# Exponential families in phylogenetics

Marc Manceau

August 24, 2020

**Abstract**

This document intends to present thoughts on exponential families in phylogenetics. We first describe what exponential families are, and why they are so appreciated by statisticians. Then we turn to drawing up an inventory of exponential families in phylogenetics, which, we hope, could foster the development of new statistical methods.

## Contents

# 1 Basics on exponential families

## 1.1 Lexicon

Let's simply start with the definition.

**Definition 1** *A family of probability distributions parametrized by a parameter $\theta$ is called* an exponential family *if its probability mass function, or density, can be written as*

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

*where, in the vectorial case, $\eta(\theta)^t T(x)$ denotes the vectorial product $\sum_i \eta_i(\theta) T_i(x)$.*

This definition comes with its own lexicon. The quantity $\eta(\theta)$ is called *natural parameter*, and the family is said to be *in its canonical form* if $\eta_i(\theta) = \theta_i$. It is said to be *curved* if the dimension of $\theta$ is less than the dimension of $\eta(\theta)$.

The quantity $T(x)$ is called *sufficient statistic*, for this is all we need to know about a realisation to compute its probability.

Finally, $A(\eta)$ is called *log-partition function*, which is a term coming from physics, where the partition function refers to the normalizing factor that ensures that $f$ is a density, i.e.

$$A(\eta) = \ln\left(\int_x h(x)e^{\eta^t T(x)}dx\right)$$

## 1.2 Why they are so used

First, exponential families have this really nice property that allows to summarize an arbitrary amount of iid information with only a fixed number of values, through their sufficient statistics $T(x)$.

Second, these families have conjugate priors (i.e. there exists a family of distributions such that $f(\theta)$ is in the same family as $f(\theta|x)$. In a Bayesian framework, this makes them the perfect building blocks to get the posterior distribution analytically.

In general, the conjugate prior of an exponential family which density is given in the same form as Definition 1 has a density of the form

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

where $\chi, \nu$ are hyperparameters. Indeed, one can check that the posterior is in the same family,

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

There is also a nice general result for the moment generating function of the sufficient statistics, namely

$$\begin{aligned} M_T(u) &:= \mathbb{E}\left(e^{u^t T(x)|\eta}\right) \\ &= \int_x h(x)e^{(\eta+u)^t T(x) - A(\eta)}dx \\ &= e^{A(\eta+u) - A(\eta)} \end{aligned}$$

which allows one to easily derive the moments of the sufficient statistics through successive differentiation (and evaluation at the origin $u = 0$) of $M_t$.

(Plus, write something about Generalized Linear Model).

## 1.3 A few examples of discrete distributions

The best way to get a feeling of what these families are is to go through a few examples, first in a discrete setting.

**Example 1** *The family of binomial distributions $(\mathcal{B}(n,p))_{p\in(0,1)}$, i.e. with known $n$ and only $p$ as a parameter, is an exponential family.*

*The natural parameter is,*

$$\eta(p) := \ln \frac{p}{1-p} \quad \textit{(a.k.a. logit function)} \quad \Longleftrightarrow \quad p = \frac{e^{\eta}}{1+e^{\eta}} \quad \textit{(a.k.a. logistic function)}$$

*and the sufficient statistic is the number of successes $T(x) = x$.*

*Finally its conjugate prior is the Beta distribution,*

$$\begin{aligned} & p|\alpha,\beta \ \sim \textit{Beta}(\alpha,\beta) \\ \Longrightarrow \ & p|x,\alpha,\beta \ \sim \textit{Beta}(\alpha+x,\beta+(n-x)) \end{aligned}$$

It's sufficient to rewrite the probability mass function of the binomial in the form given in Definition 1.

$$\begin{aligned} f(x|p) &= \binom{n}{x} p^x (1-p)^{n-x} \\ &= \binom{n}{x} e^{x \ln p + (n-x)\ln(1-p)} \\ &= \binom{n}{x} e^{x \ln \frac{p}{1-p} + n \ln(1-p)} \end{aligned}$$

Note that we see on this expression that it couldn't work with $n$ as a parameter. As a general rule, it never works when a parameter changes the support of the distribution.

Finally, let's have a look at the conjugate prior. Suppose $p \sim \text{Beta}(\alpha,\beta)$, i.e.

$$f(p|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}$$

We can derive the posterior distribution of $p$, simply keeping all parts that depend on $p$ for simplicity,

$$\begin{aligned} f(p|x,\alpha,\beta) &\propto f(x|p) f(p) \\ &\propto p^x (1-p)^{n-x} p^{\alpha-1}(1-p)^{\beta-1} \\ &\propto p^{\alpha-1+x}(1-p)^{\beta-1+(n-x)} \end{aligned}$$

Note that whenever it's possible, it helps to have an interpretation of hyperparameters. Here, $p$ is estimated observing $\alpha-1$ successes and $\beta-1$ failures in a series of Bernoulli experiments. If one wants to have a non-informative prior, Beta(1,1) is a good candidate.

Finally, the law of $x|\alpha,\beta$ is called *Beta-binomiale*.

**Example 2** *The family of geometric distributions with support in $\mathbb{N}$, $(\mathcal{G}(p))_{p\in(0,1)}$ is an exponential family.*

*The natural parameter is $\eta(p) := \ln(1-p)$ and the sufficient statistic is the number of failures before the first success, $T(x) = x$.*

*Finally, the conjugate prior is again a Beta distribution,*

$$\begin{aligned} & p|\alpha,\beta \ \sim \textit{Beta}(\alpha,\beta) \\ \Longrightarrow \ & p|x,\alpha,\beta \ \sim \textit{Beta}(\alpha+1,\beta+x) \end{aligned}$$

We go quickly through this example, noting first that,

$$f(x|p) = (1-p)^x p = e^{x\ln(1-p)+\ln p} \quad .$$

If $p \sim \text{Beta}(\alpha,\beta)$, we get the posterior distribution

$$\begin{aligned} f(p|x,\alpha,\beta) &\propto (1-p)^x p p^{\alpha-1}(1-p)^{\beta-1} \\ &\propto p^{\alpha-1+1}(1-p)^{\beta-1+x} \end{aligned}$$

which is very coherent with the previous interpretation of hyperparameters $\alpha,\beta$.

**Example 3** *The family of negative binomial distributions $(\mathcal{NB}(r,p))_{p\in(0,1)}$, providing the distribution of the number of successes before a fixed number of failures $r$ in a series of Bernoulli experiments with probability of success $p$, is an exponential family.*

*The natural parameter is $\eta(p) := \ln p$ and the sufficient statistic is the number of successes before the $r$th failure, $T(x) = x$.*

*Finally, it won't be surprising at this point to seet a Beta distribution as conjugate prior, with,*

$$\begin{aligned} & p|\alpha,\beta \ \sim Beta(\alpha,\beta) \\ \implies \ & p|x,\alpha,\beta \ \sim Beta(\alpha+x,\beta+r) \end{aligned}$$

The negative binomial distribution is indeed characterized with the following probability mass function,

$$\begin{aligned} f(x|p) &= \binom{x+r-1}{x}(1-p)^r p^x \\ &= \binom{x+r-1}{x} e^{x\ln p + r\ln(1-p)} \end{aligned}$$

and, as for the binomial distribution, we have to fix $r$ (which governs the support of the distribution) to get an exponential family.

If $p \sim \text{Beta}(\alpha,\beta)$, we get the posterior distribution

$$\begin{aligned} f(p|x,\alpha,\beta) &\propto (1-p)^r p^x p^{\alpha-1}(1-p)^{\beta-1} \\ &\propto p^{\alpha-1+x}(1-p)^{\beta-1+r} \end{aligned}$$

which is again coherent with the previous interpretation of hyperparameters $\alpha-1, \beta-1$ as respectively the number of successes and failures in a series of Bernoulli experiments.

**Example 4** *The family of Poisson distribution $(\mathcal{P}(\lambda))_{\lambda\in\mathbb{R}^+}$ is an exponential family.*

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

*The conjugate prior is a Gamma distribution, with*

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

Indeed, we have the following probability mass function,

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

Further, if $\lambda \sim \Gamma(\alpha,\beta)$, we have

$$f(\lambda) = \frac{1}{\Gamma(\alpha)}\beta^\alpha \lambda^{\alpha-1}e^{-\beta\lambda}$$

which leads to the following posterior distribution,

$$f(\lambda|x) \propto \lambda^{\alpha-1+x}e^{-(\beta+1)\lambda} \implies \lambda|x\ \Gamma(\alpha+x,\beta+1)$$

The hyperparameter $\alpha$ can be interpreted as the total number of points observed, while $\beta$ is the total number of intervals.

**Example 5** *The family of multinomial distributions with fixed number of categories $k$ and fixed number of trials $n$, and parametrized by the vector of individual events probabilities $p = (p_1, p_2, ..., p_k)$ such that $\sum p_i = 1$, is an exponential family.*

*Its natural parameter is $\eta(p) = (\ln p_1, ..., \ln p_k)$ and the sufficient statistic is the number of observations in each categories, $T(x) = x = (x_1, x_2, ..., x_k)$.*

*The conjugate prior of $p$ is a Dirichlet distribution, with*

$$\begin{aligned} & p|\alpha \ \sim Dirichlet(\alpha) \\ \implies \ & p|x,\alpha \ \sim Dirichlet(\alpha+x) \end{aligned}$$

The probability mass function of the distribution is

$$f(x) = \frac{n!}{x_1!x_2!...x_k!}\prod_{i=1}^{k} p_i^{x_i} = \frac{n!}{x_1!x_2!...x_k!}e^{\sum_{i=1}^{k} x_i \ln p_i}$$

which gives to the sufficient statistic and natural parameter.

If $\alpha = (\alpha_1, \alpha_2, ...\alpha_k)$ with $\forall i, \alpha_i \geq 0$, and $p|\alpha$ Dirichlet($\alpha$), this means that we have the following prior mass function on $p$,

$$f(p) = \frac{\Gamma\left(\sum_{i=1}^{k} \alpha_i\right)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} p_i^{\alpha_i - 1}$$

which in turn allows us to express the posterior as

$$f(p|x) \propto \prod_{i=1}^{k} p_i^{\alpha_i - 1 + x_i} \quad \Longrightarrow \quad p|x \ \sim \text{Dirichlet}(\alpha + x).$$

Note that the multinomial distribution with $n = 1$ is sometimes called *categorical distribution*. The categorical distribution being the generalization of a Bernoulli distribution with $k \geq 2$ categories, and the multinomiale distribution being the generalization of a binomiale with $k \geq 2$ categories (i.e. the sum of $n$ categorical variables). The Dirichlet distribution is also a generalization of a Beta distribution in higher dimension. As for the Beta-binomial case, it is possible to integrate out the parameter $p$, in which case $x|\alpha \sim$Dirichlet-multinomial($\alpha$).

Finally, I find it interesting to end up with a discrete example that is NOT an exponential family.

**Example 6** *The hypergeometric distribution, describing the distribution of the number of successes for a fixed number $n$ of draws without replacement in an urn containing a fixed total number $N$ of balls among which $M$ lead to success, is not an exponential family. It nevertheless has a conjugate prior for the number $M$ of balls leading to success,*

$$\begin{aligned} & M|N, \alpha, \beta \ \sim \textit{Beta-binomial}(N, \alpha, \beta) \\ \Longrightarrow \ & M|x, N, \alpha, \beta \ \sim \textit{Beta-binomial}(N, \alpha + x, \beta + (n - x)) \end{aligned}$$

One can convince himself/herself that it is not an exponential family by looking at

$$f(x|M) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{n}{N}} \quad \forall x \in [\max(0, n + M - N), \min(n, M)]$$

Finally, checking that the Beta-binomial distribution is a conjugate prior is quite messy, because the probability mass function is full of combinatorial terms that we would have to arrange. Actually, I have never checked it myself, so I hope the result claimed above really holds.

## 1.4 A few examples of continuous distributions

We start with an example which is not an exponential family, before turning to classical laws that are exponential families.

**Example 7** *The family of uniform distributions $(\mathcal{U}(0, \theta))_{\theta \in R^+}$ is not an exponential family. It nevertheless admits a Pareto distribution as a conjugate prior.*

The density of the distribution is given by

$$f(x) = \frac{1}{\theta}\mathbb{1}_{x \in (0,\theta)}$$

where we find back this general rule, that a family of distributions which support depends on a parameter value cannot be an exponential family.

Nevertheless, assume that $\theta \sim$ Pareto($x_m, \alpha$). This means that

$$f(\theta|x_m, \alpha) = \alpha x_m^{\alpha} \theta^{\alpha+1} \mathbb{1}_{\theta \geq x_m)}$$

which in turn yields,

$$\begin{aligned} f(\theta|x) &\propto \theta^{-(\alpha+1)}\mathbb{1}_{\theta\geq x_m)}\theta^{-1}\mathbb{1}_{x\in(0,\theta)} \\ &\propto \theta^{-(\alpha+2)}\mathbb{1}_{\theta\geq x_m}\mathbb{1}_{\theta\geq x} \end{aligned}$$

which is another Pareto distribution with parameters $(\max(x_m, x), \alpha + 1)$. Note also that the family of Pareto distributions with known $x_m$ is an exponential family.

**Example 8** *The family of exponential distributions $(\mathcal{E}(\lambda))_{\lambda\in\mathbb{R}^+}$ is an exponential family, with natural parameter $\eta := -\lambda$ and sufficient statistic $T(x) = x$.*

*It is conjugate to a Gamma distribution: if $\lambda|\alpha, \beta \sim \Gamma(\alpha, \beta)$, then $\lambda|x, \alpha, \beta \sim \Gamma(\alpha + 1, \beta + x)$*

This comes quickly by rewriting the density as $f(x|\lambda) = \lambda e^{-\lambda x} = e^{-\lambda x + \ln\lambda}$.

Moreover, the Gamma distribution has probability density $f(\lambda|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\beta\lambda}$, which leads to the following Gamma posterior,

$$f(\lambda|x, \alpha, \beta) \propto \lambda^{\alpha-1+1}e^{-(\beta+x)\lambda} \quad .$$

Note that $\alpha - 1$ corresponds to the number of observations made, and $\beta$ corresponds to the total sum of observations made so far.

**Example 9** *The family of Gamma distributions $(\Gamma(\alpha, \beta))_{\alpha>0,\beta>0}$ is an exponential family, with natural parameter $\eta = (\alpha - 1, -\beta)$ and sufficient statistic $T(x) = (\ln x, x)$.*

*When the shape parameter $\alpha$ is fixed, a conjugate prior for $\beta$ is another Gamma distribution. When it's not fixed, the conjugate prior is something that does not have a name, but is easily expressed.*

Indeed, recall that the density of the distribution is,

$$\begin{aligned} f(x|\alpha, \beta) &= \frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x} \\ &= e^{(\alpha-1)\ln x - \beta x + \alpha\ln\beta - \ln\Gamma(\alpha)} \end{aligned}$$

In case $\alpha$ is fixed, suppose $\beta|\alpha_0, \beta_0 \sim \Gamma(\alpha_0, \beta_0)$. Then,

$$f(\beta|x, \alpha, \alpha_0, \beta_0) \propto \beta^{\alpha_0-1+\alpha}e^{-(\beta_0+x)\beta}$$

which is a $\Gamma(\alpha_0 + \alpha, \beta_0 + x)$.

In case $\alpha$ is not fixed, the conjugate prior on $\alpha, \beta$ has the following density with 4 hyperparameters $a, b, c, d$,

$$f(\alpha, \beta|a, b, c, d) \propto a^{\alpha-1}e^{-\beta b}\Gamma(\alpha)^{-c}\beta^{\alpha d}$$

in which case the posterior is obtained by updating $(a, b, c, d)$ with $(ax, b + x, c + 1, d + 1)$.

**Example 10** *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 exponential families.*

*Their natural parameters are, respectively, $\eta = (\mu/\sigma^2)$, $\eta = -1/(2\sigma^2)$, $\eta = (\mu/\sigma^2, -1/(2\sigma^2))$, and their sufficient statistics are $T(x) = x$, $T(x) = x^2 - 2\mu x$, $T(x) = (x, x^2)$.*

*Finally, they are conjugate to the following priors: $\mu \sim \mathcal{N}(\mu_0, \sigma_0^2)$, $\sigma^2 \sim \Gamma^{-1}(\alpha, \beta)$, $\sigma^2 \sim \mathcal{N}\Gamma^{-1}(\mu_0, \lambda, \alpha, \beta)$.*

It is sufficient to rewrite carefully the density of the distribution depending on which parameters are fixed or not,

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}} = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2\sigma^2} + \frac{\mu x}{\sigma^2} - \frac{\mu^2}{2\sigma^2} - \ln|\sigma|\right)$$

Suppose now that $\mu \sim \mathcal{N}(\mu_0, \sigma_0^2)$. To ease algebra here, we introduce the precision of the distribution to be $\tau_0 = 1/\sigma_0^2$. The posterior is then given by,

$$\begin{aligned} f(\mu|x, \tau, \mu_0, \tau_0^2) &\propto \exp\left(-\tau_0\frac{\mu^2}{2} + \tau_0\mu_0\mu\right)\exp\left(\tau\mu x - \tau\frac{\mu^2}{2}\right) \\ &\propto \exp\left(-\mu^2\frac{\tau_0 + \tau}{2} + \mu(\tau_0\mu_0 + \tau x)\right) \end{aligned}$$

where we recognize another normal distribution with precision $\tau' = \tau_0 + \tau$ and product $\mu'\tau' = \tau_0\mu_0 + \tau x \implies \mu' = (\tau_0\mu_0 + \tau x)/(\tau_0 + \tau)$.

We now consider the case where $\sigma^2$ is a parameter and $\mu$ is fixed. Again, for simplicity, we will consider a parametrization in terms of the precision $\tau = 1/\sigma^2$. Assume that $\tau \sim \Gamma(\alpha, \beta)$, then,

$$\begin{aligned} f(\tau|x, \mu, \alpha, \beta) &\propto \tau^{\alpha-1} e^{-\beta\tau} \tau^{\frac{1}{2}} e^{-\tau\frac{(x-\mu)^2}{2}} \\ &\propto \tau^{\alpha+\frac{1}{2}-1} e^{-\tau(\beta+\frac{(x-\mu)^2}{2})} \end{aligned}$$

where we recognize another Gamma distribution with parameters $\left(\alpha + \frac{1}{2}, \beta + \frac{(x-\mu)^2}{2}\right)$. Note that if $\tau \sim \Gamma(\alpha, \beta)$, we call *inverse Gamma* with the same parameters (and denote $\Gamma^{-1}(\alpha, \beta)$) the law of $1/\tau = \sigma^2$.

Finally, we consider the case where $\mu, \tau$ are not fixed. Suppose that $\tau|\alpha, \beta \sim \Gamma(\alpha, \beta)$, and that $\mu|\mu_0, \tau, \lambda \sim \mathcal{N}\left(\mu_0, \frac{1}{\lambda\tau}\right)$. When this is the case, we say that $\mu, \tau$ follows a *Normal-Gamma* distribution with parameters $(\mu_0, \lambda, \alpha, \beta)$. We anyway get the following posterior,

$$\begin{aligned} f(\mu, \tau|\mu_0, \lambda, \alpha, \beta) &\propto \tau^{\alpha-1} e^{-\beta\tau} \tau^{\frac{1}{2}} e^{-\frac{\lambda\tau}{2}(\mu-\mu_0)^2} \tau^{\frac{1}{2}} e^{-\frac{\tau}{2}(x-\mu)^2} \\ &\propto \tau^{\alpha+\frac{1}{2}-1} e^{-\beta\tau} \tau^{\frac{1}{2}} e^{-\mu^2\frac{\tau(\lambda+1)}{2} + \mu\tau(\lambda\mu_0+x) - \frac{\tau}{2}(\lambda\mu_0^2+x^2)} \\ &\propto \tau^{\alpha+\frac{1}{2}-1} e^{-\tau(\beta+\frac{\lambda\mu_0^2+x^2}{2})} \tau^{\frac{1}{2}} e^{-\mu^2\frac{\tau(\lambda+1)}{2} + \mu\tau(\lambda+1)\frac{\lambda\mu_0+x}{\lambda+1}} \end{aligned}$$

which is another Normal-Gamma distribution with updated parameters $\left(\frac{\lambda\mu_0+x}{\lambda+1}, \lambda+1, \alpha+\frac{1}{2}, \beta+\frac{\lambda\mu_0^2+x^2}{2}\right)$. Again, we call *Normal Inverse Gamma* distribution the distribution of $\mu, \sigma^2$ instead of $\mu, \tau$. Note that, when we observe $n$ values $(x_i)_{i=1}^n$, the updates of $\lambda$ and $\alpha$ are easily written as $\lambda^{(n)} = \lambda + n$ and $\alpha^{(n)} = \alpha + \frac{n}{2}$. However, we then have to carefully write what happens to $\mu_0$ and $\beta$. Provided that what is written above is right, one can show with a short recursion that, in fact, $\mu_0^{(n)} = \frac{\lambda\mu_0+\sum_i x_i}{\lambda+n}$ and $\beta^{(n)} = \beta + \frac{1}{2}\sum(x_i - \bar{x})^2 + \frac{\lambda n}{\lambda+n}\frac{(\bar{x}-\mu_0)^2}{2}$ (according to Wikipedia, I didn't double check).

Now that we described quite in details what happens with the normal distribution, we turn to its multivariate extension, without so many details.

**Example 11** *The family of multivariate normal distributions $(\mathcal{N}_d(\mu, \Sigma))$ with mean vector $\mu \in \mathbb{R}^d$ and covariance matrix $\Sigma$ is an exponential family. It is here also interesting to parametrize it using the inverse of the covariance matrix, called the precision matrix, $\Lambda := \Sigma^{-1}$, so that we get a natural parameter $(\Lambda\mu, -\frac{1}{2}\Lambda)$ and sufficient statistics $T(x) = (x, x^t x)$.*

*When $\Lambda$ is fixed, it is conjugate to another multivariate normal distribution for $\mu$. When $\mu$ is fixed, it is conjugate to a Wishart distribution for $\Lambda$. When both are parameters, it is conjugate to a* Normal-Wishart *distribution for $(\mu, \Lambda)$.*

I simply recall here the density of a multivariate normal distribution,

$$f(x|\mu, \Sigma) = (2\pi)^{-\frac{d}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}{}^t(x-\mu)\Sigma^{-1}(x-\mu)\right)$$

The Wishart distribution is a generalization of the Gamma distribution to higher dimension. It has support in positive definite matrices. This is also another exponential family. When combined with a multivariate normal distribution in the same way as Gamma and Normal distributions where combined previsouly, it gives a Normal-Wishart distribution. Details are quite ugly, so I refer to Wikipedia pages for the distribution definitions and the updates of the hyperparameters of the conjugate prior.

# 2 Phylodynamics models

## 2.1 The Kingman coalescent

If $\mathcal{T}$ is a tree with $n$ leaves and successive coalescence intervals $(T_2, T_3, ..., T_n)$, where $T_i :=$ time elapsed to go from $i$ to $i-1$ lineages, then it is said to follow a Kingman coalescent with parameter $\theta$ if the density is,

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

We thus have here an exponential family with natural parameter $\eta = -\theta$ and sufficient statistic $\sum_{i=2}^{n} \binom{i}{2} T_i$.
It turns out that this is conjugate to a Gamma distribution. Assume $\theta \sim \Gamma(\alpha, \beta)$, then,

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

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

### 2.2 The pure-birth (Yule) tree

If $x_i$ is the depth of leaf $i$, then we have the following density for a Yule tree,

$$f(T|\lambda) = \lambda^{n-1} \prod_{i=0}^{n-1} e^{-\lambda x_i}$$

It's more or less equivalent to observing $n-1$ iid $\mathcal{E}(\lambda)$ random variables, and we thus have the following Gamma conjugate prior.

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

### 2.3 Birth-death reconstructed trees

With the same notation as above, the density of the reconstructed tree is,

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

$$\text{xhere} \quad p(x_i|\lambda, \mu) = \left( \frac{\lambda - \mu}{\lambda - \mu e^{-(\lambda-\mu)x_i}} \right)^2 e^{-(\lambda-\mu)x_i} = \exp\left( -(\lambda - \mu)x_i + 2\ln(\lambda - \mu) - 2\ln(\lambda - \mu e^{-(\lambda-\mu)x_i}) \right)$$

This function $p$ is not an exponential family, since we cannot factorize $\eta(\theta)T(x)$ within the exponential. Finding a conjugate prior seems thus very optimistic. On the one hand, we could easily deal with the numerator part with some components, say, $(\lambda-\mu)^\alpha e^{-(\lambda-\mu)\beta}$. But unfortunately, it seems very optimistic to deal with the denominator.
Some hope could have come from the special case $\lambda = \mu$. However, here again, the function $p$ cannot be factorized, see the following expression,

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

Finally, another option consists in looking for the probability distribution of $U, D, T_{part}$, the sufficient statistics of the fully observed linear birth-death process (respectively, number of steps up, number of steps down, and total particle time). Knowing these, the fully observed birth-death process is an exponential family.

## 3 Trait evolution along a fixed tree

### 3.1 Molecular evolution with a 4 states Markov process

Let's take as first example a JC69 model, with rate of transition from any two different nucleotides $\alpha$. If we only observe $X_1$ at time $t_1$ and $X_2$ at time $t_2$, then the probability of the observation is,

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

which is not an exponential family.
What about the continuously observed process $(X_t)_{t \in (t_1, t_2)}$ then ?

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

which is an exponential family with sufficient statistics,

$$T_i := \int_{t_1}^{t_2} \mathbb{1}_{X_t=i} dt \quad \text{(total time spent in state } i\text{)}$$
$$U_{ij} := \text{ number of steps from } i \text{ to } j$$

Depending on the complexity of the model, we get different $T(x)$ with different dimensions. For example, the JC69 model simplifies to $e^{-3\alpha t}\alpha^U$ where $U$ is simply the number of steps.

Note that this seems a bit weird because $t$ is considered to be fixed and we observe the full trajectory of $X$. The analogy for a Poisson process would be to have the density $\lambda^{N_t}e^{-\lambda t}$ for the full process. But then, we get

$$\mathbb{P}(N_t = n) = \int_{t_1,t_2,...t_n} \mathbb{P}(X_t)dt_1dt_2...dt_n = \lambda^n e^{-\lambda t} \int_{t_1...t_n} dt_1...dt_n = \frac{(\lambda t)^n}{n!}e^{-\lambda t}$$

and by analogy for JC69, we get,

$$\mathbb{P}(U_t = n) = e^{-3\alpha t}\frac{(\alpha t)^n}{n!}$$

Let's have a quick look at what it gives us on a very simple tree, with 3 leaves labeled 1, 2, 3, one internal node 4 joining leaves 1 and 2, and the last internal node 5 joining 3 and 4. We fix this topology and the time at which leaves 1, 2, 3 live. We would like to have three remaining parameters: $(t_4, t_5, \alpha)$. We get,

$$\begin{aligned}
&\mathbb{P}(U_{14}, U_{24}, U_{45}, U_{35}|t_4, t_5, \alpha) \\
=&e^{-3\alpha((t_4-t_1)+(t_4-t_2)+(t_5-t_4)+(t_5-t_3))}\frac{(\alpha(t_4-t_1))^{U_{14}}}{U_{14}!}\frac{(\alpha(t_4-t_2))^{U_{24}}}{U_{24}!}\frac{(\alpha(t_5-t_4))^{U_{45}}}{U_{45}!}\frac{(\alpha(t_5-t_3))^{U_{35}}}{U_{35}!} \\
\propto&\exp\left(U_{14}\ln(\alpha(t_4-t_1)) + U_{24}\ln(\alpha(t_4-t_2)) + U_{45}\ln(\alpha(t_5-t_4)) + U_{35}\ln(\alpha(t_5-t_3))\right)
\end{aligned}$$

which is an exponential family with sufficient statistic $T(x) = (U_{14}, U_{24}, U_{45}, U_{35})$ and natural parameter $\eta = (\ln(\alpha(t_4-t_1)), \ln(\alpha(t_4-t_2)), \ln(\alpha(t_5-t_4)), \ln(\alpha(t_5-t_3)))$. This is another example of a curved distribution, with the dimension of $\eta$ (4) greater than the dimension of $\theta$ (3). Note also that if we fix $(t_4, t_5)$ and keep only $\alpha$ as a parameter, this simplifies to an exponential family with natural parameter $\ln\alpha$ and sufficient statistic $U_{14} + U_{24} + U_{45} + U_{35}$.

## 3.2 Continuous trait evolution along a fixed tree

Suppose we have a BM running along a fixed tree $T$ with $n$ leaves and coalescence times between two leaves $k, l$ denoted $t_{k,l}$. It the initial value is $\mu$ and the infinitesimal variance $\sigma^2$, then, we have the following tip distribution,

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

We could first dream of not fixing $T$ at all. We would then start with $\sigma^2\Sigma \sim$ Inverse Wishart. As a result, we could get some posterior of $\sigma^2\Sigma|(X_f)$ and hope this would help us reconstruct the tree. But in fact, this would have broader support than only covariance matrices corresponding to an (ultrametric or not) tree. Which means that we would need to find a clever way of projecting this distribution in the correct matrix subspace, which finally does not sound like a great idea. Or at least not a simple one.

However, considering that $T$ is fixed, we can look for a conjugate prior for $(\mu, \sigma^2)$. This task seems much simpler, and in fact a Normal-inverse-Gamma seems to do the job. Indeed, we have,

$$\begin{aligned}
f(x|\mu,\sigma) &= (2\pi)^{-\frac{n}{2}}|\sigma^2\Sigma|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}{}^t(x-\mu V)(\sigma^2\Sigma_T^{-1})(x-\mu V)\right) \\
&= (2\pi)^{-\frac{n}{2}}\sigma^{-n}|\Sigma_T|^{-\frac{1}{2}}\exp\left(-\frac{1}{2\sigma^2}{}^t x\Sigma_T^{-1}x - \mu V\Sigma_T^{-1}x - \mu^t x\Sigma_T^{-1}V + \mu^{2t}V\Sigma_T^{-1}V\right)
\end{aligned}$$

Assuming then that $(\mu, \sigma^2)$ follows a Normal inverse-Gamma with parameters $(\mu_0, \lambda, \alpha, \beta)$ means that we have the following prior density,

$$f(\mu, \sigma^2) \propto \sigma^{-1}\sigma^{-2(\alpha+1)}\exp\left(-\frac{2\beta + \lambda(\mu-\mu_0)^2}{2\sigma^2}\right)$$

which leads to the following posterior,

$$f(\mu,\sigma^2|x) \propto \sigma^{-1}\sigma^{-2(\alpha+1)-n}\exp\left(-\frac{1}{2\sigma^2}\left((2\beta+{}^tx\Sigma_T^{-1}x)+\mu^2(\lambda+{}^tV\sigma_T^{-1}V)-\mu(2\lambda\mu_0+{}^tV\Sigma_T^{-1}x-{}^tx\Sigma_T^{-1}V)+\lambda\mu_0\right)\right)$$

Provided there are not too many errors, this is likely to be another Normal inverse-gamma distribution with updated parameters,

$$\begin{aligned}
&-2(\alpha'+1) = -2(\alpha+1)-n \implies \alpha' = \alpha+\frac{n}{2}\\
&\lambda' = \lambda+{}^tV\Sigma_T^{-1}V\\
&2\lambda'\mu_0' = 2\lambda\mu_0+2{}^tV\Sigma_T^{-1}x \implies \mu_0' = \frac{\lambda\mu_0+{}^tV\Sigma_T^{-1}x}{\lambda+{}^tV\Sigma_T^{-1}V}\\
&\lambda'\mu_0'^2+2\beta' = 2\beta+{}^tx\Sigma_T^{-1}x+\lambda\mu_0^2 \implies \beta' = \beta+\frac{1}{2}{}^tx\Sigma_T^{-1}x+\frac{1}{2}\lambda\mu_0^2-\frac{1}{2}(\lambda\mu_0+{}^tV\Sigma_T^{-1}x)^2
\end{aligned}$$

Note that we could build up slowly from simpler cases first, $\sigma^2$ known, then $\mu$ known, and then both unknown.

It would be interesting to then write clearly what happens to the setting with multiple traits. This has already been described in Tolkoff et al. (2017), page 3.

# 4 Gibbs sampling in a phylogenetic setting

Lartillot (2006) introduced a Gibbs sampler in the context of molecular phylogenetics, using these conjugacy properties with data augmentation, for various parts of the model. He shows that it is more efficient than a Metropolis-Hastings MCMC sampling the same posterior.

Here is a short summary of the various components of the model he considers:

1. sites can either all have the same substitution process (SUB) or each has its own substitution process (MAX).
2. the substitution model $Q$ depends on $(\rho_{ij})$, the relative exchangeabilities and $(\pi_i)$, the stationary profile: $Q_{ij} = \rho_{ij}\pi_j$ In WAG, $\rho$ is fixed throughout the sequence and $\pi$ varies at each locus. In *Poisson*, $\rho_{ij} = 1\forall i,j$ and $\pi$ varies at each locus. In GTR, $\pi$ and $\rho$ vary at each locus.
3. $\pi \sim \text{Dirichlet}(w)$, where $w$ has total weight $\delta = \sum_i w_i$ and center $\pi_0(a) = w_a/\delta$.
4. the tree topology is fixed, but there is an exponential prior on branch-lengths: all $l_j$ are independent and $l_j \sim \mathcal{E}(\beta)$.
5. all sites-specific rates are independent and Gamma distributed with mean one and parameter $\alpha$, with $p(r_i) \propto r_i^{\alpha-1}e^{-\alpha r_i}$.
6. when they vary, relative exchangeabilities are all independent with $\rho_{ij} \sim \mathcal{E}(1)$.

The data-augmentation and Gibbs sampling steps correspond to, along each branch $j$, at any site $i$,

1. sampling $n_{ij}$ the total number of substitutions, $(t_{ij}^k)$ the times at which substitutions occur, and $(\sigma_{ij}^k)$ the successive states.
2. sampling $l$ given everything else. This is a Gamma distribution with known shape and scale parameters.
3. sampling $r$ given everything else. This is a known Gamma distribution.
4. sampling $\pi$ given everything else. This is a Dirichlet with known parameters (or product of Dirichlet for MAX).
5. update the hyperparameters with a MH step, but using the posterior integrated over the variables for which we have an analytic distribution $(l, r, \pi)$.

Even though each step requires the costly data-agumentation step, the resulting decorrelation time for the Gibbs-MCMC is always estimated to be at least one order of magnitude smaller than for the MH-MCMC.

# References

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

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