STAT 802: Multivariate Analysis

Course outline:

- Multivariate Distributions.
- The Multivariate Normal Distribution.
- The 1 sample problem.
- Paired comparisons.
- Repeated measures: 1 sample.
- One way MANOVA.
- Two way MANOVA.
- Profile Analysis.

- Multivariate Multiple Regression.
- Discriminant Analysis.
- Clustering.
- Principal Components.
- Factor analysis.
- Canonical Correlations.

Basic structure of typical multivariate data set:

Case by variables: data in matrix. Each row is a case, each column is a variable.

Example: Fisher's iris data: 5 rows of 150 by 5 matrix:

Case #	Variety	Sepal Length	Sepal Width	Petal Length	Petal Width
1	Setosa	5.1	3.5	1.4	0.2
2	Setosa	4.9	3.0	1.4	0.2
⋮	⋮	⋮	⋮	⋮	⋮
51	Versicolor	7.0	3.2	4.7	1.4
⋮	⋮	⋮	⋮	⋮	⋮

Usual model: rows of data matrix are independent random variables.

Vector valued random variable: function $\mathbf{X}: \Omega \mapsto \mathbb{R}^p$ such that, writing $\mathbf{X} = (X_1, \ldots, X_p)^T$,

$$P(X_1 \leq x_1, \ldots, X_p \leq x_p)$$

defined for any const's $(x_1, \ldots, x_p)$.

Cumulative Distribution Function (CDF) of $\mathbf{X}$: function $F_{\mathbf{X}}$ on $\mathbb{R}^p$ defined by

$$F_{\mathbf{X}}(x_1, \ldots, x_p) = P(X_1 \leq x_1, \ldots, X_p \leq x_p)\,.$$

Defn: Distribution of rv $\mathbf{X}$ is **absolutely continuous** if there is a function f such that

$$P(\mathbf{X} \in A) = \int_A f(x)dx \qquad (1)$$

for any (Borel) set A. This is a p dimensional integral in general. Equivalently

$$F(x_1, \ldots, x_p) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_p} f(y_1, \ldots, y_p)\, dy_p, \ldots, dy_1 \,.$$

Defn: Any f satisfying (1) is a **density** of $\mathbf{X}$.

For most x F is differentiable at x and

$$\frac{\partial^p F(x)}{\partial x_1 \cdots \partial x_p} = f(x)\,.$$

Building Multivariate Models

Basic tactic: specify density of

$$\mathbf{X} = (X_1, \ldots, X_p)^T.$$

Tools: marginal densities, conditional densities, independence, transformation.

Marginalization: Simplest multivariate problem

$$\mathbf{X} = (X_1, \ldots, X_p), \qquad Y = X_1$$

(or in general Y is any X_j).

Theorem 1 *If* $\mathbf{X}$ *has density* $f(x_1, \ldots, x_p)$ *and* $q < p$ *then* $\mathbf{Y} = (X_1, \ldots, X_q)$ *has density*

$$f_{\mathbf{Y}}(x_1, \ldots, x_q) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, \ldots, x_p)\, dx_{q+1} \ldots dx_p$$

$f_{X_1,\ldots,X_q}$ is the **marginal** density of $X_1, \ldots, X_q$ and $f_{\mathbf{X}}$ the **joint** density of $\mathbf{X}$ but they are both just densities. "Marginal" just to distinguish from the joint density of $\mathbf{X}$.

Independence, conditional distributions

Def'n: Events A and B are independent if

$$P(AB) = P(A)P(B)\,.$$

(Notation: AB is the event that both A and B happen, also written $A \cap B$.)

Def'n: A_i, $i = 1, \ldots, p$ are **independent** if

$$P(A_{i_1} \cdots A_{i_r}) = \prod_{j=1}^{r} P(A_{i_j})$$

for any $1 \le i_1 < \cdots < i_r \le p$.

Def'n: $\mathbf{X}$ and $\mathbf{Y}$ are **independent** if

$$P(\mathbf{X} \in A; \mathbf{Y} \in B) = P(\mathbf{X} \in A)P(\mathbf{Y} \in B)$$

for all A and B.

Def'n: Rvs $\mathbf{X}_1, \ldots, \mathbf{X}_p$ **independent**:

$$P(\mathbf{X}_1 \in A_1, \cdots, \mathbf{X}_p \in A_p) = \prod P(\mathbf{X}_i \in A_i)$$

for any $A_1, \ldots, A_p$.

Theorem:

1. If $\mathbf{X}$ and $\mathbf{Y}$ are independent with joint density $f_{\mathbf{X},\mathbf{Y}}(x,y)$ then $\mathbf{X}$ and $\mathbf{Y}$ have densities $f_{\mathbf{X}}$ and $f_{\mathbf{Y}}$, and
$$f_{\mathbf{X},\mathbf{Y}}(x,y) = f_{\mathbf{X}}(x) f_{\mathbf{Y}}(y)\,.$$

2. If $\mathbf{X}$ and $\mathbf{Y}$ independent with marginal densities $f_{\mathbf{X}}$ and $f_{\mathbf{Y}}$ then $(\mathbf{X},\mathbf{Y})$ has joint density
$$f_{\mathbf{X},\mathbf{Y}}(x,y) = f_{\mathbf{X}}(x) f_{\mathbf{Y}}(y)\,.$$

3. If $(\mathbf{X},\mathbf{Y})$ has density $f(x,y)$ and there exist $g(x)$ and $h(y)$ st $f(x,y) = g(x)h(y)$ for (almost) **all** (x,y) then $\mathbf{X}$ and $\mathbf{Y}$ are independent with densities given by
$$f_{\mathbf{X}}(x) = g(x) / \int_{-\infty}^{\infty} g(u) du$$
$$f_{\mathbf{Y}}(y) = h(y) / \int_{-\infty}^{\infty} h(u) du\,.$$

Theorem: If $\mathbf{X}_1, \ldots, \mathbf{X}_p$ are independent and $\mathbf{Y}_i = g_i(\mathbf{X}_i)$ then $\mathbf{Y}_1, \ldots, \mathbf{Y}_p$ are independent. Moreover, $(\mathbf{X}_1, \ldots, \mathbf{X}_q)$ and $(\mathbf{X}_{q+1}, \ldots, \mathbf{X}_p)$ are independent.

Conditional densities

Conditional density of $\mathbf{Y}$ given $\mathbf{X} = x$:

$$f_{\mathbf{Y}|\mathbf{X}}(y|x) = f_{\mathbf{X},\mathbf{Y}}(x, y)/f_{\mathbf{X}}(x)\,;$$

in words "conditional = joint/marginal".

Change of Variables

Suppose $\mathbf{Y} = g(\mathbf{X}) \in \mathbb{R}^p$ with $\mathbf{X} \in \mathbb{R}^p$ having density $f_{\mathbf{X}}$. **Assume g is a one to one ("injective") map,** i.e., $g(x_1) = g(x_2)$ if and only if $x_1 = x_2$. Find $f_{\mathbf{Y}}$:

Step 1: Solve for x in terms of y: $x = g^{-1}(y)$.

Step 2: Use basic equation:

$$f_{\mathbf{Y}}(y)dy = f_{\mathbf{X}}(x)dx$$

and rewrite it in the form

$$f_{\mathbf{Y}}(y) = f_{\mathbf{X}}(g^{-1}(y))\frac{dx}{dy}$$

Interpretation of derivative $\frac{dx}{dy}$ when $p > 1$:

$$\frac{dx}{dy} = \left|\text{det}\left(\frac{\partial x_i}{\partial y_j}\right)\right|$$

which is the so called **Jacobian**.

Equivalent formula inverts the matrix:

$$f_{\mathbf{Y}}(y) = \frac{f_{\mathbf{X}}(g^{-1}(y))}{\left|\frac{dy}{dx}\right|}.$$

This notation means

$$\left|\frac{dy}{dx}\right| = \left|\det \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \cdots & \frac{\partial y_1}{\partial x_p} \\ & \vdots & & \\ \frac{\partial y_p}{\partial x_1} & \frac{\partial y_p}{\partial x_2} & \cdots & \frac{\partial y_p}{\partial x_p} \end{bmatrix}\right|$$

but with x replaced by the corresponding value of y, that is, replace x by $g^{-1}(y)$.

Example: The density

$$f_{\mathbf{X}}(x_1, x_2) = \frac{1}{2\pi}\exp\left\{-\frac{x_1^2 + x_2^2}{2}\right\}$$

is the **standard bivariate normal density**. Let $\mathbf{Y} = (Y_1, Y_2)$ where $Y_1 = \sqrt{X_1^2 + X_2^2}$ and $0 \le Y_2 < 2\pi$ is angle from the positive x axis to the ray from the origin to the point (X_1, X_2). I.e., $\mathbf{Y}$ is $\mathbf{X}$ in polar co-ordinates.

Solve for x in terms of y:

$$\begin{aligned} X_1 &= Y_1 \cos(Y_2) \\ X_2 &= Y_1 \sin(Y_2) \end{aligned}$$

so that

$$\begin{aligned} g(x_1, x_2) &= (g_1(x_1, x_2), g_2(x_1, x_2)) \\ &= (\sqrt{x_1^2 + x_2^2}, \text{argument}(x_1, x_2)) \\ g^{-1}(y_1, y_2) &= (g_1^{-1}(y_1, y_2), g_2^{-1}(y_1, y_2)) \\ &= (y_1 \cos(y_2), y_1 \sin(y_2)) \\ \left|\frac{dx}{dy}\right| &= \left|\det \begin{pmatrix} \cos(y_2) & -y_1 \sin(y_2) \\ \sin(y_2) & y_1 \cos(y_2) \end{pmatrix}\right| \\ &= y_1 \,. \end{aligned}$$

It follows that

$$\begin{aligned} f_{\mathbf{Y}}(y_1, y_2) &= \frac{1}{2\pi} \exp\left\{-\frac{y_1^2}{2}\right\} y_1 \times \\ & \quad 1(0 \le y_1 < \infty) 1(0 \le y_2 < 2\pi)\,. \end{aligned}$$

Next: marginal densities of Y_1, Y_2?

Factor $f_{\mathbf{Y}}$ as $f_{\mathbf{Y}}(y_1, y_2) = h_1(y_1)h_2(y_2)$ where

$$h_1(y_1) = y_1 e^{-y_1^2/2} 1(0 \le y_1 < \infty)$$

and

$$h_2(y_2) = 1(0 \le y_2 < 2\pi)/(2\pi)\,.$$

Then

$$\begin{aligned} f_{Y_1}(y_1) &= \int_{-\infty}^{\infty} h_1(y_1)h_2(y_2)\,dy_2 \\ &= h_1(y_1)\int_{-\infty}^{\infty} h_2(y_2)\,dy_2 \end{aligned}$$

so marginal density of Y_1 is a multiple of h_1. Multiplier makes $\int f_{Y_1} = 1$ but in this case

$$\int_{-\infty}^{\infty} h_2(y_2)\,dy_2 = \int_0^{2\pi} (2\pi)^{-1} dy_2 = 1$$

so that

$$f_{Y_1}(y_1) = y_1 e^{-y_1^2/2} 1(0 \le y_1 < \infty)\,.$$

(Special Weibull or Rayleigh distribution.)

Similarly

$$f_{Y_2}(y_2) = 1(0 \le y_2 < 2\pi)/(2\pi)$$

which is the **Uniform**$(0, 2\pi)$ density. Exercise: $W = Y_1^2/2$ has standard exponential distribution. Recall: by definition $U = Y_1^2$ has a χ^2 distribution on 2 degrees of freedom. Exercise: find χ_2^2 density.

Remark: easy to check $\int_0^\infty ye^{-y^2/2}dy = 1$.

Thus: have proved original bivariate normal density integrates to 1.

Put $I = \int_{-\infty}^\infty e^{-x^2/2}dx$. Get

$$\begin{aligned} I^2 &= \int_{-\infty}^\infty e^{-x^2/2}dx \int_{-\infty}^\infty e^{-y^2/2}dy \\ &= \int_{-\infty}^\infty \int_{-\infty}^\infty e^{-(x^2+y^2)/2}dydx \\ &= 2\pi. \end{aligned}$$

So $I = \sqrt{2\pi}$.

Linear Algebra Review

Notation:

- Vectors $x \in \mathbb{R}^n$ are column vectors

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$$

- An $m \times n$ matrix A has m rows, n columns and entries A_{ij}.

- Matrix and vector addition defined componentwise:

$$(A+B)_{ij} = A_{ij} + B_{ij}; \qquad (x+y)_i = x_i + y_i$$

- If A is $m \times n$ and B is $n \times r$ then AB is the $m \times r$ matrix

$$(AB)_{ij} = \sum_{k=1}^{n} A_{ik} B_{kj}$$

- The matrix I or sometimes $I_{n\times n}$ which is an $n \times n$ matrix with $I_{ii} = 1$ for all i and $I_{ij} = 0$ for any pair $i \neq j$ is called the $n \times n$ **identity matrix**.

- The **span** of a set of vectors $\{x_1, \ldots, x_p\}$ is the set of all vectors x of the form $x = \sum c_i x_i$. It is a vector space. The **column space** of a matrix, A, is the span of the set of columns of A. The **row space** is the span of the set of rows.

- A set of vectors $\{x_1, \ldots, x_p\}$ is **linearly independent** if $\sum c_i x_i = 0$ implies $c_i = 0$ for all i. The **dimension** of a vector space is the cardinality of the largest possible set of linearly independent vectors.

Defn: The **transpose**, A^T, of an $m \times n$ matrix A is the $n \times m$ matrix whose entries are given by

$$(A^T)_{ij} = A_{ji}$$

so that A^T is $n \times m$. We have

$$(A + B)^T = A^T + B^T$$

and

$$(AB)^T = B^T A^T .$$

Defn: **rank** of matrix A, rank(A): # of linearly independent columns of A. We have

$$\begin{aligned} \text{rank}(A) &= \text{dim(column space of } A) \\ &= \text{dim(row space of } A) \\ &= \text{rank}(A^T) \end{aligned}$$

If A is $m \times n$ then rank(A) $\leq$ min(m, n).

Matrix inverses

For now: all matrices square $n \times n$.

If there is a matrix B such that $BA = I_{n \times n}$ then we call B the inverse of A. If B exists it is unique and $AB = I$ and we write $B = A^{-1}$. The matrix A has an inverse if and only if $\text{rank}(A) = n$.

Inverses have the following properties:

$$(AB)^{-1} = B^{-1}A^{-1}$$

(if one side exists then so does the other) and

$$(A^T)^{-1} = (A^{-1})^T$$

Determinants

Again A is $n \times n$. The determinant if a function on the set of $n \times n$ matrices such that:

1. $\det(I) = 1$.

2. If A' is the matrix A with two columns interchanged then

$$\det(A') = -\det(A)\,.$$

(So: two equal columns implies $\det(A) = 0$.)

3. $\det(A)$ is a linear function of each column of A. If $A = (a_1, \ldots, a_n)$ with a_i denoting the ith column of the matrix then

$$\begin{aligned}\det(a_1, \ldots, &a_i + b_i, \ldots, a_n) \\ &= \det(a_1, \ldots, a_i, \ldots, a_n) \\ &\quad + \det(a_1, \ldots, b_i, \ldots, a_n)\end{aligned}$$

Here are some properties of the determinant:

1. $\det(A^T) = \det(A)$.

2. $\det(AB) = \det(A)\det(B)$.

3. $\det(A^{-1})$=$1/\det(A)$.

4. A is invertible if and only if $\det(A) \neq 0$ if and only if $\text{rank}(A) = n$.

5. Determinants can be computed (slowly) by expansion by minors.

Special Kinds of Matrices

1. A is symmetric if $A^T = A$.

2. A is orthogonal if $A^T = A^{-1}$ (or $AA^T = A^T A = I$).

3. A is idempotent if $AA \equiv A^2 = A$.

4. A is diagonal if $i \neq j$ implies $A_{ij} = 0$.

Inner Products, orthogonal, orthonormal vectors

Defn: Two vectors x and y are **orthogonal** if $x^T y = \sum x_i y_i = 0$.

Defn: The **inner product** or **dot product** of x and y is

$$< x, y >= x^T y = \sum x_i y_i$$

Defn: x and y are **orthogonal** if $x^T y = 0$.

Defn: The **norm** (or length) of x is $||x|| = (x^T x)^{1/2} = (\sum x_i^2)^{1/2}$

A is orthogonal if each column of A has length 1 and is orthogonal to each other column of A.

Quadratic Forms

Suppose A is an $n \times n$ matrix. The function

$$g(x) = x^T A x$$

is called a quadratic form. Now

$$\begin{aligned} g(x) &= \sum_{ij} A_{ij} x_i x_j \\ &= \sum_i A_{ii} x_i^2 + \sum_{i<j} (A_{ij} + A_{ji}) x_i x_j \end{aligned}$$

so that $g(x)$ depends only on the total $A_{ij} + A_{ji}$. In fact

$$x^T A x = x^T A^T x = x^T \left(\frac{A + A^T}{2} \right) x$$

Thus we will assume that A is symmetric.

Eigenvalues and eigenvectors

If A is $n \times n$ and $v \neq 0 \in \mathbb{R}^n$ and $\lambda \in \mathbb{R}$ such that

$$Av = \lambda v$$

then λ is **eigenvalue** (or characteristic or latent value) of A; v is corresponding **eigenvector**. Since $Av - \lambda v = (A - \lambda I)v = 0$ matrix $A - \lambda I$ is singular.

Therefore $\det(A - \lambda I) = 0$.

Conversely: if $A - \lambda I$ singular then there is $v \neq 0$ such that $(A - \lambda I)v = 0$.

Fact: $\det(A - \lambda I)$ is polynomial in λ of degree n.

Each root is an eigenvalue.

General A the roots could be multiple roots or complex valued.

Diagonalization

Matrix A is **diagonalized** by a non-singular matrix P if $P^{-1}AP \equiv D$ is diagonal.

If so then $AP = PD$ so each column of P is eigenvector of A with the ith column having eigenvalue D_{ii}.

Thus to be diagonalizable A must have n linearly independent eigenvectors.

Symmetric Matrices

If A is symmetric then

1. Every eigenvalue of A is real (not complex).

2. A is diagonalizable; columns of P may be taken unit length, mutually orthogonal: A is diagonalizable by an orthogonal matrix P; in symbols $P^T AP = D$.

3. Diagonal entries in $D =$ eigenvalues of A.

4. If $\lambda_1 \neq \lambda_2$ are two eigenvalues of A and v_1 and v_2 are corresponding eigenvectors then

$$v_1^T A v_2 = v_1^T \lambda_2 v_2 = \lambda_2 v_1^T v_2$$

and

$$\begin{aligned}(v_1^T A v_2) = (v_1^T A v_2)^T &= v_2^T A^T v_1 \\ &= v_2^T A v_1 = v_2^T \lambda_1 v_1 = \lambda_1 v_2^T v_1\end{aligned}$$

Since $(\lambda_1 - \lambda_2) v_1^T v_2 = 0$ and $\lambda_1 \neq \lambda_2$ we see $v_1^T v_2 = 0$. Eigenvectors corresponding to distinct eigenvalues are orthogonal.

Positive Definite Matrices

Defn: A symmetric matrix $\mathbf{A}$ is non-negative definite if $x^T\mathbf{A}x \ge 0$ for all x. It is positive definite if in addition $x^T\mathbf{A}x = 0$ implies $x = 0$.

$\mathbf{A}$ is non-negative definite iff all its eigenvalues are non-negative.

$\mathbf{A}$ is positive definite iff all eigenvalues positive.

A non-negative definite matrix has a symmetric non-negative definite square root. If

$$\mathbf{PDP}^T = \mathbf{A}$$

for $\mathbf{P}$ orthogonal and $\mathbf{D}$ diagonal then

$$\mathbf{A}^{1/2} = \mathbf{PD}^{1/2}\mathbf{P}^T$$

is symmetric, non-negative definite and

$$\mathbf{A}^{1/2}\mathbf{A}^{1/2} = \mathbf{A}$$

Here $\mathbf{D}^{1/2}$ is diagonal with

$$(\mathbf{D}^{1/2})_{ii} = (\mathbf{D}_{ii})^{1/2}.$$

Many other square roots possible. If $\mathbf{AA}^T = \mathbf{M}$ and $\mathbf{P}$ is orthogonal and $\mathbf{A}^* = \mathbf{AP}$ then $\mathbf{A}^*(\mathbf{A}^*)^T = \mathbf{M}$.

Orthogonal Projections

Suppose S vector subspace of $\mathbb{R}^n$, $a_1, \ldots, a_m$ basis for S. Given any $x \in \mathbb{R}^n$ there is a unique $y \in S$ which is closest to x; y minimizes

$$(x-y)^T(x-y)$$

over $y \in S$. Any y in S is of the form

$$y = c_1 a_1 + \cdots + c_m a_m = Ac$$

A, $n \times m$, columns $a_1, \ldots, a_m$; c column with ith entry c_i. Define

$$Q = A(A^T A)^{-1} A^T$$

(A has rank m so $A^T A$ is invertible.) Then

$$\begin{aligned}
&(x - Ac)^T(x - Ac) \\
&\quad = (x - Qx + Qx - Ac)^T(x - Qx + Qx - Ac) \\
&\quad = (x - Qx)^T(x - Qx) + (Qx - Ac)^T(x - Qx) \\
&\qquad + (x - Qx)^T(Qx - Ac) \\
&\qquad + (Qx - Ac)^T(Qx - Ac)
\end{aligned}$$

Note that $x - Qx = (I - Q)x$ and that

$$QAc = A(A^TA)^{-1}A^TAc = Ac$$

so that

$$Qx - Ac = Q(x - Ac)$$

Then

$$(Qx - Ac)^T(x - Qx) = (x - Ac)^TQ^T(I - Q)x$$

Since $Q^T = Q$ we see that

$$\begin{aligned} Q^T(I - Q) &= Q(I - Q) \\ &= Q - Q^2 \\ &= Q - A(A^TA)^{-1}A^TA(A^TA)^{-1}A^T \\ &= Q - Q = 0 \end{aligned}$$

This shows that

$$\begin{aligned} (x - Ac)^T(x - Ac) =& (x - Qx)^T(x - Qx) \\ &+ (Qx - Ac)^T(Qx - Ac) \end{aligned}$$

Choose Ac to minimize: minimize second term.

Achieved by making $Qx = Ac$.

Since $Qx = A(A^T A)^{-1} A^T x$ can take

$$c = (A^T A)^{-1} A^T x.$$

Summary: closest point y in S is

$$y = Qx = A(A^T A)^{-1} A^T x$$

call y the orthogonal projection of x onto S.

Notice that the matrix Q is idempotent:

$$Q^2 = Q$$

We call Qx the orthogonal projection of x on S because Qx is perpendicular to the residual $x - Qx = (I - Q)x$.

Partitioned Matrices

Suppose A_{11} $p \times r$ matrix, $A_{1,2}$ $p \times s$, $A_{2,1}$ $q \times r$ and $A_{2,2}$ $q \times s$. Make $(p+q) \times (r+s)$ matrix by putting A_{ij} in 2 by 2 matrix:

$$A = \left[\begin{array}{cc} A_{11} & A_{12} \\ A_{21} & A_{22} \end{array} \right]$$

For instance if

$$A_{11} = \left[\begin{array}{cc} 1 & 0 \\ 0 & 1 \end{array} \right]$$

$$A_{12} = \left[\begin{array}{c} 2 \\ 3 \end{array} \right]$$

$$A_{21} = \left[\begin{array}{cc} 4 & 5 \end{array} \right]$$

and

$$A_{22} = [6]$$

then

$$A = \left[\begin{array}{cc|c} 1 & 0 & 2 \\ 0 & 1 & 3 \\ \hline 4 & 5 & 6 \end{array} \right]$$

Lines indicate partitioning.

We can work with partitioned matrices just like ordinary matrices always making sure that in products we never change the order of multiplication of things.

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} + \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} = \begin{bmatrix} A_{11}+B_{11} & A_{12}+B_{12} \\ A_{21}+B_{21} & A_{22}+B_{22} \end{bmatrix}$$

and

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} = \begin{bmatrix} A_{11}B_{11}+A_{12}B_{21} & A_{11}B_{12}+A_{12}B_{22} \\ A_{21}B_{11}+A_{22}B_{21} & A_{21}B_{12}+A_{22}B_{22} \end{bmatrix}$$

Note partitioning of A and B must match.

Addition: dimensions of A_{ij} and B_{ij} must be the same.

Multiplication formula A_{12} must have as many columns as B_{21} has rows, etc.

In general: need $A_{ij}B_{jk}$ to make sense for each i, j, k.

Works with more than a 2 by 2 partitioning.

Defn: **block diagonal** matrix: partitioned matrix A for which $A_{ij} = 0$ if $i \neq j$. If

$$A = \left[\begin{array}{cc} A_{11} & 0 \\ 0 & A_{22} \end{array}\right]$$

then A is invertible iff each A_{ii} is invertible and then

$$A^{-1} = \left[\begin{array}{cc} A_{11}^{-1} & 0 \\ 0 & A_{22}^{-1} \end{array}\right]$$

Moreover $\det(A) = \det(A_{11})\det(A_{22})$. Similar formulas work for larger matrices.

Partitioned inverses. Suppose $\mathbf{A}$, $\mathbf{C}$ are symmetric positive definite. Look for inverse of

$$\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^T & \mathbf{C} \end{bmatrix}$$

of form

$$\begin{bmatrix} \mathbf{E} & \mathbf{F} \\ \mathbf{F}^T & \mathbf{G} \end{bmatrix}$$

Multiply to get equations

$$\begin{aligned} \mathbf{AE} + \mathbf{BF}^T &= \mathbf{I} \\ \mathbf{AF} + \mathbf{BG} &= \mathbf{0} \\ \mathbf{B}^T\mathbf{E} + \mathbf{CF}^T &= \mathbf{0} \\ \mathbf{B}^T\mathbf{F} + \mathbf{CG} &= \mathbf{I} \end{aligned}$$

Solve to get

$$\begin{aligned} \mathbf{F}^T &= -\mathbf{C}^{-1}\mathbf{B}^T\mathbf{E} \\ \mathbf{AE} - \mathbf{BC}^{-1}\mathbf{B}^T\mathbf{E} &= \mathbf{I} \\ \mathbf{E} &= (\mathbf{A} - \mathbf{BC}^{-1}\mathbf{B}^T)^{-1} \\ \mathbf{G} &= (\mathbf{C} - \mathbf{B}^T\mathbf{A}^{-1}\mathbf{B})^{-1} \end{aligned}$$

The Multivariate Normal Distribution

Defn: $Z \in \mathbb{R}^1 \sim N(0,1)$ iff

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\,.$$

Defn: $\mathbf{Z} \in \mathbb{R}^p \sim MVN_p(0,I)$ if and only if $\mathbf{Z} = (Z_1,\ldots,Z_p)^T$ with the Z_i independent and each $Z_i \sim N(0,1)$.

In this case according to our theorem

$$\begin{aligned} f_{\mathbf{Z}}(z_1,\ldots,z_p) &= \prod \frac{1}{\sqrt{2\pi}} e^{-z_i^2/2} \\ &= (2\pi)^{-p/2} \exp\{-z^T z/2\}\,; \end{aligned}$$

superscript t denotes matrix transpose.

Defn: $\mathbf{X} \in \mathbb{R}^p$ has a multivariate normal distribution if it has the same distribution as $\mathbf{AZ}+\boldsymbol{\mu}$ for some $\boldsymbol{\mu} \in \mathbb{R}^p$, some $p \times q$ matrix of constants $\mathbf{A}$ and $Z \sim MVN_q(0,I)$.

$p = q$, $\mathbf{A}$ singular: $\mathbf{X}$ does not have a density.

$\mathbf{A}$ invertible: derive multivariate normal density by change of variables:

$$\mathbf{X} = \mathbf{A}\mathbf{Z} + \boldsymbol{\mu} \Leftrightarrow \mathbf{Z} = \mathbf{A}^{-1}(\mathbf{X} - \boldsymbol{\mu})$$

$$\frac{\partial \mathbf{X}}{\partial \mathbf{Z}} = \mathbf{A} \qquad \frac{\partial \mathbf{Z}}{\partial \mathbf{X}} = \mathbf{A}^{-1}.$$

So

$$\begin{aligned} f_{\mathbf{X}}(x) &= f_{\mathbf{Z}}(\mathbf{A}^{-1}(x - \boldsymbol{\mu}))|\det(\mathbf{A}^{-1})| \\ &= \frac{\exp\{-(x - \boldsymbol{\mu})^T(\mathbf{A}^{-1})^T\mathbf{A}^{-1}(x - \boldsymbol{\mu})/2\}}{(2\pi)^{p/2}|\det \mathbf{A}|}. \end{aligned}$$

Now define $\boldsymbol{\Sigma} = \mathbf{A}\mathbf{A}^T$ and notice that

$$\boldsymbol{\Sigma}^{-1} = (\mathbf{A}^T)^{-1}\mathbf{A}^{-1} = (\mathbf{A}^{-1})^T\mathbf{A}^{-1}$$

and

$$\det \boldsymbol{\Sigma} = \det \mathbf{A} \det \mathbf{A}^T = (\det \mathbf{A})^2.$$

Thus $f_{\mathbf{X}}$ is

$$\frac{\exp\{-(x - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(x - \boldsymbol{\mu})/2\}}{(2\pi)^{p/2}(\det \boldsymbol{\Sigma})^{1/2}};$$

the $MVN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ density. Note density is the same for all $\mathbf{A}$ such that $\mathbf{A}\mathbf{A}^T = \boldsymbol{\Sigma}$. This justifies the notation $MVN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

For which $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ is this a density?

Any $\boldsymbol{\mu}$ but if $x \in \mathbb{R}^p$ then

$$\begin{aligned} x^T \boldsymbol{\Sigma} x &= x^T \mathbf{A}\mathbf{A}^T x \\ &= (\mathbf{A}^T x)^T (\mathbf{A}^T x) \\ &= \sum_1^p y_i^2 \ge 0 \end{aligned}$$

where $y = \mathbf{A}^T x$. Inequality strict except for $y = 0$ which is equivalent to $x = 0$. Thus $\boldsymbol{\Sigma}$ is a positive definite symmetric matrix.

Conversely, if $\boldsymbol{\Sigma}$ is a positive definite symmetric matrix then there is a square invertible matrix $\mathbf{A}$ such that $\mathbf{A}\mathbf{A}^T = \boldsymbol{\Sigma}$ so that there is a $MVN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution. ($\mathbf{A}$ can be found via the Cholesky decomposition, e.g.)

When $\mathbf{A}$ is singular $\mathbf{X}$ will not have a density: $\exists a$ such that $P(a^T \mathbf{X} = a^T \boldsymbol{\mu}) = 1$; $\mathbf{X}$ is confined to a hyperplane.

Still true: distribution of $\mathbf{X}$ depends only on $\boldsymbol{\Sigma} = \mathbf{A}\mathbf{A}^T$: if $\mathbf{A}\mathbf{A}^T = BB^T$ then $\mathbf{A}\mathbf{Z} + \boldsymbol{\mu}$ and $B\mathbf{Z} + \boldsymbol{\mu}$ have the same distribution.

Expectation, moments

Defn: If $\mathbf{X} \in \mathbb{R}^p$ has density f then

$$\text{E}(g(\mathbf{X})) = \int g(x) f(x)\, dx\,.$$

any g from $\mathbb{R}^p$ to $\mathbb{R}$.

FACT: if $Y = g(X)$ for a smooth g (mapping $\mathbb{R} \to \mathbb{R}$)

$$\begin{aligned}\text{E}(Y) &= \int y f_Y(y)\, dy \\ &= \int g(x) f_Y(g(x)) g'(x)\, dx \\ &= \text{E}(g(X))\end{aligned}$$

by change of variables formula for integration. This is good because otherwise we might have two different values for $\text{E}(e^X)$.

Linearity: $\text{E}(aX + bY) = a\text{E}(X) + b\text{E}(Y)$ for real X and Y.

Defn: The r^{th} moment (about the origin) of a real rv X is $\mu'_r = \text{E}(X^r)$ (provided it exists). We generally use μ for $\text{E}(X)$.

Defn: The r^{th} central moment is

$$\mu_r = \text{E}[(X - \mu)^r]$$

We call $\sigma^2 = \mu_2$ the variance.

Defn: For an $\mathbb{R}^p$ valued random vector $\mathbf{X}$

$$\boldsymbol{\mu}_{\mathbf{X}} = \text{E}(\mathbf{X})$$

is the vector whose i^{th} entry is $\text{E}(X_i)$ (provided all entries exist).

Fact: same idea used for random matrices.

Defn: The $(p \times p)$ variance covariance matrix of $\mathbf{X}$ is

$$\text{Var}(\mathbf{X}) = \text{E}\left[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T\right]$$

which exists provided each component X_i has a finite second moment.

Example moments: If $Z \sim N(0,1)$ then

$$\begin{aligned}
\mathrm{E}(Z) &= \int_{-\infty}^{\infty} z e^{-z^2/2} dz/\sqrt{2\pi} \\
&= \left. \frac{-e^{-z^2/2}}{\sqrt{2\pi}} \right|_{-\infty}^{\infty} \\
&= 0
\end{aligned}$$

and (integrating by parts)

$$\begin{aligned}
\mathrm{E}(Z^r) &= \int_{-\infty}^{\infty} z^r e^{-z^2/2} dz/\sqrt{2\pi} \\
&= \left. \frac{-z^{r-1} e^{-z^2/2}}{\sqrt{2\pi}} \right|_{-\infty}^{\infty} \\
&\quad + (r-1) \int_{-\infty}^{\infty} z^{r-2} e^{-z^2/2} dz/\sqrt{2\pi}
\end{aligned}$$

so that

$$\mu_r = (r-1)\mu_{r-2}$$

for $r \ge 2$. Remembering that $\mu_1 = 0$ and

$$\mu_0 = \int_{-\infty}^{\infty} z^0 e^{-z^2/2} dz/\sqrt{2\pi} = 1$$

we find that

$$\mu_r = \begin{cases} 0 & r \text{ odd} \\ (r-1)(r-3)\cdots 1 & r \text{ even}. \end{cases}$$

If now $X \sim N(\mu, \sigma^2)$, that is, $X \sim \sigma Z + \mu$, then $\mathsf{E}(X) = \sigma \mathsf{E}(Z) + \mu = \mu$ and

$$\mu_r(X) = \mathsf{E}[(X-\mu)^r] = \sigma^r \mathsf{E}(Z^r)$$

In particular, we see that our choice of notation $N(\mu, \sigma^2)$ for the distribution of $\sigma Z + \mu$ is justified; σ is indeed the variance.

Similarly for $\mathbf{X} \sim MVN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ we have $\mathbf{X} = \mathbf{AZ} + \mu$ with $\mathbf{Z} \sim MVN(0, I)$ and

$$\mathsf{E}(\mathbf{X}) = \boldsymbol{\mu}$$

and

$$\begin{aligned}
\mathsf{Var}(\mathbf{X}) &= \mathsf{E}\left\{(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})^T\right\} \\
&= \mathsf{E}\left\{\mathbf{AZ}(\mathbf{AZ})^T\right\} \\
&= \mathbf{A}\mathsf{E}(\mathbf{ZZ}^T)\mathbf{A}^T \\
&= \mathbf{A}I\mathbf{A}^T = \boldsymbol{\Sigma}\,.
\end{aligned}$$

Note use of easy calculation: $\mathsf{E}(\mathbf{Z}) = 0$ and

$$\mathsf{Var}(\mathbf{Z}) = \mathsf{E}(\mathbf{ZZ}^T) = I\,.$$

Moments and independence

Theorem: If $X_1, \ldots, X_p$ are independent and each X_i is integrable then $X = X_1 \cdots X_p$ is integrable and

$$\mathsf{E}(X_1 \cdots X_p) = \mathsf{E}(X_1) \cdots \mathsf{E}(X_p)\,.$$

Moment Generating Functions

Defn: The moment generating function of a real valued X is

$$M_X(t) = \mathsf{E}(e^{tX})$$

defined for those real t for which the expected value is finite.

Defn: The moment generating function of $\mathbf{X} \in \mathbb{R}^p$ is

$$M_{\mathbf{X}}(u) = \mathsf{E}[e^{u^T\mathbf{X}}]$$

defined for those vectors u for which the expected value is finite.

Example: If $Z \sim N(0,1)$ then

$$\begin{aligned}
M_Z(t) &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{tz-z^2/2}dz \\
&= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-(z-t)^2/2+t^2/2}dz \\
&= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-u^2/2+t^2/2}du \\
&= e^{t^2/2}
\end{aligned}$$

Theorem: $(p = 1)$ If M is finite for all t in a neighbourhood of $\mathbf{0}$ then

1. Every moment of X is finite.
2. M is C^∞ (in fact M is analytic).
3. $\mu_k' = \frac{d^k}{dt^k}M_X(0)$.

Note: C^∞ means has continuous derivatives of all orders. Analytic means has convergent power series expansion in neighbourhood of each $t \in (-\epsilon, \epsilon)$.

The proof, and many other facts about mgfs, rely on techniques of complex variables.

Characterization & MGFs

Theorem: Suppose $\mathbf{X}$ and $\mathbf{Y}$ are $\mathbb{R}^p$ valued random vectors such that

$$M_{\mathbf{X}}(\mathbf{u}) = M_{\mathbf{Y}}(\mathbf{u})$$

for $\mathbf{u}$ in some open neighbourhood of $\mathbf{0}$ in $\mathbb{R}^p$. Then $\mathbf{X}$ and $\mathbf{Y}$ have the same distribution.

The proof relies on techniques of complex variables.

MGFs and Sums

If $X_1, \ldots, X_p$ are independent and $Y = \sum X_i$ then mgf of Y is product mgfs of individual X_i:

$$\mathrm{E}(e^{tY}) = \prod_i \mathrm{E}(e^{tX_i})$$

or $M_Y = \prod M_{X_i}$. (Also for multivariate X_i.)

Example: If $Z_1, \ldots, Z_p$ are independent $N(0,1)$ then

$$\begin{aligned}\mathrm{E}(e^{\sum a_i Z_i}) &= \prod_i \mathrm{E}(e^{a_i Z_i}) \\ &= \prod_i e^{a_i^2/2} \\ &= \exp(\sum a_i^2/2)\end{aligned}$$

Conclusion: If $\mathbf{Z} \sim MNV_p(0, I)$ then

$$M_Z(u) = \exp(\sum u_i^2/2) = \exp(\mathbf{u}^T\mathbf{u}/2).$$

Example: If $X \sim N(\mu, \sigma^2)$ then $X = \sigma Z + \mu$ and

$$M_X(t) = \mathrm{E}(e^{t(\sigma Z+\mu)}) = e^{t\mu} e^{\sigma^2 t^2/2}.$$

Theorem: Suppose $\mathbf{X} = \mathbf{A}\mathbf{Z} + \boldsymbol{\mu}$ and $\mathbf{Y} = \mathbf{A}^*\mathbf{Z}^* + \boldsymbol{\mu}^*$ where $\mathbf{Z} \sim MVN_p(0, I)$ and $\mathbf{Z}^* \sim MVN_q(0, I)$. Then $\mathbf{X}$ and $\mathbf{Y}$ have the same distribution if and only iff the following two conditions hold:

1. $\boldsymbol{\mu} = \boldsymbol{\mu}^*$.
2. $\mathbf{A}\mathbf{A}^T = \mathbf{A}^*(\mathbf{A}^*)^T$.

Alternatively: if $\mathbf{X}$, $\mathbf{Y}$ each MVN then $\text{E}(\mathbf{X}) = \text{E}(\mathbf{Y})$ and $\text{Var}(\mathbf{X}) = \text{Var}(\mathbf{Y})$ imply that $\mathbf{X}$ and $\mathbf{Y}$ have the same distribution.

Proof: If 1 and 2 hold the mgf of $\mathbf{X}$ is

$$
\begin{aligned}
\text{E}\left(e^{t^T\mathbf{X}}\right) &= \text{E}\left(e^{t^T(\mathbf{A}\mathbf{Z}+\boldsymbol{\mu}}\right) \\
&= e^{t^T\boldsymbol{\mu}}\text{E}\left(e^{(\mathbf{A}^Tt)^T\mathbf{Z}}\right) \\
&= e^{t^T\boldsymbol{\mu}+(\mathbf{A}^Tt)^T(\mathbf{A}^Tt)} \\
&= e^{t^T\boldsymbol{\mu}+t^T\boldsymbol{\Sigma}t}
\end{aligned}
$$

Thus $M_{\mathbf{X}} = M_{\mathbf{Y}}$. Conversely if $\mathbf{X}$ and $\mathbf{Y}$ have the same distribution then they have the same mean and variance.

Thus mgf is determined by $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$.

Theorem: If $\mathbf{X} \sim MVN_p(\mu, \boldsymbol{\Sigma})$ then there is $\mathbf{A}$ a $p \times p$ matrix such that $\mathbf{X}$ has same distribution as $\mathbf{AZ} + \boldsymbol{\mu}$ for $\mathbf{Z} \sim MVN_p(0, I)$.

We may assume that $\mathbf{A}$ is symmetric and non-negative definite, or that $\mathbf{A}$ is upper triangular, or that Ba is lower triangular.

Proof: Pick any $\mathbf{A}$ such that $\mathbf{A}\mathbf{A}^T = \boldsymbol{\Sigma}$ such as $\mathbf{P}\mathbf{D}^{1/2}\mathbf{P}^T$ from the spectral decomposition. Then $\mathbf{AZ} + \boldsymbol{\mu} \sim MVN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

From the symmetric square root can produce an upper triangular square root by the Gram Schmidt process: if $\mathbf{A}$ has rows $a_1^T, \ldots, a_p^T$ then let v_p be $a_p/\sqrt{a_p^T a_p}$. Choose v_{p-1} proportional to $a_{p-1} - bv_p$ where $b = a_{p-1}^T v_p$ so that v_{p-1} has unit length. Continue in this way; you automatically get $a_j^T v_k = 0$ if $j < k$. If $\mathbf{P}$ has columns $v_1, \ldots, v_p$ then $\mathbf{P}$ is orthogonal and $\mathbf{AP}$ is an upper triangular square root of $\boldsymbol{\Sigma}$.

Variances, Covariances, Correlations

Defn: The covariance between $\mathbf{X}$ and $\mathbf{Y}$ is

$$\text{Cov}(\mathbf{X},\mathbf{Y}) = \text{E}\left\{(\mathbf{X}-\boldsymbol{\mu}_{\mathbf{X}})(\mathbf{Y}-\boldsymbol{\mu}_{\mathbf{Y}})^T\right\}$$

This is a matrix.

Properties:

- $\text{Cov}(\mathbf{X},\mathbf{X}) = \text{Var}(\mathbf{X})$.
- Cov is bilinear:

  $$\begin{aligned}\text{Cov}(\mathbf{A}\mathbf{X}+\mathbf{B}\mathbf{W},\mathbf{Y}) =& \mathbf{A}\text{Cov}(\mathbf{X},\mathbf{Y}) \\ &+ \mathbf{B}\text{Cov}(\mathbf{W},\mathbf{Y})\end{aligned}$$

  and

  $$\begin{aligned}\text{Cov}(\mathbf{X},\mathbf{C}\mathbf{Y}+\mathbf{D}\mathbf{Z}) =& \text{Cov}(\mathbf{X},\mathbf{Y})\mathbf{C}^T \\ &+ \text{Cov}(\mathbf{X},\mathbf{Z})\mathbf{D}^T\end{aligned}$$

Properties of the MVN distribution

1: All margins are multivariate normal: if

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix}$$

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$$

and

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}$$

then $\mathbf{X} \sim MVN(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \Rightarrow \mathbf{X}_1 \sim MVN(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$.

2: $\mathbf{MX} + \boldsymbol{\nu} \sim MVN(\mathrm{M}\boldsymbol{\mu} + \boldsymbol{\nu}, \mathrm{M}\Sigma\mathrm{M}^T)$: affine transformation of MVN is normal.

3: If

$$\boldsymbol{\Sigma}_{12} = \text{Cov}(\mathbf{X}_1, \mathbf{X}_2) = \mathbf{0}$$

then $\mathbf{X}_1$ and $\mathbf{X}_2$ are independent.

4: All conditionals are normal: the conditional distribution of $\mathbf{X}_1$ given $\mathbf{X}_2 = x_2$ is $MVN(\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(x_2 - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21})$

Proof of ($\mathbf{1}$): If $\mathbf{X} = \mathbf{AZ} + \boldsymbol{\mu}$ then

$$\mathbf{X}_1 = [I|\mathbf{0}]\,\mathbf{X}$$

for I the identity matrix of correct dimension.

So

$$\mathbf{X}_1 = ([I|\mathbf{0}]\,\mathbf{A})\,\mathbf{Z} + [I|\mathbf{0}]\,\boldsymbol{\mu}$$

Compute mean and variance to check rest.

Proof of ($\mathbf{2}$): If $\mathbf{X} = \mathbf{AZ} + \boldsymbol{\mu}$ then

$$\mathbf{MX} + \boldsymbol{\nu} = \mathbf{MAZ} + \boldsymbol{\nu} + \mathbf{M}\boldsymbol{\mu}$$

Proof of ($\mathbf{3}$): If

$$\mathbf{u} = \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \end{bmatrix}$$

then

$$M_{\mathbf{X}}(u) = M_{\mathbf{X}_1}(\mathbf{u}_1) M_{\mathbf{X}_2}(\mathbf{u}_2)$$

Proof of (**4**): first case: assume $\mathbf{\Sigma}_{22}$ has an inverse.

Define

$$\mathbf{W} = \mathbf{X}_1 - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{X}_2$$

Then

$$\begin{bmatrix} \mathbf{W} \\ \mathbf{X}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{I} & -\mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix}$$

Thus $(\mathbf{W}, \mathbf{X}_2)^T$ is $MVN(\boldsymbol{\mu}_1 - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\boldsymbol{\mu}_2, \mathbf{\Sigma}^*)$ where

$$\mathbf{\Sigma}^* = \begin{bmatrix} \mathbf{\Sigma}_{11} - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma}_{22} \end{bmatrix}$$

Now joint density of $\mathbf{W}$ and $\mathbf{X}$ factors

$$f_{\mathbf{W},\mathbf{X}_2}(w, x_2) = f_{\mathbf{W}}(w) f_{\mathbf{X}_2}(x_2)$$

By change of variables joint density of $\mathbf{X}$ is

$$f_{\mathbf{X}_1,\mathbf{X}_2}(x_1, x_2) = c f_{\mathbf{W}}(x_1 - \mathbf{M}x_2) f_{\mathbf{X}_2}(x_2)$$

where $c = 1$ is the constant Jacobian of the linear transformation from $(\mathbf{W}, \mathbf{X}_2)$ to $(\mathbf{X}_1, \mathbf{X}_2)$ and

$$\mathbf{M} = \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}$$

Thus conditional density of $\mathbf{X}_1$ given $\mathbf{X}_2 = x_2$ is

$$\frac{f_{\mathbf{W}}(x_1 - \mathbf{M}x_2) f_{\mathbf{X}_2}(x_2)}{f_{\mathbf{X}_2}(x_2)} = f_{\mathbf{W}}(x_1 - \mathbf{M}x_2)$$

As a function of x_1 this density has the form of the advertised multivariate normal density.

Specialization to bivariate case:

Write

$$\boldsymbol{\Sigma} = \left[\begin{array}{cc} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{array} \right]$$

where we define

$$\rho = \frac{\text{Cov}(X_1, X_2)}{\sqrt{\text{Var}(X_1)\text{Var}(X_2)}}$$

Note that

$$\sigma_i^2 = \text{Var}(X_i)$$

Then

$$W = X_1 - \rho\frac{\sigma_1}{\sigma_2}X_2$$

is independent of X_2. The marginal distribution of W is $N(\mu_1 - \rho\sigma_1\mu_2/\sigma_2, \tau^2)$ where

$$\begin{aligned} \tau^2 = & \text{Var}(X_1) - 2\rho\frac{\sigma_1}{\sigma_2}\text{Cov}(X_1, X_2) \\ & + \left(\rho\frac{\sigma_1}{\sigma_2}\right)^2 \text{Var}(X_2) \end{aligned}$$

This simplifies to

$$\sigma_1^2(1-\rho^2)$$

Notice that it follows that

$$-1 \le \rho \le 1$$

More generally: any X and Y:

$$\begin{aligned} 0 &\le \text{Var}(X-\lambda Y) \\ &= \text{Var}(X) - 2\lambda\text{Cov}(X,Y) + \lambda^2\text{Var}(Y) \end{aligned}$$

RHS is minimized at

$$\lambda = \frac{\text{Cov}(X,Y)}{\text{Var}(Y)}$$

Minimum value is

$$0 \le \text{Var}(X)(1-\rho_{XY}^2)$$

where

$$\rho_{XY} = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$$

defines the correlation between X and Y.

Multiple Correlation

Now suppose $\mathbf{X}_2$ is scalar but $\mathbf{X}_1$ is vector.

Defn: Multiple correlation between $\mathbf{X}_1$ and X_2

$$R^2(\mathbf{X}_1, X_2) = \max |\rho_{\mathbf{a}^T\mathbf{X}_1,X_2}|^2$$

over all $\mathbf{a} \neq 0$.

Thus: maximize

$$\frac{\text{Cov}^2(\mathbf{a}^T\mathbf{X}_1, X_2)}{\text{Var}(\mathbf{a}^T\mathbf{X}_1)\text{Var}(X_2)} = \frac{\mathbf{a}^T\mathbf{\Sigma}_{12}\mathbf{\Sigma}_{21}\mathbf{a}}{\left(\mathbf{a}^T\mathbf{\Sigma}_{11}\mathbf{a}\right)\mathbf{\Sigma}_{22}}$$

Put $b = \mathbf{\Sigma}_{11}^{1/2}\mathbf{a}$. For $\mathbf{\Sigma}_{11}$ invertible problem is equivalent to maximizing

$$\frac{\mathbf{b}^T\mathbf{Q}\mathbf{b}}{\mathbf{b}^T\mathbf{b}}$$

where

$$\mathbf{Q} = \mathbf{\Sigma}_{11}^{-1/2}\mathbf{\Sigma}_{12}\mathbf{\Sigma}_{21}\mathbf{\Sigma}_{11}^{-1/2}$$

Solution: find largest eigenvalue of $\mathbf{Q}$.

Note

$$\mathbf{Q} = \mathbf{v}\mathbf{v}^T$$

where

$$\mathbf{v} = \boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{\Sigma}_{12}$$

is a vector. Set

$$\mathbf{v}\mathbf{v}^T\mathbf{x} = \lambda\mathbf{x}$$

and multiply by $\mathbf{v}^T$ to get

$$\mathbf{v}^T\mathbf{x} = 0 \text{ or } \lambda = \mathbf{v}^T\mathbf{v}$$

If $\mathbf{v}^T\mathbf{x} = 0$ then we see $\lambda = 0$ so largest eigenvalue is $\mathbf{v}^T\mathbf{v}$.

Summary: maximum squared correlation is

$$R^2(\mathbf{X}_1, X_2) = \frac{\mathbf{v}^T\mathbf{v}}{\boldsymbol{\Sigma}_{22}} = \frac{\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}}{\boldsymbol{\Sigma}_{22}}$$

Achieved when eigenvector is $\mathbf{x} = \mathbf{v} = \mathbf{b}$ so

$$\mathbf{a} = \boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}$$

Notice: since R^2 is squared correlation between two scalars ($\mathbf{a}^t\mathbf{X}_1$ and X_2) we have

$$0 \le R^2 \le 1$$

Equals 1 iff X_2 is linear combination of $\mathbf{X}_1$.

Correlation matrices, partial correlations:

Correlation between two scalars X and Y is

$$\rho_{XY} = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$$

If $\mathbf{X}$ has variance $\boldsymbol{\Sigma}$ then the correlation matrix of $\mathbf{X}$ is $\mathbf{R}_{\mathbf{X}}$ with entries

$$R_{ij} = \frac{\text{Cov}(X_i, X_j)}{\sqrt{\text{Var}(X_i)\text{Var}(X_j)}} = \frac{\Sigma_{ij}}{\sqrt{\Sigma_{ii}\Sigma_{jj}}}$$

If $\mathbf{X}_1, \mathbf{X}_2$ are MVN with the usual partitioned variance covariance matrix then the conditional variance of $\mathbf{X}_1$ given $\mathbf{X}_2$ is

$$\boldsymbol{\Sigma}_{11\cdot 2} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$$

From this define **partial correlation matrix**

$$\mathbf{R}_{11\cdot 2} = \frac{(\boldsymbol{\Sigma}_{11\cdot 2})_{ij}}{\sqrt{\boldsymbol{\Sigma}_{11\cdot 2})_{ii}\boldsymbol{\Sigma}_{11\cdot 2})_{jj}}}$$

Note: these are used even when $\mathbf{X}_1, \mathbf{X}_2$ are NOT MVN

Likelihood Methods of Inference

Given data X with model $\{f_\theta(x); \theta \in \Theta\}$:

Definition: The likelihood function is map L: domain Θ, values given by

$$L(\theta) = f_\theta(X)$$

Key Point: think about how the density depends on θ not about how it depends on X.

Notice: X, observed value of the data, has been plugged into the formula for density.

We use likelihood for most inference problems:

1. Point estimation: we must compute an estimate $\hat{\theta} = \hat{\theta}(X)$ which lies in Θ. The **maximum likelihood estimate (MLE)** of θ is the value $\hat{\theta}$ which maximizes $L(\theta)$ over $\theta \in \Theta$ if such a $\hat{\theta}$ exists.

2. Point estimation of a function of θ: we must compute an estimate $\hat{\phi} = \hat{\phi}(X)$ of $\phi = g(\theta)$. We use $\hat{\phi} = g(\hat{\theta})$ where $\hat{\theta}$ is the MLE of θ.

3. Interval (or set) estimation. We must compute a set $C = C(X)$ in Θ which we think will contain θ_0. We will use
$$\{\theta \in \Theta : L(\theta) > c\}$$
for a suitable c.

4. Hypothesis testing: decide whether or not $\theta_0 \in \Theta_0$ where $\Theta_0 \subset \Theta$. We base our decision on the likelihood ratio
$$\frac{\sup\{L(\theta); \theta \in \Theta_0\}}{\sup\{L(\theta); \theta \in \Theta \setminus \Theta_0\}}$$

Maximum Likelihood Estimation

To find MLE maximize L.

Typical function maximization problem:

Set gradient of L equal to 0

Check root is maximum, not minimum or saddle point.

Often L is product of n terms (given n independent observations).

Much easier to work with logarithm of L: log of product is sum and logarithm is monotone increasing.

Definition: The **Log Likelihood** function is

$$\ell(\theta) = \log\{L(\theta)\}\,.$$

Samples from MVN Population

Simplest problem: collect replicate measurements $\mathbf{X}_1, \ldots, \mathbf{X}_n$ from single population.

Model: X_i are iid $MVN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Parameters (θ): $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Parameter space: $\boldsymbol{\mu} \in \mathbb{R}^p$ and $\boldsymbol{\Sigma}$ is some positive definite $p \times p$ matrix.

Log likelihood is

$$\begin{aligned}\ell(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = -np\log(\pi)/2 - n\log\det\boldsymbol{\Sigma}/2 \\ - \sum(\mathbf{X}_i - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{X}_i - \boldsymbol{\mu})/2\end{aligned}$$

Take derivatives.

$$\begin{aligned}\frac{\partial \ell}{\partial \boldsymbol{\mu}} &= \boldsymbol{\Sigma}^{-1}\left\{\sum(\mathbf{X}_i - \boldsymbol{\mu})\right\} \\ &= n\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{X}} - \boldsymbol{\mu})\end{aligned}$$

where $\bar{\mathbf{X}} = \sum \mathbf{X}_i/n$.

Second derivative wrt $\boldsymbol{\mu}$ is a matrix:

$$-n\boldsymbol{\Sigma}^{-1}$$

Fact: if second derivative matrix is negative definite at critical point then critical point is a maximum.

Fact: if second derivative matrix is negative definite everywhere then function is concave; no more than 1 critical point.

Summary: ℓ is maximized at

$$\hat{\boldsymbol{\mu}} = \bar{\mathbf{X}}$$

(regardless of choice of $\boldsymbol{\Sigma}$).

More difficult: differentiate ℓ wrt $\boldsymbol{\Sigma}$.

Somewhat simpler: set $\mathbf{D} = \boldsymbol{\Sigma}^{-1}$

First derivative wrt $\mathbf{D}$ is matrix with entries

$$\frac{\partial \ell}{\partial \mathbf{D}_{ij}}$$

Warning: method used ignores symmetry of $\boldsymbol{\Sigma}$.

Need: derivative of two functions:

$$\frac{\partial \log \det \mathbf{A}}{\partial \mathbf{A}} = \mathbf{A}^{-1}$$

and

$$\frac{\partial \mathbf{x}^T \mathbf{A} \mathbf{x}}{\partial \mathbf{A}} = \mathbf{x}\mathbf{x}^T$$

Fact: ij^{th} entry of $\mathbf{A}^{-1}$ is

$$(-1)^{i+j} \frac{\det(\mathbf{A}^{(ij)})}{\det \mathbf{A}}$$

where $\mathbf{A}^{(ij)}$ denotes matrix obtained from $\mathbf{A}$ by removing column j and row i.

Fact: $\det(\mathbf{A}) = \sum_k (-1)^{i+k} A_{ik} \det(\mathbf{A}^{(ik)})$; expansion by minors.

Conclusion

$$\frac{\partial \log \det \mathbf{A}}{\partial A_{ij}} = (\mathbf{A}^{-1})_{ij}$$

and

$$\frac{\partial \log \det \mathbf{A}^{-1}}{\partial A_{ij}} = -(\mathbf{A}^{-1})_{ij}$$

Implication

$$\frac{\partial \ell}{\partial \mathbf{D}} = -n\boldsymbol{\Sigma}/2 - \sum_i (\mathbf{X}_i - \boldsymbol{\mu})(\mathbf{X}_i - \boldsymbol{\mu})^T/2$$

Set $=0$ and find only critical point is

$$\hat{\boldsymbol{\Sigma}} = \sum_i (\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^T/n$$

Usual sample covariance matrix is

$$\mathbf{S} = \sum_i (\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^T/(n-1)$$

Properties of MLEs:

1) $\bar{\mathbf{X}} \sim MVN_p(\boldsymbol{\mu}, n^{-1}\boldsymbol{\Sigma})$

2) $\mathrm{E}(\mathbf{S}) = \boldsymbol{\Sigma}$.

Distribution of $\mathbf{S}$? Joint distribution of $\bar{\mathbf{X}}$ and $\mathbf{S}$?

Univariate Normal samples: Distribution Theory

Theorem: Suppose $X_1, \ldots, X_n$ are independent $N(\mu, \sigma^2)$ random variables. Then

1. $\bar{X}$ (sample mean)and s^2 (sample variance) independent.

2. $n^{1/2}(\bar{X} - \mu)/\sigma \sim N(0, 1)$.

3. $(n - 1)s^2/\sigma^2 \sim \chi^2_{n-1}$.

4. $n^{1/2}(\bar{X} - \mu)/s \sim t_{n-1}$.

Proof: Let $Z_i = (X_i - \mu)/\sigma$.

Then $Z_1, \ldots, Z_p$ are independent $N(0,1)$.

So $Z = (Z_1, \ldots, Z_p)^T$ is multivariate standard normal.

Note that $\bar{X} = \sigma\bar{Z} + \mu$ and $s^2 = \sum(X_i - \bar{X})^2/(n-1) = \sigma^2 \sum(Z_i - \bar{Z})^2/(n-1)$ Thus

$$\frac{n^{1/2}(\bar{X}-\mu)}{\sigma} = n^{1/2}\bar{Z}$$

$$\frac{(n-1)s^2}{\sigma^2} = \sum(Z_i - \bar{Z})^2$$

and

$$T = \frac{n^{1/2}(\bar{X}-\mu)}{s} = \frac{n^{1/2}\bar{Z}}{s_Z}$$

where $(n-1)s_Z^2 = \sum(Z_i - \bar{Z})^2$.

So: reduced to $\mu = 0$ and $\sigma = 1$.

Step 1: Define

$$Y = (\sqrt{n}\bar{Z}, Z_1 - \bar{Z}, \ldots, Z_n - \bar{Z})^T \,.$$

(So Y has dimension $n+1$.) Now

$$Y = \begin{bmatrix} \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \cdots & \frac{1}{\sqrt{n}} \\ 1-\frac{1}{n} & -\frac{1}{n} & \cdots & -\frac{1}{n} \\ -\frac{1}{n} & 1-\frac{1}{n} & \cdots & -\frac{1}{n} \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix} \begin{bmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_n \end{bmatrix}$$

or letting $\mathbf{M}$ denote the matrix

$$Y = \mathbf{M}Z \,.$$

It follows that $Y \sim MVN(0, \mathbf{M}\mathbf{M}^T)$ so we need to compute $\mathbf{M}\mathbf{M}^T$:

$$\mathbf{M}\mathbf{M}^T = \left[\begin{array}{c|cccc} 1 & 0 & 0 & \cdots & 0 \\ \hline 0 & 1-\frac{1}{n} & -\frac{1}{n} & \cdots & -\frac{1}{n} \\ \vdots & -\frac{1}{n} & \ddots & \cdots & -\frac{1}{n} \\ 0 & \vdots & \cdots & & 1-\frac{1}{n} \end{array}\right]$$

$$= \left[\begin{array}{c|c} 1 & 0 \\ \hline & \\ 0 & \mathbf{Q} \end{array}\right] .$$

Put $\mathbf{Y}_2 = (Y_2, \ldots, Y_{n+1})$. Since

$$\text{Cov}(Y_1, \mathbf{Y}_2) = 0$$

conclude Y_1 and $\mathbf{Y}_2$ are independent and each is normal.

Thus $\sqrt{n}\bar{Z}$ is independent of $Z_1 - \bar{Z}, \ldots, Z_n - \bar{Z}$.

Since s_Z^2 is a function of $Z_1 - \bar{Z}, \ldots, Z_n - \bar{Z}$ we see that $\sqrt{n}\bar{Z}$ and s_Z^2 are independent.

Also, see $\sqrt{n}\bar{Z} \sim N(0, 1)$.

First 2 parts done.

Consider $(n-1)s^2/\sigma^2 = \mathbf{Y}_2^T\mathbf{Y}_2$. Note that $\mathbf{Y}_2 \sim MVN(0, \mathbf{Q})$.

Now: distribution of quadratic forms:

Suppose $Z \sim MVN(0, \mathbf{I})$ and $\mathbf{A}$ is symmetric. Put $\mathbf{A} = \mathbf{P}\mathbf{D}\mathbf{P}^T$ for $\mathbf{D}$ diagonal, $\mathbf{P}$ orthogonal.

Then

$$\mathbf{Z}^T\mathbf{A}\mathbf{Z} = (\mathbf{Z}^*)^T\mathbf{D}\mathbf{Z}^*$$

where

$$\mathbf{Z}^* = \mathbf{P}^T\mathbf{Z}$$

But $\mathbf{Z}^* \sim MVN(0, \mathbf{P}^T\mathbf{P} = \mathbf{I})$ is standard multivariate normal.

So: $\mathbf{Z}^T\mathbf{A}\mathbf{Z}$ has same distribution as

$$\sum_i \lambda_i Z_i^2$$

where $\lambda_1, \ldots, \lambda_n$ are eigenvalues of $\mathbf{A}$.

Special case: if all λ_i are either 0 or 1 then $\mathbf{Z}^T\mathbf{A}\mathbf{Z}$ has a chi-squared distribution with df = number of λ_i equal to 1.

When are eigenvalues all 1 or 0?

Answer: if and only if $\mathbf{A}$ is idempotent.

1) If $\mathbf{A}$ idempotent and λ, x is an eigenpair the

$$\mathbf{A}x = \lambda x$$

and

$$\mathbf{A}x = \mathbf{A}\mathbf{A}x = \lambda \mathbf{A}x = \lambda^2 x$$

so

$$(\lambda - \lambda^2)x = 0$$

proving λ is 0 or 1.

2) Conversely if all eigenvalues of $\mathbf{A}$ are 0 or 1 then $\mathbf{D}$ has 1s and 0s on diagonal so

$$\mathbf{D}^2 = \mathbf{D}$$

and

$$\mathbf{A}\mathbf{A} = \mathbf{P}\mathbf{D}\mathbf{P}^T\mathbf{P}\mathbf{D}\mathbf{P}^T = \mathbf{P}\mathbf{D}^2\mathbf{P}^t = \mathbf{P}\mathbf{D}\mathbf{P} = \mathbf{A}$$

Next case: $\mathbf{X} \sim MVN_p(0, \mathbf{\Sigma})$. Then $\mathbf{X} = \mathbf{AZ}$ with $\mathbf{AA}^T = \mathbf{\Sigma}$.

Since $\mathbf{X}^T\mathbf{X} = \mathbf{Z}^T\mathbf{A}^T\mathbf{AZ}$ it has the law

$$\sum \lambda_i Z_i^2$$

λ_i are eigenvalues of $\mathbf{A}^T\mathbf{A}$. But

$$\mathbf{A}^T\mathbf{A}x = \lambda x$$

implies

$$\mathbf{AA}^T\mathbf{A}x = \mathbf{\Sigma}\mathbf{A}x = \lambda \mathbf{A}x$$

So eigenvalues are those of $\mathbf{\Sigma}$ and $\mathbf{X}^T\mathbf{X}$ is χ^2_ν iff $\mathbf{\Sigma}$ is idempotent and trace$(\mathbf{\Sigma}) = \nu$.

Our case: $\mathbf{A} = \mathbf{Q} = \mathbf{I} - \mathbf{1}\mathbf{1}^T/n$. Check $\mathbf{Q}^2 = \mathbf{Q}$.
How many degrees of freedom: trace($\mathbf{D}$).

Defn: The trace of a square matrix $\mathbf{A}$ is

$$\text{trace}(\mathbf{A}) = \sum \mathbf{A}_{ii}$$

Property: trace($\mathbf{AB}$) = trace($\mathbf{BA}$).

So:

$$\begin{aligned}\text{trace}(\mathbf{A}) &= \text{trace}(\mathbf{PDP}^T) \\ &= \text{trace}(\mathbf{DP}^T\mathbf{P}) = \text{trace}(\mathbf{D})\end{aligned}$$

Conclusion: df for $(n-1)s^2/\sigma^2$ is

$$\text{trace}(\mathbf{I} - \mathbf{1}\mathbf{1}^T/n) = n - 1.$$

Derivation of the χ^2 density:

Suppose $Z_1, \ldots, Z_n$ independent $N(0,1)$. Define χ^2_n distribution to be that of $U = Z_1^2 + \cdots + Z_n^2$. Define angles $\theta_1, \ldots, \theta_{n-1}$ by

$$\begin{aligned} Z_1 &= U^{1/2} \cos\theta_1 \\ Z_2 &= U^{1/2} \sin\theta_1 \cos\theta_2 \\ \vdots &= \vdots \\ Z_{n-1} &= U^{1/2} \sin\theta_1 \cdots \sin\theta_{n-2} \cos\theta_{n-1} \\ Z_n &= U^{1/2} \sin\theta_1 \cdots \sin\theta_{n-1}\,. \end{aligned}$$

(Spherical co-ordinates in n dimensions. The θ values run from 0 to π except last θ from 0 to 2π.) Derivative formulas:

$$\frac{\partial Z_i}{\partial U} = \frac{1}{2U} Z_i$$

and

$$\frac{\partial Z_i}{\partial \theta_j} = \begin{cases} 0 & j > i \\ -Z_i \tan\theta_i & j = i \\ Z_i \cot\theta_j & j < i\,. \end{cases}$$

Fix $n = 3$ to clarify the formulas. Use shorthand $R = \sqrt{U}$.

Matrix of partial derivatives is

$$\begin{bmatrix} \frac{\cos\theta_1}{2R} & -R\sin\theta_1 & 0 \\ \frac{\sin\theta_1\cos\theta_2}{2R} & R\cos\theta_1\cos\theta_2 & -R\sin\theta_1\sin\theta_2 \\ \frac{\sin\theta_1\sin\theta_2}{2R} & R\cos\theta_1\sin\theta_2 & R\sin\theta_1\cos\theta_2 \end{bmatrix}.$$

Find determinant:

$$U^{1/2}\sin(\theta_1)/2$$

(non-negative for all U and θ_1).

General n: every term in the first column contains a factor $U^{-1/2}/2$ while every other entry has a factor $U^{1/2}$.

FACT: multiplying a column in a matrix by c multiplies the determinant by c.

SO: Jacobian of transformation is

$$u^{(n-2)/2}u^{-1/2}/2 \times h(\theta_1, \theta_{n-1})$$

for some function, h, which depends only on the angles.

Thus joint density of $U, \theta_1, \ldots \theta_{n-1}$ is

$$(2\pi)^{-n/2} \exp(-u/2) u^{(n-2)/2} h(\theta_1, \cdots, \theta_{n-1})/2\,.$$

To compute the density of U we must do an $n-1$ dimensional multiple integral $d\theta_{n-1} \cdots d\theta_1$.

Answer has the form

$$cu^{(n-2)/2} \exp(-u/2)$$

for some c.

Evaluate c by making

$$\int f_U(u)du = c\int_0^\infty u^{(n-2)/2}\exp(-u/2)du$$
$$= 1.$$

Substitute $y = u/2$, $du = 2dy$ to see that

$$c2^{n/2}\int_0^\infty y^{(n-2)/2}e^{-y}dy = c2^{n/2}\Gamma(n/2)$$
$$= 1.$$

CONCLUSION: the χ_n^2 density is

$$\frac{1}{2\Gamma(n/2)}\left(\frac{u}{2}\right)^{(n-2)/2}e^{-u/2}1(u > 0)\,.$$

Fourth part: consequence of first 3 parts and def'n of t_ν distribution.

Defn: $T \sim t_\nu$ if T has same distribution as

$$Z/\sqrt{U/\nu}$$

for $Z \sim N(0,1)$, $U \sim \chi^2_\nu$ and Z, U independent.

Derive density of T in this definition:

$$\begin{aligned} P(T \le t) &= P(Z \le t\sqrt{U/\nu}) \\ &= \int_0^\infty \int_{-\infty}^{t\sqrt{u/\nu}} f_Z(z) f_U(u) dz du \end{aligned}$$

Differentiate wrt t by differentiating inner integral:

$$\frac{\partial}{\partial t}\int_{at}^{bt} f(x)dx = bf(bt) - af(at)$$

by fundamental thm of calculus. Hence

$$\frac{d}{dt}P(T \le t) = \int_0^\infty \frac{f_U(u)}{\sqrt{2\pi}} \left(\frac{u}{\nu}\right)^{1/2} \exp\left(-\frac{t^2 u}{2\nu}\right) du\,.$$

Plug in

$$f_U(u) = \frac{1}{2\Gamma(\nu/2)}(u/2)^{(\nu-2)/2}e^{-u/2}$$

to get

$$f_T(t) = \frac{\int_0^\infty (u/2)^{(\nu-1)/2}e^{-u(1+t^2/\nu)/2}du}{2\sqrt{\pi\nu}\Gamma(\nu/2)}.$$

Substitute $y = u(1+t^2/\nu)/2$, to get

$$dy = (1+t^2/\nu)du/2$$

$$(u/2)^{(\nu-1)/2} = [y/(1+t^2/\nu)]^{(\nu-1)/2}$$

leading to

$$f_T(t) = \frac{(1+t^2/\nu)^{-(\nu+1)/2}}{\sqrt{\pi\nu}\Gamma(\nu/2)}\int_0^\infty y^{(\nu-1)/2}e^{-y}dy$$

or

$$f_T(t) = \frac{\Gamma((\nu+1)/2)}{\sqrt{\pi\nu}\Gamma(\nu/2)}\frac{1}{(1+t^2/\nu)^{(\nu+1)/2}}.$$

Multivariate Normal samples: Distribution Theory

Theorem: Suppose $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are independent $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ random variables. Then

1. $\bar{\mathbf{X}}$ (sample mean)and $\mathbf{S}$ (sample variance-covariance matrix) are independent.

2. $n^{1/2}(\bar{\mathbf{X}} - \boldsymbol{\mu}) \sim MVN(0, \mathbf{I})$.

3. $(n-1)\mathbf{S} \sim \text{Wishart}_p(n-1, \boldsymbol{\Sigma})$.

4. $T^2 = n(\bar{\mathbf{X}} - \boldsymbol{\mu})^T \mathbf{S}^{-1} (\bar{\mathbf{X}} - \boldsymbol{\mu})$ is Hotelling's T^2. $(n-p)T^2/(p(n-1))$ has an $F_{p,n-p}$ distribution.

Proof: Let $\mathbf{X}_i = \mathbf{A}\mathbf{Z}_i + \boldsymbol{\mu}$ where $\mathbf{A}\mathbf{A}^T = \boldsymbol{\Sigma}$ and $\mathbf{Z}_1, \ldots, \mathbf{Z}_p$ are independent $MVN(0, \mathbf{I})$.

So $\mathbf{Z} = (\mathbf{Z}_1^T, \ldots, \mathbf{Z}_p^T)^T \sim MVN_p(0, \mathbf{I})$.

Note that $\bar{\mathbf{X}} = \mathbf{A}\bar{\mathbf{Z}} + \boldsymbol{\mu}$ and

$$\begin{aligned}(n-1)\mathbf{S} &= \sum(\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^T \\ &= \mathbf{A}\sum(\mathbf{Z}_i - \bar{\mathbf{Z}})(\mathbf{Z}_i - \bar{\mathbf{Z}})^T\mathbf{A}^T\end{aligned}$$

Thus

$$n^{1/2}(\bar{X} - \mu) = \mathbf{A}n^{1/2}\bar{\mathbf{Z}}$$

and

$$T^2 = \left(n^{1/2}\bar{\mathbf{Z}}\right)^T \mathbf{S}_{\mathbf{Z}}^{-1} \left(n^{1/2}\bar{\mathbf{Z}}\right)$$

where

$$\mathbf{S}_Z = \sum(\mathbf{Z}_i - \bar{\mathbf{Z}})(\mathbf{Z}_i - \bar{\mathbf{Z}})^T/(n-1).$$

Consequences. In 1, 2 and 4: can assume $\boldsymbol{\mu} = 0$ and $\boldsymbol{\Sigma} = \mathbf{I}$. In 3 can take $\boldsymbol{\mu} = 0$.

Step 1: Do general $\mathbf{\Sigma}$. Define

$$\mathbf{Y} = (\sqrt{n}\bar{\mathbf{Z}}^T, \mathbf{Z}_1^T - \bar{\mathbf{Z}}^T, \ldots, \mathbf{Z}_n^T - \bar{\mathbf{Z}}^T)^T .$$

(So $\mathbf{Y}$ has dimension $p(n+1)$.) Clearly $\mathbf{Y}$ is MVN with mean 0.

Compute variance covariance matrix

$$\begin{bmatrix} \mathbf{I}_{p\times p} & 0 \\ 0 & \mathbf{Q}^* \end{bmatrix}$$

where $\mathbf{Q}^*$ has a pattern. It is a $p \times p$ patterned matrix with entry ij being

$$\begin{aligned} \text{Cov}(\mathbf{Z}_i - \bar{\mathbf{Z}}, \mathbf{Z}_j - \bar{\mathbf{Z}}) &= \begin{cases} -\mathbf{\Sigma}/n & i \neq j \\ (n-1)\mathbf{\Sigma}/n & i = j \end{cases} \\ &= \mathbf{Q}_{ij}\mathbf{\Sigma} \end{aligned}$$

Kronecker Products

Defn: If $\mathbf{A}$ is $p \times q$ and $\mathbf{B}$ is $r \times s$ then $\mathbf{A} \bigotimes \mathbf{B}$ is the $pr \times qs$ matrix with the pattern

$$\left[\begin{array}{cccc} \mathbf{A}_{11}\mathbf{B} & \mathbf{A}_{12}\mathbf{B} & \cdots & \mathbf{A}_{1q}\mathbf{B} \\ \vdots & \vdots & \cdots & \vdots \\ \mathbf{A}_{p1}\mathbf{B} & \mathbf{A}_{p2}\mathbf{B} & \cdots & \mathbf{A}_{pq}\mathbf{B} \end{array}\right]$$

So our variance covariance matrix is

$$\mathbf{Q}^* = \mathbf{Q} \bigotimes \mathbf{\Sigma}$$

Conclusions so far:

1) $\bar{\mathbf{X}}$ and $\mathbf{S}$ are independent.

2) $\sqrt{n}(\bar{\mathbf{X}} - \boldsymbol{\mu}) \sim MVN(0, \mathbf{\Sigma})$

Next: Wishart law.

Defn: The $\text{Wishart}_p(n, \mathbf{\Sigma})$ distribution is the distribution of

$$\sum_1^n \mathbf{Z}_i \mathbf{Z}_i^T$$

where $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ are iid $MVN_p(0, \mathbf{\Sigma})$.

Properties of Wishart.

1) If $\mathbf{A}\mathbf{A}^t = \mathbf{\Sigma}$ then

$$\text{Wishart}_p(0, \mathbf{\Sigma}) = \mathbf{A}\text{Wishart}_p(0, \mathbf{I})\mathbf{A}^T$$

2) if $\mathbf{W}_i, i = 1, 2$ independent $\text{Wishart}_p(n_i, \mathbf{\Sigma})$ then

$$\mathbf{W}_1 + \mathbf{W}_2 \sim \text{Wishart}_p(n_1 + n_2, \mathbf{\Sigma}).$$

Proof of part 3: rewrite

$$\sum (\mathbf{Z}_i - \bar{\mathbf{Z}})(\mathbf{Z}_i - \bar{\mathbf{Z}})^T$$

in form

$$\sum_{j=1}^{n-1} \mathbf{U}_i \mathbf{U}_i^T$$

for $\mathbf{U}_i$ iid $MVN_p(0, \mathbf{\Sigma})$. Put $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ as cols in matrix $\mathbf{Z}$ which is $p \times n$. Then check that

$$\mathbf{Z}\mathbf{Q}\mathbf{Z}^T = \sum (\mathbf{Z}_i - \bar{\mathbf{Z}})(\mathbf{Z}_i - \bar{\mathbf{Z}})^T$$

Write $\mathbf{Q} = \sum \mathbf{v}_i \mathbf{v}_i^T$ for $n-1$ orthogonal unit vectors $\mathbf{v}_1, \ldots, \mathbf{v}_{n-1}$. Define

$$\mathbf{U}_i = \mathbf{Z}\mathbf{v}_i$$

and compute covariances to check that the $\mathbf{U}_i$ are iid $MVN_p(0, \mathbf{\Sigma})$. Then check that

$$\mathbf{Z}\mathbf{Q}\mathbf{Z}^T = \sum \mathbf{U}_i \mathbf{U}_i^T$$

Proof of 4: suffices to have $\mathbf{\Sigma} = \mathbf{I}$.

Uses further props of Wishart distribution.

3: If $\mathbf{W} \sim Wishart_p(n, \mathbf{\Sigma})$ and $\mathbf{a} \in \mathbb{R}$ then

$$\frac{\mathbf{a}^T \mathbf{W} \mathbf{a}}{\mathbf{a}^T \mathbf{\Sigma} \mathbf{a}} \sim \chi^2_n$$

4: If $\mathbf{W} \sim Wishart_p(n, \mathbf{\Sigma})$ and $n \geq p$ then

$$\frac{\mathbf{a}^T \mathbf{\Sigma}^{-1} \mathbf{a}}{\mathbf{a}^T \mathbf{W}^{-1} \mathbf{a}} \sim \chi^2_{n-p+1}$$

5: If $\mathbf{W} \sim Wishart_p(n, \mathbf{\Sigma})$ then

$$\text{trace}(\mathbf{\Sigma}^{-1} \mathbf{W}) \sim \chi^2_{np}$$

6: If $\mathbf{W} \sim Wishart_{p+q}(n, \mathbf{\Sigma})$ is partitioned into components then

$$\mathbf{W}_{11} - \mathbf{W}_{12} \mathbf{W}_{22}^{-1} \mathbf{W}_{21} \sim Wishart_p(n-q, \mathbf{\Sigma}_{11.2})$$

One sample tests on mean vectors

Given data $\mathbf{X}_1, \ldots, \mathbf{X}_n$ iid $MVN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ test

$$H_o : \boldsymbol{\mu} = \boldsymbol{\mu}_o$$

by computing

$$T^2 = n(\bar{\mathbf{X}} - \boldsymbol{\mu}_o)^T \mathbf{S}^{-1} (\bar{\mathbf{X}} - \boldsymbol{\mu}_o)$$

and getting P-values from F distribution using theorem.

Example: no realistic ones. This hypothesis is not intrinsically useful. However: other tests can sometimes be reduced to it.

Example: Ten water samples split in half. One half of each to each of two labs. Measure biological oxygen demand (BOD) and suspended solids (SS). For sample i let X_{i1} be BOD for lab A, X_{i2} be SS for lab A, X_{i3} be BOD for lab B and X_{i4} be SS for lab B. Question: are labs measuring the same thing? Is there bias in one or the other?

Notation $\mathbf{X}_i$ is vector of 4 measurements on sample i.

Data:

	Lab A		Lab B	
Sample	BOD	SS	BOD	SS
1	6	27	25	15
2	6	23	28	13
3	18	64	36	22
4	8	44	35	29
5	11	30	15	31
6	34	75	44	64
7	28	26	42	30
8	71	124	54	64
9	43	54	34	56
10	33	30	29	20
11	20	14	39	21

Model: $\mathbf{X}_1, \ldots, \mathbf{X}_{11}$ are iid $MVN_4(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Multivariate problem because: not able to assume independence between any two measurements on same sample.

Potential sub-model: each measurement is

true mmnt + lab bias + mmnt error.

Model for measurement error vector $\mathbf{U}_i$ is multivariate normal mean 0 and diagonal covariance matrix $\mathbf{\Sigma_U}$.

Lab bias is unknown vector $\boldsymbol{\beta}$.

True measurement should be same for both labs so has form

$$[Y_{i1}, Y_{i2}, Y_{i1}, Y_{i2}]$$

where Y_{i1}, Y_{i2} are iid bivariate normal with unknown means θ_1, θ_2 and unknown 2×2 variance covariance $\mathbf{\Sigma_Y}$.

This would give structured model

$$\mathbf{X}_i = \mathbf{CY} + \boldsymbol{\beta} + \mathbf{U}$$

where

$$\mathbf{C} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$$

This model has variance covariance matrix

$$\mathbf{\Sigma_X} = \mathbf{C}\mathbf{\Sigma_Y}\mathbf{C}^T + \mathbf{\Sigma_U}$$

Notice that this matrix has only 7 parameters: four for the diagonal entries in $\mathbf{\Sigma_U}$ and 3 for the entries in $\mathbf{\Sigma_Y}$.

We skip this model and let $\mathbf{\Sigma_X}$ be unrestricted.

Question of interest:

$$H_o : \mu_1 = \mu_3 \text{ and } \mu_2 = \mu_4$$

Reduction: partition $\mathbf{X}_i$ as

$$\begin{bmatrix} \mathbf{U}_i \\ \mathbf{V}_i \end{bmatrix}$$

where $\mathbf{U}_i$ and $\mathbf{V}_i$ each have two components.

Define $\mathbf{W}_i = \mathbf{U}_i$. Then our model makes $\mathbf{W}_i$ iid $MVN_2(\boldsymbol{\mu}_\mathbf{W}, \mathbf{\Sigma_W})$. Our hypothesis is

$$H_o : \boldsymbol{\mu}_\mathbf{W} = 0$$

Carrying out our test in `SPlus`:

Working on CSS unix workstation:

Start `SPlus` then read in, print out data:

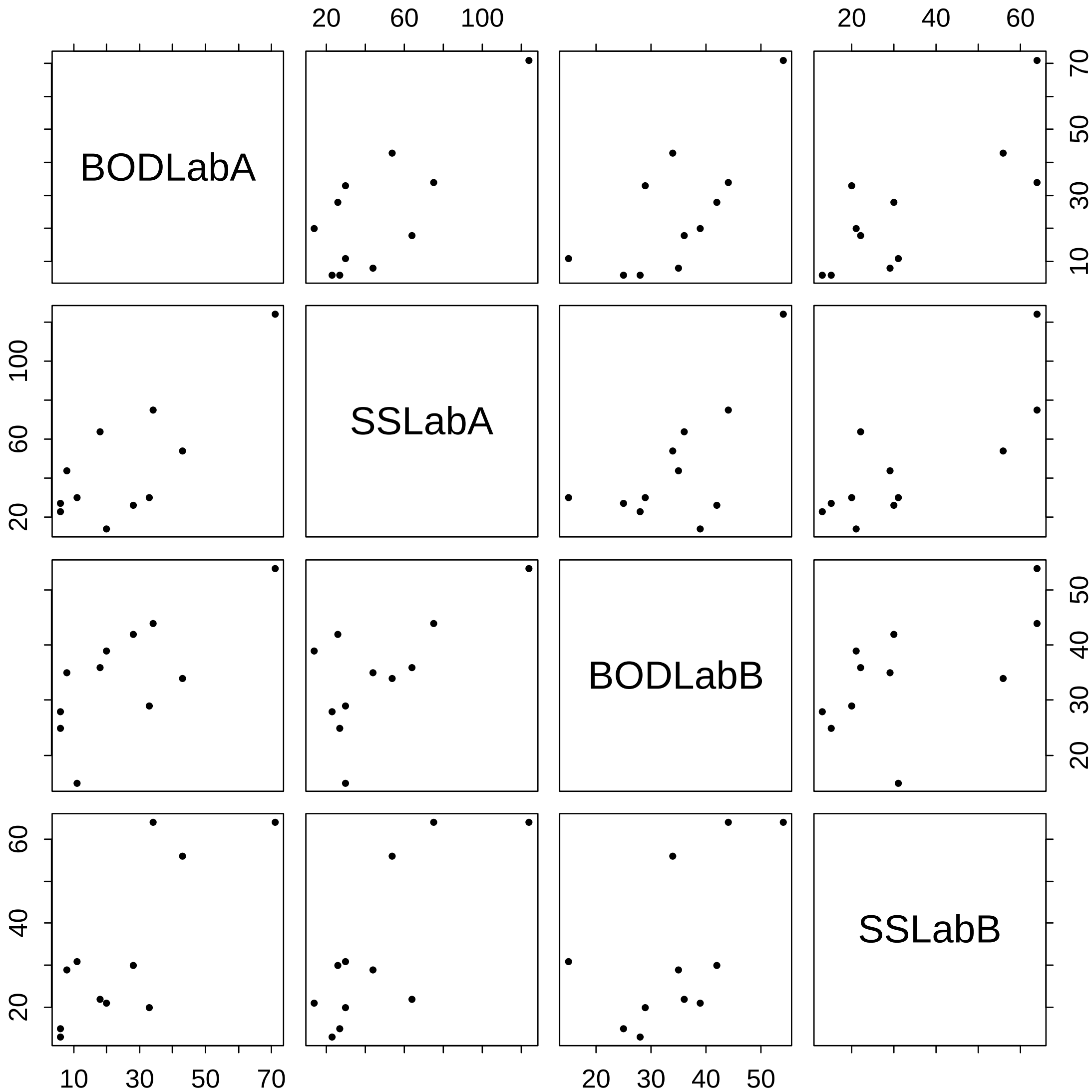
```
[61]ehlehl% mkdir .Data
[62]ehlehl% Splus
S-PLUS : Copyright (c) 1988, 1996 MathSoft, Inc.
S : Copyright AT&T.
Version 3.4 Release 1 for Sun SPARC, SunOS 5.3 : 1996
Working data will be in .Data
> # Read in and print out data
> eff <- read.table("effluent.dat",header=T)
> eff
   BODLabA SSLabA BODLabB SSLabB
 1       6     27      25     15
 2       6     23      28     13
 3      18     64      36     22
 4       8     44      35     29
 5      11     30      15     31
 6      34     75      44     64
 7      28     26      42     30
 8      71    124      54     64
 9      43     54      34     56
10      33     30      29     20
11      20     14      39     21
```

Do some graphical preliminary analysis.

Look for non-normality, non-linearity, outliers.

Make plots on screen or saved in file.

```
> # Make pairwise scatterplots on screen using
> # motif graphics device and then in a postscript
> # file.
> motif()
> pairs(eff)
> postscript("pairs.ps",horizontal=F,
+   height=6,width=6)
> pairs(eff)
> dev.off()
Generated postscript file "pairs.ps".
 motif
     2
```



BODLabA
SSLabA
BODLabB
SSLabB
20
60
100
20
40
60
70
50
30
10
100
60
20
50
40
30
20
60
40
20
10
30
50
70
20
30
40
50

Examine correlations

```
> cor(eff)
          BODLabA    SSLabA   BODLabB    SSLabB
BODLabA 0.9999999 0.7807413 0.7228161 0.7886035
 SSLabA 0.7807413 1.0000000 0.6771183 0.7896656
BODLabB 0.7228161 0.6771183 1.0000001 0.6038079
 SSLabB 0.7886035 0.7896656 0.6038079 1.0000001
```

Notice high correlations.

Mostly caused by variation in true levels from sample to sample.

Get partial correlations.

Adjust for overall BOD and SS content of sample.

```
> aug <- cbind(eff,(eff[,1]+eff[,3])/2,
+                   (eff[,2]+eff[,4])/2)
> aug
   BODLabA SSLabA BODLabB SSLabB   X2   X3
 1       6     27      25     15 15.5 21.0
 2       6     23      28     13 17.0 18.0
 3      18     64      36     22 27.0 43.0
 4       8     44      35     29 21.5 36.5
 5      11     30      15     31 13.0 30.5
 6      34     75      44     64 39.0 69.5
 7      28     26      42     30 35.0 28.0
 8      71    124      54     64 62.5 94.0
 9      43     54      34     56 38.5 55.0
10      33     30      29     20 31.0 25.0
11      20     14      39     21 29.5 17.5
> bigS <- var(aug)
```

Now compute partial correlations for first four variables given means of BOD and SS:

```
> S11 <- bigS[1:4,1:4]
> S12 <- bigS[1:4,5:6]
> S21 <- bigS[5:6,1:4]
> S22 <- bigS[5:6,5:6]
> S11dot2 <- S11 - S12 %*% solve(S22,S21)
> S11dot2
           BODLabA     SSLabA    BODLabB     SSLabB
BODLabA  24.804665  -7.418491 -24.804665   7.418491
 SSLabA  -7.418491  59.142084   7.418491 -59.142084
BODLabB -24.804665   7.418491  24.804665  -7.418491
 SSLabB   7.418491 -59.142084  -7.418491  59.142084
> S11dot2SD <- diag(sqrt(diag(S11dot2)))
> S11dot2SD
         [,1]     [,2]     [,3]     [,4]
[1,] 4.980428 0.000000 0.000000 0.000000
[2,] 0.000000 7.690389 0.000000 0.000000
[3,] 0.000000 0.000000 4.980428 0.000000
[4,] 0.000000 0.000000 0.000000 7.690389
> R11dot2 <- solve(S11dot2SD)%*%
+           S11dot2%*%solve(S11dot2SD)
> R11dot2
          [,1]      [,2]      [,3]      [,4]
[1,]  1.000000 -0.193687 -1.000000  0.193687
[2,] -0.193687  1.000000  0.193687 -1.000000
[3,] -1.000000  0.193687  1.000000 -0.193687
[4,]  0.193687 -1.000000 -0.193687  1.000000
```

Notice little residual correlation.

Carry out Hotelling's T^2 test of $H_0 : \boldsymbol{\mu}_{\mathbf{W}} = 0$.

```
> w <- eff[,1:2]-eff[3:4]
> dimnames(w)<-list(NULL,c("BODdiff","SSdiff"))
> w
      BODdiff SSdiff
 [1,]     -19     12
 [2,]     -22     10
  etc
 [8,]      17     60
  etc
> Sw <- var(w)
> cor(w)
          BODdiff    SSdiff
BODdiff 1.0000001 0.3057682
 SSdiff 0.3057682 1.0000000
> mw <- apply(w,2,mean)
> mw
   BODdiff   SSdiff
 -9.363636 13.27273
> Tsq <- 11*mw%*%solve(Sw,mw)
> Tsq
         [,1]
[1,] 13.63931
> FfromTsq <- (11-2)*Tsq/(2*(11-1))
> FfromTsq
        [,1]
[1,] 6.13769
> 1-pf(FfromTsq,2,9)
[1] 0.02082779
```

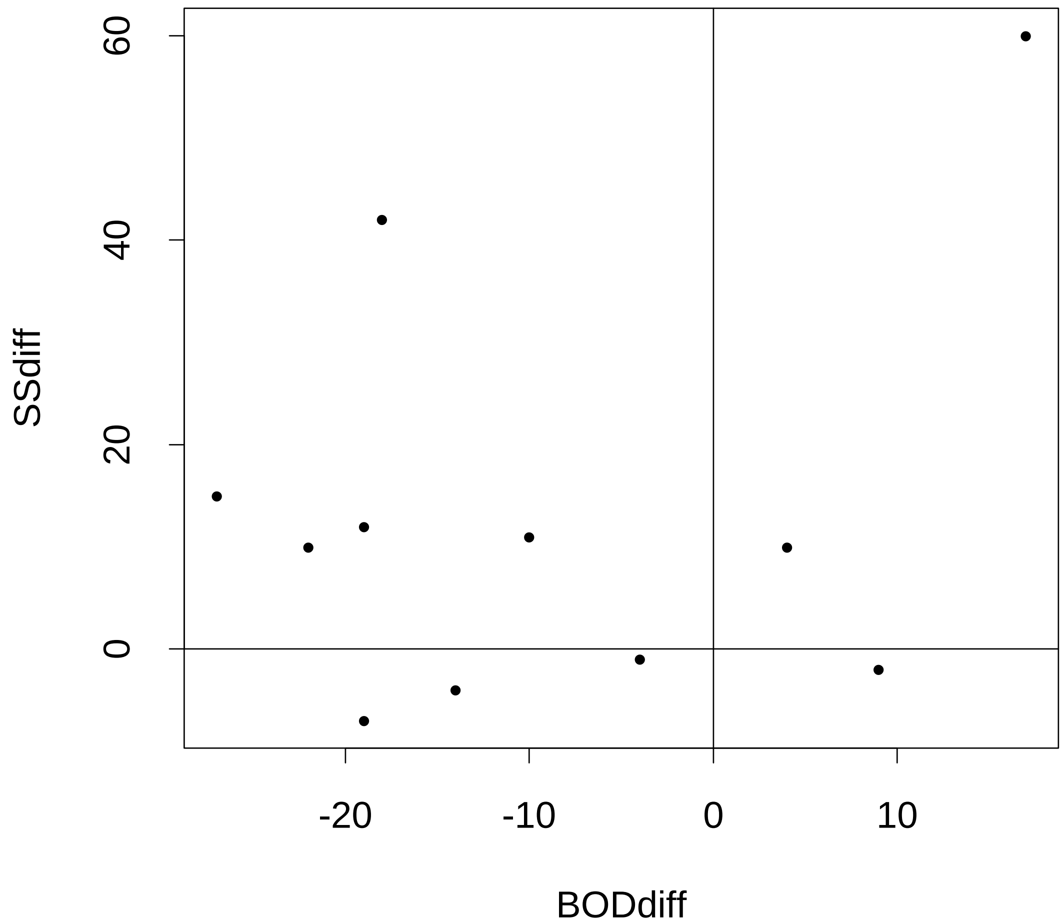
Conclusion: Pretty clear evidence of difference in mean level between labs.

Which measurement causes the difference?

```
> TBOD <- sqrt(11)*mw[1]/sqrt(Sw[1,1])
> TBOD
   BODdiff
 -2.200071
> 2*pt(TBOD,1)
   BODdiff
 0.2715917
> 2*pt(TBOD,10)
    BODdiff
 0.05243474
> TSS <- sqrt(11)*mw[2]/sqrt(Sw[2,2])
> TSS
  SSdiff
 2.15153
> 2*pt(-TSS,10)
     SSdiff
 0.05691733
> postscript("differences.ps",
+      horizontal=F,height=6,width=6)
> plot(w)
> abline(h=0)
> abline(v=0)
> dev.off()
```

Conclusion? Neither? Not a problem – summarizes evidence!

Problem: several tests at level 0.05 on same data. **Simultaneous** or **Multiple** comparisons.



In general can test hypothesis $H_o : \mathbf{C}\boldsymbol{\mu} = 0$ by computing $\mathbf{Y}_i = \mathbf{C}\mathbf{X}_i$ and then testing $H_o : \boldsymbol{\mu}_{\mathbf{Y}} = 0$ using Hotelling's T^2.

Simultaneous confidence intervals

Confidence interval for $\mathbf{a}^T\boldsymbol{\mu}$:

$$\mathbf{a}^T\bar{\mathbf{X}} \pm t_{n-1,\alpha/2}\sqrt{\mathbf{a}^T\mathbf{S}\mathbf{a}}$$

Based on t distribution.

Give coverage intervals for 6 parameters of interest: 4 entries in $\boldsymbol{\mu}$ and $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_3$ and $\boldsymbol{\mu}_2 - \boldsymbol{\mu}_4$

$$\begin{array}{cc}
\boldsymbol{\mu}_1 & 25.27 \pm 2.23 \times 19.68/\sqrt{11} \\
\boldsymbol{\mu}_2 & 46.45 \pm 2.23 \times 31.84/\sqrt{11} \\
\boldsymbol{\mu}_3 & 34.64 \pm 2.23 \times 10.45/\sqrt{11} \\
\boldsymbol{\mu}_4 & 33.18 \pm 2.23 \times 19.07/\sqrt{11} \\
\boldsymbol{\mu}_1 - \boldsymbol{\mu}_3 & -9.36 \pm 2.23 \times 14.12/\sqrt{11} \\
\boldsymbol{\mu}_2 - \boldsymbol{\mu}_4 & 13.27 \pm 2.23 \times 20.46/\sqrt{11}
\end{array}$$

Problem: each confidence interval has 5% error rate. Pick out last interval (on basis of looking most interesting) and ask about error rate?

Solution: adjust 2.23, t multiplier to get

$$P(\text{all intervals cover truth}) \geq 0.95$$

Rao or Scheffé type intervals

Based on inequality:

$$\left|\mathbf{a}^T\mathbf{b}\right|^2 \le \mathbf{a}^T\mathbf{M}\mathbf{a}\mathbf{b}^T\mathbf{M}^{-1}\mathbf{b}$$

for any symmetric non-singular matrix $\mathbf{M}$.

Proof by Cauchy Schwarz: inner product of vectors $\mathbf{M}^{1/2}\mathbf{a}$ and $\mathbf{M}^{-1/2}\mathbf{B}$.

Put $\mathbf{b} = n^{-1/2}(\bar{\mathbf{X}} - \boldsymbol{\mu})$ and $\mathbf{M} = \mathbf{S}$ to get

$$\left|n^{1/2}(\mathbf{a}^T\bar{\mathbf{X}} - \mathbf{a}^T\boldsymbol{\mu})\right|^2 \le \mathbf{a}^T\mathbf{S}\mathbf{a}T^2$$

This inequality is true for all $\mathbf{a}$. Thus the event that there is any $\mathbf{a}$ such that

$$\frac{(\mathbf{a}^T\bar{\mathbf{X}} - \mathbf{a}^T\boldsymbol{\mu})^2}{\mathbf{a}^T\mathbf{S}\mathbf{a}} > c$$

is a subset of the event

$$T^2 > c$$

Adjust c to make the latter event have probability α by taking

$$c = \frac{p(n-1)}{n-p} F_{p,n-p,\alpha}.$$

Then the probability that every one of the uncountably many confidence intervals

$$\mathbf{a}^T \bar{\mathbf{X}} \pm \sqrt{c}\sqrt{\mathbf{a}^T \mathbf{S} \mathbf{a}}$$

covers the corresponding true parameter value is at least $1-\alpha$.

In fact the probability of this happening is exactly equal to $1-\alpha$ because for each data set the supremum of

$$\frac{(\mathbf{a}^T \bar{\mathbf{X}} - \mathbf{a}^T \boldsymbol{\mu})^2}{\mathbf{a}^T \mathbf{S} \mathbf{a}}$$

over all $\mathbf{a}$ is T^2.

Our case

$$\sqrt{c} = \frac{4(10)}{7} F_{4,7,0.05} = 4.85$$

Coverage probability of single interval using $\sqrt{c}4.85$? From t distribution:

$$99.93\%$$

Probability all 6 intervals would cover using $\sqrt{c}4.85$?

Use Bonferroni inequality:

$$P(\cup A_i) \le \sum P(A_i)$$

Simultaneous coverage probability of 6 intervals using $\sqrt{c}4.85$

$$\ge 1 - 6 * (1 - 0.9993) = 99.59\%$$

Usually just use

$$\sqrt{c} = t_{n-1,\alpha/12} = 3.28$$

General Bonferroni strategy. If we want intervals for $\theta_1, \ldots, \theta_k$ get interval for θ_i at level $1 - \alpha/k$. Simultaneous coverage probability is at least $1 - \alpha$. Notice that Bonferroni narrower in our example unless $0.0007k = 0.5$ giving $k > 71$.

Motivations for T^2:

1: Hypothesis H_o : $\boldsymbol{\mu} = 0$ is true iff all hypotheses $H_{o\mathbf{a}}$: $\mathbf{a}^T\boldsymbol{\mu} = 0$ are true. Natural test for $H_{o\mathbf{a}}$ rejects if

$$t(\mathbf{a}) = \frac{n^{1/2}\mathbf{a}^T(\bar{\mathbf{X}} - \boldsymbol{\mu})}{\mathbf{a}^T\mathbf{S}\mathbf{a}}$$

large. So take largest test statistic.

Fact:

$$\sup_{\mathbf{a}} t^2(\mathbf{a}) = T^2$$

Proof: like calculation of maximal correlation.

2: likelihood ratio method.

Compute

$$\ell(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}) - \ell(\hat{\boldsymbol{\mu}}_o, \hat{\boldsymbol{\Sigma}}_o)$$

where the subscript o indicates estimation assuming H_o.

In our case to test $H_o : \boldsymbol{\mu} = 0$ find

$$\hat{\boldsymbol{\mu}}_o = 0 \quad \hat{\boldsymbol{\Sigma}}_o = \sum \mathbf{X}_i \mathbf{X}_i^T / n$$

and

$$\ell(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}) - \ell(\hat{\boldsymbol{\mu}}_o, \hat{\boldsymbol{\Sigma}}_o) = n \log\{\det(\hat{\boldsymbol{\Sigma}}) / \det(\hat{\boldsymbol{\Sigma}}_0)\}/2$$

Now write

$$\sum \mathbf{X}_i \mathbf{X}_i^T = \sum (\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^T + n\bar{\mathbf{X}}\bar{\mathbf{X}}^T$$

Use formula:

$$\det(\mathbf{A} + \mathbf{v}\mathbf{v}^T) = \det(\mathbf{A})(1 + \mathbf{v}^T \mathbf{A}^{-1} \mathbf{v})$$

to get

$$\det(n\hat{\boldsymbol{\Sigma}}_0) = \det(n\hat{\boldsymbol{\Sigma}})(1 + n\bar{\mathbf{X}}^T \left(n\hat{\boldsymbol{\Sigma}}\right)^{-1} \bar{\mathbf{X}})$$

so that the ratio of determinants is a monotone increasing function of T^2.

Again conclude: likelihood ratio test rejects for $T^2 > c$ where c chosen to make level α.

3: compare estimates of $\mathbf{\Sigma}$.

In univariate regression F tests to compare a restricted model with a full model have form

$$\frac{ESS_{\text{Restricted}} - ESS_{\text{Full}}}{ESS_{\text{Full}}} \frac{df_{\text{Error}}}{df_{\text{difference}}}$$

This is a monotone function of

$$\frac{\hat\sigma^2_{\text{Restricted}}}{\hat\sigma^2_{\text{Full}}}$$

where ESS denotes an error sum of squares and $\hat\sigma^2$ and estimate of the residual variance – ESS/df.

Here: substitute matrices.

Analogue of ESS for full model:

$$\mathbf{E}$$

Analogue of ESS for reduced model:

$$\mathbf{E} + \mathbf{H}$$

(This defined $\mathbf{H}$ to be the change in the **Sum of Squares and Cross Products matrix**.)

In 1 sample example:

$$\mathbf{E} = \sum (\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^T$$

and

$$\mathbf{E} + \mathbf{H} = \sum \mathbf{X}_i \mathbf{X}_i^T$$

Test of $\boldsymbol{\mu} = 0$ based on comparing

$$\mathbf{H} = \sum \mathbf{X}_i \mathbf{X}_i^T - \sum (\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^T = n\bar{\mathbf{X}}\bar{\mathbf{X}}^T$$

to

$$\mathbf{E} = \sum (\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^T = (n-1)\mathbf{S}$$

To make comparison. If null true

$$\mathsf{E}(n\bar{\mathbf{X}}\bar{\mathbf{X}}^T) = \mathbf{\Sigma}$$

and

$$\mathsf{E}(\mathbf{S}) = \mathbf{\Sigma}$$

so try tests based on closeness of

$$\mathbf{S}^{-1}n\bar{\mathbf{X}}\bar{\mathbf{X}}^T$$

to identity matrix.

Measures of size based on eigenvalues of

$$\mathbf{S}^{-1}n\bar{\mathbf{X}}\bar{\mathbf{X}}^T$$

which are same as eigenvalues of

$$\mathbf{S}^{-1/2}n\bar{\mathbf{X}}\bar{\mathbf{X}}^T\mathbf{S}^{-1/2}$$

Suggested size measures for $\mathbf{A} - \mathbf{I}$:

- trace$(\mathbf{A} - \mathbf{I})$ (= sum of eigenvalues).

- det$(\mathbf{A} - \mathbf{I})$ (= product of eigenvalues).

- maximum eigenvalue of $\mathbf{A} - \mathbf{I}$.

For our matrix: eigenvalues all 0 except for one. (So really–matrix not close to $\mathbf{I}$.)

Largest eigenvalue is

$$T^2 = n\bar{\mathbf{X}}\mathbf{S}^{-1}\bar{\mathbf{X}}$$

But: see two sample problem for precise tests based on suggestions.

Two sample problem

Given data $\mathbf{X}_{ij}; j = 1, \ldots, n_i; i = 1, 2$. Model $\mathbf{X}_{ij} \sim MVN_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, independent.

Test $H_o : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$.

Case 1: for motivation only. $\boldsymbol{\Sigma}_i$ known $i = 1, 2$.

Natural test statistic: based on

$$\mathbf{D} = \bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2$$

which has $MVN(\boldsymbol{\mu}_{\mathbf{D}}, \boldsymbol{\Sigma}_{\mathbf{D}})$ where

$$\boldsymbol{\mu}_{\mathbf{D}} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$$

and

$$\boldsymbol{\Sigma}_{\mathbf{D}} = n_1^{-1} \boldsymbol{\Sigma}_1 + n_2^{-1} \boldsymbol{\Sigma}_2$$

So

$$\mathbf{D}^T \left(\boldsymbol{\Sigma}_1 / n_1 + \boldsymbol{\Sigma}_2 / n_2 \right)^{-1} \mathbf{D}$$

has a χ_p^2 distribution if null true.

If $\boldsymbol{\Sigma}_i$ not known must estimate. No universally agreed best procedure (even for $p = 1$ — called Behrens-Fisher problem).

Usually: assume $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$.

If so: MLE of $\boldsymbol{\mu}_i$ is $\bar{\mathbf{X}}_i$ and of $\boldsymbol{\Sigma}$ is

$$\frac{\sum_{ij}(\mathbf{X}_{ij} - \bar{\mathbf{X}}_i)(\mathbf{X}_{ij} - \bar{\mathbf{X}}_i)^T}{n_1 + n_2}$$

Usual estimate of $\boldsymbol{\Sigma}$ is

$$\mathbf{S}_{\text{Pooled}} = \frac{\sum_{ij}(\mathbf{X}_{ij} - \bar{\mathbf{X}}_i)(\mathbf{X}_{ij} - \bar{\mathbf{X}}_i)^T}{n_1 + n_2 - 2}$$

which is unbiased.

Possible test developments:

1) By analogy with 1 sample:

$$T^2 = \frac{n_1 n_2}{n_1 + n_2} \mathbf{D}^T \mathbf{S}_{\text{Pooled}}^{-1} \mathbf{D}$$

which has the distribution

$$\frac{n_1 + n_2 - 1 - p}{p(n_1 + n_2 - 2)} T^2 \sim F_{p, n_1 + n_2 - 1 - p}$$

2) Union-intersection: test of $\mathbf{a}^t(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = 0$ based on

$$t_a = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \frac{\mathbf{a}^T \mathbf{D}}{\sqrt{\mathbf{a}^T \mathbf{S} \mathbf{a}}}$$

Maximize t^2 over all $\mathbf{a}$.

Get

$$T^2 = \frac{n_1 n_2}{n_1 + n_2} \mathbf{D}^T \mathbf{S}^{-1} \mathbf{D}$$

3) Likelihood ratio: the MLE of $\boldsymbol{\Sigma}$ for the unrestricted model is

$$\frac{n_1 + n_2 - 2}{n_1 + n_2} \mathbf{S}$$

Under the hypothesis $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ the mle of $\boldsymbol{\Sigma}$ is

$$\frac{\sum_{ij} (\mathbf{X}_{ij} - \bar{\bar{\mathbf{X}}})(\mathbf{X}_{ij} - \bar{\bar{\mathbf{X}}})^T}{n_1 + n_2}$$

where

$$\bar{\bar{\mathbf{X}}} = \frac{n_1 \bar{\mathbf{X}}_1 + n_2 \bar{\mathbf{X}}_2}{n_1 + n_2}$$

This simplifies to

$$\frac{\mathbf{E}+\mathbf{H}}{n_1+n_2}$$

The log-likelihood ratio is a multiple of

$$\log\det\hat{\mathbf{\Sigma}}_{\text{Full}} - \log\det\hat{\mathbf{\Sigma}}_{\text{Restricted}}$$

which is a function of

$$\log\{\det\mathbf{E}/\det(\mathbf{E}+\mathbf{H})\}$$

or equivalently a function of **Wilk's** Λ:

$$\Lambda = \frac{\det\mathbf{E}}{\det(\mathbf{E}+\mathbf{H})} = \frac{1}{\det(\mathbf{H}\mathbf{E}^{-1}+\mathbf{I})}$$

Compute det: multiply together eigenvalues.

If λ_i are the eigenvalues of $\mathbf{H}\mathbf{E}^{-1}$ then

$$\Lambda = \frac{1}{\prod(1+\lambda_i)}$$

Two sample analysis in SAS on css network

- Type `sas` to start system.

- Several windows open. Go to Program Editor.

- Under file menu open file with sas code. Contents of `sas2sample.sas`

```
    data long;
    infile 'tab57sh';
    input group a b c;
run;
proc print;
run;
proc glm;
    class group;
    model a b c = group;
    manova h=group / printh printe;
run;
```

Notes:

1) First 4 lines form `DATA` step:

a) creates data set named long by reading in 4 columns of data from file named `tab57sh` stored in same directory as I was in when I typed `sas`.

b) Calls variables group (=1 or 2 as label for the two groups) and a, b, c which are names for the 3 test scores for each subject.

2) Next two lines: print out data: result is (slightly edited)

```
Obs     group      a      b      c
  1       1       19     20     18
  2       1       20     21     19
  3       1       19     22     22
 etc till
 11       2       15     17     15
 12       2       13     14     14
 13       2       14     16     13
```

3) Then use `proc glm` to do analysis:

a) `class group` declares that the variable group defines levels of a categorical variable.

b) `model` statement says to regress the variables `a, b, c` on variable group.

c) `manova` statement says to do both 3 univariate regressions and a mulivariate regression and to print out the $\mathbf{H}$ and $\mathbf{E}$ matrices where $\mathbf{H}$ is the matrix corresponding to the presence of the factor `group` in the model.

Output of MANOVA: First univariate results

```
    The GLM Procedure
 Class Level Information
       Class         Levels    Values
       group             2    1 2
 Number of observations    13
Dependent Variable: a
                Sum of
Source    DF   Squares  Mean Square F Value Pr > F
Model      1  54.276923  54.276923   19.38 0.0011
Error     11  30.800000   2.800000
Corrd Tot 12  85.076923
R-Square   Coeff Var      Root MSE        a Mean
0.637975    10.21275      1.673320      16.38462
Source    DF  Type ISS  Mean Square F Value Pr > F
group      1  54.276923  54.276923   19.38 0.0011
Source    DF  TypeIIISS Mean Square F Value Pr > F
group      1  54.276923  54.276923   19.38 0.0011

Dependent Variable: b
                Sum of
Source    DF   Squares  Mean Square F Value Pr > F
Model      1  70.892308  70.892308  34.20 0.0001
Error     11  22.800000   2.072727
Corrd Tot 12  93.692308

Dependent Variable: c
                Sum of
Source    DF   Squares Mean Square F Value Pr > F
Model      1   94.77692  94.77692   39.64 <.0001
Error     11   26.30000   2.39090
Corrd Tot 12  121.07692
```

The matrices $\mathbf{E}$ and $\mathbf{H}$.

```
  E = Error SSCP Matrix
   a     b     c
a 30.8  12.2  10.2
b 12.2  22.8   3.8
c 10.2   3.8  26.3
Partial Correlation Coefficients from
    the Error SSCP Matrix / Prob > |r|
DF = 11     a              b              c
a    1.000000       0.460381       0.358383
                      0.1320         0.2527
b    0.460381       1.000000       0.155181
       0.1320                        0.6301
c    0.358383       0.155181       1.000000
       0.2527         0.6301
H = Type III SSCP Matrix for group
     a                 b                 c
a   54.276923077   62.030769231   71.723076923
b   62.030769231   70.892307692   81.969230769
c   71.723076923   81.969230769   94.776923077
```

The eigenvalues of $\mathbf{E}^{-1}\mathbf{H}$.

```
Characteristic Roots and Vectors of: E Inverse * H
H = Type III SSCP Matrix for group
E = Error SSCP Matrix
Characteristic       Characteristic Vector V'EV=1
Root     Percent        a           b           c
5.816159 100.00  0.00403434  0.12874606 0.13332232
0.000000   0.00 -0.09464169 -0.10311602 0.16080216
0.000000   0.00 -0.19278508  0.16868694 0.00000000
MANOVA Test Criteria and Exact F Statistics
 for the Hypothesis of No Overall group Effect
H = Type III SSCP Matrix for group
E = Error SSCP Matrix
S=1    M=0.5    N=3.5
Statistic              Value   F   NumDF DenDF Pr > F
Wilks' Lambda         0.1467 17.45    3     9  0.0004
Pillai's Trace        0.8533 17.45    3     9  0.0004
Hotelling-Lawley Tr 5.8162 17.45    3     9  0.0004
Roy's Greatest Root 5.8162 17.45    3     9  0.0004
```

Things to notice:

1. The conclusion is clear. The mean vectors for the two groups are not the same.

2. The four statistics have the following definitions in terms of eigenvalues of $\mathbf{E}^{-1}\mathbf{H}$:

   Wilk's Lambda:

$$\frac{1}{\prod(1+\lambda_i)} = \frac{1}{6.816}$$

   Pillai's trace:

$$\text{trace}(\mathbf{H}(\mathbf{H}+\mathbf{E})^{-1}) = \sum \frac{\lambda_i}{1+\lambda_i} = \frac{5.816}{6.816}$$

   Hotelling-Lawley trace:

$$\text{trace}(\mathbf{H}\mathbf{E}^{-1}) = \sum \lambda_i = 5.816$$

   Roy's greatest Root:

$$\max\{\lambda_i\} = 5.816$$

1 way layout

Also called m sample problem.

Data $\mathbf{X}_{ij}, j = 1, \ldots, n_i; i = 1, \ldots, m$.

Model $\mathbf{X}_{ij}$ independent $MVN_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$.

First problem of interest: test

$$H_o : \boldsymbol{\mu}_1 = \cdots = \boldsymbol{\mu}_m$$

Based on $\mathbf{E}$ and $\mathbf{H}$. MLE of $\boldsymbol{\mu}_i$ is $\bar{\mathbf{X}}_i$.

$$\mathbf{E} = \sum_{ij} (\mathbf{X}_{ij} - \bar{\mathbf{X}}_i)(\mathbf{X}_{ij} - \bar{\mathbf{X}}_i)^T$$

Under H_o MLE of $\boldsymbol{\mu}$, the common value of the $\boldsymbol{\mu}_i$ is

$$\bar{\bar{\mathbf{X}}} = \frac{\sum_{ij} \mathbf{X}_{ij}}{\sum_i n_i}$$

So

$$\mathbf{E} + \mathbf{H} = \sum_{ij} (\mathbf{X}_{ij} - \bar{\bar{\mathbf{X}}})(\mathbf{X}_{ij} - \bar{\bar{\mathbf{X}}})^T$$

This makes

$$\mathbf{H} = \sum_{ij} (\bar{\mathbf{X}}_i - \bar{\bar{\mathbf{X}}})(\bar{\mathbf{X}}_i - \bar{\bar{\mathbf{X}}})^T$$

Notice can do sum over j to get factor of n_i:

$$\mathbf{H} = \sum_{i} n_i (\bar{\mathbf{X}}_i - \bar{\bar{\mathbf{X}}})(\bar{\mathbf{X}}_i - \bar{\bar{\mathbf{X}}})^T$$

Note: rank of $\mathbf{H}$ is minimum of p and $m-1$. The data

```
1 19 20 18
1 20 21 19
1 19 22 22
1 18 19 21
1 16 18 20
1 17 22 19
1 20 19 20
1 15 19 19
2 12 14 12
2 15 15 17
2 15 17 15
2 13 14 14
2 14 16 13
3 15 14 17
3 13 14 15
3 12 15 15
3 12 13 13
4  8  9 10
4 10 10 12
4 11 10 10
4 11  7 12
```

Code

```
    data three;
    infile 'tab57for3sams';
    input group a b c;
run;
proc print;
run;
proc glm;
    class group;
    model a b c = group;
    manova h=group / printh printe;
run;
    data four;
    infile 'table5.7';
    input group a b c;
run;
proc print;
run;
proc glm;
    class group;
    model a b c = group;
    manova h=group / printh printe;
run;
```

Pieces of output: first set of code does first 3 groups.

So: $\mathbf{H}$ has rank 2.

```
Characteristic Roots & Vectors of: E Inverse * H
 Characteristic     Characteristic Vector V'EV=1
      Root Percent     a        b        c
6.90568180  96.94   0.01115  0.14375 0.08795
0.21795125   3.06  -0.07763 -0.09587 0.16926
0.00000000   0.00  -0.18231  0.13542 0.02083
S=2    M=0    N=5
Statistic         Value    F   NumDF  Den DF Pr > F
Wilks'           0.1039  8.41   6       24   <.0001
Pillai's         1.0525  4.81   6       26   0.0020
Hotelling-Lawley 7.1236 13.79   6   14.353   <.0001
Roy's            6.9057 29.92   3       13   <.0001
NOTE: F Statistic for Roy's is an upper bound.
NOTE: F Statistic for Wilks'is exact.
```

Notice two eigenvalues not 0. Note that exact distribution for Wilk's Lambda is available. Now 4 groups

```
      Root Percent      a        b       c
15.3752900  98.30   0.01128  0.13817 0.08126
0.2307260    1.48  -0.04456 -0.09323 0.15451
0.0356937    0.23  -0.17289  0.09020 0.04777
S=3    M=-0.5    N=6.5
Statistic           Value     F    NumDF Den DF Pr > F
Wilks'         0.04790913 10.12    9   36.657 <.0001
Pillai's       1.16086747  3.58    9       51 0.0016
Hot'ng-Lawley 15.64170973 25.02    9   20.608 <.0001
Roy's         15.37528995 87.13    3       17 <.0001
NOTE: F Statistic for Roy's is an upper bound.
```

Other Hypotheses

How do mean vectors differ? One possibility:

$$\boldsymbol{\mu}_{ik} - \boldsymbol{\mu}_{jk} = c_i - c_j$$

for constants c_i and c_j which do not depend on k. This is an additive model for the means.

Test ?

Define $\alpha = \sum_{ij} \boldsymbol{\mu}_{ij}/(pk)$ Then put

$$\begin{aligned}
\beta_i &= \sum_j \boldsymbol{\mu}_{ij}/p - \alpha, \\
\gamma_j &= \sum_i \boldsymbol{\mu}_{ij}/k - \alpha \\
\tau_{ij} &= \boldsymbol{\mu}_{ij} - \beta_i - \gamma_j - \alpha
\end{aligned}$$

If the τ_{ij} are all 0 then

$$\boldsymbol{\mu}_{ik} - \boldsymbol{\mu}_{jk} = \beta_i - \beta_j$$

so we test the hypothesis that all τ_{ij} are 0.

Univariate Two Way Anova

Data Y_{ijk}

$k = 1, \ldots, n_{ij}; j = 1, \ldots, p; i = 1, \ldots, m.$

Model: independent, $Y_{ijk} \sim N(\mu_{ij}, \sigma^2)$.

Note: this is the **fixed effects** model.

Usual approach: define grand mean, main effects, interactions:

$$\begin{aligned}
\mu &= \sum_{ijk} \mu_{ij} / \sum_{ij} n_{ij} \\
\alpha_i &= \sum_{jk} \mu_{ij} / \sum_{j} n_{ij} - \mu \\
\beta_j &= \sum_{ik} \mu_{ij} / \sum_{i} n_{ij} - \mu \\
\gamma_{ij} &= \mu_{ij} - (\mu + \alpha_i + \beta_j)
\end{aligned}$$

Test additive effects: $\gamma_{ij} = 0$ for all i, j.

Usual test based on ANOVA:

Stack observations Y_{ijk} into vector $\mathbf{Y}$, say.

Estimate μ, α_i, etc by least squares.

Form vectors with entries $\hat{\mu}$, $\hat{\alpha}_i$ etc.

Write

$$\mathbf{Y} = \hat{\boldsymbol{\mu}} + \hat{\boldsymbol{\alpha}} + \hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\gamma}} + \hat{\boldsymbol{\epsilon}}$$

This defines the vector of fitted residuals $\hat{\boldsymbol{\epsilon}}$.

Fact: all vectors on RHS are independent and orthogonal. So:

$$||\mathbf{Y}||^2 = ||\hat{\boldsymbol{\mu}}||^2 + ||\hat{\boldsymbol{\alpha}}||^2 + ||\hat{\boldsymbol{\beta}}||^2 + ||\hat{\boldsymbol{\gamma}}||^2 + ||\hat{\boldsymbol{\epsilon}}||^2$$

This is the ANOVA table. Usually we defined the corrected total sum of squares to be

$$||\mathbf{Y}||^2 - ||\hat{\boldsymbol{\mu}}||^2$$

Our problem is like this one BUT the errors are not modeled as independent.

In the analogy:

i labels group.

j labels the columns: ie j is a, b, c.

k runs from 1 to $n_{ij} = n_i$.

But

$$\text{Cov}(Y_{ijk}, Y_{i'j'k'}) = \begin{cases} \Sigma_{jj'} & i = i', k = k' \\ 0 & \text{otherwise} \end{cases}$$

Now do analysis in SAS.

Tell SAS that the variables A, B and C are **repeated measurements** of the same quantity.

```
proc glm;
    class group;
    model a b c = group;
    repeated scale;
run;
```

The results are as follows:

```
General Linear Models Procedure
Repeated Measures Analysis of Variance
Repeated Measures Level Information
Dependent Variable  A   B    C
Level of SCALE      1   2    3

Manova Test Criteria and Exact F
   Statistics for the Hypothesis of no
   SCALE Effect
H = Type III SS&CP Matrix for SCALE
               E = Error SS&CP Matrix
S=1    M=0    N=7
Statistic
                  Value     F   NumDF DenDF Pr > F
Wilks' Lambda    0.56373  6.1912   2    16  0.0102
Pillai's Trace   0.43627  6.1912   2    16  0.0102
Hotelling-Lawley 0.77390  6.1912   2    16  0.0102
Roy's            0.77390  6.1912   2    16  0.0102
```

Note: should look at interactions first.

```
Manova Test Criteria and F Approximations
 for the Hypothesis of no SCALE*GROUP Effect
S=2    M=0    N=7
Statistic          Value   F   NumDF DenDF Pr > F
Wilks' Lambda    0.56333 1.7725  6    32  0.1364
Pillai's Trace   0.48726 1.8253  6    34  0.1234
Hotelling-Lawley 0.68534 1.7134  6    30  0.1522
Roy's            0.50885 2.8835  3    17  0.0662
NOTE: F Statistic for Roy's Greatest
        Root is an upper bound.
NOTE: F Statistic for Wilks' Lambda is exact.
```

Only weak evidence of interaction. Repeated statement: univariate anova. Results:

```
Repeated Measures Analysis of Variance
Tests of Hypotheses for Between Subjects Effects
Source DF Type III SS Mean Square  F    Pr > F
GROUP   3  743.900000  247.966667 70.93 0.0001
Error  17   59.433333    3.496078
Repeated Measures Analysis of Variance
Univariate Tests of Hypotheses for
          Within Subject Effects
Source: SCALE                       Adj Pr > F
DF TypeIIISS MS      F   Pr > F G - G  H - F
 2  16.624   8.312  5.39 0.0093 0.0101 0.0093
Source: SCALE*GROUP
DF TypeIII   MS    F   Pr > F G - G  H - F
6   18.9619 3.160 2.05 0.0860 0.0889 0.0860
Source: Error(SCALE)
DF TypeIII SS     Mean Square
34  52.4667     1.54313725
Greenhouse-Geisser Epsilon = 0.9664
       Huynh-Feldt Epsilon = 1.2806
```

Greenhouse-Geisser, Huynh-Feldt test to see if Σ has certain structure.

Return to 2 way anova model. Express as:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}$$

For fixed effects model is ϵ_{ijk} iid $N(0, \sigma^2)$.

For MANOVA model vector of ϵ_{ijk} is MVN but with covariance as for Y.

Intermediate model. Put in *subject effect*.

Assume

$$\epsilon_{ijk} = \delta_{ik} + u_{ijk}$$

where u_{ijk} iid $N(0, \sigma^2)$ and δ_{ik} are iid $N(0, \tau^2)$. Then

$$\text{Cov}(Y_{ijk}, Y_{i'j'k'}) = \begin{cases} \sigma^2 + \tau^2 & i', j = j', k = k' \\ \tau^2 & i = i', k = k', j \neq j' \\ 0 & \text{otherwise} \end{cases}$$

This model is usually not fitted by maximum likelihood but by analyzing the behaviour of the ANOVA table under this model.

Essentially model says

$$\boldsymbol{\Sigma} = \tau^2 \mathbf{1}\mathbf{1}^T + \sigma^2 \mathbf{I}$$

GG, HF test for slightly more general pattern for $\boldsymbol{\Sigma}$.

Do univariate anova: The data reordered:

```
1 1 1 19
1 1 2 20
1 1 3 18
2 1 1 20
2 1 2 21
2 1 3 19
 et cetera
2 4 2 10
2 4 3 12
3 4 1 11
3 4 2 10
3 4 3 10
4 4 1 11
4 4 2 7
4 4 3 12
```

The four columns are now labels for subject number, group, scale (a, b or c) and the response.

The sas commands:

```
    data long;
    infile 'table5.7uni';
    input subject group scale score;
run;
proc print;
run;
proc glm;
    class group;
    class scale;
    class subject;
    model score =group subject(group)
                 scale group*scale;
    random subject(group) ;
run;
```

Some of the output:

```
Dependent Variable: SCORE
               Sum of    Mean
Source   DF    Squares  Square F    Pr > F
Model    28   843.5333 30.126 19.52 0.0001
Error    34    52.4667  1.543
Total  62   896.0000
Root MSE      SCORE Mean
1.242231        15.33333


Source          DF  TypeISS    MS      F     Pr > F

GROUP            3 743.9000 247.9667 160.69  0.0001
SUBJECT(GROUP) 17  59.4333   3.4961   2.27  0.0208
SCALE            2  21.2381  10.6190   6.88  0.0031
GROUP*SCALE      6  18.9620   3.1603   2.05  0.0860

Source          DF TypeIIISS     MS     F     Pr > F
GROUP            3 743.9000  247.9667 160.69 0.0001
SUBJECT(GROUP) 17  59.4333    3.4961   2.27 0.0208
SCALE            2  16.6242    8.3121   5.39 0.0093
GROUP*SCALE      6  18.9619    3.1603   2.05 0.0860

Source          Type III Expected Mean Square

GROUP           Var(Error) + 3 Var(SUBJECT(GROUP))
                     + Q(GROUP,GROUP*SCALE)
SUBJECT(GROUP) Var(Error) + 3 Var(SUBJECT(GROUP))
SCALE           Var(Error) + Q(SCALE,GROUP*SCALE)
GROUP*SCALE     Var(Error) + Q(GROUP*SCALE)
```

Type I Sums of Squares:

- *Sequential* sums of squares.
- Each line is a sum of squares comparing the model with effects listed above to one with one extra effect.
- Depend on order terms listed in model.

Type III Sums of Squares:

- Roughly: each line compares model with all other effects in model.
- In unbalanced designs be careful about the differences between Types II, III and IV.

Notice hypothesis of no group by scale interaction is acceptable.

Under the assumption of no such group by scale interaction the hypothesis of no group effect is tested by dividing group ms by subject(group) ms.

Value is 70.9 on 3,17 degrees of freedom.

This is NOT the F value in the table above since the table above is for FIXED effects.

Notice that the sums of squares in this table match those produced in the repeated measures ANOVA. This is not accidental.