3.5 Referring to example 3.6:

a. Verify the maximum of the likelihood ratio statistic.

Exmp. 3.6 gives the results of a study comparing radiation therapy with surgery in treating cancer of the larynx. The full parameter space is as follows,

	cancer controlled	cancer not controlled	
surgery	p_{11}	p_{12}	p_1
radiation	p_{21}	p_{22}	$1-p_1$
	p_2	$1-p_2$	1

Under this model, each observation X_i comes from a multinomial distribution with four cells & cells probabilities $\vec{p}=(p_{11},p_{12},p_{21},p_{22})$, with $\sum_{ij} p_{ij}=1$, that is $X_i \sim M_4(1,\vec{p})$, $i=1,\dots,n$. If we denote y_{ij} as the number of x_i that are in cell ij, the likelihood function can be written as $\ell(\vec{p}|\vec{y}) \propto \prod_{ij} p_{ij}^{y_{ij}}$.

The null hypothesis to be tested is one of independence, that is the treatment has no bearing on the control of cancer, $H_0: p_{11}=p_1p_2$. The likelihood ratio statistic for testing this hypothesis is $\lambda(\vec{y})=\frac{\max_{\vec{p}:p_{11}=p_1p_2}\ell(\vec{p}|\vec{y})}{\max_{\vec{p}}\ell(\vec{p}|\vec{y})}$.

The hypothesis space H is constrained by $0\le p_{ij}\le 1$ and $p_{i1}+p_{i2}=1$, thus $\dim(\Theta)=2$ where p_{1j}, p_{2j} are free parameters. The space of null hypothesis H_0 is the subspace where $p_{11}=p_1p_2$, thus $\dim(\Theta_0)=1$ where only one of p_{ij} is considered free parameter under H_0. The MLE of $\hat{p}_{ij}$ is given by the Lagrange multiplier.

MLE PMF$_{M_4}$ $f(\vec{y}|\vec{p})=\frac{n!}{y_{11}!\cdots y_{ij}!}p_{11}^{y_{11}}\cdots p_{ij}^{y_{ij}}$. Taking logarithm $\log\ell(\vec{p}|\vec{y})=\log n!(\prod_{ij}\frac{p_{ij}^{y_{ij}}}{y_{ij}!})$ $=\log n!+\sum_{ij}y_{ij}\log p_{ij}-\sum_{ij}\log y_{ij}!$. Using Lagrange multiplier with $\sum_{ij}p_{ij}=1$, $L(\vec{p},\lambda)=\log\ell(\vec{p}|\vec{y})+\lambda(1-\sum_{ij}p_{ij})$. To find maximum, we differentiate the lagrangian wrt. p_{ij}, $\frac{\partial}{\partial p_{ij}}L(\vec{p},\lambda)=\frac{\partial}{\partial p_{ij}}\log\ell(\vec{p}|\vec{y})+\frac{\partial}{\partial p_{ij}}\lambda(1-\sum_{ij}p_{ij})$ $=\frac{\partial}{\partial p_{ij}}\log\ell(\vec{p}|\vec{y})-\lambda=\frac{\partial}{\partial p_{ij}}(\log n!+\sum_{ij}y_{ij}\log p_{ij}-\sum_{ij}\log y_{ij}!)-\lambda=\frac{y_{ij}}{p_{ij}}-\lambda$. By setting the Lagrangian equal to zero, we can compute extremum $p_{ij}=\frac{y_{ij}}{\lambda}$. Solving λ, $\sum_{ij}p_{ij}=\sum_{ij}\frac{y_{ij}}{\lambda}=1$, $1=\frac{1}{\lambda}\sum_{ij}y_{ij}$, $1=\frac{n}{\lambda} \Rightarrow n=\lambda$. By MLE $\hat{p}_{ij}=\frac{y_{ij}}{n}$ under H. Since $\Theta_0\subseteq\Theta$, MLE for $\hat{p}_1=\frac{y_{11}+y_{12}}{n}$ under H_0. By 2^{nd} derivative, $\frac{\partial^2}{\partial p_{ij}^2}L(\vec{p},\lambda)$ $=\frac{\partial}{\partial p_{ij}}(\frac{y_{ij}}{p_{ij}}-\lambda)=-\frac{y_{ij}}{p_{ij}^2}<0$ at $\hat{p}_{ij}=\frac{y_{ij}}{n}$, $\therefore$ it is maximum.

Appendix

The log-likelihood ratio statistic for exmp. 3.6 is $-2\log\lambda=2\sum_{ij}y_{ij}\log(\frac{\hat{p}_{ij}}{\hat{p}_1})$. By Wilk's Theorem, as sample size $n\to\infty$, distribution of test statistic $-2\log\lambda$ asymptotically approaches χ^2 under H_0 with $df=df_{alt}-df_{null}$ (number of free parameters of models alternative & null respectively).

use if H_i variance too high.

$$\arg\min \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{m}\sum_{j=1}^{m} H_{ij} - v(\tau_i, X_i)\right)^2 \simeq v^*$$

$v^* \simeq \arg\min \frac{1}{n}\sum_{i=1}^{n}(H_i - v(\tau_i, X_i))^2$

$Cov(X,Y) = E(XY) - E(X)E(Y)$

3.21 Monte Carlo marginalisation is a technique for calculating a marginal density when simulating from a joint density. Let $(X_i, Y_i) \sim f_{XY}(x,y)$, independent, and the corresponding marginal distribution $f_X(x) = \int f_{XY}(x,y)\,dy$.

a. Let $w(x)$ be an arbitrary density. Show that

MC generation.

$$\lim_n \frac{1}{n}\sum_{i=1}^{n} \frac{f_{XY}(x^*, y_i)\, w(x_i)}{f_{XY}(x_i, y_i)} = \iint \frac{f_{XY}(x^*, y)\, w(x)}{f_{XY}(x,y)} f_{XY}(x,y)\,dx\,dy = f_X(x^*)$$

and so we have a Monte Carlo estimate of f_X, the marginal distribution of X, from only knowing the form of the joint distribution.

Soln • LHS: Given $\bar{h}_n = \frac{1}{n}\sum_{i=1}^{n} \frac{f_{XY}(x^*, y_i) w(x_i)}{f_{XY}(x_i, y_i)}$ where $h(x,y) = \frac{f_{XY}(x^*, y)\, w(x)}{f_{XY}(x,y)}$, $\bar{h}_n$ converges to $E_{f_{XY}}[h(x,y)]$ by Strong Law of Large Number.

$$\therefore E_{f_{XY}}[h(x,y)] = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \frac{f_{XY}(x^*, y_i)\, w(x_i)}{f_{XY}(x_i, y_i)} = LHS$$

• RHS: WLOG, $E_{f_{XY}}[h(x,y)] = \iint \frac{f_{XY}(x^*, y)\, w(x)}{f_{XY}(x,y)} f_{XY}(x,y)\,dx\,dy = \iint f_{XY}(x^*, y)\, w(x)\,dx\,dy$

$= \int k_1 g(y)\,dy = \int k_2 g(y)\,dy = \int f_{XY}(x^*, y)\,dy = f_X(x^*) = RHS$ for arbitrary $f_{XY}(x^*, y)$ and $w(x)$. $\therefore$ LHS = RHS (proven).

b. Let $X|Y=y \sim Gamma(y,1)$ and $Y \sim Exp(1)$. Use the technique of part a. to plot the marginal density of X. Compare it to the exact marginal.

Soln $f_{XY}(x,y) = f_X(x|y)\cdot f_Y(y) = \frac{x^{Y-1} e^{-x}}{\Gamma(1)}\cdot e^{-Y} = x^{Y-1} e^{-(x+y)}$

Let $w(x_i)$ be arbitrary density $X \sim Exp(1)$.

create linspace

$$h(x^*, x_i; y_i) = \frac{f_{XY}(x^*, y_i)\, w(x_i)}{f_{XY}(x_i, y_i)} = \frac{x^{*Y-1} e^{-(x^*+y)}}{x^{Y-1} e^{-(x+y)}}\cdot e^{-x} = \frac{x^{*Y-1}\cdot e^{-x^*}}{x^{Y-1}\cdot e^{-x}}\cdot e^{-x} = \frac{x^{*Y-1}\cdot e^{-x^*}}{x^{Y-1}} = \left(\frac{x^*}{x}\right)^{Y-1}\cdot e^{-x^*}, x^* \neq 0$$

We can train a neural network that minimises the loss function $x^* \simeq \arg\min \frac{1}{n}\sum_{i=1}^{n}(h(x_i, y_i) - x_i)^2$ from the samples drawn; $(X_i, Y_i) \sim f_{XY}(x,y)$. Let $f_{XY}(x,y)$ be a bivariate normal, WLOG of an arbitrary $Cov(x,y) = 0.5$; $\binom{X}{Y} \sim N_2\left(\binom{0}{0}, \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}\right)$. For a 4-layer neural network..

The following code is ran on Anaconda environment (Python 3.9).

```
import torch
import torch.nn as nn
import torch.optim as optim
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import norm

#Function to generate samples from true bivariate normal distribution
def generate_true_samples(num_samples):
    mean = [0,0]
    covariance_matrix = [[1,0.5],[0.5,1]]
    return np.random.multivariate_normal(mean, covariance_matrix, num_samples)

#Marginal distribution PDF; h(x,y)
def custom_function(v1,v2,v3):
    result_tensor = (v3/v1)**(v2-1) * torch.exp(-v3)
    return result_tensor
```

Using the function $h(x^*, x_i, y_i) = \left(\frac{x^*}{x_i}\right)^{Y_i - 1}\cdot e^{-x^*}$ from above case.

```
#Neural network model
class NeuralNetwork(nn.Module):
    def __init__(self):
```

4.1 (Chen & Shao 1997) As mentioned, normalising constants are superfluous in Bayesian inference except in the case when several models are considered at once (as in the computation of Bayes factors). In such cases, where $\pi_1(\theta) = \frac{\tilde{\pi}_1(\theta)}{c_1}$ and $\pi_2(\theta) = \frac{\tilde{\pi}_2(\theta)}{c_2}$, & only $\tilde{\pi}_1$ and $\tilde{\pi}_2$ are known, the quantity to approximate is $\rho = \frac{c_1}{c_2}$ or $\xi = \log\left(\frac{c_1}{c_2}\right)$.

a. Show that the ratio ρ can be approximated by $\frac{1}{n}\sum_{i=1}^{n} \frac{\tilde{\pi}_1(\theta_i)}{\tilde{\pi}_2(\theta_i)}$, $\theta_1, \dots, \theta_n \sim \pi_2$.
(Hint: Use an importance sampling argument).

~~Soln~~ $\cancel{\frac{1}{n}\sum_{i=1}^{n}\frac{\tilde{\pi}_1(\theta_i)}{\tilde{\pi}_2(\theta_i)} \approx E_{\tilde{\pi}_2}\left[\frac{\tilde{\pi}_1(\theta)}{\tilde{\pi}_2(\theta)}\right] = \int \frac{\tilde{\pi}_1}{\tilde{\pi}_2}\cdot \tilde{\pi}_2 \, d\theta = \int \tilde{\pi}_1 \cdot \frac{\tilde{\pi}_2}{\tilde{\pi}_2} d\theta = \int \tilde{\pi}_1 \, d\theta = c_1}$, ~~$c_1$ is normalising constant~~

~~$\therefore$ For some known $\tilde{\pi}_1, \tilde{\pi}_2$, ratio $\rho = \frac{c_j}{c_k}$ can be estimated, $j, k \in \mathbb{Z}^+$.~~

~~b. Show that $\frac{\int \tilde{\pi}_1(\theta)\alpha(\theta)\pi_2(\theta)\,d\theta}{\int \tilde{\pi}_2(\theta)\alpha(\theta)\pi_1(\theta)\,d\theta} = \frac{c_1}{c_2}$ holds for every function $\alpha(\theta)$ such that both integrals are finite.~~

~~Soln~~

Soln Under common support $\Omega_1 \cap \Omega_2$ for π_1, π_2.

$$\frac{c_1}{c_2} = \frac{\int_{\Omega_1\cap\Omega_2} \tilde{\pi}_1(\theta)\,d\theta}{\int_{\Omega_1\cap\Omega_2} \tilde{\pi}_2(\theta)\,d\theta} = E_{\pi_2}\left[\frac{\tilde{\pi}_1}{\tilde{\pi}_2}\right], \text{ by Bridge representation of Bayes factor.}$$

$$\simeq \frac{1}{n}\sum_{i=1}^{n} \frac{\tilde{\pi}_1(\theta_i)}{\tilde{\pi}_2(\theta_i)}, \quad \theta_1 \dots \theta_n \sim \pi_2 \text{ (shown)}$$

b. Show that $\frac{\int \tilde{\pi}_1(\theta)\alpha(\theta)\pi_2(\theta)\,d\theta}{\int \tilde{\pi}_2(\theta)\alpha(\theta)\pi_1(\theta)\,d\theta} = \frac{c_1}{c_2}$ holds for every function $\alpha(\theta)$ such that both integrals are finite.

Soln Under common support $\Omega_1 \cap \Omega_2$ for π_1, π_2. Let $\alpha(\theta)$ be arbitrary function defined on $\Omega_1 \cap \Omega_2$. Consider the identity $0 < \left|\int_{\Omega_1\cap\Omega_2} \alpha(\theta)\pi_1(\theta)\pi_2(\theta)\,d\theta\right| < \infty$.

Using identity $-|f| \leq f \leq |f|$, $0 < \int_{\Omega_1\cap\Omega_2} \alpha(\theta)\pi_1(\theta)\pi_2(\theta)\,d\theta \leq \left|\int_{\Omega_1\cap\Omega_2}\alpha(\theta)\pi_1(\theta)\pi_2(\theta)\,d\theta\right| < \infty$.
Given $\int_{\Omega_1\cap\Omega_2} \pi_1(\theta)\pi_2(\theta)\,d\theta > 0$ for pdf. π_1, π_2, there exists arbitrary positive function $\alpha(\theta)$.

Given any positive $\alpha(\theta)$, $\frac{\int_{\Omega_2} \tilde{\pi}_1(\theta)\alpha(\theta)\pi_2(\theta)\,d\theta}{\int_{\Omega_1} \tilde{\pi}_2(\theta)\alpha(\theta)\pi_1(\theta)\,d\theta} = \frac{c_1}{c_2} \times \frac{\int_{\Omega_1\cap\Omega_2} \alpha(\theta)\pi_1(\theta)\pi_2(\theta)\,d\theta}{\int_{\Omega_1\cap\Omega_2}\alpha(\theta)\pi_1(\theta)\pi_2(\theta)\,d\theta} = \frac{c_1}{c_2}$.
$\therefore$ the fraction holds. (shown)

c. Deduce that $\frac{\frac{1}{n_2}\sum_{i=1}^{n_2} \tilde{\pi}_1(\theta_{2_i})\alpha(\theta_{2_i})}{\frac{1}{n_1}\sum_{i=1}^{n_1} \tilde{\pi}_2(\theta_{1_i})\alpha(\theta_{1_i})}$, with $\theta_{1_i} \sim \pi_1$ and $\theta_{2_i} \sim \pi_2$, is convergent estimator of $\rho = \frac{c_1}{c_2}$

Soln Using generalised representation of Bridge sampling, and under the common support $\Omega_1 \cap \Omega_2$ for π_1, π_2.

$$\frac{c_1}{c_2} = \frac{\int_{\Omega_2} \tilde{\pi}_1(\theta)\alpha(\theta)\pi_2(\theta)\,d\theta}{\int_{\Omega_1}\tilde{\pi}_2(\theta)\alpha(\theta)\pi_1(\theta)\,d\theta} = \frac{E_{\pi_2}[\tilde{\pi}_1(\theta)\alpha(\theta)]}{E_{\pi_1}[\tilde{\pi}_2(\theta)\alpha(\theta)]}$$

~~We introduce $\pi_0(\theta) = \frac{\tilde{\pi}_0(\theta)}{c_0}$ such that,~~

$$\cancel{\frac{c_1}{c_2} = \frac{\left(\frac{c_0}{c_2}\right)}{\left(\frac{c_0}{c_1}\right)} = \frac{E_{\pi_2}\left[\frac{\tilde{\pi}_0}{\tilde{\pi}_2}\right]}{E_{\pi_1}\left[\frac{\tilde{\pi}_0}{\tilde{\pi}_1}\right]}}$$

$$\simeq \frac{\frac{1}{n_2}\sum_{i=1}^{n_2}\tilde{\pi}_1(\theta_{2_i})\alpha(\theta_{2_i})}{\frac{1}{n_1}\sum_{i=1}^{n_1}\tilde{\pi}_2(\theta_{1_i})\alpha(\theta_{1_i})}, \text{ MC estimate.}$$

Given $\bar{h}_m = \frac{1}{m}\sum_{i=1}^{m} h(x_i)$, $\bar{h}_m$ converges to $E_f[h(x)]$ by Strong Law of Large Number. $\therefore$ the Bridge Estimator is convergent to ρ.

d. Show that part b. covers the case of the Newton & Raftery (1994) representation $\frac{c_1}{c_2} = \frac{E_{\pi_2}[\tilde{\pi}_2(\theta)^{-1}]}{E_{\pi_1}[\tilde{\pi}_1(\theta)^{-1}]}$.

Soln For any positive $\alpha(\theta)$, taking $\alpha(\theta) = \frac{1}{\tilde{\pi}_1(\theta)\tilde{\pi}_2(\theta)}$ under common support $\Omega_1 \cap \Omega_2$

$$\frac{c_1}{c_2} = \frac{\int_{\Omega_2} \tilde{\pi}_1(\theta)\left(\frac{1}{\tilde{\pi}_1(\theta)\tilde{\pi}_2(\theta)}\right)\pi_2(\theta)\,d\theta}{\int_{\Omega_1} \tilde{\pi}_2(\theta)\left(\frac{1}{\tilde{\pi}_1(\theta)\tilde{\pi}_2(\theta)}\right)\pi_1(\theta)\,d\theta} = \frac{\int_{\Omega_2} \frac{\pi_2(\theta)}{\tilde{\pi}_2(\theta)}\,d\theta}{\int_{\Omega_1} \frac{\pi_1(\theta)}{\tilde{\pi}_1(\theta)}\,d\theta} = \frac{E_{\pi_2}[\tilde{\pi}_2(\theta)^{-1}]}{E_{\pi_1}[\tilde{\pi}_1(\theta)^{-1}]} \quad \text{(shown).}$$

e. Show that the optimal choice (in terms of mean square error) of $\alpha(\theta)$ in part c is $\alpha(\theta) = c\,\frac{n_1+n_2}{n_1\pi_1(\theta)+n_2\pi_2(\theta)}$, where c is constant.

Soln Under common support $\Omega_1 \cap \Omega_2$ for π_1, π_2. Given $\hat{r} \simeq \frac{\int_{\Omega_2} \tilde{\pi}_1(\theta)\alpha(\theta)\pi_2(\theta)\,d\theta}{\int_{\Omega_1} \tilde{\pi}_2(\theta)\alpha(\theta)\pi_1(\theta)\,d\theta}$ and the relative mean square error $MSE(\hat{r}) = \frac{E(\hat{r}-r)^2}{r^2}$.

Using the identity $\frac{\int_{\Omega_2} \tilde{\pi}_1(\theta)\alpha(\theta)\pi_2(\theta)\,d\theta}{\int_{\Omega_1} \tilde{\pi}_2(\theta)\alpha(\theta)\pi_1(\theta)\,d\theta} = \frac{c_1}{c_2} \times \frac{\int_{\Omega_1\cap\Omega_2} \alpha(\theta)\pi_1(\theta)\pi_2(\theta)\,d\theta}{\int_{\Omega_1\cap\Omega_2} \alpha(\theta)\pi_1(\theta)\pi_2(\theta)\,d\theta}$. Let d_1, d_2 be numerator & denominator of $\hat{r}$ respectively.

$$\bar{d}_i = E(d_i) = c_i \int_{\Omega_1\cap\Omega_2} \alpha(\theta)\pi_1(\theta)\pi_2(\theta)\,d\theta, \quad i=1,2.$$

$$Var(d_1) = \frac{c_1^2}{n_2}\left(\int_{\Omega_1\cap\Omega_2} \pi_1^2(\theta)\pi_2(\theta)\alpha(\theta)\,d\theta - \left(\int_{\Omega_1\cap\Omega_2} \alpha(\theta)\pi_1(\theta)\pi_2(\theta)\,d\theta\right)^2\right)$$

$$Var(d_2) = \frac{c_2^2}{n_1}\left(\int_{\Omega_1\cap\Omega_2} \pi_2^2(\theta)\pi_1(\theta)\alpha(\theta)\,d\theta - \left(\int_{\Omega_1\cap\Omega_2} \alpha(\theta)\pi_1(\theta)\pi_2(\theta)\,d\theta\right)^2\right)$$

By the δ-method,

$$MSE(\hat{r}) = \frac{E\left(\frac{d_1}{d_2} - \frac{\bar{d}_1}{\bar{d}_2}\right)^2}{\left(\frac{\bar{d}_1}{\bar{d}_2}\right)^2} = \frac{Var(d_1)}{\bar{d}_1^2} + \frac{Var(d_2)}{\bar{d}_2^2} + O\left(\frac{1}{n^2}\right)$$

$$= \frac{1}{ns_1s_2}\left(\frac{\int_{\Omega_1\cap\Omega_2} \pi_1\pi_2(s_1\pi_1+s_2\pi_2)\alpha^2\,d\theta}{\left(\int_{\Omega_1\cap\Omega_2} \pi_1\pi_2\alpha\,d\theta\right)^2} - 1\right)$$

$$= \frac{1}{n}\left(\frac{\int_{\Omega_1\cap\Omega_2} s_1\pi_1 s_2\pi_2(s_1\pi_1+s_2\pi_2)\alpha^2\,d\theta}{\left(\int_{\Omega_1\cap\Omega_2} s_1\pi_1 s_2\pi_2\alpha\,d\theta\right)^2}\right) - \frac{1}{n_1} - \frac{1}{n_2},$$

Where $n = n_1+n_2$, $s_i = n_i/n$ and s_i $(i=1,2)$ are assumed to be asymptotically between 0 and 1. $\therefore$ $MSE(\hat{r})$ is minimised at $\alpha(\theta) = c\,\frac{n_1+n_2}{n_1\pi_1+n_2\pi_2}$, $\theta \in \Omega_1\cap\Omega_2$, & that α is optimal choice (shown).

4.2 (Continuation of Problem 4.1) When the priors π_1 and π_2 belong to a parameterised family (that is, $\pi_i(\theta) = \pi(\theta|\lambda_i)$), the corresponding constants are denoted by $c(\lambda_i)$.

a. Verify the identity $-\log\left(\frac{c(\lambda_1)}{c(\lambda_2)}\right) = E\left[\frac{U(\theta,\lambda)}{\pi(\lambda)}\right]$, where $U(\theta,\lambda) = \frac{d}{d\lambda}\log(\tilde{\pi}(\theta|\lambda))$ and $\pi(\lambda)$ is an arbitrary distribution on λ.

Soln Given two unnormalised densities with same support $\Omega_1 \cap \Omega_2$, we can use the geometric path to link $\tilde{\pi}_1$ and $\tilde{\pi}_2$: $\tilde{\pi}(\theta|\lambda) = \tilde{\pi}_1(\theta)^{1-\lambda}\tilde{\pi}_2(\theta)^{\lambda}$, $\lambda \in [0,1]$.

WLOG taking $\lambda_1 = 0$, $\lambda_2 = 1$ under $\lambda \in [0,1]$ for geometric path.

$$-\log\left(\frac{c(0)}{c(1)}\right) = \log\left(\frac{c(1)}{c(0)}\right) = \log\left(\frac{\int\tilde{\pi}(\theta|1)\,d\theta}{\int\tilde{\pi}(\theta|0)\,d\theta}\right), \text{ by definition of normalising constant}$$

$$= \int_0^1\left(\frac{\int\frac{d}{d\lambda}\tilde{\pi}(\theta|\lambda)\,d\theta}{\int\tilde{\pi}(\theta|\lambda)\,d\theta}\right)d\lambda, \text{ from } \int_0^1\frac{\int\frac{d}{d\lambda}\tilde{\pi}(\theta|\lambda)\,d\theta}{\int\tilde{\pi}(\theta|\lambda)\,d\theta}\,d\lambda = \left[\log\int\tilde{\pi}(\theta|\lambda)d\theta\right]_0^1$$

$$= \int_0^1\frac{\int\frac{d}{d\lambda}\tilde{\pi}(\theta|\lambda)\,d\theta}{c(\lambda)}\,d\lambda$$

$$= \int_0^1\left(\int\frac{\left(\frac{d}{d\lambda}\tilde{\pi}(\theta|\lambda)\right)\pi(\theta|\lambda)}{\tilde{\pi}(\theta|\lambda)}\,d\theta\right)d\lambda, \text{ from } \pi(\theta|\lambda) = \frac{\tilde{\pi}(\theta|\lambda)}{c(\lambda)}$$

$$= \int_0^1\left(\int\left(\frac{d}{d\lambda}\log\tilde{\pi}(\theta|\lambda)\right)\pi(\theta|\lambda)\,d\theta\right)d\lambda, \text{ using } (\log f)' = \frac{f'}{f}$$

$$= \int_0^1 E_{\theta|\lambda}\left[\frac{d}{d\lambda}\log\tilde{\pi}(\theta|\lambda)\right]d\lambda$$

$$= \int_0^1 E_{\theta|\lambda}\left[U(\theta,\lambda)\right]d\lambda, \text{ using identity } U(\theta,\lambda) = \frac{d}{d\lambda}\log(\tilde{\pi}(\theta|\lambda))$$

Considering $\lambda \sim U_{[0,1]}$, we can interpret $\int_0^1 E_{\theta|\lambda}[U(\theta,\lambda)]\,d\lambda$ as expectation of $U(\theta,\lambda)$ over the joint distribution $\pi(\theta,\lambda)$. That is

$$-\log\left(\frac{c(\lambda_1)}{c(\lambda_2)}\right) = \int_0^1 E_{\theta|\lambda}[U(\theta,\lambda)]\,d\lambda = E_{\theta|\lambda}[U(\theta,\lambda)] = E_{\theta,\lambda}\left[\frac{U(\theta,\lambda)}{\pi(\lambda)}\right]. \text{ (shown).}$$

b. Show that $\xi = \log\left(\frac{c_1}{c_2}\right)$ can be estimated with the Bridge estimator of Gelman and Meng (1998), $\hat{\xi} = \frac{1}{n}\sum_{i=1}^{n}\frac{U(\theta_i,\lambda_i)}{\pi(\lambda_i)}$, when the (θ_i,λ_i)'s are simulated from the joint density induced by $\pi(\lambda)$ and $\pi(\theta|\lambda_i)$.

Soln By MC estimate, $E_{\theta,\lambda}\left[\frac{U(\theta,\lambda)}{\pi(\lambda)}\right] \simeq \frac{1}{n}\sum_{i=1}^{n}\frac{U(\theta_i,\lambda_i)}{\pi(\lambda_i)}$.

4.12 In the setting of Section 4.3, examine whether the substituition of $\sum_{i=1}^{n-1}(X_{(i+1)}-X_{(i)})\frac{h(X_{(i)})+h(X_{(i+1)})}{2}$ into $\sum_{i=1}^{n-1}(X_{(i+1)}-X_{(i)})h(X_{(i)})$ improves the speed of convergence. (Hint: Examine the influence of the remainder terms $\int_{-\infty}^{X_{(1)}} h(x)f(x)\,dx$ and $\int_{X_{(n)}}^{+\infty} h(x)f(x)\,dx$.)

Soln Let $Z_1 = X_1$, $Z_i = X_i - X_{i-1}$, and $Z_{n+1} = 1 - X_n$.

For $\theta_1 = \sum_{i=1}^{n-1}(X_{i+1}-X_i)h(X_i)$, we take a, b such that $0 \le a < b \le 1$ for the error bound calculation of the Left-end point, such that

$$\int_a^b f(t)\,dt - (b-a)h(a) = \frac{(b-a)^2}{2} f'(\xi) \quad \text{where } a \le \xi \le b.$$

If $c_1 \ge |f'(x)|$, $0 \le x \le 1$, then

$$|\theta - \theta_1| \le \left|\int_0^{X_1} f(t)\,dt - X_1 h(0)\right| + \left|\sum_{i=1}^{n-1}\int_{X_i}^{X_{i+1}} f(t)\,dt - (X_{i+1}-X_i)h(X_i)\right| + \left|\int_{X_n}^1 f(t)\,dt - (1-X_n)h(X_n)\right| \le \frac{c_1}{2}\sum_{i=1}^{n+1} Z_i^2$$

$$Var(\theta_1) = E[(\theta-\theta_1)^2] \le E\left[\left(\frac{c_1}{2}\sum_{i=1}^{n+1} Z_i^2\right)^2\right] = \frac{c_1^2}{4}\left[\sum_{i=1}^{n+1} E[Z_i^4] + \sum_{i\ne j} E[Z_i^2 Z_j^2]\right]$$

For $\theta_2 = \sum_{i=1}^{n-1}(X_{i+1}-X_i)\frac{h(X_i)+h(X_{i+1})}{2}$, we take similar boundary conditions for the error bound calculation of Trapezoidal rule such that

$$\int_a^b f(t)\,dt - \frac{b-a}{2}[f(b)+f(a)] = -\frac{(b-a)^3}{12} f''(\xi) \quad \text{where } a \le \xi \le b.$$

If $c_2 \ge |f''(x)|$, $0 \le x \le 1$, then

$$|\theta - \theta_2| \le \left|\int_0^{X_1} f(t)\,dt - \tfrac{1}{2}X_1(h(0)+h(X_1))\right| + \left|\sum_{i=1}^{n-1}\int_{X_i}^{X_{i+1}} f(t)\,dt - \tfrac{1}{2}(X_{i+1}-X_i)(h(X_{i+1})+h(X_i))\right| + \left|\int_{X_n}^1 f(t)\,dt - \tfrac{1}{2}(1-X_n)(h(1)+h(X_n))\right| \le \frac{c_2}{12}\sum_{i=1}^{n+1} Z_i^3.$$

$$Var(\theta_2) = E[(\theta-\theta_2)^2] \le E\left[\left(\frac{c_2}{12}\sum_{i=1}^{n+1} Z_i^3\right)^2\right] = \frac{c_2^2}{144}\left[\sum_{i=1}^{n+1} E[Z_i^6] + \sum_{i\ne j} E[Z_i^3 Z_j^3]\right]$$

Given $0 < Z_i = \frac{1}{n}$, $n \in \mathbb{Z}^+$. $Var(\theta_2) \le Var(\theta_1)$ when θ_2 is substituited into θ_1 & therefore improves the speed of convergence.

1.40 Consider the mixture density $X \sim f(x|p) = \sum_{i=1}^{k} p_i f_i(x)$, where $p_i > 0$, $\sum_i p_i = 1$ and the densities f_i are known. The prior $\pi(p)$ is a Dirichlet distribution $D(\alpha_1, \dots, \alpha_k)$.

a. Explain why the computing time could get prohibitive as sample size increases.

Soln:

Sampling from the Dirichlet process mixture is unable to scale to large datasets due to high computational costs associated with Bayesian inference. The Gibbs sampling iteratively updates the posterior ; $\pi(p|x) \propto f(x|p)\pi(p)$, and given that the posterior distribution considers all possible partitions of the sample from the mixture model, the computational time is expensive.

b. A sequential alternative which approximates the Bayes estimator is to replace $\pi(p|x_1 .. x_n)$ by $D(\alpha_1^{(n)}, \dots, \alpha_k^{(n)})$ with $\alpha_1^{(n)} = \alpha_1^{(n-1)} + P(Z_{n1} = 1|x_n), \dots, \alpha_k^{(n)} = \alpha_k^{(n-1)} + P(Z_{nk} = 1|x_n)$ and Z_{ni} $(1 \le i \le k)$ is the component indicator vector of X_n. Justify this approximation and compare with the updating of $\pi(p|x_1..x_{n-1})$ when x_n is observed.

Soln: Under Multinomial Mixture Model with n datasets with k clusters.

Let $X \sim \text{Multinomial}(n, p)$ where $p = (p_1, p_2, \dots, p_k)^T$ with Dirichlet prior $\pi(p)$ under probability simplex $\Delta_k = \{p = (p_1, p_2, \dots, p_k)^T \in \mathbb{R}^k \mid p_i \geq 0 \text{ for } \forall i \text{ and } \sum_{i=1}^{k} p_i = 1\}$.

Dirichlet distribution has pdf. $\pi(p) = \dfrac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} p_i^{\alpha_i - 1}$.

The sample space of multinomial with k clusters is the set of vertices of the k-dimensional hypercube $\mathbb{H}_k$ with component indicator vector $\vec{x} = (0, 0, \dots, 0, 1, 0, \dots, 0)^T$ for $\vec{x} \in \mathbb{R}^k$. For n dataset, $1 \le j \le n$, let $X_j = (X_{j1}, \dots, X_{jk})^T$. Using the prior & distribution function,

$$p \sim Dir(\alpha_1, \dots, \alpha_k) \text{ and } X_j | p \sim \text{Multinomial}(p) \text{ for } j = 1, \dots, n.$$

This gives the posterior, $\pi(p|x_1 .. x_n) \propto \prod_{j=1}^{n} f(x_j|p)\pi(p) \propto \prod_{j=1}^{n}\prod_{i=1}^{k} p_i^{x_{ji}} \prod_{i=1}^{k} p_i^{\alpha_i - 1} = \prod_{i=1}^{k} p_i^{\sum_{j=1}^{n} x_{ji} + \alpha_i - 1}$

$\therefore$ the posterior $\pi(p|x_1 .. x_n)$ is also a dirichlet distribution given by

$$p | X_1 \dots X_n \sim Dir(\alpha + n\bar{X}) \text{ where } \bar{X} = \frac{1}{n}\sum_{j=1}^{n} X_j \in \Delta_k. \text{ (justified)}$$

The updating of $\pi(p|x_1 .. x_n)$ when x_n is observed adds to the α_i parameter of the Dirichlet distribution ~~, thus skewing the distribution to the parameter of choice ;~~ which the cluster x_n is classified in. Thus it skews the distribution to the parameter of choice ; α_i.