STK4021 Solutions - Exercises 11 and 12

By Dennis Christensen

Exercise 11. This exercise introduces the binomial distribution, and looks at conjugate priors.

(a) Let $y = (y_1, \ldots, y_n)$, where all the y_i are iid Bernoulli variables with parameter θ, with density

$$f_\theta(y_i) = \theta^{y_i}(1-\theta)^{1-y_i},$$

for $y_i \in \{0, 1\}$. Verify that $\mathbb{E}[y_i] = \theta$ and that $\mathrm{Var}(y_i) = \theta(1-\theta)$.

(b) Write down the log-likelihood for the full data y and verify that the maximum likelihood estimator $\hat{\theta}$ for θ is the sample mean.

(c) Now let m be the number of observations with $y_i = 1$. In other words, $m = \sum_{i=1}^n y_i$, a sum of independent Bernoulli trials. We know that m follows the Binomial distribution, $m \mid \theta \sim \mathrm{Binomial}(\theta, \mathrm{n})$, with density

$$f_\theta(m) = \binom{n}{m}\theta^m(1-\theta)^{n-m},$$

for $m = 0, 1, \ldots, n$.

Verify that $\mathbb{E}[m] = \theta n$ and $\mathrm{Var}(m) = \theta(1-\theta)n$.

(d) Now verify that the conjugate prior for θ is the Beta distribution $\mathrm{Beta(a, b)}$, with density

$$\pi(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1},$$

where $a, b > 0$. Show also that

$$\mathbb{E}[\theta] = \frac{a}{a+b}, \qquad \mathrm{Var}(\theta) = \frac{ab}{(a+b)^2(a+b+1)}.$$

(e) (Based on Bishop (2006, chapter 2)) Although the Beta prior and the binomial likelihood clearly share the same functional form as a function of θ, it is less obvious where the normalisation constant in the Beta distribution comes from. In this exercise, we verify that this normalisation constant is correct. We need to show that

$$\int_0^1 s^{a-1}(1-s)^{b-1}\,\mathrm{d}s = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)},$$

where we recall the definition of the Gamma function,

$$\Gamma(a) = \int_0^\infty x^{a-1} e^{-x} \, \mathrm{d}x.$$

Now, first note that

$$\begin{aligned}
\Gamma(a)\Gamma(b) &= \int_0^\infty x^{a-1} \exp(-x) \, \mathrm{d}x \int_0^\infty y^{b-1} \exp(-y) \, \mathrm{d}y \\
&= \int_0^\infty x^{a-1} \left(\int_0^\infty y^{b-1} \exp(-(x+y)) \, \mathrm{d}y \right) \mathrm{d}x.
\end{aligned}$$

Prove the normalisation by first substituting $t = y + x$ (holding x fixed), then change the order of integration between x and t, and finally make the substitution $x = ts$ (holding t fixed).

(f) Find the posterior distribution $\pi(\theta \mid m)$. Show that we can write

$$\mathbb{E}[\theta \mid m] = \lambda \mathbb{E}[\theta] + (1 - \lambda)\hat{\theta}$$

for some $0 \leq \lambda \leq 1$ which you should identify. What happens to $\mathbb{E}[\theta \mid m]$ as the number of observations goes to infinity?

Solution.

(a) We have

$$\begin{aligned}
\mathbb{E}[y_i] &= 0 \times \mathbb{P}(y_i = 0) + 1 \times \mathbb{P}(y_i = 1) = 0 + f_\theta(1) = \theta \\
\mathbb{E}[y_i^2] &= 0^2 \times \mathbb{P}(y_i = 0) + 1^2 \times \mathbb{P}(y_i = 1) = 0 + f_\theta(1) = \theta,
\end{aligned}$$

so that

$$\mathrm{Var}(y_i) = \mathbb{E}[y_i^2] - \mathbb{E}[y_i^2] = \theta - \theta^2 = \theta(1 - \theta).$$

(b) The log-likelihood is given by

$$\begin{aligned}
\ell(\theta) &= \log \left\{ \prod_{i=1}^n \theta^{y_i} (1 - \theta)^{1 - y_i} \right\} \\
&= \sum_{i=1}^n \log \left\{ \theta^{y_i} (1 - \theta)^{1 - y_i} \right\} \\
&= \sum_{i=1}^n \left\{ y_i \log \theta + (1 - y_i) \log(1 - \theta) \right\} \\
&= n\bar{y} \log \theta + (n - n\bar{y}) \log(1 - \theta).
\end{aligned}$$

To find the MLE, we solve $\ell'(\theta) = 0$.

$$\begin{aligned}
\frac{n\bar{y}}{\theta} - \frac{n - n\bar{y}}{1 - \theta} &= 0 \\
n\bar{y}(1 - \theta) - (n - n\bar{y})\theta &= 0 \\
n\bar{y} - n\theta &= 0 \\
\theta &= \bar{y},
\end{aligned}$$

so the MLE is indeed $\hat{\theta} = \bar{y}$. If we want to be scrupulous here, we can also verify that indeed,

$$\ell''(\theta) = \sum_{i=1}^{n} \left\{ -\frac{y_i}{\theta^2} - \frac{1-y_i}{(1-\theta)^2} \right\} = -\frac{1}{\theta^2(1-\theta)^2} \sum_{i=1}^{n} \left\{ y_i(1-\theta)^2 + (1-y_i)\theta^2 \right\} \le 0,$$

where this inequality follows from the fact that

$$y(1-\theta)^2 + (1-y)\theta^2 = \begin{cases} (1-\theta)^2 & \text{if } y = 1 \\ \theta^2 & \text{if } y = 0. \end{cases},$$

which is non-negative in either case.

(c) By linearity of expectation, we have

$$\mathbb{E}[m] = \mathbb{E}\left[\sum_{i=1}^{n} y_i \right] = \sum_{i=1}^{n} \mathbb{E}[y_i] = \theta n,$$

and by independence, we have

$$\mathrm{Var}(m) = \mathrm{Var}\left(\sum_{i=1}^{n} y_i \right) = \sum_{i=1}^{n} \mathrm{Var}(y_i) = \theta(1-\theta)n.$$

(d) From (b), we know that the log-likelihood is a linear combination of $\log\theta$ and $\log(1-\theta)$, and so we know that the conjugate prior must satisfy

$$\pi(\theta) \propto \theta^{a-1}(1-\theta)^{b-1},$$

for some parameters a, b, where we need $a, b > 0$ for the prior density to be normalisable. Hence the prior is of the same functional form as a Beta density, which forces $\pi(\theta) = \mathrm{Beta}(\theta; a, b)$.

(e) Following the hint, we have $\mathrm{d}t = \mathrm{d}y$ and $\mathrm{d}x = t\,\mathrm{d}s$, and so

$$\begin{aligned}
\Gamma(a)\Gamma(b) &= \int_{x=0}^{\infty} x^{a-1} \left(\int_{y=0}^{\infty} y^{b-1} \exp(-(x+y))\,\mathrm{d}y \right) \mathrm{d}x \\
&= \int_{x=0}^{\infty} x^{a-1} \left(\int_{t=x}^{\infty} (t-x)^{b-1} \exp(-t)\,\mathrm{d}t \right) \mathrm{d}x \\
&= \int_{t=0}^{\infty} \exp(-t) \left(\int_{x=0}^{t} x^{a-1}(t-x)^{b-1}\,\mathrm{d}x \right) \mathrm{d}t \\
&= \int_{t=0}^{\infty} \exp(-t) \left(\int_{s=0}^{1} (ts)^{a-1}(t-ts)^{b-1} t\,\mathrm{d}s \right) \mathrm{d}t \\
&= \underbrace{\int_{t=0}^{\infty} t^{a+b-1} \exp(-t)\,\mathrm{d}t}_{\Gamma(a+b)} \times \int_{s=0}^{1} s^{a-1}(1-s)^{b-1}\,\mathrm{d}s,
\end{aligned}$$

and so

$$\int_0^1 s^{a-1}(1-s)^{b-1}\,\mathrm{d}s = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}.$$

(f) We have

$$\begin{aligned}\log \pi(\theta \mid m) &= \log \pi(\theta) + \log \pi(m \mid \theta) + \text{constant}\\ &= (a-1)\log\theta + (b-1)\log(1-\theta) + n\bar{y}\log\theta + (n-n\bar{y})\log(1-\theta) + \text{constant}\\ &= (a+n\bar{y}-1)\log\theta + (b+n-n\bar{y}-1)\log(1-\theta),\end{aligned}$$

which forces, by functional form,

$$\pi(\theta \mid m) = \text{Beta}(\theta; a+n\bar{y}, b+n-n\bar{y}).$$

Thus, by the formula of the mean of a Beta distribution,

$$\mathbb{E}[\theta \mid m] = \frac{a+n\bar{y}}{a+n\bar{y}b+n-n\bar{y}} = \frac{a+n\bar{y}}{a+b+n}.$$

This can be written as

$$\begin{aligned}\mathbb{E}[\theta \mid m] &= \left\{\frac{a+b}{a+b+n}\right\}\frac{a}{a+b} + \left\{\frac{n}{a+b+n}\right\}\bar{y}\\ &= \underbrace{\left\{\frac{a+b}{a+b+n}\right\}}_{\lambda}\frac{a}{a+b} + \underbrace{\left\{1-\frac{a+b}{a+b+n}\right\}}_{1-\lambda}\bar{y}\\ &= \lambda\mathbb{E}[\theta] + (1-\theta)\bar{y}.\end{aligned}$$

We note that as $n \to \infty$, $\lambda \to 0$, and so the distance between the posterior mean and the MLE decreases. The prior is "washed out" as the number of data points increases.

Exercise 12 (Based on Nils Lid Hjort's exercises, # 13)**.** The Beta-binomial model, with a Beta distribution for the binomial probability parameter, is on the 'Nice List' where the Bayesian machinery works particularly well: Prior elicitation is easy, as is the updating mechanism. This exercise concerns the generalisation to the Dirichlet-multinomial model, which is certainly also on the Nice List and indeed in broad and frequent use for a number of statistical analyses.

(a) Let $(y_1, \ldots, y_m)$ be the count vector associated with n independent experiments having m different outcomes $A_1, \ldots, A_m$. In other words, y_j is the number of events of type A_j, for $j = 1, \ldots, m$. Show that if the vector of $\mathbb{P}(A_j) = p_j$ is constant across the n independent experiments, then the probability distribution governing the count data is

$$f(y_1, \ldots, y_m) = \frac{n!}{y_1! \cdots y_m!} p_1^{y_1} \cdots p_m^{y_m},$$

for $y_1 \geq 0, \ldots, y_m \geq 0, y_1 + \cdots + y_m = n$. This is the multinomial model. Explain how it generalises the binomial model.

(b) Show that

$$\mathbb{E}\, Y_j = np_j, \quad \operatorname{Var} Y_j = np_j(1-p_j), \quad \operatorname{cov}(Y_j, Y_k) = -np_jp_k \text{ for } j \neq k.$$

(c) Now define the Dirichlet distribution over m cells with parameters $(a_1, \ldots, a_m)$ as having probability density

$$\pi(p_1, \ldots, p_{m-1}) = \frac{\Gamma(a_1 + \cdots + a_m)}{\Gamma(a_1) \cdots \Gamma(a_m)} p_1^{a_1 - 1} \cdots p_{m-1}^{a_{m-1}-1} (1 - p_1 - \cdots - p_{m-1})^{a_m - 1},$$

over the simplex where each $p_j \geq 0$ and $p_1 + \cdots + p_{m-1} \leq 1$. Of course we may choose to write this as

$$\pi(p_1, \ldots, p_{m-1}) \propto p_1^{a_1-1} \cdots p_{m-1}^{a_{m-1}-1} p_m^{a_m - 1},$$

with $p_m = 1 - p_1 - \cdots - p_{m-1}$; the point is however that there are only $m - 1$ unknown parameters in the model as one knows the mth once one learns the values of the other $m-1$. Show that the marginals are Beta distributed,

$$p_j \sim \mathrm{Beta}(a_j, a - a_j) \quad \text{where } a = a_1 + \cdots + a_m.$$

(d) Infer from this that

$$\mathbb{E}\, p_j = p_{0,j} \qquad \mathrm{Var}\, p_j = \frac{1}{a+1} p_{0,j}(1 - p_{0,j}),$$

in terms of $a_j = a p_{0,j}$. Show also that

$$\mathrm{cov}(p_j, p_k) = -\frac{1}{a+1} p_{0,j} p_{0,k} \quad \text{for } j \neq k.$$

For the 'flat Dirichlet', with parameters $(1, \ldots, 1)$ and prior density $(m-1)!$ over the simplex, find the means, variances, covariances.

(e) Now for the basic Bayesian updating result. When $(p_1, \ldots, p_m)$ has a $\mathrm{Dir}(a_1, \ldots, a_m)$ prior, then, given the multinomial data $y = (y_1, \ldots, y_m)$, show that

$$(p_1, \ldots, p_m) \mid y \sim \mathrm{Dir}(a_1 + y_1, \ldots, a_m + y_m).$$

Give formulae for the posterior means, variances, and covariances. In particular, explain why

$$\hat{p}_j = \frac{a_j + y_j}{a + n}$$

is a natural Bayes estimate of the unknown p_j. Also find an expression for the posterior standard deviation of the p_j.

(f) In order to carry out easy and flexible Bayesian inference for $p_1, \ldots, p_m$ given observed counts $y_1, \ldots, y_m$, one needs a recipe for simulating from the Dirichlet distribution. One such is as follows: Let $X_1, \ldots, X_m$ be independent with $X_j \sim \mathrm{Gamma}(a_j, 1)$ for $j = 1, \ldots, m$. Then the ratios

$$Z_1 = \frac{X_1}{X_1 + \cdots + X_m}, \ldots, Z_m = \frac{X_m}{X_1 + \cdots + X_m}$$

are in fact $\mathrm{Dir}(a_1, \ldots, a_m)$. Try to show this from the transformation law for probability distributions: If X has density $f(x)$, and $Z = h(X)$ is a one-to-one transformation with inverse $X = h^{-1}(Z)$, then the density of Z is

$$g(z) = f(h^{-1}(z)) \left| \frac{\partial h^{-1}(z)}{\partial z} \right|$$

(featuring the determinant of the Jacobian of the transformation). Use in fact this theorem to find the joint distribution of $(Z_1, \dots, Z_{m-1}, S)$, where $S = X_1 + \cdots + X_m$ (one discovers that the Dirichlet vector of Z_j is independent of their sum S).

(g) The Dirichlet distribution has a nice 'collapsibility' property: If say $(p_1, \dots, p_8)$ is $\mathrm{Dir}(a_1, \dots, a_8)$, show that then the collapsed vector $(p_1 + p_2, p_3 + p_4 + p_5, p_6, p_7 + p_8)$ is $\mathrm{Dir}(a_1 + a_2, a_3 + a_4 + a_5, a_6, a_7 + a_8)$.

Solution.

(a) Setting $m = 2$, we obtain

$$f(y_1, y_2) = \frac{n!}{y_1! y_2!} p_1^{y_1} p_2^{y_2},$$

where $y_1 + y_2 = n$ and $p_1 + p_2 = 1$. In other words, we can parametrise this as

$$f(y_1) = \frac{n!}{y_1!(n - y_1)!} p_1^{y_1} (1 - p_1)^{n - y_1} = \binom{n}{y_1} p_1^{y_1} (1 - p_1)^{n - y_1},$$

which is the binomial model.

(b) Just like the binomial distribution, we can write $Y = (Y_1, \dots, Y_m)$ as the sum of the results of n independent experiments, say $Y = X_1 + \cdots + X_n$, where each $X_i \in \{0, 1\}^m$ has exactly one nonzero component. Note that

$$\mathbb{E}[X_i] = (p_1, \dots, p_n),$$

and therefore the results for $\mathbb{E}Y_j$ and $\mathrm{Var}Y_j$ follow directly from linearity of expectation and linearity of variance for independent variables.

Alternatively, we can prove the result directly. Without loss of generality, let $j = 1$. Then we have

$$\mathbb{E}\, Y_1 = \sum_{y_1, \dots, y_m} y_1 \frac{n!}{y_1! \cdots y_m!} p_1^{y_1} \cdots p_m^{y_m} = n p_1 \sum_{\substack{y_1, \dots, y_m \\ y_1 \geq 1}} \frac{(n-1)!}{(y_1 - 1)! y_2! \cdots y_m!} p_1^{y_1 - 1} p_2^{y_2} \cdots p_m^{y_m} = n p_1.$$

Similarly,

$$\begin{aligned} \mathbb{E}[Y_1(Y_1 - 1)] &= \sum_{y_1, \dots, y_m} y_1 (y_1 - 1) \frac{n!}{y_1! \cdots y_m!} p_1^{y_1} \cdots p_m^{y_m} \\ &= n(n-1) p_1^2 \sum_{\substack{y_1, \dots, y_m \\ y_1 \geq 2}} \frac{(n-2)!}{(y_1 - 2)! y_2! \cdots y_n!} p_1^{y_1 - 2} p_2^{y_2} \cdots p_m^{y_m} = n(n-1) p_1^2. \end{aligned}$$

Hence

$$\mathrm{Var}\, Y_1 = \mathbb{E}[Y_1^2] - \mathbb{E}[Y_1]^2 = \mathbb{E}[Y_1(Y_1 - 1)] + \mathbb{E}[Y_1] - \mathbb{E}[Y_1]^2 = n(n-1)p_1^2 + np_1 - (np_1)^2 = np_1(1 - p_1).$$

Using the same trick as above for the covariance, we obtain $\mathbb{E}[Y_j Y_k] = n(n-1) p_j p_k$, and thus

$$\mathrm{cov}(Y_j, Y_k) = \mathbb{E}[Y_j Y_k] - \mathbb{E}[Y_j]\mathbb{E}[Y_k] = n(n-1) p_j p_k - (np_j)(np_k) = -n p_j p_k.$$

(c) Again, let $j = 1$ without loss of generality. Marginalising, we have

$$\begin{aligned}\pi(p_1) &\propto \int p_1^{a_1-1} \cdots p_{m-1}^{a_{m-1}-1}(1-p_1-\cdots-p_{m-1})^{a_m-1}\,\mathrm{d}p_2\cdots\mathrm{d}p_{m-1}\\ &= \int p_1^{a_1-1}\cdots p_{m-2}^{a_{m-2}-1}\left\{\int_0^{1-p_1-\cdots-p_{m-2}} p_{m-1}^{a_{m-1}-1}(1-p_1-\cdots-p_{m-1})^{a_m-1}\,\mathrm{d}p_{m-1}\right\}\mathrm{d}p_2\cdots\mathrm{d}p_{m-2}.\end{aligned}$$

We do the inner integral first, working our way outwards. The trick is to make the substitution

$$p_{m-1} = (1-p_1-\cdots-p_{m-2})u,$$

so that the limits of integration turn into 0 and 1. The inner integral is, as a function of p_1.

$$\begin{aligned}&\int_0^1 (1-p_1-\cdots-p_{m-2})^{a_{m-1}-1}u^{a_{m-1}-1}(1-u)^{a_m-1}(1-p_1-\cdots-p_{m-2})^{a_m-1}(1-p_1-\cdots-p_{m-2})\,\mathrm{d}u\\ &\qquad\propto (1-p_1-\cdots-p_{m-2})^{a_{m-1}+a_m-1}.\end{aligned}$$

Repeating this argument for $p_{m-2}, p_{m-3}, \ldots, p_2$, we end up with

$$\pi(p_1) \propto p_1^{a_1-1}(1-p_1)^{a_2+\cdots+a_m-1} = p_1^{a_1-1}(1-p_1)^{a-a_1-1},$$

which, by functional form, forces $p_1 \sim \mathrm{Beta}(a_1, a-a_1)$.

(d) The mean and variance follow directly from the formulae for the mean and variance of the Beta distribution. For the covariance, we let $j = 1$ and $k = 2$ without loss of generality. Write $a = a_1 + \cdots + a_m$. Then, normalising, we have

$$\begin{aligned}\mathbb{E}[p_1p_2] &= \frac{\Gamma(a)}{\Gamma(a_1)\cdots\Gamma(a_m)}\int p_1^{a_1}p_2^{a_2}p_3^{a_3-1}\cdots p_{m-1}^{a_{m-1}-1}(1-p_1-\cdots-p_{m-1})^{a_m-1}\,\mathrm{d}p_1\cdots\mathrm{d}p_{m-1}\\ &= \frac{\Gamma(a)}{\Gamma(a_1)\Gamma(a_1)}\frac{\Gamma(a_1+1)\Gamma(a_2+1)}{\Gamma(a+2)}\int\{\text{normalised density}\}\,\mathrm{d}p_1\cdots\mathrm{d}p_{m-1}\\ &= \frac{a_1a_2}{a(a+1)}.\end{aligned}$$

Hence

$$\begin{aligned}\mathrm{cov}(p_1,p_2) = \mathbb{E}[p_1p_2]-\mathbb{E}[p_1]\mathbb{E}[p_2] &= \frac{a_1a_2}{a(a+1)} - \frac{a_1a_2}{a^2} = a_1a_2\frac{a^2-a(a+1)}{a(a+1)}\\ &= -\frac{1}{a+1}\frac{a_1}{a}\frac{a_2}{a} = -\frac{1}{a+1}p_{0,1}p_{0,2}.\end{aligned}$$

For the flat Dirichlet, we plug in $a_1 = \cdots = a_m = 1$ in the formulae derived and obtain

$$\mathbb{E}\,p_j = \frac{1}{m}, \quad \mathrm{Var}\,p_j = \frac{1}{2}\frac{1}{m}\left(1-\frac{1}{m}\right), \quad \mathrm{cov}(p_j,p_k) = -\frac{1}{2}\frac{1}{m^2}\text{ for } j \neq k.$$

(e) Using the normalisation trick, we have

$$\begin{aligned}\pi(p_1,\ldots,p_m \mid y) \propto \pi(y \mid p_1,\ldots,p_m)\pi(p_1,\ldots,p_m) &\propto p_1^{y_1}\cdots p_m^{y_m} \times p_1^{a_1-1}\cdots p_m^{a_m-1}\\ &= p_1^{a_1+y_1-1}\cdots p_m^{a_m+y_m-1},\end{aligned}$$

so by functional form, $(p_1, \dots, p_m) \mid y \sim \text{Dir}(a_1 + y_1, \dots, a_m + y_m)$.

Using the formulae for mean, variance and covariance for the Dirichlet distribution (derived earlier), we have

$$\mathbb{E}[p_j \mid y] = \frac{a_j + y_j}{a + n}, \quad \text{Var}[p_j \mid y] = \frac{1}{a_j + y_j + 1} \frac{a_j + y_j}{a + n} \left(1 - \frac{a_j + y_j}{a + n}\right),$$

$$\text{cov}(p_j, p_k \mid y) = -\frac{1}{a + n + 1} \frac{(a_j + y_j)(a_k + y_k)}{(a + n)^2} \text{ for } j \neq k.$$

Recalling that the Bayes estimator when using the quadratic loss is the posterior mean, we see that $\hat{p}_j = (a_j + y_j)/(a + n)$ is a natural Bayes estimator for p_j.

The posterior standard deviation is given by

$$\text{sd}(p_j \mid y) = \left\{ \frac{1}{a_j + y_j + 1} \hat{p}_j (1 - \hat{p}_j) \right\}^{1/2}.$$

(f) Write $z_{1:m-1} = (z_1, \dots, z_{m-1})$. By the transformation law, we have

$$\begin{aligned}
\pi(z_1, \dots, z_{m-1}, s) &= \pi(x_1(z_{1:m-1}, s), \dots, x_m(z_{1:m-1}, s)) \left| \frac{\partial(x_1, \dots, x_m)}{\partial(z_1, \dots, z_{m-1}, s)} \right| \\
&= \pi_{x_{1:m}}(sz_1, \dots, sz_{m-1}, s(1 - z_1 - \dots - z_{m-1})) \left| \frac{\partial(x_1, \dots, x_m)}{\partial(z_1, \dots, z_{m-1}, s)} \right|
\end{aligned}$$

Now, the Jacobian is

$$\left| \frac{\partial(x_1, \dots, x_m)}{\partial(z_1, \dots, z_{m-1}, s)} \right| = \begin{vmatrix} s & 0 & \cdots & 0 & -s \\ 0 & s & \cdots & 0 & -s \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & s & -s \\ z_1 & z_2 & \cdots & z_{m-1} & 1 - z_1 - \dots - z_{m-1} \end{vmatrix}.$$

Now, let D_j be the determinant of the $(m - j + 1 : m) \times (m - j + 1 : m)$ submatrix of the above. That is, D_m is the full determinant, $D_1 = 1 - z_1 - \dots - z_{m-1}$, and so on. Then we have the recurrence relation

$$\begin{aligned}
D_m &= sD_{m-1} + (-1)^{m+1}(-s)z_1 \times (-1)^m s^{m-2} = sD_{m-1} + s^{m-1} z_1 \\
&= s\{sD_{m-2} + s^{m-2} z_2\} + s^{m-1} z_1 \\
&= \cdots \\
&= s^{m-1} D_1 + s^{m-1}(z_1 + \dots + z_{m-1}) \\
&= s^{m-1}(1 - z_1 - \dots - z_{m-1}) + s^{m-1}(z_1 + \dots + z_{m-1}) \\
&= s^{m-1}.
\end{aligned}$$

Hence we obtain

$$\begin{aligned}
\pi(z_1, \dots, z_{m-1}, s) &\propto (sz_1)^{a_1 - 1} \exp\{-sz_1\} \cdots (sz_{m-1})^{a_{m-1} - 1} \exp\{-sz_{m-1}\} \\
&\quad \times (s(1 - z_1 - \dots - z_{m-1}))^{a_m - 1} \exp(-s(1 - z_1 - \dots - z_{m-1})) \times s^{m-1} \\
&= s^{a-1} \exp(-s) \times z_1^{a_1 - 1} \cdots z_{m-1}^{a_{m-1} - 1} (1 - z_1 - \dots - z_{m-1})^{a_m - 1},
\end{aligned}$$

where $a = a_1 + \cdots + a_m$. Hence, by functional form, and since the domains of S and $(Z_1, \ldots, Z_{m-1})$ are independent, we see that S and $(Z_1, \ldots Z_{m-1})$ are independent, and that

$$S \sim \text{Gamma}(a, 1), \quad (Z_1, \ldots, Z_{m-1}) \sim \text{Dir}(a_1, \ldots, a_m).$$

(g) It suffices to show that if $(p_1, \ldots, p_{m-1}) \sim \text{Dir}(a_1, \ldots, a_m)$, then $(p_1, \ldots, p_{m-2} + p_{m-1}) \sim \text{Dir}(a_1, \ldots, a_{m-1} + a_m)$. Indeed, then the rest follows by reordering the variables if necessary and applying mathematical induction.

Now, write $p = (p_1, \ldots, p_{m-1})$ and let $q = (q_1, \ldots, q_{m-1})$, where

$$\begin{aligned} q_1 &= p_1, \\ &\vdots \\ q_{m-3} &= p_{m-3}, \\ q_{m-2} &= p_{m-2} + p_{m-1}, \\ q_{m-1} &= p_{m-1}. \end{aligned}$$

We want to show that $(q_1, \ldots, q_{m-2})$ is $\text{Dir}(a_1, \ldots, a_{m-3}, a_{m-2} + a_{m-1})$. Now,

$$\pi(q) = \pi(p(q)) \propto q_1^{a_1 - 1} \cdots q_{m-3}^{a_{m-3}-1} (q_{m-2} - q_{m-1})^{a_{m-2}-1} q_{m-1}^{a_{m-1}-1} (1 - q_1 - \cdots - q_{m-2})^{a_m - 1},$$

and so marginalising out q_{m-1}, we have

$$\pi(q_1, \ldots q_{m-2}) \propto q_1^{a_1-1} \cdots q_{m-3}^{a_{m-3}-1} (1 - q_1 - \cdots - q_{m-2})^{a_m - 1} \int_0^{q_{m-2}} (q_{m-2} - q_{m-1})^{a_{m-2}-1} q_{m-1}^{a_{m-1}-1} \, \mathrm{d}q_{m-1}.$$

Therefore, we are done if we can show that

$$\int_0^{q_{m-2}} (q_{m-2} - q_{m-1})^{a_{m-2}-1} q_{m-1}^{a_{m-1}-1} \, \mathrm{d}q_{m-1} \propto q_{m-2}^{a_{m-2}+a_{m-1}-1}.$$

Write $r = q_{m-2}, s = q_{m-1}, \alpha = a_{m-2}, \beta = a_{m-1}$ to simplify notation. We need to show that

$$\int_0^r (r - s)^{\alpha - 1} s^{\beta - 1} \, \mathrm{d}r \propto r^{\alpha + \beta - 1}.$$

Using the substitution $s = ur$ (again, to shift the integration limits to 0 and 1), we obtain

$$\int_0^1 (r - ur)^{\alpha-1} u^{\beta-1} r^{\beta-1} r \, \mathrm{d}u = r^{\alpha+\beta-1} \int_0^1 (1-u)^{\alpha-1} u^{\beta-1} \, \mathrm{d}u \propto r^{\alpha+\beta-1},$$

and we are done.

References

Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.