smfd8.tex
Day 8. 12.11.2014.

We turn now to a technical result, which is important in reducing n-dimensional problems to one-dimensional ones.

Theorem (Cramér-Wold device). The distribution of a random n-vector $\mathbf{X}$ is completely determined by the set of all one-dimensional distributions of linear combinations $\mathbf{t}^T\mathbf{X} = \sum_i t_i X_i$, where $\mathbf{t}$ ranges over all fixed n-vectors.

Proof. $Y := \mathbf{t}^T\mathbf{X}$ has CF

$$\phi_Y(t) := E\exp\{itY\} = E\exp\{it\mathbf{t}^T\mathbf{X}\}.$$

If we know the distribution of each Y, we know its CF $\phi_Y(t)$. In particular, taking $t = 1$, we know $E\exp\{i\mathbf{t}^T\mathbf{X}\}$. But this is the CF of $\mathbf{X} = (X_1, \cdots, X_n)^T$ evaluated at $\mathbf{t} = (t_1, \cdots, t_n)^T$. But this determines the distribution of $\mathbf{X}$. //

Thus by the Cramér-Wold device, to define an n-dimensional distribution it suffices to define the distributions of *all linear combinations*.

The Cramér-Wold device suggests a way to *define* the multivariate normal distribution. The definition below seems indirect, but it has the advantage of handling the full-rank and singular cases together ($\rho = \pm 1$ as well as $-1 < \rho < 1$ for the bivariate case).

Definition. An n-vector $\mathbf{X}$ has an *n-variate normal* distribution iff $\mathbf{a}^T\mathbf{X}$ has a univariate normal distribution for all constant n-vectors $\mathbf{a}$.

Proposition. (i) Any linear transformation of a multinormal n-vector is multinormal,
(ii) Any vector of elements from a multinormal n-vector is multinormal. In particular, the components are univariate normal.

Proof. (i) If $\mathbf{Y} = \mathbf{AX} + \mathbf{c}$ ($\mathbf{A}$ an $m \times n$ matrix, $\mathbf{c}$ an m-vector) is an m-vector, and $\mathbf{b}$ is any m-vector,

$$\mathbf{b}^T\mathbf{Y} = \mathbf{b}^T(\mathbf{AX} + \mathbf{c}) = (\mathbf{b}^T\mathbf{A})\mathbf{X} + \mathbf{b}^T\mathbf{c}.$$

If $\mathbf{a} = \mathbf{A}^T\mathbf{b}$ (an m-vector), $\mathbf{a}^T\mathbf{X} = \mathbf{b}^T\mathbf{AX}$ is univariate normal as $\mathbf{X}$ is multinormal. Adding the constant $\mathbf{b}^T\mathbf{c}$, $\mathbf{b}^T\mathbf{Y}$ is univariate normal. This holds for

all $\mathbf{b}$, so $\mathbf{Y}$ is m-variate normal.
(ii) Take a suitable matrix $\mathbf{A}$ of 1s and 0s to pick out the required sub-vector.

Theorem 1. If $\mathbf{X}$ is n-variate normal with mean μ and covariance matrix $\mathbf{\Sigma}$, its CF is

$$\phi(\mathbf{t}) := E\exp\{i\mathbf{t}^T\mathbf{X}\} = \exp\{i\mathbf{t}^T\mu - \frac{1}{2}\mathbf{t}^T\mathbf{\Sigma}\mathbf{t}\}.$$

Proof. By Proposition 1, $Y := \mathbf{t}^T\mathbf{X}$ has mean $\mathbf{t}^T\mu$ and variance $\mathbf{t}^T\mathbf{\Sigma}\mathbf{t}$. By definition of multinormality, $Y = \mathbf{t}^T\mathbf{X}$ is univariate normal. So Y is $N(\mathbf{t}^T\mu, \mathbf{t}^T\mathbf{\Sigma}\mathbf{t})$, so Y has CF

$$\phi_Y(t) := E\exp\{itY\} = E\exp\{it\mathbf{t}^T\mathbf{X}\} = \exp\{it\mathbf{t}^T\mu - \frac{1}{2}t^2\mathbf{t}^T\mathbf{\Sigma}\mathbf{t}\}.$$

Taking $t = 1$ (as in the proof of the Cramér-Wold device),

$$E\exp\{i\mathbf{t}^T\mathbf{X}\} = \exp\{i\mathbf{t}^T\mu - \frac{1}{2}\mathbf{t}^T\mathbf{\Sigma}\mathbf{t}\}. \qquad //$$

Corollary. The components of $\mathbf{X}$ are independent iff $\mathbf{\Sigma}$ is diagonal.

Proof. The components are independent iff the joint CF factors into the product of the marginal CFs. This factorization takes place, into $\Pi_j \exp\{i\mu_j t_j - \frac{1}{2}\sigma_{jj}t_j^2\}$, in the diagonal case only. //

Recall that a covariance matrix $\mathbf{\Sigma}$ is always
(a) symmetric ($\sigma_{ij} = \sigma_{ji}$, as $\sigma_{ij} = cov(X_i, X_j)$),
(b) non-negative definite, written $\mathbf{\Sigma} \geq 0$: $\mathbf{a}^T\mathbf{\Sigma}\mathbf{a} \geq 0$ for all n-vectors $\mathbf{a}$.
Suppose that $\mathbf{\Sigma}$ is, further, *positive definite*, written $\mathbf{\Sigma} > 0$:

$$\mathbf{a}^T\mathbf{\Sigma}\mathbf{a} > 0 \qquad \text{unless} \qquad \mathbf{a} = \mathbf{0}.$$

The Multinormal Density.
If $\mathbf{X}$ is n-variate normal, $N(\mu, \mathbf{\Sigma})$, its density (in n dimensions) need not exist (e.g. the singular case $\rho = \pm 1$ with $n = 2$). But if $\mathbf{\Sigma} > \mathbf{0}$ (so $\mathbf{\Sigma}^{-1}$ exists), $\mathbf{X}$ has a density. The link between the multinormal density below and the multinormal MGF above is due to the English statistician F. Y. Edgeworth (1845-1926) in 1893.

Theorem (Edgeworth). If μ is an n-vector, $\mathbf{\Sigma} > \mathbf{0}$ a symmetric positive definite $n \times n$ matrix, then
(i)

$$f(\mathbf{x}) := \frac{1}{(2\pi)^{\frac{1}{2}n}|\mathbf{\Sigma}|^{\frac{1}{2}}} \exp\{-\frac{1}{2}(\mathbf{x}-\mu)^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)\}$$

is an n-dimensional probability density function (of a random n-vector $\mathbf{X}$, say),
(ii) $\mathbf{X}$ has CF $\phi(\mathbf{t}) = \exp\{i\mathbf{t}^T\mu - \frac{1}{2}\mathbf{t}^T\mathbf{\Sigma}\mathbf{t}\}$,
(iii) $\mathbf{X}$ is multinormal $N(\mu, \mathbf{\Sigma})$.

Proof. Write $\mathbf{Y} := \mathbf{\Sigma}^{-\frac{1}{2}}\mathbf{X}$ ($\mathbf{\Sigma}^{-\frac{1}{2}}$ exists as $\mathbf{\Sigma} > \mathbf{0}$, by above). Then $\mathbf{Y}$ has covariance matrix $\mathbf{\Sigma}^{-\frac{1}{2}}\mathbf{\Sigma}(\mathbf{\Sigma}^{-\frac{1}{2}})^T$. Since $\mathbf{\Sigma} = \mathbf{\Sigma}^T$ and $\mathbf{\Sigma} = \mathbf{\Sigma}^{\frac{1}{2}}\mathbf{\Sigma}^{\frac{1}{2}}$, $\mathbf{Y}$ has covariance matrix $\mathbf{I}$ (the components Y_i of $\mathbf{Y}$ are uncorrelated).

Change variables as above, with $\mathbf{y} = \mathbf{\Sigma}^{-\frac{1}{2}}\mathbf{x}$, $\mathbf{x} = \mathbf{\Sigma}^{\frac{1}{2}}\mathbf{y}$. The Jacobian is (taking $\mathbf{A} = \mathbf{\Sigma}^{-\frac{1}{2}}$) $J = \partial\mathbf{x}/\partial\mathbf{y} = det(\mathbf{\Sigma}^{\frac{1}{2}}), = (det\mathbf{\Sigma})^{\frac{1}{2}}$ by the product theorem for determinants. Substituting, the integrand is

$$\exp\{-\frac{1}{2}(\mathbf{x}-\mu)^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)\} = \exp\{-\frac{1}{2}(\mathbf{\Sigma}^{\frac{1}{2}}\mathbf{y}-\mathbf{\Sigma}^{\frac{1}{2}}(\mathbf{\Sigma}^{-\frac{1}{2}}\mu))^T\mathbf{\Sigma}^{-1}(\mathbf{\Sigma}^{\frac{1}{2}}\mathbf{y}-\mathbf{\Sigma}^{\frac{1}{2}}(\mathbf{\Sigma}^{-\frac{1}{2}}\mu))\}.$$

Writing $\nu := \mathbf{\Sigma}^{-\frac{1}{2}}\mu$, this is

$$\exp\{-\frac{1}{2}(\mathbf{y}-\nu)^T\mathbf{\Sigma}^{\frac{1}{2}}\mathbf{\Sigma}^{-1}\mathbf{\Sigma}^{\frac{1}{2}}(\mathbf{y}-\nu)\} = \exp\{-\frac{1}{2}(\mathbf{y}-\nu)^T(\mathbf{y}-\nu)\}.$$

So by the change-of-density formula, $\mathbf{Y}$ has density

$$g(\mathbf{y}) = \frac{1}{(2\pi)^{\frac{1}{2}n}|\mathbf{\Sigma}|^{\frac{1}{2}}}.|\mathbf{\Sigma}|^{\frac{1}{2}}.\exp\{-\frac{1}{2}(\mathbf{y}-\nu)^T(\mathbf{y}-\nu)\}.$$

This factorises as

$$\Pi_{i=1}^n \frac{1}{(2\pi)^{\frac{1}{2}}} \exp\{-\frac{1}{2}(y_i-\nu_i)^2\}.$$

So the components Y_i of $\mathbf{Y}$ are independent $N(\nu_i, 1)$. So $\mathbf{Y}$ is multinormal, $N(\nu, I)$.
(i) Taking $A = B = \mathbb{R}^n$, $\int_{\mathbb{R}^n} f(\mathbf{x})d\mathbf{x} = \int_{\mathbb{R}^n} g(\mathbf{y})d\mathbf{y}, = 1$ as g is a probability density, as above. So f is also a probability density (non-negative and integrates to 1).
(ii) $\mathbf{X} = \Sigma^{\frac{1}{2}}\mathbf{Y}$ is a linear transformation of $\mathbf{Y}$, and $\mathbf{Y}$ is multivariate normal,

$N(\nu, I)$. So $\mathbf{X}$ is multivariate normal.
(iii) $E\mathbf{X} = \mathbf{\Sigma}^{\frac{1}{2}} E\mathbf{Y} = \mathbf{\Sigma}^{\frac{1}{2}}\nu = \mathbf{\Sigma}^{\frac{1}{2}}.\mathbf{\Sigma}^{-\frac{1}{2}}\mu = \mu$, $cov\mathbf{X} = \mathbf{\Sigma}^{\frac{1}{2}} cov\mathbf{Y}(\mathbf{\Sigma}^{\frac{1}{2}})^T = \mathbf{\Sigma}^{\frac{1}{2}}\mathbf{I}\mathbf{\Sigma}^{\frac{1}{2}} = \mathbf{\Sigma}$. So $\mathbf{X}$ is multinormal $N(\mu, \mathbf{\Sigma})$. So its CF is

$$\phi(\mathbf{t}) = \exp\{i\mathbf{t}^T\mu - \frac{1}{2}\mathbf{t}^T\mathbf{\Sigma}\mathbf{t}\}. \qquad //$$

Note. The inverse $\mathbf{\Sigma}^{-1}$ of the covariance matrix $\mathbf{\Sigma}$ is called the *concentration matrix*, K.

Conditional independence of two components X_i, X_j of a multinormal vector given the others can be identified by vanishing of the (off-diagonal) (i, j) entry k_{ij} in the concentration matrix K. The proof needs the results on conditioning and regression in IV.6 D6 below, and the formula for the inverse of a partitioned matrix; see Problems 6.

Independence of Linear Forms

Given a normally distributed random vector $\mathbf{x} \sim N(\mu, \Sigma)$ and a matrix A, one may form the *linear form* $A\mathbf{x}$. One often encounters several of these together, and needs their joint distribution – in particular, to know when these are independent.

Theorem 3. Linear forms $A\mathbf{x}$ and $B\mathbf{x}$ with $\mathbf{x} \sim N(\mu, \Sigma)$ are independent iff

$$A\Sigma B^T = 0.$$

In particular, if A, B are symmetric and $\Sigma = \sigma^2 I$, they are independent iff

$$AB = 0.$$

Proof. The joint CF is

$$\phi(\mathbf{u}, \mathbf{v}) := E\exp\{i\mathbf{u}^T A\mathbf{x} + i\mathbf{v}^T B\mathbf{x}\} = E\exp\{i(A^T\mathbf{u} + B^T\mathbf{v})^T\mathbf{x}\}.$$

This is the CF of $\mathbf{x}$ at argument $\mathbf{t} = A^T\mathbf{u} + B^T\mathbf{v}$, so

$$\phi(\mathbf{u}, \mathbf{v}) = \exp\{i(\mathbf{u}^T A + \mathbf{v}^T B)\mu - \frac{1}{2}(A^T\mathbf{u} + B^T\mathbf{v})^T\Sigma(A^T\mathbf{u} + B^T\mathbf{v})\}$$

$$= \exp\{i(\mathbf{u}^T A + \mathbf{v}^T B)\mu - \frac{1}{2}[\mathbf{u}^T A\Sigma A^T\mathbf{u} + \mathbf{u}^T A\Sigma B^T\mathbf{v} + \mathbf{v}^T B\Sigma A^T\mathbf{u} + \mathbf{v}^T B\Sigma B^T\mathbf{v}]\}.$$

This factorises into a product of a function of **u** and a function of **v** iff the two cross-terms in **u** and **v** vanish, that is, iff $A\Sigma B^T = 0$ and $B\Sigma A^T = 0$; by symmetry of Σ, the two are equivalent.

4. Quadratic forms in normal variates

We give a brief treatment of this important material; for full detail see e.g. [BF], 3.4 – 3.6. Recall (IV.3, D5)
(i) with $x \sim N(\mu, \Sigma)$, linear forms Ax, BX are independent iff $A\Sigma B^T = 0$;
(ii) for a projection, $P^2 = P$ (P is *idempotent*); for a symmetric projection, $P^T P = P$.
We restrict attention, for simplicity, to $\mu = 0$, $\Sigma = \sigma^2 I$, $x \sim N(0, \sigma^2 I)$.

It turns out that the distribution theory relevant to regression depends on *quadratic forms in normal variates*, $x^T Ax$ for a normally distributed random vector x, and that we can confine attention to projection matrices. For P a symmetric projection,

$$x^T Px = x^T P^T Px = (Px)^T (Px),$$

which reduces from *quadratic forms* to *linear* forms – which are much easier! So: if xP_1x, xP_2x are quadratic forms in normal vectors x, with P_1, P_2 projections, $x^T P_1 x$ and $x^T P_2 x$ are independent iff

$$P_1 P_2 = 0 :$$

P_1, P_2 are *orthogonal projections*. Recall that projections P_1, P_2 are *orthogonal* if their ranges are orthogonal subspaces, i.e.

$$(P_1x).(P_2x) = 0 \quad \forall\ x : \quad x^T P_1^T P_2 x = 0 \quad \forall x; \quad P_1^T P_2 = 0 \quad \forall x; \quad P_1 P_2 = 0$$

for P_i symmetric. Note that for P a projection, $I - P$ is a projection orthogonal to it:

$$(I-P)^2 = I-2P+P^2 = 1-2P+P = I-P; \quad P(I-P) = P-P^2 = P-P = 0.$$

If λ is an eigenvalue of A, λ^2 is an eigenvalue of A^2 (check). So if a projection P has eigenvalue λ, $\lambda^2 = \lambda$: $\lambda = 0$ or 1. Also, the trace is the sum of the eigenvalues; for a projection, this is the number of non-zero eigenvalues; this is the rank. So:

For a projection, the eigenvalues are 0 or 1, and the trace is the rank.

By Spectral Decomposition (III.1 D4), a symmetric projection matrix P can be diagonalised by an orthogonal transformation O to a diagonal matrix D:

$$O^T P O = D, \qquad P = O D O^T;$$

as above, the diagonal entries d_{ii} are 0 or 1, and we may re-order so that the 1s come first. So with $y := O^T x$,

$$x^T P x = x^T O D O^T x = y^T D y = y_1^2 + \ldots + y_r^2.$$

Normality is preserved under orthogonal transformations (check!), so also $y \sim N(0, \sigma^2 I)$. So $y_1^2 + \ldots + y_r^2$ is σ^2 times the sum of r independent squares of standard normal variates, and this sum is $\chi^2(r)$ (by definition of chi-square):

$$x^T P x \sim \sigma^2 \chi^2(r).$$

If P has rank r, $I - P$ has rank $n - r$ (where n is the sample size – the dimension of the vector space we are working in):

$$x^T (I - P) x \sim \sigma^2 \chi^2(n - r),$$

and the two quadratic forms are independent.

It turns out that all this can be generalised, to the sum of several projections, not just two. This result – the key to all the distribution theory in Regression – is *Cochran's theorem* (William G. COCHRAN (1909-1980) in 1934); [BF] Th. 3.27):

Theorem (Cochran's Theorem). If

$$I = P_1 + \ldots + P_k$$

with each P_i a symmetric projection with rank n_i, then
(i) the ranks sum: $n = n_1 + \ldots + n_k$;
(ii) each quadratic form $Q_i := x^T P_i x \sim \sigma^2 \chi^2(n_i)$;
(iii) $Q_1, \ldots, Q_k$ are mutually independent;
(iv) $P_1, \ldots, P_k$ are mutually orthogonal: $P_i P_j = 0$ for $i \neq j$.

The quadratic forms that we encounter in Statistics are called *sums of squares (SS)* – for *regression* (SSR), for *error* (SSE), for the *hypothesis* (SSH), etc.

Recall the definition of the *Fisher F-distribution* with degrees of freedom (df) m and n (note the order): $F(m,n)$ is the distribution of the ratio

$$F := \frac{U/m}{V/n},$$

where U, V are independent chi-square random variables with df m, n (see e.g. [BF] 2.3 for the explicit formula for the density, but we shall not need this).

Recall also (or, if you have not met these, take a look at a textbook):
(i) *Analysis of variance (ANOVA)* (see e.g. [BF] Ch. 2). Here one tests for differences between the *means* of different (normal) populations by analysing *variances.* Specifically, one looks at *within-groups* variability and *between-groups* variability, and *rejects* the null hypothesis of no difference between the group means if the second is *too big* compared to the first. As above, one forms the relevant F-statistic, and rejects if F is too big. Here one has *qualitative* factors (which group?).
(ii) *Analysis of Covariance (ANCOVA)* (see e.g. [BF] Ch. 5. Similarly for ANCOVA, where one has both qualitative factors (as with ANOVA) and quantitative ones (covariates), as with Regression.
(iii) Tests of linear hypotheses in Regression (see e.g. [BF] Ch. 6). Here we reject if SSH is too big compared to SSE.

5. Estimation theory for the multivariate normal.

Given a sample $x_1, \ldots, x_n$ from the multivariate normal $N_p(\mu, \Sigma)$, form the *sample mean* (vector) and the *sample covariance matrix* as in the one-dimensional case:

$$\bar{x} := \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad S := \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^T(x_i - \bar{x}).$$

The likelihood for a sample of size 1 is

$$L(x|\mu, \Sigma) = (2\pi)^{-p/2}|\Sigma|^{-1/2}\exp\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\},$$

so the likelihood for a sample of size n is

$$L = (2\pi)^{-np/2}|\Sigma|^{-n/2}\exp\{-\frac{1}{2}\sum_{1}^{n}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)\}.$$

Writing

$$x_i - \mu = (x_i - \bar{x}) - (\mu - \bar{x}),$$

$$\sum_1^n (x_i - \mu)^T \Sigma^{-1} (x_i - \mu) = \sum_1^n (x_i - \bar{x})^T \Sigma^{-1} (x_i - \bar{x}) + n(\bar{x} - \mu)^T \Sigma^{-1} (\bar{x} - \mu)$$

(the cross-terms cancel as $\sum_1^n (x_i - \bar{x}) = 0$). The summand in the first term on the right is a scalar, so is its own trace. Since $trace(AB) = trace(BA)$ and $trace(A + B) = trace(B + A)$,

$$trace(\sum_1^n (x_i - \bar{x})^T \Sigma^{-1} (x_i - \bar{x})) = trace(\Sigma^{-1} \sum_1^n (x_i - \bar{x})^T (x_i - \bar{x}))$$

$$= trace(\Sigma^{-1}.nS) = n\ trace(\Sigma^{-1} S).$$

Combining,

$$L = (2\pi)^{-np/2} |\Sigma|^{-n/2} \exp\{-\frac{1}{2} n\ trace(\Sigma^{-1} S) - \frac{1}{2} n(\bar{x} - \mu)^T \Sigma^{-1} (\bar{x} - \mu)\}.$$

Write

$$V := \Sigma^{-1}$$

('V for variance'); then

$$\ell = const - \frac{1}{2} n\ trace(VS) - (\bar{x} - \mu)^T V (\bar{x} - \mu).$$

So by the Fisher-Neyman Theorem, $(\bar{x}, S)$ is sufficient for (μ, Σ). It is in fact minimal sufficient (Problems 2 Q2).

These natural estimators are in fact the MLEs:

Theorem. For the multivariate normal $N_p(\mu, \Sigma)$, $\bar{x}$ and S are the maximum likelihood estimators for μ, Σ.

Proof. Write $V = (v_{ij}) := \Sigma^{-1}$. By above, the likelihood is

$$L = const.|V|^{n/2} \exp\{-\frac{1}{2} n\ trace(VS) - \frac{1}{2} n(\bar{x} - \mu)^T V (\bar{x} - \mu)\},$$

so the log-likelihood is

$$\ell = c + \frac{1}{2} n \log |V| - \frac{1}{2} n\ trace(VS) - \frac{1}{2} n(\bar{x} - \mu)^T V (\bar{x} - \mu).$$