smfd6.tex
Day 6. 25.5.2012

6. FURTHER BAYESIAN ASPECTS.

1. *Posterior means* [O'H] 1.25, p.15]. If t is an estimate of θ given data x, the mean squared error is

$$E[(t-\theta)^2|x] = E[t^2|x] - 2E[t\theta|x] + E[\theta^2|x] = t^2 - 2tE[\theta|x] + E[\theta^2|x]$$

(t is a statistic, that is, a function of the data x, so is known when x is known, and can be taken out of the expectation signs). Add and subtract $(E[\theta|x])^2$:

$$E[(t-\theta)^2|x] = (t - E[\theta|x])^2 + var(\theta|x).$$

Thus the value of t which minimises the posterior expected squared error is

$$t = E[\theta|x],$$

the *posterior mean*. This now has two roles:
(i) minimising mean square error,
(ii) location summary of the posterior distribution.

2. *Multimodal distributions* [O'H] 2.8, p. 31]. One should graph the posterior distribution, to check on shape characteristics, such as the number of modes and skewness properties.

A bimodal density may indicate a non-homogeneous population, which could - or should - be broken down.
E.g.: Adult height is bimodal. For, males are several inches taller than females on average. In such cases, it is usually better to decompose into more homogeneous sub-populations and analyse these separately.

If $f(\theta)$ has k modes, separated by $k-1$ antimodes t_i, let $C_1 = (-\infty, t_1)$, $C_2 = [t_1, t_2), \cdots C_{k-1} = [t_{k-2}, t_{k-1})$, $C_k = [t_{k-1}, \infty)$. Then f induces a density f_i on each C_i by

$$f_i(\theta) := f(\theta)/c_i \quad \text{if} \quad \theta \in C_i, \quad 0 \quad \text{else}, \qquad c_i := \int_{C_i} f(\theta)d\theta$$

(so $\Sigma_1^k c_i = 1$). This represents the k-modal f as a mixture of the k unimodal f_i:

$$f(\theta) = \Sigma_1^k c_i f_i(\theta).$$

Now let ϕ be a random variable taking values $1, \cdots, k$ with probabilities $P(\phi = i) = c_i$: $\theta|\phi = i$ has density f_i. The Conditional Variance Formula,

$$var(\theta) = E[var(\theta|\phi)] + var[E(\theta|\phi)]$$

decomposes the variance of θ into the within-mode variance (the first term on the RHS) and between-mode variance (2nd term on RHS).

For multimodal densities, overall summary statistics provide limited insight, and it is usually better to decompose by modes, to reduce to the unimodal case.

3. *Density plots for bivariate densities* [O'H] 2.9, p. 31]. A full density plot for two-dimensional variables needs three dimensions, so computer graphics. A partial density plot can be provided in two dimensions by drawing *contours*. One can learn to read a contour plot as one can learn to read a map, of which one already has some experience. Here, multimodality (above) shows up as the presence of several different peaks or summits (local maxima). These may be highly revealing. For instance, in B. W. SILVERMAN's book *Density Estimation* (Chapman & Hall), a contour plot arising in a case study in medical statistics is shown. Here the presence of two peaks correctly suggested that two different forms of the disease existed, for which two different clinical treatments were adopted.

The Bivariate Normal distribution (V.2) is a classic case in which the contour plot is unimodal: there is only one peak, and the contours are elliptical. This can be generalised, to weaken the strong assumption of bivariate normality: the class of *elliptically contoured distributions* has many of the desirable properties of the bivariate normal, but is much more general, so more flexible. It can be used, for instance, to model densities with thicker tails than the normal.

4. *Repeated use of Bayes' Theorem* [O'H] 3.5, p. 66]. Suppose now our data x is partitioned into (x_1, x_2), where we observe x_1 first and x_2 second. With prior $f(\theta)$, we have two stages:
Stage 1. Posterior

$$f(\theta|x_1) = f(\theta)f(x_1|\theta)/f(x_1), \qquad f(x_1) = \int f(\theta)f(x_1|\theta)d\theta. \qquad (i)$$

Stage 2. The *prior* density for stage 2 is the *posterior* density above after stage 1. The *likelihood* for stage 2 is $f(x_2|\theta, x_1)$. So the posterior density

after stage 2 is

$$f(\theta|x_1,x_2)=f(\theta|x_1)f(x_2|\theta,x_1)/f(x_2|x_1),\qquad f(x_2|x_1):=\int f(\theta|x_1)f(x_2|\theta,x_1)d\theta. \tag{ii}$$

Substitute $f(\theta|x_1)$ from (i) into (ii):

$$f(\theta|x_1,x_2)=\frac{f(\theta)f(x_1|\theta)f(x_2|\theta,x_1)}{f(x_1)f(x_2|x_1)}.$$

Now $f(x_2|x_1):=f(x_1,x_2)/f(x_1)$, so the denominator is $f(x_1,x_2)$. Similarly, the numerator is

$$f(\theta).\frac{f(\theta,x_1)}{f(\theta)}.\frac{f(\theta,x_1,x_2)}{f(\theta,x_1)}=f(\theta,x_1,x_2)=f(\theta)f(x_1,x_2|\theta).$$

So

$$f(\theta|x_1,x_2)=f(\theta).f(x_1,x_2|\theta)/f(x_1,x_2),$$

the usual result of Bayes' Theorem for updating by the whole data $x=(x_1,x_2)$ in one step. So:

Proposition. If data $x=(x_1,x_2)$ arrives in two stages, with x_1 first and x_2 second, two applications of Bayes' Theorem, updating by x_1 first, then by x_2 given x_1, is equivalent to one application of Bayes' Theorem updating by $x=(x_1,x_2)$.

Corollary. If data $x=(x_1,\cdots,x_n)$ arrives successively in n stages, n applications of Bayes' Theorem - updating by x_i given $x_1,\cdots,x_{i-1}$ $(i=1,\cdots,n)$ are equivalent to one application of Bayes' Theorem, updating by $x=(x_1,\cdots,x_n)$.

The systematic repeated use of Bayes' theorem is important in the subjects of Time Series (Ch. VI) and Forecasting. In particular, the repeated *recursive* use of Bayes' theorem occurs in the *Kalman filter*, which is widely use - for instance, in engineering applications [on-line, or real-time, control of spacecraft, etc.] and in econometric time-series.

5. *Sufficiency* [O'H] 3.9, 69]. Suppose now that $x=(x_1,x_2)$, where x_1 is informative about θ, x_2 is uninformative. This is the idea of *sufficiency*, already encountered in classical statistics. We give a Bayesian treatment. To

say that x_2 is uninformative means that x_2 cannot affect our views on θ, that is,
(i) $f(\theta|x) = f(\theta|x_1, x_2)$ does not depend on x_2, i.e.

$$f(\theta|x_1, x_2) = f(\theta|x_1), \qquad \text{or} \qquad \frac{f(\theta, x_1, x_2)}{f(x_1, x_2)} = \frac{f(\theta, x_1)}{f(x_1)} :$$

$$\frac{f(\theta, x_1, x_2)}{f(\theta, x_1)} = \frac{f(x_1, x_2)}{f(x_1)}, \qquad \text{i.e.} \qquad f(x_2|x_1, \theta) = f(x_2|x_1) :$$

(ii) $f(x_2|x_1, \theta)$ does not depend on θ.
Either of (i), (ii), which are equivalent, can be used as the definition of sufficiency in a Bayesian treatment. Notice that (i) is essentially a Bayesian statement: it is meaningless in classical statistics, as there θ cannot have a density.

Now recall the classical Fisher-Neyman Factorisation Criterion for sufficiency: the likelihood $f(x|\theta)$ factorises as
(iii) $f(x|\theta)$, or $f(x_1, x_2|\theta)$, $= g(x_1, \theta)h(x_1, x_2)$,
for some functions g, h. As before:

Proposition. x_1 is sufficient for θ iff the Factorisation Criterion (iii) holds.

Proof. (ii) $\Rightarrow$ (iii):

$$\begin{aligned} f(x|\theta) = f(x_1, x_2|\theta) &= f(x_1|\theta) f(x_2|x_1, \theta) \qquad \text{(as in 4 above)} \\ &= f(x_1|\theta) f(x_2|x_1) \qquad \text{(by (ii))}, \end{aligned}$$

giving (iii).
(iii) $\Rightarrow$ (i): By Bayes' Theorem in the form 'posterior proportional to prior times likelihood', the factor $h(x_1, x_2)$ in (iii) can be absorbed into the constant of proportionality [which is unimportant: it can be recovered from the remaining terms, its role being merely to make these integrate to one]. Then x_2 vanishes from the analysis, so does not appear in the posterior, giving (i).
•
Note. This proof is easier than the classical one! To a Bayesian, it is also more intuitive and revealing.

6. *Exponential families.* A likelihood $f(x|\theta)$ belongs to the *exponential family* if it is of the form

$$f(x|\theta) = \exp\{a(\theta)u(x) + b(\theta) + k(x)\}$$

(as usual, we use vector notation: x, θ may be several-dimensional; see below). Exponential families (introduced in 1935-36 by Darmois, Pitman and Koopman) arise naturally in classical statistics. We quote: if a statistic $u(x)$ is minimum-variance ('efficient') and unbiased for θ, then the likelihood can be written in the above form (this follows from the conditions for equality in the Cramér-Rao inequality giving the minimum-variance bound, or 'information bound'). By the Fisher-Neyman Factorisation Criterion, $u(x)$ is sufficient for θ. So *efficiency implies sufficiency and membership of an exponential family.*

Now efficiency is not a Bayesian concept (it looks at the distribution of the statistic, so at values we could have seen but didn't, not just at the likelihood), nor is unbiasedness (for the same reason). However, sufficiency is important in Bayesian statistics also (above), and so too are exponential families.

First, we generalise the exponential family approach to cover several parameters and several sufficient statistics: call $f(x|\theta)$ a member of the *k-parameter exponential family* if

$$f(x|\theta) = \exp\{\Sigma_1^k A_j(\theta)B_j(x) + C(x) + D(\theta)\}.$$

Then by the Fisher-Neyman Factorisation Criterion, $B_1(x), \cdots, B_k(x)$ are sufficient statistics for the k parameters $A_1(\theta), \cdots, A_k(\theta)$. Suppose the prior is of the form

$$f(\theta) = f(\theta; a_1, \cdots, a_k, d) = \exp\{\Sigma_1^k a_j A_j(\theta) + dD(\theta) + c(a_1, \cdots, a_k, d)\}.$$

Then the posterior $f(\theta|x) \propto f(\theta)f(x|\theta)$, i.e. to

$$\exp\{\Sigma_1^k A_j(\theta)(a_j + B_j(x)) + (d+1)D(\theta)\},$$

i.e. to

$$f(\theta; a_1 + B_1(x), \cdots, a_k + B_k(x); d+1).$$

This is a $(k+1)$-dimensional exponential family. Its importance is that if the prior belongs to this family, so too does the posterior: the family is *closed under sampling.* This property is of crucial importance, partly because it is so mathematically convenient, partly because it shows up the structure of the problem. For instance, we shall return below to two of the examples we met in S2, where the relationship between prior and likelihood can now be seen in this light to be natural. The prior above is called the *natural conjugate*

family to the exponential family above.
Example 1. *Bernoulli distribution.*

$$\begin{aligned} f(x|\theta) &= \theta^x(1-\theta)^{1-x} \qquad (x=0,1) \\ &= \Big(\frac{\theta}{1-\theta}\Big)^x(1-\theta) \\ &= \exp\{x\log\Big(\frac{\theta}{1-\theta}\Big)+\log(1-\theta)\}: \end{aligned}$$

here $k=1, A_1(\theta)=\log\Big(\frac{\theta}{1-\theta}\Big), B_1(x)=x, C(x)=0, D(\theta)=\log(1-\theta)$.
The natural conjugate family is

$$\begin{aligned} f(\theta;a_1,d) &= \exp\{a_1A_1(\theta)+dD(\theta)+c(a_1,d)\} \\ &= \exp\{a_1\log\Big(\frac{\theta}{1-\theta}\Big)+d\log(1-\theta)+c(a_1,d)\} \\ &= \theta^{a_1}(1-\theta)^{d-a_1}\exp\{c(a_1,d)\}, \end{aligned}$$

which is *Beta* $B(a_1, d-a_1)$ as in S2.
2. *Normal distribution*, $N(\mu,\sigma^2)$: $\theta=(\mu,\sigma^2)$,

$$f(x|\theta)=\exp\{-\frac{1}{2}\frac{x^2}{\sigma^2}+\frac{x\mu}{\sigma^2}-\frac{1}{2}\frac{\mu^2}{\sigma^2}-\log\sigma-\frac{1}{2}\log 2\pi\},$$

$k=2, A_1(\theta)=1/\sigma^2, B_1(x)=-\frac{1}{2}x^2, A_2(\theta)=\mu/\sigma^2, B_2(x)=x, C(x)=0, D(\theta)=-\frac{1}{2}[\log(2\pi\sigma^2)+\mu^2/\sigma^2]$. The natural conjugate family is

$$f(\theta;a_1,a_2,d)=\exp\{a_1A_1(\theta)+a_2A_2(\theta)+dD(\theta)+c(a_1,a_2,d)\}$$

$$\propto(\sigma^2)^{-\frac{1}{2}d}\exp\{\frac{a_1}{\sigma^2}+\frac{a_2\mu}{\sigma^2}-\frac{1}{2}d\mu^2\sigma^2\}.$$

The exponent is σ^2 times

$$-\frac{1}{2}d(\mu^2-\frac{2a_2\mu}{d}+a_1)=-\frac{1}{2}d[(\mu-\frac{a_2}{d})^2-a_1-\frac{{a_2}^2}{d^2}].$$

Writing $m:=a_2/d, b:=-a_1-{a_2}^2/2d$,

$$f(\theta;a_1,a_2,d)\propto(\sigma^2)^{-\frac{1}{2}d}\exp\{-\frac{1}{2}d(\mu-m)^2/\sigma^2-b/\sigma^2\}.$$

For σ known, this is a normal prior for μ, as in S2. With neither σ nor μ known (both parameters), this is the natural conjugate prior to the normal

$N(\mu, \sigma^2)$. More generally, one can work with $(\sigma^2)^{-t}$ in place of $(\sigma^2)^{-\frac{1}{2}d}$. Here m, d, b (and t if present) are *hyperparameters* for the parameters μ, σ.

7. *Asymptotic normality* [O'H] 3.18, p. 74]. We recall (or quote) that in classical statistics, the maximum-likelihood estimator $\hat{\theta}$ of θ based on n i.i.d. readings $x_1, \cdots, x_n$ is asymptotically normal, with mean θ and variance $1/(nI(\theta))$, where $I(\theta)$ is the Fisher information per reading:

$$I(\theta) := E[(\ell'(\theta))^2] = -E[\ell''(\theta)], \qquad \ell(\theta) := \log f(x|\theta)$$

the log-likelihood (the likelihood itself is usually written $L(\theta)$ in classical statistics). This result needs some regularity conditions, the principal ones being
(i) enough smoothness to justify differentiating under the integral sign twice with respect to θ (as in the derivation of the above equation for the information, and in the proof of the Cramér-Rao inequality),
(ii) that the support of the likelihood (the region where it is positive) should not depend on θ.
Now the above is a large-sample result, in which the sample size n increases. It is thus natural to expect that in this situation, the data information will swamp the prior information, and the same result will hold in the Bayesian case also. This is indeed so; see O'Hagan SS3.18-26 for details.

8. *Shrinkage* [O'H] 6.42, p. 159]. We have seen that in the Bayesian paradigm the posterior gives a compromise between the prior and the likelihood. The effect is to 'pull' the likelihood towards the prior. Thus a Bayesian estimate typically 'pulls' a classical estimate towards a prior estimate. With several parameters - with the same prior mean, say - their classical estimates will all be pulled towards the same prior estimate. It is thus typical of the Bayesian paradigm that estimators are less spread out than in the classical paradigm, a phenomenon known as *shrinkage*.

Similar shrinkage effects occur in higher dimensions (Ch. VII) – the *James-Stein phenomenon*.

9. *Bayes Linear Estimates* [O'H] 6.48 p. 163]. Recall 1 - *Posterior means*. Take expectations of 1 over x, and use the Conditional Mean Formula ($E[E(.|x)] = E$):

$$D := E[(d(x) - \theta)^2] = E[(d(x) - E[\theta|x])^2] + Evar(\theta|x), \qquad (*)$$

which is minimised by the posterior mean $d(x) = E(\theta|x)$. Suppose now that $d(x)$ is a *linear* function, $a + b'z$, where $z = z(x)$ and b are vectors:

$$\begin{aligned} D &= E[(a + b'z - \theta)^2] \\ &= E[a^2 + 2ab'z + b'zz'b - 2a\theta - 2b'z\theta + \theta^2] \\ &= a^2 + 2ab'Ez + b'E(zz')b - 2aE\theta - 2b'E(z\theta) + E(\theta^2). \end{aligned}$$

Add and subtract $[E(\theta)]^2$, $(b'Ez)^2 = b'EzEz'b$ and $2b'EzE\theta$, and observe that $E(zz') - EzEz'$ is the covariance matrix $V := varz$ of z and $E(z\theta) - EzE\theta$ is the vector c of covariances between θ and the elements of the vector z. We obtain

$$D = (a + b'Ez - E\theta)^2 + b'(varz)b - 2b'cov(z, \theta) + var\theta. \tag{1}$$

Write $b^* := V^{-1}c$. We show that

$$D = (a + b'Ez - E\theta)^2 + (b - b^*)'V(b - b^*) + D^*, \tag{2}$$

where $D^* := V - c'V^{-1}c$. For, the second term on the right in (2) is

$$[b - V^{-1}c]'V[b - V^{-1}c] = b'Vb - 2b'VV^{-1}c + c'(V^{-1})'VV^{-1}c = b'Vb - 2b'c + c'V^{-1}c;$$

this and the definition of D^* give (1).

The third term on the right in (2) does not involve a, b, while the first two are non-negative (the first is a square, the second a quadratic form with matrix V, non-negative definite as V is a covariance matrix). So the expected quadratic loss D is minimised by choosing $b = b^*$, $a = -b^{*\prime}Ez + E\theta$. This choice gives

$$d(x) = E\theta + cV^{-1}(z - Ez), \qquad c := cov(z, \theta), \quad V := var(z).$$

This gives the *Bayes linear estimator* of θ based on data $z = z(x)$. From (*), minimising D means minimising the mean square error in $d(x)$ as an approximation to the posterior mean $E(\theta|x)$. Thus the Bayes linear estimator is the best approximation to the posterior mean (in the sense of mean-square error) among the class of linear estimators (in $z = z(x)$).

Note that the Bayes linear estimator depends only on $E\theta$, Ez, $c = cov(z, \theta)$, $V = var(z)$, that is, only on first and second moments. So to construct it, we do not need to know the full likelihood, only the first and second moments of $(\theta, z(x))$, the parameter and the function z in which we

want the estimator to be linear.

Note that the Bayes linear estimator violates the Likelihood Principle: it depends on the distribution of $z = z(x)$, not just on the observed likelihood.

10. *Odds* [O'H] 6.23 p. 150]. While Bayesian statistics avoids tests of hypotheses as such, one may well wish to compare two possible values of θ, say θ_0 and θ_1, against each other. Write down the proportionality form of Bayes' Theorem for each value of θ and divide: the constants of proportionality cancel, giving

$$\frac{P(\theta = \theta_1|x)}{P(\theta = \theta_0|x)} = \frac{P(\theta = \theta_1)}{P(\theta = \theta_0)} . \frac{f(x|\theta = \theta_1)}{f(x|\theta = \theta_0)}.$$

The first term on the right is the (prior) *odds* (ratio) in favour of θ_1 (or against θ_0) [we use the term odds here in the same sense as its everyday use in gambling]. The second term on the right is the *likelihood ratio*. The left is the *posterior odds*. Thus a use of Bayes' Theorem updates the prior odds to the posterior odds by the *Bayes factor* given by the likelihood ratio .

11. *Invariance and Jeffreys priors*. Suppose we work with a parameter θ, with information per reading $I(\theta) = E[(\ell'(\theta)^2] = \int ((\log f)_\theta)^2 f(\theta)$. If we reparametrise to $\phi := g(\theta)$, then as $\partial/\partial\phi = (d\theta/d\phi)(\partial/\partial\phi)$,

$$I(\phi) = (d\theta/d\phi)^2 I(\theta).$$

The idea of choosing a prior which is large where the information is large is very attractive (and reminiscent of maximum-likelihood estimation!). Jeffreys suggested choosing a prior of the form

$$\pi(\theta) \propto \sqrt{I(\theta)}.$$

The square root is explained by requiring the prior to be *invariant under reparametrisation*, as by above,

$$\pi(\phi)d\phi \propto \sqrt{I(\phi)}d\phi = \sqrt{I(\theta)}d\theta \propto \pi(\theta)d\theta : \qquad \pi(\phi)d\phi = \pi(\theta)d\theta$$

(both sides integrate to 1, so we can take equality here). There is an extension to higher dimensions, using the Fisher information matrix and the square root of the modulus of its determinant.

Bayesian procedures are in general not invariant under reparametrisation! This can be seen as a drawback, but Bayesians argue that one needs

to incorporate a loss function (or utility function), and one should seek a parametrisation that suits this loss function.
Note. Sir Harold JEFFREYS (1891-1989) was primarily a geophysicist, and wrote an influential book *The Earth: Its Origin, History and Physical Constitution*, 1924[1]. He was also a pioneer of Bayesian statistics, and wrote an early book on it, *Theory of probability* (1st ed. 1939, 2nd ed. 1960, 3rd ed. 1983). He also wrote (with his wife) 'Jeffreys and Jeffreys', *Methods of mathematical physics*, CUP, 1946.

Postscript: Bayesian solution of the equity premium puzzle.

Following Markowitz (I.5), we should diversify our financial savings into a range of assets in our portfolio – including cash (invested risklessly – e.g., by buying Government bonds, or 'gilts', or putting it in the bank or building society – which we suppose riskless here, discounting such disasters as the Icelandic banking crisis, Northern Rock, RBS etc.) and risky stock. There is no point in taking risk unless we are paid for it, so there will be an excess return – equity premium – $\mu - r$ of the risky stock (return μ) over the riskless cash (return r), to be compared with the volatility σ of the risky stock via the *Sharpe ratio* (or *market price of risk*) $\lambda := (\mu - r)/\sigma$. Historical data show that the observed excess return seems difficult to explain.

A Bayesian solution to this 'equity premium puzzle' has been put forward by Jobert, Platania and Rogers. They conclude that there is no equity premium puzzle, if one uses a Bayesian analysis to reflect fully one's uncertainty in modelling this situation. See
[JPR] A. JOBERT, A. PLATANIA & L. C. G. ROGERS, *A Bayesian solution to the equity premium puzzle*. Preprint, Cambridge (available from Chris Rogers' homepage: Cambridge University, Statistical Laboratory).
The Twenties Example [JPR]. One observes daily prices of a stock for T years, with an annual return rate of 20% and an annual volatility of 20%. How large must T be to give confidence intervals of $\pm 1\%$ for (i) the volatility, (ii) the mean? Answers: (i) about 11; (ii) about 1,550!!

This illustrates what is called *mean blur*; see e.g.
D. G. LUENBERGER, *Investment Science*, OUP, 1997.
Rough explanation: for the mean, only the starting and final values matter (so effective sample size is 2); for the volatility, everything matters.

[1] Jeffreys was the first to suggest that the earth's core is liquid – but he was a strong opponent of continental drift!

V. REGRESSION

1. LEAST SQUARES

The idea of regression is to take some sample of size n from some unknown population (typically n is large – the larger the better), and seek how best to represent it in terms of a smaller number of variables, typically involving p parameters (p to be kept as small as possible, to give a parsimonious representation of the data – so p is much smaller than n, $p << n$). Usually we will have p explanatory variables, and represent the data as a linear combination of them (the coefficients being the parameters) plus some random error, as best we can. To do this, we use the *method of least squares*, and choose the coefficients so as to minimise the sum of squares (SS) of the differences between the observed data points and the linear combination. This gives us a fitted value; what is left over is called a residual; thus

$$\textit{data} = \textit{true value} + \textit{error} = \textit{fitted value} + \textit{residual.}$$

If the data forms an n-vector y and the parameters form a p-vector β, the model equation is

$$y = A\beta + \epsilon,$$

where A is an $n \times p$ matrix of constants (the *design matrix*), and ϵ is an n-vector of errors. In the full-rank case (where A has rank p), it can be shown ([BF], 3.1) that the *least-squares estimates* (LSEs) of β are

$$\hat{\beta} = (A^T A)^{-1} A^T y,$$

and (Gauss-Markov Theorem) that this gives the minimum-variance unbiased (= 'best') linear estimator (or BLUE): in this sense *least-squares is best.*

Geometrically, the Method of Least Squares projects n-dimensional reality onto the best approximating p-dimensional subspace. Indeed, the key role is played by the *projection matrix* $P = A(A^T A)^{-1} A^T$ (or $P = AC^{-1}A^T$ with $C := A^T A$ the *information matrix*; P is $n \times n$, C is $p \times p$). P is also called the *hat matrix*, H, as it projects the data y onto the fitted values $\hat{y} = A\hat{\beta}$.

To make good statistical sense of this, we need a statistical model for the error structure. We will use the *multivariate normal* distribution (Section 3), whose estimation theory follows in Section 4.

The most basic case is $p = 2$, where one fits a line (two parameters, slope

and intercept) through n data points in the plane. One can show (see e.g. [BF], 1.2) that the least-squares (best) line is

$$y = a + bx, \quad b = \frac{\overline{xy} - \overline{x}.\overline{y}}{\overline{x^2} - \overline{x}^2} = s_{xy}/s_{xx} = r_{xy}s_y/s_x, \quad a = \overline{y} - b\overline{x}.$$

(here s_{xy} is the sample covariance between x and y, $s_{xx} = s_x^2$ is the sample variance of x, $r_{xy} = s_{xy}/(s_x s_y)$ the sample correlation coefficient). This is the *sample regression line*. By LLN, its large-sample limit is the *(population) regression line*,

$$y = \alpha + \beta x, \quad \beta = \rho\sigma_2/\sigma_1, \quad \alpha = Ey - \beta Ex: \quad y - Ey = (\rho\sigma_2/\sigma_1)(x - Ex).$$

The multivariate normal reduces in this case to the *bivariate normal* in Section 2; we treat this in full because of its fundamental importance and of how well it illustrates the general case.

Motivating examples:
1. *CAPM* (I.5). The capital asset pricing model looks at individual risky assets and compares them with 'the market', or some proxy for it such as an index. One seeks to 'pick winners' by maximising 'beta', or the slope of the linear trend of asset price versus market price.
2. *Examination scores* (BF, 1.4). Here x is the 'incoming score' of an entrant to an elite academic programme, y is the 'graduating score'; the question is how well does the institution pick its intake (i.e., how well does x predict y).
3. *Galton's height data* (BF, 1.3). Here y = offspring's height (adult sons, say), x = average of parents' heights.