10424/624: Recitation 2

January 24, 2025

1 Conjugate Priors

Suppose k has a Poisson distribution with unknown rate parameter λ

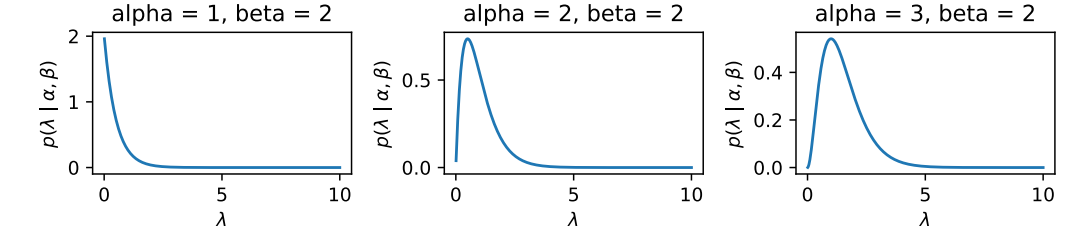
$$\Pr(k \mid \lambda) = \frac{\lambda^k e^{-\lambda}}{k!} \qquad k = 1, 2, \ldots$$

The Poisson distribution is used for modeling the number of times an event occurs within a fixed time interval given a mean occurrence rate assuming that the occurrences are independent.
Let the prior for λ be a gamma distribution with shape parameter α and rate parameter β:

$$p(\lambda \mid \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda} \qquad \lambda > 0$$

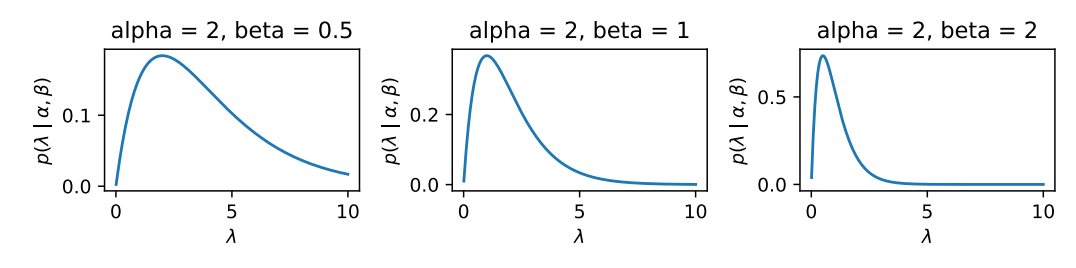
where Γ is the gamma function.

1. Plot the gamma distribution over the domain $(0, 10]$ with the following parameters: $(\alpha = 1, \beta = 2), (\alpha = 2, \beta = 2)$ and $(\alpha = 3, \beta = 2)$. How does increasing α affect our belief about λ?



As α increases, our belief increasingly moves towards greater values of λ. Our prior belief is also more spread out, and this is also shown by the fact that the mode of the distribution $p(\lambda \mid \alpha, \beta)$ decreases as α increases.

2. Now, plot the gamma distribution over the domain $(0, 10]$ with the following parameters: $(\alpha = 2, \beta = 0.5), (\alpha = 2, \beta = 1)$ and $(\alpha = 2, \beta = 2)$. How does increasing β affect our belief about λ?



As β increases, our prior belief on λ moves closer to zero. The mode of the prior distribution also becomes larger, indicating that our belief is more concentrated. The different plots are more similar to each other when compared to those obtained by changing α.

3. Show that given an observation k, the posterior $p(\lambda \mid k, \alpha, \beta)$ is a gamma distribution with updated parameters (α', β').

We have

$$\begin{aligned}\Pr(k \mid \lambda) \cdot p(\lambda \mid \alpha, \beta) &= \frac{\lambda^k e^{-\lambda}}{k!} \cdot \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda} \\ &= \frac{\beta^\alpha}{\Gamma(\alpha) \cdot k!} \lambda^{\alpha+k-1} e^{-(\beta+1)\lambda}\end{aligned}$$

We observe that this resembles to a Gamma distribution with parameters $\alpha' = \alpha + k$ and $\beta' = \beta + 1$. Indeed,

$$\int_0^\infty \lambda^{\alpha+k-1} e^{-(\beta+1)\lambda} = \frac{\Gamma(\alpha+k)}{(\beta+1)^{\alpha+k}}$$

$$\implies \Pr(k \mid \alpha, \beta) = \int_0^\infty \Pr(k \mid \lambda) \cdot p(\lambda \mid \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha) \cdot k!} \cdot \frac{\Gamma(\alpha+k)}{(\beta+1)^{\alpha+k}}$$

Thus, by Bayes' theorem,

$$\begin{aligned}p(\lambda \mid k, \alpha, \beta) &= \frac{\Pr(k \mid \lambda) \cdot p(\lambda \mid \alpha, \beta)}{\Pr(k \mid \alpha, \beta)} \\ &= \frac{(\beta+1)^{\alpha+k}}{\Gamma(\alpha+k)} \lambda^{\alpha+k-1} e^{-(\beta+1)\lambda}\end{aligned}$$

Thus, the posterior is a Gamma distribution with parameters $\alpha' = \alpha + k$ and $\beta' = \beta + 1$.

4. Suppose we receive a set of n observations $\mathcal{D} = \{k_1, k_2, \ldots, k_n\}$. We again start with a gamma prior for λ with parameters α and β and we update our belief on λ after each observation $k_i \in \mathcal{D}$.

   What is the posterior $p(\lambda \mid \mathcal{D}, \alpha, \beta)$? What is the posterior mean? What is the posterior mode?

Since the samples are iid, the likelihood is

$$\begin{aligned}\Pr(\mathcal{D} \mid \alpha) &= \prod_{i=1}^{n} \Pr(k_i \mid \alpha, \beta) \\ &= \prod_{i=1}^{n} \frac{\lambda^{k_i}}{k_i!} e^{-\lambda} \\ &= \frac{\lambda^{\sum_{i=1}^{n} k_i}}{K} e^{-n\lambda},\end{aligned}$$

where $K = \prod_{i=1}^{n} k_i!$. By Bayes' theorem,

$$\begin{aligned}p(\lambda \mid \mathcal{D}, \alpha, \beta) &\propto \Pr(\mathcal{D} \mid \lambda) \cdot p(\lambda \mid \alpha, \beta) \\ &= \frac{\lambda^{\sum_{i=1}^{n} k_i}}{K} e^{-n\lambda} \cdot \frac{\beta^{\alpha}}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda} \\ &= \frac{\beta^{\alpha}}{\Gamma(\alpha) \cdot K} \lambda^{\alpha-1+\sum_{i=1}^{n} k_i} e^{-(\beta+n)\lambda},\end{aligned}$$

which corresponds to a Gamma distribution with updated parameters $\alpha' = \alpha + \sum_{i=1}^{n} k_i$ and $\beta' = \beta + n$.
From the Wikipedia entry on Gamma distribution, we have

$$\begin{aligned}\text{Posterior mean:} \quad & \frac{\alpha + \sum_{i=1}^{n} k_i}{\beta + n} \\ \text{Posterior mode:} \quad & \frac{(\alpha + \sum_{i=1}^{n} k_i - 1)_+}{\beta + n},\end{aligned}$$

where $(x)_+ = \max\{x, 0\}$

5. In light of these results, can you give an interpretation of the prior parameters α and β? What happens in the limit as $n \to \infty$?

We can interpret β as the number of 'pseudo-samples' seen before the experiment, and α to be the sum of the k_i's observed in those pseudo-experiments. Our prior belief is as if we have observed some instances of the experiment beforehand.
As $n \to \infty$, the denominator is increasingly dominated by n, and the numerator is increasingly dominated by $\sum_{i=1}^{n} k_i$. Thus, the posterior mean and mode both become close to the empirical mean. Our posterior for λ tends to the empirical mean.

2 Maximum likelihood - Interpretation

Short answer: Suppose you flip a coin with unknown bias θ, $\Pr(x = \text{H} \mid \theta) = \theta$, three times and observe the outcome HHH. What is the maximum likelihood estimator for θ? Do you think this is a good estimator? Would you want to use it to make predictions?

Let h_1, h_2, h_3 be the random variables for the outcomes of the three tosses. Let $h_1 = 1$ if $x_1 = \text{H}$ and 0 otherwise. The likelihood for the coin is given by

$$\Pr(h_1, h_2, h_3 \mid \theta) = \theta^{\sum_{i=1}^{3} h_i}(1-\theta)^{3-\sum_{i=1}^{3} h_i}$$

Since we obtain three heads, the likelihood of this outcome is

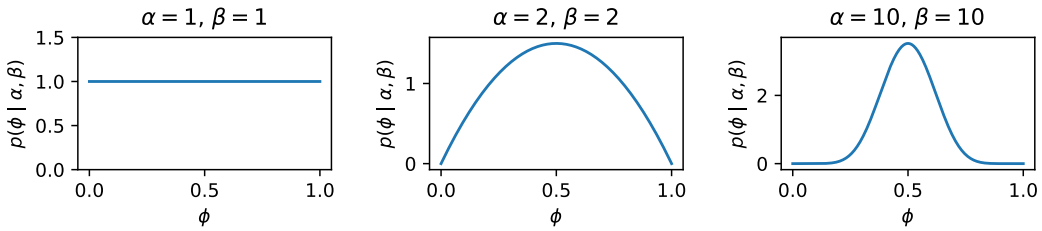
$$\Pr(h_1 = 1, h_2 = 1, h_3 = 1 \mid \theta) = \theta^3$$

The likelihood is maximized by $\theta = 1$, which would be our maximum likelihood estimator. This estimator is not very trustworthy, since we have tossed the coin only thrice, and obtaining three heads is not that unlikely. For instance, even with an unbiased coin, we can obtain three consecutive heads with probability 1/8, which is quite significant.

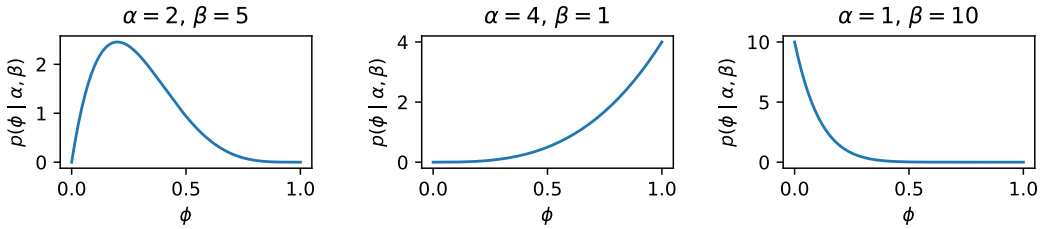
3 Beta distribution - Intuition

The PDF for a Beta distribution is given by

$$P(\phi \mid \alpha, \beta) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

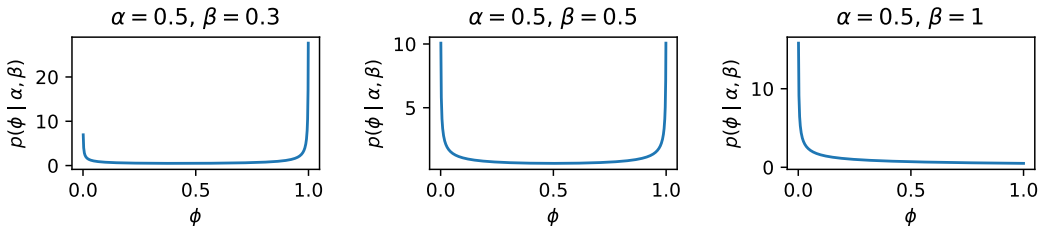


For $\alpha = \beta$, we observe that the PDF is symmetric about 0.5. As $\alpha = \beta$ increases, the function becomes more concentrated about 0.5.



For $\alpha < \beta$, we observe that the PDF is skewed towards lower values of ϕ, whereas for $\beta > \alpha$, the PDF is skewed towards higher values. As α becomes smaller and β increases, the skew to the left becomes more prominent. A similar property holds for a skew to the right, with the skew increasing as α becomes larger and β smaller.

Intuition about α and β. What does a Beta prior for the bias θ of a coin *mean*? Here, the bias of the coin is simply taken to be the probability of obtaining heads. A Beta prior $\mathcal{B}(\alpha, \beta)$ is like having information about $\alpha + \beta$ 'pseudo' coin tosses, out of which α were obtained to be heads and the remaining were tails. Indeed, a smaller α and larger β moves our prior belief about the bias towards lower probabilities. On the other hand, larger α and β means that our belief is skewed towards a high probability of obtaining heads. Let us now take a look at some pathological cases of (α, β):



We observe that for $\alpha < 1$ or $\beta < 1$, the distribution is highly skewed. If both α and β are less than 1, the distribution is bimodal, with very concentrated modes at zero and one.

4 Conditional Distribution for 2-dimensional Gaussian distribution

Consider a multivariate Gaussian distribution in two variables. That is, we have

$$X \sim \mathcal{N}(\mu, \Sigma),$$

with $X, \mu \in \mathbb{R}^2$, and $\Sigma \in \mathbb{R}^{2\times 2}$. Let

$$\mu = \begin{bmatrix}\mu_1 \\ \mu_2\end{bmatrix}, \quad \text{and} \quad \Sigma = \begin{bmatrix}\sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22}\end{bmatrix}$$

Here, $\sigma_{11} = \mathbb{V}(X_1)$, $\sigma_{22} = \mathbb{V}(X_2)$, $\sigma_{12} = \mathrm{Cov}(X_1, X_2) = \mathrm{Cov}(X_2, X_1)$.
For a two dimensional matrix, we know that

$$\Sigma^{-1} = \frac{1}{(\sigma_{11}\sigma_{22} - \sigma_{12}^2)} \begin{bmatrix}\sigma_{22} & -\sigma_{12} \\ -\sigma_{12} & \sigma_{11}\end{bmatrix}$$

We thus have

$$\begin{aligned}
P(\mathbf{x}; \mu, \Sigma) &= \frac{1}{2\pi|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}\begin{bmatrix}x_1 - \mu_1 \\ x_2 - \mu_2\end{bmatrix}^T \Sigma^{-1} \begin{bmatrix}x_1 - \mu_1 \\ x_2 - \mu_2\end{bmatrix}\right) \\
&= \frac{1}{2\pi|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}\begin{bmatrix}\Delta_1 \\ \Delta_2\end{bmatrix}^T \Sigma^{-1} \begin{bmatrix}\Delta_1 \\ \Delta_2\end{bmatrix}\right) \\
&= \frac{1}{2\pi|\Sigma|^{1/2}} \exp\left(-\frac{1}{2(\sigma_{11}\sigma_{22} - \sigma_{12}^2)} \begin{bmatrix}\Delta_1 & \Delta_2\end{bmatrix} \begin{bmatrix}\sigma_{22} & -\sigma_{12} \\ -\sigma_{12} & \sigma_{11}\end{bmatrix} \begin{bmatrix}\Delta_1 \\ \Delta_2\end{bmatrix}\right) \\
&= \frac{1}{2\pi|\Sigma|^{1/2}} \exp\left(-\frac{1}{2(\sigma_{11}\sigma_{22} - \sigma_{12}^2)} \begin{bmatrix}\Delta_1\sigma_{22} - \Delta_2\sigma_{12} & \Delta_2\sigma_{11} - \Delta_1\sigma_{12}\end{bmatrix} \begin{bmatrix}\Delta_1 \\ \Delta_2\end{bmatrix}\right) \\
&= \frac{1}{2\pi|\Sigma|^{1/2}} \exp\left(-\frac{1}{2(\sigma_{11}\sigma_{22} - \sigma_{12}^2)} \left(\Delta_1^2\sigma_{22} - \Delta_1\Delta_2\sigma_{12} - \Delta_1\Delta_2\sigma_{12} + \Delta_2^2\sigma_{11}\right)\right) \\
&= \frac{1}{2\pi|\Sigma|^{1/2}} \exp\left(-\frac{1}{2} \cdot \frac{\left(\Delta_1^2\sigma_{22} - 2\Delta_1\Delta_2\sigma_{12} + \Delta_2^2\sigma_{11}\right)}{(\sigma_{11}\sigma_{22} - \sigma_{12}^2)}\right)
\end{aligned}$$

where $\Delta_i = x_i - \mu_i$ for $i \in \{1, 2\}$.
By the marginal distribution of a Gaussian, we also have

$$P(x_2; \mu, \Sigma) = \frac{1}{\sqrt{2\pi\sigma_{22}}} \exp\left(-\frac{(x_2 - \mu_2)^2}{2\sigma_{22}}\right) \quad = \frac{1}{\sqrt{2\pi\sigma_{22}}} \exp\left(-\frac{\Delta_2^2}{2\sigma_{22}}\right)$$

The conditional distribution of x_1 given x_2 is now given as

$$\begin{aligned}
P(x_1 \mid x_2; \mu, \Sigma) &= \frac{P(\mathbf{x}; \mu, \Sigma)}{P(x_2; \mu, \Sigma)} \\
&= \frac{\frac{1}{2\pi|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}\begin{bmatrix}\Delta_1 \\ \Delta_2\end{bmatrix}^T \Sigma^{-1} \begin{bmatrix}\Delta_1 \\ \Delta_2\end{bmatrix}\right)}{\frac{1}{\sqrt{2\pi\sigma_{22}}} \exp\left(-\frac{\Delta_2^2}{2\sigma_{22}}\right)} \\
&= \frac{\sqrt{2\pi\sigma_{22}}}{2\pi|\Sigma|^{1/2}} \cdot \frac{\exp\left(-\frac{1}{2}\begin{bmatrix}\Delta_1 \\ \Delta_2\end{bmatrix}^T \Sigma^{-1} \begin{bmatrix}\Delta_1 \\ \Delta_2\end{bmatrix}\right)}{\exp\left(-\frac{\Delta_2^2}{2\sigma_{22}}\right)}
\end{aligned}$$

Looking at the non-exponential part:

$$\frac{\sqrt{2\pi\sigma_{22}}}{2\pi|\Sigma|^{1/2}} = \sqrt{\frac{\sigma_{22}}{2\pi(\sigma_{11}\sigma_{22}-\sigma_{12}^2)}} = \sqrt{\frac{1}{2\pi(\sigma_{11}-\sigma_{12}^2/\sigma_{22})}}$$

Simplifying the exponential part:

$$\frac{\exp\left(-\frac{1}{2}\begin{bmatrix}\Delta_1\\ \Delta_2\end{bmatrix}^T \Sigma^{-1}\begin{bmatrix}\Delta_1\\ \Delta_2\end{bmatrix}\right)}{\exp\left(-\frac{\Delta_2^2}{2\sigma_{22}}\right)}$$
$$= \exp\left(-\frac{1}{2}\left[\frac{\left(\Delta_1^2\sigma_{22}-2\Delta_1\Delta_2\sigma_{12}+\Delta_2^2\sigma_{11}\right)}{(\sigma_{11}\sigma_{22}-\sigma_{12}^2)} - \frac{\Delta_2^2}{\sigma_{22}}\right]\right)$$

Focusing on the part inside the square brackets

$$\begin{aligned}
&\frac{\left(\Delta_1^2\sigma_{22}-2\Delta_1\Delta_2\sigma_{12}+\Delta_2^2\sigma_{11}\right)}{(\sigma_{11}\sigma_{22}-\sigma_{12}^2)} - \frac{\Delta_2^2}{\sigma_{22}}\\
&= \frac{\Delta_1^2\sigma_{22}^2-2\Delta_1\Delta_2\sigma_{12}\sigma_{22}+\Delta_2^2\sigma_{11}\sigma_{22}}{(\sigma_{11}\sigma_{22}-\sigma_{12}^2)\sigma_{22}} - \frac{\Delta_2^2\sigma_{11}\sigma_{22}-\Delta_2^2\sigma_{12}^2}{(\sigma_{11}\sigma_{22}-\sigma_{12}^2)\sigma_{22}}\\
&= \frac{\Delta_1^2\sigma_{22}^2-2\Delta_1\Delta_2\sigma_{12}\sigma_{22}+\Delta_2^2\sigma_{12}^2}{(\sigma_{11}\sigma_{22}-\sigma_{12}^2)\sigma_{22}}\\
&= \frac{(\Delta_1\sigma_{22}-\Delta_2\sigma_{12})^2}{(\sigma_{11}\sigma_{22}-\sigma_{12}^2)\sigma_{22}}\\
&= \frac{(\Delta_1-\Delta_2\sigma_{12}/\sigma_{22})^2}{\sigma_{11}-\sigma_{12}^2/\sigma_{22}}
\end{aligned}$$

Thus, the conditional distribution is

$$P(x_1 \mid x_2;\mu,\Sigma) = \frac{1}{\sqrt{2\pi(\sigma_{11}-\sigma_{12}^2/\sigma_{22})}}\exp\left(-\frac{1}{2}\cdot\frac{\left(x_1-\left\{\mu_1+(x_2-\mu_2)\frac{\sigma_{12}}{\sigma_{22}}\right\}\right)^2}{\sigma_{11}-\sigma_{12}^2/\sigma_{22}}\right)$$

We observe that this is a Gaussian with mean $\tilde{\mu}$ and variance $\tilde{\sigma}$, where

$$\tilde{\mu} = \mu_1 + (x_2-\mu_2)\frac{\sigma_{12}}{\sigma_{22}}, \quad \text{and}$$
$$\tilde{\sigma} = \sigma_{11} - \frac{\sigma_{12}^2}{\sigma_{22}}$$