Multiparameter models (cont.)

Dr. Jarad Niemi

STAT 544 - Iowa State University

January 24, 2024

Outline

- Multinomial
- Multivariate normal
  - Unknown mean
  - Unknown mean and covariance

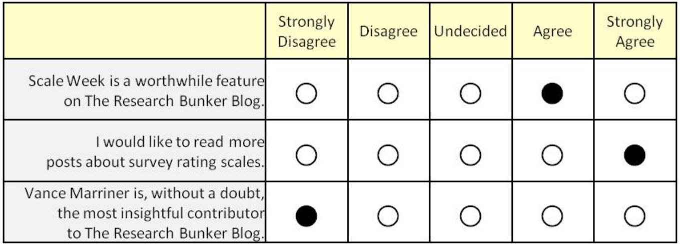
In the process, we'll introduce the following distributions

- Multinomial
- Dirichlet
- Multivariate normal
- Inverse Wishart (and Wishart)
- normal-inverse Wishart distribution

Motivating examples

Multivariate count data:

- Item-response (Likert scale)



	Strongly Disagree	Disagree	Undecided	Agree	Strongly Agree
Scale Week is a worthwhile feature on The Research Bunker Blog.	○	○	○	●	○
I would like to read more posts about survey rating scales.	○	○	○	○	●
Vance Marriner is, without a doubt, the most insightful contributor to The Research Bunker Blog.	●	○	○	○	○

- Voting

Multinomial distribution

Suppose there are K categories and each individual independently chooses category k with probability π_k such that $\sum_{k=1}^K \pi_k = 1$. Let

- $Y_k \in \{0, 1, \ldots, n\}$ be the number of individuals who choose category k
- with $n = \sum_{k=1}^K Y_k$ being the total number of individuals.

Then $Y = (Y_1, \ldots, Y_K)$ has a multinomial distribution, i.e. $Y \sim Mult(n, \pi)$, with probability mass function (pmf)

$$p(y) = n! \prod_{k=1}^{k} \frac{\pi_k^{y_k}}{y_k!}.$$

Properties of the multinomial distribution

The multinomial distribution with pmf:

$$p(y) = n! \prod_{k=1}^{k} \frac{\pi_k^{y_k}}{y_k!}$$

has the following properties:

- $E[Y_k] = n\pi_k$
- $Var[Y_k] = n\pi_k(1 - \pi_k)$
- $Cov[Y_k, Y_{k'}] = -n\pi_k\pi_{k'}$ for $k \neq k'$

Marginally, each component of a multinomial distribution is a binomial distribution with $Y_k \sim Bin(n, \pi_k)$.

Dirichlet distribution

Let $\pi = (\pi_1, \ldots, \pi_K)$ have a Dirichlet distribution, i.e. $\pi \sim Dir(a)$, with concentration parameter $a = (a_1, \ldots, a_K)$ where $a_k > 0$ for all k.

The probability density function (pdf) for π is

$$p(\pi) = \frac{1}{\text{Beta}(a)} \prod_{k=1}^{K} \pi_k^{a_k - 1}$$

with $\sum_{k=1}^K \pi_k = 1$ and $Beta(a)$ is the beta function, i.e.

$$\text{Beta}(a) = \frac{\prod_{k=1}^K \Gamma(a_k)}{\Gamma(\sum_{k=1}^K a_k)}.$$

Properties of the Dirichlet distribution

The Dirichlet distribution with pdf

$$p(\pi) \propto \prod_{k=1}^{K} \pi_k^{a_k - 1}$$

has the following properties (where $a_0 = \sum_{k=1}^{K} a_k$):

- $E[\pi_k] = \frac{a_k}{a_0}$
- $Var[\pi_k] = \frac{a_k(a_0 - a_k)}{a_0^2(a_0+1)}$
- $Cov[\pi_k, \pi_{k'}] = \frac{-a_k a_{k'}}{a_0^2(a_0+1)}$

Marginally, each component of a Dirichlet distribution is a beta distribution with $\pi_k \sim Be(a_k, a_0 - a_k)$.

Bayesian inference

The conjugate prior for a multinomial distribution, i.e. $Y \sim Mult(n, \pi)$, with unknown probability vector π is a Dirichlet distribution. The Jeffreys prior is a Dirichlet distribution with $a_k = 0.5$ for all k. Some argue that for large K, this prior will put too much mass on rare categories and would suggest the Dirichlet prior with $a_k = 1/K$ for all k.

The posterior under a Dirichlet prior is

$$\begin{aligned} p(\pi|y) &\propto p(y|\pi)p(\pi) \\ &\propto \left[\prod_{k=1}^K \pi_k^{y_k}\right]\left[\prod_{k=1}^K \pi_k^{a_k-1}\right] \\ &= \prod_{k=1}^K \pi_k^{a_k+y_k-1} \end{aligned}$$

Thus $\pi|y \sim Dir(a+y)$.

Multivariate normal distribution

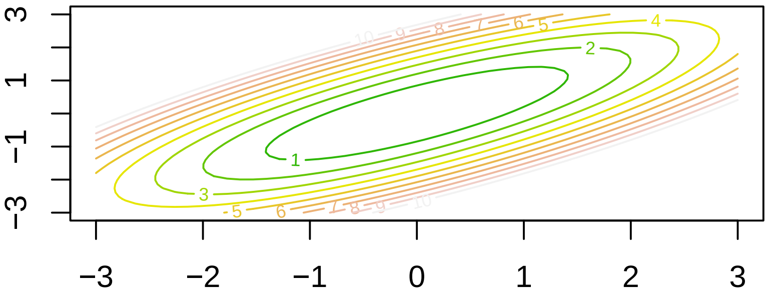
Let $Y = (Y_1, \ldots, Y_K)$ have a multivariate normal distribution, i.e. $Y \sim N_K(\mu, \Sigma)$ with mean μ and variance-covariance matrix Σ.

The probability density function (pdf) for Y is

$$p(y) = (2\pi)^{-k/2} |\Sigma|^{-1/2} \exp\left(-\frac{1}{2}(y-\mu)^\top \Sigma^{-1} (y-\mu) \right)$$

Bivariate normal contours

Contours of a bivariate normal with correlation of 0.



Properties of the multivariate normal distribution

The multivariate normal distribution has the following properties:

- For any subvector $Y_\mathbf{k}$ of Y where $\mathbf{k} \subset \{1, 2, \ldots, K\}$ with $|\mathbf{k}| = d$, we have $Y_\mathbf{k} \sim N_d(\mu_\mathbf{k}, \Sigma_{\mathbf{k},\mathbf{k}})$ where
  - $\mu_\mathbf{k}$ contains the corresponding elements from μ and
  - $\Sigma_{\mathbf{k},\mathbf{k}}$ is the submatrix of Σ constructed by extracting rows $\mathbf{k}$ and columns $\mathbf{k}$.
  - $Cov[Y_\mathbf{k}, Y_{\mathbf{k}'}] = \Sigma_{\mathbf{k},\mathbf{k}'}$ is the submatrix of Σ constructed by extracting rows $\mathbf{k}$ and columns $\mathbf{k}'$.
- Conditional distributions are also normal, i.e. for $\mathbf{k} \cap \mathbf{k}' = \emptyset$

$$\left(\begin{array}{c} Y_\mathbf{k} \\ Y_{\mathbf{k}'} \end{array} \right) \sim N\left(\left[\begin{array}{c} \mu_\mathbf{k} \\ \mu_{\mathbf{k}'} \end{array} \right], \left[\begin{array}{cc} \Sigma_{\mathbf{k},\mathbf{k}} & \Sigma_{\mathbf{k},\mathbf{k}'} \\ \Sigma_{\mathbf{k}',\mathbf{k}} & \Sigma_{\mathbf{k}',\mathbf{k}'} \end{array} \right] \right)$$

then

$$Y_\mathbf{k} | Y_{\mathbf{k}'} = y_{\mathbf{k}'} \sim N\left(\mu_\mathbf{k} + \Sigma_{\mathbf{k},\mathbf{k}'} \Sigma_{\mathbf{k}',\mathbf{k}'}^{-1} (y_{\mathbf{k}'} - \mu_{\mathbf{k}'}), \Sigma_{\mathbf{k},\mathbf{k}} - \Sigma_{\mathbf{k},\mathbf{k}'} \Sigma_{\mathbf{k}',\mathbf{k}'}^{-1} \Sigma_{\mathbf{k}',\mathbf{k}} \right).$$

Representing independence in a multivariate normal

Let $Y \sim N(\mu, \Sigma)$ with precision matrix $\Omega = \Sigma^{-1}$.

- If $\Sigma_{k,k'} = 0$, then Y_k and $Y_{k'}$ are independent of each other.
- If $\Omega_{k,k'} = 0$, then Y_k and $Y_{k'}$ are conditionally independent of each other given Y_j for $j \neq k, k'$.

Default inference with an unknown mean

Let $Y_i \overset{ind}{\sim} N_K(\mu, S)$ with default prior $p(\mu) \propto 1$ where $Y_i = (Y_{i1}, \ldots, Y_{iK})$, then

$$\begin{array}{rl} p(\mu|y) & \propto p(y|\mu)p(\mu) \\ & \propto \exp\left(-\frac{1}{2}\sum_{i=1}^n (y_i - \mu)^\top S^{-1}(y_i - \mu)\right) \\ & = \exp\left(-\frac{1}{2} tr(S^{-1}S_0)\right) \end{array}$$

where

$$S_0 = \sum_{i=1}^n (y_i - \mu)(y_i - \mu)^\top.$$

This posterior is proper if $n \geq 1$ (text has a typo) and, in that case, is

$$\mu|y \sim N_K\left(\overline{y}, \frac{1}{n}S\right).$$

where this $\overline{y} = (\overline{y}_1, \ldots, \overline{y}_K)$ has elements

$$\overline{y}_k = \frac{1}{n}\sum^{n} \overline{y}_{ik}.$$

Conjugate inference with an unknown mean

Let $Y_i \stackrel{ind}{\sim} N(\mu, S)$ with conjugate prior $\mu \sim N_K(m, C)$

$$
\begin{array}{rl}
p(\mu|y) \propto & p(y|\mu)p(\mu) \\
\propto & \exp\left(-\frac{1}{2}\sum_{i=1}^n (y_i - \mu)^\top S^{-1}(y_i - \mu)\right) \\
 & \times \exp\left(-\frac{1}{2}\mu - m)^\top C^{-1}(\mu - m)\right) \\
= & \exp\left(-\frac{1}{2}(\mu - m')^\top C'^{-1}(\mu - m')\right)
\end{array}
$$

and thus

$$\mu|y \sim N(m', C')$$

where

$$
\begin{array}{rl}
C' & = \left[C^{-1} + nS^{-1}\right]^{-1} \\
m' & = C'\left[C^{-1}m + nS^{-1}\overline{y}\right].
\end{array}
$$

Inverse Wishart distribution

Let the $K \times K$ matrix Σ have an inverse Wishart distribution, i.e. $\Sigma \sim IW(v, W^{-1})$, with degrees of freedom $v > K - 1$ and positive definite scale matrix W.

The pdf for Σ is

$$p(\Sigma) \propto |\Sigma|^{-(v+K+1)/2} \exp\left(-\frac{1}{2} tr\left(W\Sigma^{-1}\right)\right).$$

Properties of the inverse Wishart distribution

The inverse Wishart distribution with pdf

$$p(\Sigma) \propto |\Sigma|^{-(v+K+1)/2} \exp\left(-\frac{1}{2} tr\left(W\Sigma^{-1}\right)\right).$$

has the following properties:

- $E[\Sigma] = (v-K-1)^{-1}W$ for $v > K+1$.
- Marginally, $\sigma_k^2 = \Sigma_{kk} \sim Inv-\chi^2(v, W_{kk})$.
- If a $K \times K$ matrix Σ^{-1} has a Wishart distribution, i.e. $\Sigma^{-1} \sim Wishart(v, W)$, then $\Sigma \sim IW(v, W^{-1})$.

Normal-inverse Wishart distribution

A multivariate generalization of the normal-scaled-inverse-χ^2 distribution is the normal-inverse Wishart distribution. For a vector $\mu \in \mathbb{R}^K$ and $K \times K$ matrix Σ, the normal-inverse Wishart distribution is

$$\begin{aligned} \mu|\Sigma &\sim N(m, \Sigma/c) \\ \Sigma &\sim IW(v, W^{-1}) \end{aligned}$$

The marginal distribution for μ, i.e.

$$p(\mu) = \int p(\mu|\Sigma)p(\Sigma)d\Sigma,$$

is a multivariate t-distribution, i.e.

$$\mu \sim t_{v-K+1}(m, W/[c(v-K+1)]).$$

Conjugate inference with unknown mean and covariance

Let $Y_i \stackrel{ind}{\sim} N(\mu, \Sigma)$ with conjugate prior

$$\mu|\Sigma \sim N(m, \Sigma/c) \quad \Sigma \sim IW(v, W^{-1})$$

which has pdf

$$p(\mu, \Sigma) \propto |\Sigma|^{-((v+K)/2+1)} \exp\left(-\frac{1}{2}tr(W\Sigma^{-1}) - \frac{c}{2}(\mu - m)^\top \Sigma^{-1}(\mu - m)\right).$$

The posterior is a normal-inverse Wishart with parameters

$$\begin{aligned}
c' &= c + n \\
v' &= v + n \\
m' &= \tfrac{c}{c'}m + \tfrac{n}{c'}\overline{y} \\
W' &= W + S + \tfrac{cn}{c'}(\overline{y} - m)(\overline{y} - m)^\top
\end{aligned}$$

where

$$S = \sum^{n}(y_i - \overline{y})(y_i - \overline{y})^\top.$$

Default inference with unknown mean and covariance

- The prior $\Sigma \sim IW(K+1, \mathrm{I})$ is non-informative in the sense that marginally each correlation has a uniform distribution on (-1,1).
- The prior

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

  which can be thought of as a normal-inverse-Wishart distribution with $c \to 0$, $v \to -1$, and $|W| \to 0$, results in the posterior distribution

  $$\begin{aligned} \mu|\Sigma, y &\sim N(\overline{y}, \Sigma/n) \\ \Sigma|y &\sim IW(n-1, S^{-1}). \end{aligned}$$

Issues with the inverse Wishart distribution

- Marginals of the IW have an IG (or scaled-inverse-χ^2) distribution and therefore inherit the low density near zero resulting in a (possible) bias for small variances toward larger values.
- Due to the above issue, and the relationship between the variances and the correlations (`http://www.themattsimpson.com/2012/08/20/prior-distributions-for-covariance-matrices-the-scaled-inverse-wishart-prior/`), the correlations can be biased:
  - small variances imply small correlations
  - large variances imply large correlations

Remedies:

- Don't blindly use I for the scale matrix in an IW, instead use a reasonable diagonal matrix for your data set.
- Use the scaled Inverse wishart distribution (see pg 74)
- Use the separation strategy, i.e. $\Sigma = \Delta\Lambda\Delta$ where Δ is diagonal and Λ is a correlation matrix, where you specify the standard deviations (or variances) and correlations separately. In this case, Gelman recommends putting the LKJ prior (see page 582) on the correlation matrix.