Bayesian Statistics
Prior

Shaobo Jin

Department of Mathematics

Prior Distribution

The main difference between a frequentist model and a Bayesian model is that the parameter of the data generating distribution is random and follows a known distribution (prior distribution). The parameters in a prior distribution are called the hyperparameters.

1. A subjective prior incorporates our prior knowledge.
2. An objective prior fulfills some desired (theoretical) properties.

It is in general very difficult to specify an exact prior distribution. Most critiques of Bayesian methods is specifying a prior distribution.

Subjective Prior: Expert Advise

Example

Suppose that we are interested in the effectiveness $\theta \in [0, 1]$ of a vaccine.

- An expert expects a 80% decrease in the number of disease cases among the group of vaccinated people compared to non-vaccinated group of people.
- Suppose that we would like to use a Beta (a, b) prior.
- The hyperparameters can be set such that the expectation of the beta distribution $\frac{a}{a+b}$ is close to 80%.

Subjective Prior: Previous Experiences

Example

Suppose that we want to predict the number of sold cups of coffee during midsommar celebration.

- Suppose that the sales records from previous years show that the number ranges between 600 and 800 cups.
- We can choose a prior distribution such that the majority of mass is close/within such range.

Mixture Prior Distribution: Example

Example

Suppose that we are interested in the temperature θ at the midsommar celebration.

- One expert guesses that the temperature is around 22°C, and another expert guesses 10°C.
- One example is to specify the temperate as

$$wN\left(22, \sigma_1^2\right) + (1-w)\, N\left(10, \sigma_2^2\right).$$

Conjugate Prior

Definition

Let $\mathcal{F}$ be a family of probability distributions on Θ. If $\pi(\cdot) \in \mathcal{F}$ and $\pi(\cdot \mid x) \in \mathcal{F}$ for every x, then the family of distributions $\mathcal{F}$ is conjugate. The prior distribution that is an element in a conjugate family is called a conjugate prior.

The main benefit of a conjugate prior is tractability, that is, we only need to update the hyperparameters without changing the family of distributions. It makes Bayesian computation much easier.

Conjugate Prior: Example

Example

1. Suppose that we have an iid sample $X_i \mid \theta \sim \text{Bernoulli}(\theta)$. Show that $\theta \sim \text{Beta}(a_0, b_0)$ is conjugate.
2. Suppose that we have an iid sample $X_i \mid \mu \sim N\left(\mu, \sigma^2\right)$, $i = 1, ..., n$, where σ^2 is known. Show that $\mu \sim N\left(\mu_0, \sigma_0^2\right)$ is conjugate.
3. Suppose that we have an iid sample $X_i \mid \mu, \sigma^2 \sim N\left(\mu, \sigma^2\right)$, $i = 1, ..., n$. Show that $\mu \mid \sigma^2 \sim N\left(\mu_0, \sigma^2/\lambda_0\right)$ and $\sigma^2 \sim \text{InvGamma}(a_0, b_0)$ form a conjugate prior, where

$$\pi\left(\sigma^2\right) = \frac{b_0^{a_0}}{\Gamma(a_0)} \frac{1}{\left(\sigma^2\right)^{a_0+1}} \exp\left(-\frac{b_0}{\sigma^2}\right).$$

Find Conjugate Prior

The likelihood $f(x \mid \theta)$ entirely determines the class of conjugate priors.

Example

Find the conjugate prior.

1. Suppose that we have an iid sample $X_i \mid \theta \sim \text{Poisson}(\theta)$.
2. Suppose that we have an iid sample

$$X_i \mid \theta \sim \text{Multinomial}(m, \theta_1, ..., \theta_k).$$

Exponential Family

Definition

A class of probability distributions $\mathcal{P} = \{\mathrm{P}_\theta : \theta \in \Theta\}$ is called an exponential family, if there exist a number $k \in \mathbb{N}$, real-valued functions A, ζ_1, .., ζ_k on Θ, real-valued statistics T_1, ..., T_k, and a function h on the sample space $\mathcal{X}$ such that

$$f(x \mid \theta) = A(\theta) \exp\left\{\sum_{j=1}^{k} \zeta_j(\theta) T_j(x)\right\} h(x),$$

where $A(\theta) > 0$ depends only on θ and $h(x) \geq 0$ depends only on x. We often denote the real valued functions by

$$\begin{aligned} \zeta(\theta) &= (\zeta_1 \quad \cdots \quad \zeta_k)^T, \\ T(x) &= (T_1 \quad \cdots \quad T_k)^T. \end{aligned}$$

Exponential Family: Example

Example (Normal distribution)

Normal distribution with $\theta = (\mu, \sigma^2)$:

$$f\left(x \mid \mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{\mu^2}{2\sigma^2}\right\} \exp\left\{-\frac{x^2}{2\sigma^2} + \frac{x\mu}{\sigma^2}\right\}.$$

Example (Binomial distribution)

Binomial distribution:

$$\mathrm{P}\left(X = x \mid \theta\right) = \exp\left\{x \log\left(\theta\right) + \left(n - x\right) \log\left(1 - \theta\right)\right\} \binom{n}{x}.$$

Exponential Family: Counterexample

1. Exponential distribution:

$$\begin{aligned} f(x \mid \theta) &= \theta \exp\{-\theta x\}, \quad x \geq 0 \\ &= \theta \exp\{-\theta x\} 1(x \geq 0), \end{aligned}$$

   where $1(\cdot)$ is the indicator function.

2. Shifted exponential distribution with $\theta = (\lambda, \mu)$:

$$\begin{aligned} f(x \mid \theta) &= \lambda \exp\{-\lambda(x-\mu)\}, \quad x \geq \mu \\ &= \lambda \exp\{\lambda\mu\} \exp\{-\lambda x\} 1(x \geq \mu). \end{aligned}$$

Natural Parameter

We can parameterize the probability function as

$$f(x \mid \zeta) = C(\zeta) \exp\left\{\sum_{j=1}^{k} \zeta_j T_j(x)\right\} h(x),$$

where ζ is called the natural parameter.

Example (Binomial distributin)

For $\theta \in (0,1)$,

$$f(x \mid \theta) = (1-\theta)^n \exp\left\{x \log\left(\frac{\theta}{1-\theta}\right)\right\} \binom{n}{x}.$$

Define $\zeta = \log\left(\frac{\theta}{1-\theta}\right) \in \mathbb{R}$. Then,

$$f(x \mid \zeta) = \left(1 - \frac{\exp(\zeta)}{1+\exp(\zeta)}\right)^n \exp\{x\zeta\} \binom{n}{x}.$$

Conjugate Prior for Exponential Family

Theorem

Suppose that

$$f(x \mid \zeta) = \exp\left\{\sum_{j=1}^{k} \zeta_j T_j(x) + \log C(\zeta)\right\} h(x).$$

Then the conjugate family for ζ is given by

$$\pi(\zeta) = K(\mu_0, \lambda_0) \exp\left\{\zeta^T \mu_0 + \lambda_0 \log C(\zeta)\right\},$$

where μ and λ are hyperparameters. The posterior satisfies

$$\pi(\zeta \mid x) \propto \exp\left\{\zeta^T [\mu_0 + T(x)] + (\lambda_0 + 1) \log C(\zeta)\right\}.$$

Conjugate Prior: Example

Example

Using the exponential family for the following examples.

1. Let $X_1, .., X_n$ be an iid from $N\left(\theta, \sigma^2\right)$, where σ^2 is known. Show that $\theta \sim N\left(\mu_0, \sigma_0^2\right)$ is conjugate.
2. Let $X_1, ..., X_n$ be an iid sample from Bernoulli (θ). Show that $\theta \sim \text{Beta}\,(a, b)$ is conjugate.
3. Suppose that $\{Y_i\}_{i=1}^n$ are independent observations such that

$$\mathrm{P}\left(Y_i = 1 \mid X_i = x_i\right) = \frac{\exp\left\{y_i\left(x_i^T\theta\right)\right\}}{1 + \exp\left(x_i^T\theta\right)}.$$

Find the conjugate prior for θ.

No Prior Information

When no prior information is available, we still need to specify a prior in order to use Bayesian modelling.

Definition

The Laplace prior is $\pi(\theta)$ is a constant over Θ.

Disambiguation: The name Laplace prior is often referred to as that θ follows a Laplace distribution

$$\pi(\theta) = \frac{1}{2b_0} \exp\left(-\frac{|\theta - a_0|}{b_0}\right), \quad -\infty < \theta < \infty.$$

The prior in the above definition is often referred to as the flat prior, uniform prior, among others.

Uniform Prior as Non-Informative Prior

Intuitively speaking, a constant $\pi(\theta)$ means that we treat all θ equally.

- The posterior depends only on the likelihood.

For a distribution P with density p, its entropy is

$$S(P) = -\mathrm{E}[\log p].$$

The entropy is often called the Shannon entropy if the random variable is discrete and the differential entropy if the random variable is continuous.

Example

Find the entropy of the following distributions.

1. $X \sim N(0, \sigma^2)$.
2. X is uniform on the finite discrete set $\{1, 2, ..., n\}$.

Uniform Distribution Maximizes Entropy

The entropy of a random variable measures its uncertainty.

- If a random variable puts majority of probability mass on one value, then the uncertainty is small.
- If the possible values of a random variable are equally alike, then the uncertainty is large.

Example

1. Suppose that X is a discrete random variable with a finite sample space $\{1, 2, ..., n\}$. Show that the discrete uniform distribution maximizes the Shannon entropy.
2. Suppose that X is a continuous random variable with a closed sample space $[a, b]$. Show that the continuous uniform distribution maximizes the differential entropy.

Improper Prior

The uniform prior is proportional to a density of a probability measure if the parameter space Θ is bounded.

However, in many cases, the prior is not a probability measure. Instead it yields

$$\int_{\Theta} \pi(\theta)\, d\theta = \infty.$$

Such prior distribution is said to be an improper prior.

- The uniform prior is an improper prior if Θ is not bounded.

But as long as the posterior distribution is well defined, the Bayesian methods still apply.

Improper Posterior

One risk of using improper prior is that the posterior can be undefined.

Example

Let $X \sim \text{Binomial}(n, \theta)$ and $\pi(\theta) \propto \frac{1}{\theta(1-\theta)}$. The posterior satisfies

$$\begin{aligned} \pi(\theta \mid x) &\propto \theta^x (1-\theta)^{n-x} \frac{1}{\theta(1-\theta)} \\ &= \theta^{x-1} (1-\theta)^{n-x-1}, \end{aligned}$$

which is not defined for $x = 0$ or $x = n$.

In order to have a well-defined posterior, we need

$$\int f(x \mid \theta) \pi(\theta) \, d\theta < \infty.$$

But this may not be an easy task to check.

Marginalization Paradox

Since the improper prior is not a probability density, the posterior, even exists, may not follow the rules of probability. One example is the marginalization paradox.

- Consider a model $f(x \mid \alpha, \beta)$ and a prior $\pi(\alpha, \beta)$. Suppose that the marginal posterior $\pi(\alpha \mid x)$ satisfies

$$\pi(\alpha \mid x) = \pi(\alpha \mid z(x))$$

  for some function $z(x)$.
- Suppose that $f(z \mid \alpha, \beta) = f(z \mid \alpha)$, that is, does not depend on β.
- If $\pi(\alpha, \beta)$ is a proper prior, we can recover $\pi(\alpha \mid x)$ from $f(z \mid \alpha)$ and some $\pi(\alpha)$ as $\pi(\alpha \mid x) \propto f(z \mid \alpha)\pi(\alpha)$.
- However, if $\pi(\alpha, \beta)$ is not a proper prior, it can happen that $f(z \mid \alpha)\pi(\alpha)$ is not proportional to $\pi(\alpha \mid x)$ for any $\pi(\alpha)$.

Marginalization Paradox: Example

Example

Let $X_1, ..., X_n$ be independent exponential random variables. The first m have mean η^{-1} and the rest have mean $(c\eta)^{-1}$, where $c \neq 1$ is a known constant and $m \in \{1, ..., n-1\}$.

- We consider the improper prior $\pi(\eta) = 1$ such that $\pi(\eta, m) = \pi(\eta)\pi(m) = \pi(m)$.
- The marginal posterior distribution satisfies

$$\pi(m \mid x) \quad \propto \quad \frac{c^{n-m}\pi(m)}{\left(\sum_{i=1}^{m} z_i + c\sum_{i=m+1}^{n} z_i\right)^{n+1}},$$

where $z_i = x_i/x_1$. Hence, the marginal posterior depends only on $z = (z_2, ..., z_n)$, since $z_1 = 1$.

Marginalization Paradox: Example

Example

$$\pi(m \mid x) \propto \frac{c^{n-m}\pi(m)}{\left(\sum_{i=1}^{m} z_i + c\sum_{i=m+1}^{n} z_i\right)^{n+1}}.$$

- The density of z is

$$f(z \mid \eta, m) = \frac{c^{n-m}\Gamma(n)}{\left(\sum_{i=1}^{m} z_i + c\sum_{i=m+1}^{n} z_i\right)^{n}} \equiv f(z \mid m),$$

  which only depends on m, not η.
- However, it is not possible to find a $\pi^*(m)$ such that

$$\pi(m \mid x) \propto f(z \mid m)\pi^*(m).$$

Invariance?

Another issue of the uniform prior is that it is not invariant against reparametrization.

- Suppose that we choose the uniform prior for $\theta \in \Theta$.
- Now we reparameterize to $\eta = \eta(\theta)$, which is one-to-one, such that $\theta = h(\eta)$. Then,

$$\pi_\eta(\eta) \quad = \quad \pi_\theta(h(\eta)) \left| \det \left[\frac{\partial h(\eta)}{\partial \eta^T} \right] \right|,$$

  which is not a constant.

A constant prior on θ does not always yield a constant prior on $\eta(\theta)$, even though η is a strictly monotone transformation.

Invariance: Example

Example (Binomial distributin)

Suppose that $X \mid \theta \sim \text{Binomial}(n, \theta)$. We have no information regarding θ. Hence we let $\theta \sim \text{Uniform}(0, 1)$.

- Consider the odds ratio $\zeta = \frac{\theta}{1-\theta}$.
- By change of variables,

$$\begin{aligned} \pi_\zeta(\zeta) &= \pi_\theta(h(\zeta)) \left| \det \left[\frac{\partial h(\eta)}{\partial \eta^T} \right] \right| = \left| \frac{\partial}{\partial \zeta} \frac{\zeta}{1+\zeta} \right| \\ &= \frac{1}{(1+\zeta)^2}. \end{aligned}$$

- Further, $\theta \sim \text{Uniform}(0, 1)$ is the same as $\theta \sim \text{Beta}(1, 1)$. The prior is conjugate. But the resulting prior for ζ is not.

Invariance Under Monotone Transformation

Suppose that a procedure of finding prior yields the prior density $\pi_\theta(\theta)$ for θ.

- Let h be a smooth and monotone transformation. By the change of variables $\eta = \eta(\theta)$ and $\theta = h(\eta)$, the density of η induced from $\pi_\theta(\theta)$ is

$$\pi_\theta(h(\eta)) \left| \frac{dh(\eta)}{d\eta} \right|.$$

- If we use the same procedure of finding prior as we used for θ, it should yield the prior density $\pi_\eta(\eta)$ for η.

Invariance means that such two densities should be the same:

$$\pi_\theta(h(\eta)) \left| \frac{dh(\eta)}{d\eta} \right| = \pi_\eta(\eta).$$

Motivation: Use of Fisher Information

Suppose that P and Q are two probability measures with densities p and q, respectively.

- The Kullback-Leibler divergence is

$$\mathrm{KL}(P,Q) = \int \log\left[\frac{p(x)}{q(x)}\right] p(x)\,dx.$$

- We consider the symmetric metric

$$\mathrm{KL}(P_\theta, P_{\theta'}) + \mathrm{KL}(P_{\theta'}, P_\theta).$$

- If we change the parametrization such that $\theta = h(\eta)$ using Fisher information, then parametrization leaves the distance between distributions approximately unchanged:

$$\mathrm{KL}(P_\theta, P_{\theta'}) + \mathrm{KL}(P_{\theta'}, P_\theta) = \mathrm{KL}\left(P_\eta, P_{\eta'}\right) + \mathrm{KL}\left(P_{\eta'}, P_\eta\right).$$

Jeffreys Prior

Definition

Consider a statistical model $f(x \mid \theta)$ with Fisher information matrix $\mathcal{I}(\theta)$. The Jeffreys prior is

$$\pi(\theta) \quad \propto \quad [\det(\mathcal{I}(\theta))]^{1/2}.$$

The Jeffreys prior is invariant to reparametrization under smooth monotone transformation, because we can show

$$\pi_\theta(h(\eta)) \left| \frac{dh(\eta)}{d\eta} \right| \quad = \quad \pi_\eta(\eta).$$

Jeffreys Prior: Example

Example

Find the Jeffreys prior for θ.

1. Suppose that $X \mid \theta \sim \text{Binomial}(n, \theta)$. Show also that the Jeffreys prior is invariant to the transformation $\eta = \frac{\theta}{1-\theta}$.
2. Suppose that $X_i \mid \theta \sim N(\theta, 1)$, $i = 1, ..., n$.
3. Suppose that $X_i \mid \theta$ belongs to a location family with density $f(x_i - \theta)$, where $f(x)$ is a density function.
4. Suppose that $X_i \mid \theta$ belongs to a scale family with density $\theta^{-1} f\left(\theta^{-1} x_i\right)$, where $f(x)$ is a density function and $\theta \in \mathbb{R}_+$.

Jeffreys Prior is Non-Informative

The Jeffreys prior is derived in order to achieve invariance. It turns out that it is also non-informative.

- Under the Jeffreys prior, the posterior can be approximated by

$$\begin{aligned}\pi(\theta \mid x) &\propto \pi(\theta) f(x \mid \theta) \\ &\approx [\det(\mathcal{I}(\theta))]^{1/2} \exp\left(-\frac{1}{2}\left(\theta-\hat{\theta}\right)^T \mathcal{I}(\theta)\left(\theta-\hat{\theta}\right)\right),\end{aligned}$$

  that is, $\theta \mid x \approx N\left(\hat{\theta}, \mathcal{I}^{-1}(\theta)\right)$.

- The frequentist approach yields $\hat{\theta}-\theta \approx N\left(0, \mathcal{I}^{-1}(\theta)\right)$.

Inference using the Jeffreys prior coincides approximately with the inference from the likelihood function.

Independent Jeffreys Prior

Example

Suppose that $X_i \mid \mu, \sigma^2 \sim N\left(\mu, \sigma^2\right)$, $i = 1, ..., n$. Find the Jeffreys prior for $\theta = \left(\mu, \sigma^2\right)$.

When we have multiple parameters, it is also common to use the independent Jeffreys prior.

- Obtain the Jeffeys prior for each parameter separately by fixing the others.
- Multiple the single parameter Jeffreys prior together.

Example

Suppose that $X_i \mid \mu, \sigma^2 \sim N\left(\mu, \sigma^2\right)$, $i = 1, ..., n$. Find the independent Jeffreys prior for $\theta = \left(\mu, \sigma^2\right)$.

A Cautious Note

The Jeffreys prior do not necessarily perform satisfactorily for all inferential purposes.

Example

Suppose that we observe one observation $X \mid \theta \sim N_p(\theta, I)$.

- The Jeffreys prior is the uniform prior and the posterior is $\theta \mid x \sim N_p(x, I)$.
- Suppose that we are interested in the parameter $\eta = \theta^T\theta$. The posterior distribution of η is noncentral χ^2 with p degrees of freedom. The posterior expected value is $x^Tx + p$.
- If we consider a quadratic loss, the loss of another estimator $x^Tx - p$ is no greater than the loss of $x^Tx + p$ for all θ.
- This means that for any θ, we can always find an estimator that is better than the estimator using the Jeffreys prior.

Reference Prior

Consider the Kullback-Leibler divergence

$$\mathrm{KL}\left(\pi\left(\theta \mid x\right), \pi\left(\theta\right)\right) = \int \pi\left(\theta \mid x\right) \log\left(\frac{\pi\left(\theta \mid x\right)}{\pi\left(\theta\right)}\right) d\theta \geq 0.$$

A large KL means that a lot information has come from the data.

The expected KL under the marginal of x is then

$$\begin{aligned}
\mathrm{E}\left[\mathrm{KL}\left(\pi\left(\theta \mid x\right), \pi\left(\theta\right)\right)\right] &= \int m\left(x\right) \left[\int \pi\left(\theta \mid x\right) \log\left(\frac{\pi\left(\theta \mid x\right)}{\pi\left(\theta\right)}\right) d\theta\right] dx \\
&= \int \int f\left(x, \theta\right) \log\left(\frac{\pi\left(\theta \mid x\right)}{\pi\left(\theta\right)}\right) d\theta dx \\
&= \int \int f\left(x, \theta\right) \log\left(\frac{f\left(x, \theta\right)}{\pi\left(\theta\right) m\left(x\right)}\right) d\theta dx,
\end{aligned}$$

where $m\left(x\right)$ is the marginal density of x.

Mutual Information

In probability theory, the mutual information of two random variables X and Y is defined as

$$\mathrm{MI}(X,Y) = \int\int f(x,y) \log\left(\frac{f(x,y)}{f(x)\,f(y)}\right) dxdy \quad \geq \quad 0.$$

It is a measure to quantity the information in $f(x,y)$ instead of $f(x)\,f(y)$.

- The expected KL in the previous slide

$$\mathrm{E}\left[\mathrm{KL}\left(\pi(\theta \mid x), \pi(\theta)\right)\right] \quad = \quad \int\int f(x,\theta) \log\left(\frac{f(x,\theta)}{\pi(\theta)\,m(x)}\right) d\theta dx$$

  is the mutual information of X and θ.

The reference prior aims to maximize the mutual information of the prior and posterior.

Reference Prior and Entropy

Result

Let $p(x)$ be the density of a distribution P. Then,

$$\mathrm{MI}(X, \theta) = S(\pi(\theta)) - \int m(x) S(\pi(\theta \mid x)) dx,$$

where

$$S(P) = -\mathrm{E}[\log p(X)]$$

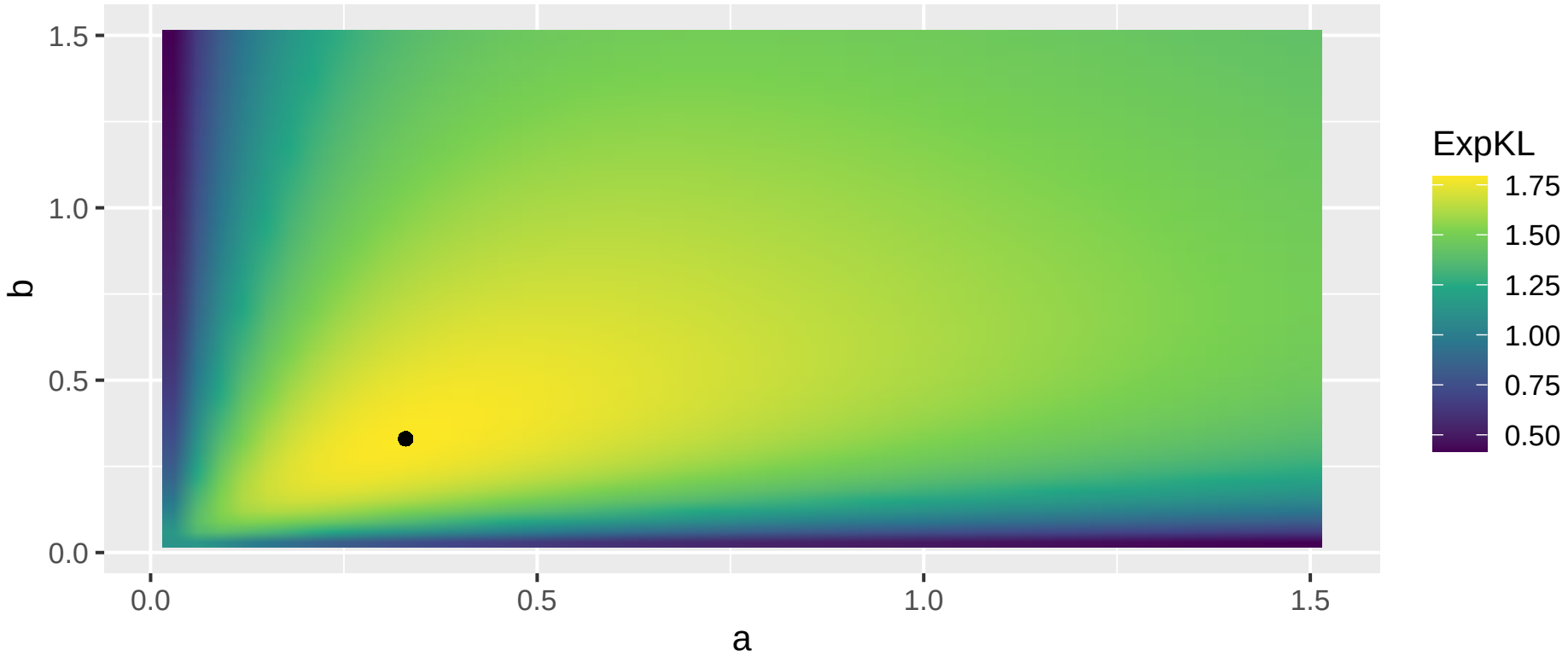
is the entropy.

Thus, a reference prior that generates a large mutual information corresponds to a prior with large entropy and a posterior with low expected entropy.

Reference Prior: Example

Example

Suppose that $X \mid \theta \sim \text{Binomial}(n, \theta)$ and we consider the class of conjugate priors $\theta \sim \text{Beta}(a, b)$. Find the expected KL.



Explicit Form of Reference Prior

Suppose that we can replicate the experiment independently k times. Each time we observe a data set of sample size n. Denote all realizations by $x = \left(x^{(1)}, ..., x^{(k)}\right)$.

Let $\pi^*(\theta)$ be a continuous and strictly positive function such that the posterior $\pi^*(\theta \mid x)$ is proper and asymptotically consistent. For any interior point θ_0 of Θ, define

$$\begin{aligned} p_k(\theta) &= \exp\left\{\int f(x \mid \theta) \log \pi^*(\theta \mid x)\, dx\right\}, \\ p(\theta) &= \lim_{k\to\infty} \frac{p_k(\theta)}{p_k(\theta_0)}, \end{aligned}$$

Suppose $p_k(\theta)$ is continuous for all k. Then, under some extra assumptions on the ratio $p_k(\theta)/p_k(\theta_0)$ and on $p(\theta)$, $p(\theta)$ is a reference prior.

Approximate Reference Prior

Find the reference prior is not an easy task, since the integrals can be difficult to evaluate.

Algorithm 1: Approximate reference prior

Choose an arbitrary continuous and positive function $\pi^*(\theta)$, e.g., $\pi^*(\theta) = 1$;
for *any* θ *of interest including a* θ_0 **do**
  for j *from* 1 *to* m **do**
    Simulate independently $\left\{x_j^{(1)}, ..., x_j^{(k)}\right\}$ from $f(x \mid \theta)$;
    Compute the integral $c_j = \int_\Theta \left[\prod_{i=1}^k f\left(x_j^{(i)} \mid \theta\right)\right] \pi^*(\theta)\, d\theta$ analytically or approximate numerically ;
    Evaluate $r_j(\theta) = \log\left\{\left[\prod_{i=1}^k f\left(x_j^{(i)} \mid \theta\right)\right] \pi^*(\theta) / c_j\right\}$;
  end
  Compute $p_k(\theta) = \exp\left\{m^{-1} \sum_{j=1}^m r_j(\theta)\right\}$;
  Let $\pi(\theta) \propto p_k(\theta) / p_k(\theta_0)$;
end

Reference Prior and Jeffreys Prior: Example

Example

Suppose that $X \mid \theta \sim \text{Binomial}(n, \theta)$. Approximate the reference prior.

In fact, if the distribution of MLE $\sqrt{n}\left(\hat{\theta} - \theta\right)$ can be approximated by $N\left(\theta, \mathcal{I}^{-1}(\theta)\right)$, and the posterior distribution of $\sqrt{n}\left(\theta - \hat{\theta}\right)$ can be approximately $N\left(0, \mathcal{I}^{-1}(\theta)\right)$, then, the reference prior and the joint Jeffreys prior are asymptotically equivalent.

Reference Prior With Presence of Nuisance Parameter

Suppose that $x \mid \theta \sim f(x \mid \theta_1, \theta_2)$ and $\theta = (\theta_1, \theta_2)$, where θ_1 is the parameter of interest and θ_2 is the nuisance parameter. The reference prior is obtained as follows.

- First, treating θ_1 as fixed. Use the Jeffreys prior associated with $f(x \mid \theta_2)$ as $\pi(\theta_2 \mid \theta_1)$.
- Then, derive the marginal distribution

$$f(x \mid \theta_1) = \int f(x \mid \theta_1, \theta_2)\, \pi(\theta_2 \mid \theta_1)\, d\theta_2.$$

Compute the Jeffreys prior $\pi(\theta_1)$ associated with $f(x \mid \theta_1)$.

Neyman-Scott Problem: Example

Consider the Neyman-Scott problem, where $X_{ij} \mid \theta \sim N\left(\mu_{ij}, \sigma^2\right)$, $i = 1, ..., n$ and $j = 1, 2$. We are interested in σ and μ_{ij}'s are nuisance parameters.

- The usual Jeffreys prior is $\pi(\theta) \propto \sigma^{-n-1}$. The posterior mean of σ^2 is

$$\mathrm{E}\left[\sigma^2 \mid x\right] = \frac{\sum_{i=1}^{n}\left(x_{i1} - x_{i2}\right)^2}{4n-4} \xrightarrow{P} \frac{\sigma^2}{2} \neq \sigma^2,$$

  where $\xrightarrow{P}$ means convergence in probability.
- The reference prior is $\pi(\theta) \propto \sigma^{-1}$. The posterior mean of σ^2 is

$$\mathrm{E}\left[\sigma^2 \mid x\right] = \frac{\sum_{i=1}^{n}\left(x_{i1} - x_{i2}\right)^2}{2n-4} \xrightarrow{P} \sigma^2.$$

Berger-Bernardo Method

The idea of deriving the prior conditioning on a subset of parameter can be applied to a general setting with more than two sets of parameters. The resulting method is the Berger-Bernado method.

Suppose the $p \times 1$ vector θ is partitioned into m groups, denoted by θ_1,..., θ_m. The reference prior is obtained in a similar manner to

$$\pi(\theta) \propto \pi(\theta_m \mid \theta_1, ..., \theta_{m-1}) \pi(\theta_{m-1} \mid \theta_1, ..., \theta_{m-2}) \cdots \pi(\theta_2 \mid \theta_1) \pi(\theta_1).$$

Berger-Bernardo Method: Algorithm

Algorithm 2: Berger-Bernardo method

1 Initiate some $\pi_m(\theta_m \mid \theta_1, ..., \theta_{m-1})$, e.g., Jeffreys prior ;

2 **for** j *in* $m-1, m-2, ..., 1$ **do**

3 Obtain the marginal distribution

$$f(x \mid \theta_1, ..., \theta_j) = \int f(x \mid \theta)\, \pi_{j+1}(\theta_{j+1}, ..., \theta_m \mid \theta_1, ..., \theta_j)\, d(\theta_{j+1}, ..., \theta_m).$$

4 Determine the reference prior $h_j(\theta_j \mid \theta_1, ..., \theta_{j-1})$ related to the model $f(x \mid \theta_1, ..., \theta_j)$, where $\theta_1, ..., \theta_{j-1}$ is treated as fixed ;

5 Compute $\pi_j(\theta_j, ..., \theta_m \mid \theta_1, ..., \theta_{j-1})$ by

$$\pi_j(\theta_j, ..., \theta_m \mid \theta_1, ..., \theta_{j-1}) \propto \pi_{j+1}(\theta_{j+1}, ..., \theta_m \mid \theta_1, ..., \theta_j)\, h_j(\theta_j \mid \theta_1, ..., \theta_{j-1}).$$

6 **end**

7 Obtain the reference prior $\pi(\theta) = \pi_1(\theta_1, ..., \theta_m)$;

Berger-Bernardo Method: Example

Example

Consider $X \mid \theta \sim \text{Multinomial}\,(n, \theta_1, ..., \theta_4)$. The likelihood is

$$f\,(x \mid \theta_1, \theta_2, \theta_3) \quad = \quad \frac{n!}{x_1!x_2!x_3!x_4!}\theta_1^{x_1}\theta_2^{x_2}\theta_3^{x_3}\theta_4^{x_4},$$

where $\theta_4 = 1 - \sum_{i=1}^{3} \theta_i$. Find the reference prior of $\theta = (\theta_1, \theta_2, \theta_3)$, where $m = 3$.

Influence of Prior

The assessment of the influence of the prior is called sensitivity analysis. In general, the prior can have a big impact for small sample sizes.

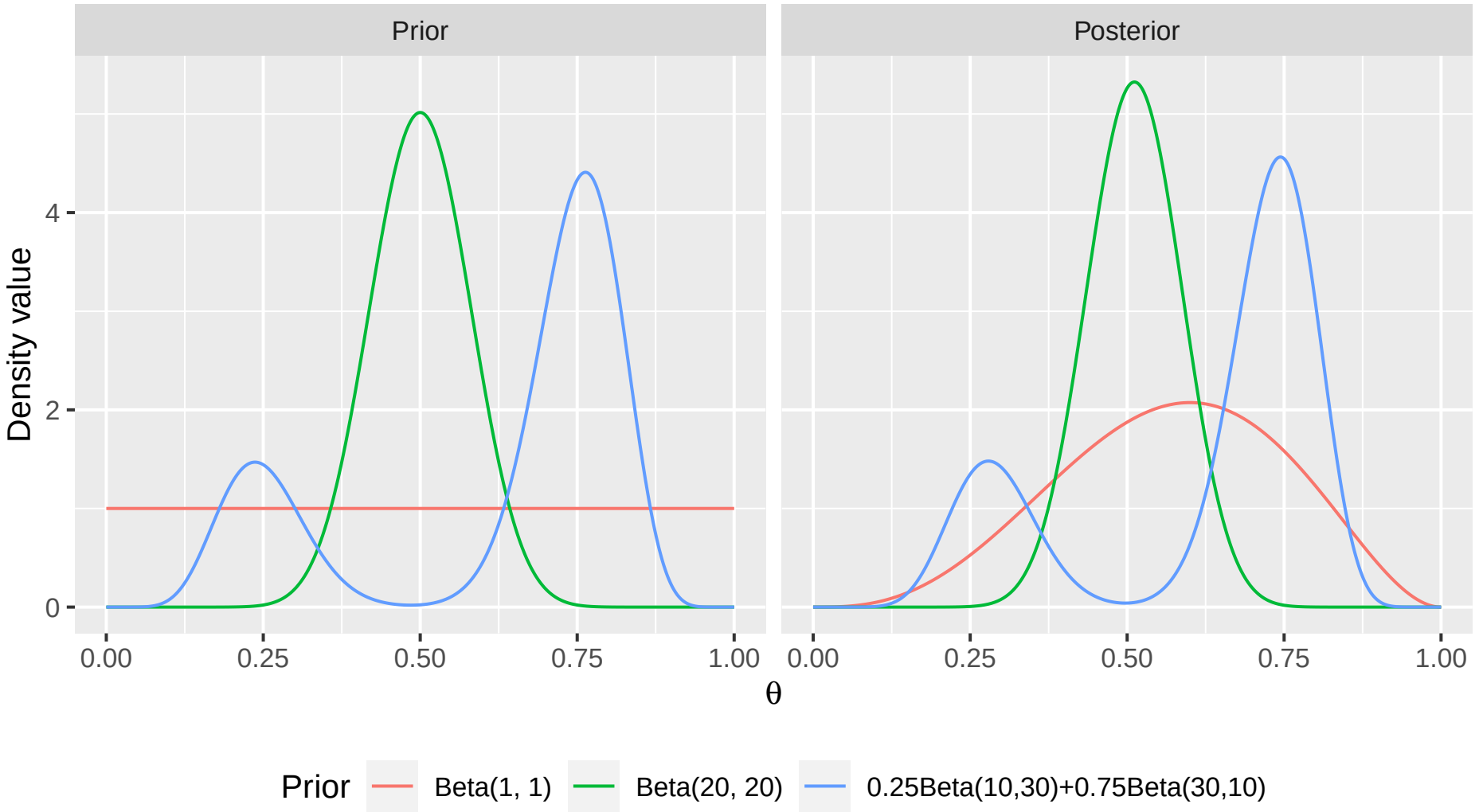
- But it becomes less important as the sample size increases. Most priors will lead to similar inference that is equivalent to the one based only on the likelihood.

Example

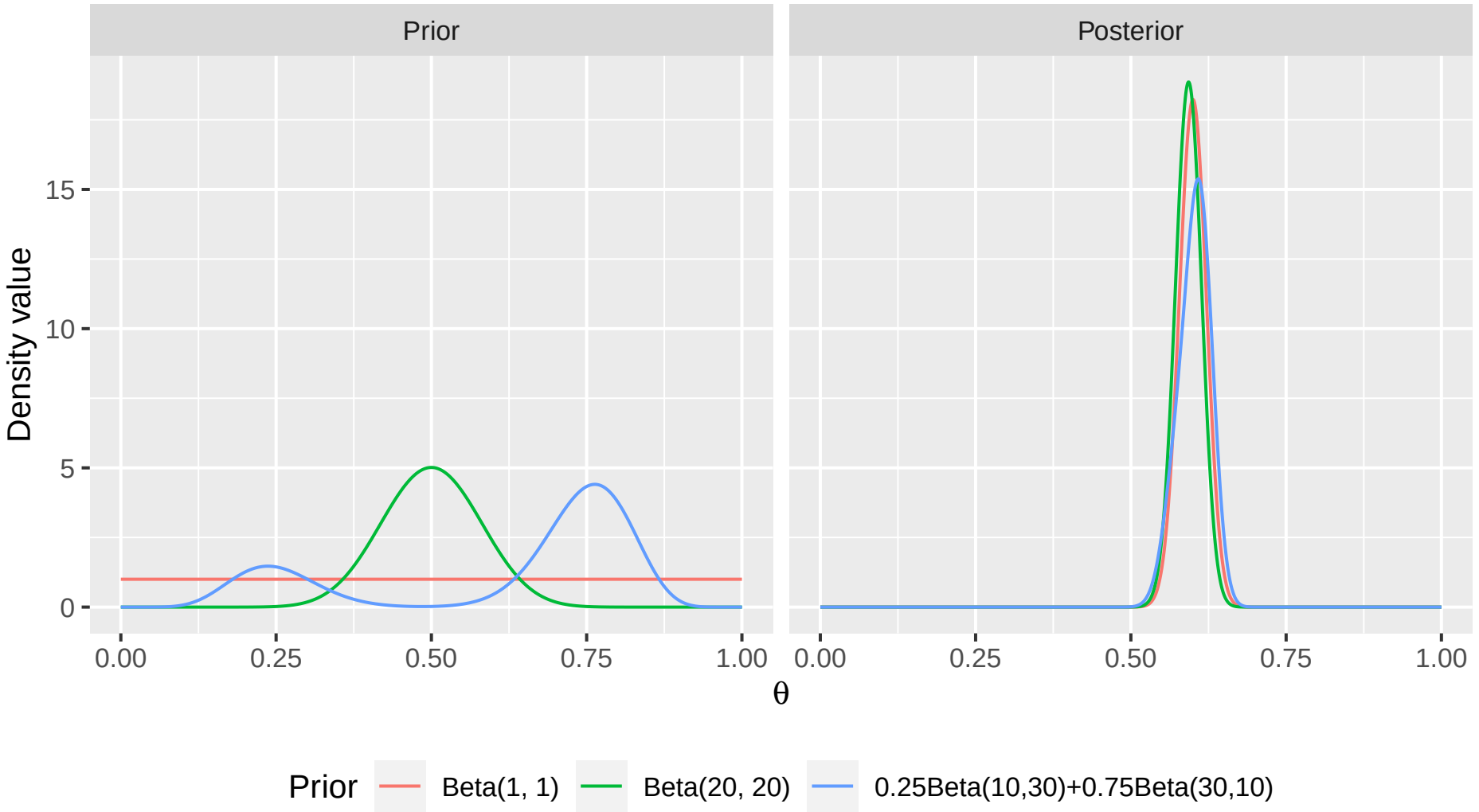
Suppose that we have an iid sample $X_i \mid \theta \sim \text{Bernoulli}(\theta)$.The conjugate prior $\theta \sim \text{Beta}(a_0, b_0)$ yields the posterior

$$\text{Beta}\left(a_0 + \sum_{i=1}^{n} x_i, b_0 + n - \sum_{i=1}^{n} x_i\right).$$

Small Sample Size



Large Sample Size



Hierarchical Prior Distribution

We can apply a hierarchical prior, applying a prior on the prior.

- Suppose that $\pi_1(\theta \mid \lambda)$ is a conjugate prior for $f(x \mid \theta)$, where λ is the hyperparameter.
- Instead of specifying the value of λ, we let

$$\lambda \sim \pi_2(\lambda), \ \theta \mid \lambda \sim \pi_1(\theta \mid \lambda), \ x \mid \theta \sim f(x \mid \theta).$$

- For example, if $X \mid z \sim N(\mu, z\sigma^2)$ and z is inverse gamma, then X follows a t distribution.
- A t distribution prior with low degrees of freedom (e.g., 3) is a popular choice.

Different Priors in Practice

We have introduced different ways of constructing the prior, e.g., conjugate prior, uniform prior, Jeffreys prior, and reference prior. Depending on how much information the priors contain, we can roughly partition the prior into the following groups according to their level of informative relative to the likelihood:

1. noninformative flat prior,
2. super-vague but proper prior, e.g., a prior with a massive variance such as $1,000,000$,
3. very weakly informative prior, e.g., a prior with a sizable variance such as 10,
4. weakly informative prior, e.g., a prior with variance 1,
5. informative prior.

The first two groups are generally not recommended.

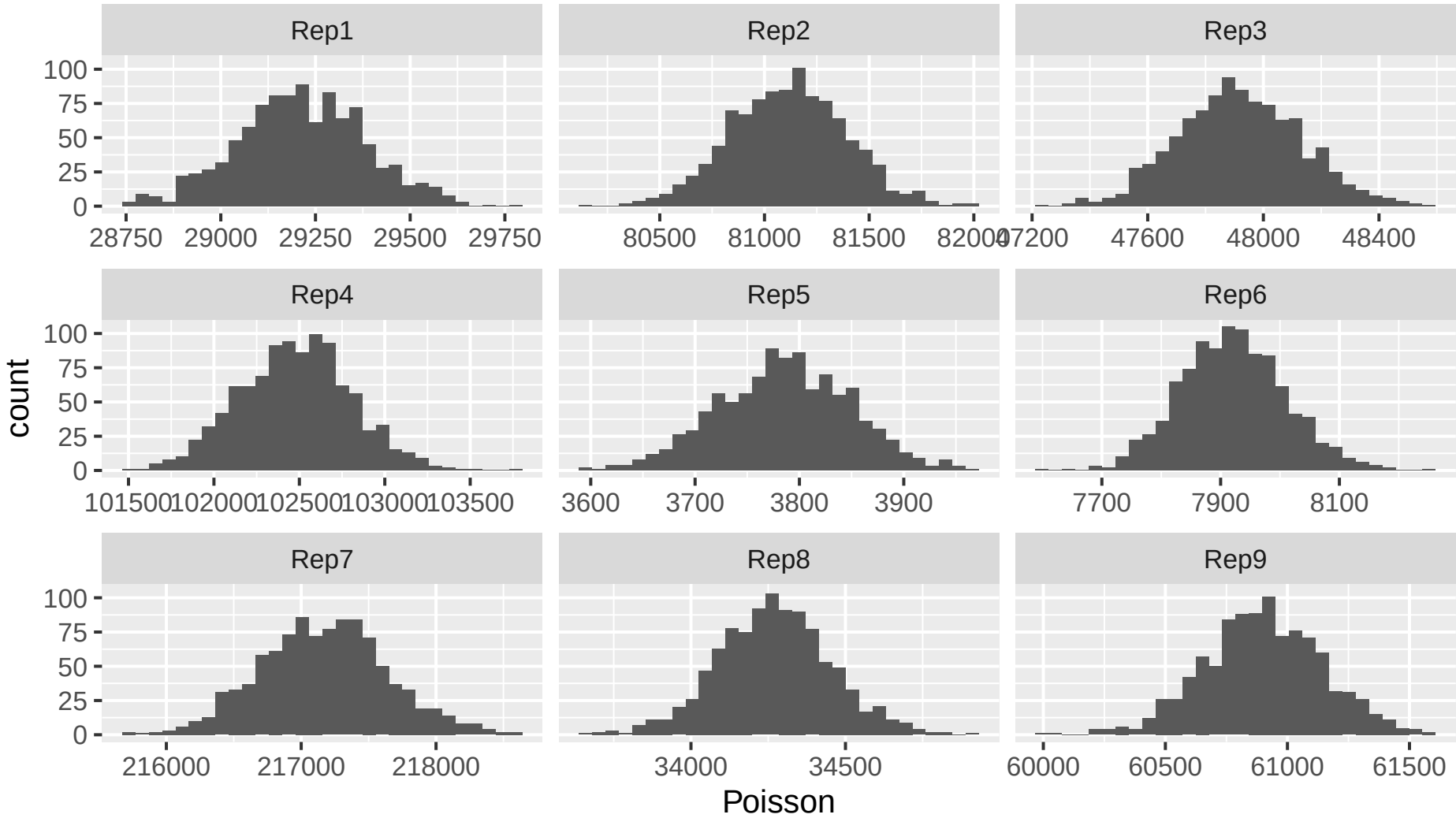
Prior Predictive Check

The prior predictive check is a way to assess whether your prior is appropriate.

Algorithm 3: Prior predictive check

1 **for** j *in* 1, 2, ..., m **do**
2 | Simulate $\theta_{\text{sim}} \sim \pi(\theta)$;
3 | Simulate $x_{\text{sim}} \sim f(x \mid \theta_{\text{sim}})$ of sample size n ;
4 **end**
5 Visualize each data set or investigate the summary statistics to judge whether the simulated data are plausible to avoid super bad priors.

Prior 1


Rep1
Rep2
Rep3
Rep4
Rep5
Rep6
Rep7
Rep8
Rep9
count
Poisson
100
75
50
25
0
28750 29000 29250 29500 29750
80500 81000 81500 82000
47200 47600 48000 48400
101500 102000 102500 103000 103500
3600 3700 3800 3900
7700 7900 8100
216000 217000 218000
34000 34500
60000 60500 61000 61500

Prior 2