Introduction to Bayesian methods

OpenBUGS

Basic Model and Priors

In Bayesian methods the following elements are used:

- **A Model.** A parametric model with parameter vector $\boldsymbol{\theta}$ is selected to represent the variable $\boldsymbol{X}$ being analyzed, that is:

$$\boldsymbol{X} \sim \boldsymbol{p}(\boldsymbol{x}|\boldsymbol{\theta})$$

- **A prior distribution of the parameters.** Since the parameters are assumed random in Bayesian methods, they have a distribution function. Initially a "prior distribution" is given:

$$\boldsymbol{\theta} \sim \boldsymbol{p}(\boldsymbol{\theta}|\boldsymbol{\alpha}),$$

  that is, a distribution with parameter $\boldsymbol{\alpha}$, **called hyperparameter** to be distinguished from the initial model parameter $\boldsymbol{\theta}$.

- **A sample.** It is assumed that we have a sample $\boldsymbol{X} = (\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_n})$ of size $\boldsymbol{n}$ coming from the model population, i.e. with likelihood

$$\boldsymbol{p}(\boldsymbol{X}|\boldsymbol{\theta}) \equiv \boldsymbol{L}(\boldsymbol{X}|\boldsymbol{\theta}).$$

Posterior and predictive distributions

- **The marginal likelihood of the evidence.** This is:

$$p(X|\alpha) = \int_{\theta} p(X|\theta)p(\theta|\alpha)d\theta.$$

- **The posterior distribution.** Combining the parameters prior distribution and the sample information we get the posterior distribution of the parameters given X, by means of

$$p(\theta|X,\alpha) = \frac{p(X|\theta)p(\theta|\alpha)}{p(X|\alpha)} \propto p(X|\theta)p(\theta|\alpha)$$

- **The prior predictive distribution.**

$$p(\tilde{x}|\alpha) = \int_{\theta} p(\tilde{x}|\theta)p(\theta|\alpha)d\theta$$

- **The posterior predictive distribution.**

$$p(\tilde{x}|X,\alpha) = \int_{\theta} p(\tilde{x}|\theta)p(\theta|X,\alpha)d\theta$$

Choice of simple prior distributions is convenient, because this simplifies things. In this sense, **the so-called conjugated distributions are of special interest**.

Definition (Conjugate distributions)

A family of prior distributions is said natural conjugate of a family of likelihoods, iff the family of posterior distributions that results coincides with the initial family of distributions.

$$p(\theta|X,\alpha) \propto p(\mathbf{X}|\theta)p(\theta|\alpha) = p(\theta|H(\mathbf{X},\alpha)),$$

that is, the posterior belongs to the prior family where the posterior hyperparameters are a function of the data and the prior hyperparameters.

Thus, **obtaining the posterior distribution reduces to calculating the posterior hyperparameters as a function of the prior hyperparameters and the sample**.

The main advantage is that **we use the same family for priors and posteriors**.

Tables of conjugate priors can be found in:

`https://en.wikipedia.org/wiki/Conjugate_prior`

Some illustrative examples are given next.

Conjugate of the Bernoulli family $Be(\theta)$

The natural conjugate of the Bernoulli family $Be(\theta)$ is the beta family $Beta(r, n)$.

Consider a Bernoulli variable with beta $Beta(r', n')$ prior $Be(\theta)$;

$$p(\theta; \alpha) = p(\theta; r', n') \propto \theta^{r'-1}(1-\theta)^{n'-r'-1},; \quad 0 \leq \theta \leq 1. \tag{1}$$

where $\propto$ means "proportional to".

If we repeat n Bernoulli's experiment and get x successes, then, the likelihood $p(x; \theta)$, becomes:

$$p(x; \theta) = \begin{pmatrix} n \\ x \end{pmatrix} \theta^x (1-\theta)^{n-x}; \quad 0 \leq x \leq n, \tag{2}$$

that is,

$$p(x; \theta) \propto \theta^x (1-\theta)^{n-x}.$$

Therefore, **the posterior distribution becomes a $Beta(r' + x, n' + n)$**, because

$$\begin{aligned} p(\theta|x, r', n') &\propto p(x; \theta)p(\theta; r', n') = \theta^x (1-\theta)^{n-x} \theta^{r'-1}(1-\theta)^{n'-r'-1} \\ &= \theta^{r'+x-1}(1-\theta)^{n'+n-(r'+x)-1}. \end{aligned} \tag{3}$$

Asymptotic behavior

We have seen that **the conjugate of the Bernoulli family is the Beta family**, that is, if the prior of the Bernoulli parameter θ is a $Beta(r', n')$ distribution, then, the posterior of θ is also a $Beta(r'', n'')$ distribution.

The resulting formulas for determining the posterior parameters as a function of the prior parameters and the experimental sample, consisting in repeating n times the Bernoulli experiment and observing x successes, is given by:

$$Beta(r', n') \quad \Rightarrow \quad Beta(r' + x, n' + n) \quad \Rightarrow \quad \mathbf{r'' = r' + x; \quad n'' = n' + n}.$$

When n tends to infinity, the number of expected successes x tends to infinity too, and then, the prior hyperparameters become negligible with respect to the posterior hyperparameters, so that **the objective contribution of the sample dominates**. Only **when the sample size n is relatively small, the expert opinion has some weight**.

Conjugate of the Poisson family $P(\theta)$

The conjugate distribution of the Poisson family, $P(\theta)$, is the gamma family $G(k, \lambda)$.

Consider a Poisson variable $P(\theta)$, with gamma prior $G(k', \lambda')$ with prior pdf:

$$p(\theta; k', \lambda') = \frac{e^{-\lambda'\theta}(\lambda\theta)^{k'-1}}{\Gamma(k')} \propto e^{-\lambda'\theta}\theta^{k'-1}; \quad \theta \geq 0. \tag{4}$$

If we observe x events during a period of duration λ, the likelihood $p(x; \theta)$ of observing those events assuming a process with intensity λ is

$$p(x; \theta) = \frac{e^{-\lambda\theta}(\lambda\theta)^{x}}{x!} \propto e^{-\lambda\theta}\theta^{x}. \tag{5}$$

Therefore, the posterior becomes:

$$\begin{aligned} p(\theta|x, k', \lambda') &\propto p(x; \theta)p(\theta; k', \lambda') \\ &= e^{-\lambda\theta}\theta^{x}e^{-\lambda'\theta}\theta^{k'-1} = e^{-(\lambda'+\lambda)\theta}\theta^{(k'+x)-1}, \end{aligned} \tag{6}$$

that is, **the posterior is the gamma distribution $G(k' + x, \lambda' + \lambda)$**.

Asymptotic behavior

We have seen that **the conjugate of the Poisson family is the Gamma family**, that is, if the prior of the Poisson parameter λ is a $Gamma(k', \lambda')$ distribution, then, the posterior of λ is also a $Gamma(k'', \lambda'')$ distribution.

The resulting formulas for determining the posterior parameters as a function of the prior parameters and the experimental sample, consisting in observing the event during a period of duration λ and observing x events, is given by:

$$Gamma(k', \lambda') \quad \Rightarrow \quad Gamma(k' + x, \lambda' + \lambda) \quad \Rightarrow \quad \mathrm{k}'' = \mathrm{k}' + \mathrm{x}; \quad \lambda'' = \lambda' + \lambda.$$

When λ tends to infinity, the number of expected successes x tends to infinity too, and then, the prior hyperparameters become negligible with respect to the posterior hyperparameters, so that **the objective contribution of the sample dominates**. Only **when the observation duration λ is relatively small, the expert opinion has some weight**.

Bayesian methods

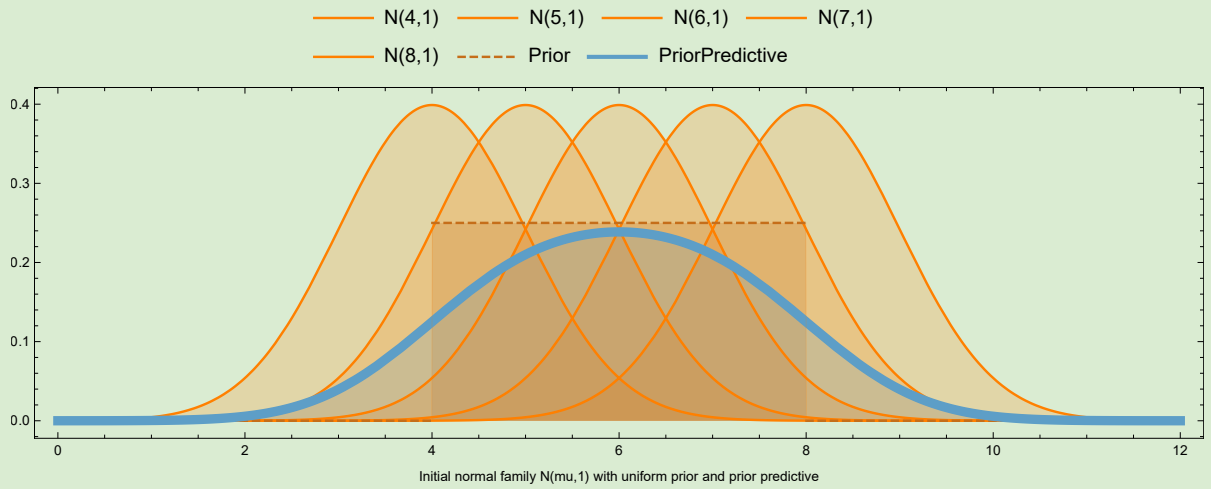
We end this part by saying that **Bayesian methods do not select one model from the initial family of distributions but any possible convex combination (predictive model)** of this family compatible with the prior.

This enlarges substantially the possible set of models, allowing the sample to provide us with the most convenient model.

Illustrative examples

Example 1

Bayesian methods **consider parameters as random variables**.

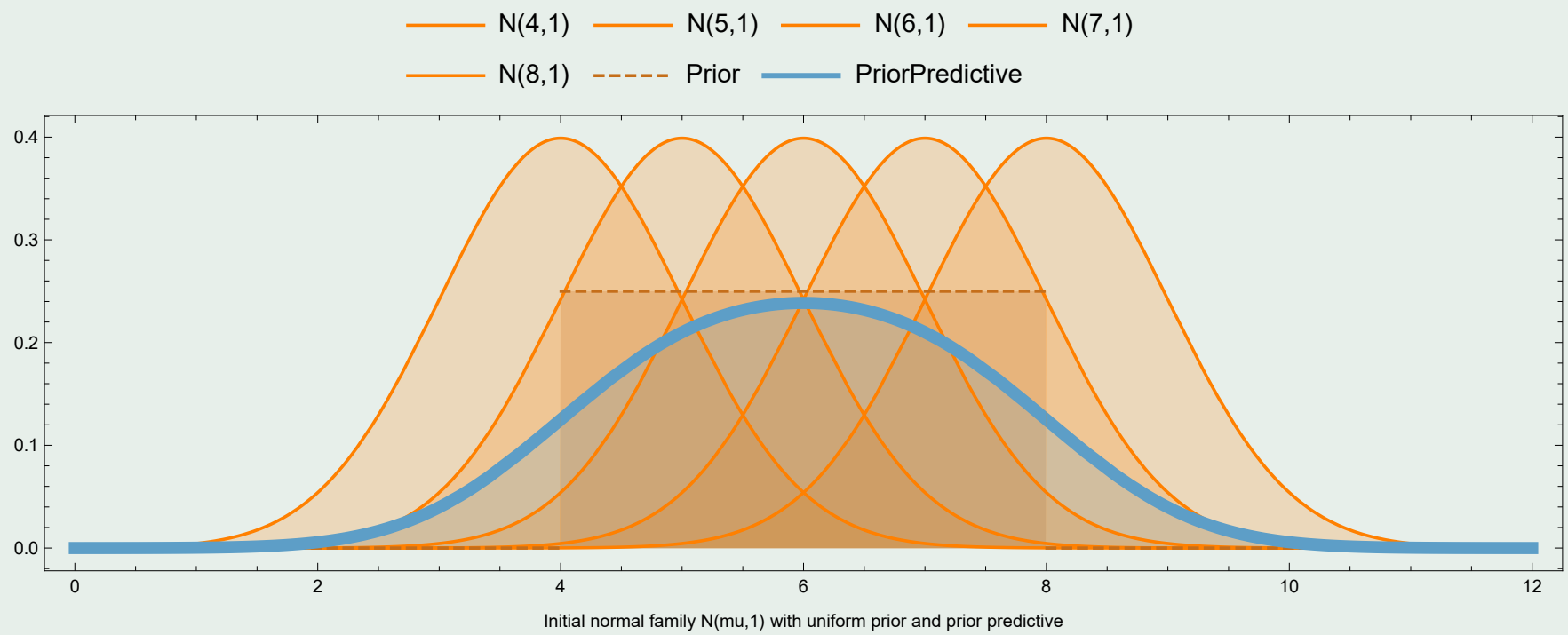


For example if we have a model with a normal $N(\theta, 1)$ random variable, **we can assume that the parameter θ is random**, for example a uniform $U(4, 8)$ distribution.

The initial random distribution of the parameters is called the prior distribution. If this distribution has parameters, **we call them hyperparameters** to distinguish them from the parameters of the initial model. Thus, the prior in the example is the uniform distribution, and the hyperparameters are, for example, 4 and 8.

The distribution we are really assuming is a weighted or mixed distribution that results after combining the prior with the initial family of distributions.

Example 1: family of normals $N(\theta, 1)$ with Uniform prior

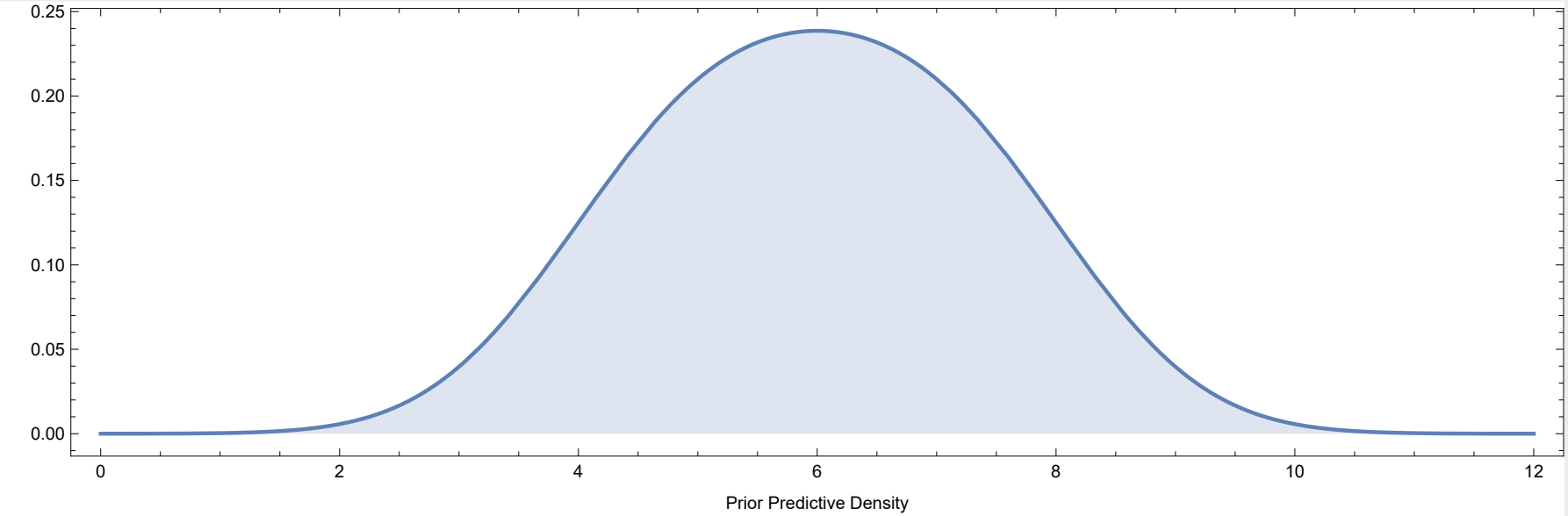


This figure shows a family of normals $N(\theta, 1)$ and a uniform $U(4, 8)$ prior, that leads to the prior predictive density in the figure, that is, to:

$$p(\tilde{x}|4, 8) = \int_{\theta} f_{U(4,8)}(\theta) f_{N(\theta,1)}(\tilde{x}) d\theta,$$

where θ is the parameter and 4 and 8 are the hyperparameters.

Example 1: family of normals $N(\theta, 1)$ with Uniform prior



This is the initially assumed prior predictive distribution:

$$p(\tilde{x}|4, 8) = \int_{\theta} f_{U(4,8)}(\theta) f_{N(\theta,1)}(\tilde{x}) d\theta,$$

which is not a normal distribution.

Example 1: family of normals $N(\theta, 1)$ with Uniform prior

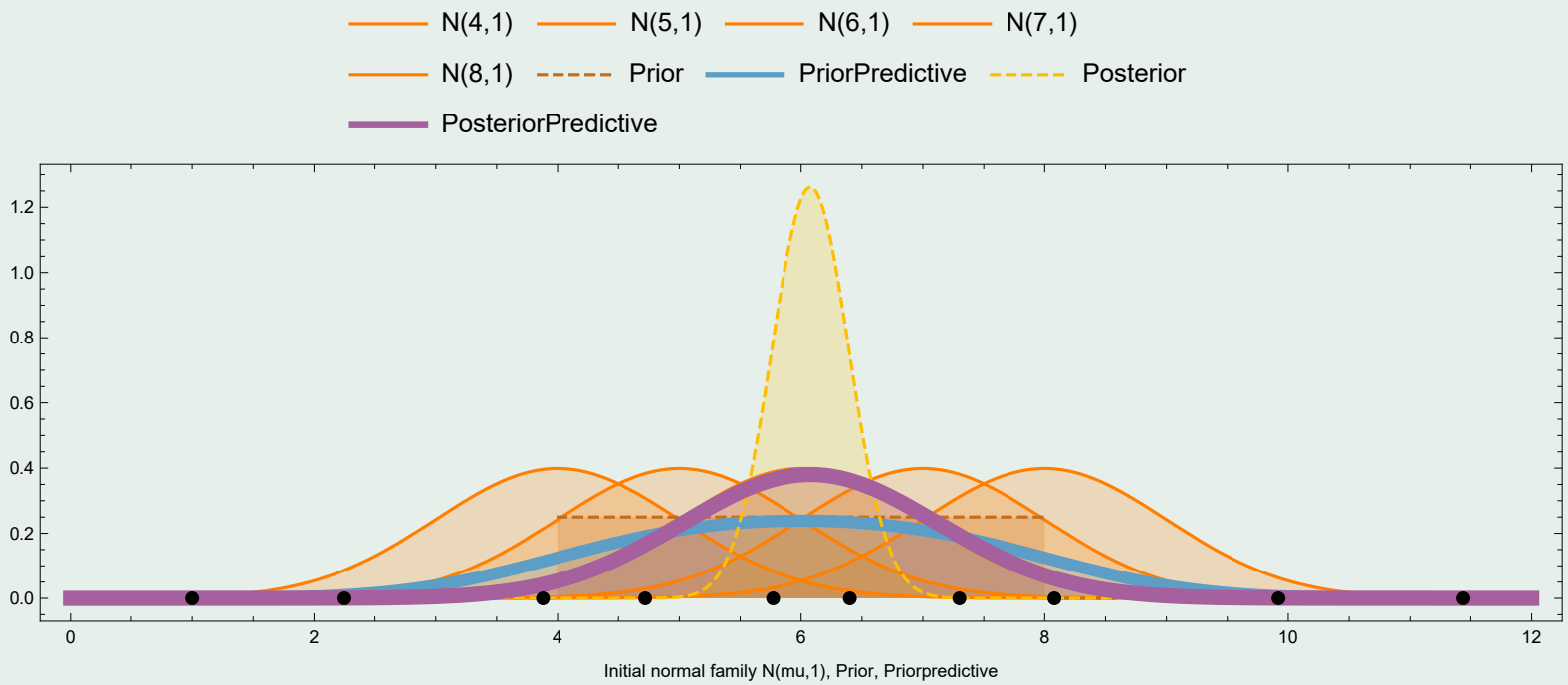


Illustration of how to update the uniform prior with one sample to obtain the posterior.

The posterior distribution is:

$$p(\theta|X, U(4,8)) \propto f_{U(4,8)}(\theta) \prod_{i=1,10} f_{N(\theta,1)}(x_i)$$

The posterior predictive distribution (dashed line) becomes

$$p(\tilde{x}|X, U(4,8)) = \int_{\theta} p(\theta|X, U(4,8)) f_{N(\theta,1)}(\tilde{x}) d\theta$$

Example 1: family of normals $N(\theta, 1)$ with Uniform prior

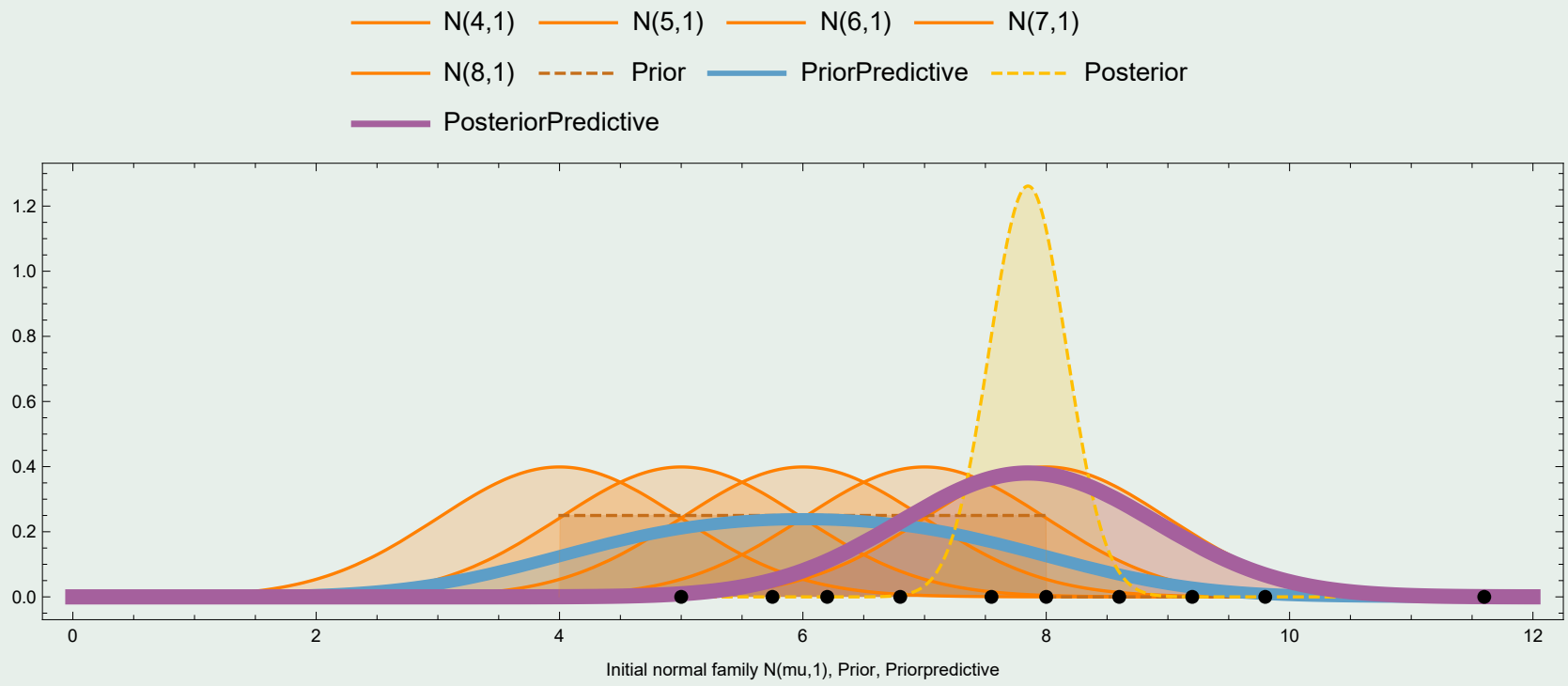


Illustration of how to update the uniform prior with one sample to obtain the posterior.

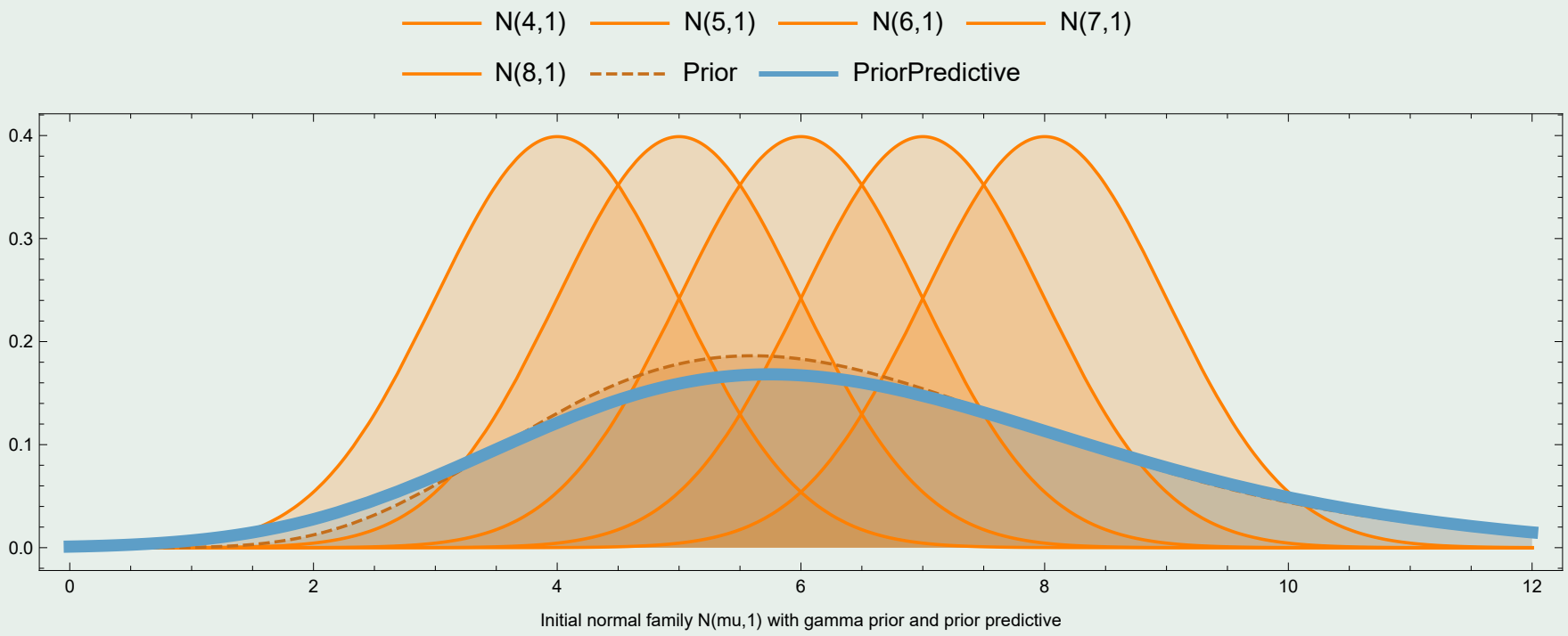
The posterior distribution is:

$$p(\theta|X, U(4,8)) \propto f_{U(4,8)}(\theta) \prod_{i=1,10} f_{N(\theta,1)}(x_i)$$

The posterior predictive distribution becomes

$$p(\tilde{x}|X, U(4,8)) = \int_{\theta} p(\theta|X, U(4,8)) f_{N(\theta,1)}(\tilde{x}) d\theta$$

EXAMPLE 2: FAMILY OF NORMALS $N(\theta, 1)$ WITH GAMMA PRIOR

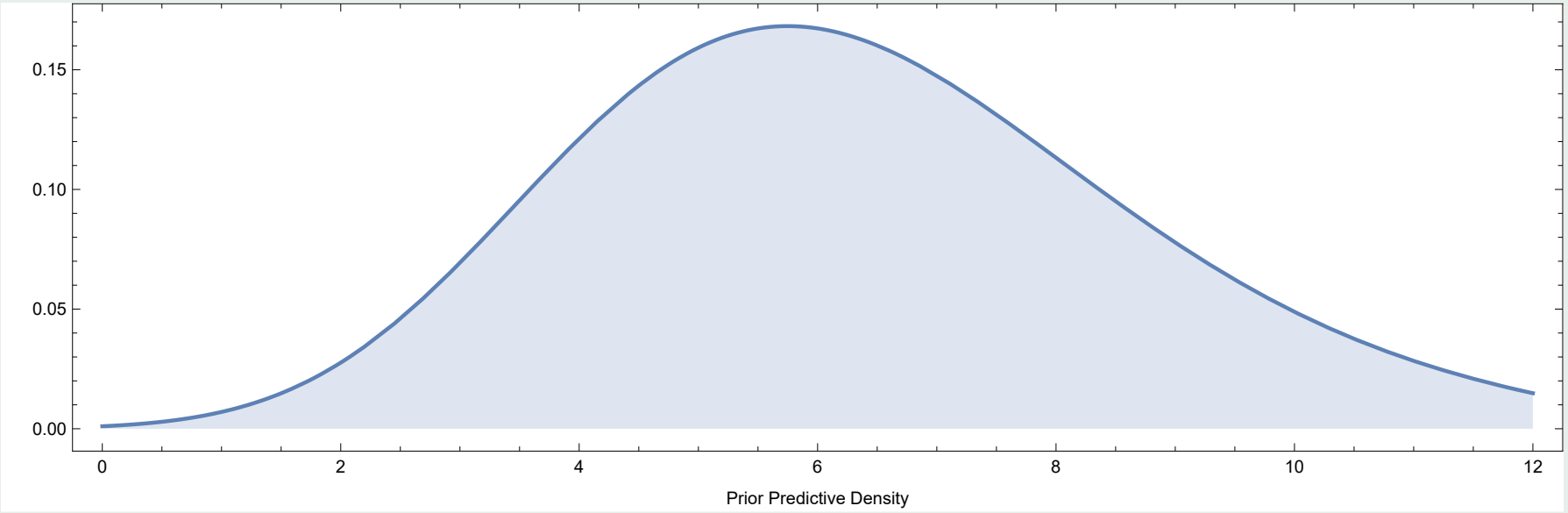


This figure shows a family of normals $N(\theta, 1)$ and a gamma $Gamma(8, 0.8)$ prior, that leads to the prior predictive density in the figure, that is, to:

$$p(\tilde{x}|8, 0.8) = \int_{\theta} f_{G(8,0.8)}(\theta) f_{N(\theta,1)}(\tilde{x}) d\theta,$$

where θ is the parameter and 8 and 0.8 are the hyperparameters.

EXAMPLE 2: FAMILY OF NORMALS $N(\theta, 1)$ WITH GAMMA PRIOR



This is the prior predictive distribution

$$p(\tilde{x}|8, 0.8) = \int_{\theta} f_{G(8,0.8)}(\theta) f_{N(\theta,1)}(\tilde{x}) d\theta,$$

which is not a gamma distribution.

EXAMPLE 2: FAMILY OF NORMALS $N(\theta, 1)$ WITH GAMMA PRIOR

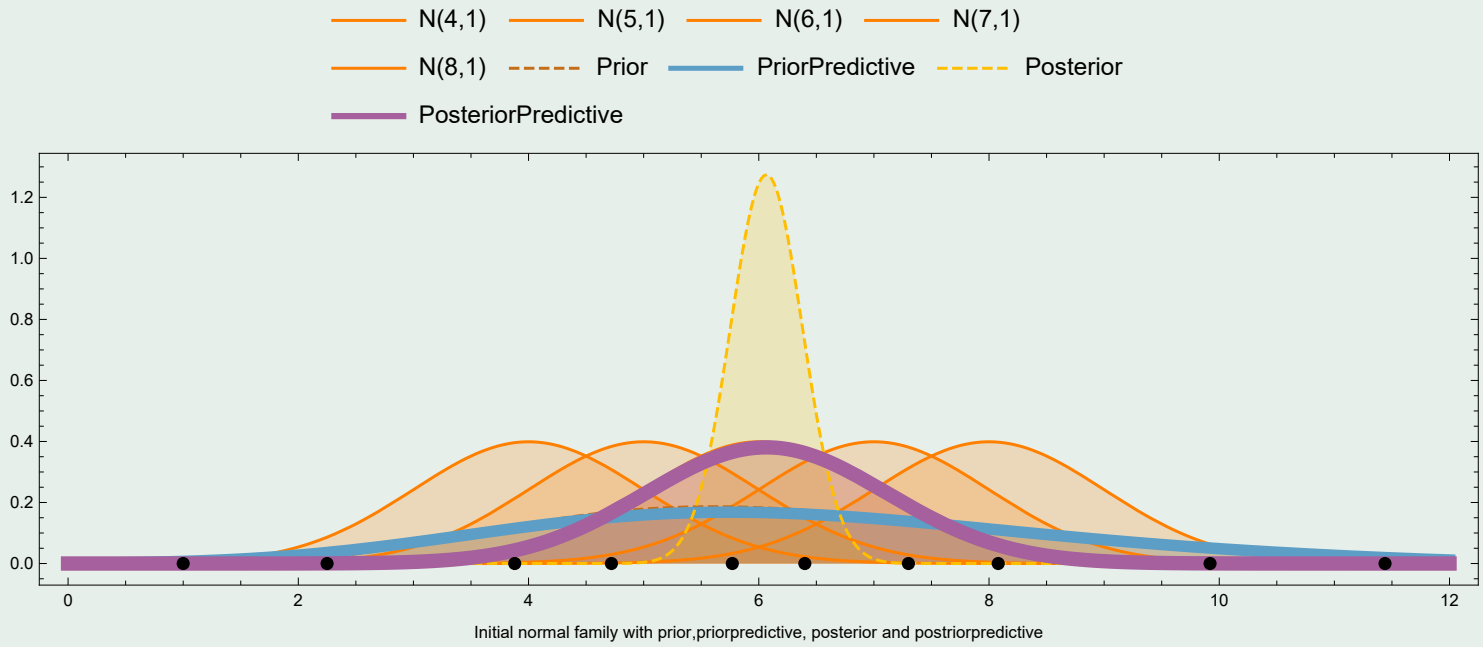


Illustration of how to update the gamma prior with one sample to obtain the posterior.

The posterior distribution is:

$$p(\theta|X, Gamma(8, 0.8)) \propto f_{Gamma(8,0.8)}(\theta) \prod_{i=1,10} f_{N(\theta,1)}(x_i)$$

The posterior predictive distribution becomes

$$p(\tilde{x}|X, Gamma(8, 0.8)) = \int_{\theta} p(\theta|X, Gamma(8, 0.8)) f_{N(\theta,1)}(\tilde{x}) d\theta$$

EXAMPLE 2: FAMILY OF NORMALS $N(\theta, 1)$ WITH GAMMA PRIOR

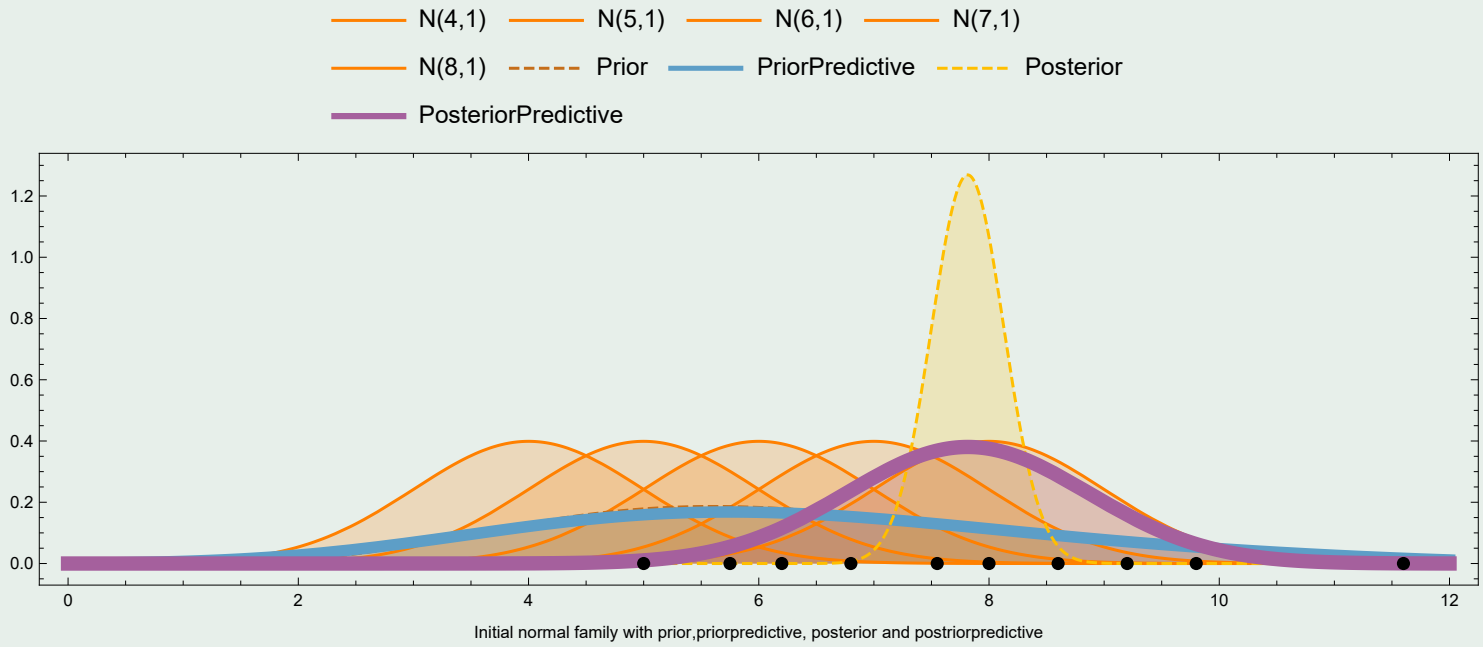


Illustration of how to update the gamma prior with one sample to obtain the posterior.

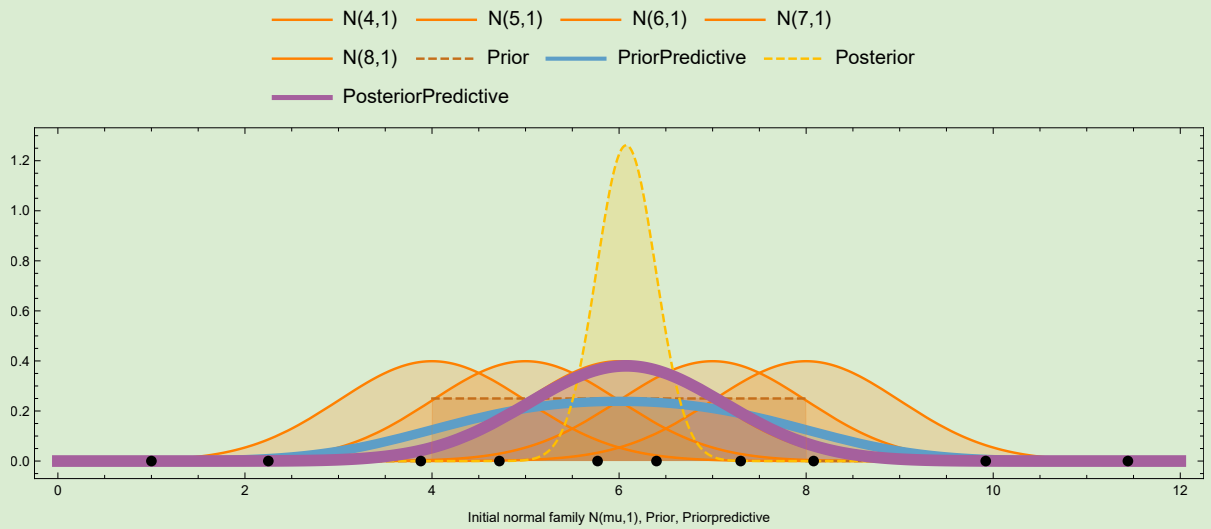
The posterior distribution is:

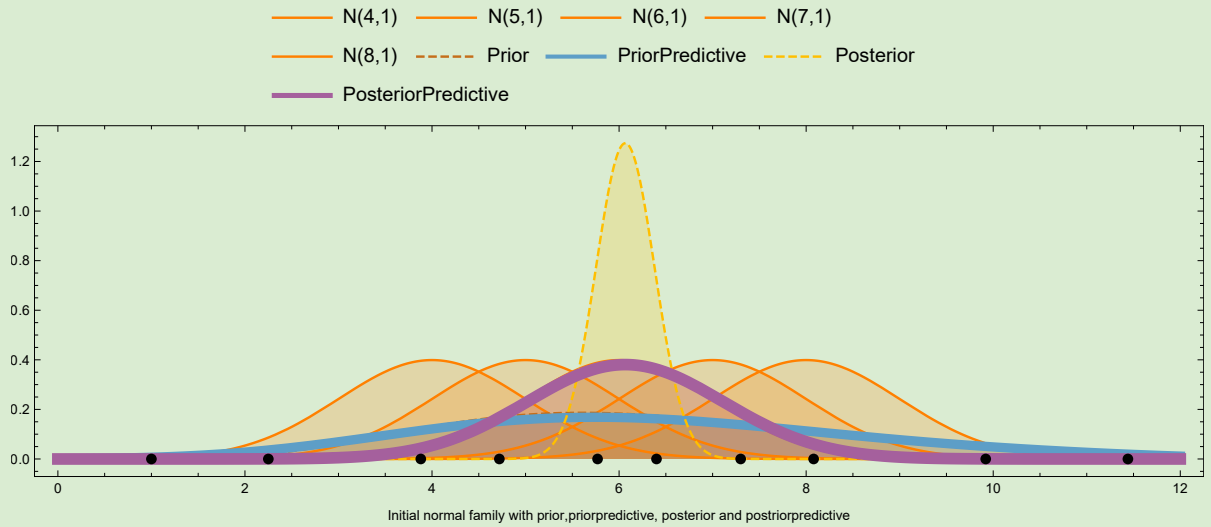
$$p(\theta|X, Gamma(8, 0.8)) \propto f_{Gamma(8,0.8)}(\theta) \prod_{i=1,10} f_{N(\theta,1)}(x_i)$$

The posterior predictive distribution becomes

$$p(\tilde{x}|X, Gamma(8, 0.8)) = \int_{\theta} p(\theta|X, Gamma(8, 0.8)) f_{N(\theta,1)}(\tilde{x}) d\theta$$

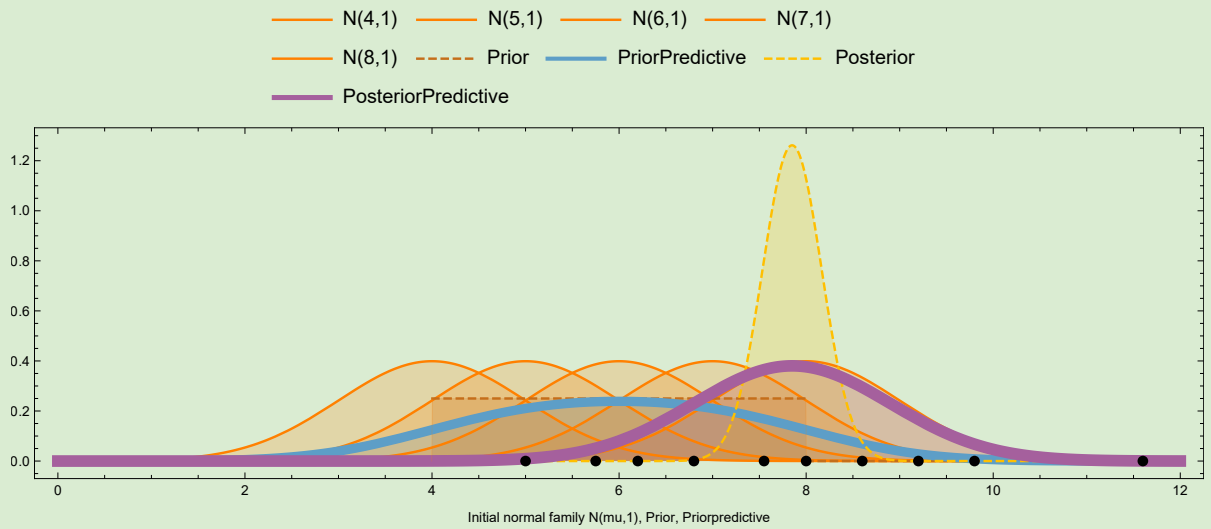
Family of normals $N(\theta, 1)$ with uniform and gamma priors

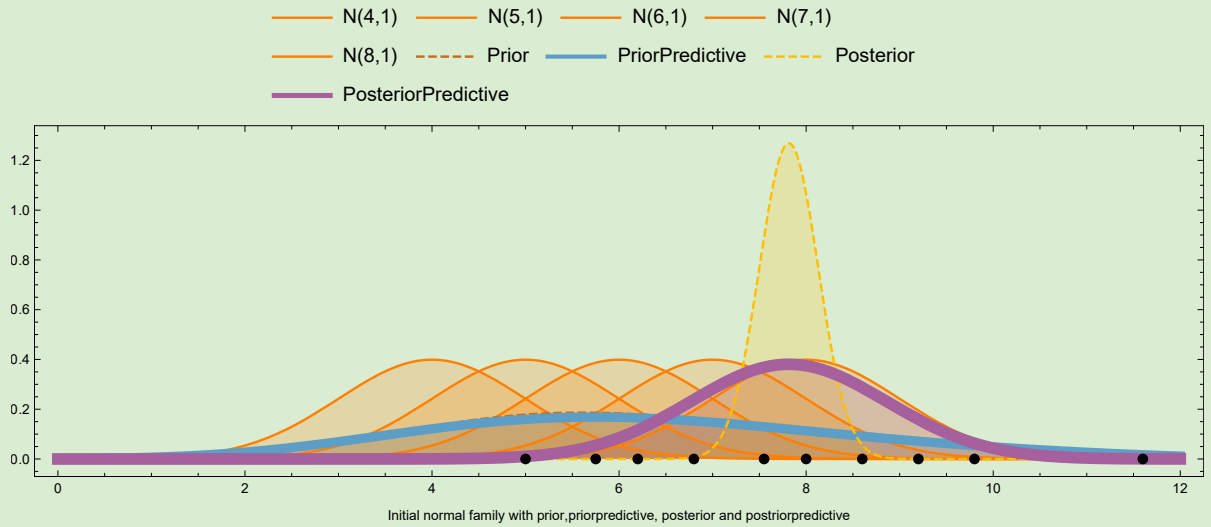




The posterior distributions tend to be independent on the priors when the sample size increases, as illustrated by the two figures corresponding to two different priors

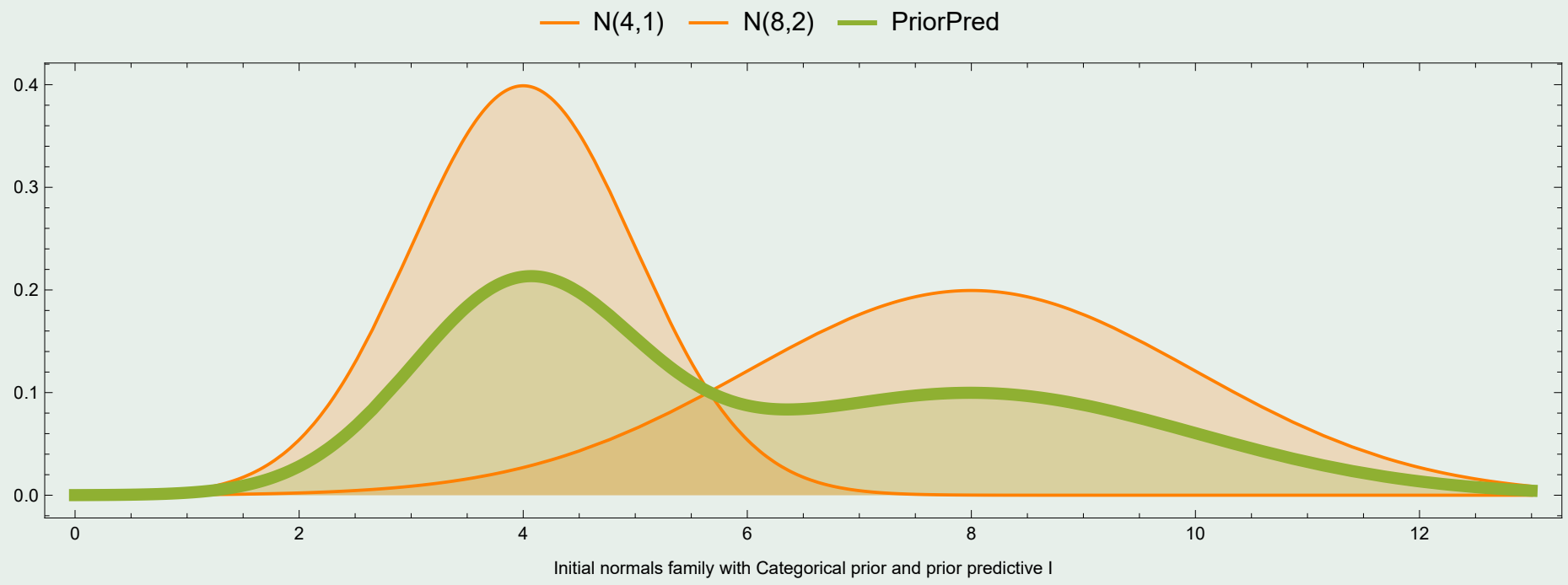
Family of normals $N(\theta, 1)$ with uniform and gamma priors





The posterior distributions tend to be independent on the priors when the sample size increases, as illustrated by the two figures corresponding to two different priors

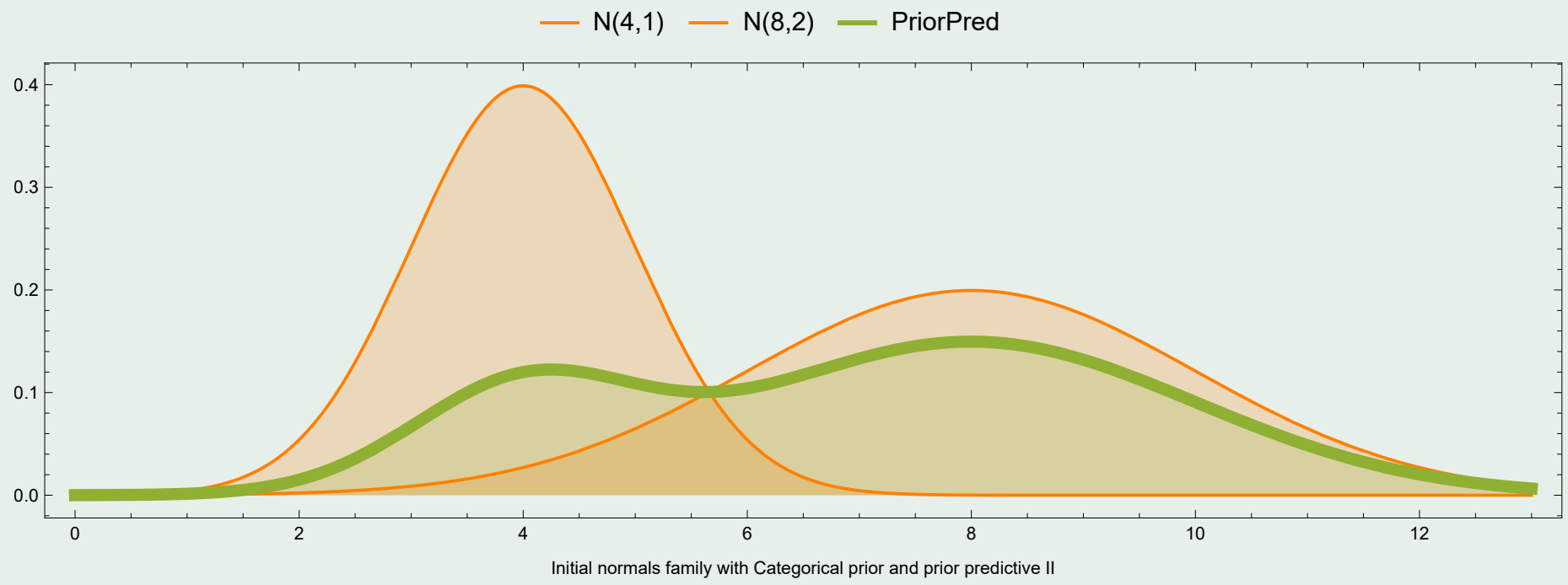
EXAMPLE 3: NORMALS $N(4,1)$ AND $N(8,2)$ WITH CATEGORICAL $cat(1/2)$ PRIOR



This figure shows normals $N(4,1)$ and $N(8,2)$ and the resulting prior predictive distribution associated with a discrete prior with weights $1/2$ and $1/2$, respectively. i. e.

$$p(\tilde{x}|1/2) = 1/2 f_{N(4,1)}(x) + 1/2 f_{N(8,2)}(x)$$

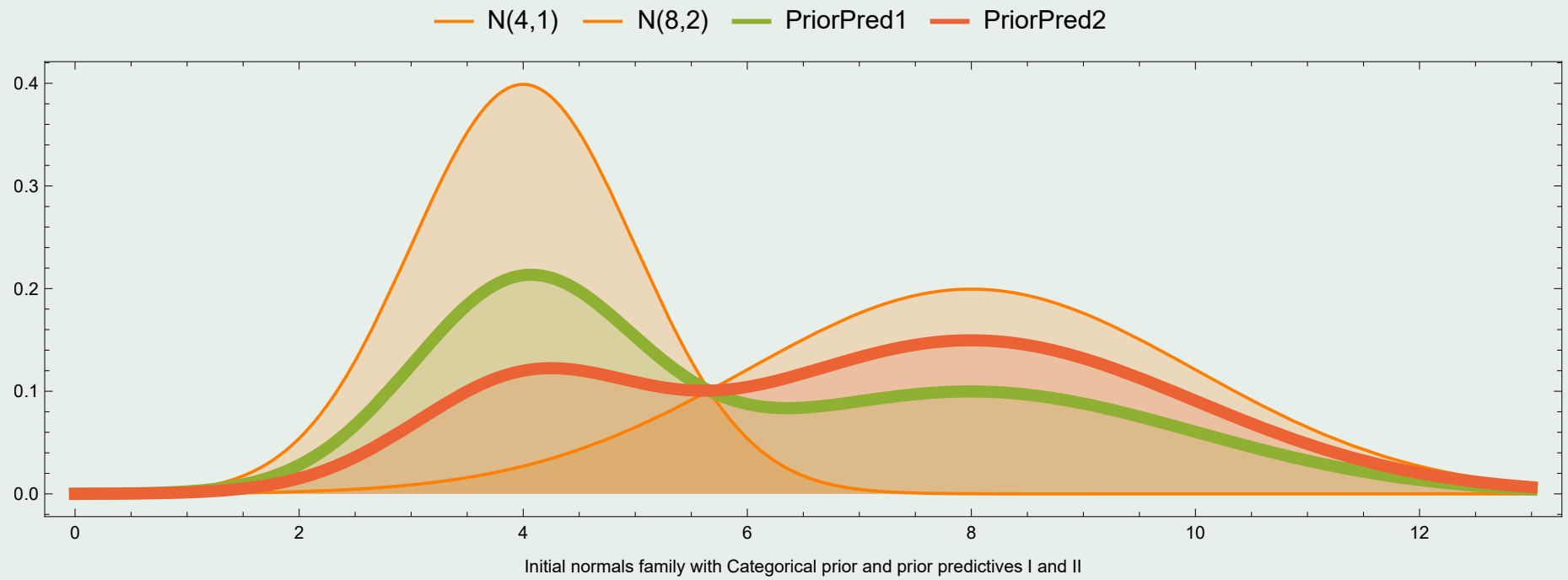
EXAMPLE 3: NORMALS $N(4,1)$ AND $N(8,2)$ WITH CATEGORICAL $cat(1/4)$ PRIOR



This figure shows normals $N(4,1)$ and $N(8,2)$ and the resulting prior predictive distribution associated with a discrete prior with weights $1/4$ and $3/4$, respectively. i. e.

$$p(\tilde{x}|1/4) = 1/4 f_{N(4,1)}(x) + 3/4 f_{N(8,2)}(x)$$

Example 3: normals $N(4,1)$ and $N(8,2)$ with categorical priors



This figure shows normals $N(4,1)$ and $N(8,2)$ and the resulting mixed distribution associated with two discrete priors with weights $1/2$ and $1/2$ and $1/4$ and $3/4$, respectively.

EXAMPLE 3: NORMALS $N(4,1)$ AND $N(8,2)$ WITH CATEGORICAL PRIORS

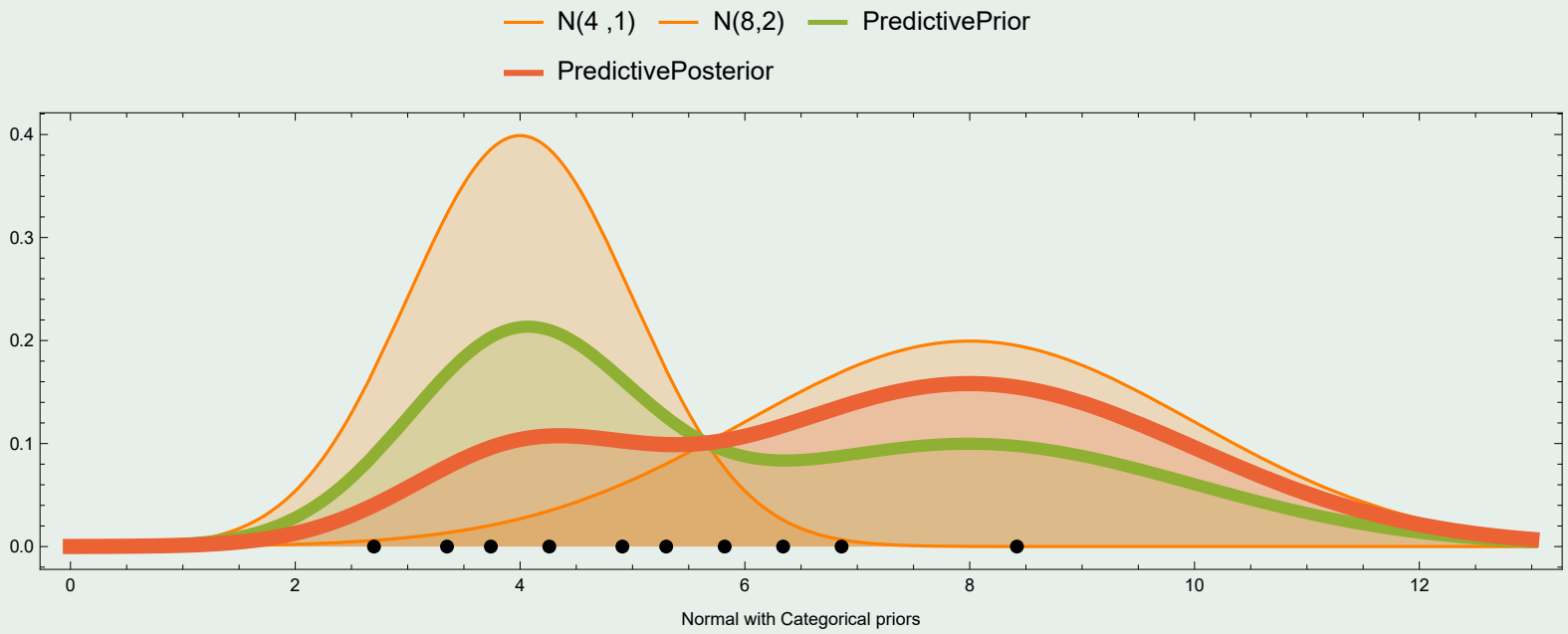


Illustration of how to obtain the posterior and the predictive prior with one sample.

The posterior distribution is:

$$p(\theta|X, Cat(\theta_i)) \propto p_{Cat(\theta_i)}(\theta) \prod_{i=1,10} f_{N(\theta,1)}(x_i)$$

The posterior predictive distribution becomes

$$p(\tilde{x}|X, Cat(\theta_i)) = \sum_{i=1}^{2} p(\theta|X, Cat(\theta_i)) f_{N(\theta,1)}(\tilde{x}) d\theta$$

EXAMPLE 3: NORMALS $N(4,1)$ AND $N(8,2)$ WITH CATEGORICAL PRIORS

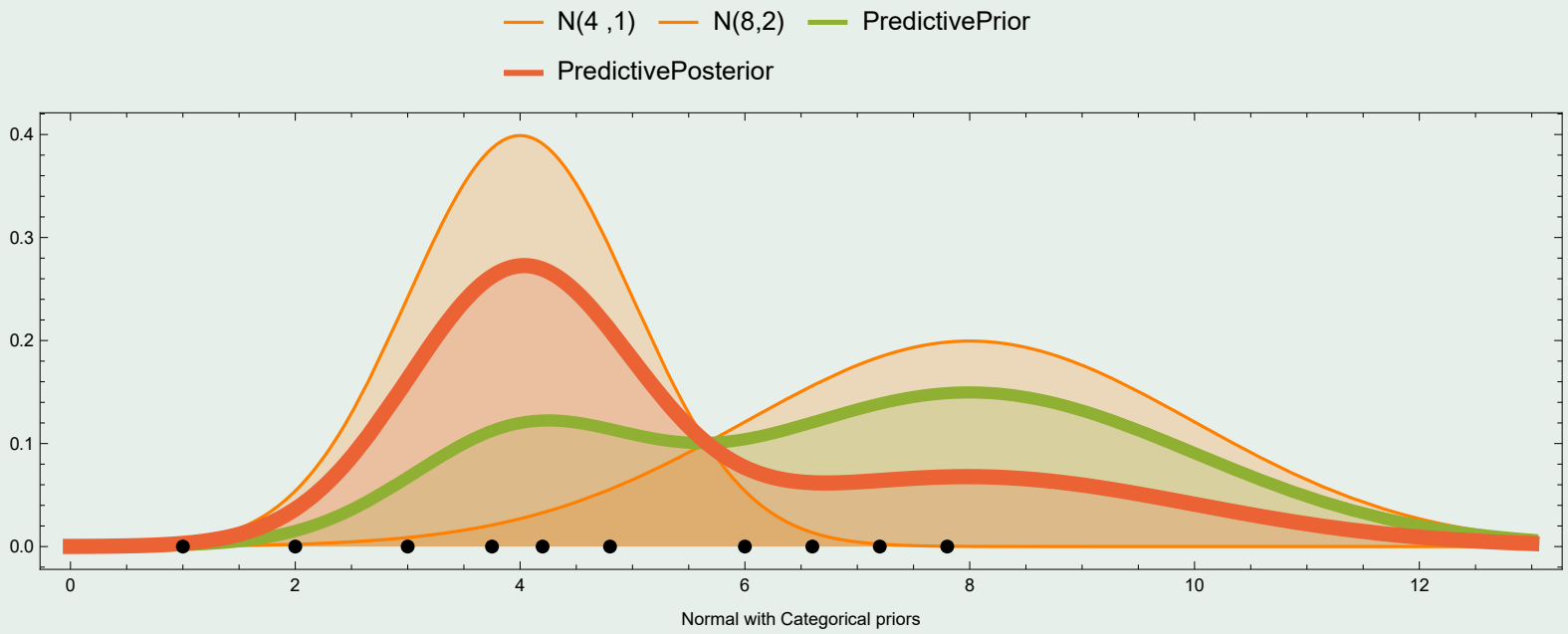


Illustration of how to update the uniform prior with one sample to obtain the posterior.

The posterior distribution is:

$$p(\theta|X, Cat(\theta_i)) \propto p_{Cat(\theta_i)}(\theta) \prod_{i=1,10} f_{N(\theta,1)}(x_i)$$

The posterior predictive distribution becomes

$$p(\tilde{x}|X, Cat(\theta_i)) = \sum_{i=1}^{2} p(\theta|X, Cat(\theta_i)) f_{N(\theta,1)}(\tilde{x}) d\theta$$