STATS 551 Homework 2

Multi-Parameter Models & Hierarchical Models

Responsible GSI: Wayne Wang

Dirichlet-Multinomial model (4×10 points).

1. The Multinomial model with a Dirichlet prior is a generalization of the Bernoulli model and Beta prior. The Dirichlet distribution for K outcomes is the exponential family distribution on the $K-1$ dimensional probability simplex given by
$$\pi_{\boldsymbol{\alpha}}(\boldsymbol{\theta}) = \frac{\Gamma(\sum_{k=1}^{K} \alpha_j)}{\prod_{j=1}^{K} \Gamma(\alpha_j)} \prod_{j=1}^{K} \theta_j^{\alpha_j - 1},$$
where $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_K)^T \in \mathbb{R}_+^K$ is a non-negative vector of scaling coefficients, which are known parameters. The multinomial distribution, denoted as Multinomial$(n, \boldsymbol{\theta})$, is a discrete distribution over K dimensional non-negative integer vectors $\mathbf{X} \in \mathbb{Z}_+^K$ and $\sum_{j=1}^{K} X_j = n$. Assuming this Dirichlet-Multinomial model, what is the posterior distribution of $\boldsymbol{\theta}$? What is the posterior mean?

2. Derive the posterior predictive density function of a future observation of $\mathbf{X}$ conditioned on the current observations.

3. Define $Y = \frac{\theta_1}{\theta_1 + \theta_2}$. Write down the marginal posterior distribution for Y. Show that this distribution is identical to the posterior distribution for Y obtained by treating X_1 as an observation from the binomial distribution with probability Y and sample size $X_1 + X_2$, ignoring the data $X_3, \dots, X_K$.

4. On September 25, 1988, the evening of a presidential campaign debate, ABC News conducted a survey of registered voters in the United States; 639 persons were polled before the debate, and 639 different persons were polled after. The results are displayed in Figure 1. Assume the surveys are independent simple random samples from the population of registered voters. Model the data with two different multinomial distributions. For $j = 1, 2$, let θ_j be the proportion

of voters who preferred Bush, out of those who had a preference for either Bush or Dukakis at the time of survey j. Plot a histogram of the posterior density for $\theta_2 - \theta_1$. What is the posterior probability that there was a shift toward Bush?

Survey	Bush	Dukakis	No opinion/other	Total
pre-debate	294	307	38	639
post-debate	288	332	19	639

Figure 1: Number of respondents in each preference category from ABC News pre- and post-debate surveys in 1988.

Guideline for Submission: For questions require coding submit a PDF or an HTML file generated by R markdown (or jupyter notebook) with annotated code and results; for other questions submit either a scanned version of the written solution or a PDF generated by R markdown or TeX or similar tools.

Analysis of proportions (6×10 points). A survey was done of bicycle and other vehicular traffic in the neighborhood of the campus of the University of California, Berkeley, in the spring of 1993. Sixty city blocks were selected at random; each block was observed for one hour, and the numbers of bicycles and other vehicles traveling along that block were recorded. The sampling was stratified into six types of city blocks: busy, fairly busy, and residential streets, with and without bike routes, with ten blocks measured in each stratum. Figure 2 displays the number of bicycles and other vehicles recorded in the study. For this problem, restrict your attention to the first two rows of the table: residential streets labeled as 'bike routes', which we will use to illustrate this computational exercise.

Type of street	Bike route?	Counts of bicycles/other vehicles
Residential	yes	16/58, 9/90, 10/48, 13/57, 19/103, 20/57, 18/86, 17/112, 35/273, 55/64
Residential	no	12/113, 1/18, 2/14, 4/44, 9/208, 7/67, 9/29, 8/154
Fairly busy	yes	8/29, 35/415, 31/425, 19/42, 38/180, 47/675, 44/620, 44/437, 29/47, 18/462
Fairly busy	no	10/557, 43/1258, 5/499, 14/601, 58/1163, 15/700, 0/90, 47/1093, 51/1459, 32/1086
Busy	yes	60/1545, 51/1499, 58/1598, 59/503, 53/407, 68/1494, 68/1558, 60/1706, 71/476, 63/752
Busy	no	8/1248, 9/1246, 6/1596, 9/1765, 19/1290, 61/2498, 31/2346, 75/3101, 14/1918, 25/2318

Figure 2: Counts of bicycles and other vehicles in one hour in each of 10 city blocks in each of six categories. (The data for two of the residential blocks were lost.) For example, the first block had 16 bicycles and 58 other vehicles, the second had 9 bicycles and 90 other vehicles, and so on. Streets were classified as 'residential', 'fairly busy', or 'busy' before the data were gathered.

1. Set up a model for the data in the table so that, for $j = 1, \ldots, 10$, the observed number of bicycles at location j is binomial with unknown probability θ_j and sample size equal to the total number of vehicles (bicycles included) in that block. The parameter θ_j can be interpreted as the underlying or 'true' proportion of traffic at location j that is bicycles. Assign a beta population

distribution for the parameters θ_j and a noninformative hyperprior distribution as in the rat tumor example of Section 5.3 of BDA. Write down the joint posterior distribution.

2. Compute the marginal posterior density of the hyperparameters and draw simulations from the joint posterior distribution of the parameters and hyperparameters.

3. Compare the posterior distributions of the parameters θ_j to the raw proportions, (number of bicycles / total number of vehicles) in location j. How do the inferences from the posterior distribution differ from the raw proportions?

4. Give a 95% posterior interval for the average underlying proportion of traffic that is bicycles.

5. A new city block is sampled at random and is a residential street with a bike route. In an hour of observation, 100 vehicles of all kinds go by. Give a 95% posterior interval for the number of those vehicles that are bicycles. Discuss how much you trust this interval in application.

6. Was the beta distribution for the θ_j's reasonable?

Guideline for Submission: For questions require coding submit a PDF or an HTML file generated by R markdown (or jupyter notebook) with annotated code and results; for other questions submit either a scanned version of the written solution or a PDF generated by R markdown or TeX or similar tools.

Optional Reading. Read one of the following papers and post your summary and thoughts on Canvas. Bonus points up to 5 will be rewarded.

1. The selection of prior distributions by formal rules, Kass, R. E, and Wasserman, L. (1996), Journal of the American Statistical Association 91, 1343-1370.

2. Parameterization and Bayesian modeling, Gelman, A. (2004), Journal of the American Statistical Association 99, 537-545.