smfd9(13a).tex
Day 9. 8.11.2013.

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. //

6. Conditioning and regression

In general, we should always *use what we know*. In Probability and Statistics, this goes by the technical term of *conditioning*. This rests ultimately on the formula $P(A|B) := P(A \cap B)/P(B)$ of elementary probability (applicable only when $P(B) > 0$!), and its analogue with sums replaced by integrals when densities exist (which they do not in general!). Both these elementary cases are handled above in our treatment of the bivariate normal distribution (IV.2, Day 5). The general approach to conditioning is due to Kolmogorov in 1933, and uses Measure Theory and σ-fields; see e.g. [SP]. We pause to make the link between 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

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

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

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. $\mathbf{x}^T \Sigma \mathbf{x} > \mathbf{0}$ for all $\mathbf{x} \neq 0$ 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 $\mathbf{x}_1 \neq 0$. This says that Σ_{11} is positive definite. //

Theorem (Normal Conditioning). 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 $\mathbf{A\Sigma B}^T = \mathbf{0}$, or as Σ is symmetric, $\mathbf{B\Sigma A}^T = \mathbf{0}$. 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

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

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

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 $\mathbf{C}^T = -\Sigma_{11}^{-1}\Sigma_{12}$:

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

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

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

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

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 $\Sigma_{2|1} := \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$ is called the *partial covariance matrix* of $\mathbf{X}_2$ given $\mathbf{X}_1$.

Elliptical models

The multinormal, or Gaussian, model is wonderfully convenient mathematically. In particular, the property of having linear regression is highly convenient. However, we note two properties of normal or Gaussian distributions, in any dimension:

(i) they are *symmetrical*, and so cannot model *skewness*;

(ii) they have *extremely thin tails* (so deviations of, say, 3 standard deviations from the mean are very rare).

But these contradict common observation in finance!

Skew.

Profit and loss are profoundly asymmetrical! Large unexpected profits are nice; large unexpected losses are lethal. Consequently, a given amount of profit gives less pleasure than a given amount of loss gives pain. One can see the same effect in prices falling below a peak once the market has turned *far faster* than they increase when the market is rising (so one can detect the *arrow of time* from time series of price data).

Tails.

Inspection (EDA) of any financial data set will reveal *much fatter* tails than Gaussian. Typically, one sees *heavy tails* – tails that decay like a power (as with the Student t-distributions).

There is a third problem, that arises in portfolio management, where we have a range of assets (balanced, by Markowitzian diversification). The tails of two different components of a multinormal vector are (asymptotically) independent. By contrast, the negative tails (downside risk) of assets are usually highly dependent: in a falling market, everything falls, and the tails are heavily dependent.

For all these reasons, it is important to seek other models, which retain as many as possible of the desirable properties of the normal but not the disadvantages above. Such models exist – the *elliptical*, or *elliptically contoured*, models. These may be characterised in several ways. An elliptically contoured distribution in n dimensions with mean vector μ and covariance matrix Σ of rank k (with Cholesky decomposition $\Sigma = A^T A$) has a *stochastic representation*

$$X = \mu + RA^T u \qquad (R: \text{ risk-driver});$$

here u is a random vector uniformly distributed over the unit sphere in k

dimensions and $R \geq 0$ is a scalar random variable independent of u. Alternatively, X has CF

$$\psi(t) = e^{it^T\mu}\phi(t^T\Sigma t)$$

for some scalar function ϕ. Thus $\phi(x) = e^{-\frac{1}{2}x}$ gives the Gaussian case, and choosing ϕ to decrease more slowly gives heavier tails, as required. For background, we refer to e.g. the book [MFE] and the paper [BFK].

Copulas.

Given a random n-vector $X = (X_1, \ldots, X_n)$, write $F(x) = F(x_1, \ldots, x_n) := P(X_{\leq} x_1, \ldots, X_n \leq x_n)$ for the *joint* distribution function, $F_i(x_i) := P(X_i \leq x_i)$ for the *marginal* distribution functions. Then by *Sklar's theorem* (Abe SKLAR (1915-) in 1958),

$$F(x) = C(F_1(x_1), \ldots, F_n(x_n))$$

for some distribution function $C(u) = C(u_1, \ldots, u_n)$ on the unit n-cube. This C is called the *copula*, as it *couples* the marginals together to give the joint distribution. The copula contains all the information on the *dependence* structure (vital for financial applications, as above!). For more on this, see e.g. [MFE] Ch. 5.

7. Generalised linear models (GLMs).

In Regression above, we took as our basic model

$$y = A\beta + \epsilon : \qquad Ey = A\beta; \qquad Ey_i = \sum_j a_{ij}\beta_j$$

– our data y (an n-vector) is modelled as a linear transformation (by a known matrix A, the *design matrix*, $n \times p$) of a p-vector β of parameters, plus an *error*. That is, we work with *linear combinations of predictors plus error*; in particular, the mean μ is given by a *linear predictor*, η. This simple procedure is surprisingly general and effective, but there are situations where it does not apply. We turn to these, seeking to use as much as possible of the approach above.

First, we generalise this by allowing the linear predictor η to be some (smooth and monotone, so invertible) function g of the mean μ:

$$\eta = g(\mu),$$

where g is called the *link function*, or *link*. Next, we need to specify the *error* structure. This is done by means of *exponential families* (I.6.4, D2): the y_i

are independent, with densities

$$f(y_i) = \exp\{\frac{\omega_i(y_i\theta_i - b(\theta_i))}{\phi} + c(y,\phi)\};$$

here b, c are known functions, ω_i are known weights, ϕ is a scale parameter (known or unknown), and the parameter θ_i depends on η.

The case where this dependence is given by the identity,

$$\theta = \eta,$$

is particularly important; here the link is called *canonical*.

GLMs were introduced by Nelder and Wedderburn in 1972; our treatment here follows [BF] Ch. 8. The standard work is
[McN] P. McCULLAGH and J. A. NELDER, *Generalised linear models*, 2nd ed., 1989, Chapman and Hall (1st ed. 1983).
They have been extended to *hierarchical GLMs* (see Ch. VII):
[NLP] J. A. NELDER, Y. LEE and Y. PAWITAN, *Generalised linear models with random effects: unified analysis via H-likelihood*. Chapman and Hall, 2006.
Examples.
1. *Normal*. Here $g(\mu) = \mu$, the errors are normal, and the GLM reduces to the ordinary Linear Model above – as was to be expected!
2. *Poisson*. For the Poisson distribution $P(\lambda)$, writing y for the usual $k = 0, 1, 2, \ldots$,

$$f(y,k) = e^{-\lambda}\lambda^y/y! = \exp\{y\log\lambda - \lambda - \log y!\}.$$

So $\theta = \eta = \log\lambda$: the canonical link is the *logarithm*:

$$\eta = \log\lambda.$$

The Poisson distribution is the default option for *count data*. The logarithm here explains the use of logs in *log-linear models* for count data – contingency tables, etc. (Pearson's chi-square goodness-of-fit test, 1900). For details, see e.g. [BF] 8.3 – 8.5.
3. *Gamma*. The Gamma density $\Gamma(\lambda,\alpha)$ $(\lambda,\alpha > 0)$ has density $f(x) = \lambda^\alpha e^{-\lambda x}x^{\alpha-1}/\Gamma(\lambda)$ on $(0,\infty)$. The mean is $\mu = \alpha/\lambda$, and the canonical link is $\eta = 1/\mu$. The Gamma is the default option for error structure on $(0,\infty)$; here it is often used with the log-link $\eta = \log\mu$. See e.g. [BF] 8.2.3 for an application (to athletics times).