14
Multivariate Models

In this chapter we revisit the Multinomial model and the multivariate Normal. Let us first review some notation from linear algebra. In what follows, x and y are vectors and A is a matrix.

Linear Algebra Notation

$x^T y$	inner product $\sum_j x_j y_j$
$\|A\|$	determinant
A^T	transpose of A
A^{-1}	inverse of A
I	the identity matrix
$\text{tr}(A)$	trace of a square matrix; sum of its diagonal elements
$A^{1/2}$	square root matrix

The trace satisfies $\text{tr}(AB) = \text{tr}(BA)$ and $\text{tr}(A) + \text{tr}(B)$. Also, $\text{tr}(a) = a$ if a is a scalar. A matrix is **positive definite** if $x^T \Sigma x > 0$ for all nonzero vectors x. If a matrix A is symmetric and positive definite, its square root $A^{1/2}$ exists and has the following properties: (1) $A^{1/2}$ is symmetric; (2) $A = A^{1/2} A^{1/2}$; (3) $A^{1/2} A^{-1/2} = A^{-1/2} A^{1/2} = I$ where $A^{-1/2} = (A^{1/2})^{-1}$.

L. Wasserman, *All of Statistics*

14.1 Random Vectors

Multivariate models involve a random vector X of the form

$$X = \begin{pmatrix} X_1 \\ \vdots \\ X_k \end{pmatrix}.$$

The mean of a random vector X is defined by

$$\mu = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_k \end{pmatrix} = \begin{pmatrix} E(X_1) \\ \vdots \\ E(X_k) \end{pmatrix}. \tag{14.1}$$

The **covariance matrix** Σ, also written $\mathbb{V}(X)$, is defined to be

$$\Sigma = \begin{bmatrix} \mathbb{V}(X_1) & \mathsf{Cov}(X_1, X_2) & \cdots & \mathsf{Cov}(X_1, X_k) \\ \mathsf{Cov}(X_2, X_1) & \mathbb{V}(X_2) & \cdots & \mathsf{Cov}(X_2, X_k) \\ \vdots & \vdots & \vdots & \vdots \\ \mathsf{Cov}(X_k, X_1) & \mathsf{Cov}(X_k, X_2) & \cdots & \mathbb{V}(X_k) \end{bmatrix}. \tag{14.2}$$

This is also called the variance matrix or the variance–covariance matrix. The inverse Σ^{-1} is called the **precision matrix**.

14.1 Theorem. *Let a be a vector of length k and let X be a random vector of the same length with mean μ and variance Σ. Then $\mathbb{E}(a^T X) = a^T \mu$ and $\mathbb{V}(a^T X) = a^T \Sigma a$. If A is a matrix with k columns, then $\mathbb{E}(AX) = A\mu$ and $\mathbb{V}(AX) = A\Sigma A^T$.*

Now suppose we have a random sample of n vectors:

$$\begin{pmatrix} X_{11} \\ X_{21} \\ \vdots \\ X_{k1} \end{pmatrix}, \begin{pmatrix} X_{12} \\ X_{22} \\ \vdots \\ X_{k2} \end{pmatrix}, \ldots, \begin{pmatrix} X_{1n} \\ X_{2n} \\ \vdots \\ X_{kn} \end{pmatrix}. \tag{14.3}$$

The sample mean $\overline{X}$ is a vector defined by

$$\overline{X} = \begin{pmatrix} \overline{X}_1 \\ \vdots \\ \overline{X}_k \end{pmatrix}$$

where $\overline{X}_i = n^{-1}\sum_{j=1}^n X_{ij}$. The sample variance matrix, also called the covariance matrix or the variance–covariance matrix, is

$$S = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1k} \\ s_{12} & s_{22} & \cdots & s_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ s_{1k} & s_{2k} & \cdots & s_{kk} \end{bmatrix} \tag{14.4}$$

where

$$s_{ab} = \frac{1}{n-1}\sum_{j=1}^n (X_{aj} - \overline{X}_a)(X_{bj} - \overline{X}_b).$$

It follows that $\mathbb{E}(\overline{X}) = \mu.$ and $\mathbb{E}(S) = \Sigma$.

14.2 Estimating the Correlation

Consider n data points from a bivariate distribution:

$$\begin{pmatrix} X_{11} \\ X_{21} \end{pmatrix}, \begin{pmatrix} X_{12} \\ X_{22} \end{pmatrix}, \cdots, \begin{pmatrix} X_{1n} \\ X_{2n} \end{pmatrix}.$$

Recall that the correlation between X_1 and X_2 is

$$\rho = \frac{\mathbb{E}((X_1 - \mu_1)(X_2 - \mu_2))}{\sigma_1 \sigma_2} \tag{14.5}$$

where $\sigma_j^2 = \mathbb{V}(X_{ji})$, $j = 1, 2$. The nonparametric plug-in estimator is the sample correlation [1]

$$\widehat{\rho} = \frac{\sum_{i=1}^n (X_{1i} - \overline{X}_1)(X_{2i} - \overline{X}_2)}{s_1 s_2} \tag{14.6}$$

where

$$s_j^2 = \frac{1}{n-1}\sum_{i=1}^n (X_{ji} - \overline{X}_j)^2.$$

We can construct a confidence interval for ρ by applying the delta method. However, it turns out that we get a more accurate confidence interval by first constructing a confidence interval for a function $\theta = f(\rho)$ and then applying

[1] More precisely, the plug-in estimator has n rather than $n-1$ in the formula for s_j but this difference is small.

the inverse function f^{-1}. The method, due to Fisher, is as follows: Define f and its inverse by

$$\begin{aligned} f(r) &= \frac{1}{2}\Big(\log(1+r) - \log(1-r)\Big) \\ f^{-1}(z) &= \frac{e^{2z}-1}{e^{2z}+1}. \end{aligned}$$

Approximate Confidence Interval for The Correlation

1. Compute
$$\widehat{\theta} = f(\widehat{\rho}) = \frac{1}{2}\Big(\log(1+\widehat{\rho}) - \log(1-\widehat{\rho})\Big).$$
2. Compute the approximate standard error of $\widehat{\theta}$ which can be shown to be
$$\widehat{\mathsf{se}}(\widehat{\theta}) = \frac{1}{\sqrt{n-3}}.$$
3. An approximate $1-\alpha$ confidence interval for $\theta = f(\rho)$ is
$$(a,b) \equiv \left(\widehat{\theta} - \frac{z_{\alpha/2}}{\sqrt{n-3}}, \widehat{\theta} + \frac{z_{\alpha/2}}{\sqrt{n-3}}\right).$$
4. Apply the inverse transformation $f^{-1}(z)$ to get a confidence interval for ρ:
$$\left(\frac{e^{2a}-1}{e^{2a}+1}, \frac{e^{2b}-1}{e^{2b}+1}\right).$$

Yet another method for getting a confidence interval for ρ is to use the bootstrap.

14.3 Multivariate Normal

Recall that a vector X has a multivariate Normal distribution, denoted by $X \sim N(\mu, \Sigma)$, if its density is

$$f(x;\, \mu, \Sigma) = \frac{1}{(2\pi)^{k/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right\} \tag{14.7}$$

where μ is a vector of length k and Σ is a $k \times k$ symmetric, positive definite matrix. Then $\mathbb{E}(X) = \mu$ and $\mathbb{V}(X) = \Sigma$.

14.2 Theorem. *The following properties hold:*

1. *If $Z \sim N(0,1)$ and $X = \mu + \Sigma^{1/2} Z$, then $X \sim N(\mu, \Sigma)$.*
2. *If $X \sim N(\mu, \Sigma)$, then $\Sigma^{-1/2}(X - \mu) \sim N(0,1)$.*
3. *If $X \sim N(\mu, \Sigma)$ a is a vector of the same length as X, then $a^T X \sim N(a^T \mu, a^T \Sigma a)$.*
4. *Let*
$$V = (X - \mu)^T \Sigma^{-1} (X - \mu).$$
Then $V \sim \chi^2_k$.

14.3 Theorem. *Given a random sample of size n from a $N(\mu, \Sigma)$, the log-likelihood is (up to a constant not depending on μ or Σ) given by*

$$\ell(\mu, \Sigma) = -\frac{n}{2}(\overline{X} - \mu)^T \Sigma^{-1} (\overline{X} - \mu) - \frac{n}{2} \text{tr}(\Sigma^{-1} S) - \frac{n}{2} \log |\Sigma|.$$

The MLE *is*

$$\widehat{\mu} = \overline{X} \quad \text{and} \quad \widehat{\Sigma} = \left(\frac{n-1}{n} \right) S. \tag{14.8}$$

14.4 Multinomial

Let us now review the Multinomial distribution. The data take the form $X = (X_1, \ldots, X_k)$ where each X_j is a count. Think of drawing n balls (with replacement) from an urn which has balls with k different colors. In this case, X_j is the number of balls of the k^{th} color. Let $p = (p_1, \ldots, p_k)$ where $p_j \geq 0$ and $\sum_{j=1}^k p_j = 1$ and suppose that p_j is the probability of drawing a ball of color j.

14.4 Theorem. *Let $X \sim \text{Multinomial}(n, p)$. Then the marginal distribution of X_j is $X_j \sim \text{Binomial}(n, p_j)$. The mean and variance of X are*

$$\mathbb{E}(X) = \begin{pmatrix} np_1 \\ \vdots \\ np_k \end{pmatrix}$$

and

$$\mathbb{V}(X) = \begin{pmatrix} np_1(1-p_1) & -np_1 p_2 & \cdots & -np_1 p_k \\ -np_1 p_2 & np_2(1-p_2) & \cdots & -np_2 p_k \\ \vdots & \vdots & \vdots & \vdots \\ -np_1 p_k & -np_2 p_k & \cdots & np_k(1-p_k) \end{pmatrix}.$$

PROOF. That $X_j \sim \text{Binomial}(n, p_j)$ follows easily. Hence, $\mathbb{E}(X_j) = np_j$ and $\mathbb{V}(X_j) = np_j(1-p_j)$. To compute $\mathsf{Cov}(X_i, X_j)$ we proceed as follows: Notice that $X_i + X_j \sim \text{Binomial}(n, p_i + p_j)$ and so $\mathbb{V}(X_i + X_j) = n(p_i + p_j)(1 - p_i - p_j)$. On the other hand,

$$\begin{aligned} \mathbb{V}(X_i + X_j) &= \mathbb{V}(X_i) + \mathbb{V}(X_j) + 2\mathsf{Cov}(X_i, X_j) \\ &= np_i(1-p_i) + np_j(1-p_j) + 2\mathsf{Cov}(X_i, X_j). \end{aligned}$$

Equating this last expression with $n(p_i + p_j)(1 - p_i - p_j)$ implies that $\mathsf{Cov}(X_i, X_j) = -np_ip_j$. ∎

14.5 Theorem. *The maximum likelihood estimator of p is*

$$\widehat{p} = \begin{pmatrix} \widehat{p}_1 \\ \vdots \\ \widehat{p}_k \end{pmatrix} = \begin{pmatrix} \frac{X_1}{n} \\ \vdots \\ \frac{X_k}{n} \end{pmatrix} = \frac{X}{n}.$$

PROOF. The log-likelihood (ignoring a constant) is

$$\ell(p) = \sum_{j=1}^{k} X_j \log p_j.$$

When we maximize ℓ we have to be careful since we must enforce the constraint that $\sum_j p_j = 1$. We use the method of Lagrange multipliers and instead maximize

$$A(p) = \sum_{j=1}^{k} X_j \log p_j + \lambda\left(\sum_j p_j - 1\right).$$

Now

$$\frac{\partial A(p)}{\partial p_j} = \frac{X_j}{p_j} + \lambda.$$

Setting $\frac{\partial A(p)}{\partial p_j} = 0$ yields $\widehat{p}_j = -X_j/\lambda$. Since $\sum_j \widehat{p}_j = 1$ we see that $\lambda = -n$ and hence $\widehat{p}_j = X_j/n$ as claimed. ∎

Next we would like to know the variability of the MLE. We can either compute the variance matrix of $\widehat{p}$ directly or we can approximate the variability of the MLE by computing the Fisher information matrix. These two approaches give the same answer in this case. The direct approach is easy: $\mathbb{V}(\widehat{p}) = \mathbb{V}(X/n) = n^{-2}\mathbb{V}(X)$, and so

$$\mathbb{V}(\widehat{p}) = \frac{1}{n}\Sigma$$

where

$$\Sigma = \begin{pmatrix} p_1(1-p_1) & -p_1p_2 & \cdots & -p_1p_k \\ -p_1p_2 & p_2(1-p_2) & \cdots & -p_2p_k \\ \vdots & \vdots & \vdots & \vdots \\ -p_1p_k & -p_2p_k & \cdots & p_k(1-p_k) \end{pmatrix}.$$

For large n, $\widehat{p}$ has approximately a multivariate Normal distribution.

14.6 Theorem. *As* $n \to \infty$,

$$\sqrt{n}(\widehat{p} - p) \rightsquigarrow N(0, \Sigma).$$

14.5 Bibliographic Remarks

Some references on multivariate analysis are Johnson and Wichern (1982) and Anderson (1984). The method for constructing the confidence interval for the correlation described in this chapter is due to Fisher (1921).

14.6 Appendix

PROOF of Theorem 14.3. Denote the i^{th} random vector by X^i. The log-likelihood is

$$\begin{aligned} \ell(\mu, \Sigma) &= \sum_{i=1}^{n} f(X^i;\ \mu, \Sigma) \\ &= -\frac{kn}{2}\log(2\pi) - \frac{n}{2}\log|\Sigma| - \frac{1}{2}\sum_{i=1}^{n}(X^i - \mu)^T\Sigma^{-1}(X^i - \mu). \end{aligned}$$

Now,

$$\begin{aligned} &\sum_{i=1}^{n}(X^i - \mu)^T\Sigma^{-1}(X^i - \mu) \\ &\quad= \sum_{i=1}^{n}[(X^i - \overline{X}) + (\overline{X} - \mu)]^T\Sigma^{-1}[(X^i - \overline{X}) + (\overline{X} - \mu)] \\ &\quad= \sum_{i=1}^{n}\left[(X^i - \overline{X})^T\Sigma^{-1}(X^i - \overline{X})\right] + n(\overline{X} - \mu)^T\Sigma^{-1}(\overline{X} - \mu) \end{aligned}$$

since $\sum_{i=1}^{n}(X^i - \overline{X})\Sigma^{-1}(\overline{X} - \mu) = 0$. Also, notice that $(X^i - \mu)^T\Sigma^{-1}(X^i - \mu)$ is a scalar, so

$$\sum_{i=1}^{n}(X^i - \mu)^T\Sigma^{-1}(X^i - \mu) = \sum_{i=1}^{n}\text{tr}\left[(X^i - \mu)^T\Sigma^{-1}(X^i - \mu)\right]$$

$$\begin{aligned} &= \sum_{i=1}^{n} \text{tr}\left[\Sigma^{-1}(X^i - \mu)(X^i - \mu)^T\right] \\ &= \text{tr}\left[\Sigma^{-1}\sum_{i=1}^{n}(X^i - \mu)(X^i - \mu)^T\right] \\ &= n \text{ tr}\left[\Sigma^{-1}S\right] \end{aligned}$$

and the conclusion follows. ■

14.7 Exercises

1. Prove Theorem 14.1.
2. Find the Fisher information matrix for the MLE of a Multinomial.
3. (Computer Experiment.) Write a function to generate `nsim` observations from a Multinomial(n, p) distribution.
4. (Computer Experiment.) Write a function to generate `nsim` observations from a Multivariate normal with given mean μ and covariance matrix Σ.
5. (Computer Experiment.) Generate 100 random vectors from a $N(\mu, \Sigma)$ distribution where

$$\mu = \begin{pmatrix} 3 \\ 8 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}.$$

   Plot the simulation as a scatterplot. Estimate the mean and covariance matrix Σ. Find the correlation ρ between X_1 and X_2. Compare this with the sample correlations from your simulation. Find a 95 percent confidence interval for ρ. Use two methods: the bootstrap and Fisher's method. Compare.
6. (Computer Experiment.) Repeat the previous exercise 1000 times. Compare the coverage of the two confidence intervals for ρ.