Maths for Brain Imaging: Lecture 1

W.D. Penny
Wellcome Department of Imaging Neuroscience,
University College, London WC1N 3BG.

October 1, 2006

1 Probability Density Functions

The probability of a continuous variable, x, assuming a particular value or range of values is defined by a Probability Density Funcion (PDF), $p(x)$. *Probability is measured by the area under the PDF*; the total area under a PDF is therefore unity

$$\int p(x)dx = 1 \tag{1}$$

The probability of x assuming a value between a and b is given by

$$p(a \leq x \leq b) = \int_a^b p(x)dx \tag{2}$$

which is the area under the PDF between a and b. *The probability of x taking on a single value is therefore zero.* This makes sense because we are dealing with continuous values; as your value becomes more precise the probability for it decreases. It only makes sense, therefore to talk about the probability of a value being within a certain precision or being above or below a certain value.

To calculate such probabilities we need to calculate integrals like the one above. This process is simplified by the use of Cumulative Density Functions (CDF) which are defined as

$$CDF(a) = p(x \leq a) = \int_{-\infty}^a p(x)dx \tag{3}$$

Hence

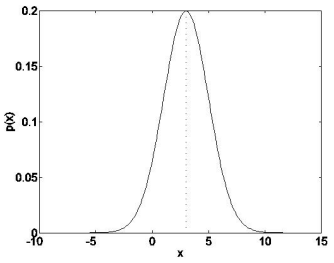
$$p(a \leq x \leq b) = CDF(b) - CDF(a) \tag{4}$$

1.1 The Gaussian Density

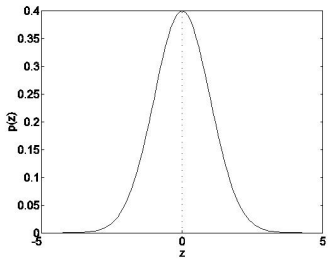
The *Normal* or *Gaussian* probability density function, for the case of a single variable, is

$$p(x) \equiv N(x; \mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{5}$$

where μ and σ^2 are known as the *mean* and *variance*, and σ (the square root of the variance) is called the *standard deviation*. The quantity in front of the



(a)



(b)

Figure 1: *(a) The Gaussian Probability Density Function with mean $\mu = 3$ and standard deviation $\sigma = 2$, (b) The standard Gaussian density, $p(z)$. This has zero mean and unit variance.*

exponential ensures that $\int p(x)dx = 1$. The above formula is often abbreviated to the shorthand $p(x) = N(x; \mu, \sigma)$. The terms Normal and Gaussian are used interchangeably.

If we subtract the mean from a Gaussian variable and then divide by that variables *standard deviation* the resulting variable, $z = (x - \mu)/\sigma$, will be distributed according the *standard* normal distribution, $p(z) = N(z; 0, 1)$ which can be written

$$p(z) = \frac{1}{(2\pi)^{1/2}} \exp\left(-\frac{z^2}{2}\right) \tag{6}$$

The probability of z being above 0.5 is given by the area to the right of 0.5. We can calculate it as

$$\begin{aligned} p(z) \geq 0.5 &= \int_{0.5}^{\infty} p(z)dz \\ &= 1 - CDF_{Gauss}(0.5) \end{aligned} \tag{7}$$

where CDF_{Gauss} is the cumulative density function for a Gaussian.

1.2 Probability relations

The same probability relations hold for continuous variables as for discrete variables ie. the conditional probability is

$$p(y|x) = \frac{p(x, y)}{p(x)} \tag{8}$$

Re-arranging gives the joint probability

$$p(x, y) = p(y|x)p(x) \tag{9}$$

which, if y does not depend on x (ie. x and y are independent) means that

$$p(x, y) = p(y)p(x) \tag{10}$$

1.3 Expectation and moments

The *expected value* of a function $f(x)$ is defined as

$$E[f(x)] \equiv < f(x) >= \int p(x)f(x)dx \tag{11}$$

and $E[]$ is referred to as the *expectation* operator, which is also sometimes written using the angled brackets $<>$. The kth *moment* of a distribution is given by

$$E[x^k] = \int p(x)x^k dx \tag{12}$$

The mean is therefore the first moment of a distribution.

$$E[x] = \int p(x)xdx = \mu \tag{13}$$

The kth *central moment* of a distribution is given by

$$E[(x - \mu)^k] = \int p(x)(x - \mu)^k dx \tag{14}$$

The variance is therefore the second central moment

$$E[(x - \mu)^2] = \int p(x)(x - \mu)^2 dx = \sigma^2 \tag{15}$$

Sometimes we will use the notation

$$Var(x) = E[(x - \mu)^2] \tag{16}$$

The third central moment is *skewness* and the fourth central moment is *kurtosis* (see later). In the appendix we give examples of various distributions and of skewness and kurtosis.

1.4 Mean and Variance

For more on the mean and variance of functions of random variables see Weisberg [7] and Bevington and Robinson [**?**].

Expectation is a *linear operator*. That is

$$E[(a_1 x + a_2 x)] = a_1 E[x] + a_2 E[x] \tag{17}$$

Therefore, given the function

$$y = ax \tag{18}$$

we can calculate the mean and variance of y as functions of the mean and variance of x.

$$\begin{aligned} E[y] &= aE[x] \\ Var(y) &= a^2 Var(x) \end{aligned} \tag{19}$$

If y is a function of many *uncorrelated* variables

$$y = \sum_i a_i x_i \tag{20}$$

we can use the results

$$E[y] = \sum_i a_i E[x_i] \tag{21}$$

$$Var[y] = \sum_i a_i^2 Var[x_i] \tag{22}$$

But if the variables are correlated then

$$Var[y] = \sum_i a_i^2 Var[x_i] + 2 \sum_i \sum_j a_i a_j Var(x_i, x_j) \tag{23}$$

where $Var(x_i, x_j)$ denotes the covariance of the random variables x_i and x_j.

1.5 Standard Error

As an example, the mean

$$m = \frac{1}{N} \sum_i x_i \tag{24}$$

of uncorrelated variables x_i has a variance

$$\begin{aligned} Var(m) &= \sum_i \frac{1}{N} Var(x_i) \\ &= \frac{\sigma_x^2}{N} \end{aligned} \tag{25}$$

where we have used the substitution $a_i = 1/N$ in equation 22.

2 Maximum Likelihood Estimation

We can learn the mean and variance of a Gaussian distribution using the Maximum Likelihood (ML) framework as follows. A Gaussian variable x_n has the PDF

$$p(x_n) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{26}$$

which is also called the likelihood of the data point. Given N Independent and Identically Distributed (IID) (it is often assumed that the data points, or errors, are independent and come from the same distribution) samples $y = [y_1, y_2, .., y_N]$ we have

$$p(y) = \prod_{n=1}^{N} p(y_n) \tag{27}$$

which is the likelihood of the data set. We now wish to set μ and σ^2 so as to maximise this likelihood. For numerical reasons (taking logs gives us bigger numbers) this is more conveniently achieved by maximising the log-likelihood (note: the maximum is given by the same values of μ and σ)

$$L \equiv \log p(y) = -\frac{N}{2}\log 2\pi - \frac{N}{2}\log \sigma^2 - \sum_{n=}^{N} \frac{(y_n - \mu)^2}{2\sigma^2} \tag{28}$$

The optimal values of μ and σ are found by setting the derivatives $\frac{dL}{d\mu}$ and $\frac{dL}{d\sigma}$ to zero. This gives

$$\mu = \frac{1}{N}\sum_{n=1}^{N} y_n \tag{29}$$

and

$$\sigma^2 = \frac{1}{N}\sum_{n=1}^{N} (y_n - \mu)^2 \tag{30}$$

We note that the last formula is different to the usual formula for estimating variance

$$\sigma^2 = \frac{1}{N-1}\sum_{n=1}^{N} (x_n - \mu)^2 \tag{31}$$

because of the difference in normalisation. The last estimator of variance is preferred as it is an *unbiased* estimator (see later section on bias and variance).

If we had an input-dependent mean, $\mu_n = wx_n$, then the optimal value for w can be found by maximising L. As only the last term in equation 28 depends on w this therefore corresponds to minimisation of the squared errors between μ_n and y_n. This provides the connection between ML estimation and Least Squares (LS) estimation; ML reduces to LS for the case of Gaussian noise.

3 Correlation and Regression

3.1 Correlation

The *covariance* between two variables x and y is measured as

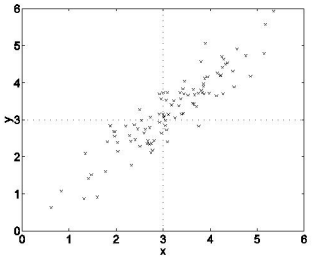
$$\sigma_{xy} = \frac{1}{N-1}\sum_{n=1}^{N} (x_i - \mu_x)(y_i - \mu_y) \tag{32}$$

where μ_x and μ_y are the means of each variable. Note that $\sigma_{yx} = \sigma_{xy}$. Sometimes we will use the notation

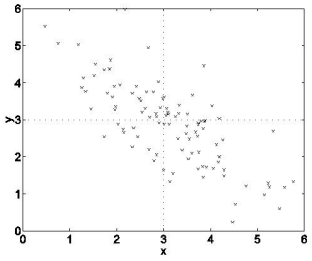
$$Var(x, y) = \sigma_{xy} \tag{33}$$

If x tends to be above its mean when y is above its mean then σ_{xy} will be positive. If they tend to be on opposite sides of their means σ_{xy} will be negative. The *correlation* or *Pearson's correlation coefficient* is a normalised covariance

$$r = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \tag{34}$$



(a)



(b)

Figure 2: *(a) Positive correlation,* $r = 0.9$ *and (b) Negative correlation,* $r = -0.7$*. The dotted horizontal and vertical lines mark* μ_x *and* μ_y*.*

such that $-1 \leq r \leq 1$, a value of -1 indicating perfect negative correlation and a value of $+1$ indicating perfect positive correlation; see Figure 2. A value of 0 indicates no correlation. The strength of a correlation is best measured by r^2 which takes on values between 0 and 1, a value near to 1 indicating strong correlation (regardless of the sign) and a value near to zero indicating a very weak correlation.

3.2 Linear regression

We now look at modelling the relationship between two variables x and y as a linear function; given a collection of N data points $\{x_i, y_i\}$, we aim to estimate y_i from x_i using a linear model

$$\hat{y}_i = ax_i + b \tag{35}$$

where we have written $\hat{y}$ to denote our estimated value. Regression with one input variable is often called *univariate* linear regression to distinguish it from *multivariate* linear regression where we have lots of inputs. The goodness of fit of the model to the data may be measured by the least squares cost function

$$E = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \tag{36}$$

The values of a and b that minimize the above cost function can be calculated by setting the first derivatives of the cost function to zero and solving the resulting

simultaneous equations (derivatives are used to find maxima and minima of functions).

The result is derived as follows. We can find the slope a and offset b by minising the cost function

$$E = \sum_{i=1}^{N} (y_i - ax_i - b)^2 \tag{37}$$

Differentiating with respect to a gives

$$\frac{\partial E}{\partial a} = -2 \sum_{i=1}^{N} x_i (y_i - ax_i - b) \tag{38}$$

Differentiating with respect to b gives

$$\frac{\partial E}{\partial b} = -2 \sum_{i=1}^{N} (y_i - ax_i - b) \tag{39}$$

By setting the above derivatives to zero we obtain the *normal equations* of the regression. Re-arranging the normal equations gives

$$a \sum_{i=1}^{N} x_i^2 + b \sum_{i=1}^{N} x_i = \sum_{i=1}^{N} x_i y_i \tag{40}$$

and

$$a \sum_{i=1}^{N} x_i + bN = \sum_{i=1}^{N} y_i \tag{41}$$

By substituting the mean observed values μ_x and μ_y into the last equation we get

$$b = \mu_y - a\mu_x \tag{42}$$

Now let

$$S_{xx} = \sum_{i=1}^{N} (x_i - \mu_x)^2 \tag{43}$$

$$= \sum_{i=1}^{N} x_i^2 - N\mu_x^2 \tag{44}$$

and

$$S_{xy} = \sum_{i=1}^{N} (x_i - \mu_x)(y_i - \mu_y) \tag{45}$$

$$= \sum_{i=1}^{N} x_i y_i - N\mu_x\mu_y \tag{46}$$

Substiting for b into the first normal equation gives

$$a\sum_{i=1}^{N} x_i^2 + (\mu_y - a\mu_x)\sum_{i=1}^{N} x_i = \sum_{i=1}^{N} x_i y_i \tag{47}$$

Re-arranging gives

$$\begin{aligned} a &= \frac{\sum_{i=1}^{N} x_i y_i - \mu_y \sum_{i=1}^{N} x_i}{\sum_{i=1}^{N} x_i^2 + \mu_x \sum_{i=1}^{N} x_i} \\ &= \frac{\sum_{i=1}^{N} x_i y_i - N\mu_x\mu_y}{\sum_{i=1}^{N} x_i^2 + N\mu_x^2} \\ &= \frac{\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)}{\sum_{i=1}^{N}(x_i - \mu_x)^2} \\ &= \frac{\sigma_{xy}}{\sigma_x^2} \end{aligned} \tag{48}$$

To summarise, the solutions are

$$a = \frac{\sigma_{xy}}{\sigma_x^2} \tag{49}$$

and

$$b = \mu_y - a\mu_x \tag{50}$$

where μ_x and μ_y are the mean observed values of the data and σ_x^2 and σ_{xy} are the input variance and input-output covariance. This enables least squares fitting of a regression line to a data set as shown in Figure 3.

The model will fit some data points better than others; those that it fits well constitute the *signal* and those that it does'nt fit well constitute the *noise*. The strength of the noise is measured by the noise variance

$$\sigma_e^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{51}$$

and the strenth of the signal is given by $\sigma_y^2 - \sigma_e^2$. The *signal-to-noise ratio* is therefore $(\sigma_y^2 - \sigma_e^2)/\sigma_e^2$.

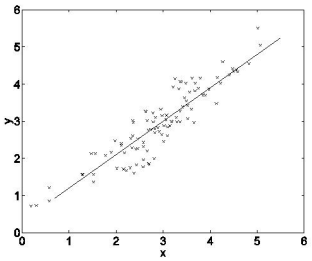
Splitting data up into signal and noise components in this manner (ie. breaking down the variance into what the model *explains* and what it does not) is at the heart of statistical procedures such as analysis of variance (ANOVA) [3].

3.3 Relation to correlation

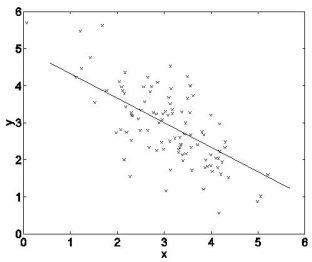
The correlation measure r is intimately related to the linear regression model. Indeed (by substituting σ_{xy} from equation 32 into equation 49) r may be expressed as

$$r = \frac{\sigma_x}{\sigma_y} a \tag{52}$$

where a is the slope of the linear regression model. Thus, for example, the sign of the slope of the regression line defines the sign of the correlation. The correlation is, however, also a function of the standard deviation of the x and



(a)



(b)

Figure 3: *The linear regression line is fitted by minimising the vertical distance between itself and each data point. The estimated lines are (a)* $\hat{y} = 0.9003x + 0.2901$ *and (b)* $\hat{y} = -0.6629x + 4.9804$.

y variables; for example, if σ_x is very large, it is possible to have a strong correlation even though the slope may be very small.

The relation between r and linear regression emphasises the fact that r is only a measure of *linear* correlation. It is quite possible that two variables have a strong nonlinear relationship (ie. are nonlinearly correlated) but that $r = 0$. Measures of nonlinear correlation will be discussed in a later lecture.

The strenth of correlation can also be expressed in terms of quantites from the linear regresssion model

$$r^2 = \frac{\sigma_y^2 - \sigma_e^2}{\sigma_y^2} \tag{53}$$

where σ_e^2 is the noise variance and σ_y^2 is the variance of the variable we are trying to predict. Thus r^2 is seen to measure the proportion of variance explained by a linear model, a value of 1 indicating that a linear model perfectly describes the relationship between x and y.

3.4 Finding the uncertainty in estimating the slope

The data points may be written as

$$\begin{aligned} y_i &= \hat{y}_i + e_i \\ &= ax_i + b + e_i \end{aligned} \tag{54}$$

where the noise, e_i has mean zero and variance σ_e^2. The mean and variance of each data point are

$$E(y_i) = ax_i + b \tag{55}$$

and

$$Var(y_i) = Var(e_i) = \sigma_e^2 \tag{56}$$

We now calculate the variance of the estimate a. From earlier we see that

$$a = \frac{\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)}{\sum_{i=1}^{N}(x_i - \mu_x)^2} \tag{57}$$

Let

$$c_i = \frac{(x_i - \mu_x)}{\sum_{i=1}^{N}(x_i - \mu_x)^2} \tag{58}$$

We also note that $\sum_{i=1}^{N} c_i = 0$ and $\sum_{i=1}^{N} c_i x_i = 1$. Hence,

$$a = \sum_{i=1}^{N} c_i(y_i - \mu_y) \tag{59}$$

$$= \sum_{i=1}^{N} c_i y_i - \mu_y \sum_{i=1}^{N} c_i \tag{60}$$

The mean estimate is therefore

$$E(a) = \sum_{i=1}^{N} c_i E(y_i) - \mu_y \sum_{i=1}^{N} c_i \tag{61}$$

$$
\begin{aligned}
&= a\sum_{i=1}^{N} c_i x_i + b\sum_{i=1}^{N} c_i - \mu_y \sum_{i=1}^{N} c_i \\
&= a
\end{aligned}
\tag{62}
$$

The variance is

$$
Var(a) = Var(\sum_{i=1}^{N} c_i y_i - \mu_y \sum_{i=1}^{N} c_i) \tag{63}
$$

The second term contains two fixed quantities so acts like a constant. Hence,

$$
\begin{aligned}
Var(a) &= Var(\sum_{i=1}^{N} c_i y_i) \\
&= \sum_{i=1}^{N} c_i^2 Var(y_i) \\
&= \sigma_e^2 \sum_{i=1}^{N} c_i^2 \\
&= \frac{\sigma_e^2}{\sum_{i=1}^{N}(x_i - \mu_x)^2} \\
&= \frac{\sigma_e^2}{(N-1)\sigma_x^2}
\end{aligned}
\tag{64}
$$

4 Inference

When we estimate the mean and variance from small samples of data our estimates may not be very accurate. But as the number of samples increases our estimates get more and more accurate and as this number approaches infinity the sample mean approaches the true mean or *population* mean. In what follows we refer to the sample means and variances as m and s and the population means and standard deviations as μ and σ.

Hypothesis Testing: Say we have a hypothesis $\boldsymbol{H}$ which is *The mean value of my signal is 32*. This is often referred to as the *null hypothesis* or H_0. We then get some data and test $\boldsymbol{H}$ which is then either *accepted* or *rejected* with a certain probability or *significance level*, p. Very often we choose $p = 0.05$ (a value used throughout science).

We can do a *one-sided* or a *two-sided* statistical test depending on exactly what the null hypothesis is. In a one-sided test our hypothesis may be (i) our parameter is less than x or (ii) our parameter is greater than x. For two-sided tests our hypothesis is of the form (iii) our parameter is x. This last hypothesis can be rejected if the sample statistic is either much smaller or much greater than it should be if the parameter truly equals x.

4.1 Regression

In a linear regression model we are often interested in whether or not the gradient is significantly different from zero or other value of interest.

To answer the question we first estimate the variance of the slope and then perform a t-test. In the appendix we show that the variance of the slope is given by [1]

$$\sigma_a^2 = \frac{\sigma_e^2}{(N-1)\sigma_x^2} \tag{65}$$

We then calculate the t-statistic

$$t = \frac{a - a_h}{\sigma_a} \tag{66}$$

where a_h is our hypothesized slope value (eg. a_h may be zero) and look up $p(t)$ with $N-2$ DF (we have used up 1DF to estimate the input variance and 1DF to estimate the noise variance). In the data plotted in Figure 3(b) the estimated slope is $a = -0.6629$. From the data we also calculate that $\sigma_a = 0.077$. Hence, to find out if the slope is significantly non-zero we compute $CDF_t(t)$ where $t = -0.6629/0.077 = -8.6$. This has a p-value of 10^{-13} ie. a very significant value. To find out if the slope is significantly different from -0.7 we calculate $CDF_t(t)$ for $t = (-0.6629 + 0.7)/0.077 = 0.4747$ which gives a p-value of 0.3553 ie. not significantly different (again, we must bear in mind that we need to do a two-sided test; see earlier).

4.2 Correlation

Because of the relationship between correlation and linear regression we can find out if correlations are significantly non-zero by using exactly the same method as in the previous section; if the slope is significantly non-zero then the corresponding correlation is also significantly non-zero.

By substituting $a = (\sigma_y/\sigma_x)r$ (this follows from equation 49 and equation 34) and $\sigma_e^2 = (1-r^2)\sigma_y^2$ (from equation 53) into equation 65 and then σ_a into equation 66 we get the test statistic [2]

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}} \tag{67}$$

which has $N-2$ DF.

For example, the two signals in Figure 4(a) have, over the $N = 50$ given samples, a correlation of $r = 0.8031$ which gives $t = 9.3383$ and a p-value of 10^{-12}. We therefore reject the hypothesis that the signals are not correlated; they clearly are. The signals in Figure 4(b) have a correlation of $r = 0.1418$ over the $N = 50$ given samples which gives $t = 0.9921$ and a p-value of $p = 0.1631$. We therefore accept the null hypothesis that the signals are not correlated.

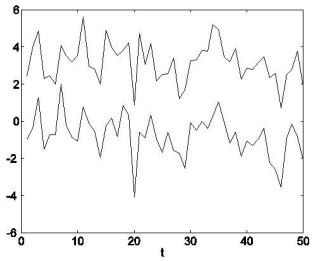
5 Linear algebra

5.1 Transposes and Inner Products

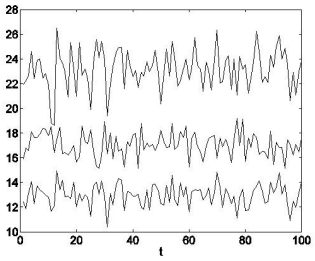
A collection of variables may be treated as a single entity by writing them as a *vector*. For example, the three variables x_1, x_2 and x_3 may be written as the

[1] When estimating σ_x^2 we should divide by $N-1$ and when estimating σ_e^2 we should divide by $N-2$.

[2] Strictly, we should use $\sigma_e^2 = \frac{N-1}{N-2}(1-r^2)\sigma_y^2$ to allow for using $N-2$ in the denominator of σ_e^2.



(a)



(b)

Figure 4: *Two signals (a) sample correlation* $r = 0.8031$ *and (b) sample correlation,* r=0.1418. *Strong correlation; by shifting and scaling one of the time series (ie. taking a linear function) we can make it look like the other time series.*

vector

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \tag{68}$$

Bold face type is often used to denote vectors (scalars - single variables - are written with normal type). Vectors can be written as *column vectors* where the variables go down the page or as *row vectors* where the variables go across the page (it needs to be made clear when using vectors whether $\boldsymbol{x}$ means a row vector or a column vector - most often it will mean a column vector and in our text it will *always* mean a column vector, unless we say otherwise). To turn a column vector into a row vector we use the *transpose* operator

$$\boldsymbol{x}^T = [x_1, x_2, x_3] \tag{69}$$

The transpose operator also turns row vectors into column vectors. We now define the *inner product* of two vectors

$$\begin{aligned} \boldsymbol{x}^T\boldsymbol{y} &= [x_1, x_2, x_3] \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} \\ &= x_1y_1 + x_2y_2 + x_3y_3 \\ &= \sum_{i=1}^{3} x_iy_i \end{aligned} \tag{70}$$

which is seen to be a scalar. The *outer product* of two vectors produces a matrix

$$\begin{aligned} \boldsymbol{x}\boldsymbol{y}^T &= \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} [y_1, y_2, y_3] \\ &= \begin{bmatrix} x_1y_1 & x_1y_2 & x_1y_3 \\ x_2y_1 & x_2y_2 & x_2y_3 \\ x_3y_1 & x_3y_2 & x_3y_3 \end{bmatrix} \end{aligned} \tag{71}$$

An $N \times M$ matrix has N rows and M columns. The ijth entry of a matrix is the entry on the jth column of the ith row. Given a matrix $\boldsymbol{A}$ (matrices are also often written in bold type) the ijth entry is written as $\boldsymbol{A}_{ij}$. When applying the transpose operator to a matrix the ith row becomes the ith column. That is, if

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \tag{72}$$

then

$$\boldsymbol{A}^T = \begin{bmatrix} a_{11} & a_{21} & a_{31} \\ a_{12} & a_{22} & a_{32} \\ a_{13} & a_{23} & a_{33} \end{bmatrix} \tag{73}$$

A matrix is *symmetric* if $\boldsymbol{A}_{ij} = \boldsymbol{A}_{ji}$. Another way to say this is that, for symmetric matrices, $\boldsymbol{A} = \boldsymbol{A}^T$.

Two matrices can be multiplied if the number of columns in the first matrix equals the number of rows in the second. Multiplying $\boldsymbol{A}$, an $N \times M$ matrix, by $\boldsymbol{B}$, an $M \times K$ matrix, results in $\boldsymbol{C}$, an $N \times K$ matrix. The ijth entry in $\boldsymbol{C}$ is the inner product between the ith row in $\boldsymbol{A}$ and the jth column in $\boldsymbol{B}$. As an example

$$\begin{bmatrix} 2 & 3 & 4 \\ 5 & 6 & 7 \end{bmatrix} \begin{bmatrix} 1 & 3 & 7 & 2 \\ 4 & 3 & 4 & 1 \\ 5 & 6 & 4 & 2 \end{bmatrix} = \begin{bmatrix} 34 & 39 & 42 & 15 \\ 64 & 75 & 87 & 30 \end{bmatrix} \tag{74}$$

Given two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ we note that

$$(\boldsymbol{AB})^T = \boldsymbol{B}^T\boldsymbol{A}^T \tag{75}$$

5.2 Properties of matrix multiplication

Matrix multiplication is associative

$$(\boldsymbol{AB})\boldsymbol{C} = \boldsymbol{A}(\boldsymbol{BC}) \tag{76}$$

distributive

$$\boldsymbol{A}(\boldsymbol{B} + \boldsymbol{C}) = \boldsymbol{AB} + \boldsymbol{AC} \tag{77}$$

but not commutative

$$\boldsymbol{AB} \neq \boldsymbol{BA} \tag{78}$$

5.3 Covariance matrices

In the previous chapter the covariance, σ_{xy}, between two variables x and y was defined. Given p variables there are $p \times p$ covariances to take account of. If we write the covariances between variables x_i and x_j as σ_{ij} then all the covariances can be summarised in a *covariance matrix* which we write below for $p = 3$

$$\boldsymbol{C} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_2^2 & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 \end{bmatrix} \tag{79}$$

The ith diagonal element is the covariance between the ith variable and itself which is simply the variance of that variable; we therefore write σ_i^2 instead of σ_{ii}. Also, note that because $\sigma_{ij} = \sigma_{ji}$ covariance matrices are symmetric.

We now look at computing a covariance matrix from a given data set. Suppose we have p variables and that a single observation $\boldsymbol{x}_i$ (a row vector) consists of measuring these variables and suppose there are N such observations. We now make a matrix $\boldsymbol{X}$ by putting each $\boldsymbol{x}_i$ into the ith row. The matrix $\boldsymbol{X}$ is therefore an $N \times p$ matrix whose rows are made up of different observation vectors. If all the variables have zero mean then the covariance matrix can then be evaluated as

$$\boldsymbol{C} = \frac{1}{N-1}\boldsymbol{X}^T\boldsymbol{X} \tag{80}$$

This is a multiplication of a $p \times N$ matrix, $\boldsymbol{X}^T$, by a $N \times p$ matrix, $\boldsymbol{X}$, which results in a $p \times p$ matrix. To illustrate the use of covariance matrices for time

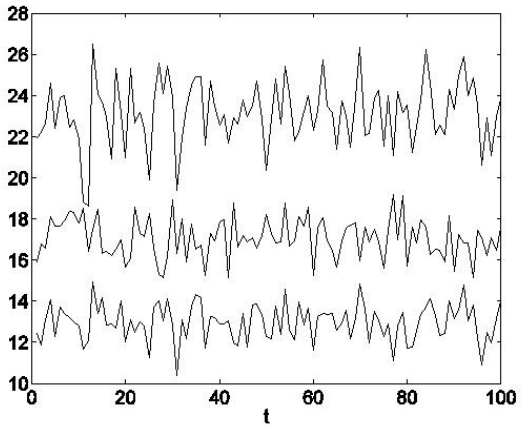


Figure 5: *Three time series having the covariance matrix* $\boldsymbol{C}_1$ *and mean vector* $\boldsymbol{m}_1$ *shown in the text. The top and bottom series have high covariance but none of the other pairings do.*

series, figure 5 shows 3 time series which have the following covariance relation

$$\boldsymbol{C}_1 = \begin{bmatrix} 1 & 0.1 & 1.6 \\ 0.1 & 1 & 0.2 \\ 1.6 & 0.2 & 2.0 \end{bmatrix} \tag{81}$$

and mean vector

$$\boldsymbol{m}_1 = [13, 17, 23]^T \tag{82}$$

5.4 Diagonal matrices

A *diagonal matrix* is a square matrix ($M = N$) where all the entries are zero except along the diagonal. For example

$$\boldsymbol{D} = \begin{bmatrix} 4 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 6 \end{bmatrix} \tag{83}$$

There is also a more compact notation for the same matrix

$$\boldsymbol{D} = diag([4, 1, 6]) \tag{84}$$

If a covariance matrix is diagonal it means that the covariances between variables are zero, that is, the variables are all uncorrelated. Non-diagonal covariance matrices are known as *full* covariance matrices. If $\boldsymbol{V}$ is a vector of variances $\boldsymbol{V} = [\sigma_1^2, \sigma_2^2, \sigma_3^2]^T$ then the corresponding diagonal covariance matrix is $\boldsymbol{V}_d = diag(\boldsymbol{V})$.

5.5 The correlation matrix

The correlation matrix, $\boldsymbol{R}$, can be derived from the covariance matrix by the equation

$$\boldsymbol{R} = \boldsymbol{BCB} \tag{85}$$

where $\boldsymbol{B}$ is a diagonal matrix of inverse standard deviations

$$\boldsymbol{B} = diag([1/\sigma_1, 1/\sigma_2, 1/\sigma_3]) \tag{86}$$

5.6 The identity matrix

The identity matrix is a diagonal matrix with ones along the diagonal. Multiplication of any matrix, $\boldsymbol{X}$ by the identity matrix results in $\boldsymbol{X}$. That is

$$\boldsymbol{IX} = \boldsymbol{X} \tag{87}$$

The identity matrix is the matrix equivalent of multiplying by 1 for scalars.

5.7 Matrix inverse

Given a matrix $\boldsymbol{X}$ its inverse $\boldsymbol{X}^{-1}$ is defined by the properties

$$\begin{aligned} \boldsymbol{X}^{-1}\boldsymbol{X} &= \boldsymbol{I} \\ \boldsymbol{X}\boldsymbol{X}^{-1} &= \boldsymbol{I} \end{aligned} \tag{88}$$

where $\boldsymbol{I}$ is the identity matrix. The inverse of a diagonal matrix with entries d_{ii} is another diagonal matrix with entries $1/d_{ii}$. This satisfies the definition of an inverse, eg.

$$\begin{bmatrix} 4 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 6 \end{bmatrix} \begin{bmatrix} 1/4 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1/6 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{89}$$

More generally, the calculation of inverses involves a lot more computation. Before looking at the general case we first consider the problem of solving simultaneous equations. These constitute relations between a set of *input or independent* variables $\boldsymbol{x}_i$ and a set of *output or dependent* variables y_i. Each input-output pair constitutes an observation. In the following example we consider just $N = 3$ observations and $p = 3$ dimensions per observation

$$\begin{array}{lll} 2w_1 & +w_2 + w_3 & = 5 \\ 4w_1 & -6w_2 & = -2 \\ -2w_1 & +7w_2 + 2w_3 & = 9 \end{array}$$

which can be written in matrix form

$$\begin{bmatrix} 2 & 1 & 1 \\ 4 & -6 & 0 \\ -2 & 7 & 2 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix} = \begin{bmatrix} 5 \\ -2 \\ 9 \end{bmatrix} \tag{90}$$

or in matrix form

$$\boldsymbol{Xw} = \boldsymbol{y} \tag{91}$$

This system of equations can be solved in a systematic way by subtracting multiples of the first equation from the second and third equations and then subtracting multiples of the second equation from the third. For example, subtracting twice the first equation from the second and -1 times the first from the third gives

$$\begin{bmatrix} 2 & 1 & 1 \\ 0 & -8 & -2 \\ 0 & 8 & 3 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix} = \begin{bmatrix} 5 \\ -12 \\ 14 \end{bmatrix} \tag{92}$$

Then, subtracting -1 times the second from the third gives

$$\begin{bmatrix} 2 & 1 & 1 \\ 0 & -8 & -2 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix} = \begin{bmatrix} 5 \\ -12 \\ 2 \end{bmatrix} \tag{93}$$

This process is known as *forward elimination*. We can then substitute the value for w_3 from the third equation into the second etc. This process is *back-substitution*. The two processes are together known as *Gaussian elimination*. Following this through for our example we get $\boldsymbol{w} = [1, 1, 2]^T$.

When we come to invert a matrix (as opposed to solve a system of equations as in the previous example) we start with the equation

$$\boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{I} \tag{94}$$

and just write down all the entries in the $\boldsymbol{A}$ and $\boldsymbol{I}$ matrices in one big matrix

$$\begin{bmatrix} 2 & 1 & 1 & 1 & 0 & 0 \\ 4 & -6 & 0 & 0 & 1 & 0 \\ -2 & 7 & 2 & 0 & 0 & 1 \end{bmatrix} \tag{95}$$

We then perform forward elimination [3] until the part of the matrix corresponding to $\boldsymbol{A}$ equals the identity matrix; the matrix on the right is then $\boldsymbol{A}^{-1}$ (this is because in equation 94 if $\boldsymbol{A}$ becomes $\boldsymbol{I}$ then the left hand side is $\boldsymbol{A}^{-1}$ and the right side must equal the left side). We get

$$\begin{bmatrix} 1 & 0 & 0 & \frac{12}{16} & \frac{-5}{16} & \frac{-6}{16} \\ 0 & 1 & 0 & \frac{4}{8} & \frac{-3}{8} & \frac{-2}{8} \\ 0 & 0 & 1 & -1 & 1 & 1 \end{bmatrix} \tag{96}$$

This process is known as the *Gauss-Jordan* method. For more details see Strang's excellent book on Linear Algebra [6] where this example was taken from.

Inverses can be used to solve equations of the form $\boldsymbol{X}\boldsymbol{w} = \boldsymbol{y}$. This is achieved by multiplying both sides by $\boldsymbol{X}^{-1}$ giving

$$\boldsymbol{w} = \boldsymbol{X}^{-1}\boldsymbol{y} \tag{97}$$

Hence,

$$\begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix} = \begin{bmatrix} \frac{12}{16} & \frac{-5}{16} & \frac{-6}{16} \\ \frac{4}{8} & \frac{-3}{8} & \frac{-2}{8} \\ -1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 5 \\ -2 \\ 9 \end{bmatrix} \tag{98}$$

[3] We do not perform back-substitution but instead continue with forward elimination until we get a diagonal matrix.

which also gives $\boldsymbol{w} = [1, 1, 2]^T$.

The inverse of a product of matrices is given by

$$(\boldsymbol{AB})^{-1} = \boldsymbol{B}^{-1}\boldsymbol{A}^{-1} \tag{99}$$

Only square matrices are invertible because, for $\boldsymbol{y} = \boldsymbol{Ax}$, if $\boldsymbol{y}$ and $\boldsymbol{x}$ are of different dimension then we will not necessarily have a one-to-one mapping between them.

5.8 Orthogonality

The length of a d-element vector $\boldsymbol{x}$ is written as $||\boldsymbol{x}||$ where

$$\begin{aligned} ||\boldsymbol{x}||^2 &= \sum_{i=1}^{d} x_i^2 \\ &= \boldsymbol{x}^T\boldsymbol{x} \end{aligned} \tag{100}$$

Two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ are *orthogonal* if

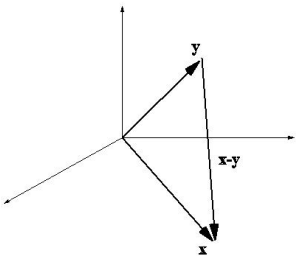


Figure 6: *Two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$. These vectors will be orthogonal if they obey Pythagoras' relation ie. that the sum of the squares of the sides equals the square of the hypoteneuse.*

$$||\boldsymbol{x}||^2 + ||\boldsymbol{y}||^2 = ||\boldsymbol{x} - \boldsymbol{y}||^2 \tag{101}$$

That is, if

$$x_1^2 + ... + x_d^2 + y_1^2 + ... + y_d^2 = (x_1 - y_1)^2 + ... + (x_d - y_d)^2 \tag{102}$$

Expanding the terms on the right and re-arranging leaves only the cross-terms

$$\begin{aligned} x_1y_1 + ... + x_dy_d &= 0 \\ \boldsymbol{x}^T\boldsymbol{y} &= 0 \end{aligned} \tag{103}$$

That is, two vectors are orthogonal if their inner product is zero.

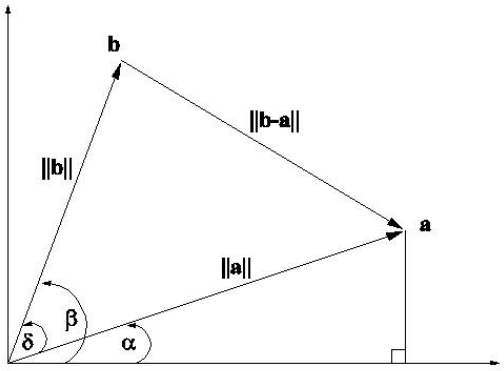


Figure 7: *Working out the angle between two vectors.*

5.9 Angles between vectors

Given a vector $\boldsymbol{b} = [b_1, b_2]^T$ and a vector $\boldsymbol{a} = [a_1, a_2]^T$ we can work out that

$$\begin{aligned} \cos\alpha &= \frac{a_1}{||\boldsymbol{a}||} \\ \sin\alpha &= \frac{a_2}{||\boldsymbol{a}||} \\ \cos\beta &= \frac{b_1}{||\boldsymbol{b}||} \\ \sin\beta &= \frac{b_2}{||\boldsymbol{b}||} \end{aligned} \tag{104}$$

(105)

Now, $cos\delta = cos(\beta - \alpha)$ which we can expand using the trig identity

$$\cos(\beta - \alpha) = \cos\beta\cos\alpha + \sin\beta\sin\alpha \tag{106}$$

Hence

$$\cos(\delta) = \frac{a_1 b_1 + a_2 b_2}{||\boldsymbol{a}||||\boldsymbol{b}||} \tag{107}$$

More generally, we have

$$\cos(\delta) = \frac{\boldsymbol{a}^T\boldsymbol{b}}{||\boldsymbol{a}||||\boldsymbol{b}||} \tag{108}$$

Because, $\cos\pi/2 = 0$, this again shows that vectors are orthogonal for $\boldsymbol{a}^T\boldsymbol{b} = 0$. Also, because $|\cos\delta| \leq 1$ where $|x|$ denotes the absolute value of x we have

$$|\boldsymbol{a}^T\boldsymbol{b}| \leq ||\boldsymbol{a}||||\boldsymbol{b}|| \tag{109}$$

which is known as the *Schwarz Inequality*.

5.10 Projections

The projection of a vector $\boldsymbol{b}$ onto a vector $\boldsymbol{a}$ results in a projection vector $\boldsymbol{p}$ which is the point on the line $\boldsymbol{a}$ which is closest to the point $\boldsymbol{b}$. Because $\boldsymbol{p}$ is a

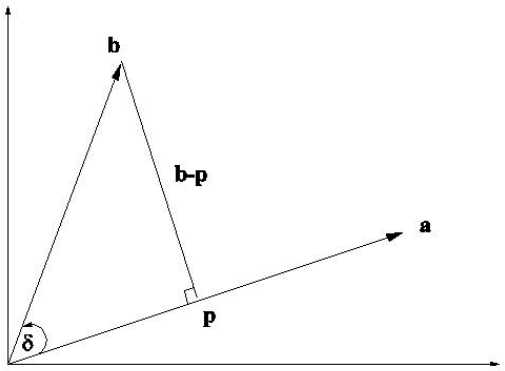


Figure 8: *The projection of* $\boldsymbol{b}$ *onto* $\boldsymbol{a}$ *is the point on* $\boldsymbol{a}$ *which is closest to* $\boldsymbol{b}$.

point on $\boldsymbol{a}$ it must be some scalar multiple of it. That is

$$\boldsymbol{p} = w\boldsymbol{a} \tag{110}$$

where w is some coefficient. Because $\boldsymbol{p}$ is the point on $\boldsymbol{a}$ *closest* to $\boldsymbol{b}$ this means that the vector $\boldsymbol{b} - \boldsymbol{p}$ is orthogonal to $\boldsymbol{a}$. Therefore

$$\begin{aligned} \boldsymbol{a}^T(\boldsymbol{b} - \boldsymbol{p}) &= 0 \\ \boldsymbol{a}^T(\boldsymbol{b} - w\boldsymbol{a}) &= 0 \end{aligned} \tag{111}$$

Re-arranging gives

$$w = \frac{\boldsymbol{a}^T\boldsymbol{b}}{\boldsymbol{a}^T\boldsymbol{a}} \tag{112}$$

and

$$\boldsymbol{p} = \frac{\boldsymbol{a}^T\boldsymbol{b}}{\boldsymbol{a}^T\boldsymbol{a}}\boldsymbol{a} \tag{113}$$

We refer to $\boldsymbol{p}$ as the *projection vector* and to w as the *projection*.

6 Multiple Regression

A good practical introduction to the material on regression is presented by Kleinbaum et al. [3]. More details of matrix manipulations are available in Weisberg [7] and Strang has a great in-depth intro to linear algebra [6]. See also

relevant material in *Numerical Recipes* [5]. See Chatfield's book on multivariate analysis for more details [1].

For a multivariate linear data set, the dependent variable y_i is modelled as a linear combination of the input variables $\boldsymbol{x}_i$ and an error term [4]

$$y_i = \boldsymbol{x}_i \boldsymbol{w} + e_i \tag{114}$$

where $\boldsymbol{x}_i$ is a row vector, $\boldsymbol{w}$ is a column vector and e_i is an error. The overall goodness of fit can be assessed by the least squares cost function

$$E = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \tag{115}$$

where $\hat{\boldsymbol{y}}_i = \boldsymbol{x}_i \boldsymbol{w}$.

6.1 Estimating the weights

The least squares cost function can be written in matrix notation as

$$E = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})^T (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}) \tag{116}$$

where $\boldsymbol{X}$ is an N-by-p matrix whose rows are made up of different input vectors and $\boldsymbol{y}$ is a vector of targets. The weight vector that minimises this cost function can be calculated by setting the first derivative of the cost function to zero and solving the resulting equation.

By expanding the brackets and collecting terms (using the matrix identity $(\boldsymbol{AB})^T = \boldsymbol{B}^T \boldsymbol{A}^T$ we get

$$E = \boldsymbol{y}^T \boldsymbol{y} - 2\boldsymbol{w}^T \boldsymbol{X}^T \boldsymbol{y} + \boldsymbol{w}^T \boldsymbol{X}^T \boldsymbol{X} \boldsymbol{w} \tag{117}$$

The derivative with respect to $\boldsymbol{w}$ is [5]

$$\frac{\partial E}{\partial \boldsymbol{w}} = -2\boldsymbol{X}^T \boldsymbol{y} + 2\boldsymbol{X}^T \boldsymbol{X} \boldsymbol{w} \tag{118}$$

Equating this derivative to zero gives

$$(\boldsymbol{X}^T \boldsymbol{X})\boldsymbol{w} = \boldsymbol{X}^T \boldsymbol{y} \tag{119}$$

which, in regression analysis, is known as the 'normal equation'. Hence,

$$\hat{\boldsymbol{w}} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y} \tag{120}$$

This is the general solution for multivariate linear regression [6]. It is a unique minimum of the least squares error function (ie. this is the only solution).

Once the weights have been estimated we can then estimate the error or noise variance from

$$\sigma_e^2 = \frac{1}{N-1} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \tag{121}$$

[4] The error term is introduced because, very often, given a particular data set it will not be possible to find an exact linear relationship between $\boldsymbol{x}_i$ and y_i for every i. We therefore cannot directly estimate the weights as $\boldsymbol{X}^{-1}\boldsymbol{y}$.

[5] From matrix calculus [4] we know that the derivative of $\boldsymbol{c}^T \boldsymbol{B} \boldsymbol{c}$ with respect to $\boldsymbol{c}$ is $(\boldsymbol{B}^T + \boldsymbol{B})\boldsymbol{c}$. Also we note that $\boldsymbol{X}^T \boldsymbol{X}$ is symmetric.

[6] In practice we can use the equivalent expression $\hat{\boldsymbol{w}} = \boldsymbol{X}^{+1}\boldsymbol{y}$ where $\boldsymbol{X}^{+1}$ is the pseudo-inverse [6]. This method is related to Singular Value Decomposition and is discussed later.

6.2 Understanding the solution

If the inputs are zero mean then the input covariance matrix multiplied by N-1 is

$$\boldsymbol{C}_x = \boldsymbol{X}^T\boldsymbol{X} \tag{122}$$

The weights can therefore be written as

$$\hat{\boldsymbol{w}} = \boldsymbol{C}_x^{-1}\boldsymbol{X}^T\boldsymbol{y} \tag{123}$$

ie. the inverse covariance matrix times the inner products of the inputs with the output (the ith weight will involve the inner product of the ith input with the output).

6.2.1 Single input

For a single input $\boldsymbol{C}_x^{-1} = 1/(N-1)\sigma_{x_1}^2$ and $\boldsymbol{X}^T\boldsymbol{y} = (N-1)\sigma_{x_1 y}$. Hence

$$\hat{w}_1 = \frac{\sigma_{x_1 y}}{\sigma_{x_1}^2} \tag{124}$$

This is *exactly* the same as the estimate for the slope in linear regression (first lecture). This is re-assuring.

6.2.2 Uncorrelated inputs

For two uncorrelated inputs

$$\boldsymbol{C}_x^{-1} = \begin{bmatrix} \frac{1}{(N-1)\sigma_{x_1}^2} & 0 \\ 0 & \frac{1}{(N-1)\sigma_{x_2}^2} \end{bmatrix} \tag{125}$$

We also have

$$\boldsymbol{X}^T\boldsymbol{y} = \begin{bmatrix} (N-1)\sigma_{x_1,y} \\ (N-1)\sigma_{x_2,y} \end{bmatrix} \tag{126}$$

The two weights are therefore

$$\begin{aligned} \hat{w}_1 &= \frac{\sigma_{x_1 y}}{\sigma_{x_1}^2} \\ \hat{w}_2 &= \frac{\sigma_{x_2 y}}{\sigma_{x_2}^2} \end{aligned} \tag{127}$$

Again, these solutions are the same as for the univariate linear regression case.

6.2.3 General case

If the inputs are correlated then a coupling is introduced in the estimates of the weights; weight 1 becomes a function of $\sigma_{x_2 y}$ as well as $\sigma_{x_1 y}$

$$\hat{\boldsymbol{w}} = \begin{bmatrix} \sigma_{x_1}^2 & \sigma_{x_1 x_2} \\ \sigma_{x_1 x_2} & \sigma_{x_2}^2 \end{bmatrix}^{-1} \begin{bmatrix} \sigma_{x_1,y} \\ \sigma_{x_2,y} \end{bmatrix} \tag{128}$$

6.3 Inference

Some of the inputs in a linear regression model may be very useful in predicting the output. Others, not so. So how do we find which inputs or *features* are useful ? This problem is known as feature selection.

The problem is tackled by looking at the coefficients of each input (ie. the weights) and seeing if they are significantly non-zero. The procedure is identical to that described for univariate linear regression.

The only added difficulty is that we have more inputs and more weights, but the procedure is basically the same. Firstly, we have to estimate the variance on each weight. This is done in the next section. We then compare each weight to zero using a t-test.

6.3.1 Functions of random vectors

For a vector of random variables, $\boldsymbol{z}$, and a matrix of constants, $\boldsymbol{C}$, and a vector of constants, $\boldsymbol{d}$, we have

$$Var(\boldsymbol{C}\boldsymbol{z} + \boldsymbol{d}) = \boldsymbol{C}[Var(\boldsymbol{z})]\boldsymbol{C}^T \tag{129}$$

where, here, Var() denotes a covariance matrix. This is a generalisation of the result for scalar random variables $Var(cz) = c^2 Var(z)$.

The covariance between a pair of random vectors is given by

$$Var(\boldsymbol{C}_1\boldsymbol{z}, \boldsymbol{C}_2\boldsymbol{z}) = \boldsymbol{C}_1[Var(\boldsymbol{z})]\boldsymbol{C}_2^T \tag{130}$$

6.3.2 The weight covariance matrix

Different instantiations of target noise will generate different estimated weight vectors according to equation 120. For the case of Gaussian noise we do not actually have to compute the weights on many instantiations of the target noise and then compute the sample covariance [7]; the corresponding weight covariance matrix is given by the equation

$$\boldsymbol{\Sigma} = Var((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}) \tag{131}$$

Substituting $\boldsymbol{y} = \boldsymbol{X}\hat{\boldsymbol{w}} + \boldsymbol{e}$ gives

$$\boldsymbol{\Sigma} = Var((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} + (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{e}) \tag{132}$$

This is in the form of $Var(\boldsymbol{C}\boldsymbol{z} + \boldsymbol{d})$ (see earlier) with $\boldsymbol{d}$ being given by the first term which is constant, $\boldsymbol{C}$ being given by $(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ and $\boldsymbol{z}$ being given by $\boldsymbol{e}$. Hence,

$$\begin{aligned} \boldsymbol{\Sigma} &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T[Var(\boldsymbol{e})][(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T]^T \\ &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T(\sigma_e^2\boldsymbol{I})[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T]^T \\ &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T(\sigma_e^2\boldsymbol{I})\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1} \end{aligned} \tag{133}$$

[7] But this type of procedure is the basis of bootstrap estimates of parameter variances. See [2].

Re-arranging further gives

$$\mathbf{\Sigma} = \sigma_e^2 (\boldsymbol{X}^T \boldsymbol{X})^{-1} \tag{134}$$

In the appendix we show that this can be evaluated as

$$\mathbf{\Sigma} = \sigma_e^2 (\boldsymbol{X}^T \boldsymbol{X})^{-1} \tag{135}$$

The correlation in the inputs introduces a correlation in the weights; for uncorrelated inputs the weights will be uncorrelated. The variance of the jth weight, w_j, is then given by the jth diagonal entry in the covariance matrix

$$\sigma_{w_j}^2 = \mathbf{\Sigma}_{jj} \tag{136}$$

To see if a weight is significantly non-zero we then compute $CDF_t(t)$ (the cumulative density function; see earlier lecture) where $t = w_j/\sigma_{w_j}$ and if it is above some threshold, say $p = 0.05$, the corresponding feature is removed.

Note that this procedure, which is based on a t-test, is exactly equivalent to a similar procedure based on a partial F-test (see, for example, [3] page 128).

If we do remove a weight then we must recompute all the other weights (and variances) *before* deciding whether or not the other weights are significantly non-zero. This usually proceeds in a stepwise manner where we start with a large number of features and reduce them as necessary (*stepwise backward selection*) or gradually build up the number of features (*stepwise forward selection*) [3].

Note that, if the weights were uncorrelated we could do feature selection in a single step; we would not have to recompute weight values after each weight removal. This provides one motivation for the use of orthogonal transforms in which the weights *are* uncorrelated. Such transforms include Fourier and Wavelet transforms as we shall see in later lectures.

6.4 Equivalence of t-test and F-test for feature selection

When adding a new variable x_p to a regression model we can test to see if the increase in the proportion of variance explained is *significant* by computing

$$F = \frac{(N-1)\sigma_y^2 \left[r^2(y, \hat{y}_p) - r^2(y, \hat{y}_{p-1}) \right]}{\sigma_e^2(p)} \tag{137}$$

where $r^2(y, \hat{y}_p)$ is the square of the correlation between y and the regression model with all p variables (ie. including x_p) and $r^2(y, \hat{y}_{p-1})$ is the square of the correlation between y and the regression model without x_p. The denominator is the noise variance from the model including x_p. This statistic is distributed according to the F-distribution with $v_1 = 1$ and $v_2 = N - p - 2$ degrees of freedom.

This test is identical to the double sided t-test on the t-statistic computed from the regression coefficient a_p, described in this lecture (see also page 128 of [3]). This test is also equivalent to seeing if the partial correlation between x_p and y is significantly non-zero (see page 149 of [3]).

6.5 Example

Suppose we wish to predict a time series x_3 from two other time series x_1 and x_2. We can do this with the following regression model [8]

$$x_3 = w_0 + w_1 x_1 + w_2 x_2 \tag{138}$$

and the weights can be found using the previous formulae. To cope with the constant, w_0, we augment the $\boldsymbol{X}$ vector with an additional column of 1's.

We analyse data having covariance matrix $\boldsymbol{C}_1$ and mean vector $\boldsymbol{m}_1$ (see equations 82 and 81 in an earlier lecture). $N = 50$ data points were generated and are shown in Figure 9. The weights were then estimated from equation 120 as

$$\begin{aligned} \hat{\boldsymbol{w}} &= [w_1, w_2, w_0]^T \\ &= [1.7906, -0.0554, 0.6293]^T \end{aligned} \tag{139}$$

Note that w_1 is much bigger than w_2. The weight covariance matrix was estimated from equation 135 as

$$\boldsymbol{\Sigma} = \begin{bmatrix} 0.0267 & 0.0041 & -0.4197 \\ 0.0041 & 0.0506 & -0.9174 \\ -0.4197 & -0.9174 & 21.2066 \end{bmatrix} \tag{140}$$

giving $\sigma_{w_1} = 0.1634$ and $\sigma_{w_2} = 0.2249$. The corresponding t-statistics are $t_1 = 10.96$ and $t_2 = -0.2464$ giving p-values of 10^{-15} and 0.4032. This indicates that the first weight is significantly different from zero but the second weight is not ie. x_1 is a good predictor of x_3 but x_2 is not. We can therefore remove x_2 from our regression model.

Question: But what does linear regression tell us about the data that the correlation/covariance matrix does'nt ? *Answer:* Partial correlations.

6.6 Partial Correlation

Remember (see eg. equation 53 from lecture 1), the square of the correlation coefficient between two variables x_1 and y is given by

$$r^2_{x_1 y} = \frac{\sigma_y^2 - \sigma_e^2(x_1)}{\sigma_y^2} \tag{141}$$

where $\sigma_e^2(x_1)$ is the variance of the errors from using a linear regression model based on x_1 to predict y. Writing $\sigma_y^2 = \sigma_e^2(0)$, ie. the error with no predictive variables

$$r^2_{x_1 y} = \frac{\sigma_e^2(0) - \sigma_e^2(x_1)}{\sigma_e^2(0)} \tag{142}$$

When we have a second predictive variable x_2, the square of the *partial correlation* between x_2 and y is defined as

$$r^2_{x_2 y|x_1} = \frac{\sigma_e^2(x_1) - \sigma_e^2(x_1, x_2)}{\sigma_e^2(x_1)} \tag{143}$$

[8] Strictly, we can only apply this model if the samples *within* each time series are independent (see later). To make them independent we can randomize the time index thus removing any correlation between lagged samples. We therefore end up with a random variables rather than time series.

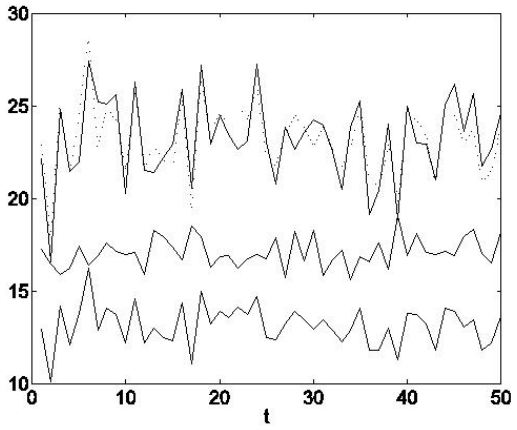


Figure 9: *Three time series having the correlation matrix $\boldsymbol{C}_1$ and mean vector $\boldsymbol{m}_1$ shown in the text. The dotted line shows the value of the third time series as predicted from the other two using a regression model.*

where $\sigma_e^2(x_1, x_2)$ is the variance of the errors from the regression model based on x_1 and x_2. It's the extra proportion of variance in y explained by x_2. It's different to $r^2_{x_2 y}$ because x_2 may be correlated to x_1 which itself explains some of the variance in y. After *controlling* for this, the resulting proportionate reduction in variance is given by $r^2_{x_2 y|x_1}$. More generally, we can define pth order partial correlations which are the correlations between two variables after controlling for p variables.

The sign of the partial correlation is given by the sign of the corresponding regression coefficient.

6.6.1 Relation to regression coefficients

Partial correlations are to regression coefficients what the correlation is to the slope in univariate linear regression. If the partial correlation is significantly non-zero then the corresponding regression coefficient will also be. And vice-versa.

References

[1] C. Chatfield. *An Introduction to Multivariate Analysis.* Chapman and Hall, 1991.

[2] B. Efron and R.J. Tibshirani. *An introduction to the bootstrap.* Chapman and Hall, 1993.

[3] D.G. Kleinbaum, L.L. Kupper, and K.E. Muller. *Applied Regression Analysis and Other Multivariable Methods.* PWS-Kent, Boston, 1988.

[4] J.R. Magnus and H. Neudecker. *Matrix Differential Calculus with Applications in Statistics and Econometrics*. John Wiley, 1997.

[5] W. H. Press, S.A. Teukolsky, W.T. Vetterling, and B.V.P. Flannery. *Numerical Recipes in C*. Cambridge, 1992.

[6] G. Strang. *Linear algebra and its applications*. Harcourt Brace, 1988.

[7] S. Weisberg. *Applied Linear Regression*. John Wiley, 1980.