smfl3.tex
Lecture 3. 21.1.2011

Proof. (i) This is just linearity of the expectation operator E: $Y_i = \sum_j a_{ij} X_j + b_i$, so

$$EY_i = \sum_j a_{ij} EX_j + b_i = \sum_j a_{ij}\mu_j + b_i,$$

for each i. In vector notation, this is $\mu_{\mathbf{Y}} = \mathbf{A}\mu + \mathbf{b}$.
(ii) $Y_i - EY_i = \sum_k a_{ik}(X_k - EX_k) = \sum_k a_{ik}(X_k - \mu_k)$, so

$$cov(Y_i, Y_j) = E[\sum_r a_{ir}(X_r - \mu_r) \sum_s a_{js}(X_s - \mu_s)] = \sum_{rs} a_{ir} a_{js} E[(X_r - \mu_r)(X_s - \mu_s)]$$

$$= \sum_{rs} a_{ir} a_{js} \sigma_{rs} = \sum_{rs} \mathbf{A}_{ir} \mathbf{\Sigma}_{rs} (\mathbf{A}^T)_{sj} = (\mathbf{A}\mathbf{\Sigma}\mathbf{A}^T)_{ij},$$

identifying the elements of the matrix product $\mathbf{A}\mathbf{\Sigma}\mathbf{A}^T$. //

COROLLARY. Covariance matrices $\mathbf{\Sigma}$ are non-negative definite.

Proof. Let $\mathbf{a}$ be any $n \times 1$ matrix (row-vector of length n); then $Y := \mathbf{aX}$ is a scalar. So $Y = Y^T = \mathbf{X}\mathbf{a}^T$. Taking $\mathbf{a} = \mathbf{A}^T, \mathbf{b} = \mathbf{0}$ above, Y has variance [$= 1 \times 1$ covariance matrix] $\mathbf{a}^T \mathbf{\Sigma} \mathbf{a}$. But variances are non-negative. So $\mathbf{a}^T \mathbf{\Sigma} \mathbf{a} \geq \mathbf{0}$ for all n-vectors $\mathbf{a}$. This says that $\mathbf{\Sigma}$ is non-negative definite. //

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. When the MGF exists (as here), $Y := \mathbf{t}^T \mathbf{X}$ has MGF

$$M_Y(s) := E \exp\{sY\} = E \exp\{s\mathbf{t}^T \mathbf{X}\}.$$

If we know the distribution of each Y, we know its MGF $M_Y(s)$. In particular, taking $s = 1$, we know $E \exp\{\mathbf{t}^T \mathbf{X}\}$. But this is the MGF 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 $\mathbf{X}$.

When MGFs do not exist, replace $\mathbf{t}$ by $i\mathbf{t}$ ($i = \sqrt{-1}$) and use characteristic functions (CFs) instead. //

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 $\mathbf{a}$.

First, some properties resulting from the definition.

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 MGF is

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

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 MGF

$$M_Y(s) := E\exp\{sY\} = \exp\{s\mathbf{t}^T\mu + \frac{1}{2}s^2\mathbf{t}^T\mathbf{\Sigma}\mathbf{t}\}.$$

But $E(e^{sY}) = E\exp\{s\mathbf{t}^T\mathbf{X}\}$, so taking $s = 1$ (as in the proof of the Cramér-Wold device),

$$E\exp\{\mathbf{t}^T\mathbf{X}\} = \exp\{\mathbf{t}^T\mu + \frac{1}{2}\mathbf{t}^T\mathbf{\Sigma}\mathbf{t}\},$$

giving the MGF of $\mathbf{X}$ as required. //

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

Proof. The components are independent iff the joint MGF factors into the product of the marginal MGFs. This factorization takes place, into $\Pi_i \exp\{\mu_i t_i + \frac{1}{2}\sigma_{ii}t_i^2\}$, 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: $\mathbf{a}^T\mathbf{\Sigma}\mathbf{a} \geq 0$ for all n-vectors $\mathbf{a}$.
Suppose that $\mathbf{\Sigma}$ is, further, *positive definite*:

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

[We write $\mathbf{\Sigma} > 0$ for '$\mathbf{\Sigma}$ is positive definite', $\mathbf{\Sigma} \geq 0$ for '$\mathbf{\Sigma}$ is non-negative definite'.]

Recall from Linear Algebra (or see III.1 below) that λ is an *eigenvalue* of a matrix $\mathbf{A}$ with *eigenvector* $\mathbf{x}$ ($\neq \mathbf{0}$) if

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

($\mathbf{x}$ is *normalized* if $\mathbf{x}^T\mathbf{x} = \Sigma_i x_i^2 = 1$, as is always possible), and
(i) a symmetric matrix has all its eigenvalues real,
(ii) a non-negative definite matrix has all its eigenvalues non-negative,
(iii) a positive definite matrix is non-singular (has an inverse), and has all its eigenvalues positive.
We quote (III.1, L12 below):

THEOREM (Spectral Decomposition, or Jordan Decomposition). If $\mathbf{A}$ is a symmetric matrix, $\mathbf{A}$ can be written

$$\mathbf{A} = \mathbf{\Gamma}\mathbf{\Lambda}\mathbf{\Gamma}^T,$$

where $\mathbf{\Lambda}$ is a diagonal matrix of eigenvalues of $\mathbf{A}$, $\mathbf{\Gamma}$ is an orthogonal matrix whose columns are normalized eigenvectors.

COROLLARY. (i) For $\mathbf{\Sigma}$ a covariance matrix, we can define its *square root* matrix $\mathbf{\Sigma}^{\frac{1}{2}}$ by $\mathbf{\Sigma}^{\frac{1}{2}} := \mathbf{\Gamma}\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{\Gamma}^T$, $\mathbf{\Lambda}^{\frac{1}{2}} := diag(\lambda_i^{\frac{1}{2}})$, with $\mathbf{\Sigma}^{\frac{1}{2}}\mathbf{\Sigma}^{\frac{1}{2}} = \mathbf{\Sigma}$.
(ii) For $\mathbf{\Sigma}$ a non-singular (i.e. positive definite) covariance matrix, we can define its *inverse square root* matrix $\mathbf{\Sigma}^{-\frac{1}{2}}$ by

$$\mathbf{\Sigma}^{-\frac{1}{2}} := \mathbf{\Gamma}\mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{\Gamma}^T, \qquad \mathbf{\Lambda}^{-\frac{1}{2}} := diag(\lambda^{-\frac{1}{2}}), \qquad \text{with} \qquad \mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{\Lambda}^{-\frac{1}{2}} = \mathbf{\Lambda}^{-1}.$$

THEOREM. If X_i are independent (univariate) normal, any linear combination of the X_i is normal. That is, $\mathbf{X} = (X_1, \cdots, X_n)^T$, with X_i independent normal, is multinormal.

Proof. If X_i are independent $N(\mu_i, \sigma_i^2)$ $\quad (i = 1, \cdots, n)$, $Y := \sum_i a_i X_i + c$ is a linear combination, Y has MGF

$$\begin{aligned}
M_Y(t) &:= E\exp\{t(c + \textstyle\sum_i a_i X_i)\} \\
&= e^{tc} E\Pi \exp\{t a_i X_i\} \qquad \text{(property of exponentials)} \\
&= e^{tc} \Pi E \exp\{t a_i X_i\} \qquad \text{(independence)} \\
&= e^{tc} \Pi \exp\{\mu_i(a_i t) + \frac{1}{2}\sigma_i^2 (a_i t)^2\} \qquad \text{(normal MGF)} \\
&= \exp\{[c + \textstyle\sum_i a_i \mu_i]t + \frac{1}{2}[\textstyle\sum_i a_i^2 \sigma_i^2]t^2\},
\end{aligned}$$

so Y is $N(c + \sum_i a_i \mu_i, \sum_i a_i^2 \sigma_i^2)$, from its MGF. //

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 MGF $M(\mathbf{t}) = \exp\{\mathbf{t}^T \mu + \frac{1}{2}\mathbf{t}^T \mathbf{\Sigma} \mathbf{t}\}$,
(iii) $\mathbf{X}$ is multinormal $N(\mu, \mathbf{\Sigma})$.