Multinomial Data

The multinomial distribution is a generalization of the binomial for the situation in which each trial results in one and only one of *several* categories, as opposed to just two, as in the case of the binomial experiment.

Let $\boldsymbol{Y} = (Y_1, \ldots, Y_k)$, where Y_i is the number of n independent trials that result in category i, $i = 1, \ldots, k$. The likelihood function is such that

$$f(\boldsymbol{y}|\boldsymbol{\theta}) \propto \prod_{i=1}^{k} \theta_i^{y_i},$$

where θ_i is the probability that a given trial results in category i, $i = 1, \ldots, k$.

The parameter space is

$$\Theta = \{\boldsymbol{\theta} : \theta_i \geq 0, i = 1, \ldots, k; \sum_{j=1}^{k} \theta_j = 1\}.$$

Of course, the vector of observations satisfies $y_1 + \cdots + y_k = n$.

Conjugate prior for multinomial data

The so-called *Dirichlet* distribution is the conjugate family of priors for the multinomial distribution. The Dirichlet distribution is such that

$$\pi(\boldsymbol{\theta}; \boldsymbol{\alpha}) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} \theta_i^{\alpha_i - 1} I_{\Theta}(\boldsymbol{\theta}),$$

where $\alpha_i > 0$, $i = 1, \ldots, k$.

Using this prior in the multinomial experiment yields a Dirichlet posterior with parameters $y_i + \alpha_i$, $i = 1, \ldots, k$.

The parameters of the Dirichlet prior have the same sort of interpretation as those of a beta prior, which of course is a special case of the Dirichlet.

The information in a prior with parameters α_1, $\ldots, \alpha_k$ is equivalent to that in a multinomial experiment with $\alpha_1 + \cdots + \alpha_k$ trials and α_i outcomes in category i, $i = 1, \ldots, k$.

A natural noninformative prior is to take $\alpha_i = 1$, $i = 1, \ldots, k$, which is uniform over Θ.

What is the Jeffreys prior?

$$\log f(\boldsymbol{y}|\boldsymbol{\theta}) = C_{\boldsymbol{y}} + \sum_{i=1}^{k} y_i \log \theta_i.$$

$$\frac{\partial}{\partial \theta_j} \log f(\boldsymbol{y}|\boldsymbol{\theta}) = \frac{y_j}{\theta_j}$$

$$\frac{\partial^2}{\partial \theta_i \partial \theta_j} \log f(\boldsymbol{y}|\boldsymbol{\theta}) = \begin{cases} -y_i/\theta_i^2, & i = j, \\ 0, & i \neq j. \end{cases}$$

The information matrix is thus diagonal with diagonal entries equal to

$$\frac{1}{\theta_i^2} E(Y_i) = \frac{n}{\theta_i}, \quad i = 1, \dots, k.$$

So, the Jeffreys prior is Dirichlet with $\alpha_i = 1/2$, $i = 1, \dots, k$, which is a proper prior.

One can verify that the marginal distributions of a Dirichlet are also Dirichlet.

Multivariate Normal Distribution

Suppose we have a random sample of size n from the d-variate normal distribution. Here the data $\boldsymbol{Y}$ are an n by d matrix. The ith row of this matrix is $\boldsymbol{Y}_i^T$, where

$$\boldsymbol{Y}_i^T = (Y_{i1}, \ldots, Y_{id}), \quad i = 1, \ldots, n.$$

The parameters of the d-variate normal are the mean vector $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$. These are defined by

$$\boldsymbol{\mu} = (\mu_1, \ldots, \mu_d)^T = E(\boldsymbol{Y}_i^T)$$

and

$$\begin{aligned} \boldsymbol{\Sigma}_{ij} = \operatorname{Cov}(Y_{ri}, Y_{rj}), \qquad & i = 1, \ldots, d, \\ & j = 1, \ldots, d. \end{aligned}$$

The likelihood function is

$$f(\boldsymbol{y}|\boldsymbol{\mu},\boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-n/2}$$

$$\times \exp\left(-\tfrac{1}{2}\textstyle\sum_{i=1}^{n}(\boldsymbol{y}_i-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_i-\boldsymbol{\mu})\right).$$

In the (unlikely) event that $\boldsymbol{\Sigma}$ is known, we only need a prior for $\boldsymbol{\mu}$. It can be verified that the multivariate normal is a conjugate prior for $\boldsymbol{\mu}$ in this case.

Suppose that a priori $\boldsymbol{\mu} \sim N(\boldsymbol{\eta},\boldsymbol{\Lambda})$. Proceeding analogously to the univariate case, it can be shown that the posterior distribution is normal with mean vector $\boldsymbol{\mu}_n$ and covariance matrix $\boldsymbol{\Lambda}_n$, where

$$\boldsymbol{\mu}_n = (\boldsymbol{\Lambda}^{-1}+n\boldsymbol{\Sigma}^{-1})^{-1}(\boldsymbol{\Lambda}^{-1}\boldsymbol{\eta}+n\boldsymbol{\Sigma}^{-1}\bar{\boldsymbol{y}})$$

and

$$\boldsymbol{\Lambda}_n^{-1} = \boldsymbol{\Lambda}^{-1}+n\boldsymbol{\Sigma}^{-1}.$$

Let $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ contain the first k and the last $d-k$ elements of $\boldsymbol{\mu}$, respectively. Similarly, define $\boldsymbol{\mu}_{n1}$ and $\boldsymbol{\mu}_{n2}$ in terms of the elements of $\boldsymbol{\mu}_n$.

Partition $\boldsymbol{\Lambda}_n$ as

$$\boldsymbol{\Lambda}_n = \begin{bmatrix} \boldsymbol{\Lambda}_n^{11} & \boldsymbol{\Lambda}_n^{12} \\ \boldsymbol{\Lambda}_n^{21} & \boldsymbol{\Lambda}_n^{22} \end{bmatrix},$$

where $\boldsymbol{\Lambda}_n^{11}$ is $k \times k$ and $\boldsymbol{\Lambda}_n^{22}$ is $(d-k) \times (d-k)$.

It follows that the conditional distribution of $\boldsymbol{\mu}_1$ given $\boldsymbol{\mu}_2$ is normal with mean vector

$$\boldsymbol{\mu}_{n1} + \boldsymbol{\Lambda}_n^{12}(\boldsymbol{\Lambda}_n^{22})^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_{n2})$$

and covariance matrix

$$\boldsymbol{\Lambda}_n^{11} - \boldsymbol{\Lambda}_n^{12}(\boldsymbol{\Lambda}_n^{22})^{-1}\boldsymbol{\Lambda}_n^{21}.$$

Of course, the marginal of, for example, $\boldsymbol{\mu}_1$ is normal with mean vector $\boldsymbol{\mu}_{n1}$ and covariance matrix $\boldsymbol{\Lambda}_n^{11}$.

By letting $|\mathbf{\Lambda}^{-1}| \to 0$, we can obtain a noninformative prior in the limit. The resulting prior is uniform over all the d-dimensional reals, and of course is improper.

If $n \geq d$, the posterior corresponding to the uniform prior for $\boldsymbol{\mu}$ is $N(\bar{\boldsymbol{y}}, \mathbf{\Sigma}/n)$.

Inadmissibility of a Bayes estimator: James-Stein theory

Suppose we observe $\boldsymbol{Y}$ that has a d-variate normal distribution with unknown mean vector $\boldsymbol{\mu}$ and known covariance matrix $\boldsymbol{I}_d$, the $d \times d$ identity.

This problem is equivalent to one where we simultaneously estimate means from independent experiments.

If we use the noninformative, uniform prior for $\boldsymbol{\mu}$, and the squared error loss

$$L(\boldsymbol{\mu}, \boldsymbol{a}) = \sum_{i=1}^{d} (\mu_i - a_i)^2,$$

then the Bayes estimator of $\boldsymbol{\mu}$ is, not surprisingly, $\boldsymbol{Y}$.

The surprising thing is that this "natural" estimator is inadmissible for $d \geq 3$. (It is admissible for $d = 1$ or 2.) This result is proven by Stein (1955), *Proceedings of the Third Berkeley Symposium*.

James and Stein (1960), *Proceedings of the Fourth Berkeley Symposium*, produced an estimator that has uniformly smaller risk than $\boldsymbol{Y}$. The estimator is

$$\delta_{\mathsf{JS}}(\boldsymbol{Y}) = \left(1 - \frac{d-2}{\sum_{i=1}^{d} Y_i^2}\right)\boldsymbol{Y}.$$

It turns out that the ratio $R(\boldsymbol{\mu}, \delta_{\text{JS}})/R(\boldsymbol{\mu}, \boldsymbol{Y})$ is very close to 1 over most of the parameter space. Only near $\boldsymbol{\mu}^T = (0, \ldots, 0)$ is the ratio of risks substantially smaller than 1.

One way of seeing why is to first prove the following fact:

> For a set of μ_i's that are all bounded in absolute value by the same constant, and when d is large, the statistic
>
> $$T_d = \sum_{i=1}^{d} Y_i^2/(d-2)$$
>
> is very close to $\theta_d = 1 + \sum_{i=1}^{d} \mu_i^2/d$.

Proof

We have, for each i,

$$E(Y_i^2) = 1 + \mu_i^2$$

and

$$\text{Var}(Y_i^2) = 2(1 + 2\mu_i^2).$$

For an arbitrarily small, positive ϵ, Markov's inequality says that

$$P\left(|T_d - \theta_d| < \epsilon\right) \geq 1 - E(T_d - \theta_d)^2/\epsilon^2.$$

Now,

$$\begin{aligned} E(T_d - \theta_d)^2 &= \text{Var}(T_d) + [E(T_d) - \theta_d]^2 \\ &= \text{Var}(T_d) + \frac{4\theta_d^2}{(d-2)^2}. \end{aligned}$$

Since the Y_is are independent,

$$\text{Var}(T_d) = \frac{2}{(d-2)^2} \sum_{i=1}^{d} (1 + 2\mu_i^2).$$

Using the fact that $|\mu_1|, \ldots, |\mu_d|$ are all less than or equal to the same constant, we have

$$E(T_d - \theta_d)^2 \leq \frac{C}{d}$$

for some positive constant C. It follows that when d is sufficiently big, $P\left(|T_d - \theta_d| < \epsilon\right)$ is arbitrarily close to 1.

Q.E.D.

To get a better understanding of the James-Stein estimator, we now consider

$$\hat{\delta}_{\text{JS}}(\boldsymbol{Y}) = \left(1 - \frac{1}{\theta_d}\right)\boldsymbol{Y} = \left(\frac{\bar{\mu}_d^2}{1+\bar{\mu}_d^2}\right)\boldsymbol{Y},$$

where $\bar{\mu}_d^2 = \sum_{i=1} \mu_i^2/d$.

The result proved on the previous pages shows that, for large d, $\delta_{\text{JS}} \approx \hat{\delta}_{\text{JS}}$.

The random variable $\hat{\delta}_{\text{JS}}$ provides us with some intuition about the James-Stein estimator. If the vector $\boldsymbol{\mu}$ is close to the origin, i.e., $\mathbf{0} = (0, \ldots, 0)^T$, then $\bar{\mu}_d^2$ is close to 0, and hence $\hat{\delta}_{\text{JS}} \approx \mathbf{0}$. This is good!!

On the other hand, if $\boldsymbol{\mu}$ is far from the origin, then

$$\frac{\bar{\mu}_d^2}{1+\bar{\mu}_d^2} \approx 1,$$

and $\widehat{\delta}_{\mathsf{JS}} \approx \boldsymbol{Y}$. This is good, because if $\boldsymbol{\mu}$ is not close to the origin, then there's no rationale for shrinking the estimate towards the origin.

Shrinkage towards $\mathbf{0}$ is arbitrary

Suppose we have a rationale for shrinking the estimate towards a point $\boldsymbol{\alpha}$ in d-space. For example, some theory may suggest that $\boldsymbol{\mu} = \boldsymbol{\alpha}$.

We may define an estimate

$$\delta_{\mathsf{JS}}(\boldsymbol{Y};\boldsymbol{\alpha}) = \boldsymbol{Y} - \frac{d-2}{\sum_{i=1}^d (Y_i - \alpha_i)^2}(\boldsymbol{Y} - \boldsymbol{\alpha}).$$

Using the squared error loss on p. 130N, verify that

$$R(\boldsymbol{\mu}, \delta_{\mathsf{JS}}(\boldsymbol{Y}; \boldsymbol{\alpha})) = R(\boldsymbol{\mu} - \boldsymbol{\alpha}, \delta_{\mathsf{JS}}(\boldsymbol{Y})) \qquad (*)$$

for all $\boldsymbol{\mu}$.

Because $\boldsymbol{Y}$ has constant risk and because

$$R(\boldsymbol{\mu}, \delta_{\mathsf{JS}}(\boldsymbol{Y})) \leq R(\boldsymbol{\mu}, \boldsymbol{Y})$$

for all $\boldsymbol{\mu}$, $(*)$ implies that $\delta_{\mathsf{JS}}(\boldsymbol{Y}; \boldsymbol{\alpha})$ has risk no larger than that of $\boldsymbol{Y}$ for all $\boldsymbol{\mu}$.

The part of the parameter space where $\delta_{\mathsf{JS}}(\boldsymbol{Y}; \boldsymbol{\alpha})$ has substantially smaller risk is near $\boldsymbol{\alpha}$.

Read Example 46, p. 256 of Berger (2nd edition) for more details on this problem.