Notes for STA 437/1005 — Methods for Multivariate Data

Radford M. Neal, 26 November 2010

Random Vectors

Notation:

Let X be a random vector with p elements, so that $X = [X_1, \ldots, X_p]'$, where $'$ denotes transpose. (By convention, our vectors are column vectors unless otherwise indicated.)

We denote a particular realized value of X by x.

Expection:

The expectation (expected value, mean) of a random vector X is $E(X) = \int x f(x) dx$, where $f(x)$ is the joint probability density function for the distribution of X.

We often denote $E(X)$ by μ, with $\mu_j = E(X_j)$ being the expectation of the j'th element of X.

Variance:

The variance of the random variable X_j is $\mathrm{Var}(X_j) = E[(X_j - E(X_j))^2]$, which we sometimes write as σ_j^2.

The standard deviation of X_j is $\sqrt{\mathrm{Var}(X_j)} = \sigma_j$.

Covariance and correlation:

The covariance of X_j and X_k is $\mathrm{Cov}(X_j, X_k) = E[(X_j - E(X_j))(X_k - E(X_k))]$, which we sometimes write as σ_{jk}. Note that $\mathrm{Cov}(X_j, X_j)$ is the variance of X_j, so $\sigma_{jj} = \sigma_j^2$.

The correlation of X_j and X_k is $\mathrm{Cov}(X_j, X_k)/(\sigma_j \sigma_k)$, which we sometimes write as ρ_{jk}. Note that correlations are always between -1 and $+1$, and ρ_{jj} is always one.

Covariance and correlation matrices:

The covariances for all pairs of elements of $X = [X_1, \ldots, X_p]'$ can be put in a matrix called the covariance matrix:

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{bmatrix}$$

Note that the covariance matrix is symmetrical, with the variances of the elements on the diagonal.

The covariance matrix can also be written as $\Sigma = E\left[\,(X - E(X))\,(X - E(X))'\,\right]$.

Similarly, the correlations can be put into a a symmetrical correlation matrix, which will have ones on the diagonal.

Multivariate Sample Statistics

Notation:

Suppose we have n observations, each with values for p variables. We denote the value of variable j in observation i by x_{ij}, and the vector of all values for observation i by x_i.

We often view the observed x_i as a random sample of realizations of a random vector X with some (unknown) distribution.

The is potential ambiguity between the notation x_i for observation i, and the notation x_j for a realization of the random variable X_j. (The textbook uses bold face for x_i.)

I will (try to) reserve i for indexing observations, and use j and k for indexing variables, but the textbook somtimes uses i to index a variable.

Sample means:

The sample mean of variable j is $\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$.

The sample mean vector is $\bar{x} = [\bar{x}_1, \ldots, \bar{x}_p]'$.

If the observations all have the same distribution, the sample mean vector, $\bar{x}$, is an unbiased estimate of the mean vector, μ, of the distribution from which these observations came.

Sample variances:

The sample variance of variable j is $s_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2$.

If the observations all have the same distribution, the sample variance, s_j^2, is an estimate of the variance, σ_j^2, of the distribution for X_j, and will be an unbiased estimate if the observations are independent.

Sample covariance and correlation:

The sample covariance of variable j with variable k is $\frac{1}{n-1}\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)(x_{ik}-\bar{x}_k)$.

The sample covariance is denoted by s_{jk}. Note that s_{jj} equals s_j^2, the sample variance of variable j.

The sample correlation of variable j with variable k is $s_{jk}/(s_j s_k)$, often denoted by r_{jk}.

Sample covariance and correlation matrices:

The sample covariances may be arranged as the sample covariance matrix:

$$S = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{bmatrix}$$

The sample covariance matrix can also be computed as $S = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})'$.

Similarly, the sample correlations may be arranged as the sample correlation matrix, sometimes denoted R (though the textbook also uses R for the population correlation matrix).

Linear Combinations of Random Variables

Define the random variable $Y = a_1X_1 + a_2X_2 + \cdots + a_pX_p$, which can be written as $Y = a'X$, where $a = [a_1, a_2, \ldots, a_p]'$.

Then one can show that $E(Y) = a'\mu$ and $\mathrm{Var}(Y) = a'\Sigma a$, where $\mu = E(X)$ and Σ is the covariance matrix for X.

For a random vector of dimension q defined as $Y = AX$, with A being a $q \times p$ matrix, one can show that $E(Y) = A\mu$ and $\mathrm{Var}(Y) = A\Sigma A'$, where $\mathrm{Var}(Y)$ is the covariance matrix of Y.

Similarly, if x_i is the i'th observed vector, and we define $y_i = Ax_i$, then the sample mean vector of y is $\bar{y} = A\bar{x}$ and the sample covariance matrix of y is ASA', where S is the sample covariance matrix of x.

Positive definite matrices

A symmetric square matrix, A, is said to be positive definite if $v'Av > 0$ for any non-zero vector v. A is said to be positive semi-definite (or non-negative definite) if $v'Av \geq 0$ for all v.

Covariance and correlation matrices, and sample covariance and sample correlation matrices, are always positive semi-definite.

The Multivariate Normal (Gaussian) Distribution

If X is a random vector of dimension p, the multivariate normal (also called Gaussian) distribution for X with mean μ and covariance matrix Σ will have the joint probability density function

$$f(x) = \frac{1}{(2\pi)^{p/2}} \frac{1}{(\det \Sigma)^{1/2}} \exp\left(-\tfrac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right)$$

Such a multivariate normal distribution, written as $N(\mu, \Sigma)$, is defined for any vector μ of dimension p and any $p \times p$ positive definite matrix Σ.

If X has the $N(\mu, \Sigma)$ distribution, then $D^2 = (X-\mu)'\Sigma^{-1}(X-\mu)$ has the χ^2 distribution with p degrees of freedom.

If $x_1, \ldots, x_n$ are i independent observations on p variables coming from a multivariate normal distribution, with sample mean $\bar{x}$ and sample covariance matrix S, then $d_i^2 = (x_i - \bar{x})'S^{-1}(x_i - \bar{x})$ will have approximately a χ^2 distribution with p degrees of freedom (for large n).

Eigenvectors and eigenvalues of positive definite matrices

By definition, a non-zero vector e is an eigenvector of a square matrix M, with eigenvalue λ, if $Me = \lambda e$.

The eigenvectors of a symmetric, positive definite matrix are real (no imaginary part), and their

eigenvalues are positive reals. (For symmetric, positive semi-definite matrices, the eigenvectors are real and the eigenvalues are non-negative reals).

The spectral decomposition theorem: If A is a $k \times k$ symmetric real matrix, it is possible to find a set of k eigenvectors of A that are orthogonal and have length one, and if $e_1, \ldots, e_k$ are any such set of eigenvectors, with eigenvalues $\lambda_1, \ldots, \lambda_k$, then

$$A \;=\; \lambda_1 e_1 e_1' \;+\; \cdots \;+\; \lambda_k e_k e_k'$$

Principal Component Analysis (PCA)

The principal component directions of a $p \times p$ covariance or correlation matrix are the eigenvectors, e_1, e_2, ..., e_p, ordered by decreasing eigenvalue, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$.

The principal components have the property that e_1 is the direction with greatest variance, e_2 is the direction of greatest variance subject to the constraint that e_2 be orthogonal to e_1, etc.

From n observations, x_1, ..., x_n, each a vector of p variables, we find the principal components from either the sample covariance matrix or the sample correlation matrix. The projection of observation i on the k'th principal component direction is $z_{ik} = e_k' x_i$. We may decide to reduce the observations to just the projections on the first k principal component directions, with k less than p.

Various forms of the multivariate T^2 test

These T^2 tests generalize the one-sample and two-sample t tests for univariate data. The data is a sample of n observations of p variables, or two samples of n_1 and n_2 observations of p variables.

T^2 statistic for one sample:

$$T^2 \quad = \quad n(\overline{\mathbf{x}} - \mu_0)' \mathbf{S}^{-1} (\overline{\mathbf{x}} - \mu_0)$$

The distribution of T^2 under the null hypothesis that the mean is μ_0 is $[(n-1)p/(n-p)]F_{p,n-p}$, which is approximately χ^2_p when $n - p$ and n/p are both large.

T^2 statistic for two samples, using pooled covariance estimate:

$$T^2 \quad = \quad ((\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2) \;-\; \delta_0)' \; [(1/n_1 + 1/n_2)\, \mathbf{S}_{\text{pooled}}]^{-1} \; ((\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2) \;-\; \delta_0)$$

Here, $\mathbf{S}_{\text{pooled}} \;=\; ((n_1-1)\mathbf{S}_1 + (n_2-1)\mathbf{S}_2) \;/\; (n_1+n_2-2)$. The distribution of T^2 under the null hypothesis that the difference in means of the groups is δ_0 is $[(n_1+n_2-2)p/(n_1+n_2-p-1)]F_{p,n_1+n_2-p-1}$. This distribution is approximately χ^2_p when $n_1 + n_2 - p$ and $(n_1 + n_2)\,/\,p$ are both large.

T^2 statistic for two samples, covariances not necessarily equal:

$$T^2 \quad = \quad ((\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2) \;-\; \delta_0)' \; [(1/n_1)\mathbf{S}_1 + (1/n_2)\mathbf{S}_2]^{-1} \; ((\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2) \;-\; \delta_0)$$

There is no exact formula for the distribution of T^2 in this case, but there are approximations using an F distribution, and when both $n_1 - p$ and $n_2 - p$ are large, the distribution is approximately χ^2_p.

Bonferroni correction with multiple tests

If m hypothesis tests are performed, and we reject those of the m null hypotheses for which the p-value is less than α/m, then we are guaranteed that the probability of rejecting *any* true null hypothesis is no greater than α.

Similarly, if we create m level $1 - \alpha/m$ confidence intervals, the probability that *all* confidence intervals will contain the true parameter value is at least $1 - \alpha$.

False Discovery Rates

Define the False Discovery Rate (FDR) as the expected fraction of rejected hypotheses that are false (this fraction is defined to be zero if no hypotheses are rejected).

Suppose we fix a limit on False Discovery Rate of α, in advance of analysing the data. We perform m hypothesis tests, and order the p-values obtained as $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$. We then set k as follows:

$$k \quad = \quad \max \left\{ i \ : \ p_{(i)} \ \leq \ (i/m)\, \alpha \right\}$$

and we reject all those hypotheses for which the p-value is $p_{(k)}$ or less. (We reject no hypotheses if there is no i with $p_{(i)} \leq (i/m)\alpha$).

Then the False Discovery Rate is bounded by $FDR \leq \alpha$.