

Record answers in the blue books. This exam is closed book/notes.

1. Each item in a large grocery store has two prices: the price indicated on the shelf where the item is on display, and the price which goes on your bill when you buy the item – that is, the price given by the computer scanner. For the most part the shelf price and the scan price are equal, but errors inevitably enter the system. Let $\theta = (\alpha, \beta, \gamma)$ be the vector of unknown proportions which characterize pricing errors in a given store at a given point in time. Here α is the proportion of items with a pricing error which favors the store, β is the proportion of items with an error which favors the customer, and $\gamma = 1 - \alpha - \beta$ is the proportion of items priced correctly. A government regulator is interested in testing the null hypothesis $H_0 : \alpha \leq \beta$ versus the alternative hypothesis $H_a : \alpha > \beta$. An experiment is designed to obtain a random sample of n items from the store and to record the pricing status of each item. Expecting that errors are relatively rare, the regulator takes a Dirichlet(1, 1, 98) prior distribution for $\theta = (\alpha, \beta, \gamma)$. Symmetry ensures that $P(H_0) = P(\alpha > \beta)$ is the same as $P(H_a) = P(\alpha \leq \beta)$ for this prior. That is, both probabilities equal 1/2 and the regulator is not prejudiced for or against the store.
 - (a) Sample counts (12, 8, 280) arise after price-checking $n = 300$ items from the store. What is the joint posterior distribution of $\theta = (\alpha, \beta, \gamma)$?
 - (b) Your computer is able to realize independent Gamma-distributed random variables. How can you use this facility to approximate $P(H_0|\text{data})$?
 - (c) What is the marginal posterior distribution of γ ? Comment on how the opinion of the regulator has changed with regard to the overall proportion of errors in the store.
2. In hypothesis testing problems (such as the one in the last problem), it is convenient to report the posterior odds $O(\text{post}) = P(H_0|D)/P(H_a|D)$ for data D , and to see how the odds have changed from the prior odds $O(\text{prior}) = P(H_0)/P(H_a)$. The Bayes factor BF is that amount which relates these two quantities through the relation

$$O(\text{post}) = BF \times O(\text{prior}).$$

- (a) Show that the Bayes factor is $BF = P(D|H_0)/P(D|H_a)$.
- (b) Describe an action space and a loss function which make the hypothesis testing problem a decision theory problem. In decision theory terms, why should the Bayesian base inference on the posterior odds $O(\text{post})$?

3. The i th amino acid x_i in a length- m protein sequence $x = (x_1, \dots, x_m)$ arises from the set of five possibilities $\mathcal{A} = \{C, D, E, F, G\}$. Unknown is the position b , somewhere in $\{1, 2, \dots, m-2\}$, which indicates the left-most site in a binding element comprised of 3 sites $b, b+1, b+2$. Conditionally on some parameters in θ and on the location b , we model x by asserting that x_i are independent multinomials over \mathcal{A} . More specifically, if site i is not part of the binding element (i.e., it is in the background), then x_i takes values in \mathcal{A} with probabilities $\alpha = (\alpha_1, \dots, \alpha_5)$. There is also a site-specific probability vector for each site within the binding element. So x_b arises from probability vector $\beta = (\beta_1, \dots, \beta_5)$, the second amino acid x_{b+1} is governed by $\gamma = (\gamma_1, \dots, \gamma_5)$ and the third amino acid x_{b+2} is governed by $\delta = (\delta_1, \dots, \delta_5)$. In total, the parameter set θ is comprised of four probability vectors $\theta = (\alpha, \beta, \gamma, \delta)$. The unknown state is $s = (b, \theta)$, and x is the observed data.

We consider the following prior distribution for s . First, b is uniform over $\{1, \dots, m-2\}$. Second, the four vectors comprising θ have independent Dirichlet prior distributions:

$$\begin{aligned} \alpha &\sim \text{Dirichlet}(2, 1, 1, 1, 1) & \beta &\sim \text{Dirichlet}(1, 1, 1, 1, 10) \\ \gamma &\sim \text{Dirichlet}(1, 1, 10, 1, 1) & \delta &\sim \text{Dirichlet}(1, 1, 10, 1, 1) \end{aligned}$$

and θ is independent of b .

- (a) A priori, what do we expect the binding element (x_b, x_{b+1}, x_{b+2}) to look like?
 (b) We observe the following sequence x :

site i	1	2	3	4	5	6	7	8	9	10
amino acid x_i	G	C	E	C	D	D	G	F	E	C

If $b = 1$, characterize the full conditional distribution $p(\alpha|\text{everything else})$. Similarly characterize the full conditionals of β , γ and δ .

- (c) For the same data as above, and for some fixed θ , calculate the ratio

$$r = \frac{P(b = 2|x, \theta)}{P(b = 1|x, \theta)}.$$

- (d) Suggest a Markov chain Monte Carlo algorithm for sampling states $s = (b, \theta)$ from $p(s|x)$.

FYI: Recall that if $u = (u_1, \dots, u_k)$ is a probability vector having a Dirichlet(a_1, \dots, a_k) distribution, then the density function for u is

$$\frac{\Gamma(\sum_i a_i)}{\prod_i \Gamma(a_i)} \prod_i u_i^{a_i-1}$$

Note also that if z takes one of k values with probabilities in u , then $p(z|u) = \prod_i u_i^{y_i}$ where y_i is a binary code for z . That is, $y_i = 1$ when $i = z$ and is 0 otherwise.