STAT 207 Intermediate Bayesian Statistical Modeling

Multinomial and multivariate normal models

Zehang (Richard) Li

Adapted from Bruno Sansó’s slides

The multinomial distribution

- Consider a random variable taking one of the *k* possible outcomes.
- Count the number of occurrences of each type of outcome for *n* trials.
- The vector of such counts *y* has the density

$$p(y|\theta) \propto \prod_{j=1}^{K} \theta_j^{y_j}$$

  with $\sum_{j=1}^{K} \theta_j = 1$.
- The multinomial distribution is a generalization of the binomial distribution.

The Dirichlet distribution

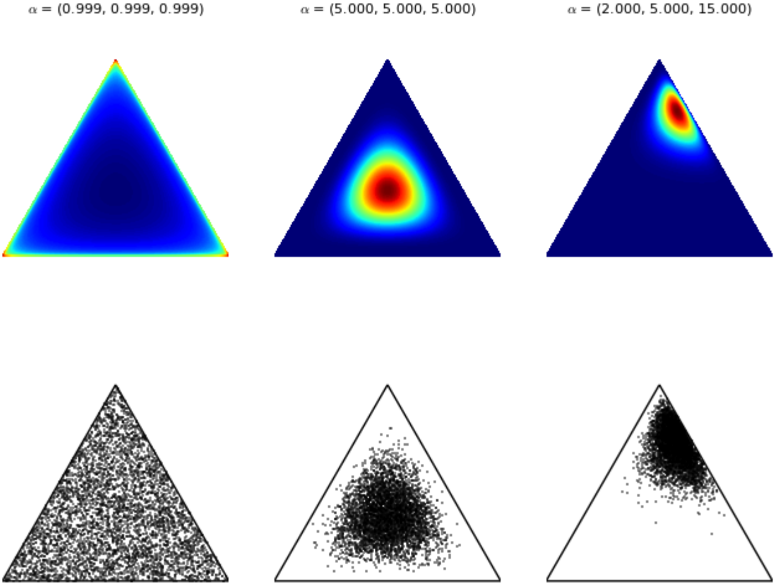
- A generalization of the beta distribution to K components is the Dirichlet distribution.

  $$p(\theta|\alpha) \propto \prod_{j=1}^{K} \theta_j^{a_j-1}$$

  with $\sum_{j=1}^{K} \theta_j = 1$ and $\alpha_j > 0, j = 1, ..., K$.
- This distribution is a conjugate prior for the multinomial likelihood. The corresponding posterior distribution is

  $$p(\theta|y) \propto \prod_{j=1}^{K} \theta_j^{y_j+\alpha_j-1}.$$

The Dirichlet distribution



α = (0.999, 0.999, 0.999)

α = (5.000, 5.000, 5.000)

α = (2.000, 5.000, 15.000)

Sampling from the Dirichlet distribution

- There is an important relationship between Dirichlet and gamma distributions. Let

  $$Z_j \sim Gamma(\alpha_j, \beta)$$

  for $j = 1, ..., K$ independently, then the vector θ with

  $$\theta_j = \frac{Z_j}{\sum_j Z_j}, j = 1, ..., K$$

  follow a Dirichlet distribution with parameter α.
- The conditional distribution of a sub-vector of θ given the remaining elements of θ is also a Dirichlet.

Slovenia opinion poll example

- In 1990, a plebiscite was held in Slovenia at which the adult citizens voted on the question of independence.
- A Slovenian public opinion survey had been conducted that included several questions concerning likely plebiscite attendance and voting. In that survey, 2074 Slovenians were asked:
  1. Are you in favor of independence?
  2. Will you attend the plebiscite?
- The rules of the plebiscite were such that only those attending and voting ‘yes’ would be counted as being in favor of independence.

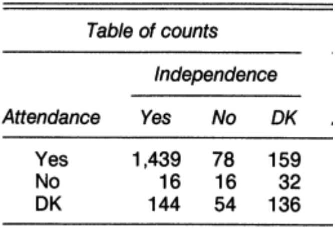
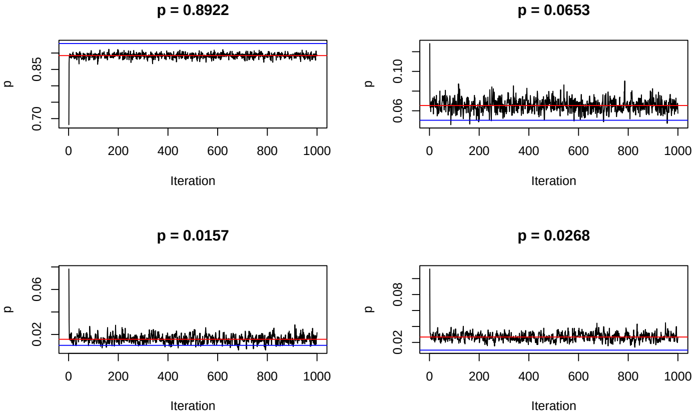


Table of counts

	Independence		
Attendance	*Yes*	*No*	*DK*
Yes	1,439	78	159
No	16	16	32
DK	144	54	136

Dirichlet model for contingency tables

- If we make the assumption that the DK's are missing at random *(we will come back to this later)*, we can use a simple Dirichlet distribution to model the probability of the four cells.
- Red line shows the posterior mean. Blue line shows the naive estimates ignoring the DK's.



Multivariate normal

- A *k*-dimensional vector *y* follows a multivariate normal distribution with mean μ and covariatence matrix Σ if

$$p(y|\mu,\Sigma)\propto|\Sigma|^{-1/2}\exp(-\frac{1}{2}(y-\mu)^T\Sigma^{-1}(y-\mu))$$

- $\mu\in\mathbb{R}^k$ and Σ is a symmetric and positive definite matrix. If Σ is not of full rank, the distribution is degenerate but does not have a density.
- A positive definite matrix Σ satisfies:
  1. $x^T\Sigma x>0$ for all nonzero vector x. $x^T\Sigma x=0$ if and only if $x=0$.
  2. Eigenvalues of Σ are all positive.
  3. The determinant of Σ is positive (since it is the product of all eigenvalues)

Properties

- Any subset of y has a normal distribution.
- Any linear combination of y is also normal.

$$Ay \sim N(A\mu, A\Sigma A^T)$$

- The conditional distributions are also normal.

$$p(y^{(1)}|y^{(2)}) \sim N(m, W)$$

where

$$\begin{aligned} m &= \mu^{(1)} + \Sigma^{(12)}(\Sigma^{(22)})^{-1}(y^{(2)} - \mu^{(2)}) \\ W &= \Sigma^{(11)} - \Sigma^{(12)}(\Sigma^{(22)})^{-1}\Sigma^{(21)} \end{aligned}$$

Likelihood

Notice we can rewrite the likelihood

$$p(y|\mu,\Sigma)\propto|\Sigma|^{-1/2}\exp(-\frac{1}{2}(y-\mu)^T\Sigma^{-1}(y-\mu))$$

into

$$p(y|\mu,\Sigma)\propto|\Sigma|^{-1/2}\exp(tr(\Sigma^{-1}S))$$

with $S=\sum_i^n(y_i-\bar{y})(y_i-\bar{y})^T$. That is the sample mean and variance are sufficient statistics for μ and Σ.

Multivariate normal with known Σ

- Let $y|\mu, \Sigma \sim N(\mu, \Sigma)$ with known covariance matrix Σ.
- We place a conjugate normal prior on $\mu \sim N(\mu_0, \Lambda_0)$.
- Then the posterior is given by

$$\mu|y, \Sigma \sim N(\mu|\mu_n, \Lambda_n)$$

where

$$\mu_n = \Lambda_n^{-1}(\Lambda_0^{-1}\mu_0 + n\Sigma^{-1}\bar{y})$$
$$\Lambda_n^{-1} = \Lambda_0^{-1} + n\Sigma^{-1}$$

- Notice for the predictive distribution of a new data point $\tilde{y}$,

$$p(\tilde{y}, \mu|y) = N(\tilde{y}|\mu, \Sigma)N(\mu|\mu_n, \Lambda_n)$$

So $(\tilde{y}, \mu)$ has a joint normal posterior distribution. That is $\tilde{y}|y$ is also normally distributed. By the law of total expectation,

$$\mathbb{E}(\tilde{y}|y) = \mathbb{E}(\mathbb{E}(\tilde{y}|\mu, y)|y) = \mathbb{E}(\mu|y) = \mu_n$$

$$\text{var}(\tilde{y}|y) = \mathbb{E}(\text{var}(\tilde{y}|\mu, y)|y) + \text{var}(\mathbb{E}(\tilde{y}|\mu, y)|y) = \mathbb{E}(\Sigma|y) + \text{var}(\mu|y) = \Sigma + \Lambda_n$$

Multivariate normal with unknown Σ

- A conjugate prior for the covariance matrix is the inverse Wishart distribution. For a $k \times k$ positive definite matrix, denote $\Sigma \sin$ Inv-Wishart(ν, Λ) if the density is

$$p(\Sigma|\nu, \Lambda) \propto |\Sigma|^{-\frac{\nu+k+1}{2}} \exp(-\frac{1}{2} tr(\Lambda\Sigma^{-1}))$$

- $\mu > k - 1$ is the degrees of freedom and Λ is the scale matrix and is positive definite.
- If μ is known and Σ is unknown, Inverse Wishart prior on $|Sigma$ leads to conjugate analysis of Σ.

Multivariate normal with unknown Σ

- More generally with both parameters unknown, we place a Normal Inverse WIshart (NIW) prior on (μ, Σ). We denote $(\mu, \Sigma) \sim \mathit{NIW}(\mu_0, \kappa_0, \nu_0, \Lambda_0)$ if

$$\Sigma \sim \text{Inv-Wishart}(\nu_0, \Lambda_0)$$

$$\mu | \Sigma_0 \sim N(\mu_0, \Sigma/\kappa_0)$$

- The posterior density of (μ, Σ) is again a NIW with

$$\begin{aligned}
\mu_n &= \frac{\kappa_0}{\kappa_0 + n}\mu_0 + \frac{n}{\kappa_0 + n}\bar{y} \\
\kappa_n &= \kappa_0 + n \\
\nu_n &= \nu_n + n \\
\Lambda_n &= \Lambda_n + \sum_i^n (y_i - \bar{y})(y_i - \bar{y})^T + \frac{\kappa_0 n}{\kappa_0 + n}(\bar{y} - \mu_0)(\bar{y} - \mu_0)^T
\end{aligned}$$

- The posterior predictive distribution of $\tilde{y}$ is a multivariate *t* distribution.

Noninformative priors

- Setting $\Sigma \sim \text{Inv-Wishart}(k+1, I)$ is a commonly used noninformative prior. It has a nice property that the marginal distribution of each off-diagonal elements in the correlation matrix is uniform (see Barnard, McCulloch, Meng, 2000).
- Another noninformative prior is the Jefferys prior,

  $$p(\mu, \Sigma) \propto |\Sigma|^{-(d+1)/2}$$

  which is the limit of NIW as $\kappa_0 \to 0$, $\nu_0 \to -1$ and $|\Lambda_0| \to 0$
- Note again, that in many situations (especially later in hierarchical models), it makes more sense to use a weakly informative prior rather than seeking the noninformativeness.

Alternatives to Inverse Wishart

- A major issue with Inverse Wishart prior for the covariance matrix is that there is only one df parameter for all dimensions. This can be seen from the marginal prior distribution of the variance components.
- Another issue is that the Inverse Wishart imposes a prior dependence between correlation and marginal variances.
- When we start to model covariance matrix in hierarchical models, we could also run into issues where the posterior distribution of Σ is heavily concentrated on a degenerate matrix (i.e., the boundary), as the density increases as $|\Sigma| \to 0$.
- Many alternatives have been proposed based on a decomposition

  $$\Sigma = diag(D) R diag(D)$$

  including the scaled inverse-Wishart, LKJ prior for the correlation matrix, etc.

Generating normal and inverse Wishart random variables

- To generate $x \sim N(\mu, \Sigma)$, we can start with the Cholesky decomposition of $\Sigma = LL^T$ where L is a lower triangular matrix.
- Then we generate $z \sim N(0, I)$ and compute $x = Lz + \mu$.

- To generate $\Sigma \sim \text{Inv-Wishart}(\nu, \Lambda)$, it is equivalent to first generate $\Sigma^{-1} = \Omega \sim \text{Wishart}(\nu, \Lambda)$ and invert the matrix.
- The Wishart distribution can be sampled by generating ν independent samples $\alpha_1, ..., \alpha_\nu \sim N(0, \Lambda)$ and let $\Omega = \sum_{i=1}^{\nu} \alpha_i \alpha_i^T$.

- A faster alternative is to use Bartlet's decomposition:
  1. Generate a lower triangular matrix A with $a_{ii} \sim \sqrt{\chi^2_{\nu-i+1}}$ and $a_{ij} \sim N(0, 1)$ for $j < i$.
  2. Compute the Cholesky decomposition $S = LL^T$.
  3. Compute $\Omega = LAA^T L^T$.

  This method requires only $k(k+1)/2$ random variable generations and automatically produces the Cholesky decomposition for Ω.

- There are R packages with highly optimized implementations, e.g., mvtnorm, MCMCpack, LaplacesDemon, ...