Problemset 3

Nikolai German (12712506)

Exercise 2.1 - Bootstrapping

(a)

> Generate one sample of size $n = 120$ (set seed 123) from the mixture distribution (i.e. generate each sampled point from the first component with probability 0.6 and from the second with probability 0.4) and compute $I_{obs} = Q_n(0.75) - Q_n(0.25)$.

```
mixture_sample <- function(n, mu1 = 0, mu2 = 5, sigma1 = 1, sigma2 = 2,
                           probs = c(.5, .5), seed = NULL) {
  checkmate::assertIntegerish(n,
                              len = 1, any.missing = FALSE, lower = 1)
  checkmate::assertNumeric(mu1,
                           any.missing = FALSE, len = 1)
  checkmate::assertNumeric(mu2,
                           any.missing = FALSE, len = 1)
  checkmate::assertNumeric(sigma1,
                           any.missing = FALSE, len = 1, lower = 0)
  checkmate::assertNumeric(sigma2,
                           any.missing = FALSE, len = 1, lower = 0)
  checkmate::assertNumeric(probs,
                           len = 2, all.missing = FALSE, lower = 0, upper = 1)

  if (any(is.na(probs))) {
    probs[is.na(probs)] <- 1 - probs[!is.na(probs)]
  }

  if (sum(probs) != 1) {
    stop("probabilities have to sum up to 1!")
  }
```

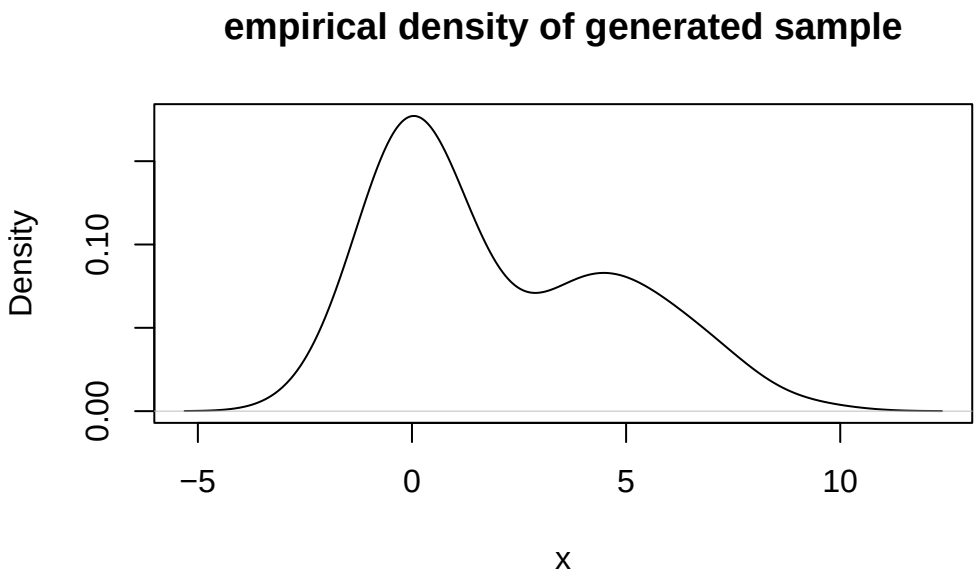
```
  n1 <- round(probs[[1]] * n)
  n2 <- n - n1

  res <- numeric(n)

  if(!is.null(seed)) {
    set.seed(seed)
  }
  res[1:n1] <- rnorm(n1, mu1, sigma1)
  res[(n1 + 1):n] <- rnorm(n2, mu2, sigma2)

  return(res)
}
```

```
samp <- mixture_sample(n = 120, probs = c(.6, .4), seed = 123)
```



Computing I_{obs}:

```
IQR(samp)
```

```
[1] 4.535644
```

(b)

From the empirical distribution generated in (a), draw $B = 500$ bootstrap samples (set seed 1234) and compute I_b^* for each. Compute the bootstrap standard error $\widehat{SE}_{boot}$ and the skewness of $\{I_b^*\}$.

```
bootstrap_IQR <- function(samp, B, seed = NULL) {
  checkmate::assertNumeric(samp, any.missing = FALSE, min.len = 1)
  checkmate::assertIntegerish(B, len = 1, lower = 1, any.missing = FALSE)
  checkmate::assertIntegerish(seed,
                              len = 1, lower = 1,
                              any.missing = FALSE, null.ok = TRUE)

  res <- numeric(B)

  if (!is.null(seed)) {
    set.seed(seed)
  }

  for (i in seq_len(B)) {
    bootstrap <- sample(samp, replace = TRUE)
    res[[i]] <- IQR(bootstrap)
  }

  structure(res,
            SE = sd(res),
            Skewness = mean(((res - mean(res)) / sd(res))^3))
}
```

We compute the Standarderror $\widehat{SE}_{boot} = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(I_b^* - \bar{I}^*)^2}$ and the Skewness $\hat{\gamma}_{boot} = \frac{1}{B-1}\sum_{b=1}^{B}(\frac{I_b^* - \bar{I}^*}{\widehat{SE}_{boot}})^3$

```
I_B <- bootstrap_IQR(samp, 500, 1234)
attr(I_B, "SE")
```

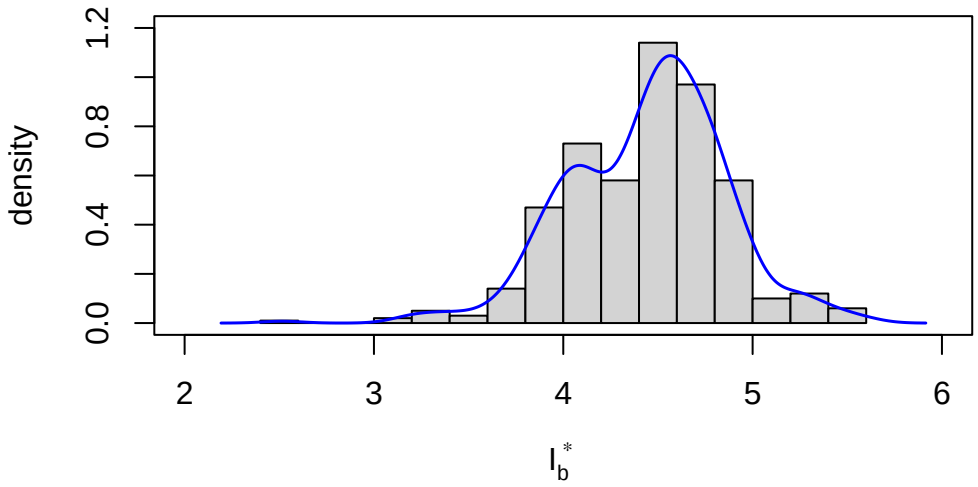
```
[1] 0.4148202
```

```
attr(I_B, "Skewness")
```

```
[1] -0.3640475
```

(c)

Plot a histogram of $\{I_b^*\}$ with an overlaid kernel density estimate.



(d)

Construct a 95% percentile bootstrap CI and a normal-approximation CI for the true IQR. Compare their width.

For the 95% percentile bootstrap CI, we just look at the 2.5% and 97.5% percentiles of $\{I_b^*\}$:

```
c(quantile(I_B, .025), quantile(I_B, .975))
```

```
    2.5%    97.5%
3.638510 5.243665
```

For the normal-approximation CI we use $I \sim \mathcal{N}(\bar{I}^*, \hat{SE}_{boot})$:

```
c("2.5%" = qnorm(.025) * attr(I_B, "SE") + mean(I_B),
  "97.5%" = qnorm(.975) * attr(I_B, "SE") + mean(I_B))
```

```
    2.5%    97.5%
3.630347 5.256412
```

We observe, that the two CIs are almost identical, the normal CI is slightly wider on the right due to the skewness.

(e)

> Discuss: How does the mixture's skew affect the bootstrap distribution and the reliability of the normal CI? What would change if n were larger?

The bigger the skew, the less reliable the normal CI, due to the very way it is contructed using mean and se. With bigger n, the se gets smaller and the effect of the skewness therefore bigger, since the mean differs from the median.

Exercise 2.2 - Bayesian Reasoning

(a)

> Compute the posterior probability of each model given the observed outcomes, using Bayes' theorem. Express the likelihood for each model as the product of Bernoulli likelihoods over the 30 days: $\mathbb{P}(x_1, ..., x_{30}|M_i) = \prod_{t=1}^{30} \pi_{it}^{x_t}(1-\pi_{it})^{1-x_t}$ Which model is most plausible after observing the data?

Using Bayes' Theorem, we obtain:

$$\mathbb{P}(M_i|x_1, ..., x_{30}) = \frac{\mathbb{P}(x_1, ..., x_{30}|M_i) \cdot \mathbb{P}(M_i)}{\sum_{j=1}^{4} \mathbb{P}(x_1, ..., x_{30}|M_j) \cdot \mathbb{P}(M_j)} \tag{1}$$

$$\propto \mathbb{P}(x_1, ..., x_{30}|M_i) \cdot \mathbb{P}(M_i) \tag{2}$$

```
prior <- c(0.3, 0.35, 0.2, 0.15)
posterior <- rep(1, nrow(data$pi_it))

for (j in seq_len(nrow(data$pi_it))) {
  for (i in seq_along(data$x_t)) {
    pi <- data$pi_it[j,i]
    x <- data$x_t[[i]]
    posterior[[j]] <- posterior[[j]] * pi^(x) * (1 - pi)^(1 - x)
  }
  posterior[[j]] * prior[[j]]
}

which.max(posterior)
```

```
[1] 3
```

```
round(posterior / sum(posterior), 4)
```

```
[1] 0.0879 0.1389 0.7730 0.0001
```

(b)

> Each model provides a probability forecast $\pi_{i,new}$ for failure tomorrow. Compute the posterior predictive probability of failure tomorrow $\hat{\pi} = \mathbb{P}(\text{failure tommorow}|\text{data}) = \mathbb{P}(x = 1|x_1, ..., x_{30})$ as the sum of $\pi_{i,new}$, weighted by the posterior reliability of the models from (a).

$$\begin{aligned} \mathbb{P}(x = 1|x_1, ..., x_{30}) &= \sum_{i=1}^{4} p(x = 1|M_i)p(M_i|x_1, ...x_{30}) \\ &= \sum_{i=1}^{4} \pi_{i,new} p(M_i|x_1, ...x_{30}) \end{aligned} \tag{3}$$

```
sum(data$pi_new * posterior / sum(posterior))
```

```
[1] 0.3620955
```

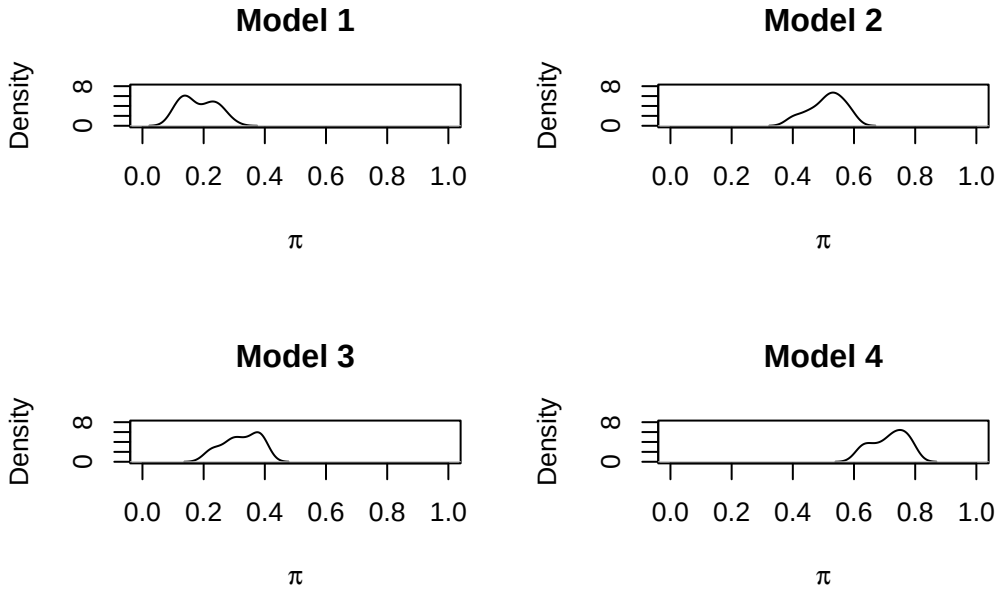
(c)

> Suppose instead that in the 30-day window, failures had occurred on 20 days. In general, how would this affect the posterior model probabilities and the posterior predictive probability of failure tomorrow? Comment on which models become more plausible and how this affects $\hat{\pi}$. For this, you may first need to look at the distribution of each model's predictions.

Since we then would have a higher count of $x_i = 1$, the term $\pi_{it}^{x_t}$ would gain influence for every Model M_t. Therefore the likelihood would be higher for models, which more frequently predict a higher chance of malfunction π_{it}.

Let's have a look at the distribution of each models predictions:

```r
par(mfrow = c(2,2))
for (i in seq_len(4)) {
  plot(density(data$pi_it[i,]),
       xlim = c(0, 1), ylim = c(0,8),
       main = sprintf("Model %d", i),
       xlab = latex2exp::TeX(r"($\pi$)"))
}
```



Models 2 and 4 predict rather high values for π. Therefore the likelihood would be higher for these models. Since there are nevertheless around 33% of cases of $x_i = 0$, the likelihood of Model 2 would be probably higher than the one of Model 4.

Looking at predictive probability, we will observe, that the predictions of models 2 & 4 get weighted more heavily in comparison to models 1 & 3. Thus resulting in the predictive probability beeing closer on those two predictions.

(d)

> The plant must decide every day whether to perform preventive maintenance. The costs for this can be described as a loss function $L(\theta, d)$, where θ is a binary variable (F = failure, NF = nofailure) indicating if failure actually (would have) occurred and d is the

decision that was made (M = maintenance, NM = no maintenance).

$$L(\theta, d) = \begin{cases} 0 & \text{if (F,M)} \\ q & \text{if (F,NM)} \\ 1 & \text{if (NF,M)} \\ 0 & \text{if (NF,NM)} \end{cases}$$

Assume $q = 20$ and $\mathbb{E}(\hat{\pi}) = \mathbb{E}(\theta)$. Compute the Bayes risk, i.e. $\mathbb{E}(L(\theta, d)|d)$ of each decision d and determine the optimal action given $\hat{\pi} = \mathbb{P}(\text{failure tomorrow}|\text{data})$ for any given day.

We rewirte $\mathbb{E}(L(\theta, d)|d)$ using Indicator functions $I(.)$:

$$\mathbb{E}(L(\theta, d)|d) = \mathbb{E}_\theta\Big(I_{d=M}(d) \cdot I_{\theta=NF}(\theta) + q \cdot I_{d=NM}(d) \cdot I_{\theta=F}(\theta)|d\Big) \tag{4}$$

$$\begin{aligned} \mathbb{E}_\theta(L(\theta, d)|d) &= \begin{cases} \mathbb{E}_\theta(I_{\theta=NF}(\theta)) & : d = M \\ \mathbb{E}_\theta(q \cdot I_{\theta=F}(\theta)) & : d = NM \end{cases} \\ &= \begin{cases} \mathbb{E}_\theta(1 - I_{\theta=F}(\theta)) & : d = M \\ \mathbb{E}_\theta(q \cdot I_{\theta=F}(\theta)) & : d = NM \end{cases} \\ &= \begin{cases} 1 - \mathbb{E}_\theta(\theta) & : d = M \\ q \cdot \mathbb{E}_\theta(\theta) & : d = NM \end{cases} \\ &= \begin{cases} 1 - \mathbb{E}(\hat{\pi}) & : d = M \\ q \cdot \mathbb{E}(\hat{\pi}) & : d = NM \end{cases} \end{aligned} \tag{5}$$

It's easy to observe, that the plant should conduct maintenance, if $q \cdot \mathbb{E}(\hat{\pi}) > 1 - \mathbb{E}(\hat{\pi})$ on any given day.

(e)

Determine the critical value q^* (depending on π) at which the plant is indifferent between maintaining and not maintaining. What does this imply about risk sensitivity?

Bei Indifferenz muss gelten $\mathbb{E}(L(\theta, d)|d = M) = \mathbb{E}(L(\theta, d)|d = NM)$, was uns zu folgender Bedingung führt:

$$\begin{aligned} &\mathbb{E}(L(\theta, d)|d = M) = \mathbb{E}(L(\theta, d)|d = NM) \\ \iff &1 - \mathbb{E}(\hat{\pi}) = q^* \cdot \mathbb{E}(\hat{\pi}) \\ \iff &q^* = \gamma(\hat{\pi}) \end{aligned} \tag{6}$$

where $\gamma(\hat{\pi}) = \frac{1-\mathbb{E}(\hat{\pi})}{\mathbb{E}(\hat{\pi})}$ are the odds of $\hat{\pi}$.

Exercise 2.3 - Conjugate Priors

(a)

Explain the concept of a conjugate prior.

The Key of Bayesian Inference is *Bayes' Theorem*:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cup B)}{\mathbb{P}(B)} \tag{7}$$

We can translate Bayes' Theorem when working with distributions:

$$p(\theta|x) = \frac{f(x;\theta) \cdot p(\theta)}{\int f(x;\tilde{\theta}) \cdot p(\tilde{\theta}) d\tilde{\theta}} \tag{8}$$

Where $p(\theta)$ is the prior information or assumption we have on the paramter θ, which itself is treated as a RV. The Likelihood of the sample given some parameter θ is $f(x;\theta)$. This way we can compute the *posterior* $p(\theta|x)$. Basically, we update the *prior* through our observations.

Since $\int f(x;\tilde{\theta}) \cdot p(\tilde{\theta}) d\tilde{\theta}$ does not depend on θ an acts as normalization constant, we can rewrite (8) as

$$p(\theta|x) \propto f(x;\theta) \cdot p(\theta) \tag{9}$$

While this is a very powerfull approach it also bares some difficulties: $p(\theta|x)$ does not neccessarily take the form of an known distribution. Furthermore, if we'd try to obtain the density of $p(\theta|x)$, $f(x;\tilde{\theta}) \cdot p(\tilde{\theta}) d\tilde{\theta}$ can be analytically unsolveable. Monte-Carlo Integration is a way to overcome this, nevertheless the first problem stays.

Therefore it would be handy, if we could find a pair of likelihood and prior in that way, that the posterior belongs to a known distribution. That's what a conjugate prior is: a prior to a likelihood in such way, that the posterior is from the same family of distributions as the prior distribution (see Definition 9.1).

(b)

Show that the gamma distribution is a conjugate prior for the exponential likelihood by computing the posterior distribution. State the setup and explain your steps.

We observe the following:

$$X_i|\lambda \sim Exp(\lambda) \tag{10}$$
$$\lambda \sim Gamma(\alpha, \beta) \tag{11}$$

So the prior is the distribution $p(\lambda; \alpha, \beta)$ and the likelihood is $f(x; \lambda) = \prod_{i=1}^{n} f(x_i; \lambda)$. Subsequently using (9), we can obtain:

$$p(\lambda; x) \propto \left[\prod_{i=1}^{n} f(x_i; \lambda)\right] \cdot p(\lambda; \alpha, \beta) \tag{12}$$

Plugging in the distributions for prior and likelihood, gives us:

$$\begin{aligned} p(\lambda; x) &\propto \left[\prod_{i=1}^{n} \lambda e^{-\lambda x_i}\right] \cdot \lambda^{\alpha-1} e^{-\beta\lambda} \\ &= \lambda^n e^{-\lambda \cdot n\bar{x}} \lambda^{\alpha-1} e^{-\beta\lambda} \\ &= \lambda^{n+\alpha-1} e^{-\lambda(\beta+n\bar{x})} \end{aligned} \tag{13}$$

We can see easily, that (13) is the core of a $Gamma(\alpha + n, \beta + n\bar{x})$ Distribution. Therefore we conclude:

$$\lambda|x \sim Gamma(\alpha + n, \beta + n\bar{x}) \tag{14}$$

(c)

Your prior knowledge on λ can be captured by $Gamma(2, 3)$. You have now collected a sample $x = (1, 1, 2, 1, 3, 1, 4, 6, 1, 1)$. Specify the updated posterior parameters α^* and β^* and compute the posterior mean $\mathbb{E}(\lambda^*)$.

Using (14), we can compute α^* and β^* easily:

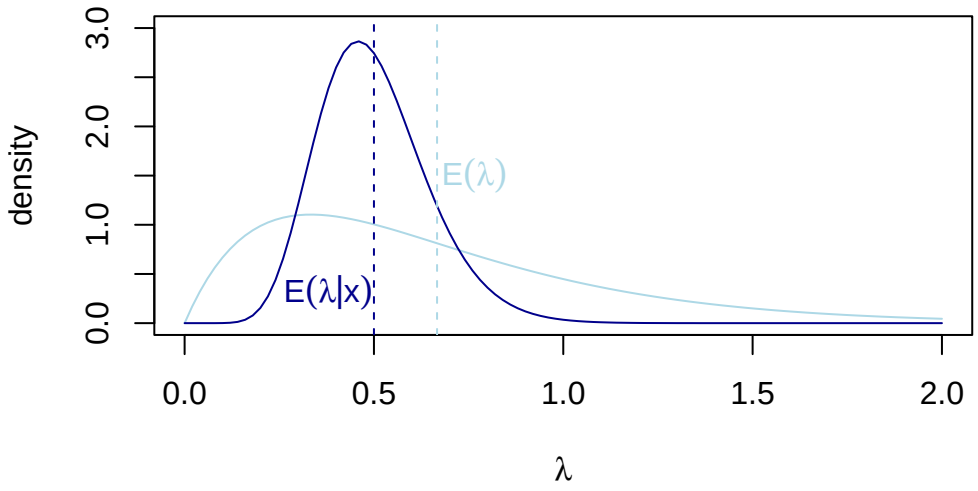
$$\alpha^* = \alpha + n = 12 \tag{15}$$
$$\beta^* = \beta + n\bar{x} = 24 \tag{16}$$

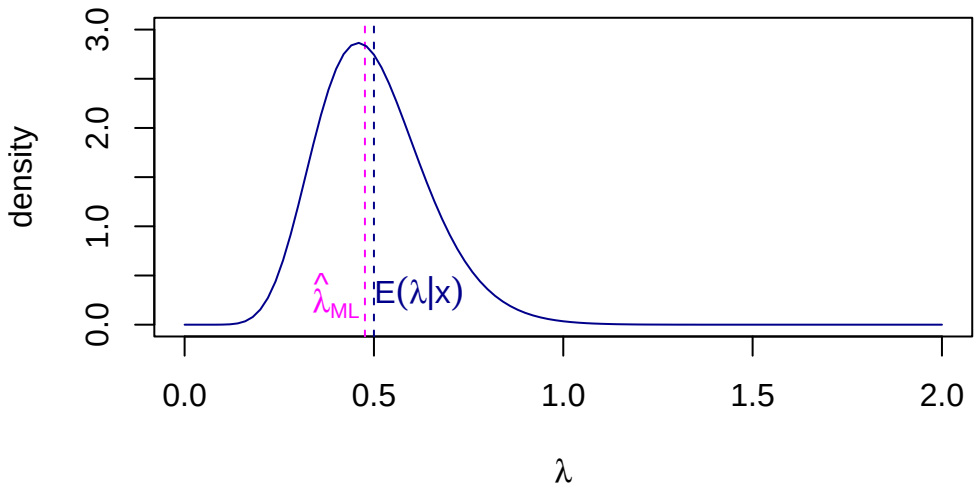
We know that for a Gamma distributed RV Y ist holds that $\mathbb{E}(Y) = \frac{\alpha}{\beta}$. Therefore we obtain

$$\begin{aligned}\mathbb{E}(\lambda^*) &= \mathbb{E}(\lambda|x) \\ &= \frac{\alpha^*}{\beta^*} \\ &= \frac{1}{2}\end{aligned}$$

(d)

What are the advantages and disadvantages of using conjugate priors? Discuss how these properties affect the choice of prior distribution and the resulting posterior inference.





Advantages

- Mathematical tractability: Conjugate priors yield closed-form posterior distributions, eliminating the need for complex numerical methods like MCMC.
- Computational efficiency: The posterior parameters can be derived through simple algebraic operations on the prior parameters and data statistics.
- Sequential updating: The posterior can serve as the prior for new data, making them ideal for online learning and streaming data scenarios.
- Interpretability: The parameters of conjugate priors often have clear interpretations as "prior observations" or "pseudo-counts."
- Consistency: As sample size increases, the influence of any proper conjugate prior diminishes, ensuring consistency of inference.

Disadvantages

- Limited flexibility: Conjugate priors cannot always express complex prior beliefs or multimodal distributions.
- Potential misspecification: Choosing a prior based on mathematical convenience rather than actual domain knowledge may misrepresent true prior beliefs.
- Restrictive assumptions: Many conjugate relationships depend on specific likelihood functions, limiting their applicability across different models.
- Oversimplification: Real-world phenomena often involve more complex relationships than conjugate models can capture.
- Sensitivity to hyperparameters: Poor choice of hyperparameters can significantly bias inference, especially with small datasets.

When selecting priors, the choice between conjugate and non-conjugate options involves several considerations:

- Problem complexity: For routine analyses with standard distributions, conjugate priors may be sufficient, while complex phenomena might require more flexible non-conjugate priors.
- Computational resources: Limited resources favor conjugate priors for their efficiency.
- Sample size: With large datasets, the prior's influence diminishes, making conjugate priors more acceptable even when imperfect.
- Prior knowledge: When strong prior information exists in a form incompatible with conjugate families, non-conjugate priors may be necessary despite computational costs.

The choice of conjugate priors affects posterior inference in several ways:

- Analytical solutions: Conjugate priors provide exact posterior distributions rather than approximations.
- Uncertainty quantification: The closed-form nature facilitates straightforward credible interval calculations.
- Posterior predictive distributions: Often available in closed form with conjugate priors, enabling efficient prediction.
- Robustness: The structured form of conjugate posteriors may be less robust to outliers compared to more flexible alternatives.
- Regularization: Conjugate priors naturally provide regularization, which can be beneficial for small sample problems.