smfd7.tex
Day 7. 30.5.2012

2. THE BIVARIATE NORMAL DISTRIBUTION

Recall two of the key ingredients of statistics:
a. *The normal distribution*, $N(\mu, \sigma^2)$:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\{-\frac{1}{2}(x-\mu)^2/\sigma^2\},$$

which has mean $EX = \mu$ and variance $varX = \sigma^2$.
b. *Linear regression by the method of least squares.* This is for *two-dimensional* (or bivariate) data $(X_1, Y_1), \ldots, (X_n, Y_n)$. Two questions arise: (i) Why linear? (ii) What (if any) is the two-dimensional analogue of the normal law?
Mathematical preliminaries. Writing

$$\phi(x) := \frac{1}{\sqrt{2\pi}} \exp\{-\frac{1}{2}x^2\}$$

for the standard normal density, $\int$ for $\int_{-\infty}^{\infty}$, we shall need
(i) *recognising normal integrals:* (a) $\int \phi(x)dx = 1$ (*'normal density'*, (b) $\int x\phi(x)dx = 0$ (*'normal mean'* - or, 'symmetry'), (c) $\int x^2\phi(x)dx = 1$ (*'normal variance'*),
(ii) *completing the square*: as for solving quadratic equations!

In view of the work above, we need an analogue in *two* dimensions of the normal distribution $N(\mu, \sigma^2)$ in one dimension. Just as in one dimension we need *two* parameters, μ and σ, in two dimensions we must expect to need *five*, by above.

Consider the following bivariate density:

$$f(x,y) = c\exp\{-\frac{1}{2}Q(x,y)\},$$

where c is a constant, Q a positive definite quadratic form in x and y:

$$c = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}, \quad Q = \frac{1}{1-\rho^2}\Big[\Big(\frac{x-\mu_1}{\sigma}{}_1\Big)^2 - 2\rho\Big(\frac{x-\mu_1}{\sigma_1}\Big)\Big(\frac{y-\mu_2}{\sigma_2}\Big) + \Big(\frac{y-\mu_2}{\sigma_2}\Big)^2\Big].$$

Here $\sigma_i > 0$, μ_i are real, $-1 < \rho < 1$. Since f is clearly non-negative, to show that f is a (probability) density (function) (in two dimensions), it suffices to show that f integrates to 1:

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dxdy = 1, \qquad \text{or} \qquad \int\int f = 1.$$

Write

$$f_1(x) := \int_{-\infty}^{\infty} f(x,y)dy, \qquad f_2(y) := \int_{-\infty}^{\infty} f(x,y)dx.$$

Then to show $\int\int f = 1$, we need to show $\int_{-\infty}^{\infty} f_1(x)dx = 1$ (or $\int_{-\infty}^{\infty} f_2(y)dy = 1$). Then f_1, f_2 are densities, in *one* dimension. If $f(x,y) = f_{X,Y}(x,y)$ is the *joint* density of *two* random variables X, Y, then $f_1(x)$ is the density $f_X(x)$ of X, $f_2(y)$ the density $f_Y(y)$ of Y (f_1, f_2, or f_X, f_Y, are called the *marginal* densities of the *joint* density f, or $f_{X,Y}$).

To perform the integrations, we have to *complete the square*. We have

$$(1-\rho^2)Q \equiv \Big[\Big(\frac{y-\mu_2}{\sigma_2}\Big) - \rho\Big(\frac{x-\mu_1}{\sigma_1}\Big)\Big]^2 + (1-\rho^2)\Big(\frac{x-\mu_1}{\sigma_1}\Big)^2$$

(reducing the number of occurrences of y to 1, as we intend to integrate out y first). Then (taking the terms free of y out through the y-integral)

$$f_1(x) = \frac{\exp(-\frac{1}{2}(x-\mu_1)^2/\sigma_1^2)}{\sigma_1\sqrt{2\pi}} \int_{-\infty}^{\infty} \frac{1}{\sigma_2\sqrt{2\pi}\sqrt{1-\rho^2}} \exp\Big(\frac{-\frac{1}{2}(y-c_x)^2}{\sigma_2^2(1-\rho^2)}\Big)dy, \tag{$*$}$$

where

$$c_x := \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x-\mu_1).$$

The integral is 1 ('normal density'). So

$$f_1(x) = \frac{\exp(-\frac{1}{2}(x-\mu_1)^2/\sigma_1^2)}{\sigma_1\sqrt{2\pi}},$$

which integrates to 1 ('normal density'), proving

Fact 1. $f(x,y)$ is a joint density function (two-dimensional), with marginal density functions $f_1(x), f_2(y)$ (one-dimensional). So we can write

$$f(x,y) = f_{X,Y}(x,y), \qquad f_1(x) = f_X(x), \qquad f_2(y) = f_Y(y).$$

Fact 2. X, Y are normal: X is $N(\mu_1, \sigma_1^2)$, Y is $N(\mu_2, \sigma_2^2)$. For, we showed $f_1 = f_X$ to be the $N(\mu_1, \sigma_1^2)$ density above, and similarly for Y by symmetry.

Fact 3. $EX = \mu_1, EY = \mu_2, varX = \sigma_1^2, varY = \sigma_2^2$.

This identifies four of the five parameters: two means μ_i, two variances σ_i^2.

Next, recall the definition of conditional probability:

$$P(A|B) := P(A \cap B)/P(B).$$

In the *discrete* case, if X, Y take possible values x_i, y_j with probabilities $f_X(x_i), f_Y(y_j)$, (X, Y) takes possible values (x_i, y_j) with probabilities $f_{X,Y}(x_i, y_j)$:

$$f_X(x_i) = P(X = x_i) = \Sigma_j P(X = x_i, Y = y_j) = \Sigma_j f_{X,Y}(x_i, y_j).$$

Then the *conditional* distribution of Y *given* $X = x_i$ is

$$f_{Y|X}(y_j|x_i) = P(Y = y_j \ \& \ X = x_i)/P(X = x_i) = f_{X,Y}(x_i, y_j)/\Sigma_j f_{X,Y}(x_i, y_j).$$

In the *density* case, we have to replace *sums* by *integrals*. Thus the conditional *density* of Y *given* $X = x$ is

$$f_{Y|X}(y|x) := f_{X,Y}(x, y)/f_X(x) = f_{X,Y}(x, y)/\int_{-\infty}^{\infty} f_{X,Y}(x, y)dy.$$

Returning to the bivariate normal:
Fact 4. The conditional distribution of y given $X = x$ is $N(\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x - \mu_1),\ \sigma_2^2(1 - \rho^2))$.

Proof. Go back to completing the square (or, return to (*) with $\int$ and dy deleted):

$$f(x, y) = \frac{\exp(-\frac{1}{2}(x - \mu_1)^2/\sigma_1^2)}{\sigma_1\sqrt{2\pi}} . \frac{\exp(-\frac{1}{2}(y - c_x)^2/(\sigma_2^2(1 - \rho^2)}{\sigma_2\sqrt{2\pi}\sqrt{1 - \rho^2}}.$$

The first factor is $f_1(x)$, by Fact 2. So, $f_{Y|X}(y|x) = f(x, y)/f_1(x)$ is the second factor:

$$f_{Y|X}(y|x) = \frac{1}{\sqrt{2\pi}\sigma_2\sqrt{1 - \rho^2}} \exp\{-\frac{1}{2}(y - c_x)^2/(\sigma_2^2(1 - \rho^2))\},$$

where c_x is the linear function of x given below (*). This not only completes the proof of Fact 4, but gives
Fact 5. The conditional mean $E(Y|X = x)$ is *linear* in x:

$$E(Y|X = x) = \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x - \mu_1).$$

Note. This simplifies when X and Y are equally variable, $\sigma_1 = \sigma_2$:

$$E(Y|X = x) = \mu_2 + \rho(x - \mu_1)$$

(recall $EX = \mu_1, EY = \mu_2$). Recall that in Galton's height example, this says: for every inch of mid-parental height above/below the average, $x - \mu_1$, the parents pass on to their child, *on average*, ρ inches, and continuing in this way: *on average*, after n generations, each inch above/below average becomes *on average* ρ^n inches, and $\rho^n \to \infty$ as $n \to \infty$, giving *regression towards the mean*.
(A regression function is a *conditional mean* – see Section 5.)
Fact 6. The conditional variance of Y given $X = x$ is

$$var(Y|X = x) = \sigma_2^2(1 - \rho^2).$$

Recall (Fact 3) that the variability (= variance) of Y is $varY = \sigma_2^2$. By Fact 5, the variability remaining in Y when X is given (i.e., not accounted for by knowledge of X) is $\sigma_2^2(1 - \rho^2)$. Subtracting: the variability of Y which *is* accounted for by knowledge of X is $\sigma_2^2\rho^2$. That is: ρ^2 is the *proportion of the variability* of Y accounted for by knowledge of X. So ρ is a measure of the *strength of association* between Y and X.

Recall that the *covariance* is defined by

$$cov(X, Y) := E[(X-EX)(Y-EY)] = E[(X-\mu_1)(Y-\mu_2)] = E(XY)-(EX)(EY),$$

and the *correlation coefficient* ρ, or $\rho(X, Y)$, defined by

$$\rho = \rho(X, Y) := cov(X, Y)/(\sqrt{varX}\sqrt{varY}) = E[(X - \mu_1)(Y - \mu_2)]/\sigma_1\sigma_2$$

is the usual measure of the strength of association between X and Y ($-1 \leq \rho \leq 1$; $\rho = \pm 1$ iff one of X, Y is a function of the other).
Fact 7. The correlation coefficient of X, Y is ρ.
Proof.

$$\rho(X, Y) := E\Big[\Big(\frac{X - \mu_1}{\sigma_1}\Big)\Big(\frac{Y - \mu_2}{\sigma_2}\Big)\Big] = \int\int \Big(\frac{x - \mu_1}{\sigma_1}\Big)\Big(\frac{y - \mu_2}{\sigma_2}\Big) f(x, y)dxdy.$$

Substitute for $f(x, y) = c\exp(-\frac{1}{2}Q)$, and make the change of variables $u := (x - \mu_1)/\sigma_1$, $v := (y - \mu_2)/\sigma_2$:

$$\rho(X, Y) = \frac{1}{2\pi\sqrt{1 - \rho^2}} \int\int uv \exp\{-\frac{1}{2}[u^2 - 2\rho uv + v^2]/(1 - \rho^2)\}dudv.$$

Completing the square, $[u^2 - 2\rho uv + v^2] = (v - \rho u)^2 + (1 - \rho^2)u^2$. So

$$\rho(X, Y) = \frac{1}{\sqrt{2\pi}} \int u \exp(-\frac{1}{2}u^2)du. \frac{1}{\sqrt{2\pi}\sqrt{1 - \rho^2}} \int v \exp\{-\frac{1}{2}(v-\rho u)^2/(1-\rho^2)\}dv.$$

Replace v in the inner integral by $(v-\rho u)+\rho u$, and calculate the two resulting integrals separately. The first is zero ('normal mean', or symmetry), the second is ρu ('normal density'). So

$$\rho(X,Y) = \frac{1}{\sqrt{2\pi}}.\rho \int u^2 \exp(-\frac{1}{2}u^2)du = \rho$$

('normal variance'), as required.

This completes the identification of all five parameters in the bivariate normal distribution: two means μ_i, two variances σ_i^2, one correlation ρ.
Note 1. The above holds for $-1 < \rho < 1$; always, $-1 \leq \rho \leq 1$. In the limiting cases $\rho = \pm 1$, one of X, Y is a linear function of the other: $Y = aX + b$, say, as with temperature (Fahrenheit and Centigrade). The situation is not really two-dimensional: we can (and should) use only *one* of X and Y, reducing to a one-dimensional problem.
Note 2. The slope of the regression line $y = c_x$ is $\rho\sigma_2/\sigma_1 = (\rho\sigma_1\sigma_2)/(\sigma_1^2)$, which can be written as $cov(X,Y)/varX = \sigma_{12}/\sigma_{11}$, or σ_{12}/σ_1^2: the line is

$$y - EY = \frac{\sigma_{12}}{\sigma_{11}}(x - EX).$$

This is the *population* version (what else?!) of the *sample regression line*

$$y - \bar{Y} = \frac{S_{XY}}{S_{XX}}(x - \bar{X}),$$

from linear regression (Section 1).

The case $\rho = \pm 1$ – apparently two-dimensional, but really one-dimensional – is *singular*; the case $-1 < \rho < 1$ - genuinely two-dimensional - is *non-singular*, or (see below) *full rank*.

We note in passing
Fact 8. The bivariate normal law has *elliptical contours*. For, the contours are $Q(x,y) = const$, which are ellipses (as Galton found).
Moment Generating Function (MGF). Recall $M(t) := E(e^{tX})$. For X $N(\mu, \sigma^2)$, $M_X(t) = \exp(\mu t + \frac{1}{2}\sigma^2 t^2)$ [SP, Problems 5]. So $M_{X-\mu}(t) = \exp(\frac{1}{2}\sigma^2 t^2)$, and the CF is $\phi_{X-\mu}(t) = \exp(-\frac{1}{2}\sigma^2 t^2)$. Then (check) $\mu = EX = M'_X(0)$, $varX = E[(X-\mu)^2] = M''_{X-\mu}(0)$.

Similarly in the bivariate case: the MGF is

$$M_{X,Y}(t_1, t_2) := E \exp(t_1 X + t_2 Y).$$

For the bivariate normal,

$$M(t_1,t_2) = E(\exp(t_1X + t_2Y)) = \int\int \exp(t_1x + t_2y)f(x,y)dxdy$$

$$= \int \exp(t_1x)f_1(x)dx \int \exp(t_2y)f(y|x)dy.$$

The inner integral is the MGF of $Y|X = x$, which is $N(c_x, \sigma_2^2, (1-\rho^2))$, so is $\exp(c_xt_2 + \frac{1}{2}\sigma_2^2(1-\rho^2)t_2^2)$. By Fact 4, $c_xt_2 = [\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x-\mu_1)]t_2$, so

$$M(t_1,t_2) = \exp(t_2\mu_2 - t_2\frac{\sigma_2}{\sigma_1}\mu_1 + \frac{1}{2}\sigma_2^2(1-\rho^2)t_2^2)\int \exp([t_1 + t_2\rho\frac{\sigma_2}{\sigma_1}]x)f_1(x)dx.$$

Since $f_1(x)$ is $N(\mu_1, \sigma_1^2)$, the inner integral is a normal MGF, which is thus $\exp(\mu_1[t_1 + t_2\rho\frac{\sigma_2}{\sigma_1}] + \frac{1}{2}\sigma_1^2[\ldots]^2)$. Combining the two terms and simplifying:
Fact 9. The joint MGF and joint CF of X, Y are

$$M_{X,Y}(t_1,t_2) = M(t_1,t_2) = \exp(\mu_1t_1 + \mu_2t_2 + \frac{1}{2}[\sigma_1^2t_1^2 + 2\rho\sigma_1\sigma_2t_1t_2 + \sigma_2^2t_2^2]),$$

$$\phi_{X,Y}(t_1,t_2) = \phi(t_1,t_2) = \exp(i\mu_1t_1 + i\mu_2t_2 - \frac{1}{2}[\sigma_1^2t_1^2 + 2\rho\sigma_1\sigma_2t_1t_2 + \sigma_2t_2^2]).$$

Fact 10. X, Y are independent if and only if $\rho = 0$.
Proof. For densities: X, Y are independent iff the joint density $f_{X,Y}(x,y)$ *factorises* as the *product* of the marginal densities $f_X(x).f_Y(y)$. For MGFs: X, Y are independent iff the joint MGF $M_{X,Y}(t_1,t_2)$ *factorises* as the *product* of the marginal MGFs $M_X(t_1).M_Y(t_2)$. From Fact 9, this occurs iff $\rho = 0$. Similarly with CFs, if we prefer to work with them.
Note. X, Y independent implies X, Y uncorrelated ($\rho = 0$) in general (when the correlation exists). The converse if *false* in general, but *true*, by Fact 10, in the bivariate normal case.

3. THE MULTIVARIATE NORMAL DISTRIBUTION.

With one regressor, we used the bivariate normal distribution as above. Similarly for two regressors, we use the trivariate normal. With any number of regressors, as here, we need a general *multivariate normal* - or *'multinormal'* - distribution in n dimensions. We must expect that in n dimensions, to handle a random n-vector $\mathbf{X} = (X_1, \cdots, X_n)^T$, we will need
(i) a *mean vector* $\mu = (\mu_1, \cdots, \mu_n)^T$ with $\mu_i = EX_i$, $\mu = E\mathbf{X}$,
(ii) a *covariance matrix* $\mathbf{\Sigma} = (\sigma_{ij})$, with $\sigma_{ij} = cov(X_i, X_j)$: $\mathbf{\Sigma} = cov\mathbf{X}$.

First, note the effect of a linear transformation:

PROPOSITION 1. If $\mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{b}$, with $\mathbf{Y}, \mathbf{b}$ m-vectors, $\mathbf{A}$ an $m \times n$ matrix and $\mathbf{X}$ an n-vector,
(i) the mean vectors are related by $E\mathbf{Y} = \mathbf{A}E\mathbf{X} + \mathbf{b} = \mathbf{A}\mu + \mathbf{b}$,
(ii) the covariance matrices are related by $\mathbf{\Sigma_Y} = \mathbf{A\Sigma A}^T$.

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

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

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

$$cov(Y_i, Y_j) = E[\sum\nolimits_r a_{ir}(X_r - \mu_r)\sum\nolimits_s a_{js}(X_s - \mu_s)] = \sum\nolimits_{rs} a_{ir}a_{js}E[(X_r - \mu_r)(X_s - \mu_s)]$$

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

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

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 $Y := \mathbf{a}\mathbf{X}$ 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 $\mathbf{y} = \mathbf{A}\mathbf{X} + \mathbf{c}$ ($\mathbf{A}$ an $m \times n$ matrix, $\mathbf{c}$ an m-vector) is an m-vector, and $\mathbf{b}$ is any m-vector,

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

If $\mathbf{a} = \mathbf{A}^T\mathbf{b}$ (an m-vector), $\mathbf{a}^T\mathbf{X} = \mathbf{b}^T\mathbf{A}\mathbf{X}$ 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 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 (for use now – though we shall prove a more general result, the Singular Values Decomposition (SVD) in VII.2, Day 12, 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 $N(\mu_i, \sigma_i^2)$ $\qquad (i = 1, \cdots, n)$, $Y := \sum_i a_i X_i + c$ is a linear combination, Y has MGF

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

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 $N(\mu, \mathbf{\Sigma})$.

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

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

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

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

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

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

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 $A = B = \mathbf{R}^n$, $\int_{\mathbf{R}^n} f(\mathbf{x})d\mathbf{x} = \int_{\mathbf{R}^n} g(\mathbf{y})d\mathbf{y}, = 1$ 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) $E\mathbf{X} = \sigma^{\frac{1}{2}} E\mathbf{Y} = \sigma^{\frac{1}{2}}\nu = \sigma^{\frac{1}{2}}.\sigma^{-\frac{1}{2}}\mu = \mu$, $cov\mathbf{X} = \sigma^{\frac{1}{2}} cov\mathbf{Y}(\sigma^{\frac{1}{2}})^T = \sigma^{\frac{1}{2}}\mathbf{I}\sigma^{\frac{1}{2}} = \sigma$. So $\mathbf{X}$ is multinormal $N(\mu, \sigma)$. So its MGF is

$$M(\mathbf{t}) = \exp\{\mathbf{t}^T\mu + \frac{1}{2}\mathbf{t}^T\sigma\mathbf{t}\}. \qquad //$$

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 MGF is

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

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

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

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