Bayesian Inference: Conjugate Priors and the Exponential Family Perspective

Søren Lund Pedersen

1 Introduction

In this article, I would like to discuss the general approach of Bayesian statistics, including what it is, the differences between frequentist and Bayesian paradigms, the nature of inference, and details on various priors. A particular focus will be placed on conjugate priors, especially in the context of exponential family distributions and dispersion models, illustrating how they simplify computations and model updating.

2 Difference Between Frequentist and Bayesian Approaches

The frequentist approach, which is most commonly taught in schools, relies solely on data to make inferences about the world. This approach works well with large and diverse data samples. However, in situations with limited data or sequential data arrival, incorporating prior knowledge becomes crucial.

Bayesian statistics allows us to incorporate prior knowledge and update that knowledge with observed data, expressing uncertainty through posterior quantities. This framework is particularly valuable when data are scarce or when expert knowledge is available to guide the analysis.

3 The Basics of Bayesian Inference

Bayesian statistics is built on Bayes' theorem. Let θ be an unknown parameter of interest and x be the observed data. The posterior distribution is given by:

$$p(\theta \mid x) = \frac{p(x \mid \theta)\, p(\theta)}{p(x)},$$

where

- $p(x \mid \theta)$ is the likelihood,
- $p(\theta)$ is the prior distribution representing our knowledge before observing the data,
- $p(x)$ is the evidence or marginal likelihood, acting as a normalizing constant,
- $p(\theta \mid x)$ is the posterior distribution, encoding updated beliefs after seeing the data.

4 Maximum a Posteriori (MAP) Estimation

In frequentist statistics, we use Maximum Likelihood Estimation (MLE) to estimate parameters by maximizing $P(x \mid \theta)$. In Bayesian statistics, we use Maximum a Posteriori (MAP) estimation, which incorporates both the data and prior beliefs:

$$\hat{\theta}_{MAP} = \arg\max_{\theta} P(\theta \mid x).$$

Using Bayes' theorem, this becomes:

$$\hat{\theta}_{MAP} = \arg\max_{\theta} [P(x \mid \theta)\, P(\theta)],$$

which balances the likelihood $P(x \mid \theta)$ and the prior $P(\theta)$.

While frequentists provide confidence intervals—interpreted as the proportion of intervals that would contain the true parameter if the study were repeated—the Bayesian approach gives credible intervals. A 95% credible interval means there is a 95% probability that the true parameter lies within the interval, given the data and the prior. Formally, we find an interval $[a, b]$ such that:

$$\int_a^b P(\theta \mid x)\, d\theta = 0.95.$$

5 Bayesian Learning and Posterior Distribution

Once a Bayesian model is established with a likelihood, a prior, and evidence, we can learn from the posterior distribution and update our model as new data arrive. Given a new observation x_{new}, the posterior predictive distribution (PPD) is:

$$p(x_{new} \mid x) = \int p(x_{new} \mid \theta)\, p(\theta \mid x)\, d\theta,$$

where

- $p(x_{new} \mid \theta)$ is the likelihood of the new observation given θ,
- $p(\theta \mid x)$ is the posterior from the previous data.

As new data arrive, we update our model:

$$p(\theta \mid x_{old}, x_{new}) = \frac{p(x_{new} \mid \theta)\, p(\theta \mid x_{old})}{p(x_{new} \mid x_{old})},$$

allowing for sequential learning where the model gradually relies more on the likelihood as data accumulates, and the influence of the prior diminishes.

6 Hyperpriors

In Bayesian statistics, selecting an appropriate prior is crucial for meaningful inference. Sometimes a prior has its own unknown parameters, called hyperparameters. A **hyperprior** is a prior placed on these hyperparameters, allowing the model to learn their values from the data. For example:

$$\theta \mid \sigma^2 \sim N(\mu, \sigma^2), \quad \sigma^2 \sim \text{Inv-Gamma}(\alpha, \beta).$$

Here, α and β are hyperparameters of the inverse-gamma prior on σ^2, and we can assign hyperpriors to α and β if needed.

7 Conjugate Priors and the Exponential Family

A **conjugate prior** is a prior distribution that, when combined with a likelihood from a particular family, yields a posterior distribution of the same family. This property greatly simplifies computation and interpretation.

Examples include:

1. **Beta-Binomial**: Binomial likelihood with a Beta prior yields a Beta posterior.
2. **Gamma-Poisson**: Poisson likelihood with a Gamma prior yields a Gamma posterior.
3. **Normal-Inverse Gamma**: Normal likelihood with unknown mean and variance, using a Normal-Inverse Gamma prior, results in a posterior of the same family.

Natural conjugate priors are particularly appealing because they align with the natural parameters of the exponential family form of the likelihood. If the likelihood $p(x \mid \theta)$ belongs to an exponential family, then the conjugate prior can often be expressed in a similar exponential form, leading to a posterior that remains in the same family.

For instance, consider a likelihood in the exponential family:

$$p(x \mid \theta) = h(x) \exp\bigl(\theta T(x) - A(\theta)\bigr).$$

A conjugate prior for θ might take the form:

$$p(\theta \mid \xi, \nu) \propto \exp\bigl(\theta \xi - \nu A(\theta)\bigr),$$

where ξ and ν are hyperparameters. When combined with the likelihood, the posterior distribution will also be of the same form, with updated hyperparameters that incorporate the data. This conjugacy ensures that Bayesian updating is mathematically tractable and computationally efficient, especially when dealing with exponential dispersion families where dispersion parameters adjust variance independently of the mean.

8 Objective Priors

Critics of Bayesian methods often point out the subjectivity inherent in choosing priors. To mitigate this, one can use **objective** or **noninformative priors**, which are designed to have minimal influence on the posterior. A common choice is a uniform prior, but its appropriateness can depend on parameterization.

Jeffreys' Prior is a popular objective prior, invariant under reparameterization. It is defined as:

$$p(\theta) \propto \sqrt{\det(I(\theta))},$$

where $I(\theta)$ is the Fisher information matrix. Jeffreys' Prior distributes prior mass in a way that is consistent across different parameterizations, making it a default choice in the absence of strong prior beliefs.

9 Improper Priors

An **improper prior** does not integrate to 1 over its domain (e.g., $p(\theta) \propto 1$ over $\mathbb{R}$). While improper priors are not valid probability distributions, they can sometimes be used in Bayesian analysis if the resulting posterior is proper (normalizable). However, care must be taken to ensure that the posterior distribution is well-defined and that inferences drawn from it are meaningful.