smfd8.tex
Day 8. 1.6.2012

4. 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 $\sum(x_i - \bar{x}) = 0$). 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,

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

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 8).

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

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

The MLE $\hat{\mu}$ for μ is $\bar{x}$, as this reduces the last term (the only one involving μ) to its minimum value, 0. For a square matrix $A = (a_{ij})$, its determinant is

$$|A| = \sum_j a_{ij} A_{ij}$$

for each i, or

$$|A| = \sum_i a_{ij} A_{ij}$$

for each j, expanding by the ith row or jth column, where A_{ij} is the *cofactor* (signed minor) of a_{ij}. From either,

$$\partial |A| / \partial a_{ij} = A_{ij},$$

so

$$\partial \log |A| / \partial a_{ij} = A_{ij} / |A| = (A^{-1})_{ji},$$

the (j, i) element of A^{-1}, recalling the formula for the matrix inverse (or $(A^{-1})_{ij}$ if A is symmetric). Also, if B is symmetric,

$$trace(AB) = \sum_i \sum_j a_{ij} b_{ji} = \sum_{i,j} a_{ij} b_{ij},$$

so

$$\partial trace(AB)/\partial a_{ij} = b_{ij}.$$

Using these, and writing $S = (s_{ij})$,

$$\partial \log |V|/\partial v_{ij} = (V^{-1})_{ij} = (\Sigma)_{ij} = \sigma_{ij} \qquad (V := \Sigma^{-1}),$$

$$\partial trace(VS)/\partial v_{ij} = s_{ij}.$$

So

$$\partial \ell/\partial v_{ij} = \frac{1}{2} n(\sigma_{ij} - s_{ij}),$$

which is 0 for all i and j iff $\Sigma = S$. This says that S is the MLE for Σ, as required. //

5. CONDITIONING AND REGRESSION

Recall that the *conditional* density of Y *given* $X = x$ is

$$f_{Y|X}(y|x) := f_{X,Y}(x,y)/\int f_{X,Y}(x,y)dy.$$

Conditional means.

The conditional mean of Y given $X = x$ is

$$E(Y|X = x),$$

a function of x called the *regression* function (of Y on x). So, if we do not specify the value x, we get $E(Y|X)$. This is *random*, because X is random (until we observe its value, x; then we get the regression function of x as above). As $E(Y|X)$ is random, we can look at its mean and variance.

Recall (SP, Ch. II)

THEOREM (Conditional Mean Formula). $E[E(Y|X)] = EY$.

Interpretation. EY takes the random variable Y, and averages out all the randomness to give a number, EY.
$E(Y|X)$ takes the random variable Y, and averages out all the randomness in Y NOT accounted for by knowledge of X.
$E[E(Y|X)]$ then averages out the remaining randomness, which IS accounted

for by knowledge of X, to give EY as above.
Example: Bivariate normal distribution, $N(\mu_1, \mu_2; \sigma_1^2, \sigma_2^2; \rho)$, or $N(\mu, \sigma)$,

$$\mu = (\mu_1, \mu_2)^T, \qquad \sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}.$$

Then

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

So

$$E[E(Y|X)] = \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(EX - \mu_1) = \mu_2 = EY, \qquad \text{as} \qquad EX = \mu_1.$$

As with the bivariate normal, we should keep some concrete instance in mind as a motivating example, e.g.:
X = incoming score of student [in medical school or university, say], Y = graduating score;
X = child's height at 2 years (say), Y = child's eventual adult height,
or X = mid-parent height, Y = child's adult height, as in Galton's study.

Recall also (SP, Ch. II)

THEOREM (Conditional Variance Formula).

$$varY = E_X var(Y|X) + var_X E(Y|X).$$

Interpretation.

$$varY = \text{total variability in } Y,$$

$$E_X var(Y|X) = \text{variability in } Y \text{ not accounted for by knowledge of } X,$$

$$var_X E(Y|X) = \text{variability in } Y \text{ accounted for by knowledge of } X.$$

Example: the bivariate normal.

$$Y|X = x \text{ is } N(\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x - \mu_1), \sigma_2^2(1 - \rho^2)), \qquad varY = \sigma_2^2,$$

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

which has variance $(\rho\sigma_2/\sigma_1)^2 varX = (\rho\sigma_2/\sigma_1)^2\sigma_1^2 = \rho^2\sigma_2^2$;

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

COROLLARY. $E(Y|X)$ has the same mean as Y and smaller variance (if anything) than Y.

Proof. From the Conditional Mean Formula, $E[E(Y|X)] = EY$. Since $var(Y|X) \geq 0$, $E_X var(Y|X) \geq 0$, so

$$varE[Y|X] \leq varY$$

from the Conditional Variance Formula. //

This result has important applications in estimation theory. Suppose we are to estimate a parameter θ, and are considering a statistic X as a possible estimator (or basis for an estimator) of θ. We would naturally want X to contain all the information on θ contained within the entire sample. What (if anything) does this mean in precise terms? The answer lies in the concept of *sufficiency* ('data reduction') - one of the most important contributions to statistics of the great English statistician R. A. (Sir Ronald) Fisher (1880-1962) in 1920. In the language of sufficiency, the Conditional Variance Formula is seen as (essentially) the Rao-Blackwell Theorem, a key result in the area (see the index in your favourite Statistics book for more).

Regression.

In the bivariate normal, with X = mid-parent height, Y = child's height, $E(Y|X = x)$ is linear in x (*regression line*). In a more detailed analysis, with U = father's height, V = mother's height, Y = child's height, one would expect $E(Y|U = u, V = v)$ to be linear in u and v (*regression plane*), etc.

In an n-variate normal distribution $N_n(\mu, \sigma)$, suppose that $\mathbf{X} = (X_1, \cdots, X_n)$ is partitioned into $\mathbf{X}_1 := (X_1, \cdots, X_r)^T$ and $\mathbf{X}_2 := (X_{r+1}, \cdots, X_n)^T$. Let the corresponding partition of the mean vector and the covariance matrix be

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix},$$

where $E\mathbf{X}_i = \mu_i$, σ_{11} is the covariance matrix of $\mathbf{X}_1$, σ_{22} that of $\mathbf{X}_2$, $\sigma_{12} = \sigma_{21}^T$ the covariance matrix of $\mathbf{X}_1$ with $\mathbf{X}_2$.

We restrict attention, for simplicity, to the non-singular case, where σ is positive definite.

LEMMA. If σ is positive definite, so is σ_{11}.

Proof. $\mathbf{x}^T\sigma\mathbf{x} > \mathbf{0}$ as σ is positive definite. Take $\mathbf{x} = (\mathbf{x}_1, \mathbf{0})^T$, where $\mathbf{x}_1$ has the same number of components as the order of σ_{11} [i.e., in matrix language, so that the partition of $\mathbf{x}$ is conformable with those of μ and σ above]. Then $\mathbf{x}_1\sigma_{11}\mathbf{x}_1 > 0$ for all $\mathbf{x}_1$. This says that σ_{11} is positive definite, as required. //

THEOREM. The conditional distribution of $\mathbf{X}_2$ given $\mathbf{X}_1 = \mathbf{x}_1$ is

$$\mathbf{X}_2|\mathbf{X}_1 = \mathbf{x}_1 \sim N(\mu_2 + \sigma_{21}\sigma_{11}^{-1}(\mathbf{x}_1 - \mu_1), \sigma_{22} - \sigma_{21}\sigma_{11}^{-1}\sigma_{12}).$$

COROLLARY. The regression of $\mathbf{X}_2$ on $\mathbf{X}_1$ is linear:

$$E(\mathbf{X}_2|\mathbf{X}_1 = \mathbf{x}_1) = \mu_2 + \sigma_{21}\sigma_{11}^{-1}(\mathbf{x}_1 - \mu_1).$$

Proof. Recall that $\mathbf{AX}, \mathbf{BX}$ are independent iff $\mathbf{A}\sigma\mathbf{B}^T = \mathbf{0}$, or as σ is symmetric, $\mathbf{B}\sigma\mathbf{A}^T = \mathbf{0}$. Now

$$\mathbf{X}_1 = \mathbf{AX} \text{ where } \mathbf{A} = (\mathbf{I}, \mathbf{0}),$$

$$\mathbf{X}_2 - \sigma_{21}\sigma_{11}^{-1}\mathbf{X}_1 = \begin{pmatrix} -\sigma_{21}\sigma_{11}^{-1} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix} = \mathbf{BX}, \text{ where } \mathbf{B} = \begin{pmatrix} -\sigma_{21}\sigma_{11}^{-1} & \mathbf{I} \end{pmatrix}.$$

Now

$$\mathbf{B}\sigma\mathbf{A}^T = \begin{pmatrix} -\sigma_{21}\sigma_{11}^{-1} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix} \begin{pmatrix} \mathbf{I} \\ \mathbf{0} \end{pmatrix} = \begin{pmatrix} -\sigma_{21}\sigma_{11}^{-1} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \sigma_{11} \\ \sigma_{21} \end{pmatrix}$$

$$= -\sigma_{21}\sigma_{11}^{-1}\sigma_{11} + \sigma_{21} = \mathbf{0},$$

so $\mathbf{X}_1$ and $\mathbf{X}_2 - \sigma_{21}\sigma_{11}^{-1}\mathbf{X}_1$ are *independent*. Since both are linear transformations of $\mathbf{X}$, which is multinormal, both are *multinormal*. Also,

$$E(\mathbf{BX}) = \mathbf{B}E\mathbf{X} = \begin{pmatrix} -\sigma_{21}\sigma_{11}^{-1} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} = \mu_2 - \sigma_{21}\sigma_{11}^{-1}\mu_1.$$

To calculate the covariance matrix, introduce $\mathbf{C} := -\sigma_{21}\sigma_{11}^{-1}$, so $\mathbf{B} = (\mathbf{C}\ \mathbf{I})$, and recall $\sigma_{12}^T = \sigma_{21}$, so $\mathbf{C}^T = -\sigma_{11}^{-1}\sigma_{12}$:

$$var(\mathbf{BX}) = \mathbf{B}\sigma\mathbf{B}^T = \begin{pmatrix} \mathbf{C} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix} \begin{pmatrix} \mathbf{C}^T \\ \mathbf{I} \end{pmatrix}$$

$$= \begin{pmatrix} \mathbf{C} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \sigma_{11}\mathbf{C}^T + \sigma_{12} \\ \sigma_{21}\mathbf{C}^T + \sigma_{22} \end{pmatrix} = \mathbf{C}\sigma_{11}\mathbf{C}^T + \mathbf{C}\sigma_{12} + \sigma_{21}\mathbf{C}^T + \sigma_{22}$$

$$= \sigma_{21}\sigma_{11}^{-1}\sigma_{11}\sigma_{11}^{-1}\sigma_{12} - \sigma_{21}\sigma_{11}^{-1}\sigma_{12} - \sigma_{21}\sigma_{11}^{-1}\sigma_{12} + \sigma_{22}$$

$$= \sigma_{22} - \sigma_{21}\sigma_{11}^{-1}\sigma_{12}.$$

By independence, the conditional distribution of $\mathbf{BX}$ given $\mathbf{X}_1 = \mathbf{AX}$ is the same as its marginal distribution, which by above is $N(\mu_2 - \sigma_{21}\sigma_{11}^{-1}\mu_1, \sigma_{22} - \sigma_{21}\sigma_{11}^{-1}\sigma_{12})$. So given $\mathbf{X}_1$, $\mathbf{X}_2 - \sigma_{21}\sigma_{11}^{-1}\mathbf{X}_1$ is $N(\mu_2 - \sigma_{21}\sigma_{11}^{-1}\mu_1, \sigma_{22} - \sigma_{21}\sigma_{11}^{-1}\sigma_{12})$.

To pass from the conditional distribution of $\mathbf{X}_2 - \sigma_{21}\sigma_{11}^{-1}\mathbf{X}_1$ given $\mathbf{X}_1$ to that of $\mathbf{X}_2$ given $\mathbf{X}_1$: just add $\sigma_{21}\sigma_{11}^{-1}\mathbf{X}_1$. Then

$$\mathbf{X}_2|\mathbf{X}_1 \sim N(\mu_2 + \sigma_{21}\sigma_{11}^{-1}(\mathbf{X}_1 - \mu_1), \sigma_{22} - \sigma_{21}\sigma_{11}^{-1}\sigma_{12}). \quad //$$

Here $\sigma_{22} - \sigma_{21}\sigma_{11}^{-1}\sigma_{12}$ is called the *partial covariance matrix* of $\mathbf{X}_2$ given $\mathbf{X}_1$.

VI. TIME SERIES (TS).

1. STATIONARY PROCESSES AND AUTOCORRELATION

A TS - a sequence of observations indexed by time - may well exhibit, on visual inspection after plotting, a *trend* - a tendency to increase or decrease with time, or *seasonality*, or both. However, the simplest case is where trend and seasonality are absent, and we begin with this. Furthermore, even if they are present, our first task may well be to remove them, by *detrending* and/or *seasonal adjustment*.

Definition. A TS, or stochastic process, is *strictly stationary* if its finite-dimensional distributions are invariant under time-shifts - that is, if for all n, $t_1, \cdots, t_n$ and h, $(X_{t_1}, \cdots, X_{t_n})$ and $(X_{t_1+h}, \cdots, X_{t_n+h})$ have the same distribution. In particular, for a stationary TS:

(i) taking $n = 1$, the marginal distribution of X_t is the same for all t, so the mean of X_t (if it is defined, as we shall assume) is constant, $= \mu$ say, and so is its variance (if defined, as we shall also assume), $= \sigma^2$ say:

$$EX_t = \mu, \qquad varX_t = \sigma^2 \qquad \text{for all } t.$$

(ii) Taking $n = 2$, the distributions of (X_{t_1}, X_{t_2}) is the same as that of (X_{t_1+h}, X_{t_2+h}), and so depends only on the *time-difference* $t_2 - t_1$, called the *lag*. With lag τ, it thus suffices to consider the distribution of $(X_t, X_{t+\tau})$, which depends only on the lag τ, not the time t. In particular, the covariance $cov(X_t, X_{t+\tau})$ is a function of τ only, $\gamma(\tau)$ say:

$$cov(X_t, X_{t+\tau}) = \gamma(\tau) \qquad \text{for all } t$$

(note that $\gamma(0) = varX_t = \sigma^2$, for all t). Similarly for the correlation:

$$corr(X_t, X_{t+\tau}) = \gamma(\tau)/\gamma(0) = \rho(\tau),$$

say (note that $\rho(0) = 1$).

Definition. The function

$$\rho(\tau) := corr(X_t, X_{t+\tau})$$

is called the *autocorrelation function* of the (strictly) stationary process (X_t).

Note. 1. If X_t is normal (Gaussian), its distribution (that is, the set of its finite-dimensional distributions) is completely determined by its means and covariances (equivalently, variances and correlations), μ and $\gamma(\tau)$ or $\rho(\tau)$.

Sometimes, however, we do not want to make the very strong assumption of normality, but only need to specify the distribution of the process as far as its means and covariances/correlations. As these involve only the one- and two-dimensional distributions, they are called the *second-order properties* of the TS or stochastic process.
2. Since covariance and correlation are commutative – $cov(X,Y) = cov(Y,X)$ and $corr(X,Y) = corr(Y,X)$ –

$$\gamma(-\tau) = \gamma(\tau), \qquad \rho(-\tau) = \rho(\tau).$$

So we can think of the lag just as a time-difference – it does not matter whether we think forwards in time or backwards in time.

Definition. A process (X_t) whose means and variances exist is called *weakly stationary* (covariance stationary, second-order stationary, wide-sense stationary) if its mean EX_t is constant over time and its covariance $cov(X_t, X_{t+\tau})$ depends only on the lag τ and not on the time t. We then use the notation $EX_t = \mu$, $cov(X_t, X_{t+\tau}) = \gamma(\tau)$, $corr(X_t, X_{t+\tau}) = \rho(\tau)$ as above.
Note. 1. A strictly stationary process is always weakly stationary. The converse is false in general but true for the normal (Gaussian) case.
2. For brevity, we now abbreviate 'weakly stationary' to 'stationary'. We will continue to say 'strictly stationary', unless the process is normal (Gaussian), when the strictness is automatic (by above), so can be understood.
White Noise. The simplest possible case of stationarity is $\mu = EX_t = 0$, $\gamma(\tau) = \sigma^2\rho(\tau)$, where $\rho(\tau) = corr(X_t, X_{t+\tau})$ is 1 for $\tau = 0$ and 0 otherwise. Such processes exist in three levels of generality:
(i) no further restriction (distinct X_t uncorrelated, but may be dependent);
(ii) distinct X_t independent;
(iii) (X_t) normal (Gaussian) - so distinct X_t are independent, because uncorrelated.

The term *white noise* (WN) is used for some/all such cases, or $WN(\sigma^2)$ if the variance σ^2 needs mention.
Note. The term shows clearly its engineering origins. The word 'noise' derives from radio engineering (for instance, spontaneous thermal fluctuations, or 'shot noise', in thermionic valves), and telephone engineering. It is also used in telecommunications, where the 'noise' – random error or disturbances – may be visual rather than aural (recall that optical fibres are used nowadays in cables for long-distance communication, with photons playing the

role of electrons in the traditional telephone cables). The term 'white' is by analogy with white rather than coloured light. In the language of spectral theory, white noise has a *flat spectrum* (a 'uniform mixture' of frequencies - just as white light is a mixture of the colours of the rainbow).
3. We shall use definition (ii) of white noise for convenience. Independence will allow us to use LLN and CLT.
4. White noise is specific to *discrete* time. A process with correlation

$$\rho(\tau) = \begin{cases} 1 & (\tau = 0) \\ 0 & (\tau \neq 0) \end{cases}$$

is realistic in discrete time (such as the white noise above), but would be pathological (and physically unrealisable) in *continuous* time, because of the discontinuity in the correlation function. However, the process corresponding to the integrated version of white noise in continuous time does exist and is extremely important: *Brownian motion* (SP, Ch. III).