STAT 4010 – Bayesian Learning

Tutorial 3: Priors

Spring 2022

Cheuk Hin (Andy) CHENG (Email | Homepage)

Di SU (Email | Homepage)

1 Different Kinds of Priors

1.1 Jeffreys Priors

Definition 1. *(Fisher Information) Let $\mathscr{F} = \{f_\theta, \theta \in \Theta\}$ be a family of distribution and is Fisher's regular (see Definition 6.4 in Ch6 lecture note from S4003 by Keith). Let $x_{1:n} \overset{\text{IID}}{\sim} f_\theta$. If $\theta \in \mathbb{R}$,*

- *(Likelihood function) $L(\theta) = \prod_{i=1}^n f_\theta(x_i)$;*
- *(Log-likelihood function) $\ell(\theta) = \log L(\theta)$;*
- *(Score function) $S(\theta) = \frac{\partial \ell(\theta)}{\partial \theta}$;*
- *(Fisher information) $I(\theta) = \mathsf{Var}\left(\frac{\partial \ell(\theta)}{\partial \theta}\right) = -\mathsf{E}\left(\frac{\partial^2 \ell(\theta)}{\partial \theta^2}\right)$.*

If $\theta \in \mathbb{R}^p$, then $I(\theta)$ is a $p \times p$ matrix with entry

$$\{I(\theta)\}_{ij} = \mathsf{E}_\theta\left(\frac{\partial \ell(\theta)}{\partial \theta_i} \cdot \frac{\partial \ell(\theta)}{\partial \theta_j}\right) = -\mathsf{E}_\theta\left(\frac{\partial^2 \ell(\theta)}{\partial \theta_i \partial \theta_j}\right).$$

Theorem 1.1. *(Fisher transformation) Let $\phi(\theta)$ be a one-to-one differentiable function and $\theta = h(\phi)$. Under the assumption in Definition 1 and suppose the fisher information of θ exists, we have*

$$I_\phi(\phi) = \left(\frac{\partial \theta}{\partial \phi}\right) I_\theta(h(\phi)) \left(\frac{\partial \theta}{\partial \phi}\right)^T.$$

Definition 2. *(Jeffreys Prior) Given the model $\{f(\cdot \mid \theta) : \theta \in \Theta\}$, the Jeffrey prior for θ is defined to be,*

$$f(\theta) \propto \sqrt{det\{I(\theta)\}}.$$

Note that if θ is univariate, then $\sqrt{det\{I(\theta)\}} = \sqrt{I(\theta)}$.

Remark 1.1. Jeffreys prior is non-informative and invariant. However, it is usually improper, non-intuitive, hard to compute, and violates the likelihood principle.

Example 1.1. (Fisher information) Let $x_{1:n} \overset{\text{IID}}{\sim} \text{Poi}(\theta)$. Let $\eta = \log(\theta)$.

1. Compute the Fisher information for θ and η.

2. Propose a Jeffreys Prior for θ.

Solution:

1. Define $S_n = \sum_{i=1}^n x_i$. Note that $\mathsf{E}(S_n) = n\theta$ and $\mathsf{Var}(S_n) = n\theta$. Then,

$$L(\theta) = \prod_{i=1}^n \frac{e^{-\theta}\theta^{x_i}}{x_i!}\mathbb{1}(x_i = 0, 1, \cdots) = e^{-n\theta}\theta^{S_n}\prod_{i=1}^n \frac{1}{x_i!}\mathbb{1}(x_i = 0, 1, \cdots);$$

$$\ell(\theta) = -n\theta + S_n \log(\theta) + \log\left(\prod_{i=1}^n \frac{1}{x_i!}\mathbb{1}(x_i = 0, 1, \cdots)\right);$$

$$\frac{\partial \ell}{\partial \theta} = -n + \frac{S_n}{\theta};$$

$$\frac{\partial^2 \ell}{\partial \theta^2} = -\frac{S_n}{\theta^2};$$

$$I_\theta(\theta) = \mathsf{Var}\left(\frac{\partial \ell}{\partial \theta}\right) = -\mathsf{E}\left(\frac{\partial^2 \ell}{\partial \theta^2}\right) = \frac{n}{\theta}.$$

Note that $\theta = e^\eta$. Then,

$$I_\eta(\eta) = \left(\frac{\partial \theta}{\partial \eta}\right)^2 I_\theta(e^\eta) = ne^\eta.$$

2. By Definition 2, the Jeffreys prior for θ is,

$$f(\theta) \propto \theta^{-1/2}\mathbb{1}(\theta > 0).$$

Example 1.2. (Fisher information for multivariate parameter) Let $x_{1:n} \overset{\text{IID}}{\sim} \mathrm{N}(\theta, \Sigma)$, where $x_{1:n}$ are two-dimensional random vectors and Σ is the known covariance matrix. Propose a Jeffreys Prior for θ.

Solution:

For $i = 1, 2, \ldots, n$ and $j = 1, 2$, denote $\mathrm{Corr}(x_{i1}, x_{i2}) = \rho$ and $\mathsf{Var}(x_{ij}) = \sigma_j^2$. Write for $i = 1, 2, \ldots, n$ and $j = 1, 2$ that, $x_i = (x_{i1}, x_{i2})^T$, $\theta = (\theta_1, \theta_2)^T$ and $z_{ij} = \frac{x_{ij} - \theta_j}{\sigma_j}$. Note that

$\frac{\partial z_{ij}}{\partial \theta_j} = \frac{-1}{\sigma_j}$. Then,

$$
\begin{aligned}
L(\theta) &= \prod_{i=1}^{n} \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\{z_{i1}^2 + z_{i2}^2 - 2\rho z_{i1} z_{i2}\}\right) \\
\ell(\theta) &= -\sum_{i=1}^{n} \frac{1}{2(1-\rho^2)}\{z_{i1}^2 + z_{i2}^2 - 2\rho z_{i1} z_{i2}\} + \log\left(\prod_{i=1}^{n} \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\right) \\
\frac{\partial \ell(\theta)}{\partial \theta_1} &= -\sum_{i=1}^{n} \frac{1}{2(1-\rho^2)}\{-2\sigma_1^{-1} z_{i1} + 2\rho\sigma_1^{-1} z_{i2}\} \\
\frac{\partial \ell(\theta)}{\partial \theta_2} &= -\sum_{i=1}^{n} \frac{1}{2(1-\rho^2)}\{-2\sigma_2^{-1} z_{i2} + 2\rho\sigma_2^{-1} z_{i1}\} \\
\frac{\partial^2 \ell(\theta)}{\partial \theta_1^2} &= -\sum_{i=1}^{n} \frac{1}{2(1-\rho^2)}\{2\sigma_1^{-2}\} = -\frac{n}{\sigma_1^2(1-\rho^2)} \\
\frac{\partial^2 \ell(\theta)}{\partial \theta_2^2} &= -\frac{n}{\sigma_2^2(1-\rho^2)} \\
\frac{\partial^2 \ell(\theta)}{\partial \theta_1 \partial \theta_2} &= \sum_{i=1}^{n} \frac{1}{2(1-\rho^2)}\{2\rho\sigma_1^{-1}\sigma_2^{-1}\} = \frac{n\rho}{\sigma_1\sigma_2(1-\rho^2)}
\end{aligned}
$$

Observe that the second derivatives are free of θ, therefore the Fisher information is,

$$
I(\theta) = \begin{bmatrix} -\frac{n}{\sigma_1^2(1-\rho^2)} & \frac{n\rho}{\sigma_1\sigma_2(1-\rho^2)} \\ \frac{n\rho}{\sigma_1\sigma_2(1-\rho^2)} & -\frac{n}{\sigma_2^2(1-\rho^2)} \end{bmatrix}.
$$

Therefore, the Jeffreys prior is

$$
\begin{aligned}
f(\theta) &\propto \sqrt{det(I(\theta))} \\
&= \left(\frac{n}{\sigma_1\sigma_2(1-\rho^2)}\right)\sqrt{(1-\rho^2)} \\
&\propto 1.
\end{aligned}
$$

As an remark, θ is a location parameter and the Jeffreys prior is also an invariant prior.

1.2 Conjugate Priors

Definition 3. *(Conjugate Prior) A family of distribution $\mathscr{F}$ on Θ is conjugate for the sampling distribution if and only if for each prior $f(\theta) \in \mathscr{F}$, the posterior $f(\theta \mid x) \in \mathscr{F}$ as well.*

Definition 4. *(Exponential Family (EF)) The exponential family is a family of distribution with density function in the (canonical) form,*

$$
f(x \mid \theta) = h(x) exp\left\{\theta^T \cdot T(x) - A(\theta)\right\}. \tag{1.1}
$$

Theorem 1.2. *(Conjugate Prior for EF) Consider the EF defined in Defintion 4, the conjugate prior of η is,*

$$f(\theta) = K(\mu, \lambda) e^{\mu^T \theta - \lambda A(\theta)} \mathbb{1}(\theta \in \Theta),$$

where μ and λ are hyper parameter.

Example 1.3. (Conjugate Prior for Beta distribution) Let β be a known constant, the Beta density has form

$$f(x \mid \theta) = \frac{x^{\theta-1}(1-x)^{\beta-1}}{B(\theta,\beta)} \mathbb{1}_{x\in(0,1)} = \left(x^{-1}(1-x)^{\beta-1}\mathbb{1}_{x\in(0,1)}\right) e^{\theta \log x - \log B(\theta,\beta)},$$

where $\eta = \log\theta$. Therefore, the conjugate prior for θ by Theorem 1.2 is

$$f(\theta \mid \mu, \lambda) \propto e^{\mu\theta - \lambda \log B(\theta,\beta)} = \frac{e^{\mu\theta}}{B(\theta,\beta)^{\lambda}}. \tag{1.2}$$

Example 1.4. (Conjugate Priors of Common Families) Define $\bar{x} = \sum_{i=1}^n x_i$ and $s^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})^2$.

Sampling Distribution	Conjugate Prior ($[\theta]$)	Posterior ($[\theta \mid x_{1:n}]$)
$\mathrm{N}(\theta, \sigma^2)$	$\mathrm{N}(\mu, \lambda^2)$	$\mathrm{N}\left(\frac{n\bar{x}/\sigma^2+\mu/\lambda^2}{n/\sigma^2+1/\lambda^2}, \frac{1}{n/\sigma^2+1/\lambda^2}\right)$
$\mathrm{N}(\mu, \theta)$	$\frac{\beta}{Ga(\alpha)}$	$\frac{\beta+(n-1)s^2/2}{Ga(\alpha+n/2)}$
$\mathrm{Bin}(m, \theta)$	$\mathrm{Beta}(\alpha, \beta)$	$\mathrm{Beta}(\alpha + n\bar{x}, \beta + n\{m - \bar{x}\})$
$\mathrm{Poi}(\theta)$	$Ga(\alpha)/\beta$	$Ga(\alpha + n\bar{x})/(\beta + b)$
$Ga(v)/\theta$	$Ga(\alpha)/\beta$	$Ga(\alpha + nv)/(\beta + b + n\bar{x})$
$\mathrm{NB}(m, \theta)$	$\mathrm{Beta}(\alpha, \beta)$	$\mathrm{Beta}(\alpha + nm, \beta + n\bar{x})$

1.3 Informative (Subjective) Prior

Approaches to construct subjective priors:

1. (PDF/CDF approach) Partition the parameter space, $\Theta = \cup_{j=1}^J \Theta_j$ where J is a fixed predetermined constant. Specify the probabilities such that,

$$0 \le \Pr(\Theta_j) \le 1 \quad \text{and} \quad \sum_{j=1}^{J} \Pr(\Theta_j) = 1.$$

   Alternatively, the pdf can be specified using relative likelihood. For example, we may

specify

$$0 \leq \Pr(\Theta_j)/\Pr(\Theta^*) \leq 1 \quad \text{and} \quad \sum_{j=1}^{J} \Pr(\Theta_j) = 1,$$

where Θ^* can be regarded as baseline parameter space subset. Moreover, we can also express our belief through CDF. Fix $\{(\theta_j, c_j)\}_{j=1}^{J}$ and set the CDF as,

$$F(\theta_j) = \Pr(\theta \leq \theta_j) = c_j.$$

The other points of the CDF can be computed using interpolation.

2. (Moment matching approach) Say we already pick a parametric model on Θ, $\{f(\cdot \mid \lambda) : \lambda \in \Lambda\}$ where λ is hyperparameter. The hyperparameter is set by matching the moment or quantile.

Remark 1.2. In constructing informative priors, the same data cannot be used twice. Use different datasets to estimate the hyper-paramters and to conduct inference on DGP parameters.

1.4 Weakly-Informative Prior

Weakly-informative prior can be used to regularize and improve non-informative prior.

1. (Regularize) To avoid improper prior, a weakly informative prior is used. For example, set $\theta \sim \mathrm{N}(0, 9999)$ instead of specifying $f(\theta) \propto 1$.
2. (Make it sensible) Set a prior to realistically reflect real life situation.

2 Comparison and Examples

Example 2.1. (Different types of priors - Exercise 2.1 Assignment 2 Spring 2021). In the last semester, we collected the number of suspected cheating cases in 21 courses offered by the the Department of Statistics. Denote them by $x_1, \ldots, x_{21}$. Assume

$$[x_1, \ldots, x_{21} \mid \theta] \overset{\text{IID}}{\sim} \mathrm{Po}(\theta).$$

1. (20%) Suggest, with a brief explanation (≤ 20 words) or mathematical derivation,
   (a) a conjugate prior on θ,
   (b) an informative prior on θ,
   (c) a non-informative prior on θ, and
   (d) a weakly-informative prior on θ.
2. (10%) Edmund believes no prior information. So, he uses a nearly flat prior $\theta \sim \mathrm{N}(0, 100^2)$. Comment.

Solution:

1. (a) By Example 2.14, I can choose $\theta \sim \mathrm{Ga}(n_0)/m_0$, where n_0 and m_0 are to be determined.

   (b) I know 5 students in a class of size 50 cheated ("succeeded") in Spring 2020. Hence I can choose $\theta \sim \mathrm{Ga}(45)/9$ by matching the number of failures and the rate of success:
   $$n_0 = 45 \quad \text{and} \quad \frac{1}{m_0+1} = \frac{1}{10} \iff m_0 = 9.$$
   (Note: see Example 2.14 for the interpretation of the hyperparameters. Be careful that we cannot use data from Fall 2020 unless it comes from other department. The "5 students cheated" data is hypothetical and I have ignored the class size difference. This answer is given to illustrate that conjugate prior can also be informative if we inject information into it.)

   (c) By Definition 2.4 and Section 2.1, I can choose the Jeffreys prior $f(\theta) \propto 1/\sqrt{\theta}$ as it is parametrization invariant.

   (d) Since the average number of suspected cheating is unlikely to be out of $(0, 20)$, I can choose $\theta \sim \mathrm{Unif}(0, 20)$.

2. Edmund's choice is not appropriate since $\theta \sim \mathrm{N}(0, 100^2)$ can be negative and such $\mathrm{Po}(\theta)$ is not well defined.

Example 2.2. (Regime-switching Bernoulli - Exercise 2.2 Spring 2021). Let $x_1, \ldots, x_n \in \mathbb{R}$ be some known constants. Assume
$$[y_i \mid \theta] \sim \begin{cases} \mathrm{Bern}(0.5) & \text{if } x_i \le 0; \\ \mathrm{Bern}(0.5 + \theta/2) & \text{if } x_i > 0, \end{cases} \quad i = 1, \ldots, n,$$
$$\theta \sim \mathrm{Beta}(\alpha, \beta).$$

The sampling distribution (2.1) is called a regime-switching Bernoulli model because the success probability for the i th observation switches between 0.5 and $0.5 + \theta/2$ depending on the sign of x_i.

1. (10%) Find the posterior $[\theta \mid y_{1:n}]$.
2. (10%) Is the prior conjugate with the posterior? If yes, explain. If no, propose a conjugate prior for θ.

Solution:

1. The posterior is

$$\begin{aligned} f(\theta \mid y_{1:n}) &\propto f(y_{1:n} \mid \theta) f(\theta) \\ &\propto \left\{ \prod_{1\le i\le n: x_i>0} f(y_i \mid \theta) \right\} \underbrace{\left\{ \prod_{1\le i\le n: x_i\le 0} f(y_i \mid \theta) \right\}}_{\text{free of } \theta} f(\theta) \\ &\propto \left\{ \prod_{1\le i\le n: x_i>0} \left(\frac{1+\theta}{2}\right)^{y_i} \left(\frac{1-\theta}{2}\right)^{1-y_i} \right\} \theta^{\alpha-1}(1-\theta)^{\beta-1} \mathbb{1}_{(0<\theta<1)} \\ &= \left(\frac{1+\theta}{2}\right)^{\sum_{i=1}^n y_i \mathbb{1}_{(x_i>0)}} \left(\frac{1-\theta}{2}\right)^{\sum_{i=1}^n (1-y_i)\mathbf{1}_{(x_i>0)}} \theta^{\alpha-1}(1-\theta)^{\beta-1} \mathbb{1}_{(0<\theta<1)} \\ &\propto (1+\theta)^{\sum_{i=1}^n y_i \mathbb{1}_{(x_i>0)}} \theta^{\alpha-1}(1-\theta)^{\beta-1+\sum_{i=1}^n (1-y_i)\mathbb{1}_{(x_i>0)}} \mathbb{1}_{(0<\theta<1)}. \end{aligned}$$

2. Comparing with Example 2.13, we cannot find α_n, β_n such that

$$\frac{f(\theta \mid y_{1:n})}{\theta^{\alpha_n-1}(1-\theta)^{\beta_n-1}} \propto 1.$$

Hence $\theta \sim \text{Beta}(\alpha, \beta)$ is not a conjugate prior. However, we can re-parametrize the prior and consider

$$\phi = \frac{1+\theta}{2} \sim \text{Beta}(\alpha, \beta) \iff \theta = 2\phi - 1 \sim 2\,\text{Beta}(\alpha, \beta) - 1.$$

Then

$$\begin{aligned} f(\phi \mid y_{1:n}) &\propto f(y_{1:n} \mid \phi) f(\phi) \\ &= f(y_{1:n} \mid \theta) f(\phi) \\ &= \phi^{\alpha-1+\sum_{i=1}^n y_i \mathbf{1}_{(x_i>0)}} (1-\phi)^{\beta-1+\sum_{i=1}^n (1-y_i)\mathbb{1}_{(x_i>0)}\mathbb{1}_{(0<\phi<1)}} \end{aligned} \tag{2.1}$$

which shows that $[\phi \mid y_{1:n}] \sim \text{Beta}(\alpha_n, \beta_n)$ where

$$\alpha_n = \alpha + \sum_{i=1}^n y_i \mathbb{1}_{(x_i>0)} \quad \text{and} \quad \beta_n = \beta + \sum_{i=1}^n (1-y_i)\,\mathbb{1}_{(x_i>0)}.$$

(Note: equation (2.1) has used the fact that $f(y_{1:n} \mid \phi) \equiv f(y_{1:n} \mid \theta)$ when ϕ and θ have one-to-one relationship. As a result, we can put $\theta = 2\phi - 1$ into $f(y_{1:n} \mid \theta)$ obtained in part 1. We remark that $f(\phi \mid y_{1:n}) \not\equiv f(\theta \mid y_{1:n})$ as the absolute value of the determinant of Jacobian is not 1.)

Example 2.3. (Prior distribution - Q2 Midterm Spring 2021). If a rv equals to zero with probability θ and follows $\text{Po}(\phi)$ with probability $1-\theta$, then it is said to follow a zero-inflated Poisson (ZIP) distribution. Let $[x \mid \theta]$ follow a ZIP distribution, whose PMF is

$$f(x \mid \theta) = \left\{\theta + (1-\theta)e^{-\phi}\right\}^{\mathbb{1}(x=0)} \left\{\frac{(1-\theta)e^{-\phi}\phi^x}{x!}\right\}^{\mathbb{1}(x>0)} \mathbb{1}(x = 0, 1, \ldots),$$

where $\phi > 0$ is non-random.

1. (10%) Derive an invariant prior for θ.
2. (10%) Is the invariant prior you found in part (1) proper?
3. (10%) We believe a priori that $\theta > 1/2$ is three times as likely as $\theta \leq 1/2$. Suggest two informative prior distributions for θ. Mathematically justify your choices. Which one do you prefer? Why? Use no more than about 50 words.

Solution:

1. By Theorem 2.2, Jeffreys prior is parametrization invariant. Hence we consider the Jeffreys prior in this question. For simplicity, we drop the indictor $\mathbb{1}(x = 0, 1, \ldots)$ as we only consider x within its support. Note that

$$\begin{aligned}
\log f(x \mid \theta) &= \log\left\{\theta + (1-\theta)e^{-\phi}\right\}\mathbb{1}(x=0) + \{\log(1-\theta) - \phi + x\log(\phi) - \log(x!)\}\mathbb{1}(x>0),\\
\frac{\partial \log f(x \mid \theta)}{\partial \theta} &= \frac{1-e^{-\phi}}{\theta + (1-\theta)e^{-\phi}}\mathbb{1}(x=0) - \frac{1}{1-\theta}\mathbb{1}(x>0),\\
\frac{\partial^2 \log f(x \mid \theta)}{\partial \theta^2} &= -\frac{\left(1-e^{-\phi}\right)^2}{\left\{\theta + (1-\theta)e^{-\phi}\right\}^2}\mathbb{1}(x=0) - \frac{1}{(1-\theta)^2}\mathbb{1}(x>0).
\end{aligned}$$

   The Fisher information is

$$\begin{aligned}
I(\theta) &= -\mathrm{E}\left\{\frac{\partial^2 \log f(x \mid \theta)}{\partial \theta^2} \mid \theta\right\}\\
&= \frac{\left(1-e^{-\phi}\right)^2}{\theta + (1-\theta)e^{-\phi}} + \frac{1-e^{-\phi}}{1-\theta}\\
&\propto \frac{1}{(1-\theta)\left\{\theta + (1-\theta)e^{-\phi}\right\}}
\end{aligned}$$

   By Definition 2.4, the Jeffreys prior is

$$\begin{aligned}
f(\theta) &\propto \sqrt{I(\theta)} \times \mathbb{1}(0<\theta<1)\\
&= \sqrt{\frac{1}{(1-\theta)\left\{\theta + (1-\theta)e^{-\phi}\right\}}}\mathbb{1}(0<\theta<1)
\end{aligned}$$

2. We try to upper bound $f(\theta)$. Since $\phi > 0$ is non-random, we have for all $\theta \in (0,1)$ that $(1-\theta)e^{-\phi} > 0$ and thus $1/(\theta + (1-\theta)\exp(-\phi)) < 1/\theta$. Therefore, the probability kernel of the prior is bounded by the probability kernel of Beta$(0.5, 0.5)$, which is

$$f(\theta) < \sqrt{\frac{1}{\theta(1-\theta)}}.$$

   Therefore,

$$\int_\Theta f(\theta)\mathrm{d}\theta < \int_0^1 \sqrt{\frac{1}{\theta(1-\theta)}}\mathrm{d}\theta = B(0.5, 0.5) = \pi < \infty.$$

   The Jeffreys prior is proper.

3. We can set priors on θ satisfying $P\left(0 < \theta \leq \frac{1}{2}\right) = 1/4$ and $P\left(\theta > \frac{1}{2}\right) = 3/4$. For example,

   - $\theta \sim \text{Beta}(2, 1)$. It's easy to get this idea by letting $\beta - 1 = 0$ to "kill" the $1 - \theta$ term:
   $$f(\theta) = \frac{1}{B(2,1)}\theta^1(1-\theta)^0\mathbb{1}(0 < \theta < 1) = \frac{\theta}{2}\mathbb{1}(0 < \theta < 1).$$
   The parameter $\alpha = 2$ is found using the relative likelihood approach; see Example 2.17.
   - Piece-wise Uniform distribution. The easiest form is:
   $$f(\theta) = \begin{cases} \frac{1}{2}, & 0 < \theta \leq \frac{1}{2} \\ \frac{3}{2}, & \frac{1}{2} < \theta < 1 \end{cases}$$
   We can also use other distributions like Truncated-Normal. (Reasonable explanations are accepted.)

Example 2.4. (Identifiability - Midterm Spring 2021). Let

$$[x_1, \ldots, x_n, x_{n+1} \mid \theta] \overset{\text{ID}}{\sim} \text{Bern}(\theta),$$
$$\theta \sim \begin{cases} \text{Beta}\left(\alpha_0, \beta_0\right) & \text{with probability } p_0; \\ \text{Beta}\left(\alpha_1, \beta_1\right) & \text{with probability } p_1, \end{cases}$$

where $\alpha_0, \alpha_1, \beta_0, \beta_1, p_0, p_1 > 0$ are non-random such that $p_0 + p_1 = 1$. Andy claimed that "we can estimate p_0 and p_1 by using $x_{1:n}$ as long as n is large enough." Comment by no more than about 50 words.

Solution: The model is problematic. Both p_0 and p_1 are not identifiable. Consider $\alpha_0 = \alpha_1$ and $\beta_0 = \beta_1$, then different choices of p_0 or p_1 gives the same prior distribution.