Review of Bayesian Methods

Jae-Kwang Kim

ISU

Two approaches

- Frequentist inference: a parameter θ is assumed to be a fixed unknown quantity.
  - Maximum likelihood estimation, Method of moments estimation
  - Confidence intervals
  - Hypothesis testing: Reject H_0/fail to reject H_0, p-value
- Bayesian inference: a parameter θ is assumed to be a random variable.
- Hence, Bayesian inference is based on the conditional probability of parameter given observed data y, $p(\theta|y)$, called *posterior probability*.
  - Posterior mean, mode
  - Credible intervals
  - $P(H_0|y)$ vs $P(H_a|y)$

Bayes' rule

- Let $p(\theta)$ be the prior probability mass or density.
- Let $p(y|\theta)$ be the data distribution (i.e., the likelihood function).
- Bayes' rule yields the posterior probability as follows:

$$p(\theta|y) = \frac{p(\theta, y)}{p(y)} = \frac{p(y|\theta)p(\theta)}{\int p(y|\theta)p(\theta)d\theta} \propto p(y|\theta)p(\theta).$$

- Note that Bayesian inference relies on both the likelihood function (data) and the prior distribution (prior knowledge, experiences, or beliefs) while frequentist inference is usually based on the likelihood.

Prediction

- After the data y have been observed, we can predict an unknown observable, $\tilde{y}$, from the same process as follows:

$$\begin{aligned} p(\tilde{y}|y) &= \int p(\tilde{y},\theta|y)d\theta \\ &= \int p(\tilde{y}|\theta,y)p(\theta|y)d\theta \\ &= \int p(\tilde{y}|\theta)p(\theta|y)d\theta \quad (\text{if } \tilde{y} \perp y|\theta) \end{aligned}$$

  which is called the *posterior predictive distribution*.

- The posterior predictive distribution can be viewed as an average of conditional predictions over the posterior distribution of θ.

Example 1

- Let $y = (y_1, \ldots, y_n)$ be the vector of recorded weights of an object weighed n times on a scale.

- Suppose that $y_i \overset{iid}{\sim} N(\theta, \sigma^2)$, where θ is the unknown true weight of the object and σ^2 is the measurement variance of the scale. We assume that σ^2 is known, say $\sigma^2 = 1$.

- The likelihood is given as

$$p(y|\theta) = \prod_{i=1}^{n} \frac{1}{(2\pi)^{1/2}} \exp\left\{-\frac{1}{2}(y_i - \theta)^2\right\} \propto \exp\left\{-\frac{n}{2}(\theta - \bar{y})^2\right\}.$$

Example 1: frequentist inference

- The MLE is given as

$$\begin{aligned}\hat{\theta} &= \arg\max_{\theta} p(y|\theta) = \arg\min_{\theta} \{-\log p(y|\theta)\} \\ &= \arg\min_{\theta} (\theta - \bar{y})^2 \\ &= \bar{y}.\end{aligned}$$

- The 95% confidence interval is given as

$$\mathrm{I}_{\mathrm{freq}} = \bar{y} \pm 1.96/\sqrt{n}.$$

- Frequentist interpretation of confidence interval:
  If we were to draw infinitely many random samples of size n from $N(\theta, 1)$, then 95% of the corresponding confidence intervals $\mathrm{I}_{\mathrm{freq}}$ would cover θ.

Example 1: Bayesian inference

- Define $p(\theta) = N(\theta|\mu_0, \tau_0^2)$, where μ_0 are τ_0 are pre-specified constants, called hyper-parameters.

- The posterior distribution is given as

$$\begin{aligned} p(\theta|y) &\propto p(y|\theta)p(\theta) \\ &= N\left(\frac{\tau_0^2}{n^{-1}+\tau_0^2}\bar{y} + \frac{n^{-1}}{n^{-1}+\tau_0^2}\mu_0, \frac{1}{n+1/\tau_0^2}\right) \end{aligned}$$

- The 95% posterior interval (also known as credible interval) is given as

$$\mathrm{I}_{\mathrm{Bayes}} = \left(\frac{\tau_0^2}{n^{-1}+\tau_0^2}\bar{y} + \frac{n^{-1}}{n^{-1}+\tau_0^2}\mu_0\right) \pm \frac{1.96}{\sqrt{n+1/\tau_0^2}}.$$

- Bayesian interpretation of credible interval:
  After observing the data y, the probability that θ is in the interval $\mathrm{I}_{\mathrm{Bayes}}$ is 0.95.

- Note that as $\tau_0 \to \infty$, we observe that $\mathrm{I}_{\mathrm{freq}} = \mathrm{I}_{\mathrm{Bayes}}$.

Binomial models

- The binomial distribution provides a natural model for data that arise from a sequence of 'Bernoulli trial'
- Consider data $y_1, \ldots, y_n$ such that $y_i \overset{iid}{\sim} \text{Ber}(\theta)$.
  - That is,

$$y_i = \begin{cases} 1 \text{ with probability } \theta \\ 0 \text{ with probability } 1-\theta \end{cases} .$$

- The data can be summarized by the total number of successes in the n trials, which we denote here by y.
- The binomial sampling model is

$$p(y|\theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y}.$$

Binomial models: Bayesian inference

- To perform Bayesian inference in the binomial model, we must specify a prior distribution for θ.
- For simplicity at this point, we assume that $\theta \sim U(0,1)$, that is,

  $$p(\theta) = \mathbb{I}(\theta \in [0,1]),$$

  where $\mathbb{I}$ denotes the indicator function.
- We can easily obtain that

  $$p(\theta|y) \propto \theta^{y}(1-\theta)^{n-y}.$$

- This leads to

  $$\theta|y \sim \text{Beta}(y+1, n-y+1).$$

- Note that $E(\theta|y) = \frac{y+1}{n+2}$ and $V(\theta|y) = \frac{(y+1)(n-y+1)}{(n+2)^2(n+3)} = \frac{E(\theta|y)\{1-E(\theta|y)\}}{n+3}$.

Prediction

- For posterior prediction from this model, we might be more interested in the outcome of one new trial, rather than another set of n new trials.
- Letting $\tilde{y}$ denote the result of a new trial, we can obtain that

$$\begin{aligned}
\Pr(\tilde{y} = 1|y) &= \int_0^1 \Pr(\tilde{y} = 1|\theta, y) p(\theta|y) d\theta \\
&= \int_0^1 \theta p(\theta|y) d\theta \\
&= E(\theta|y) = \frac{y+1}{n+2}.
\end{aligned}$$

Informative prior distribution

- In the binomial example, we have so far considered only the uniform prior distribution for θ.
- How can this specification be justified, and how in general do we approach the problem of constructing prior distributions?
- We consider two basic interpretations that can be given to prior distributions.
    - **Population** interpretation:
      the prior distribution represents a population of possible parameter value.
    - More subjective **state of knowledge** interpretation:
      the guiding principle is that we must express our knowledge (and uncertainty) about θ as if its value could be thought of as a random realization from the prior distribution.

Conjugate priors

- The property that the posterior distribution follows the same parametric form as the prior distribution is called *conjugacy*.

Definition

If $\mathcal{F}$ is a class of sampling distributions $p(y|\theta)$, and $\mathcal{P}$ is a class of prior distributions $p(\theta)$, then the class $\mathcal{P}$ is conjugate for $\mathcal{F}$ if

$$p(\theta|y) \in \mathcal{P}.$$

- Conjugate prior distributions have the practical advantage, in addition to computational convenience, of being interpretable as additional data.

Conjugate prior for binomial models

- Let $\theta \sim \text{beta}(\alpha, \beta)$. That is,

$$p(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}.$$

- It can be easily shown that the beta distribution is conjugate for binomial models as follows:

$$\begin{aligned} p(\theta|y) &\propto p(y|\theta)p(\theta) \\ &\propto \theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1} \\ &= \text{Beta}(y+\alpha, n-y+\beta). \end{aligned}$$

Remarks for binomial models with beta prior

- The uniform prior is a special case when $\alpha = \beta = 1$.
- Note that

$$\begin{aligned} E(\theta|y) &= \frac{y+\alpha}{n+\alpha+\beta} \approx \frac{y}{n}. \\ V(\theta|y) &= \frac{(y+\alpha)(n-y+\beta)}{(n+\alpha+\beta)^2(n+\alpha+\beta+1)} \approx \frac{1}{n}\frac{y}{n}\left(1-\frac{y}{n}\right). \end{aligned}$$

  for sufficiently large n.

- The parameters of the prior distribution (=hyperparameters) have no influence on the posterior distribution as $n \to \infty$.
- In addition, we can show that, as $n \to \infty$,

$$\left.\frac{\theta - E(\theta|y)}{\sqrt{V(\theta|y)}}\right| y \xrightarrow{d} N(0,1),$$

  which is known as Bayesian central limit theorem.

Normal distribution with known variance

- As the simplest case, consider a single scalar observation y from $N(\theta, \sigma^2)$, where σ^2 is known.
- The sampling distribution (=the likelihood) is

$$p(y|\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2}(y-\theta)^2 \right\}$$

- Define $\theta \sim N(\mu_0, \tau_0^2)$.
- It is easy to show that

$$p(\theta|y) \propto \exp\left\{ -\frac{1}{2\tau_1^2}(\theta-\mu_1)^2 \right\},$$

where $\mu_1 = \frac{\mu_0/\tau_0^2 + y/\sigma^2}{1/\tau_0^2 + 1/\sigma^2}$ and $\frac{1}{\tau_1^2} = \frac{1}{\tau_0^2} + \frac{1}{\sigma^2}$ and this implies that normal prior is conjugate for normal distribution.

Normal distribution with known mean but unknown variance

- Suppose that $y = (y_1, \ldots, y_n)$ such that $y_i \overset{iid}{\sim} N(\theta, \sigma^2)$, where θ is known but σ^2 is unknown
- The likelihood function is given as

$$\begin{aligned} p(y|\sigma^2) &= \prod_{i=1}^{n} \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(y_i - \theta)^2\right\} \\ &\propto (\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \theta)^2\right\}. \end{aligned}$$

Inverse gamma priors

- We define the inverse gamma distribution for σ^2 as follows:

$$p(\sigma^2) \quad = \quad \frac{\beta^\alpha}{\Gamma(\alpha)} (\sigma^2)^{-(\alpha+1)} \exp\left(-\frac{\beta}{\sigma^2}\right),$$

  where $\alpha (> 0)$ is shape parameter and $\beta (> 0)$ is scale parameter.

- Then it can be shown that

$$\sigma^2 | y \sim \text{Inv-Gamma}\left(\frac{n}{2} + \alpha, \frac{1}{2}\sum_{i=1}^{n}(y_i - \theta)^2 + \beta\right).$$

- Therefore, the inverse gamma prior is the conjugate prior for the normal model with unknown variance.

Normal data with unknown variance

- We now consider a prior for normal data as follows: $p(\mu, \sigma^2) = p(\mu|\sigma^2)p(\sigma^2)$, where

$$\begin{aligned} \mu|\sigma^2 &\sim N(\mu_0, \sigma^2\kappa_0^{-1}) \\ \sigma^2 &\sim \text{Inv-Gamma}\left(\frac{\alpha_0}{2}, \frac{\beta_0}{2}\right) \end{aligned}$$

- We can show that

$$\begin{aligned} \mu|\sigma^2, y &\sim N\left(\frac{n\bar{y} + \kappa_0\mu_0}{n + \kappa_0}, \frac{\sigma^2}{n + \kappa_0}\right) \\ \sigma^2|y &\sim \text{Inv-Gamma}\left(\frac{n + \alpha_0}{2}, \frac{(n-1)s^2 + \beta_0 + \frac{n\kappa_0}{n+\kappa_0}(\bar{y} - \mu_0)^2}{2}\right) \end{aligned}$$

- Hence, the prior is conjugate for normal data.

Proper and improper prior distribution

- In general, we call a prior density $p(\theta)$ proper if it does not depend on data and integrates to 1.
- We return to the problem of estimating the mean θ of a normal model with known variance σ^2, with a $N(\mu_0, \tau_0^2)$ prior distribution on θ. We have observed that as $\tau_0^2 \to \infty$, we have

  $$p(\theta|y) \approx N(\theta|\bar{y}, \sigma^2/n),$$

  where $N(\theta|\bar{y}, \sigma^2/n)$ denotes the normal pdf of θ with mean $\bar{y}$ and variance σ^2/n.
- Let $p(\theta) \propto 1$. Then $p(\theta|y) = N(\theta|\bar{y}, \sigma^2/n)$. In this case, while the prior is improper, the posterior is proper.

Jeffreys' invariance principle

- One approach to define noninformative prior was introduced by Jeffreys, based on considering one-to-one transformation of the parameter: $\phi = h(\theta)$.
- By transformation of variables, the prior density $p(\theta)$ is equivalent to the following prior density on ϕ:

$$p(\phi) = p(\theta) \left| \frac{d\theta}{d\phi} \right| = p(\theta)|h'(\theta)|^{-1}. \tag{1}$$

- Jeffrey's general principle is that any rule for determining the prior density $p(\theta)$ should yield an equivalent result if applied to the transformed parameter;
    - that is, $p(\phi)$ computed by determining $p(\theta)$ and applying (1) should match the distribution that is obtained by determining $p(\phi)$ directly using the transformed model, $p(y, \phi) = p(\phi)p(y|\phi)$.

Jeffreys' prior

- Jeffrey's principle leads to defining the noninformative prior density as $p(\theta) \propto [I(\theta)]^{1/2}$, where $I(\theta)$ is the Fisher information for θ:

$$I(\theta) = E\left\{ \left(\frac{d \log p(y|\theta)}{d\theta} \right)^2 \middle| \theta \right\} = -E\left(\frac{d^2 \log p(y|\theta)}{d\theta^2} \middle| \theta \right).$$

- To see that Jeffreys' prior model is invariant to parameterization, evaluate $I(\phi)$ at $\theta = h^{-1}(\phi)$:

$$\begin{aligned} I(\phi) &= -E\left(\frac{d^2 \log p(y|\phi)}{d\phi^2} \middle| \phi \right) \\ &= -E\left(\frac{d^2 \log p(y|\theta)}{d\theta^2} \left| \frac{d\theta}{d\phi} \right|^2 \middle| \theta = h^{-1}(\phi) \right) \\ &= I(\theta) \left| \frac{d\theta}{d\phi} \right|^2 ; \end{aligned}$$

thus, $I(\phi)^{1/2} = I(\theta)^{1/2} \left| \frac{d\theta}{d\phi} \right|$, satisfying (1).

Example: Jeffreys' priors for binomial models

- Consider the binomial distribution: $y \sim \text{Bin}(n, \theta)$, which has log-likelihood

  $$\log p(y|\theta) = y\log(\theta) + (n-y)\log(1-\theta) + \text{constant}.$$

  Using the fact that $E(y|\theta) = n\theta$, we can show that

  $$I(\theta) = -E\left(\frac{d^2 \log p(y|\theta)}{d\theta^2} \middle| \theta \right) = \frac{n}{\theta(1-\theta)}.$$

- Jeffreys' prior density is then $p(\theta) \propto \theta^{-1/2}(1-\theta)^{-1/2}$, which is a Beta$(1/2, 1/2)$ density.

Example: Jeffreys' priors for normal models with unknown variance

- Consider the Gaussian model: $y \sim \text{Normal}(\mu, \sigma^2)$, where μ is known. The log-likelihood is

$$\log p(y|\sigma^2) = -\frac{1}{2}\log \sigma^2 - \frac{1}{2\sigma^2}(y-\mu)^2 + \text{constant}.$$

We can easily show that

$$I(\sigma^2) = -E\left(\frac{d^2 \log p(y|\sigma^2)}{d(\sigma^2)^2} \middle| \sigma^2 \right) = \frac{1}{2(\sigma^2)^2}.$$

- Jeffreys' prior density is then $p(\sigma^2) \propto 1/\sigma^2$, which can be approximated by a Inv-Gamma$(10^{-10}, 10^{-10})$ density.

Remarks on noninformative prior

- For many problems, there is no clear choice for a vague (or flat) prior distribution, since a density that is flat or uniform in one parameterization will not be in another.

- Recall $p(\sigma^2) \propto 1/\sigma^2$. We define $\phi = \log \sigma^2$, then the prior density on ϕ is

  $$p(\phi) = p(\sigma^2) \left| \frac{d\sigma^2}{d\phi} \right| = \frac{1}{\sigma^2} \sigma^2 = 1,$$

  that is, uniform on $\phi = \log \sigma^2$.

- Nevertheless, noninformative prior and reference priors are often useful when it does not seem to be worth to the effort to quantify one's real prior knowledge as a probability distribution, as long as one is willing to perform the mathematical work to check that the posterior is proper and to determine the sensitivity of posterior inferences to modeling assumptions of convenience.

Multiparameter Problem

- Most of practical problems in statistics involve more than on unknown or unobservable quantity.
- Although a problem can include several parameters of interest, conclusions will often be drawn one, or only a few, parameters at a time.
- The ultimate aim of a Bayesian analysis is to obtain the marginal posterior distribution of the particular parameters of interest.
  - Let $\theta = (\theta_1, \theta_2)$. Suppose that we are only interested in inference for θ_1, so θ_2 is a ‘nuisance’ parameter.
  - Then, the marginal posterior distribution of θ_1 is $p(\theta_1|y)$

How to obtain marginal distribution?

(1) We first obtain the *joint* posterior distribution of *all* unknowns; e.g.,

$$p(\theta_1, \theta_2|y) \propto p(y|\theta_1, \theta_2)p(\theta_1, \theta_2),$$

(2) Then we integrate this distribution over the unknowns that are not of immediate interest to obtain the desired marginal distribution; e.g.,

$$p(\theta_1|y) = \int p(\theta_1, \theta_2|y)d\theta_2.$$

Or equivalently using simulation

(1) We draw sample from the joint posterior distribution; e.g., generate $(\theta_1^{(r)}, \theta_2^{(r)})$ from $p(\theta_1, \theta_2|y)$ for $r = 1, \ldots, R$.

(2) Then look at the parameters of interest and ignore the values of the other unknowns; e.g.,

$$\hat{E}(\theta_1|y) = R^{-1}\sum_{r=1}^{R}\theta_1^{(r)} \text{ and } v\hat{a}r(\theta_1|y) = (R-1)^{-1}\sum_{r=1}^{R}(\theta_1^{(r)} - \bar{\theta}_1)^2.$$

Remarks on marginal distribution?

- The joint posterior density can be factorized to yield

$$p(\theta_1|y) = \int p(\theta_1|\theta_2, y)p(\theta_2|y)d\theta_2.$$

- This shows that the posterior distribution of interest, $p(\theta_1|y)$, is a mixture of the conditional posterior distribution given the nuisance parameter, θ_2, where $p(\theta_2|y)$ is a weighting function for the difference possible values of θ_2.

Normal data with a noninformative prior

- Consider a vector y of n independent observations from $N(\mu, \sigma^2)$, where μ and σ^2 are unknown.
- Assuming prior independence of location and scale parameters, we define a vague prior
$$p(\mu, \sigma^2) \propto (\sigma^2)^{-1}.$$
- We discussed that this prior is uniform on $(\mu, \log \sigma^2)$.
- For multiparameter, the Jeffreys' prior is defined as $p(\theta) \propto [\det\{J(\theta)\}]^{1/2}$, where $I(\theta)$ is the Fisher information matrix.

The joint posterior

- The joint posterior is

$$
\begin{aligned}
p(\mu, \sigma^2 | y) &\propto p(y|\mu, \sigma^2) p(\mu, \sigma^2) \\
&\propto \left[\prod_{j=1}^{n} (\sigma^2)^{-1/2} \exp\left\{ -\frac{1}{2\sigma^2} (y_i - \mu)^2 \right\} \right] (\sigma^2)^{-1} \\
&\propto (\sigma^2)^{-n/2-1} \exp\left[-\frac{1}{2\sigma^2} \left\{ \sum_{i=1}^{n} (y_i - \bar{y})^2 + n(\bar{y} - \mu)^2 \right\} \right] \\
&\propto (\sigma^2)^{-n/2-1} \exp\left[-\frac{1}{2\sigma^2} \left\{ (n-1)s^2 + n(\bar{y} - \mu)^2 \right\} \right],
\end{aligned}
$$

where $s^2 = (n-1)^{-1} \sum_{i=1}^{n} (y_i - \bar{y})^2$.

- Note that $\bar{y}$ and s^2 are the sufficient statistics for μ and σ^2.

Sampling from the joint posterior

- Note that $p(\mu, \sigma^2|y) = p(\mu|\sigma^2, y)p(\sigma^2|y)$.
- We can show that
$$\mu|\sigma^2, y \sim N\left(\bar{y}, \frac{\sigma^2}{n}\right),$$
and
$$\sigma^2|y \sim \text{Inv-Gamma}\left(\frac{n-1}{2}, \frac{(n-1)s^2}{2}\right).$$
- Hence, it is easy to draw sample from the joint posterior distribution:
  (1) Draw σ^2 from Inv-Gamma $\left(\frac{n-1}{2}, \frac{(n-1)s^2}{2}\right)$.

  (2) For given σ^2, draw μ from $N\left(\bar{y}, \frac{\sigma^2}{n}\right)$.

* Note that, conditional on σ^2 (and μ), $(n-1)s^2/\sigma^2 \sim \chi^2_{n-1}$ in a classical sampling theory. We can show that $\frac{(n-1)s^2}{\sigma^2}|y \sim \chi^2_{n-1}$.

Analytic form of the marginal posterior of μ

- The population mean, μ, is typically the parameter of interest, and so the objective of the Bayesian analysis is the marginal posterior distribution of μ.

- We can drive the marginal posterior density for μ by integrating the joint posterior density over σ^2:

$$\begin{aligned}
p(\mu|y) &= \int_0^\infty p(\mu, \sigma^2|y) d\sigma^2 \\
&\propto \int_0^\infty (\sigma^2)^{-n/2-1} \exp\left[-\frac{1}{2\sigma^2}\left\{(n-1)s^2 + n(\bar{y}-\mu)^2\right\}\right] d\sigma^2 \\
&\propto \left\{(n-1)s^2 + n(\bar{y}-\mu)^2\right\}^{-n/2} \\
&\propto \left\{1 + \frac{1}{(n-1)} \frac{(\mu-\bar{y})^2}{s^2/n}\right\}^{-\frac{(n-1)+1}{2}} .
\end{aligned}$$

which is proportional to $t_{n-1}(\bar{y}, s^2/n)$ density, called non-standardized Student's t distribution.

Remarks on the marginal posterior of μ

- This is easy to see that

  $$\left.\frac{\mu-\bar{y}}{s/\sqrt{n}}\right| y \sim t_{n-1},$$

  where t_{n-1} denotes the standard t density with $n-1$ df.

- Recall that from the classical sampling theory we have shown that

  $$\left.\frac{\mu-\bar{y}}{s/\sqrt{n}}\right| \mu, \sigma^2 \sim t_{n-1}.$$

Multinomial model for categorical data

- The binomial distribution can be generalized to allow more than two possible outcomes.
- The multinomial distribution is used to describe data for which each observation is one of k possible outcomes.
- If y is the vector of counts of the number of observations of each outcome, then

$$p(y|\theta) \propto \prod_{j=1}^{k} \theta_j^{y_j},$$

where the sum of the probabilities, $\sum_{j=1}^{k} \theta_j$, is 1.

Multinomial model for categorical data (cont.)

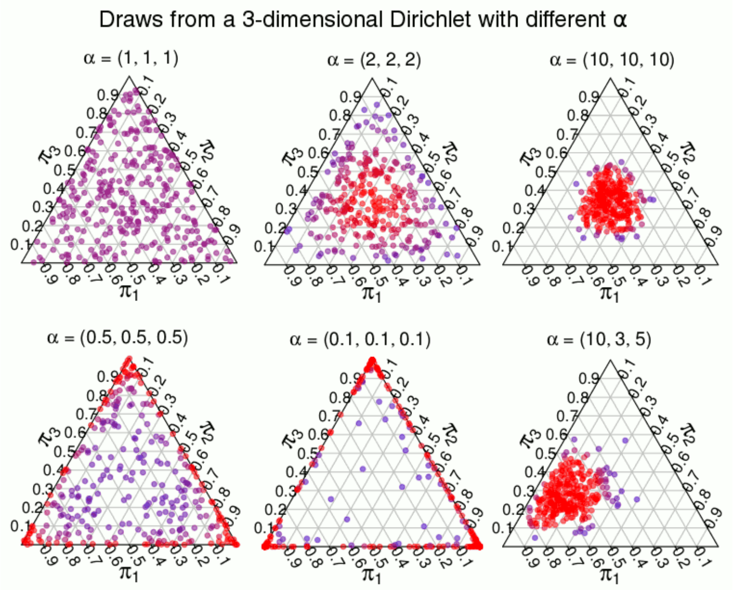
- The conjugate prior is a multivariate generalization of the beta distribution known as the Dirichlet,

$$p(\theta) \propto \prod_{j=1}^{k} \theta_j^{\alpha_j - 1},$$

  where the distribution is restricted to nonnegative θ_j's with $\sum_{j=1}^{k} \theta_j = 1$.

- It is easy to check that the posterior distribution is Dirichlet with parameters $\alpha_j + y_j$ $(j = 1, \ldots, k)$.

Dirichlet distribution with $K = 3$, $Dirichlet(\alpha_1, \alpha_2, \alpha_3)$



Draws from a 3-dimensional Dirichlet with different α

α = (1, 1, 1)

α = (2, 2, 2)

α = (10, 10, 10)

α = (0.5, 0.5, 0.5)

α = (0.1, 0.1, 0.1)

α = (10, 3, 5)

Linear regression models with non-informative prior

- Consider
$$y_i = x_i^{\mathrm{T}}\beta + \epsilon_i, \quad i = 1, \ldots, n,$$
where $x_i = (x_{i1}, \ldots, x_{ip})^{\mathrm{T}}$, $\beta = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$, and $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$.

- Let $y = (y_1, \ldots, y_n)^{\mathrm{T}}$ and $X = (x_1, \ldots, x_n)^{\mathrm{T}}$. Suppose that X has full rank and $p < n$.

- Then the likelihood is given as
$$p(y|\beta, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\|y - X\beta\|^2\right),$$
where $\|\cdot\|$ denote the Euclidean norm, i.e., $\|y\| = \sqrt{y^{\mathrm{T}}y}$.

- We define a noninformative prior for (β, σ^2),
$$p(\beta, \sigma^2) = 1/\sigma^2.$$

Independent Jeffreys prior

- Let $p(\beta, \sigma^2) = p(\beta)p(\sigma^2)$, where

$$p(\beta) = |J(\beta)|^{1/2} \quad \text{and} \quad p(\sigma^2) = |J(\sigma^2)|^{1/2}.$$

- Then we can show that

$$p(\beta, \sigma^2) \propto 1/\sigma^2.$$

Posterior inference

- The posterior is

$$p(\beta, \sigma^2 | y) \propto (\sigma^2)^{-n/2-1} \exp\left(-\frac{1}{2\sigma^2}\|y - X\beta\|^2\right),$$

- It can be shown that the full conditionals are given as follows.

$$\begin{aligned} \beta | y, \sigma^2 &\sim N\left((X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y, \sigma^2(X^{\mathrm{T}}X)^{-1}\right), \\ \sigma^2 | y, \beta &\sim \text{Inverse-Gamma}\left(\frac{n}{2}, \frac{\|y - X\beta\|^2}{2}\right) \end{aligned}$$

Linear regression models with independent normal prior

- Consider

$$y_i = x_i^{\mathrm{T}}\beta + \epsilon_i, \quad i = 1, \ldots, n,$$

  where $x_i = (x_{i1}, \ldots, x_{ip})^{\mathrm{T}}$, $\beta = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$, and $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$.

- Let $y = (y_1, \ldots, y_n)^{\mathrm{T}}$ and $X = (x_1, \ldots, x_n)^{\mathrm{T}}$. ~~Suppose that X has full rank and $p < n$.~~

- Then the likelihood is given as

$$p(y|\beta, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\|y - X\beta\|^2\right).$$

- We define a prior for (β, σ^2),

$$p(\beta, \sigma^2) = p(\beta|\sigma^2)p(\sigma^2) = N\left(\beta \,\middle|\, 0, \frac{\sigma^2}{\lambda} I_p\right) \times 1/\sigma^2.$$

  where $\lambda (> 0)$ is a deterministic hyperparameter.

Posterior inference

- The posterior is

$$p(\beta, \sigma^2 | y) \propto (\sigma^2)^{-(n+p)/2-1} \exp\left\{ -\frac{1}{2\sigma^2} \left(\|y - X\beta\|^2 + \lambda \|\beta\|^2 \right) \right\},$$

- It can be shown that the full conditionals are given as follows.

$$\begin{aligned} \beta | y, \sigma^2 &\sim N\left((X^{\mathrm{T}}X + \lambda)^{-1} X^{\mathrm{T}} y, \sigma^2 (X^{\mathrm{T}}X + \lambda)^{-1} \right), \\ \sigma^2 | y, \beta &\sim \text{Inverse-Gamma}\left(\frac{n+p}{2}, \frac{\|y - X\beta\|^2 + \lambda \|\beta\|^2}{2} \right). \end{aligned}$$

- We can show that $(X^{\mathrm{T}}X + \lambda)$ is positive definite for any $\lambda > 0$.

- In a Bayesian approach, it is natural to assign a hyperprior for λ.

Posterior inference (cont.)

- In a Bayesian approach, it is natural to assign a hyperprior for λ.
- Let $\lambda \sim \text{Gamma}(a_0/2, b_0/2)$.
- It is easy to show that

$$\lambda|\beta, \sigma^2, y \sim \text{Gamma}\left(\frac{p + a_0}{2}, \frac{\|\beta\|^2/\sigma^2 + b_0}{2}\right).$$

Lasso

- The Lasso of Tibshirani (1996) estimates linear regression coefficients through L_1-constrained least squares as follows:

$$\hat{\beta}_{\text{Lasso}} = \min_{\beta} \left\{ \|y - X\beta\|^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\},$$

where $\lambda > 0$ controls the degrees of sparsity.

Bayesian interpretation

- Tibshirani (1996) remarked that Lasso estimates can be interpreted as posterior mode estimates when the regression parameters have independent and identical Laplace (i.e., double-exponential) priors as follows:

$$p(\beta|\sigma^2) \quad = \quad \prod_{j=1}^{p} \frac{\lambda}{4\sigma^2} \exp\left(-\frac{\lambda}{2\sigma^2}|\beta_j|\right).$$

Bayesian Lasso

- Park and Casella (2008) proposed the Bayesian Lasso as follows:

$$\begin{aligned} p(\beta|\sigma^2) &= \prod_{j=1}^{p} \frac{\lambda}{2\sqrt{\sigma^2}} \exp\left(-\frac{\lambda}{\sqrt{\sigma^2}}|\beta_j|\right); \\ p(\sigma^2) &= 1/\sigma^2. \end{aligned}$$

Figure: $y_i = 0 * x_i + \epsilon_i$ for $i = 1, \ldots, 30$

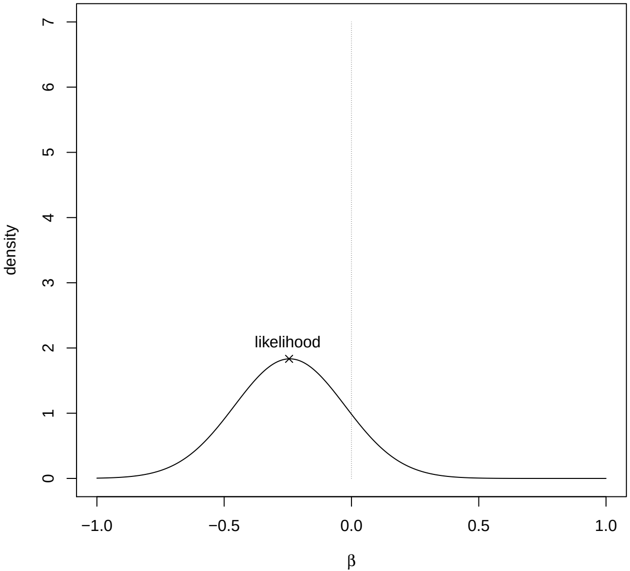


Figure: $y_i = 0 * x_i + \epsilon_i$ for $i = 1, \ldots, 30$

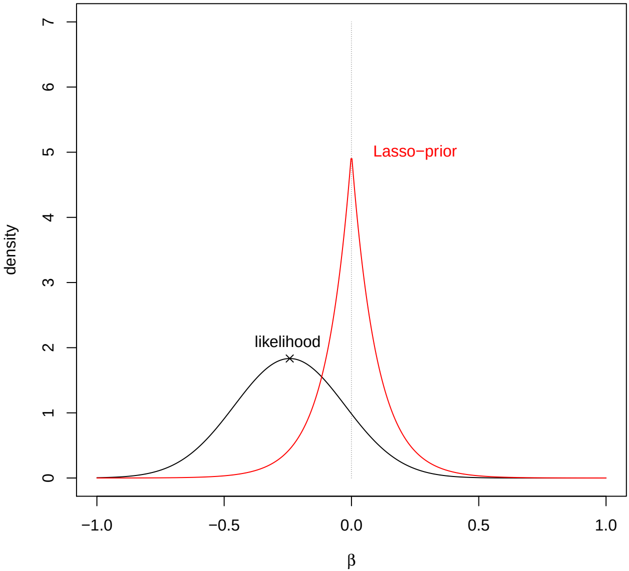


Figure: $y_i = 0 * x_i + \epsilon_i$ for $i = 1, \ldots, 30$

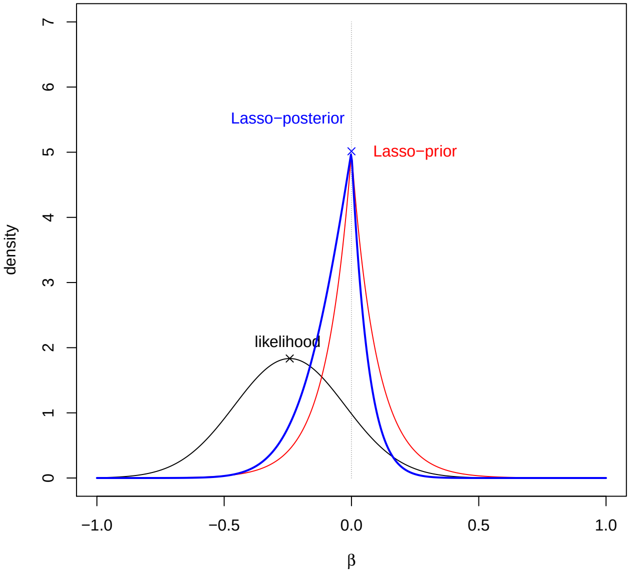


Figure: $y_i = 2 * x_i + \epsilon_i$ for $i = 1, \ldots, 30$

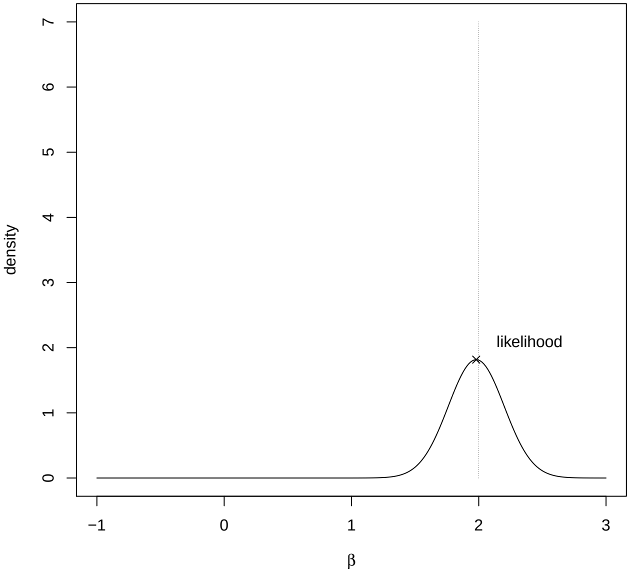


Figure: $y_i = 2 * x_i + \epsilon_i$ for $i = 1, \ldots, 30$

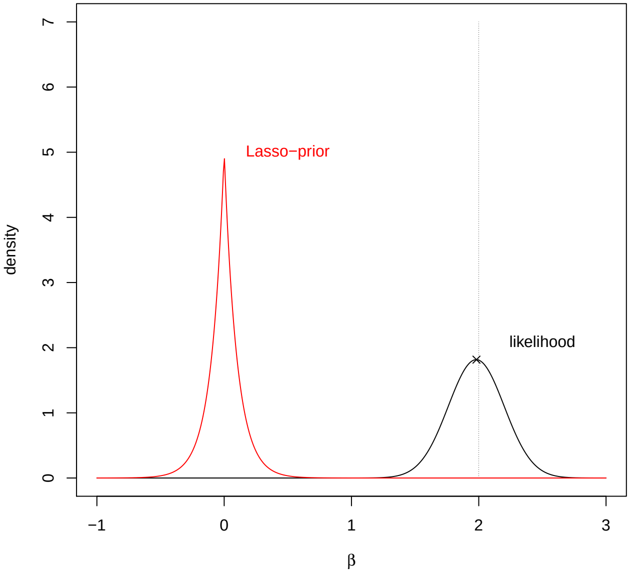
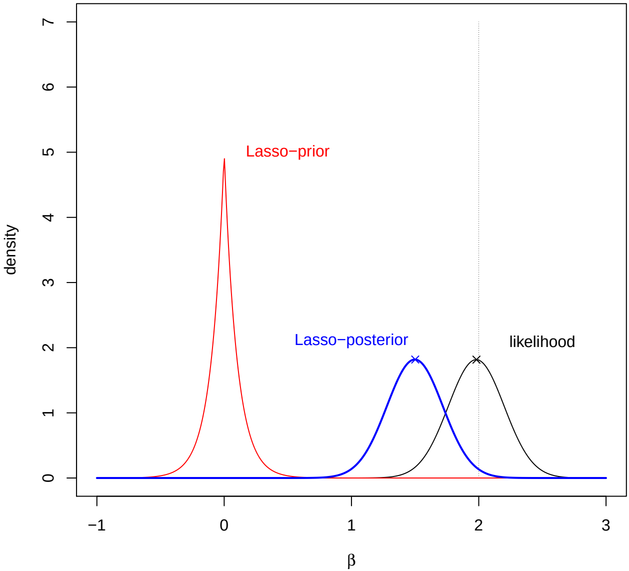


Figure: $y_i = 2 * x_i + \epsilon_i$ for $i = 1, \ldots, 30$



REFERENCES

Park, T. and G. Casella (2008), ‘The Bayesian Lasso’, *Journal of the American Statistical Association* **103**, 681–686.

Tibshirani, R. (1996), ‘Regression shrinkage and selection via the Lasso’, *Journal of the Royal Statistical Society: Series B* **58**, 267–288.