. W., Ho, L. S. T., and Suchard, M. A. (2018). Computational methods for birth-death processes. *Wiley Interdisciplinary Reviews: Computational Statistics*, 10(2):–1423.

Lartillot, N. (2006). Conjugate Gibbs sampling for Bayesian phylogenetic models. *Journal of computational biology*.

Parag, K. V., Pybus, O. G., and Wu, C.-H. (2020). Are skyline plot-based demographic estimates overly dependent on smoothing prior assumptions? *bioRxiv*.

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

Thank you for your attention !

Conclusion

Take-home messages

Exponential families are beloved by statisticians, for good reasons.

In phylogenetics we might want to think more about them.

Among models of phylodynamics Kingman's coalescents, Yule trees, the continuously observed birth-death process, are exponential families.

Among models of trait evolution any continuously observed discrete space Markov process, BM, are exponential families.

Data augmentation with Gibbs sampling could represent a promising alternative to MH-MCMC.

A few relevant papers can be found here:

Crawford, F. W., Ho, L. S. T., and Suchard, M. A. (2018). Computational methods for birth-death processes. *Wiley Interdisciplinary Reviews: Computational Statistics*, 10(2):–1423.

Lartillot, N. (2006). Conjugate Gibbs sampling for Bayesian phylogenetic models. *Journal of computational biology*.

Parag, K. V., Pybus, O. G., and Wu, C.-H. (2020). Are skyline plot-based demographic estimates overly dependent on smoothing prior assumptions? *bioRxiv*.

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

Thank you for your attention !