STATISTICS 312
MATHEMATICAL METHODS IN STATISTICS

Doug Wiens*

March 16, 2015

Contents

Part I

MATRIX ALGEBRA

1. Introduction; matrix manipulations

- Outline of this course:
  - Linear algebra – regression (linear/nonlinear), multivariate analysis, more generally linear models and linear approximations.
  - Real analysis/calculus – theory of statistical distributions, optimal selection of statistical procedures (e.g. determine a parameter estimate to minimize a certain loss function), approximations of intractable procedures with simpler ones.
  - Multivariable calculus/Optimization – find numbers or functions minimizing certain objectives, e.g. designing experiments for maximum information/minimum variance etc.; associated numerical methods.

- In Statistics, matrices are often merely convenient ways to store and refer to data. As well – in Regression for instance – there are important structural features which come from examining the algebraic properties of the *vector space* formed from all linear combinations of the columns of a matrix.

- **Example**: Within a certain range, the price (Y) of a product varies linearly with demand (x):

$$Y = \beta_0 + \beta_1 x + error.$$

The 'error' will be discussed later, but for now – why a small x and a big Y? Now we observe values (x_i, Y_i), $i = 1, ..., n$, which we aim to use to, among other things, estimate the intercept and the slope 'parameters'. Then for each i we have the data represented as

$$Y_i = \beta_0 1 + \beta_1 x_i + \varepsilon_i$$

leading to .... the more compact and useful formulation

$$\mathbf{y} = \mathbf{X}\beta + \boldsymbol{\varepsilon}$$

where $\mathbf{X}$ is ... . We will return to this.

- The regression example above employed the product of a matrix and a vector. The formula for the product of two matrices – that the $(i,j)^{th}$ element of the product $\mathbf{AB}$ of an $n \times p$ matrix $\mathbf{A}$ and a $p \times m$ matrix $\mathbf{B}$ is given by

$$[\mathbf{AB}]_{ij} = \sum_{k=1}^{p} \mathbf{A}_{ik}\mathbf{B}_{kj},$$

  is perhaps of limited usefulness – there are usually nicer expressions, easier to work with. But here is one place where it is appropriate.

  - **Markov chains**: Suppose that, at time 'm', a process is in one of several 'states' – e.g. the economy in year m might be booming, or in recession, etc. We say this process is *Markovian* if the probability of being in a certain state at time $m+1$ depends only on the state at time m. If there are s possible states then there will be an $s \times s$ 'transition matrix' $\mathbf{P}$ given by

$$P_{ij} = \mathsf{Pr}\left(X_{m+1} = j | X_m = i\right).$$

e.g. $s = 2$: 'booming', 'in recession';
Pr (booming next year|in recession this year), etc.
Define the n-step transition matrix $\mathbf{P}^{(n)}$ by

$$P_{ij}^{(n)} = \Pr\left(X_{m+n} = j | X_m = i\right).$$

Then $\mathbf{P}^{(1)} = \mathbf{P}$, and (here we'll argue informally; this can later be made more formal)

$$P_{ij}^{(2)} = \sum_{k=1}^{s} \Pr\left(i \to k \to j\right) = \sum_{k=1}^{s} P_{ik} P_{kj} = \left[\mathbf{P}^2\right]_{ij}.$$

Thus $\mathbf{P}^{(2)} = \mathbf{P}^2$; in general $\mathbf{P}^{(n)} = \mathbf{P}^n$.

- In many statistical applications, one should treat either the *rows* or the *columns* of matrices as the basic elements.

  – Define (column) vector in $\mathbb{R}^n$; sum, transpose, scalar product, outer product

  – Matrix as a column of rows, or row of columns

$$\mathbf{A}_{n\times p} = \begin{pmatrix} \mathbf{a}_1' \\ \mathbf{a}_2' \\ \vdots \\ \mathbf{a}_n' \end{pmatrix} = \left(\boldsymbol{\alpha}_1 \vdots \boldsymbol{\alpha}_2 \vdots \cdots \vdots \boldsymbol{\alpha}_p\right).$$

- If $\mathbf{X}_{n\times p}$ has rows $\left\{\mathbf{x}_i'\right\}_{i=1}^{n}$ (note: vectors are columns, rows are transposed vectors), and β is a $p \times 1$ vector, then

$$\mathbf{X}\beta = \begin{pmatrix} \mathbf{x}_1' \\ \vdots \\ \mathbf{x}_i' \\ \vdots \\ \mathbf{x}_n' \end{pmatrix} \beta = \begin{pmatrix} \mathbf{x}_1'\beta \\ \vdots \\ \mathbf{x}_i'\beta \\ \vdots \\ \mathbf{x}_n'\beta \end{pmatrix}.$$

- If $\mathbf{X}_{n\times p}$ has columns $\left\{\mathbf{z}_j\right\}_{j=1}^{p}$, and β is a $p\times 1$ vector, then

$$\mathbf{X}\beta = \begin{bmatrix} \mathbf{z}_1 \cdots \mathbf{z}_j \cdots \mathbf{z}_p \end{bmatrix} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_j \\ \vdots \\ \beta_p \end{pmatrix} = \sum_{j=1}^{p} \mathbf{z}_j\beta_j.$$

- If $\mathbf{X}_{n\times p}$ has columns $\left\{\mathbf{z}_j\right\}_{j=1}^{p}$, and $\mathbf{A}$ is a matrix with n columns, then

$$\mathbf{AX} = \mathbf{A}\begin{bmatrix} \mathbf{z}_1 \cdots \mathbf{z}_j \cdots \mathbf{z}_p \end{bmatrix} = \begin{bmatrix} \mathbf{Az}_1 \cdots \mathbf{Az}_j \cdots \mathbf{Az}_p \end{bmatrix}.$$

– If $\mathbf{X}$ is $n \times p$ and $\mathbf{A}_{m\times n}$ is a matrix with rows $\left\{\mathbf{a}_i'\right\}_{i=1}^{m}$, then

$$\mathbf{AX} = \begin{pmatrix} \mathbf{a}_1' \\ \vdots \\ \mathbf{a}_i' \\ \vdots \\ \mathbf{a}_n' \end{pmatrix} \mathbf{X} = \begin{pmatrix} \mathbf{a}_1'\mathbf{X} \\ \vdots \\ \mathbf{a}_i'\mathbf{X} \\ \vdots \\ \mathbf{a}_n'\mathbf{X} \end{pmatrix}.$$

You should become familiar with all of these, and learn to choose the most appropriate form in an application. **This is absolutely crucial in the development of a facility for matrix manipulation – start practicing it NOW**.

2. More on matrix manipulations

- Block matrices ... a particular example is, with notation as above,

$$\mathbf{A}_{n\times p}\mathbf{B}_{p\times q} = (\boldsymbol{\alpha}_1 \vdots \boldsymbol{\alpha}_2 \vdots \cdots \vdots \boldsymbol{\alpha}_p) \begin{pmatrix} \beta_1' \\ \beta_2' \\ \vdots \\ \beta_p' \end{pmatrix} = \sum_{i=1}^{p} \boldsymbol{\alpha}_i \beta_i'.$$

- More generally, suppose that two matrices are each partitioned:

$$\begin{aligned} \mathbf{P} &= \begin{pmatrix} \mathbf{A}_{p\times m} & \mathbf{B}_{p\times n} \\ \mathbf{C}_{q\times m} & \mathbf{D}_{q\times n} \end{pmatrix} : (p+q)\times(m+n)\,, \\ \mathbf{Q} &= \begin{pmatrix} \mathbf{E}_{m\times r} & \mathbf{F}_{m\times s} \\ \mathbf{G}_{n\times r} & \mathbf{H}_{n\times s} \end{pmatrix} : (m+n)\times(r+s)\,. \end{aligned}$$

  Then the product $\mathbf{PQ}$ is defined and is given by

$$\begin{aligned} \mathbf{PQ} &= \begin{pmatrix} \mathbf{A}_{p\times m} & \mathbf{B}_{p\times n} \\ \mathbf{C}_{q\times m} & \mathbf{D}_{q\times n} \end{pmatrix} \begin{pmatrix} \mathbf{E}_{m\times r} & \mathbf{F}_{m\times s} \\ \mathbf{G}_{n\times r} & \mathbf{H}_{n\times s} \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{AE}+\mathbf{BG} & \mathbf{AF}+\mathbf{BH} \\ \mathbf{CE}+\mathbf{BG} & \mathbf{CF}+\mathbf{BH} \end{pmatrix}. \end{aligned}$$

Example:

$$\mathbf{P} = \begin{pmatrix} 1/2 & \cdots & 0 & 1/2 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & 1/2 & 1/2 \\ v_1 & \cdots & v_n & w \end{pmatrix};$$

what is $\mathbf{P}^2$? [Note: if $0 \leq v_1, ..., v_n, w$ and these sum to 1, then $\mathbf{P}$ is the transition matrix of an $n+1$-state Markov chain. (Why?)]

- **A brief digression to expected values**: Let X be a random variable (r.v.) (formal definition to come later) with (i) distribution function $F(x) = P(X \leq x)$ and probability density function $f(x) = F'(x)$ or (ii) probability mass function $f(x) = P(X = x)$ for $x \in \mathbb{X}$, a finite or countable set. Then the 'expected value' is

$$E[X] = \begin{cases} \text{(i)} \ \int_{-\infty}^{\infty} x f(x) dx, \\ \text{(ii)} \ \sum_{x \in \mathbb{X}} x f(x). \end{cases}$$

Think 'average'. The extension to random vectors ('r.vecs') is immediate, involving multidimen-

sional integrals or sums. A consequence is that

$$E\left[\begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}\right] = \begin{pmatrix} E[X_1] \\ \vdots \\ E[X_n] \end{pmatrix}.$$

(Similarly with random matrices.) We *define* $E[g(\mathbf{X})]$ to be $E[Z]$, where $Z = g(\mathbf{X})$. In principle this requires the derivation of the distribution of Z. It can be shown that this can instead be obtained by integration or summation w.r.t. ('with respect to') the distribution of $\mathbf{X}$. Corresponding to the cases above, this is

$$E[g(\mathbf{X})] = \begin{cases} \text{(i)} \ \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(\mathbf{x}) f(\mathbf{x}) d\mathbf{x}, \\ \text{(ii)} \ \sum_{\mathbf{x} \in \mathbb{X}} g(\mathbf{x}) f(\mathbf{x}), \end{cases}$$

respectively.

- A special consequence is *linearity*: $E[aX+bY] = aE[X]+bE[Y]$ (a, b constants). More generally,

$$E\left[\mathbf{Ax}+\mathbf{b}\right] = \mathbf{A}E\left[\mathbf{x}\right]+\mathbf{b}$$

if $\mathbf{x}$ is a r.vec. and

$$E\left[\mathbf{AXB}+\mathbf{C}\right] = \mathbf{A}E\left[\mathbf{X}\right]\mathbf{B}+\mathbf{C}$$

for a random matrix $\mathbf{X}$. Thus, e.g., if $\boldsymbol{\mu} = E\left[\mathbf{x}\right]$ then

$$\begin{aligned} & E\left[\left(\mathbf{x}-\boldsymbol{\mu}\right)\left(\mathbf{x}-\boldsymbol{\mu}\right)'\right] \\ = \;& E\left[\mathbf{x}\mathbf{x}'-\boldsymbol{\mu}\mathbf{x}'-\mathbf{x}\boldsymbol{\mu}'+\boldsymbol{\mu}\boldsymbol{\mu}'\right] \\ = \;& \text{.... (lab problem).} \end{aligned}$$

The $(i,j)^{th}$ element is

$$E\left[\left(X_i-\mu_i\right)\left(X_j-\mu_j\right)\right] = \text{cov}\left[X_i, X_j\right]$$

($=$ the *variance* if $i=j$). The matrix ('$\boldsymbol{\Sigma}$') is called the *covariance matrix* of the r.vec. $\mathbf{x}$.

3. Regression; vector spaces and subspaces

- **Example**: linear regression. Experimenter observes a variable Y (= response to a medical treatment, say) thought to depend on type of drug used ($x_1 = 1$ for type A, 0 for type B) and amount applied (x_2). Response contains a random component as well (measurement error, model inadequacies, etc.); a tentative linear regression model might be

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

where the β's are unknown *parameters* to be estimated, ε is unobserved *random error*, assumed to have mean 0, constant variance across subjects, possibly also normally distributed.

  - Interpretation of $E[Y|x_1, x_2]$ in the two treatment groups:

$$\begin{aligned} E[Y|x_1 = 0, x_2] &= \beta_0 + \beta_2 x_2, \\ E[Y|x_1 = 1, x_2] &= \beta_0 + \beta_1 + \beta_2 x_2, \end{aligned}$$

    hence β_1 = difference in mean effects of the treatments, if the same amounts are applied.

– Then with $\mathbf{x}' = (1, x_1, x_2)$, $\beta = (\beta_0, \beta_1, \beta_2)'$:

$$Y = \mathbf{x}'\beta + \varepsilon;\ E[Y|\mathbf{x}] = \mathbf{x}'\beta.$$

- Take data:

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1'\beta \\ \mathbf{x}_2'\beta \\ \vdots \\ \mathbf{x}_n'\beta \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix};$$

more concisely

$$\mathbf{Y} = \begin{pmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \vdots \\ \mathbf{x}_n' \end{pmatrix} \beta + \boldsymbol{\varepsilon} = \mathbf{X}\beta + \boldsymbol{\varepsilon}.$$

Here the observations (<u>rows</u>) have been singled out as the relevant objects. Much of the theory will hinge on the representation of $E[Y]$ as a linear combination of the <u>columns</u> of $\mathbf{X}$, with coefficients β:

$$\mathbf{X} = (\mathbf{1} \vdots \mathbf{z}_1 \vdots \mathbf{z}_2);\ E[\mathbf{Y}] = \mathbf{1}\beta_0 + \mathbf{z}_1\beta_1 + \mathbf{z}_2\beta_2.$$

Extension to $p > 3$ columns is immediate.

- Estimation of β: Given an estimate $\hat{\beta}$ one estimates $E[Y|\mathbf{x}]$ by $\mathbf{x}'\hat{\beta}$ with *residuals* (= "*observed - expected*")

$$e_i = Y_i - \mathbf{x}_i'\hat{\beta},$$

and residual vector $\mathbf{e} = \mathbf{Y} - \mathbf{X}\hat{\beta}$.

- We define the (Euclidean) norm, i.e. the length, of a vector by $\|\mathbf{e}\| = \sqrt{\sum e_i^2} = \sqrt{\mathbf{e}'\mathbf{e}}$.

- Least Squares Principle: Choose

$$\hat{\beta} = \text{arg min}\, ||\mathbf{Y} - \mathbf{X}\beta||^2,$$

minimizing the Sum of Squares of the Residuals (or Errors, hence 'SSE').

- **Vector spaces**. A particularly important example in Statistics, and one which you ought to keep in mind as motivation for what we do next, is

$\mathbb{R}^n =$ all $n-$dimensional vectors with real elements,

and its 'subspaces'.

- We list a number of axioms to be satisfied by a structure in order that it be called a vector space; for $\mathbb{R}^n$ these are all pretty obvious. Start with a (nonempty) collection V of objects, which we can add to each other, and multiply by (real) scalars. Require that V be closed under these operations:

$$\begin{aligned} \mathbf{x}, \mathbf{y} \in V &\Rightarrow \mathbf{x} + \mathbf{y} \in V, \\ \mathbf{x} \in V, \alpha \in \mathbb{R} &\Rightarrow \alpha\mathbf{x} \in V. \end{aligned}$$

  (Is it obvious that $\mathbb{R}^n$ possesses these properties?) We say that V is a vector space ('over $\mathbb{R}$') if as well it satisfies

  1. Associativity: For all $\mathbf{x}, \mathbf{y}, \mathbf{z} \in V$, we have $\mathbf{x} + (\mathbf{y} + \mathbf{z}) = (\mathbf{x} + \mathbf{y}) + \mathbf{z}$.

  2. Commutativity: For all $\mathbf{x}, \mathbf{y} \in V$, we have $\mathbf{x} + \mathbf{y} = \mathbf{y} + \mathbf{x}$.

  3. Identity element: There is $\mathbf{0} \in V$ such that, for all $\mathbf{x} \in V$, we have $\mathbf{x} + \mathbf{0} = \mathbf{x}$.

  4. Inverse elements: For all $\mathbf{x} \in V$, there exists an element $-\mathbf{x} \in V$, called the additive inverse of $\mathbf{x}$, such that $\mathbf{x} + (-\mathbf{x}) = \mathbf{0}$.

5. Distributivity for scalar multiplication: For all $\mathbf{x}, \mathbf{y} \in V$ and $\alpha \in \mathbb{R}$ we have $\alpha(\mathbf{x} + \mathbf{y}) = \alpha\mathbf{x} + \alpha\mathbf{y}$.

6. Distributivity for scalar addition: For all $\mathbf{x} \in V$ and $\alpha, \beta \in \mathbb{R}$ we have $(\alpha + \beta)\mathbf{x} = \alpha\mathbf{x} + \beta\mathbf{x}$.

7. For all $\mathbf{x} \in V$ and $\alpha, \beta \in \mathbb{R}$ we have $(\alpha\beta)\mathbf{x} = \alpha(\beta\mathbf{x})$.

8. Scalar multiplication has an identity: $1\mathbf{x} = \mathbf{x}$.

- All properties of vector spaces follow from these axioms. For instance, the additive inverse is unique (proof: ...). Similarly, the identity element $\mathbf{0}$ is unique (lab question).

- Some other common examples of vector spaces:
  - $V =$ the set of all $n \times p$ matrices with real elements (sometimes written $\mathbb{R}^{n \times p}$). How are addition and scalar multiplication defined here?

  - V = the set of all continuous functions f on $\mathbb{R}$ (written $C(\mathbb{R})$). How are addition and scalar multiplication defined here? What theorem would one invoke to verify the closure properties? (Have you seen the proofs?)

- A nonempty subset U of V which is itself closed under addition and scalar multiplication is a vector space in its own right, called a *vector subspace* of V. (The proof consists of showing that the axioms hold in U if they hold in V and if U has these two closure properties.) Similarly $W \subset U$ closed under addition and scalar multiplication is a subspace of U.

- Definitions:
  (i) Elements $\mathbf{v}_1, ..., \mathbf{v}_m$ of V form a *spanning set* if every $\mathbf{v} \in V$ is a linear combination of them.
  (ii) Elements $\mathbf{v}_1, ..., \mathbf{v}_m$ of V are (linearly) *independent* if

$$\sum \alpha_i \mathbf{v}_i = \mathbf{0} \Rightarrow \text{all } \alpha_i = 0,$$

i.e. there is only one way in which $\mathbf{0}$ can be represented as a linear combination of them. Otherwise they are *dependent* (equivalently, at least one is a linear combination of the others).
(iii) A spanning set whose elements are independent is a *basis* of V. Thus if $\{\mathbf{v}_1, ..., \mathbf{v}_m\}$ is a basis, any $\mathbf{v} \in V$ is *uniquely* (why?) representable as a linear combination of these basis elements.

- Fact 1: Every vector space has a basis (which may be infinite). (This will not be proven here; the statement is equivalent to the Axiom of Choice in Set Theory.) No proper subset of a basis can span the entire space (why not?).

- Fact 2: If V has a basis of (finite) size r, then any $s > r$ elements of V are dependent.

  * This is obvious (is it?) if these s elements include the basis; the proof is a bit lengthy otherwise. It involves rearranging things so that one is back in the situation in which the s elements do include the basis.

  * Definition: The *dimension* of V is the unique size of a basis. Uniqueness is a consequence of Fact 2.

  * Another consequence: If $\dim(V) = r$, then any r independent vectors in V form a basis. (If not, then one can augment with elements not spanned to get $> r$ independent vectors. This contradicts Fact 2.)

- **Example**: Let $V = R^3$, and define

$$U = \left\{ \mathbf{u} = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} \in V \mid u_1 + u_2 = 0 \right\}.$$

  – Is U a subspace of V?

  – What is $\dim(U)$?

  – What is a basis for U? What is another?

  – What can one say about the matrix whose columns are the basis vectors of U?

4. Column spaces, dimension, rank

- From now on we work only with the vector space $\mathbb{R}^n$, and aim to relate matrices to (subspaces of) this vector space.

- Let V be a vector subspace of $\mathbb{R}^n$. Suppose that one forms a matrix $\mathbf{X}$ by choosing the basis elements of V to be the columns of $\mathbf{X}$. Then the interpretation of 'spanning' and 'independence' in V are, in terms of $\mathbf{X}$,

  spanning: $\mathbf{Xc} = \mathbf{y}$ is solvable (in $\mathbf{c}$) for any $\mathbf{y} \in V$;

  independence: $\mathbf{y} = \mathbf{0}$ in the above $\Rightarrow \mathbf{c} = \mathbf{0}$.

  If instead we begin with a matrix $\mathbf{X}$, then the set of all linear combinations of the columns of $\mathbf{X}$ is a vector space (why? what needs to be shown in order to verify this?), called the *column space* ($col(\mathbf{X})$), whose dimension is called the rank of $\mathbf{X}$. The independent columns of $\mathbf{X}$ form a basis for $col(\mathbf{X})$.
  Results about matrix ranks:

1) $rk(\mathbf{AB}) \leq rk(\mathbf{A})$: Since $col(\mathbf{AB}) \subseteq col(\mathbf{A})$ (why?),

   $$rk(\mathbf{AB}) = \text{dim}(col(\mathbf{AB})) \leq \text{dim}(col(\mathbf{A})) = rk(\mathbf{A}).$$

   (The inequality here is a lab problem, and follows from Fact 2 from the previous class.)

2) The rank of a matrix is at least as large as that of any of its submatrices (you should formulate and prove this).

3) Used often: $rk(\mathbf{A}'\mathbf{A}) = rk(\mathbf{A})$. The proof is being omitted, but we will look at several consequences of this very important result.

4) By 3), then 1), $rk(\mathbf{A}) = rk(\mathbf{A}'\mathbf{A}) \leq rk(\mathbf{A}')$; replacing $\mathbf{A}$ by $\mathbf{A}'$ gives $rk(\mathbf{A}') \leq rk(\mathbf{A})$ and so

   $$rk(\mathbf{A}') = rk(\mathbf{A});$$

   i.e. row rank = column rank= # of independent rows or columns. Thus, from now on, 'rank' can mean either row rank or column rank.

5) $rk(\mathbf{AB}) \leq \min(rk(\mathbf{A}), rk(\mathbf{B}))$.
Proof: That $rk(\mathbf{AB}) \leq rk(\mathbf{A})$ has been shown. Using 4) and 1),

$$rk(\mathbf{AB}) = rk(\mathbf{B}'\mathbf{A}') \leq rk\left(\mathbf{B}'\right) = rk(\mathbf{B}).$$

- A square, full rank matrix $\mathbf{A}_{n\times n}$ has an inverse, i.e. a matrix $\mathbf{B}$ such that $\mathbf{AB} = \mathbf{BA} = \mathbf{I}_n$. We write $\mathbf{B} = \mathbf{A}^{-1}$.
Proof: If $\mathbf{A}_{n\times n}$ has full rank then its n columns are independent, hence form a basis of $\mathbb{R}^n$ (why?). Thus they span: the equations

$$\mathbf{A}\left[\mathbf{b}_1 \cdots \mathbf{b}_n\right] = \left[\mathbf{e}_1 \cdots \mathbf{e}_n\right] = \mathbf{I}_n$$

are all solvable ($\mathbf{e}_1 = ...$). We write $\left[\mathbf{b}_1 \cdots \mathbf{b}_n\right] = \mathbf{B}$, then $\mathbf{AB} = \mathbf{I}_n$ and so $\mathbf{A}$ has a *right inverse*, namely $\mathbf{B}$. The matrix $\mathbf{B}$ is square, full rank (why?) and so it also has an inverse on the right: there is $\mathbf{C}_{n\times n}$ with $\mathbf{BC} = \mathbf{I}_n$. Now show (how?) that $\mathbf{C} = \mathbf{A}$; thus $\mathbf{AB} = \mathbf{BA} = \mathbf{I}_n$ and so $\mathbf{B} = \mathbf{A}^{-1}$. □

- A square matrix has full rank iff it has a non-zero determinant.

  - The determinant $|\mathbf{A}|$ is a particular sum of products of the elements of $\mathbf{A}_{n\times n}$. Each product contains n factors; there is one from each row and one from each column. It is a measure of the 'size' of the matrix, in a geometrical sense.

  - Some details are in the text. In Math 225 you might have learned various ways to compute these – cofactors, adjoints, etc.; these are in turn used to establish the statement above. We will later study some methods which are appropriate when the matrix has special structure. In the unstructured case the computations are typically done numerically.

  - Some special cases which you should remember from the 'cofactor' expansions: 2×2, triangular.

- A consequence of the preceding is that if $\mathbf{X}_{n\times p}$ has independent columns, so rank p, then $\mathbf{X}'\mathbf{X}$ is invertible. In a regression framework this can be interpreted in terms of information duplicated by dependent columns.

5. Orthogonality; projections

- Here we introduce a matrix – the 'hat matrix' – of special importance in regression. It will also be used to motivate some of what follows. Consider a regression model $\mathbf{y} = \mathbf{X}\beta + \varepsilon$ with $\mathbf{X}_{n\times p}$ of full rank p. We will later show that the LSEs are

$$\hat{\beta} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y},$$

so that the estimate of $E[\mathbf{y}] = \mathbf{X}\beta$ is $\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \mathbf{H}\mathbf{y}$, where

$$\mathbf{H}_{n\times n} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'$$

is the 'hat' matrix – it 'places the hat on $\mathbf{y}$'. Properties:

$$\begin{aligned}
\mathbf{H} &= \mathbf{H}' = \mathbf{H}^2 \text{ ('idempotent')} \\
\mathbf{H}\mathbf{X} &= \mathbf{X} \\
(\mathbf{I}-\mathbf{H})\mathbf{X} &= \mathbf{0} \\
(\mathbf{I}-\mathbf{H})^2 &= (\mathbf{I}-\mathbf{H}) \\
\mathbf{H}(\mathbf{I}-\mathbf{H}) &= \mathbf{0}.
\end{aligned}$$

  - What is $\mathbf{H}$ if $p = 1$ and $\mathbf{X}$ is merely a column of n ones? Relate this to Lab 1, Q6.

- The angle θ between non-zero vectors $\mathbf{x}, \mathbf{y}$ is defined by

$$\cos\theta = \frac{\mathbf{x}'\mathbf{y}}{\|\mathbf{x}\| \, \|\mathbf{y}\|}.$$

(What is this in $\mathbb{R}^2$?) That such an angle exists is equivalent to the statement that $|\mathbf{x}'\mathbf{y}| \leq \|\mathbf{x}\| \, \|\mathbf{y}\|$. This in turn is a version of the famous Cauchy-Schwarz Inequality, to be studied later.
Proof of this version: For any real number λ,

$$0 \leq \ \|\mathbf{x} + \lambda\mathbf{y}\|^2 = \|\mathbf{y}\|^2 \lambda^2 + 2\mathbf{x}'\mathbf{y}\lambda + \|\mathbf{x}\|^2 ,$$

so that there is at most one real zero. Thus '$B^2 - 4AC$'≤ 0, i.e.

$$4\left(\left(\mathbf{x}'\mathbf{y}\right)^2 - \|\mathbf{x}\|^2 \|\mathbf{y}\|^2\right) \leq 0.$$

□

If $|\mathbf{x}'\mathbf{y}| = \|\mathbf{x}\| \, \|\mathbf{y}\|$ then $\mathbf{x}'\mathbf{y} = \pm \|\mathbf{x}\| \, \|\mathbf{y}\|$ and $\|\mathbf{x} + \lambda_0\mathbf{y}\|^2 = 0$ for $\lambda_0 = \pm \|\mathbf{x}\| / \|\mathbf{y}\|$). In particular $\mathbf{x} + \lambda_0\mathbf{y} = \mathbf{0}$ (why?), hence $\mathbf{x} = -\lambda_0\mathbf{y}$, i.e.

we have equality in the C-S inequality if the two vectors are multiples of each other. The converse (what is it?) holds as well – verify this on your own.

- Two vectors are *orthogonal* if the angle between them $= \pm\pi/2$, equivalently if their scalar product $= 0$. We write $\mathbf{x} \perp \mathbf{y}$.

  - **Example**: If $\mathbf{z}$ is any $n \times 1$ vector, and $\mathbf{H}$ is a hat matrix, then

$$\mathbf{z} = \mathbf{H}\mathbf{z} + (\mathbf{I} - \mathbf{H})\mathbf{z} = \mathbf{z}_1 + \mathbf{z}_2,$$

    say, where $\mathbf{z}_1 \perp \mathbf{z}_2$. The first is in col$(\mathbf{X})$ (why?) and the second is in the space of vectors orthogonal to every vector in col$(\mathbf{X})$. We write $\mathbf{z}_2 \in$ col$(\mathbf{X})^{\perp}$. You should verify that this is a vector space (i.e. is closed under addition and scalar multiplication).

- Suppose $\mathbf{X}_{n\times p}$ has independent columns, so $col(\mathbf{X})$ has dimension p. Recall that the orthogonal complement to this space is

$$col(\mathbf{X})^{\perp} = \left\{ \mathbf{y} \,\middle|\, \mathbf{z}'\mathbf{y} = \mathbf{0} \text{ for every } \mathbf{z} \in col(\mathbf{X}) \right\}.$$

More simply (how? what is being asserted here?)

$$col(\mathbf{X})^{\perp} = \left\{ \mathbf{y} \,\middle|\, \mathbf{X}'\mathbf{y} = \mathbf{0} \right\}.$$

Then $col(\mathbf{X})^{\perp} = col\,(\mathbf{I} - \mathbf{H})$:

$$\begin{aligned} \mathbf{y} &\in col(\mathbf{X})^{\perp} \Rightarrow \mathbf{X}'\mathbf{y} = \mathbf{0} \Rightarrow (\mathbf{I} - \mathbf{H})\,\mathbf{y} = \mathbf{y}; \\ \mathbf{y} &\in col(\mathbf{I} - \mathbf{H}) \Rightarrow \mathbf{H}\mathbf{y} = \mathbf{0} \Rightarrow \mathbf{X}'\mathbf{y} = \mathbf{0}. \end{aligned}$$

Thus dim $\left(col(\mathbf{X})^{\perp}\right) = rk\,(\mathbf{I} - \mathbf{H})$.

6. Gram-Schmidt method; QR-decomposition

- The trace of a square matrix is the sum of its diagonal elements. A useful identity is (how?)

$$tr(\mathbf{AB}) = tr(\mathbf{BA}).$$

Thus products within traces can be rearranged cyclically:

$$\begin{aligned} tr(\mathbf{ABC}) = tr(\mathbf{CAB}) &= tr(\mathbf{BCA}) \\ \text{but not necessarily} &= tr(\mathbf{ACB}). \end{aligned}$$

It will be shown that for an idempotent matrix, $rk = tr$. A consequence is that

$$\begin{aligned} \dim\left(col(\mathbf{X})^{\perp}\right) &= rk(\mathbf{I}-\mathbf{H}) \\ &= tr(\mathbf{I}-\mathbf{H}) \\ &= n - tr(\mathbf{H}) \\ &= n - tr\left(\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\right) \\ &= n - tr\left(\mathbf{X}'\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\right) \\ &= n - p. \end{aligned}$$

Similarly $col(\mathbf{X}) = col(\mathbf{H})$, $rk(\mathbf{H}) = p$.

- A matrix $\mathbf{Q}_{n\times n}$ is *orthogonal* if the columns are mutually orthogonal, and have unit norm. Equivalently (why?)

$$\mathbf{Q}\mathbf{Q}' = \mathbf{Q}'\mathbf{Q} = \mathbf{I}_n.$$

  If $\mathbf{Q}$ is orthogonal then $\|\mathbf{Q}\mathbf{y}\| = \|\mathbf{y}\|$ for any $n \times 1$ vector $\mathbf{y}$ – 'norms are preserved'. Similarly, angles between vectors are also preserved (why?). Geometrically, an orthogonal transformation is a 'rigid motion' – it corresponds to a rotation and/or an interchange of two or more axes. Rotation through an angle θ in the plane:

$$\mathbf{Q} = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}.$$

  Interchange of axes in the plane:

$$\mathbf{Q} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

- **Gram-Schmidt Theorem**: Every m-dimensional vector space V has an orthogonal basis.

  Proof: Start with any basis $\mathbf{v}_1, ..., \mathbf{v}_m$. Normalize $\mathbf{v}_1$ to get a unit vector (i.e. a vector with unit norm) $\mathbf{q}_1 \in V$; in general suppose that mutually orthogonal unit vectors $\mathbf{q}_1, ...\mathbf{q}_j$ have been constructed, with $\mathbf{q}_i$ a linear combination of $\mathbf{v}_1, ...\mathbf{v}_i$. Define

$$\begin{aligned} \mathbf{H}_j &= \sum_{i=1}^{j} \mathbf{q}_i \mathbf{q}_i', \\ \mathbf{q}_{j+1} &= \frac{\left(\mathbf{I} - \mathbf{H}_j\right) \mathbf{v}_{j+1}}{\left\| \left(\mathbf{I} - \mathbf{H}_j\right) \mathbf{v}_{j+1} \right\|}. \end{aligned}$$

  For instance, $\mathbf{q}_2 = ...$.

  Then: (i) the denominator is non-zero (why?); (ii) $\mathbf{q}_{j+1}$ is a linear combination of $\mathbf{v}_1, ...\mathbf{v}_{j+1}$; (iii) $\mathbf{q}_{j+1} \perp \mathbf{q}_1, ...\mathbf{q}_j$ (why?). Continuing this process results in mutually orthogonal unit vectors $\mathbf{q}_1, ...\mathbf{q}_m$. Since these are orthogonal they are independent (problem on Lab 2) and so form a basis of V.

  - There is a nice geometric interpretation. The matrix $\mathbf{H}_j$ is idempotent (in fact it is the 'hat' matrix arising from the $n \times j$ matrix with columns $\mathbf{q}_1, ...\mathbf{q}_j$), and $\mathbf{H}_j \mathbf{v}_{j+1}$ is the 'projection of $\mathbf{v}_{j+1}$ onto the space spanned by $\left\{\mathbf{q}_1, ...\mathbf{q}_j\right\}$'. Thus we say that $\mathbf{q}_{j+1}$ is formed by 'subtracting from $\mathbf{v}_{j+1}$ its projection onto the space spanned by $\left\{\mathbf{q}_1, ...\mathbf{q}_j\right\}$', so as to make what is left orthogonal to this space (and then normalizing).

- **QR-decomposition**. In the previous construction, at each stage, $\mathbf{q}_j$ was obtained as a linear combination of $\mathbf{v}_1, ...\mathbf{v}_j$. Thus if $\mathbf{V}_{n \times m}$ has these vectors as its columns, and $\mathbf{Q}_{n \times m} = (\mathbf{q}_1, ...\mathbf{q}_m)$, we have $\mathbf{V}_{n \times m} \mathbf{U}_{m \times m} = \mathbf{Q}_{n \times m}$ for $\mathbf{U}$ upper triangular. Also $\mathbf{Q}'\mathbf{Q} = \mathbf{I}_m = \mathbf{U}' \left[\mathbf{V}'\mathbf{V}\right] \mathbf{U}$ so all three have non-zero determinants and so are nonsingular. Thus

$$\mathbf{V} = \mathbf{QR}$$

  for $\mathbf{R} = \mathbf{U}^{-1}$.

- **Example**: Let $\mathbf{X}_{n\times p}$ have rank $p < n$ and consider the system

$$\mathbf{X}\mathbf{t}_{p\times 1} = \mathbf{c}_{n\times 1} \tag{6.1}$$

of n equations in the p unknowns $t_1, ..., t_p$. Assume there is a solution, i.e. that $\mathbf{c} \in col\,(\mathbf{X})$. Now

$$\mathbf{X}\mathbf{t} = \mathbf{c} \Rightarrow \mathbf{X}'\mathbf{X}\mathbf{t} = \mathbf{X}'\mathbf{c} \Rightarrow \mathbf{t} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{c};$$

(check that this is a solution) and

$$\begin{aligned}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}' &= \left(\mathbf{R}'\mathbf{Q}'\mathbf{Q}\mathbf{R}\right)^{-1}\mathbf{R}'\mathbf{Q}' \\ &= \left(\mathbf{R}'\mathbf{R}\right)^{-1}\mathbf{R}'\mathbf{Q}' \\ &= \mathbf{R}^{-1}\mathbf{Q}'.\end{aligned}$$

So

$$\mathbf{t} = \mathbf{R}^{-1}\mathbf{Q}'\mathbf{c}.$$

For numerical work, this can be much more stable, since it can be computed without actually inverting a matrix – something which is numerically very unstable. For this, first compute $\mathbf{Q}'\mathbf{c} = \mathbf{b}$,

say. Then 'backsolve' the equations $\mathbf{Rt} = \mathbf{b}$, i.e.

$$\begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1p} \\ 0 & \ddots & \ddots & \\ & \ddots & r_{p-1,p-1} & r_{p-1,p} \\ 0 & & 0 & r_{pp} \end{pmatrix} \begin{pmatrix} t_1 \\ \vdots \\ t_{p-1} \\ t_p \end{pmatrix} = \begin{pmatrix} b_1 \\ \vdots \\ b_{p-1} \\ b_p \end{pmatrix},$$

starting with

$$r_{pp}t_p = b_p \Rightarrow t_p = \frac{b_p}{r_{pp}},$$

etc.

7. Least squares; eigenvalues and eigenvectors

- Another application of the QR decomposition is in regression. Let $\mathbf{X}_{n\times p}$ have rank p. Write $\mathbf{X} = \mathbf{Q}_1\mathbf{R}_1$, where $\mathbf{Q}_1 : n \times p$ has orthogonal columns ($\mathbf{Q}_1'\mathbf{Q}_1 = \mathbf{I}_p$), and $\mathbf{R}_1 : p \times p$ is upper triangular and non-singular (and $\mathbf{X}'\mathbf{X} = \mathbf{R}_1'\mathbf{R}_1$). Apply Gram-Schmidt once again to $col(\mathbf{X})^{\perp}$, a basis for which is the $n-p$ independent columns of $\mathbf{I} - \mathbf{H}$, to obtain $\mathbf{Q}_2 : n \times (n-p)$ whose columns are orthogonal to each other ($\mathbf{Q}_2'\mathbf{Q}_2 = \mathbf{I}_{n-p}$) and (why?) to those of $\mathbf{Q}_1$. Then $\mathbf{Q} = (\mathbf{Q}_1 \vdots \mathbf{Q}_2)$ has orthogonal columns and is square, hence is an orthogonal matrix ($\mathbf{Q}\mathbf{Q}' = \mathbf{Q}'\mathbf{Q} = \mathbf{I}_n$). We have

$$\mathbf{X} = (\mathbf{Q}_1 \vdots \mathbf{Q}_2) \begin{pmatrix} \mathbf{R}_1 \\ \mathbf{0} \end{pmatrix} \stackrel{def}{=} \mathbf{Q}\mathbf{R},$$

and you should verify that:

$$\begin{aligned} \mathbf{X}'\mathbf{X} &= \mathbf{R}_1'\mathbf{R}_1 = \mathbf{R}'\mathbf{R}, \\ \left(\mathbf{X}'\mathbf{X}\right)^{-1} &= \mathbf{R}_1^{-1}\mathbf{R}_1^{-1\prime}, \\ \mathbf{H} &= \mathbf{Q}_1\mathbf{Q}_1', \\ \mathbf{I} - \mathbf{H} &= \mathbf{Q}_2\mathbf{Q}_2'. \end{aligned}$$

- Least squares estimation in terms of hat matrix decomposition of norm of residuals: Note that $\mathbf{x} \perp \mathbf{y} \Rightarrow \|\mathbf{x}+\mathbf{y}\|^2 = \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2$. We aim to minimize

$$\begin{aligned}\left\|\mathbf{y}-\mathbf{X}\hat{\beta}\right\|^2 &= \left\|\mathbf{H}\left(\mathbf{y}-\mathbf{X}\hat{\beta}\right)\right\|^2 + \left\|(\mathbf{I}-\mathbf{H})\left(\mathbf{y}-\mathbf{X}\hat{\beta}\right)\right\|^2 \\ &= \left\|\mathbf{H}\mathbf{y}-\mathbf{X}\hat{\beta}\right\|^2 + \|(\mathbf{I}-\mathbf{H})\,\mathbf{y}\|^2 \\ &\geq \|(\mathbf{I}-\mathbf{H})\,\mathbf{y}\|^2,\end{aligned}$$

with equality iff $\mathbf{H}\mathbf{y} = \mathbf{X}\hat{\beta}$ iff ('if and only if')

$$\hat{\beta} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}$$

(how?), the LS estimator. The *fitted values* are

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \mathbf{H}\mathbf{y},$$

and are orthogonal to the *residuals*

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = (\mathbf{I}-\mathbf{H})\,\mathbf{y}.$$

We say that $\mathbf{H}$ and $\mathbf{I}-\mathbf{H}$ *project* the data $(\mathbf{y})$ onto the estimation space and error space, respectively, and that these spaces are orthogonal.

- In terms of the QR-decomposition: we have that $\hat{\beta} = \mathbf{R}_1^{-1}\mathbf{R}_1^{-1\prime}\mathbf{R}_1'\mathbf{Q}_1'\mathbf{y}$; i.e.

$$\mathbf{R}_1\hat{\beta} = \mathbf{Q}_1'\mathbf{y}.$$

  Backsolve this triangular system – no matrix inversions.

- The usual estimate of the variance σ_{ε}^2 of the random errors ε is

$$\begin{aligned}\hat{\sigma}^2 &= \frac{\text{SS of residuals}}{n-p} = \frac{\|\mathbf{e}\|^2}{n-p} = \frac{\|(\mathbf{I}-\mathbf{H})\,\mathbf{y}\|^2}{n-p} \\ &= \frac{\|(\mathbf{I}-\mathbf{H})\,\boldsymbol{\varepsilon}\|^2}{n-p} \overset{how?}{=} \frac{\boldsymbol{\varepsilon}'\,(\mathbf{I}-\mathbf{H})\,\boldsymbol{\varepsilon}}{n-p},\end{aligned}$$

  the *mean squared error*. Apply Lab 2 #2:

$$E\left[\boldsymbol{\varepsilon}'\,(\mathbf{I}-\mathbf{H})\,\boldsymbol{\varepsilon}\right] = tr\left\{(\mathbf{I}-\mathbf{H})\,\text{cov}\,[\boldsymbol{\varepsilon}]\right\} + \mu_{\boldsymbol{\varepsilon}}'\,(\mathbf{I}-\mathbf{H})\,\mu_{\boldsymbol{\varepsilon}}.$$

  Since (Lab 1 #5) $\text{cov}[\boldsymbol{\varepsilon}] = \sigma_{\varepsilon}^2\mathbf{I}_n$ and $\mu_{\boldsymbol{\varepsilon}} = \mathbf{0}$, we get

$$E\left[\hat{\sigma}^2\right] = \frac{E\left[\boldsymbol{\varepsilon}'\,(\mathbf{I}-\mathbf{H})\,\boldsymbol{\varepsilon}\right]}{n-p} = \sigma_{\varepsilon}^2\frac{tr\,(\mathbf{I}-\mathbf{H})}{n-p} = \sigma_{\varepsilon}^2;$$

  thus $\hat{\sigma}^2$ is *unbiased*.

- Let $\mathbf{M}_{n\times n}$ be any square matrix. For a variable λ the determinant $|\mathbf{M} - \lambda\mathbf{I}_n|$ is a polynomial in λ of degree n, called the *characteristic polynomial*. The equation

$$|\mathbf{M} - \lambda\mathbf{I}_n| = 0 \tag{7.1}$$

  is the *characteristic equation*. The Fundamental Theorem of Algebra states that there are then n (real or complex) roots of this equation. Any such root is called an *eigenvalue* (or 'characteristic root') of $\mathbf{M}$. If λ is an eigenvalue then $\mathbf{M} - \lambda\mathbf{I}_n$ is singular, so the columns are dependent:

$$(\mathbf{M} - \lambda\mathbf{I}_n)\,\mathbf{v} = 0$$

  for some non-zero vector $\mathbf{v}$, called the *eigenvector* corresponding to, or belonging to, λ. Thus

$$\mathbf{M}\mathbf{v} = \lambda\mathbf{v}. \tag{7.2}$$

  - You should work through the details for 2×2 matrices; note trace and determinant are simple functions of the eigenvalues.

– It is true in general that the trace of a matrix $\mathbf{M}$ is the sum of its eigenvalues, and the determinant is the product. You will be able to prove these facts very simply in the special case that $\mathbf{M}$ is *symmetric*.

8. Spectral decomposition

- Now suppose that $\mathbf{M}$ is *symmetric*. Then **the eigenvalues are real**.

  – We won't cover the proof of this in class, but here it is if you're interested: Define an operation $\mathbf{A}^*$ by taking a transpose and a complex conjugate:

  $$(\mathbf{A}^*)_{\alpha\beta} = \bar{a}_{\beta\alpha}.$$

  Note that $(\mathbf{AB})^* = \mathbf{B}^*\mathbf{A}^*$ and that $\mathbf{v}^*\mathbf{v} = \sum |v_\alpha|^2$ is real. For a real symmetric matrix $\mathbf{M}$ we have (why?) $\mathbf{M}^* = \mathbf{M}$. Thus in (7.2),

  $$\mathbf{v}^*\mathbf{M}\mathbf{v} = \lambda\mathbf{v}^*\mathbf{v};$$

  taking the conjugate transpose of each side gives

  $$\mathbf{v}^*\mathbf{M}\mathbf{v} = \bar{\lambda}\mathbf{v}^*\mathbf{v}.$$

  Thus $\left(\lambda - \bar{\lambda}\right)\mathbf{v}^*\mathbf{v} = 0$; so that (why?) λ is real.

- We can, and usually will, **assume that an eigenvector has unit norm** (how?).

- **Eigenvectors corresponding to distinct eigenvalues are orthogonal**. Reason: If $\mathbf{M}\mathbf{v}_i = \lambda_i \mathbf{v}_i$ for $i = 1, 2$ and $\lambda_1 \neq \lambda_2$ then

$$\begin{aligned} \mathbf{v}_1' \mathbf{M} \mathbf{v}_2 &= \mathbf{v}_1' \left(\mathbf{M}\mathbf{v}_2\right) = \lambda_2 \mathbf{v}_1' \mathbf{v}_2 \\ \text{and} &= \left(\mathbf{v}_1' \mathbf{M}\right) \mathbf{v}_2 = \lambda_1 \mathbf{v}_1' \mathbf{v}_2; \end{aligned}$$

  thus $(\lambda_1 - \lambda_2)\, \mathbf{v}_1' \mathbf{v}_2 = 0$ and so $\mathbf{v}_1' \mathbf{v}_2 = 0$.

- If λ is a multiple root of the characteristic equation, with multiplicity r, then the set of corresponding eigenvectors is a vector space (why?). It can be shown that, **for a real symmetric matrix, this vector space has dimension** r (this is the only part of this development that requires some work, and is being omitted here). We can then apply Gram-Schmidt to the r eigenvectors which form a basis of this space, to conclude that **there are r orthogonal eigenvectors corresponding to an eigenvalue λ with multiplicity r.**

- We have shown that, if $\mathbf{M}_{n\times n}$ is symmetric, with eigenvalues $\lambda_1, ..., \lambda_n$ obtained as solutions to the characteristic equation (7.1), then to each λ_i there corresponds an eigenvector $\mathbf{v}_i$ satisfying (7.2). These eigenvectors can be chosen so as to be mutually orthogonal, and to have unit norm.

- **Spectral Decomposition** of real, symmetric matrices: Let $\mathbf{M}_{n\times n}$ be real and symmetric, with eigenvalues $\lambda_1, ..., \lambda_n$ and corresponding orthogonal eigenvectors $\mathbf{v}_1, ..., \mathbf{v}_n$. Put

$$\mathbf{V}_{n\times n} = (\mathbf{v}_1 \cdots \ \mathbf{v}_n),$$

an orthogonal matrix. Let $\mathbf{D}_{\boldsymbol{\lambda}}$ be the diagonal matrix with $\lambda_1, ..., \lambda_n$ on the diagonal. Since

$$\begin{aligned}
\mathbf{MV} &= (\lambda_1\mathbf{v}_1 \cdots \ \lambda_n\mathbf{v}_n) \\
&= (\mathbf{v}_1 \cdots \ \mathbf{v}_n) \begin{pmatrix} \lambda_1 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \lambda_n \end{pmatrix} \\
&= \mathbf{VD}_{\boldsymbol{\lambda}},
\end{aligned}$$

we have

$$\mathbf{M} = \mathbf{VD}_{\boldsymbol{\lambda}}\mathbf{V}'. \tag{8.1}$$

We say that 'a real symmetric matrix is orthogonally similar to a diagonal matrix'. In a sense that will become clear, the importance of this result is that a real, symmetric matrix is 'almost' diagonal. Thus when solving problems concerning real symmetric matrices it is very often useful to solve them first for diagonal matrices. This is frequently quite simple, and then extends to the general case via (8.1).

- In the construction above we could have assumed, and sometimes will assume, that the eigenvalues were ordered before being labelled: $\lambda_1 \geq ... \geq \lambda_n$.

- Use standard properties of the trace and determinant to express $tr\,(\mathrm{M})$ and $\det(\mathrm{M})$ as $tr\,(\mathrm{D}_{\boldsymbol{\lambda}})$ and $\det(\mathrm{D}_{\boldsymbol{\lambda}})$, hence as the sum and product of the eigenvalues. These are but two examples of the adage that 'a symmetric matrix is almost diagonal'; more to come.

- **Bounds on eigenvalues**. We have

$$\begin{aligned}\max_{\|\mathbf{x}\|=1} \mathbf{x}'\mathbf{M}\mathbf{x} &= \max_{\|\mathbf{x}\|=1} \mathbf{x}'\mathbf{V}\mathbf{D}_{\boldsymbol{\lambda}}\mathbf{V}'\mathbf{x} \\ &= \max_{\|\mathbf{y}\|=1} \mathbf{y}'\mathbf{D}_{\boldsymbol{\lambda}}\mathbf{y} \text{ (why?)} \\ &= \max\left\{\sum_{i=1}^{n} \lambda_i y_i^2 \mid \sum_{i=1}^{n} y_i^2 = 1\right\}.\end{aligned}$$

  This is a weighted average $\sum_{i=1}^{n} \lambda_i w_i$ – the weights $w_i = y_i^2$ sum to 1 – of the $\{\lambda_i\}$. How would you choose the weights in order to get the maximum average? .. Then the maximizing $\mathbf{y} =$ (what?); hence the maximizing $\mathbf{x}$ is the corresponding eigenvector. An analogous result holds for $\min_{\|\mathbf{x}\|=1} \mathbf{x}'\mathbf{M}\mathbf{x}$. (You should write it out and prove it.)

9. Spectral decomposition: examples

- From the last example, we see that for any matrix $\mathbf{M}$, and vector $\mathbf{x}$,

$$\|\mathbf{x}\|^2 \, ch_{\mathsf{min}}(\mathbf{M}) \leq \mathbf{x}'\mathbf{M}\mathbf{x} \leq \|\mathbf{x}\|^2 \, ch_{\mathsf{max}}(\mathbf{M}),$$

  where $ch_{\mathsf{min}}(\mathbf{M})$ and $ch_{\mathsf{max}}(\mathbf{M})$ denote the smallest and largest eigenvalues ("characteristic roots") of $\mathbf{M}$. These bounds are attained by the corresponding eigenvectors.

- **Positive definite matrices**. If a symmetric matrix $\mathbf{M}$ is such that $\mathbf{x}'\mathbf{M}\mathbf{x} \geq 0$ for all $\mathbf{x}$, we say that $\mathbf{M}$ is *positive semidefinite* (p.s.d.) or *nonnegative definite* (n.n.d.). We write $\mathbf{M} \geq \mathbf{0}$. The preceding discussion shows (how?) that $\mathbf{M} \geq \mathbf{0}$ iff all eigenvalues are non-negative.
  If $\mathbf{x}'\mathbf{M}\mathbf{x} > 0$ for all $\mathbf{x} \neq \mathbf{0}$, we say that $\mathbf{M}$ is *positive definite* (p.d.). We write $\mathbf{M} > \mathbf{0}$. Equivalently, all eigenvalues are positive.

– Geometric interpretation: If $\mathbf{M} > \mathbf{0}$ then the set

$$\left\{\mathbf{x} \,|\, \mathbf{x}'\mathbf{M}^{-1}\mathbf{x} = c^2\right\}$$

is transformed, via the (orthogonal) transformation $\mathbf{y} = \mathbf{V}'\mathbf{x}$ (where $\mathbf{M} = \mathbf{V}\mathbf{D}_{\boldsymbol{\lambda}}\mathbf{V}'$), into the set

$$\left\{\mathbf{y} \,|\, \sum_{i=1}^{n} \frac{y_i^2}{\lambda_i} = c^2\right\}.$$

This is the ellipsoid in $\mathbb{R}^n$ with semi-axes of lengths $c\sqrt{\lambda_i}$ along the coordinate axes (and volume $\propto \sqrt{|\mathbf{M}|}$). Thus (why?) the original set, obtained from the second via the transformation $\mathbf{x} = \mathbf{V}\mathbf{y}$, is an ellipsoid as well, whose semi-axes have the same lengths but are now in the directions of the eigenvectors of $\mathbf{M}$.

- **Matrix square roots**. Can we define a notion of the square root of a (p.s.d.) matrix? Start by thinking of a diagonal matrix, in which case the method is obvious – how? Now extend to the

general case. If $\mathbf{M} \geq \mathbf{0}$ we write $\mathbf{M} = \mathbf{V}\mathbf{D}_{\boldsymbol{\lambda}}\mathbf{V}'$, where $\mathbf{V}$ is orthogonal and $\mathbf{D}_{\boldsymbol{\lambda}}$ has a non-negative diagonal. We define a symmetric, p.s.d. square root of $\mathbf{M}$ by

$$\mathbf{M}^{1/2} = \mathbf{V}\mathbf{D}_{\boldsymbol{\lambda}}^{1/2}\mathbf{V}'.$$

  - There are other roots, for instance $\mathbf{P} = \mathbf{V}\mathbf{D}_{\boldsymbol{\lambda}}^{1/2}\mathbf{W}$ for any orthogonal $\mathbf{W}$ (then $\mathbf{P}\mathbf{P}' = \mathbf{M}$) but we will generally mean the one above.

- The rank of a symmetric matrix equals the number of non-zero eigenvalues. **Reason**: Write $\mathbf{M} = \mathbf{V}\mathbf{D}_{\boldsymbol{\lambda}}\mathbf{V}'$, then the rank of $\mathbf{M}$ equals the rank of $\mathbf{D}_{\boldsymbol{\lambda}}$ (why?), and the latter is clearly (is it?) the number of non-zero diagonal elements.

  - Note also that if $\mathbf{M} = \mathbf{V}\mathbf{D}\mathbf{V}'$ is the spectral decomposition then $\mathbf{M}$ and $\mathbf{D}$ have the same eigenvalues, namely the diagonal elements of $\mathbf{D}$. This is because the characteristic polynomials are the same:

$$|\mathbf{M} - \lambda\mathbf{I}| = \left|\mathbf{V}\left(\mathbf{D} - \lambda\mathbf{I}\right)\mathbf{V}'\right| = |\mathbf{D} - \lambda\mathbf{I}|.$$

- If $\mathbf{H}$ is idempotent then (i) all eigenvalues are 0 or 1, and (ii) rank $=$ trace. **Reason**: (i) It is clearly true (how?) for diagonal idempotents. But if $\mathbf{H}$ is idempotent then $\mathbf{H} = \mathbf{V}\mathbf{D}_{\boldsymbol{\lambda}}\mathbf{V}'$ for $\mathbf{V}$ orthogonal and $\mathbf{D}_{\boldsymbol{\lambda}}$ idempotent (why?), and $\mathbf{H}$ has the same eigenvalues as $\mathbf{D}_{\boldsymbol{\lambda}}$. (ii) $rk(\mathbf{H}) = rk(\mathbf{D}_{\boldsymbol{\lambda}}) = tr(\mathbf{D}_{\boldsymbol{\lambda}}) = tr(\mathbf{H})$ (how are these steps justified?).

  - Another interesting property: in the above we can partition $\mathbf{D}_{\boldsymbol{\lambda}}$, and compatibly partition $\mathbf{V}$, as

$$\mathbf{D}_{\boldsymbol{\lambda}} = \begin{pmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}, \mathbf{V} = (\mathbf{V}_1 \vdots \mathbf{V}_2),$$

    where $rk(\mathbf{H}) = r$ and $\mathbf{V}_1$ is $n \times r$. This results in the decomposition of an idempotent matrix as

$$\mathbf{H} = \mathbf{V}_1\mathbf{V}_1', \text{ where } \mathbf{V}_1'\mathbf{V}_1 = \mathbf{I}_r.$$

    (So $\mathbf{V}_1$ is what we called $\mathbf{Q}_1$ at p. 40.)

- **An application**. By the Cauchy-Schwarz Inequality,

$$\begin{aligned} \max_{\mathbf{y}} \frac{|\mathbf{x}'\mathbf{M}\mathbf{y}|}{\|\mathbf{y}\|} &= \left\|\mathbf{M}'\mathbf{x}\right\| = \sqrt{\mathbf{x}'\mathbf{M}\mathbf{M}'\mathbf{x}} \\ &\leq \|\mathbf{x}\| \sqrt{ch_{\max}\mathbf{M}\mathbf{M}'}. \end{aligned}$$

Related facts: Note that $\mathbf{M}\mathbf{M}' \geq 0$ (why?). Conversely, any p.s.d. matrix can be represented as $\mathbf{M}\mathbf{M}'$ (in many ways). In particular, if $\mathbf{S}$ is a $p \times p$ p.s.d. matrix of rank $q \leq p$, then one can find $\mathbf{M}_{p\times q}$ such that $\mathbf{S} = \mathbf{M}\mathbf{M}'$ and $\mathbf{M}'\mathbf{M}$ is the $q \times q$ diagonal matrix of the positive eigenvalues of $\mathbf{S}$. **Reason**:
Write $\mathbf{S} = \mathbf{V}\mathbf{D}\mathbf{V}'$, where

$$\mathbf{D}_{p\times p} = \begin{pmatrix} \mathbf{D}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}, \mathbf{V} = \left(\underbrace{\mathbf{V}_1}_{q} \vdots \underbrace{\mathbf{V}_2}_{p-q} \right)$$

and $\mathbf{D}_1$ is the $q\times q$ diagonal matrix containing the positive eigenvalues. Then $\mathbf{S} = \mathbf{V}_1\mathbf{D}_1\mathbf{V}_1'$ and so $\mathbf{M}_{p\times q} = \mathbf{V}_1\mathbf{D}_1^{1/2}$ has the desired properties. (Note also that this is a version of $\mathbf{S}^{1/2}$.)

10. More examples; block matrices

- **Another application**. Illustration of preceding theory: two-population classification problem. Suppose we are given lengths and widths of n prehistoric skulls, of type A or B (the 'training sample'). We know that n_1 of these, say $\mathbf{x}_1, ...\mathbf{x}_{n_1}$, are of type A, and $n_2 = n - n_1$, say $\mathbf{y}_1, ...\mathbf{y}_{n_2}$, are of type B. Now we find a new skull, with length and width the components of $\mathbf{z}$. We are to classify it as A or B. (Others applications: rock samples in geology, risk data in an actuarial analysis, etc.)

  - Reduce to univariate problem: $u_i = \boldsymbol{\alpha}'\mathbf{x}_i$, $v_i = \boldsymbol{\alpha}'\mathbf{y}_i$ for some vector $\boldsymbol{\alpha}$. Put $w = \boldsymbol{\alpha}'\mathbf{z}$ and classify new skull as A if $|w - \bar{u}| < |w - \bar{v}|$.

  - Choose $\boldsymbol{\alpha}$ for 'maximal separation': $|\bar{u} - \bar{v}|$ should be large relative to the underlying vari-

ation. Put

$$\begin{aligned} s_1^2 &= \frac{1}{n_1-1}\sum_{i=1}^{n_1}(u_i-\bar{u})^2 \\ &= \frac{1}{n_1-1}\sum\left(\boldsymbol{\alpha}'(\mathbf{x}_i-\bar{\mathbf{x}})\right)^2 \\ &= \frac{1}{n_1-1}\boldsymbol{\alpha}'\sum(\mathbf{x}_i-\bar{\mathbf{x}})(\mathbf{x}_i-\bar{\mathbf{x}})'\boldsymbol{\alpha} \\ &= \boldsymbol{\alpha}'\mathbf{S}_1\boldsymbol{\alpha} \end{aligned}$$

and similarly define s_2^2 as the variation in the other sample. Define the two-sample covariance matrix

$$\mathbf{S} = \frac{(n_1-1)\,\mathbf{S}_1 + (n_2-1)\,\mathbf{S}_2}{n-2}$$

and choose $\boldsymbol{\alpha}$ to maximize

$$\begin{aligned} &\frac{(\bar{u}-\bar{v})^2}{\left[(n_1-1)\,s_1^2+(n_2-1)\,s_2^2\right]/(n-2)} \\ = \;&\frac{\boldsymbol{\alpha}'(\bar{\mathbf{x}}-\bar{\mathbf{y}})(\bar{\mathbf{x}}-\bar{\mathbf{y}})'\boldsymbol{\alpha}}{\boldsymbol{\alpha}'\mathbf{S}\boldsymbol{\alpha}}. \qquad (*) \end{aligned}$$

– Put $\beta = \mathbf{S}^{1/2}\boldsymbol{\alpha}$, $\boldsymbol{\alpha} = \mathbf{S}^{-1/2}\beta$ so (*) is

$$\frac{\beta'\mathbf{S}^{-1/2}(\bar{\mathbf{x}}-\bar{\mathbf{y}})(\bar{\mathbf{x}}-\bar{\mathbf{y}})'\mathbf{S}^{-1/2}\beta}{\beta'\beta},$$

which is a maximum if β is the eigenvector corresponding to

$$ch_{\max}\mathbf{S}^{-1/2}\left(\bar{\mathbf{x}}-\bar{\mathbf{y}}\right)\left(\bar{\mathbf{x}}-\bar{\mathbf{y}}\right)'\mathbf{S}^{-1/2}=ch_{\max}\mathbf{a}\mathbf{a}',$$

where $\mathbf{a}=\mathbf{S}^{-1/2}\left(\bar{\mathbf{x}}-\bar{\mathbf{y}}\right)$. Note $\mathbf{a}\mathbf{a}'$ has rank 1, hence has 1 non-zero eigenvalue – necessarily the trace of $\mathbf{a}\mathbf{a}'$:

$$\lambda=tr\left(\mathbf{a}\mathbf{a}'\right)=\mathbf{a}'\mathbf{a}=\left(\bar{\mathbf{x}}-\bar{\mathbf{y}}\right)'\mathbf{S}^{-1}\left(\bar{\mathbf{x}}-\bar{\mathbf{y}}\right).$$

Now solve $\mathbf{a}\mathbf{a}'\beta=\lambda\beta$, i.e.

$$\mathbf{a}\mathbf{a}'\beta=\beta\mathbf{a}'\mathbf{a}$$

to get ($\beta=$ what? – guess at a solution); any multiple will do. Then

$$\boldsymbol{\alpha}=\mathbf{S}^{-1/2}\beta=\mathbf{S}^{-1}\left(\bar{\mathbf{x}}-\bar{\mathbf{y}}\right)$$

and we classify as A if

$$|w-\bar{u}|=\left|\boldsymbol{\alpha}'(\mathbf{z}-\bar{\mathbf{x}})\right|<\left|\boldsymbol{\alpha}'(\mathbf{z}-\bar{\mathbf{y}})\right|=|w-\bar{v}|.$$

- A very useful identity: If $\mathbf{P}$ and $\mathbf{Q}$ are nonsingular, then

$$\begin{aligned} & \det \begin{pmatrix} \mathbf{P} & \mathbf{S} \\ \mathbf{R} & \mathbf{Q} \end{pmatrix} \\ = \; & |\mathbf{P}| \cdot |\mathbf{Q} - \mathbf{R}\mathbf{P}^{-1}\mathbf{S}| \qquad (10.1a) \\ = \; & |\mathbf{Q}| \cdot |\mathbf{P} - \mathbf{S}\mathbf{Q}^{-1}\mathbf{R}|. \qquad (10.1b) \end{aligned}$$

  To prove the first identity, factor the matrix as

$$\begin{pmatrix} \mathbf{P} & \mathbf{S} \\ \mathbf{R} & \mathbf{Q} \end{pmatrix} = \begin{pmatrix} \mathbf{P} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{R} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{I} & \mathbf{P}^{-1}\mathbf{S} \\ \mathbf{0} & \mathbf{Q} - \mathbf{R}\mathbf{P}^{-1}\mathbf{S} \end{pmatrix},$$

  and compute the determinant of each factor by using cofactors. For the second, write

$$\begin{pmatrix} \mathbf{P} & \mathbf{S} \\ \mathbf{R} & \mathbf{Q} \end{pmatrix} = \begin{pmatrix} \mathbf{0} & \mathbf{I} \\ \mathbf{I} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{Q} & \mathbf{R} \\ \mathbf{S} & \mathbf{P} \end{pmatrix} \begin{pmatrix} \mathbf{0} & \mathbf{I} \\ \mathbf{I} & \mathbf{0} \end{pmatrix},$$

  note that

$$\begin{pmatrix} \mathbf{0} & \mathbf{I} \\ \mathbf{I} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{0} & \mathbf{I} \\ \mathbf{I} & \mathbf{0} \end{pmatrix} = \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{pmatrix},$$

  so that

$$\det \begin{pmatrix} \mathbf{P} & \mathbf{S} \\ \mathbf{R} & \mathbf{Q} \end{pmatrix} = \det \begin{pmatrix} \mathbf{Q} & \mathbf{R} \\ \mathbf{S} & \mathbf{P} \end{pmatrix} = |\mathbf{Q}| \cdot |\mathbf{P} - \mathbf{S}\mathbf{Q}^{-1}\mathbf{R}|,$$

  by the first identity. □

- **Example**:

$$\det \begin{pmatrix} \mathbf{I}_n & \mathbf{1}_n \\ \mathbf{1}_n' & -1 \end{pmatrix} = |\mathbf{I}_n| \cdot |-1 - \mathbf{1}_n' \mathbf{1}_n| = -1 - n.$$

- A consequence: if $\mathbf{A}$ is $m \times n$ and $\mathbf{B}$ is $n \times m$ then

$$\begin{aligned} & \det \begin{pmatrix} \mathbf{I}_m & \mathbf{A} \\ \mathbf{B} & \mathbf{I}_n \end{pmatrix} \\ = \; & |\mathbf{I}_m| \cdot |\mathbf{I}_n - \mathbf{B}\mathbf{A}| = |\mathbf{I}_n| \cdot |\mathbf{I}_m - \mathbf{A}\mathbf{B}|, \end{aligned}$$

  so that

$$|\mathbf{I}_m - \mathbf{A}\mathbf{B}| = |\mathbf{I}_n - \mathbf{B}\mathbf{A}|.$$

  – A particular case of special interest:

$$\left|\mathbf{I} - \mathbf{a}\mathbf{b}'\right| = 1 - \mathbf{b}'\mathbf{a}.$$

- Another useful identity:

$$(\mathbf{I} - \mathbf{A}\mathbf{B})^{-1} = \mathbf{I} + \mathbf{A}\,(\mathbf{I} - \mathbf{B}\mathbf{A})^{-1}\,\mathbf{B}.$$

  Few people (not me!) remember this formula; it can be motivated as follows:

  ...

and then proven by merely multiplying $\mathbf{I} - \mathbf{AB}$ by its alleged inverse. The formula is most useful when $\mathbf{I} - \mathbf{BA}$ is smaller, or more easily inverted, than $\mathbf{I} - \mathbf{AB}$.

– A particular case of special interest: if $\mathbf{b}'\mathbf{a} \neq 1$ then

$$\left(\mathbf{I} - \mathbf{ab}'\right)^{-1} = \mathbf{I} + \frac{\mathbf{ab}'}{1 - \mathbf{b}'\mathbf{a}}.$$

11. Block matrices II, LU decomposition

- The inverse of a non-singular block matrix can be computed by

$$\begin{pmatrix} \mathbf{P} & \mathbf{S} \\ \mathbf{R} & \mathbf{Q} \end{pmatrix}^{-1}$$

$$= \begin{pmatrix} \left(\mathbf{P}-\mathbf{S}\mathbf{Q}^{-1}\mathbf{R}\right)^{-1} & -\mathbf{P}^{-1}\mathbf{S}\cdot \\ & \left(\mathbf{Q}-\mathbf{R}\mathbf{P}^{-1}\mathbf{S}\right)^{-1} \\ -\left(\mathbf{Q}-\mathbf{R}\mathbf{P}^{-1}\mathbf{S}\right)^{-1}\cdot & \\ \mathbf{R}\mathbf{P}^{-1} & \left(\mathbf{Q}-\mathbf{R}\mathbf{P}^{-1}\mathbf{S}\right)^{-1} \end{pmatrix}.$$

  This is established merely by multiplying the original block matrix by its alleged inverse, and verifying that the product is indeed the identity matrix. The previous identity is used here. This is a very useful result if one of $\mathbf{P}$ or $\mathbf{Q}$ is much smaller than the other, which in turn has special structure which makes it easy to invert.

- **Example**:

$$\left(\mathbf{I}_n + \mathbf{1}\mathbf{1}'\right)^{-1} = \mathbf{I}_n - \frac{\mathbf{1}\mathbf{1}'}{1+n},$$

so that

$$\begin{pmatrix} \mathbf{I}_n & \mathbf{1} \\ \mathbf{1}' & -1 \end{pmatrix}^{-1} = \begin{pmatrix} (\mathbf{I}+\mathbf{1}\mathbf{1}')^{-1} & \mathbf{1}/(n+1) \\ \mathbf{1}'/(n+1) & -1/(n+1) \end{pmatrix}$$
$$= \frac{1}{n+1}\begin{pmatrix} (n+1)\mathbf{I}_n - \mathbf{1}\mathbf{1}' & \mathbf{1} \\ \mathbf{1}' & -1 \end{pmatrix}.$$

- Another consequence of (10.1) is: If $\mathbf{A}$ is an $m \times n$ matrix and $\mathbf{B}$ is a $n \times m$ matrix $(m \leq n)$, then the eigenvalues of $\mathbf{BA}$ are those of $\mathbf{AB}$ together with $n-m$ zeros.

  Proof: Using the two formulas for the determinant of a block matrix, we obtain

$$\begin{aligned} \det\begin{pmatrix} \mathbf{I}_m & \mathbf{A} \\ \mathbf{B} & \lambda\mathbf{I}_n \end{pmatrix} &= |\lambda\mathbf{I}_n| \left|\mathbf{I}_m - \lambda^{-1}\mathbf{AB}\right| \\ &= \lambda^n \left|\mathbf{I}_m - \lambda^{-1}\mathbf{AB}\right| \\ &= \lambda^{n-m} |\lambda\mathbf{I}_m - \mathbf{AB}| \\ \text{and} \quad &= |\mathbf{I}_m| \, |\lambda\mathbf{I}_n - \mathbf{BA}| = |\lambda\mathbf{I}_n - \mathbf{BA}| . \end{aligned}$$

  Thus, if P_{AB} and P_{BA} are the characteristic polynomials, we have

$$(-\lambda)^{n-m} P_{AB}(\lambda) = P_{BA}(\lambda)$$

and the roots of the characteristic equation $P_{BA}(\lambda) = 0$ for $\mathbf{BA}$ are those of the characteristic equation $P_{AB}(\lambda) = 0$ for $\mathbf{AB}$ together with $n - m$ 0's - the matrices $\mathbf{AB}$ and $\mathbf{BA}$ **have the same non-zero eigenvalues**; if $m = n$ then all n eigenvalues of $\mathbf{AB}$ equal those of $\mathbf{BA}$.

- A special case of the following was used in the 'classifying skulls' example. If $\mathbf{A} \geq \mathbf{0}$ and $\mathbf{B} > \mathbf{0}$, both $n \times n$, then the ratio $\frac{\mathbf{x}'\mathbf{A}\mathbf{x}}{\mathbf{x}'\mathbf{B}\mathbf{x}}$ of quadratic forms has a maximum value equal to the largest eigenvalue of $\mathbf{AB}^{-1}$. (In that example we applied this to maximize

$$\frac{\boldsymbol{\alpha}'(\bar{\mathbf{x}} - \bar{\mathbf{y}})(\bar{\mathbf{x}} - \bar{\mathbf{y}})'\boldsymbol{\alpha}}{\boldsymbol{\alpha}'\mathbf{S}\boldsymbol{\alpha}};$$

the same technique is used here.)

Proof: Let $\mathbf{B}^{1/2}$ be a square root (necessarily non-singular): $\mathbf{B} = \mathbf{B}^{1/2}\mathbf{B}^{1/2}$ and write

$$\mathbf{y} = \mathbf{B}^{1/2}\mathbf{x},\ \mathbf{x} = \mathbf{B}^{-1/2}\mathbf{y}.$$

Then $\mathbf{x}$ and $\mathbf{y}$ both range over all of $\mathbb{R}^n$ and so

$$\begin{aligned}\max_{\mathbf{x}} \frac{\mathbf{x}'\mathbf{A}\mathbf{x}}{\mathbf{x}'\mathbf{B}\mathbf{x}} &= \max_{\mathbf{y}} \frac{\mathbf{y}'\mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2}\mathbf{y}}{\mathbf{y}'\mathbf{y}} \\ &= ch_{\max}\mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2} \ (\geq 0) \\ &= ch_{\max}\mathbf{A}\mathbf{B}^{-1} \text{ (why?)}.\end{aligned}$$

- The **LU-decomposition**: A 'leading principal minor' of a square matrix is the determinant of the submatrix formed from the first m rows and columns. Suppose that a matrix $\mathbf{M}_{n\times n}$ is such that *all* leading principal minors are non-zero. Thus if $\mathbf{M}$ is partitioned as

$$\mathbf{M}_{n\times n} = \begin{pmatrix} \mathbf{A}_{m\times m} & \mathbf{B} \\ \mathbf{C} & \mathbf{D}_{(n-m)\times(n-m)} \end{pmatrix}, \tag{11.1}$$

(for any $m \leq n$), then $|\mathbf{A}| \neq 0$. In particular $|\mathbf{M}| \neq 0$. Then we can represent $\mathbf{M}$ as

$$\mathbf{M} = \mathbf{L}\mathbf{U},$$

where $\mathbf{L}$ and $\mathbf{U}$ are nonsingular, with $\mathbf{L}$ lower triangular and $\mathbf{U}$ upper triangular (both nonsingular).

Proof: By induction. The claim is obvious if $n = 1$; assuming it is true for some n we prove it for $n + 1$. Write

$$\mathbf{M}_{(n+1)\times(n+1)} = \begin{pmatrix} \mathbf{A}_{n\times n} & \mathbf{b} \\ \mathbf{c}' & d \end{pmatrix},$$

where $\mathbf{A}$ has all leading principal minors non-zero and hence (our induction hypothesis) we can factor it is $\mathbf{A} = \mathbf{LU}$. Then

$$\begin{aligned} \mathbf{M} &= \begin{pmatrix} \mathbf{LU} & \mathbf{b} \\ \mathbf{c}' & d \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{L} & \mathbf{0} \\ \mathbf{c}'\mathbf{U}^{-1} & d - \mathbf{c}'\mathbf{U}^{-1}\mathbf{L}^{-1}\mathbf{b} \end{pmatrix} \begin{pmatrix} \mathbf{U} & \mathbf{L}^{-1}\mathbf{b} \\ \mathbf{0}' & 1 \end{pmatrix}, \end{aligned}$$

and this is a product of a lower triangular matrix with an upper triangular matrix. □

12. Further examples and applications

- Under what assumptions is Least Squares an optimal estimation method? This is answered by the **Gauss-Markov Theorem**: Consider the linear model $\mathbf{y} = \mathbf{X}\beta + \boldsymbol{\varepsilon}$, with uncorrelated, equally varied errors ε and with $\mathbf{X}_{n\times p}$ having full rank p. So

$$\begin{aligned} E[\mathbf{y}] &= \mathbf{X}\beta, \\ \text{cov}[\mathbf{y}] &= \sigma_\varepsilon^2 \mathbf{I}_n. \end{aligned}$$

  Suppose that we seek to estimate a linear combination $\alpha = \mathbf{a}'\beta$ by an *unbiased, linear* estimate $\hat{\alpha} = \mathbf{c}'\mathbf{y}$. Then the minimum variance estimate in this class, i.e. the 'Best Linear Unbiased Estimate' (BLUE), is

$$\hat{\alpha}_{BLUE} = \mathbf{a}'\hat{\beta}_{OLS}$$

  where $\hat{\beta}_{OLS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. Thus

$$\begin{aligned} \hat{\alpha}_{BLUE} &= \mathbf{a}'\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}, \\ &= \mathbf{c}'\mathbf{y} \text{ with } \mathbf{c} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{a}. \end{aligned}$$

Proof: We are to show that $\hat{\alpha}_{BLUE}$ is unbiased (this is immediate) and that no unbiased estimate $\mathbf{c}'\mathbf{y}$ has a smaller variance. That $\mathbf{c}'\mathbf{y}$ be unbiased entails (how?)

$$\mathbf{X}'\mathbf{c} = \mathbf{a}, \tag{12.1}$$

and so we must show that, for any $\mathbf{c}$ satisfying (12.1),

$$\begin{aligned} \text{var}\left[\hat{\alpha}_{BLUE}\right] &\leq \text{var}\left[\mathbf{c}'\mathbf{y}\right], \text{ i.e.} \\ \mathbf{a}'\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{a} &\leq \mathbf{c}'\mathbf{c}. \end{aligned} \tag{12.2}$$

But by (12.1),

$$\mathbf{a}'\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{a} = \mathbf{c}'\mathbf{H}\mathbf{c},$$

and so (12.2) becomes $\mathbf{c}'\left(\mathbf{I}-\mathbf{H}\right)\mathbf{c} \geq 0$. This is true since (lab 4) any idempotent matrix is p.s.d. $\square$

- **Kronecker products**. This is (another) useful matrix operation, in which a big block matrix is formed from two smaller matrices:

$$\mathbf{A}_{m\times n}\otimes\mathbf{B}_{p\times q}=\left(a_{ij}\mathbf{B}\right)\ :mp\times nq.$$

Thus $\mathbf{A}\otimes\mathbf{B}$ has as elements all $mnpq$ products $a_{ij}b_{kl}$, arranged in a particular manner. In particular,

$$\mathbf{a}_{m\times 1}\otimes\mathbf{b}_{p\times 1}=\begin{pmatrix}a_1\mathbf{b}\\ \vdots\\ a_m\mathbf{b}\end{pmatrix}.$$

Useful identities are

$$\begin{aligned}(\mathbf{A}\otimes\mathbf{B})(\mathbf{C}\otimes\mathbf{D}) &= (\mathbf{AC}\otimes\mathbf{BD}),\\ (\mathbf{A}\otimes\mathbf{B})' &= \left(\mathbf{A}'\otimes\mathbf{B}'\right).\end{aligned}$$

The first of these implies that

$$(\mathbf{A}\otimes\mathbf{B})^{-1}=\left(\mathbf{A}^{-1}\otimes\mathbf{B}^{-1}\right),$$

if $\mathbf{A}$ and $\mathbf{B}$ are invertible. The Kronecker product is related to the *vec* operator being studied in the lab: if $\mathbf{P}_{r\times s}$ has columns $\mathbf{c}_1,...,\mathbf{c}_s$, i.e.

$$\mathbf{P}=(\mathbf{c}_1\vdots\cdots\vdots\mathbf{c}_s),$$

then $vec(\mathbf{P})$ consist of these columns stretched out as one long column:

$$vec(\mathbf{P}) = \begin{pmatrix} \mathbf{c}_1 \\ \vdots \\ \mathbf{c}_s \end{pmatrix} : rs \times 1.$$

Note that

$$\begin{aligned} vec\left(\mathbf{b}_{p\times 1}\mathbf{a}'_{m\times 1}\right) &= vec\left(\mathbf{b}a_1 \vdots \cdots \vdots \mathbf{b}a_m\right) \\ &= \begin{pmatrix} \mathbf{b}a_1 \\ \vdots \\ \mathbf{b}a_m \end{pmatrix} \\ &= \mathbf{a} \otimes \mathbf{b}. \end{aligned}$$

Now here is the relationship referred to above:

$$vec(\mathbf{ABC}) = \left(\mathbf{C}' \otimes \mathbf{A}\right) vec\mathbf{B}.$$

To see this, first suppose that $\mathbf{B}$ is 'rank one', i.e. is of the form $\mathbf{B} = \mathbf{xy}'$. Then

$$\begin{aligned} vec\,(\mathbf{ABC}) &= vec\left(\mathbf{Axy}'\mathbf{C}\right) \\ &= vec\left((\mathbf{Ax})\left(\mathbf{C}'\mathbf{y}\right)'\right) \\ &= \left(\mathbf{C}'\mathbf{y}\right) \otimes \mathbf{Ax} \\ &= \left(\mathbf{C}' \otimes \mathbf{A}\right)(\mathbf{y} \otimes \mathbf{x}) \\ &= \left(\mathbf{C}' \otimes \mathbf{A}\right) vec\mathbf{B}. \end{aligned}$$

The rest of it is being carried out as a lab problem.

– An application: Consider the equation

$$\mathbf{A}_{m\times n}\mathbf{X}_{n\times p}\mathbf{B}_{p\times q} = \mathbf{C}_{m\times q},$$

in which $\mathbf{X}$ is the unknown quantity. This is converted to a system of linear equations written in a more conventional manner if we take the vec's:

$$\left(\mathbf{B}' \otimes \mathbf{A}\right) vec\mathbf{X} = vec\mathbf{C};$$

now solve this in any of the usual ways.

Part II

UNIVARIATE CALCULUS & REAL ANALYSIS

13. Limits & continuity

- Open and closed sets in $\mathbb{R}^n$; limits:
  - Neighbourhood of a point '$\mathbf{a}$', of radius δ:

$$N_\delta(\mathbf{a}) = \{\mathbf{x}|\ ||\mathbf{x}-\mathbf{a}|| < \delta\}\,.$$

  - $A \subset \mathbb{R}^n$ is *open* if

$$\mathbf{a} \in A \Rightarrow N_\delta(\mathbf{a}) \subset A$$

    for all sufficiently small $\delta > 0$.

    * Example: $(0, 1)$.

  - A sequence $\{\mathbf{x}_n\}$ tends to a point $\mathbf{a}$: '$\mathbf{x}_n \to \mathbf{a}$' as $n \to \infty$ if $\mathbf{x}_n$ gets arbitrarily close to $\mathbf{a}$ as n gets larger and larger. More formally, any neighbourhood of $\mathbf{a}$, no matter how small, will eventually contain $\mathbf{x}_n$ from some point onward. More formally yet, 'for any radius δ, we can find an N large enough that, once $n > N$,

all of the $\mathbf{x}_n$ lie in $N_\delta(\mathbf{a})$'. This required N will typically get larger as δ gets smaller. Finally,

$$\forall\delta\exists N = N(\delta)\,(n > N \Rightarrow \mathbf{x}_n \in N_\delta(\mathbf{a}))\,,$$

read 'for all δ there exists an N, that depends on δ, such that $n > N$ implies that $\mathbf{x}_n \in N_\delta(\mathbf{a})$'.

* Equivalently (why?): $\mathbf{x}_n \to \mathbf{a} \Leftrightarrow ||\mathbf{x}_n - \mathbf{a}|| \to 0$.

* This is for '$\mathbf{a}$' finite; obvious modifications otherwise. You should derive an appropriate definition of '$x_n \to \infty$' (x_n *scalars*, not vectors).

* Example $x_n = 1 - \frac{1}{n} \to 1$ as $n \to \infty$.

– A point $\mathbf{a}$ is a *limit point* of $A \subset \mathbb{R}^n$ if there is a sequence $\{\mathbf{x}_n\} \subset A$ such that $\mathbf{x}_n \to \mathbf{a}$.

* Example $A = (0,1), x_n = 1 - \frac{1}{n};\ a = 1\ (\notin A)$.

– $A \subset \mathbb{R}^n$ is *closed* if it contains all of its limit points.

  * Examples $A = [0,1]$ (if $a < 0$ or > 1 it cannot be the limit of a sequence in $[0,1]$).

– A set A is open iff A^c is closed.
  Proof: You will show (lab problem) that

$$A \text{ open } \Rightarrow A^c \text{ closed.}$$

  Conversely, suppose A^c is closed; we are to show that A is open. We will derive a contradiction from the supposition that A is *not* open. Suppose it isn't; then for some $a \in A$, no $N_\delta(a) \subset A$ (no matter how small we choose δ). Then in particular $N_{1/n}(a)$ contains points $x_n \in A^c$. Since $|x_n - a| < 1/n \to 0$, we have $x_n \to a$ and so a is a limit point of A^c, hence a member of A^c (why?). This contradicts the fact that $a \in A$, thus completing the proof.

– Arbitrary unions of open sets are open, and finite intersections are open.

Proof: If $\{A_i\}_{i\in I}$ are open, and $\mathbf{a} \in \cup_{i\in I} A_i$ then $\mathbf{a} \in A_{i^*}$ for some i^*. Then there is $N_\delta(\mathbf{a}) \subset A_{i^*} \subset \cup_{i\in I} A_i$. The second statement is a lab problem.

- A function $f(\mathbf{x}) \to L$ as $\mathbf{x} \to \mathbf{a}$ ('$f(\mathbf{x})$ tends to L as $\mathbf{x}$ tends to $\mathbf{a}$') if we can force $f(\mathbf{x})$ to be arbitrarily close to L by choosing $\mathbf{x}$ ($\neq \mathbf{a}$) sufficiently close to $\mathbf{a}$. Formally,

  $$\forall\varepsilon\exists\delta = \delta(\varepsilon, \mathbf{a})\,(0 < ||\mathbf{x} - \mathbf{a}|| < \delta \Rightarrow |f(\mathbf{x}) - L| < \varepsilon).$$

  The '$= \delta(\varepsilon, \mathbf{a})$' is often omitted (but understood, unless stated otherwise). Note the '$0 < ||\mathbf{x} - \mathbf{a}||$': $f(\mathbf{a})$ need not exist.

- Suppose $f(\mathbf{x})$ is defined for $\mathbf{x} \in D$, the *domain* of f. Then f is *continuous* at a point $\mathbf{a} \in D$ if $f(\mathbf{x}) \to f(\mathbf{a})$ as $\mathbf{x} \to \mathbf{a}$.

  - The definition requires f to be defined at $\mathbf{a}$.

  - Equivalently, $\forall\varepsilon\exists\delta = \delta(\varepsilon, \mathbf{a})\,(||\mathbf{x} - \mathbf{a}|| < \delta \Rightarrow |f(\mathbf{x}) - f(\mathbf{a})| < \varepsilon)$.

14. Continuity and Differentiation

- Suppose f maps $\mathbf{x} \in D \subset \mathbb{R}^n$ to a real number $f(\mathbf{x})$. We write $f : D \subset \mathbb{R}^n \to \mathbb{R}$; D is the *domain* of f. Recall the definition: The function f is *continuous* at a point $\mathbf{a} \in D$ if, as $\mathbf{x} \to \mathbf{a}$, we have that $f(\mathbf{x}) \to f(\mathbf{a})$ (equivalently, $|f(\mathbf{x}) - f(\mathbf{a})| \to 0$ as $\|\mathbf{x} - \mathbf{a}\| \to 0$).

- **Example**: $f(x) = x^2$, $D = (0, \infty)$. Then if $x, a > 0$ we have

$$\begin{aligned} 0 &\leq |f(x) - f(a)| \\ &= |x - a||x - a + 2a| \\ &< |x - a| \cdot (|x - a| + 2a)| \\ &\to 0, \end{aligned}$$

  as $|x - a| \to 0$.
  Here we used the 'triangle inequality':

$$|a + b| \leq |a| + |b|.$$

- Infimum and suprema (think 'min' and 'max', but ...): For any set A, q is a *lower bound* if $q \leq a$ for all $a \in A$. If there is a finite lower bound then there are many; the largest of them is the *greatest lower bound* ($g.l.b$) or *infimum* (inf). Otherwise the inf is $-\infty$. Similarly with *upper bound*, *least upper bound* ($l.u.b.$) or *supremum* (sup).

  - Here is a simple but very useful result (Lab 5). If $\{x_n\}$ is increasing: ... $x_n \leq x_{n+1}$ ... and bounded above: $x_n \leq B < \infty$, then $S = \sup x_n$ is finite (why?) and $x_n \to S$ as $n \to \infty$.

- Further properties of continuous functions $f : D \subset \mathbb{R} \to \mathbb{R}$:

  - If f is continuous on a closed and bounded set D then it is bounded there. Thus the inf and sup are finite, and *are attained*: there are points $\alpha, \beta \in D$ with $f(\alpha) \leq f(x) \leq f(\beta)$ for all $x \in D$. (What can fail on an open domain?)

– If f is continuous on a closed and bounded set D then it is 'uniformly continuous' there. A function f defined on a subset D of the real line is said to be uniformly continuous on D if, for every $\varepsilon > 0$ there is a δ such that

$$\forall x, y \in D\text{: } |x - y| < \delta \Rightarrow |f(x) - f(y)| < \varepsilon,$$

(so that f is continuous at every point of D) **AND the same** δ works for all x and y, i.e. it does not depend on which points are being compared.

- Let $f : D \subset \mathbb{R} \to \mathbb{R}$ be defined in a neighbourhood $N_r(x_0) \subset D$; put

$$\phi(h) = \frac{f(x_0 + h) - f(x_0)}{h}$$

('Newton's quotient'). If $\phi(h)$ has a limit as $h \to 0$ (in *any manner*) we call it the *derivative* $f'(x_0)$ of f at x_0, also written $(df(x)/dx)_{|x=x_0}$.

- Examples $f(x) = x^2$, $f(x) = |x|$. The former is differentiable everywhere in $\mathbb{R}$; the latter everywhere except $x = 0$.

- Differentiability $\Rightarrow$ Continuity: If $f'(x_0)$ exists then f is continuous at x_0.
Proof:

$$\begin{aligned} |f(x_0 + h) - f(x_0)| &= |h\phi(h)| = |h|\,|\phi(h)| \\ &\to 0 \cdot \left|f'(x_0)\right| = 0, \end{aligned}$$

as $|h| \to 0$. □

- Linearity, product, quotient, chain rules - read in (any) text. They allow us to build up a stock of differentiable functions from simpler ones, and also show how the derivative of the more complicated function can be gotten from those of the simpler ones.

15. Mean Value Theorem

- Recall that

  $$f'(x_0) = \lim_{h \to 0} \phi(h), \text{ for } \phi(h) = \frac{f(x_0 + h) - f(x_0)}{h},$$

  provided this limit exists.

- Relation to monotonicity: if $f \nearrow$ ('weakly increasing': $x < y \Rightarrow f(x) \leq f(y)$) on (a, b) and differentiable there then $f'(x) \geq 0$ on (a, b).

  Proof: As $h \downarrow 0$ the numerator of $\phi(h)$ is ≥ 0 and continuous, hence $f'(x) = \lim_{h \downarrow 0} \phi(h) \geq 0$. (Similarly $\lim_{h \uparrow 0} \phi(h) \geq 0$.) $\square$

- Lab problem: If f is differentiable on (a, b) and attains a maximum (or minimum) at $c \in (a, b)$ then $f'(c) = 0$.

- **Mean Value Theorem**: If f is continuous on $[a, b]$ and differentiable on (a, b) then $\exists c \in (a, b)$ with

$$f(b) = f(a) + f'(c)(b - a). \tag{15.1}$$

This is a result of crucial importance in the approximation of functions.

  - An interpretation is that 'differentiable functions are locally almost linear': If b and a are very close, and f' is continuous, we can approximate $f'(c)$ by $f'(a)$:

$$f(b) \approx f(a) + f'(a)(b - a);$$

here the rhs is a straight line (as a function of b), with slope $f'(a)$.

  - A consequence of the MVT is that if $f'(x) \geq 0$ on (a, b) then $f \nearrow$ there: suppose $a < x_1 < x_2 < b$, then

$$f(x_2) = f(x_1) + f'(c)(x_2 - x_1) \geq f(x_1).$$

- **Proof of the MVT**: Define

$$\psi(x) = f(x) - f(a) - \left(\frac{f(b) - f(a)}{b - a}\right)(x - a).$$

This is the difference between $f(x)$ and its approximation on the line between $(a, f(a))$ and $(b, f(b))$. It is enough if we can find a point $c \in (a, b)$ with $0 = \psi'(c)$, since then

$$0 = f'(c) - \left(\frac{f(b) - f(a)}{b - a}\right),$$

and this is equivalent to (15.1). Note that $\psi(a) = \psi(b) = 0$. If $\psi(x) = 0$ for all $x \in (a, b)$ then $0 = \psi'(c)$ for *any* c. If there is $x \in (a, b)$ with $\psi(x) \neq 0$ then either the sup or the inf is non-zero and is attained at some point $c \in (a, b)$, so that $\psi'(c) = 0$. □

- l'Hospital's Rule: Read in any text.
  - Rough idea: If $f(a) = g(a) = 0$, then

$$\lim_{x \to a} \frac{f(x)}{g(x)} = \lim_{x \to a} \frac{\frac{f(x)-f(a)}{x-a}}{\frac{g(x)-g(a)}{x-a}} = \lim_{x \to a} \frac{f'(x)}{g'(x)}.$$

  - Example: $\lim_{x \to 0} \frac{\sin x}{x} = \lim_{x \to 0} \frac{\cos x}{1} = 1$.
  - Others: $\lim_{x \to 0} \frac{1-\cos x}{x^2}$, $\lim_{x \to \infty} \frac{a^x}{x}$ $(a > 1)$, $\lim_{x \to \infty} \frac{a^x}{x^m}$ $(a > 1,\ m > 0)$, $\lim_{x \to 0} x^x$.

16. Probability spaces and random variables

- We start with a *sample space* Ω, whose elements are all possible outcomes of an experiment (e.g. toss a coin three times, Ω is all possible sequences of three Ts and Hs). A *Borel field* or σ-algebra of events is a collection $\mathbb{B}$ of subsets E ('events') of Ω such that one of its elements is Ω itself, it is closed under complementation, and closed under the taking of countable unions.

- A *probability* is a function P defined on $\mathbb{B}$ such that $P(\Omega) = 1$, $0 \leq P(E) \leq 1$, and probabilities of disjoint countable unions are additive. The triple $(\Omega, \mathbb{B}, P)$ is called a *probability space*. All the usual rules for manipulating probabilities follow from these axioms. e.g. $P(E^c) = 1 - P(E)$, $P(\phi) = 0$, $P(E) \leq P(F)$ if $E \subset F$.

- A (real valued, finite) random variable (r.v.) is a function $X : \Omega \to \mathbb{R}$ with the property that

if A is any open subset of $\mathbb{R}$, then $X^{-1}(A) = \{\omega \mid X(\omega) \in A\}$ is an event, i.e. a member of $\mathbb{B}$. E.g. $X(\omega) = \#$ of heads in the sequence ω of tosses. (For a finite sample space Ω we generally take $\mathbb{B}$ to be the set of *all* subsets of Ω.)

- Note that $X^{-1}(A^c) = \left\{X^{-1}(A)\right\}^c$:

$$\begin{aligned} X^{-1}(A^c) &= \{\omega \,|\, X(\omega) \in A^c\} \\ &= \{\omega \,|\, X(\omega) \notin A\} \\ &= \{\omega \,|\, X(\omega) \in A\}^c \\ &= \left\{X^{-1}(A)\right\}^c. \end{aligned}$$

- By the preceding points, if B is closed then $A = B^c$ is open and so $X^{-1}(B) = \left\{X^{-1}(A)\right\}^c \in \mathbb{B}$: the inverse images of closed sets must also be events.

- Since the set $A = (-\infty, x]$ is closed, so also $X^{-1}(A) = \{\omega \mid X(\omega) \leq x\}$ is a member of $\mathbb{B}$, hence has a probability. We write

$$F(x) = P\left(\{\omega \mid X(\omega) \leq x\}\right) = P\left(X \leq x\right)$$

and call F the *distribution function* (d.f.) of the r.v. X.

– A d.f. is then a function $F : \mathbb{R} \to [0,1]$ satisfying (i) $F(-\infty) = 0$, $F(\infty) = 1$ [why? $F(-\infty) = P(\{\omega \mid X(\omega) \leq -\infty\}) = P(\phi) = 0$; similarly $F(\infty) = P(\Omega) = 1$] (ii) F is *weakly increasing*: $x < y \Rightarrow F(x) \leq F(y)$ (lab problem) and (iii) F is *right continuous*, in that $x_n \downarrow x \Rightarrow F(x_n) \to F(x)$.

- Here is the proof that a distribution function is right continuous. It is based on the 'continuity of probabilities':

$$\begin{aligned} E_n &\supseteq E_{n+1} \supseteq \cdots \text{ and } \cap_{n=1}^{\infty} E_n = E \\ &\Rightarrow P(E_n) \to P(E). \end{aligned} \tag{16.1}$$

Define $A = (-\infty, x]$ and $A_n = (-\infty, x_n]$. Then $F(x) = P\left(X^{-1}(A)\right)$ and $F(x_n) = P\left(X^{-1}(A_n)\right)$. So define events $E = X^{-1}(A)$ and $E_n = X^{-1}(A_n)$. Then as $x_n \downarrow x$ we have

$$\cdots A_n \supseteq A_{n+1} \cdots \supseteq \cap_{n=1}^{\infty} A_n = A,$$

and so

$$\cdots E_n \supseteq E_{n+1} \supseteq \cdots \supseteq \cap_{n=1}^{\infty} E_n \stackrel{*}{=} E,$$

($*$: $E = X^{-1}(\cap_{n=1}^{\infty} A_n) = \cap_{n=1}^{\infty} X^{-1}(A_n) = \cap_{n=1}^{\infty} E_n$) hence by (16.1),

$$F(x) = P(E) = \lim_{n \to \infty} P(E_n) = \lim_{x_n \downarrow x} F(x_n).$$

Recall the notion of expected value, which we defined in terms of a density or probability mass function. If $F(x)$ is differentiable then $f = F'$ is the 'probability density function' density (p.d.f.) and expectations, probabilities etc. are obtained by (Riemann -) integration of f. If F is a step function with jumps of height p_n at points x_n ($n = 0, 1, 2, ...$) then the 'probability mass function' (p.m.f.) is the function $f(x_n) = p_n$ and expectations, probabilities etc. are obtained by summation over f. In the former case we say that X is *continuous*; in the latter X is *discrete*. These notions can be unified via the Riemann-Stieltjes integral, (possibly) to be discussed later.

17. Convergence in probability, Jensen's inequality

- Limits and continuity in probability: Let $\{X_n\}$ be a sequence of r.v.s, e.g. toss a fair coin n times and let X_n denote the proportion of heads in the n tosses. Then $E[X_n] = 1/2$ and we also expect X_n to be near $1/2$, with high probability, for n large. We say that 'X_n converges to a constant c *in probability*', and write $X_n \xrightarrow{pr} c$, if

$$\lim_{n\to\infty} P\left(|X_n - c| \geq \varepsilon\right) = 0 \text{ for any } \varepsilon > 0.$$

  The *Weak Law of Large Numbers* states that if X_n is the average of n independent r.v.s $Z_1, ..., Z_n$, all with finite mean μ, then $X_n \xrightarrow{pr} \mu$.

  - e.g. $Z_i = I(\text{i}^{th}$ toss results in a head), $X_n = \sum Z_i/n$. Then $Z_i = 1, 0$ with probability $1/2$ each; $\mu = 1/2$; by the WLLN $X_n \xrightarrow{pr} 1/2$.

- This is a basic notion required for the theory of estimation in Statistics.

  - e.g. The variance σ^2 of a population is estimated, via a *sample* $X_1, ..., X_n$ from the population, by the sample variance S_n^2. The WLLN can be used to show that $S_n^2 \xrightarrow{pr} \sigma^2$. We say S_n^2 is a *consistent estimate* of σ^2. Then also $S_n \xrightarrow{pr} \sigma$; this is a consequence of the following result.

- If $X_n \xrightarrow{pr} c$ and the function g is continuous at c, then $g(X_n) \xrightarrow{pr} g(c)$.
  Proof: We want to show that for $\varepsilon > 0$,

$$P(|g(X_n) - g(c)| \geq \varepsilon) \to 0.$$

  Use the continuity of g to find $\delta > 0$ such that

$$|X_n - c| < \delta \Rightarrow |g(X_n) - g(c)| < \varepsilon.$$

  Then

$$P(|X_n - c| < \delta) \leq P(|g(X_n) - g(c)| < \varepsilon).$$

  Here we use the fact that if one event implies another, it has a smaller probability (i.e. $E \subset F \Rightarrow P(E) \leq P(F)$). Since the first probability $\to 1$, so does the second (why?). $\square$

- **Convex functions**: A function $f : D \to \mathbb{R}$ is convex if

$$f((1-\lambda)x + \lambda y) \leq (1-\lambda)f(x) + \lambda f(y)$$

for all $x, y \in D$. Examples: x^2, $|x|$ on $\mathbb{R}$, $-\log x$ on $(0, \infty)$.

- Convex functions are necessarily continuous. If a function has a derivative $f'(x)$ on D which is $\nearrow$, then it is convex. In particular $f''(x) \geq 0$ implies convexity.

- **Jensen's Inequality**: If X is a r.v. taking values in D, has a finite mean $E[X]$, and if f is convex on D, then $E[f(X)] \geq f(E[X])$.

  - Some applications: $E\left[X^2\right] \geq (E[X])^2$ (duh!); $E[1/X] \geq 1/E[X]$ if $X > 0$ (in other words, the r.v.s X and $1/X$ are negatively correlated - obvious?).

– The arithmetic/geometric mean inequality:

$$\text{if } x_1, ..., x_n > 0 \text{ then } \left(\prod x_i\right)^{1/n} \leq \bar{x}.$$

Proof: Define a r.v. X by $P(X = x_i) = 1/n$ and apply Jensen's Inequality using the convex function $f(x) = -\log x$. □

– Sketch of proof of Jensen's Inequality: let l be the linear function, tangent to f at the point $(E[X], f(E[X]))$. Then $l \leq f$, and so

$$f(E[X]) = l(E[X]) = E[l(X)] \leq E[f(X)].$$

18. Taylor's Theorem

- **Taylor's Theorem**: 'Sufficiently smooth functions can be approximated locally by polynomials.' Suppose $f(x)$ has n derivatives on (a, b) with $f^{(n-1)}(x)$ continuous on $[a, b]$. (We put $f^{(0)}(x) = f(x)$; the assumptions imply existence and continuity of $f^{(k)}(x)$ on (a, b) for $k < n$.) Then for $x \in [a, b]$ there is a point ξ between a and x such that

$$f(x) = \sum_{k=0}^{n-1} f^{(k)}(a)\frac{(x-a)^k}{k!} + f^{(n)}(\xi)\frac{(x-a)^n}{n!}.$$

  What does this say when $n = 1$?

- **Example**: $f(x) = e^x$; expand around $x = 0$ ('Maclaurin series'):

$$\begin{aligned} f^{(k)}(x) &= e^x, \text{ so that} \\ f^{(k)}(0) &= 1. \end{aligned}$$

Then for some ξ between 0 and x (i.e. $|\xi| \leq |x|$)

$$\begin{aligned} e^x &= \sum_{k=0}^{n-1} f^{(k)}(0)\frac{x^k}{k!} + f^{(n)}(\xi)\frac{x^n}{n!} \\ &= \sum_{k=0}^{n-1} \frac{x^k}{k!} + e^{\xi}\frac{x^n}{n!}. \end{aligned}$$

Write this as

$$e^x = p_n(x) + R_n(x).$$

We can do this since ξ is a, generally quite complicated, function of x. If $R_n(x) \to 0$ as $n \to \infty$ we say that the series $\lim_{n\to\infty} p_n(x) = \sum_{k=1}^{\infty} x^k/k!$ 'represents the function' e^x.

– It is indeed true in this example that $R_n(x) \to 0$ as $n \to \infty$. We will later show that it does so in a nice 'uniform' way; for now it is enough to show that it does so for any particular point x. Note that

$$|R_n(x)| = \frac{\left|e^{\xi}x^n\right|}{n!} \leq \frac{e^{|x|}\,|x|^n}{n!}.$$

Choose any integer $M > |x|$ and keep it fixed. Then for $n > M$ we have that

$$\begin{aligned}
|R_n(x)| &\leq \frac{e^M M^n}{n!} \\
&= \frac{e^M M^M M^{n-M}}{M!\,(M+1)\,(M+2)\cdots(M+n-M)} \\
&= \frac{e^M M^M}{M!} \cdot \frac{M}{M+1} \cdot \frac{M}{M+2} \cdots \frac{M}{M+(n-M)} \\
&< \frac{e^M M^M}{M!} \left(\frac{M}{M+1}\right)^{n-M} \\
&= cr^n,
\end{aligned}$$

for $c = \frac{e^M M^M}{M!} \left(\frac{M}{M+1}\right)^{-M}$ (a constant) and $r = \frac{M}{M+1} \in (0,1)$. Now note that

$$cr^n \to 0 \text{ as } n \to \infty,$$

so that $R_n(x) \to 0$ as $n \to \infty$.

- **Example**: $f(x) = \log(1+x)$ with $|x| < 1$; expand around 0: $f(0) = 0$ and for $k > 0$:

$$\begin{aligned} f^{(k)}(x) &= (-1)^{k+1} \frac{(k-1)!}{(1+x)^k}, \text{ so that} \\ f^{(k)}(0) &= (-1)^{k+1} (k-1)!. \end{aligned}$$

Then for some ξ with $|\xi| < |x|$,

$$\begin{aligned} & \log(1+x) \\ = & \sum_{k=1}^{n-1} (-1)^{k+1} \frac{x^k}{k} + \frac{(-1)^{n+1}}{(1+\xi)^n} \frac{x^n}{n} \\ = & x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \ldots + (-1)^n \frac{x^{n-1}}{n-1} \\ & + \frac{(-1)^{n+1}}{(1+\xi)^n} \frac{x^n}{n} \end{aligned}$$

Write this as

$$\log(1+x) = p_n(x) + R_n(x);$$

we will later derive a more refined expression for $R_n(x)$ ("integral form of the remainder") from which it follows that $R_n(x) \to 0$ as $n \to \infty$, so that the series $\lim_{n\to\infty} p_n(x) = \sum_{k=1}^{\infty} (-1)^{k+1} x^k/k$ represents the function $\log(1+x)$ if $|x| < 1$.

19. Examples I - transforming r.v.s; order statistics

- **Distribution of functions of r.v.s.** Suppose a r.v. X has a differentiable d.f. $F(x)$, density $f(x) = F'(x)$. Consider the r.v. $Y = \psi(X)$. (e.g. $Y = \log X$.) First assume ψ is *strictly monotonic* ($\uparrow$ or $\downarrow$) with inverse ψ^{-1}. The d.f. of Y is

$$\begin{aligned} G(y) &= P(Y \le y) = P(\psi(X) \le y) \\ &= \begin{cases} P(X \le \psi^{-1}(y)), & \text{if } \psi \uparrow, \\ P(X \ge \psi^{-1}(y)), & \text{if } \psi \downarrow; \end{cases} \\ &= \begin{cases} F(\psi^{-1}(y)), & \text{if } \psi \uparrow, \\ 1 - F(\psi^{-1}(y)), & \text{if } \psi \downarrow . \end{cases} \end{aligned}$$

To get the density $g(y)$ of $G(y)$ we must differentiate $\psi^{-1}(y)$. Write $x = \psi^{-1}(y)$, then $\left(\psi^{-1}(y)\right)' = dx/dy$ can be obtained by differentiating the relationship $y = \psi(x)$:

$$1 = \frac{dy}{dy} = \psi'(x)\frac{dx}{dy};$$

hence

$$\frac{dx}{dy} = \frac{1}{\psi'(x)} = \frac{1}{\psi'(\psi^{-1}(y))}.$$

In the above,

$$g(y) = \begin{cases} f(\psi^{-1}(y))/\left[\psi'(\psi^{-1}(y)\right], & \text{if } \psi \uparrow, \\ f(\psi^{-1}(y))/\left[-\psi'(\psi^{-1}(y)\right], & \text{if } \psi \downarrow . \end{cases}$$

The form of this which is worth remembering is:

$$g(y) = f(x)\left|\frac{dx}{dy}\right| \text{ if } Y = \psi(X) \text{ is strictly monotone,}$$

with x expressed in terms of y on the RHS.

- **Example**: $X > 0$, $Y = -\log X$. Then $g(y) = f(x)\left|\frac{dx}{dy}\right| = f(e^{-y})\left|\frac{de^{-y}}{dy}\right| = f(e^{-y})e^{-y}$. Thus if $X \sim Unif(0,1)$ with $f(x) = I(0 < x < 1)$, Y has density e^{-y} $(y > 0)$; we say Y has the exponential density with mean 1. (The function $g(y) = \lambda e^{-\lambda y}$ is the exponential p.d.f. with mean $1/\lambda$.)

- When ψ is non-monotonic one generally applies these ideas, but perhaps more directly. **Example**: $X \sim N(0,1)$, $Y = X^2 \sim$ how? Then $G(y) = P\left(-\sqrt{y} \le X \le \sqrt{y}\right) = \Phi\left(\sqrt{y}\right) - \Phi\left(-\sqrt{y}\right)$; continue with $\Phi' = \phi$, the $N(0,1)$ density. (Lab problem).

- A reminder: random variables $X_1, ..., X_n$ are *independent* if the distribution of any of them, given the values of the others, does not depend on these values. More on this later; for now recall that an equivalent formulation is

$$P\left(X_1 \leq x_1 \text{ and } \cdots \text{ and } X_n \leq x_n\right) = \prod_{i=1}^{n} P\left(X_i \leq x_i\right).$$

- **Distribution of order statistics**. Suppose we take a random sample (i.e., i.i.d.) $X_1, ..., X_n$. Arranged in increasing order these are the 'order statistics':

$$X_{(1)} \leq \cdots \leq X_{(n)}.$$

In particular

$$X_{(1)} = \min\left\{X_i\right\}, \; X_{(n)} = \max\left\{X_i\right\}.$$

The order statistics are not independent (why not?). They are however commonly analyzed using a trick that allows us to use techniques for independent r.v.s. In what follows we assume that the d.f. $F\left(x\right) = P\left(X_i \leq x\right)$ is continuous, and in fact

has a density $f(x)$. Then, for instance, the d.f. of $X_{(1)}$ is

$$\begin{aligned} G_{(1)}(x) &= P\left(X_{(1)} \leq x\right) \\ &= 1 - P\left(X_{(1)} > x\right) \\ &= 1 - P\left(\text{all } X_i \text{ are } > x\right). \end{aligned}$$

That was the 'trick'; now we are back to talking about independent r.v.s:

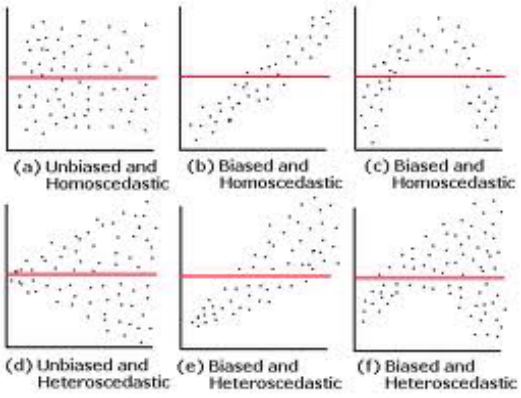
$$G_{(1)}(x) = 1 - \prod_{i=1}^{n} P\left(X_i > x\right) = 1 - \bar{F}^n(x),$$

where we use the common notation $\bar{F} = 1 - F$. Then the density is

$$g_{(1)}(x) = n\bar{F}^{n-1}(x)\, f(x).$$

The distribution of $X_{(n)}$ can be obtained in a similar manner.

20. Examples II - variance stabilization, convergence in law



Residual plots: e_i vs. $\hat{y}_i$.

- Suppose we fit a regression model, which typically assumes that the observations $Y_1, ..., Y_n$ are equally varied $(Y|\mathbf{x} \sim N\left(\mathbf{x}'\boldsymbol{\beta}, \sigma^2_Y\right))$, but a plot of the residuals $e = Y - \mathbf{x}'\hat{\boldsymbol{\beta}}$ against $\hat{Y} = \mathbf{x}'\hat{\boldsymbol{\beta}}$ indicates that the variance σ^2_Y changes with $\mu_Y = \mathbf{x}'\boldsymbol{\beta}$ (as estimated by $\hat{Y}$). See (d) above.

  - Can we find a transformation $Z = \psi(Y)$ such that a regression of Z on $\mathbf{x}$ results in a 'stable' variance, i.e. σ^2_Z is constant (as in (a))?

- Apply Taylor's Theorem: if $Z = \psi(Y)$, and if Y is near μ_Y, then expanding $\psi(Y)$ around μ_Y gives

$$\begin{aligned} Z &= \psi(\mu_Y) + \psi'(\mu_Y)(Y - \mu_Y) + \psi''(\xi)\frac{(Y-\mu_Y)^2}{2!} \\ &\approx \psi(\mu_Y) + \psi'(\mu_Y)(Y - \mu_Y), \end{aligned}$$

with

$$\begin{aligned} E[Z] &\approx \psi(\mu_Y), \\ \text{var}[Z] &\approx E\left[(Z - \psi(\mu_Y))^2\right] \approx \left(\psi'(\mu_Y)\sigma_Y\right)^2. \end{aligned}$$

Thus if, as is usually assumed, $Y \sim N\left(\mu_Y, \sigma_Y^2\right)$ then, approximately, $Z \sim N\left(\psi(\mu_Y), \left(\psi'(\mu_Y)\sigma_Y\right)^2\right)$.

  - **The 'delta method' of 'variance stabilization'.** We choose the transformation ψ so that σ_Z^2 will be constant: if $\sigma_Y^2 = h(\mu_Y)$, then we choose ψ so that

    $\psi'(\mu_Y)\sigma_Y = \psi'(\mu_Y)\sqrt{h(\mu_Y)} = c$ (a constant).

    We solve $\psi'(\mu_Y) \propto 1/\sqrt{h(\mu_Y)}$, obtaining the indefinite integral $\psi(\mu_Y) \propto \int^{\mu_Y} \frac{1}{\sqrt{h(y)}} dy$.

  - Example: if it appears that $\sigma_Y^2 \propto \mu_Y$ (so that $h(y) = y$), then the appropriate transformation would be ... $\psi(\mu_Y) \propto \sqrt{\mu_Y}$. So we would run the regression again, this time with $Z = \sqrt{Y}$ as the dependent variable.

- These kinds of approximations all become exact 'asymptotically', i.e. as the sample size $n \to \infty$. A basic tool is that of *convergence in law*, or 'in distribution'. Suppose $\{X_n\}$ is a sequence of r.v.s with d.f.s $F_n(x) = P(X_n \leq x)$; we say that $X_n \xrightarrow{L} X \sim F$ (or just $X_n \xrightarrow{L} F$) if

$$F_n(x) = P(X_n \leq x) \to P(X \leq x) = F(x)$$

  at every continuity point of F.

- The Central Limit Theorem (CLT; we'll prove it later) refers to this kind of convergence: if $X_n = \sqrt{n}\left(\bar{Y}_n - \mu\right)$, where $\bar{Y}_n = \sum_{i=1}^n Y_i/n$ and $Y_1, ..., Y_n$ are a sample with mean μ and variance σ^2, then $E[X_n] = 0$, $\text{var}[X_n] = \sigma^2$ and $X_n \xrightarrow{L} N(0, \sigma^2)$.

  - Example: a $bin\left(n,p\right)$ r.v., properly scaled and centred, is 'asymptotically $N(0,p\left(1-p\right))$'.

    * **Proof**: If $Y_i = I\left(i^{\text{th}} \text{ experiment is a 'success'}\right)$ then $S_n = \sum_{i=1}^{n} Y_i \sim bin\left(n,p\right)$ ...

- The CLT, WLLN and Taylor's Theorem together are sufficient to derive a vast array of large sample approximations in Mathematical Statistics. How? We very often study a functions of averages, say $\psi\left(\bar{Y}_n\right)$, in order to make inferences. Taylor's Theorem allows us to treat this as approximately a linear function of $\bar{Y}_n$:

$$\psi\left(\bar{Y}_n\right) = \psi\left(\mu_Y\right) + \psi'\left(\mu_Y\right)\left(\bar{Y}_n - \mu_Y\right) + R_n,$$

where the remainder is $R_n = \psi''\left(\xi\right)\frac{\left(\bar{Y}_n-\mu_Y\right)^2}{2}$. Then (apart from $\sqrt{n}R_n$, which is typically 'asymptotically negligible'),

$$\sqrt{n}\left(\psi\left(\bar{Y}_n\right) - \psi\left(\mu_Y\right)\right) \approx \psi'\left(\mu_Y\right)\cdot\sqrt{n}\left(\bar{Y}_n - \mu_Y\right),$$

which, by the CLT, $\xrightarrow{L} N\left(0,\left(\psi'\left(\mu_Y\right)\sigma_Y\right)^2\right)$.

– The WLLN, which states that $\bar{Y}_n \xrightarrow{pr} \mu_Y$, can be used here to show that $\sqrt{n}R_n \xrightarrow{pr} 0$, and so has no effect on the limiting normal distribution of $\sqrt{n}\left(\psi\left(\bar{Y}_n\right) - \psi\left(\mu_Y\right)\right)$. This is what we mean when we say that $\sqrt{n}R_n$ is 'asymptotically negligible'.

21. Sequences and Series

- **Convergence of a sequence** $\{a_n\}_{n=1}^{\infty}$: We say that $a_n \to a$ as $n \to \infty$ if

$$\forall \varepsilon > 0 \exists N (n > N \Rightarrow |a_n - a| < \varepsilon).$$

(So the *function* $f(n) = a_n \to a$.)

  - **Example**: Let $a_n = r^n$, $|r| < 1$. Then $a_n \to 0$. ($N = \log \varepsilon / \log |r|$.)

- A monotonic, bounded sequence is convergent. (**Proof**: We did this earlier ...)
  Example: $a_1 = \sqrt{2}$, $a_{n+1} = \sqrt{2 + a_n}$. Since the function $f(x) = \sqrt{2 + x}$ is continuous, and $a_{n+1} = f(a_n)$, if the sequence is convergent to a limit 'a' then $a = f(a)$. Thus $(a - 2)(a + 1) = 0$.
  Claim: the sequence is increasing and bounded, hence it is convergent and so $a = 2$. For this, suppose that $0 \leq a_n \leq 2$, as is true for $n = 1$.

Then also $a_{n+1} = \sqrt{2 + a_n} \leq \sqrt{2 + 2} = 2$ – this shows that the sequence is bounded. Furthermore

$$\begin{aligned} a_{n+1} \geq a_n &\Leftrightarrow \sqrt{2 + a_n} \geq a_n \\ &\Leftrightarrow 2 + a_n \geq a_n^2 \\ &\Leftrightarrow (a_n - 2)(a_n + 1) \leq 0, \end{aligned}$$

which holds for $a_n \in [0, 2]$. By induction we conclude that the sequence is increasing and bounded, hence convergent (to $a = 2$).

- **Series**: Put $s_n = \sum_{i=1}^{n} a_i$, the n^{th} *partial sum* of the *series* $\sum_{i=1}^{\infty} a_i$. We say that $\sum_{i=1}^{\infty} a_i = s$ if $s_n \longrightarrow s$.

  - **Example**: Geometric series $\sum_{i=0}^{\infty} x^i$ for $|x| < 1$. We have

$$s_n = \sum_{i=0}^{n} x^i = \frac{1 - x^{n+1}}{1 - x} \longrightarrow s = \frac{1}{1 - x}.$$

- Numerous tests of convergence are available - see any text (e.g. your MATH 214 text) to review

some of them. We have seen one of the most useful ones for sequences - 'monotonic + bounded'. Another is the 'ratio test'. For this, suppose that the terms in a series are all non-negative, and that for all sufficiently large n, say $n > N$, one has $(a_{n+1}/a_n) \leq r < 1$. Then $s_n \nearrow$ and

$$\begin{aligned} s_n &= \sum_{i=1}^{N} a_i + \sum_{i=N+1}^{n} a_i \\ &\leq \sum_{i=1}^{N} a_i + a_N \left(r + r^2 + \cdots + r^{n-N} \right) \\ &\leq \sum_{i=1}^{N} a_i + a_N \frac{r}{1-r}; \end{aligned}$$

hence the partial sums are monotonic and bounded.

- **Example**: Let X be a discrete r.v. with $P(X = x_n) = p_n$, $n = 0, 1, 2, \ldots$. If $\sum x_n^k p_n$ converges absolutely (i.e. $\sum \left| x_n^k p_n \right|$ converges; it can be shown that this implies convergence of $\sum x_n^k p_n$), we call it the k^{th} moment $E[X^k]$ of X. Thus $E[X^k]$ exists iff $E[|X|^k]$ exists.

Suppose X has the Poisson distribution $\mathbb{P}(\lambda)$:

$$P(X = n) = e^{-\lambda}\frac{\lambda^n}{n!},\ n = 0, 1, 2, \ldots\ .$$

Then the k^{th} moments exist for all $k > 0$. To see this, consider the partial sums

$$S_N = \sum_{n=0}^{N} n^k p_n = \sum_{n=0}^{N} n^k e^{-\lambda}\frac{\lambda^n}{n!} = \sum_{n=0}^{N} a_n, \text{ say.}$$

We must show that S_N converges; this follows from the ratio test:

$$\frac{a_{n+1}}{a_n} = \frac{\lambda}{n+1}\left(1 + \frac{1}{n}\right)^k \to 0 \text{ as } n \to \infty;$$

now choose any positive $r < 1$ – there is N such that

$$n > N \Rightarrow \frac{a_{n+1}}{a_n} \leq r.$$

- Another common test of convergence is the comparison test - if $|a_n| \leq |b_n|$ (eventually!), and if $\sum |b_n|$ converges, then so does $\sum |a_n|$. (Lab problem).

22. Sequences/series of functions; Power series

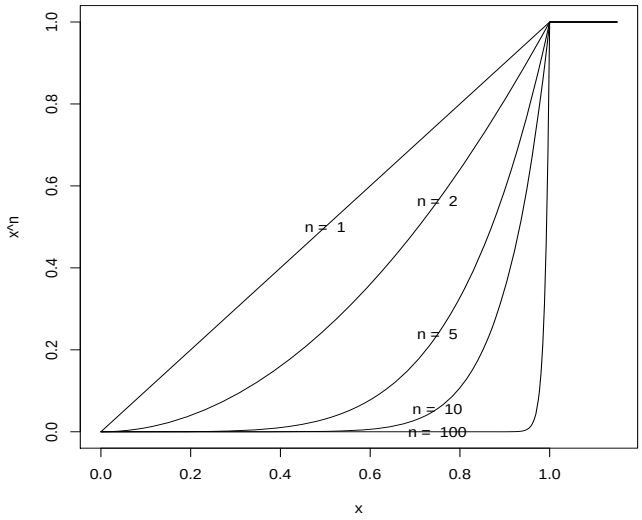
- Here we extend these convergence notions to functions. Suppose $\{f_n\}_{n=1}^{\infty}$ are real-valued functions, each defined on a domain $D \subset \mathbb{R}$. If the sequence $f_n(x)$ has a limit for every $x \in D$, denoted $f(x)$, then we say that $f_n \to f$ on D. Formally, for each $x \in D$,

$$\forall \varepsilon > 0 \exists N = N(\varepsilon, x)\, (n > N \Rightarrow |f_n(x) - f(x)| < \varepsilon)\,. \tag{22.1}$$

- Similarly, consider $s_n(x) = \sum_{i=1}^{n} f_i(x)$. If the sequence $s_n(x) \to s(x)$ for $x \in D$ we say that $s(x) = \sum_{i=1}^{\infty} f_i(x)$ and that $\sum_{i=1}^{\infty} f_i(x)$ *converges to* $s(x)$.

- **Uniform convergence**. If, in (22.1), the same $N = N(\varepsilon)$ works for all $x \in D$ we say the convergence is *uniform* on D: $f_n \rightrightarrows f$ on D. Equivalently,

$$f_n \rightrightarrows f \text{ on } D \Leftrightarrow \sup_{x \in D} |f_n(x) - f(x)| \to 0.$$



Example of non-uniformity of convergence:

$$f_n(x) = \begin{cases} x^n, & 0 \le x < 1, \\ 1, & x \ge 1 \end{cases}$$

$$\to \begin{cases} 0, & 0 \le x < 1, \\ 1, & x \ge 1 \end{cases} = f(x).$$

Then for each n,

$$\sup_{[0,\infty)} |f_n(x) - f(x)| \ge \sup_{[0,1)} |x|^n = 1$$

so that $\sup_{[0,\infty)} |f_n(x) - f(x)| \not\to 0$. Below we give a useful example of a uniformly convergent series of functions.

- **Power series**: Put $s_n(x) = \sum_{i=0}^{n} a_i(x-c)^i$; if $s_n(x) \to s(x)$ we say that $\sum_{i=0}^{\infty} a_i(x-c)^i$ is the *power series representing* s.

- **Example**: By Taylor's Theorem, if

$$s_n(x) = \sum_{k=0}^{n} f^{(k)}(c)\frac{(x-c)^k}{k!}$$

then

$$f(x) = s_n(x) + f^{(n+1)}(\xi)\frac{(x-c)^{n+1}}{(n+1)!},$$

so that if

$$f^{(n+1)}(\xi)\frac{(x-c)^{n+1}}{(n+1)!} \to 0,$$

then $\sum_{k=0}^{\infty} f^{(k)}(c)\frac{(x-c)^k}{k!}$ is the power series ('Taylor series', or 'Maclaurin's series' if $c = 0$) 'representing f'.

- The series $\sum_{i=0}^{\infty} \frac{x^i}{i!}$ represents the function $s(x) = e^x$.

Proof: We showed earlier, using Taylor's Theorem, that there is ξ between 0 and x with:

$$\begin{aligned} s(x) &= \sum_{i=0}^{n} s^{(i)}(0)\frac{x^i}{i!} + s^{(n+1)}(\xi)\frac{x^{n+1}}{(n+1)!} \\ &= \sum_{i=0}^{n} \frac{x^i}{i!} + e^{\xi}\frac{x^{n+1}}{(n+1)!} \\ &= s_n(x) + r_n(x), \text{ say.} \end{aligned}$$

Then $|s(x) - s_n(x)| = |r_n(x)|$ and so $s(x) = \sum_{i=0}^{\infty} \frac{x^i}{i!}$, i.e. the series represents the function, if $|r_n(x)| \to 0$. We show the stronger result that, on any closed interval $[a, b]$, we have that $r_n \rightrightarrows 0$. Equivalently, $s_n \rightrightarrows s$ on $[a, b]$.

– We are to show that $\sup_{x\in[a,b]} |r_n(x)| \to 0$. For this, let M be any integer that exceeds both $|a|$ and $|b|$, hence exceeds $|x|$. Let $n > M$. Then

$$\sup_{x\in[a,b]} |r_n(x)| < e^{M}\frac{M^{n+1}}{(n+1)!}$$

and exactly as in Lecture 18, this $\to 0$ as $n \to \infty$.

- **Theorem**: Suppose a power series $\sum_{n=0}^{\infty} a_n x^n$ converges for one value $x_0 \neq 0$. Then it converges absolutely (i.e. $\sum_{n=0}^{\infty} |a_n x^n|$ converges) for $|x| < |\ x_0|$. (Proof omitted.)

- If $\sum_{n=0}^{\infty} a_n x^n$ converges for $|x| < \rho$ and diverges for $|x| > \rho$ we call ρ the *radius of convergence.*

- **Example**: Put

$$s_n(x) = \sum_{i=0}^{n} (-x)^i = \frac{(1-(-x)^{n+1})}{1+x},$$

then with $s(x) = 1/(1+x)$ we have

$$|s_n(x) - s(x)| = |x|^{n+1}/|1+x|.$$

If $|x| < 1$ then $|s_n(x) - s(x)| \to 0$; if $|x| > 1$ it $\to \infty$. Thus $\rho = 1$ is the radius of convergence. By the last Theorem, if $|x| < 1$ the series is absolutely convergent:

$$\sum_{i=0}^{n} |x|^i \to \frac{1}{1-|x|}.$$

In this case when $|x| = 1$ the series diverges (i.e. the partial sums do not converge to a finite limit).

23. Power series II; Probability generating functions

- Recall the notion of radius of convergence. We have:

  Theorem: Suppose a power series $\sum_{n=0}^{\infty} a_n x^n$ has a radius of convergence $\rho > 0$. Let $0 < r < \rho$. Then ($\sum_{n=0}^{\infty} |a_n| \, r^n$ converges and):
  (i) $\sum_{n=0}^{\infty} a_n x^n$ converges uniformly on $[-r, r]$;
  (ii) For $|x| < r$ the limit function $s(x) = \sum_{n=0}^{\infty} a_n x^n$ is continuous and differentiable, and the derivative is represented by the convergent series

$$s'(x) = \sum_{n=1}^{\infty} n a_n x^{n-1}.$$

  - The derived series $s'(x) = \sum_{n=1}^{\infty} n a_n x^{n-1}$ has a radius of convergence ρ (why? – let $|x| < \rho$, then there is r such that $|x| < r < \rho$ and (ii) states that $\sum_{n=1}^{\infty} n a_n x^{n-1}$ converges). Thus we can repeat the process:

$$s''(x) = \sum_{n=2}^{\infty} n(n-1) a_n x^{n-2},$$

etc. Among other things, this implies the uniqueness of power series representations (lab problem).

- A special type of series: the *probability generating function* of a r.v. X is the function $\phi(z) = E[z^X]$, provided this exists. In particular, if X has support $\mathbb{N} = \{0, 1, 2, ...\}$ (i.e. $P(X \in \mathbb{N}) = 1$) then

$$\phi(z) = E[z^X] = \sum_{n=0}^{\infty} z^n P(X = n).$$

Since this converges for $z = 1$ (why?) it has radius of convergence $\rho \geq 1$. We can then differentiate term-by-term near $z = 0$:

$$\begin{aligned} & \phi^{(k)}(0) \\ = & \sum_{n=k}^{\infty} n(n-1)\cdots(n-k+1) z^{n-k} P(X = n)_{|z=0} \\ = & k! P(X = k). \end{aligned}$$

  - Note that, by uniqueness of power series, if we can expand $\phi(z)$ as $\sum_{n=0}^{\infty} z^n p_n$ then, necessarily, $p_n = P(X = n) = \phi^{(n)}(0)/n!$. (So we don't necessarily have to compute $\phi^{(k)}(0)$.)

  - The p.g.f. uniquely determines the distribution: two r.v.s with the same p.g.f. have the same distribution.

- **Example 1**: If $X \sim \mathbb{P}(\lambda)$ then $\phi(z) = E[z^X] = \ldots = e^{-\lambda(1-z)}$.

- **Example 2**: In this example we use the fact that a characterization of the independence of r.v.s (X, Y) is that $E[f(X)g(Y)] = E[f(X)]E[g(Y)]$ for all functions f, g such that $f(X)$ and $g(Y)$ are also r.v.s. Equivalently, $f(X)$ and $g(Y)$ are uncorrelated for all such f, g.

  If $X \sim bin(n, p)$ then (how?)

$$\phi(z) = (1 - p + pz)^n .$$

  Thus if the X_i, $i = 1, ..., I$ are independent, $X_i \sim bin(n_i, p)$ and $X = \sum_i X_i$ we have

$$\begin{aligned} E\left[z^X\right] &= E[z^{\sum_i X_i}] = E\left[\prod_i z^{X_i}\right] = \prod_i E\left[z^{X_i}\right] \\ &= \prod_i (1-p+pz)^{n_i} = (1-p+pz)^{\sum n_i}, \\ &= \text{the } bin(\textstyle\sum n_i, p) \text{ p.g.f.} \end{aligned}$$

The uniqueness then shows that $X \sim bin(N,p)$ with $N = \sum_{i=1}^{I} n_i$ (as we might anticipate?).

Example 3: Recall the hyperbolic functions

$$\cosh(z) = \frac{e^z + e^{-z}}{2}, \sinh(z) = \frac{e^z - e^{-z}}{2}.$$

A r.v. X has p.g.f. $\phi(z) = \cosh(z)/\cosh(1)$. What is the probability distribution?

Solution: Expand the exponentials to obtain

$$\phi(z) = \sum_{j=0}^{\infty} z^j c_j, \text{ with } c_j = \begin{cases} 0, & j \text{ odd}, \\ \frac{1}{j!\cosh(1)}, & j \text{ even}. \end{cases}$$

But also $\phi(z) = E\left[z^X\right] = \sum_{j=0}^{\infty} z^j P(X=j)$, so $P(X=j) = (j!\cosh(1))^{-1}$ if j is even, $= 0$ otherwise. [(i) Does $\sum P(X=j) = 1$? (ii) For practice, try $\phi(z) = \sinh(z)/\sinh(1)$.]

24. Moment generating functions I

- The *moment generating function* of a r.v. X is the function $\psi(t) = E[e^{tX}]$, provided this exists (i.e. is finite) in an open neighbourhood of 0 (i.e. for $|t| < \delta$ and some $\delta > 0$). (Replacing t by it gives the *characteristic function*, which always exists: it is $E[\cos(tX)] + iE[\sin(tX)]$.) If $P(X \in \mathbb{N}) = 1$,

$$\psi(t) = \sum_{n=0}^{\infty} e^{tn} P(X = n).$$

Note that $\psi(t) = \phi(e^t)$, so that it converges (absolutely) in a neighbourhood of $t = 0$ iff ϕ has a radius of convergence $\rho > 1$. Assume this. Then for $|t| < \log \rho$ we have, by the preceding theorem,

$$\begin{aligned} \psi'(t) &= \phi'(e^t)e^t \\ &= \sum_{n=0}^{\infty} n \left(e^t\right)^{n-1} P(X = n) \cdot e^t \\ &= \sum_{n=0}^{\infty} n e^{tn} P(X = n) \\ &= E[X e^{tX}], \end{aligned}$$

with $\psi'(0) = E[X]$. Continuing, $\psi^{(k)}(t) = E[X^k e^{tX}]$ with

$$\psi^{(k)}(0) = E[X^k].$$

(i.e. we can differentiate within the $E[\cdot]$.)

– e.g. $X \sim \mathbb{P}(\lambda)$ with $P(X = n) = e^{-\lambda}\frac{\lambda^n}{n!}$ has

$$\begin{aligned}\psi(t) &= \sum_{n=0}^{\infty} e^{tn} e^{-\lambda}\frac{\lambda^n}{n!} = e^{-\lambda}\sum_{n=0}^{\infty}\frac{\left(\lambda e^t\right)^n}{n!} \\ &= e^{-\lambda}\cdot e^{\lambda e^t} = e^{\lambda(e^t-1)}.\end{aligned}$$

Thus

$$\begin{aligned}E[X] &= \psi'(0) = \lambda, \\ E[X^2] &= \psi''(0) = \lambda^2 + \lambda, \text{ hence} \\ \text{var}[X] &= \lambda.\end{aligned}$$

– The *cumulants* κ_j of a distribution are defined as the coefficients κ_j in the expansion of the 'cumulant generating function' (c.g.f.)

$$\xi(t) = \log E[e^{tX}] = \sum_{j=1}^{\infty}\kappa_j\frac{t^j}{j!}.$$

Thus (how?) the Poisson distribution has all cumulants $= \lambda$. In general κ_1 is the mean:

$$\kappa_1 = \xi'(0) = \frac{\psi'(0)}{\psi(0)} = E[X]$$

and κ_2 is the variance (lab problem); after that they get more complicated. We will later see that the Normal distribution has all $\kappa_j = 0$ for $j > 2$.

- **Example**: Define a random variable N as the number of 'failures' which occur before the r^{th} 'success', in a sequence of trials with only these two possible outcomes. If the probability of a success on each trial is p, then the event '$N = n$' occurs if and only if (i) there are $r-1$ successes and n failures in the first $n+r-1$ trials, then (ii) one more success. The first of these events is given by the binomial probability distribution, and the second has probability p and is independent of the first. Thus

$$\begin{aligned} P(N &= n) = \binom{n+r-1}{r-1} p^{r-1}(1-p)^n \cdot p \\ &= \binom{n+r-1}{r-1} p^r (1-p)^n, \ n = 0, 1, 2, \ldots . \end{aligned}$$

This is the Negative Binomial distribution; we write $N \sim NB(r,p)$. What is the m.g.f.? First, since $\sum_n P(N=n) = 1$, we have the following useful identity for any $q \in [0,1)$:

$$\frac{1}{(1-q)^r} = \sum_{n=0}^{\infty} \binom{n+r-1}{r-1} q^n. \tag{24.1}$$

Now

$$\begin{aligned} E[e^{tN}] &= \sum_{n=0}^{\infty} e^{tn} \binom{n+r-1}{r-1} p^r (1-p)^n \\ &= p^r \sum_{n=0}^{\infty} \binom{n+r-1}{r-1} \left((1-p)e^t\right)^n \\ &= \left(\frac{p}{1-(1-p)e^t}\right)^r \end{aligned}$$

by (24.1), as long as $0 \le q = (1-p)e^t < 1$, and in particular if $|t| < -\ln(1-p)$.

The mean number of failures before the r^{th} success is

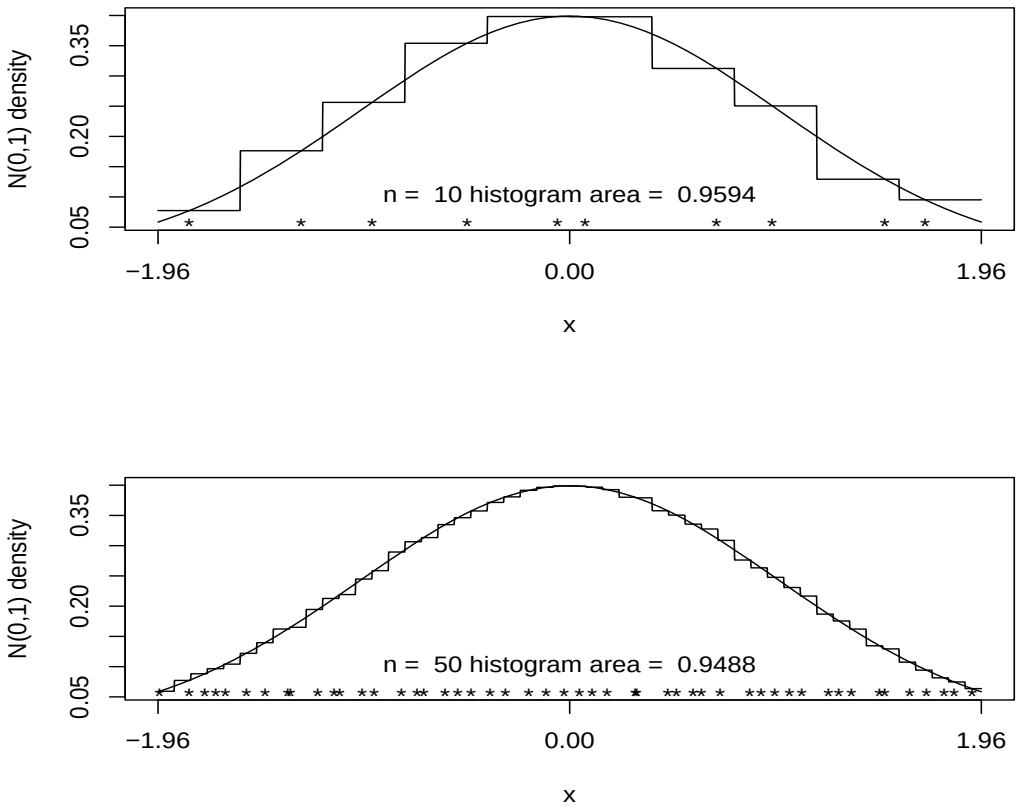
$$\frac{d}{dt} E[e^{tN}]_{|t=0} = r\frac{1-p}{p}.$$

25. Riemann Integration I

- **Riemann integration**. First consider $f : [a, b] \to \mathbb{R}$, a *bounded* function. Consider the approximating histogram ("P"), with breaks at $\{a = x_0 < x_1 < \cdots < x_n = b\}$ and heights $\{f(t_1), ..., f(t_n)\}$, where t_i is any point in $[x_{i-1}, x_i]$. A first approximation to the area under f is the 'Riemann sum'

$$S_P(f) = \sum_{i=1}^{n} f(t_i)\Delta_i,$$

where $\Delta_i = x_i - x_{i-1}$. Now let the *norm* $\Delta_P = \max_i(\Delta_i)$ shrink to 0. If $S_P(f)$ has a limit as we do this, we call this limit the Riemann integral of f, and say that f is Riemann-integrable on $[a, b]$ ('$f \in RI[a, b]$').



Approximations of area under the Normal density $\phi(x)$ between -1.96 and 1.96 using Riemann sums (= areas under histograms); target value = .95. Each of the n points t_i ("*") was randomly chosen in $[x_{i-1}, x_i]$; these intervals each have width $\Delta_i = (2 \cdot 1.96/n)$ and then $S_P(f) = \sum \phi(t_i)\,\Delta_i = 3.92 \cdot \frac{1}{n} \sum_{i=1}^{n} \phi(t_i)$.

- It follows from the definitions that continuous functions on $[a, b]$ are R-integrable there, as are bounded, monotonic functions. All the usual rules from first-year calculus follow too: if $f, g \in RI[a, b]$ then so are $f + g$, af, fg and $|f|$; in the first two cases the integral is linear; in the last we have $\left|\int_a^b f(x)dx\right| \leq \int_a^b |f(x)|dx$. If $f \leq g$ then $\int_a^b f(x)dx \leq \int_a^b g(x)dx$. If $c \in [a, b]$ then $\int_a^b f(x)dx = \int_a^c f(x)dx + \int_c^b f(x)dx$.

- An important result is the *Mean Value Theorem for Riemann integrals*: If f is continuous on $[a, b]$ then there is $c \in [a, b]$ for which

$$\int_a^b f(x)dx = f(c)(b - a).$$

Proof: Let m and M be the inf and sup of f on $[a, b]$, then

$$m \leq \frac{1}{b-a} \int_a^b f(x)dx \leq M.$$

Since f is continuous it attains m, M and every point between ('Intermediate Value Theorem'), hence there is $c \in [a, b]$ for which $f(c) = \frac{1}{b-a} \int_a^b f(x)dx$.

- One consequence: $\int_a^a f(x)dx = 0$.

- Now define

$$F(x) = \int_a^x f(t)dt,\ a \leq x \leq b,$$

the *indefinite integral* of f. We have the *Fundamental Theorem of Calculus*: If f is continuous on $[a, b]$ then F is differentiable there, with $F'(x) = f(x)$ (we call F an 'antiderivative').
Proof:

$$\begin{aligned} F'(x) &= \lim_{h \to 0} \frac{1}{h} \left[\int_a^{x+h} f(t)dt - \int_a^x f(t)dt \right] \\ &= \lim_{h \to 0} \frac{1}{h} \int_x^{x+h} f(t)dt \\ &= \lim_{h \to 0} \frac{1}{h} \cdot hf(c_h) \end{aligned}$$

with $c_h \in [x, x+h]$, by the previous MVT. Since $c_h \to x$ and f is continuous, $f(c_h) \to f(x)$. $\square$

By definition, $\int_a^b f(t)dt = F(b) = F(b) - F(a)$; if also $G'(x) = f(x)$ for $x \in [a, b]$ then

$$\int_a^b f(x)dx = G(b) - G(a).$$

Why? Define $H(x) = (G(x) - G(a)) - (F(x) - F(a))$; note that $H(a) = 0$ and $H'(x) \equiv 0$; thus (how?) $H(b) = 0$.

- **Example 1**: $f(x) = x$, $G(x) = x^2/2$ has $G'(x) = f(x)$, hence $\int_a^b f(t)dt = G(b) - G(a) = \left(b^2 - a^2\right)/2$.

- **Example 2**: the substitution $x = \tan\theta$, with $(dx/d\theta) = \sec^2\theta$, gives

$$\int_a^b \frac{1}{1+x^2}dx = \int_{\arctan a}^{\arctan b} \cos^2\theta \sec^2\theta d\theta$$
$$= \theta \Big|_{\arctan a}^{\arctan b} = \arctan b - \arctan a.$$

- A useful alternate form of Taylor's Theorem, with the **remainder in integral form**: Suppose that $f(x)$ is n times continuously differentiable on $[a, b]$. Then

$$f(b) = \sum_{k=0}^{n-1} f^{(k)}(a)\frac{(b-a)^k}{k!} + \frac{1}{(n-1)!}\int_a^b (b-x)^{n-1} f^{(n)}(x)dx.$$

Proof: Put $h_n(x) = f\ (b) - \sum_{k=0}^{n-1} f^{(k)}(x)\frac{(b-x)^k}{k!}$ and note that $h_n(b) = 0$. By the Fundamental Theorem of Calculus, $h_n(a) = \int_a^b \left[-h_n'(x)\right] dx$, i.e.

$$f(b) = \sum_{k=0}^{n-1} f^{(k)}(a)\frac{(b-a)^k}{k!} + \int_a^b \left[-h_n'(x)\right] dx.$$

When $h_n'(x)$ is calculated almost all of the terms cancel each other out; the result is

$$\left[-h_n'(x)\right] = \frac{(b-x)^{n-1}}{(n-1)!} f^{(n)}(x),$$

as required. □

26. Riemann Integration II

- **Example**: Recall Taylor's Theorem with the remainder in integral form. Suppose we wish to approximate the cube root of $b = 9$ in terms of that of $a = 8$; we will do this by expanding $f(x) = x^{1/3}$ around a. How close do we get if we stop after the quadratic term?

$$\begin{aligned} f(b) &= f(a) + f'(a)(b-a) + f''(a)\frac{(b-a)^2}{2!} \\ &\quad + \frac{1}{2!}\int_a^b (b-x)^2 f'''(x)dx. \end{aligned}$$

Here $f(a) = 2$, $f'(a) = 1/12$, $f''(a) = -1/144$:

$$9^{1/3} \approx 2 + \frac{1}{12} - \frac{1}{288} = 2.07986.$$

With a calculator, $\sqrt[3]{9} = 2.08008$ with an error of .00022; the estimate of the error using Taylor's Theorem is $R = \frac{1}{2}\int_8^9 (9-x)^2 f'''(x)dx$, with (lab question)

$$|R| \leq \frac{5}{20,736} = .00024. \qquad (26.1)$$

- *Improper* Riemann integrals, in which one or both endpoints are infinite, or at which f is unbounded, are defined by taking appropriate limits:

$$\begin{aligned}
\int_a^\infty f(x)dx &= \lim_{b\to\infty}\int_a^b f(x)dx, \\
\int_{-\infty}^\infty f(x)dx &= \int_{-\infty}^a f(x)dx + \int_a^\infty f(x)dx \text{ for any } a, \\
\int_a^b f(x)dx &= \lim_{\varepsilon\downarrow 0}\int_a^{b-\varepsilon} f(x)dx \text{ if } f(b) = \pm\infty.
\end{aligned}$$

- An application of the Fundamental Theorem of Calculus is the formula for integration by parts. If f, g are differentiable, and $f'g$, fg' are integrable, then

$$\begin{aligned}
\int_a^b [f(x)g(x)]'\,dx &= f(b)g(b) - f(a)g(a) \text{ and also} \\
&= \int_a^b \Big[f'(x)g(x) + f(x)g'x)\Big]\,dx;
\end{aligned}$$

hence

$$\int_a^b f'(x)g(x)dx = f(b)g(b) - f(a)g(a) - \int_a^b f(x)g'(x)dx.$$

- **Example**. Define $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}dx$, $(\alpha > 0)$, the Gamma integral.

  - Existence? Here is an outline: If $\alpha \geq 1$ we need only show that $F(T) = \int_0^T x^{\alpha-1}e^{-x}dx$ has a finite limit as $T \to \infty$; since $F(T)$ is increasing it is enough for it to be bounded. But for large enough x we have that $x^{\alpha-1}e^{-x} < e^{-x/2}$, whose integral is bounded. If $\alpha < 1$ ...

  - Evaluation:

$$\begin{aligned}\Gamma(\alpha) &= \int_0^\infty \left(\frac{x^\alpha}{\alpha}\right)' e^{-x}dx \\ &= \left(\frac{x^\alpha}{\alpha}\right)e^{-x}\Big|_0^\infty - \int_0^\infty \left(\frac{x^\alpha}{\alpha}\right)\frac{d}{dx}e^{-x}dx \\ &= \frac{1}{\alpha}\int_0^\infty x^\alpha e^{-x}dx = \frac{1}{\alpha}\Gamma(\alpha+1),\end{aligned}$$

    hence $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$.

    In particular for n an integer,

    $\Gamma(n+1) = n\Gamma(n) = \cdots = n(n-1)\cdots 1\cdot\Gamma(1) = n!$.

  - Lab problem: $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$.

- If X has distribution function F with density f, then $F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt$ and (Fundamental Theorem of Calculus) $F'(x) = f(x)$. **Example**: Let ϕ and Φ represent the $N\left(\mu, \sigma^2\right)$ density and distribution:

$$\phi\left(x; \mu, \sigma^2\right) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} = \Phi'\left(x; \mu, \sigma^2\right).$$

But the only 'antiderivative' is the indefinite integral itself:

$$\Phi\left(x; \mu, \sigma^2\right) = \int_{-\infty}^{x} \phi\left(t; \mu, \sigma^2\right) dt.$$

- If X has d.f. F with a density f, the m.g.f. is

$$\psi(t) = E[e^{tX}] = \int_{-\infty}^{\infty} e^{tx} f(x)dx,$$

provided this exists in some neighbourhood $|t| < \delta$ (for some $\delta > 0$). Some useful properties (shared by the m.g.f. for a discrete r.v.):

1. $\psi^{(k)}(0) = E[X^k]$ (so if the m.g.f. exists, so do all moments). In other words we can interchange the differentiation and integration.

Then if we can find an expansion of the form $\psi(t) = \sum \frac{\mu_k}{k!} t^k$, by the uniqueness of power series this must be the MacLaurin series, and so we must have $\mu_k = \psi^{(k)}(0) = E[X^k]$.

2. If $\psi_X(t) = \psi_Y(t)$ for all $|t| < \delta$ then $X \sim Y$, i.e. the distribution of a r.v. is uniquely determined by the m.g.f.

3. Sums of independent r.v.s. If $X_1, X_2, ...$ are independent r.v.s with m.g.f.s $\psi_1(t), \psi_2(t), ...$ and if $S_n = \sum_{i=1}^n X_i$, then

$$\begin{aligned} \psi_{S_n}(t) &= E\left[e^{t\sum_{i=1}^n X_i}\right] = E\left[\prod_{i=1}^n e^{tX_i}\right] \\ &= \prod_{i=1}^n E\left[e^{tX_i}\right] = \prod_{i=1}^n \psi_i(t). \end{aligned}$$

In particular, if all X_i are distributed in the same way, with m.g.f. $\psi(t)$, then the m.g.f. of their sum is $\psi_{S_n}(t) = \psi^n(t)$ and the m.g.f. of their average is

$$\psi_{\bar{X}}(t) = E\left[e^{tS_n/n}\right] = \psi_{S_n}(t/n) = \psi^n(t/n).$$

We will use this in proving the CLT.

4. If $\{X_n\}$ is a sequence of r.v.s with m.g.f.s $\psi_n(t)$, and if $\psi_n(t) \to \psi(t)$ for all t in a neighbourhood of 0, where $\psi(t)$ is the m.g.f. of a r.v. X, then $X_n \xrightarrow{L} X$:

$$P(X_n \leq x) \to P(X \leq x).$$

27. Moment generating functions II

- **Example 1**: If $X_n \sim bin(n, p_n)$ with $p_n \to 0$ and $np_n \to \mu > 0$ (as $n \to \infty$) then $X_n \xrightarrow{L} \mathbb{P}(\mu)$ (Poisson, mean μ). (For this reason the Poisson is sometimes known as the distribution of the number of 'rare events'.)

  Proof: First calculate

$$\begin{aligned}\psi_{X_n}(t) &= \sum_x e^{tx}\binom{n}{x}p_n^x\,(1-p_n)^{n-x} \\ &= \left(1-p_n+p_ne^t\right)^n \text{ (how?).}\end{aligned}$$

  We aim to show (why?) that this $\to \exp\left\{\mu\left(e^t-1\right)\right\}$ as $n \to \infty$. Take logs:

$$\begin{aligned} & n\log\left(1-p_n+p_ne^t\right) \\ = \;& n\log\left(1+p_n\left(e^t-1\right)\right) \\ = \;& \frac{\log\left(1+p_n\left(e^t-1\right)\right)}{p_n\left(e^t-1\right)}\cdot np_n\left(e^t-1\right) \\ \to \;& \lim_{x\to 0}\frac{\log(1+x)}{x}\cdot\lim_{np_n\to\mu} np_n\left(e^t-1\right) \\ = \;& \mu\left(e^t-1\right).\end{aligned}$$

- **Example 2**: Suppose $X_1, ..., X_n$ are i.i.d., each exponentially distributed with density $f(x) = \lambda e^{-\lambda x}$, $x > 0$. The m.g.f. of each is

$$M(t) = \int_0^\infty e^{tx} \lambda e^{-\lambda x} dx = \cdots = \left(1 - \frac{t}{\lambda}\right)^{-1}$$

for $|t| < \lambda$. This density is often used to model the time to the occurrence of an event like a random shock (such as might cause an electrical component to fail), and so the time to the n^{th} such shock is $S_n = \sum_{i=1}^n X_i$. We will see later how to obtain the density of S_n (by induction for instance); it will turn out to be the 'Erlang' density

$$f_n(s) = \frac{(\lambda s)^{n-1}}{(n-1)!} \lambda e^{-\lambda s}, s > 0.$$

Now we have an easy proof of this:

$$\begin{aligned} E\left[e^{tS_n}\right] &= \left(1 - \frac{t}{\lambda}\right)^{-n}, \\ \int_0^\infty e^{ts} f_n(s)\, ds &= \cdots = \left(1 - \frac{t}{\lambda}\right)^{-n}. \end{aligned}$$

And this finishes it - why?

- **Example 3**: What are the moments of S_n in the previous example? We know that if we can expand the m.g.f. as

$$\psi_{S_n}(t) = \sum_{k=0}^{\infty} \frac{\mu_k}{k!} t^k,$$

then $\mu_k = E\left[S_n^k\right]$. Using (24.1) we get, for $|t| < \lambda$,

$$\psi_{S_n}(t) = \sum_{k=0}^{\infty} \binom{k+n-1}{n-1} \left(\frac{t}{\lambda}\right)^k,$$

whence

$$\mu_k = k! \binom{k+n-1}{n-1} \lambda^{-k} = \frac{(k+n-1)!}{(n-1)!} \lambda^{-k}.$$

- **Example 4**: If $X \sim N\left(\mu, \sigma^2\right)$ then (lab problem)

$$M_X(t) = E\left[e^{tX}\right] = e^{\mu t + \frac{\sigma^2 t^2}{2}}.$$

If $X_1, ..., X_n \overset{ind.}{\sim} N\left(\mu_i, \sigma_i^2\right)$ and $\{a_i\}$ are constants, then $S = \sum_{i=1}^n a_i X_i$ has

$$E\left[e^{tS}\right] = \cdots = e^{\alpha t + \frac{\beta^2 t^2}{2}},$$

for $\alpha = \sum_{i=1}^{n} a_i \mu_i$ and $\beta^2 = \sum_{i=1}^{n} a_i^2 \sigma_i^2$. From this,

$$\sum_{i=1}^{n} a_i X_i \sim N\left(\sum_{i=1}^{n} a_i \mu_i, \sum_{i=1}^{n} a_i^2 \sigma_i^2\right).$$

In particular, $aX \sim N\left(a\mu, a^2\sigma^2\right)$.

- **Example 5**: With S_n as in Example 2,

  $$\left(S_n - n/\lambda\right)/\sqrt{n} = \sqrt{n}\left(\bar{X} - \mu_X\right) \xrightarrow{L} N\left(0, \sigma_X^2 = \lambda^{-2}\right),$$

  by the CLT. Alternate proof using m.g.f.s:

  $$\log E\left[e^{t\frac{S_n - n/\lambda}{\sqrt{n}}}\right] = -n \log\left(1 - \frac{t}{\lambda\sqrt{n}}\right) - \frac{t\sqrt{n}}{\lambda} \to \frac{t^2}{2\lambda^2},$$

  the $N\left(0, \sigma_X^2 = \lambda^{-2}\right)$ c.g.f.

Part III

ASYMPTOTICS; OPTIMALITY

28. Cauchy-Schwarz, Chebyshev, WLLN

- **Cauchy-Schwarz inequality** for r.v.s: If X and Y are any r.v.s with finite second moments, then

$$(E[XY])^2 \leq E\left[X^2\right] E\left[Y^2\right].$$

Proof: Essentially identical to the vector version:

$$\begin{aligned} 0 &\leq E\left[(X+\lambda Y)^2\right] \\ &= \lambda^2 E\left[Y^2\right] + 2\lambda E[XY] + E\left[X^2\right], \end{aligned}$$

hence '$B^2 - 4AC$' ≤ 0, i.e.

$$4(E[XY])^2 - 4E\left[X^2\right] E\left[Y^2\right] \leq 0.$$

[Equality iff $E\left[(X+\lambda_0 Y)^2\right] = 0$ for some λ_0 ... iff Y is a multiple of X.] □

  - **Example**: $\text{var}[X] \geq 0$.
  - **Example**: If $\rho = \text{corr}[X, Y]$, then $\rho^2 \leq 1$. (Equality iff ...)
  - **Example**: $E[X^3] \leq \sqrt{E[X^2]E[X^4]}$.

- **Chebyshev's Inequality**: If a r.v. X has mean μ and variance σ^2 then

$$P(|X-\mu| \geq k\sigma) \leq 1/k^2.$$

Proof: We use the following device. Let F be any event and define the *indicator* of F by

$$I(F) = \begin{cases} 1, & \text{if } F \text{ occurs,} \\ 0, & \text{otherwise.} \end{cases}$$

Note that

$$\begin{aligned} E\left[I(F)\right] &= 1 \cdot P(I=1) + 0 \cdot P(I=0) \\ &= P(I=1) \\ &= P(F). \end{aligned}$$

Now put $Z = (X-\mu)/\sigma$, then Z has mean 0 and variance 1 and we are to show that $P(|Z| \geq k) \leq 1/k^2$. But

$$Z^2 \geq k^2 I\left(|Z| \geq k\right),$$

(why?) so that

$$\begin{aligned} 1 &= \text{var}[Z] = E\left[Z^2\right] \\ &\geq k^2 E\left[I\left(|Z| \geq k\right)\right] \\ &= k^2 P(|Z| \geq k). \end{aligned}$$

□

- Chebyshev's Inequality furnishes an easy proof of the **Weak Law of Large Numbers**: *If $\bar{X}_n$ is the average of n independent r.v.s, each with mean μ and variance σ^2, then $\bar{X}_n \xrightarrow{pr} \mu$ as $n \to \infty$.*
  Proof: Recall (Lab 1) that $\bar{X}_n$ has mean μ and variance σ^2/n. Represent $\varepsilon > 0$ as $\varepsilon = k\sigma\left(\bar{X}_n\right)$ by setting $k = \varepsilon\sqrt{n}/\sigma$; then

$$\begin{aligned} P\left(\left|\bar{X}_n - \mu\right| \geq \varepsilon\right) &= P\left(\left|\bar{X}_n - \mu\right| \geq k\sigma\left(\bar{X}_n\right)\right) \\ &\leq \frac{1}{k^2} = \frac{\sigma^2}{n\varepsilon^2} \\ &\to 0 \text{ as } n \to \infty. \end{aligned}$$

□

- **Central Limit Theorem**. This is probably the most significant theorem in mathematical statistics. It gives the approximate normality of averages of r.v.s and, when combined with the MVT (or Taylor's Theorem), the WLLN and Slutsky's Theorem (next lecture), forms the basis for approximating the distributions of many other statistics of interest.

- **Theorem**: Let $X_1, X_2, ..., X_n$ be independent r.v.s, with common d.f. $F(x) = P(X_i \leq x)$, mean μ, variance σ^2 ($0 < \sigma^2 < \infty$). Put

$$Z_n = \sqrt{n}\left(\bar{X}_n - \mu\right) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - \mu)\,;$$

then Z_n has mean zero and variance σ^2, AND $Z_n \xrightarrow{L} N(0, \sigma^2)$.

  - To apply, since the statements '$Z_n \sim N(0, \sigma^2)$' and '$\bar{X}_n \sim N(\mu, \sigma^2/n)$' are equivalent, we treat $\bar{X}_n$ as if it were distributed approximately as $N(\mu, \sigma^2/n)$. Then, e.g. if we can also estimate σ^2, we have the basis for making inferences about μ ('t-test').

- The proof of the CLT will utilize the fact (lab problem) that the m.g.f. of $Z \sim N(\mu, \sigma^2)$ is

$$E[e^{tZ}] = e^{\mu t + \frac{\sigma^2 t^2}{2}}.$$

A consequence is a fact noted earlier - all cumulants of the Normal distribution, after the second, are zero.

29. Central Limit Theorem

- **Proof of CLT**: We make the additional assumption that the X_i have an m.g.f. Define

$$\psi(t) = E[e^{t(X_i-\mu)}] \ (= e^{-t\mu}E[e^{tX_i}]).$$

We are to show that the m.g.f. of

$$Z_n = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(X_i - \mu)$$

tends to that of a $N(0, \sigma^2)$ r.v., i.e. that

$$\begin{aligned} E\left[e^{t\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(X_i-\mu)}\right] &= \prod_{i=1}^{n} E\left[e^{\frac{t}{\sqrt{n}}(X_i-\mu)}\right] \\ &= \psi^n(t/\sqrt{n}) \end{aligned}$$

tends to $e^{\frac{\sigma^2 t^2}{2}}$. Equivalently, we show that

$$n \log \psi(t/\sqrt{n}) \to \frac{\sigma^2 t^2}{2} \text{ as } n \to \infty.$$

We use the following notation: '$f(x) = o(g(x))$ as $x \to L$' means '$f(x)/g(x) \to 0$ as $x \to L$'. As an example, by l'Hospital's Rule,

$$\log(1 + x) = x + o(x) \text{ as } x \to 0.$$

Let t be fixed but arbitrary. Expand $\psi(t)$ as

$$\begin{aligned}\psi(t) &= \psi(0) + \psi'(0)t + \psi''(0)\frac{t^2}{2} + \psi'''(\xi_t)\frac{t^3}{6}, \\ &\quad (0 \le |\xi_t| \le |t|) \\ &= 1 + E[X-\mu]t + E[(X-\mu)^2]\frac{t^2}{2} + \psi'''(\xi_t)\frac{t^3}{6} \\ &= 1 + \frac{\sigma^2 t^2}{2} + o(t^2) \text{ as } t \to 0.\end{aligned}$$

Why $o(t^2)$? - because $\psi'''(\xi_t)$ has a finite limit as t, hence ξ_t, tends to 0. Now

$$n \log \psi(t/\sqrt{n}) = n \log \left(1 + \underbrace{\frac{\sigma^2 t^2}{2n} + o\left(\frac{t^2}{n}\right)}_{x_n} \right)$$

$$= n(x_n + o(x_n)), \text{ where } x_n = \frac{\sigma^2 t^2}{2n} + o\left(\frac{t^2}{n}\right).$$

As $n \to \infty$, so that $\frac{t^2}{n} \to 0$,

$$nx_n = \frac{\sigma^2 t^2}{2} + \frac{o\left(\frac{t^2}{n}\right)}{\frac{t^2}{n}} t^2 \to \frac{\sigma^2 t^2}{2};$$

In particular, $x_n \to 0$ and

$$no(x_n) = nx_n \frac{o(x_n)}{x_n} \to \frac{\sigma^2 t^2}{2} \cdot 0 = 0.$$

Thus $n \log \psi(t/\sqrt{n}) \to \frac{\sigma^2 t^2}{2}$, as required. □

- **Slutsky's Theorem**: If $X_n \xrightarrow{L} X$ and $Y_n \xrightarrow{pr} c$ (constant) then:

  1. $X_n \pm Y_n \xrightarrow{L} X \pm c$,

  2. $X_n \cdot Y_n \xrightarrow{L} X \cdot c$,

  3. $X_n / Y_n \xrightarrow{L} X/c$ if $c \neq 0$.

If $X = d$ (constant) then all occurrences of $\xrightarrow{L}$ can be replaced by $\xrightarrow{pr}$. (In this case (1) was proven in Lab 6.)

- **Application**: We often make inferences about a population mean μ using the t-statistic

$$t = \frac{\sqrt{n}\left(\bar{X} - \mu\right)}{S},$$

where $\bar{X}$ is as in the CLT and S is the sample standard deviation. If the data are normally distributed then t follows a 'Student's t' distribution on $n-1$ degrees of freedom; it is well known that this distribution is closely approximated by the $N(0,1)$ when n is reasonably large. *This latter fact holds even for non-normal parent distributions*:

The WLLN + several applications of Slutsky's Theorem yield (lab problem) $S^2 \xrightarrow{pr} \sigma^2$; thus $S \xrightarrow{pr} \sigma$ (since S is a continuous function of S^2) and so $S/\sigma \xrightarrow{pr} 1$ (for the same reason). Now again by Slutsky, and the CLT,

$$t = \frac{\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma}}{\frac{S}{\sigma}} \xrightarrow{L} \frac{Z}{1} = Z \sim N(0,1).$$

30. Multidimensional calculus

- $\mathbf{f} : D \subset \mathbb{R}^n \to \mathbb{R}^m$ can be represented as

$$\begin{pmatrix} f_1(\mathbf{x}) \\ f_2(\mathbf{x}) \\ \vdots \\ f_m(\mathbf{x}) \end{pmatrix} \text{ for } f_i(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R}.$$

- **Partial derivatives**. Let $f : D \subset \mathbb{R}^n \to \mathbb{R}^1$. The *partial derivative* $\partial f(\mathbf{x})/\partial x_i$ is the ordinary derivative with respect to x_i, treating all others as constants. E.g. $\partial\left(x_1^2 + x_2^2\right)/\partial x_1 = 2x_1$.

- The *Jacobian* matrix is the $m \times n$ matrix $\mathbf{J_f}(\mathbf{x}) = \left(\frac{\partial \mathbf{f}}{\partial \mathbf{x}}\right)$ with $(i,j)^{th}$ element $\partial f_i/\partial x_j$, evaluated at $\mathbf{x} = (x_1, ..., x_n)$. This arrangement of partial derivatives ensures that the chain rule is easily represented: if $\mathbf{x}_{n\times 1} \xrightarrow{\mathbf{f}} \mathbf{y}_{m\times 1} \xrightarrow{\mathbf{g}} \mathbf{z}_{p\times 1}$ then (writing $\left(\frac{\partial \mathbf{y}}{\partial \mathbf{x}}\right)$ for $\mathbf{J_f}(\mathbf{x})$ etc.)

$$\left(\frac{\partial \mathbf{z}}{\partial \mathbf{x}}\right)_{p\times n} = \left(\frac{\partial \mathbf{z}}{\partial \mathbf{y}}\right)_{p\times m} \left(\frac{\partial \mathbf{y}}{\partial \mathbf{x}}\right)_{m\times n}. \quad (30.1)$$

- The Jacobian matrix of $f : \mathbb{R}^n \to \mathbb{R}$ is a row vector whose transpose is the *gradient*:

$$\boldsymbol{\nabla}_f(\mathbf{x}) = (\partial f/\partial x_1, \cdots, \partial f/\partial x_n)' .$$

- The Jacobian of $\boldsymbol{\nabla}_f : \mathbb{R}^n \to \mathbb{R}^n$ is called the *Hessian* of $f : \mathbb{R}^n \to \mathbb{R}$. This $n \times n$ matrix $\mathbf{H}_f(\mathbf{x})$ has $(i,j)^{th}$ element $\frac{\partial (\boldsymbol{\nabla}_f)_i}{\partial x_j} = \frac{\partial}{\partial x_j}\frac{\partial f}{\partial x_i}$. In most cases these differentiations can be carried out in either order, so that the Hessian is a *symmetric* matrix.

- If $f : D \subset \mathbb{R}^n \to \mathbb{R}$ then the *directional derivative* at $\mathbf{a}$ in the direction $\mathbf{v}$ (with $\|\mathbf{v}\| = 1$) is

$$\lim_{h \to 0} \frac{f(\mathbf{a} + h\mathbf{v}) - f(\mathbf{a})}{h} = \boldsymbol{\nabla}'_{\mathbf{f}}(\mathbf{a})\mathbf{v}.$$

Proof: Using (30.1), the limit is

$$\begin{aligned} \frac{\partial f(\mathbf{a} + h\mathbf{v})}{\partial h}_{|h=0} &= \frac{\partial f(\mathbf{a} + h\mathbf{v})}{\partial (\mathbf{a} + h\mathbf{v})} \frac{\partial (\mathbf{a} + h\mathbf{v})}{\partial h}_{|h=0} \\ &= \boldsymbol{\nabla}'_{\mathbf{f}}(\mathbf{a} + h\mathbf{v})\mathbf{v}_{|h=0}. \end{aligned}$$

- We say $\mathbf{a}$ is a 'stationary point' of $f : \mathbb{R}^n \to \mathbb{R}$ if $\boldsymbol{\nabla}_{\mathbf{f}}(\mathbf{a}) = \mathbf{0}$; this means the directional derivative is 0 in <u>all</u> directions $\mathbf{v}$.

- **Integration** in $\mathbb{R}^n$: If $f : \mathbb{R}^n \to \mathbb{R}$ then if $D \subset \mathbb{R}^n$ we write $\int_D f(\mathbf{x})\, d\mathbf{x}$ to mean the 'volume' of the region 'under' the graph of the function $f(\mathbf{x})$ for $\mathbf{x} \in D$. It is generally evaluated as an n-fold iterated integral, e.g. $\int \{\int f(x_1, x_2)\, dx_1\}\, dx_2$.

- **Change of variables**. Suppose we want to integrate $f(\mathbf{x})$, but first change variables to $\mathbf{y} = \mathbf{h}(\mathbf{x})$. Then (assuming the Jacobian is non-singular)

$$\int_D f(\mathbf{x}) d\mathbf{x} = \int_{\mathbf{h}(D)} f(\mathbf{x}(\mathbf{y})) \left| \left(\frac{\partial \mathbf{x}}{\partial \mathbf{y}} \right) \right|_+ d\mathbf{y},$$

where $|\cdot|_+$ denotes the absolute value of the determinant.

- **Example**: A problem on Lab 10 uses a transformation to polar coordinates. There $D = \mathbb{R}^2$, and the (inverse) transformation to $(y_1, y_2) = (\rho, \theta)$ is

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} \rho \cos\theta \\ \rho \sin\theta \end{pmatrix},$$

with $\theta \in [0, 2\pi)$ and $\rho > 0$ (defining $\mathbf{h}(D)$). Then

$$\int_{\mathbb{R}^2} f(x_1, x_2) d\mathbf{x} = \int_0^\infty \int_0^{2\pi} f(\rho\cos\theta, \rho\sin\theta)\, \rho d\theta d\rho,$$

since

$$\left|\frac{\partial \mathbf{x}}{\partial \mathbf{y}}\right|_+ = abs\left(\det\begin{pmatrix} \cos\theta & -\rho\sin\theta \\ \sin\theta & \rho\cos\theta \end{pmatrix}\right) = \rho.$$

31. Extrema, Lagrange multipliers

- Extrema of $f : D \subset \mathbb{R}^n \to \mathbb{R}$. There is a multivariate version of Taylor's Theorem, by which we have the expansion

$$f(\mathbf{x}_0 + \mathbf{v}) = f(\mathbf{x}_0) + \boldsymbol{\nabla}'_f(\mathbf{x}_0)\mathbf{v} + \frac{1}{2}\mathbf{v}'\mathbf{H}_f(\xi)\mathbf{v}.$$

for some ξ between $\mathbf{x}_0$ and $\mathbf{x}_0 + \mathbf{v}$. Let $\mathbf{x}_0$ be a stationary point – $\boldsymbol{\nabla}_f(\mathbf{x}_0) = \mathbf{0}$ – so that the directional derivative $\boldsymbol{\nabla}'_f(\mathbf{x}_0)\mathbf{v}$ is 0 in <u>any</u> direction $\mathbf{v}$. Then:

1. If $\mathbf{H}_f(\xi) > \mathbf{0}$ for ξ in a neighbourhood of $\mathbf{x}_0$ then $\mathbf{x}_0$ furnishes a local minimum of f: $f(\mathbf{x}_0+\mathbf{v}) > f(\mathbf{x}_0)$ in this neighbourhood (i.e. for sufficiently small $\mathbf{v} \neq \mathbf{0}$).

2. If $\mathbf{H}_f(\xi) < \mathbf{0}$ for ξ in a neighbourhood of $\mathbf{x}_0$ then $\mathbf{x}_0$ furnishes a local maximum of f: $f(\mathbf{x}_0+\mathbf{v}) < f(\mathbf{x}_0)$ for sufficiently small $\mathbf{v} \neq \mathbf{0}$.

3. If neither (1) nor (2) holds then $f(\mathbf{x}_0 + \mathbf{v}) - f(\mathbf{x}_0)$ changes sign as $\mathbf{v}$ varies; we say that $\mathbf{x}_0$ is a *saddlepoint*.

- Often we seek extrema of multivariate functions, subject to certain side conditions ('constraints'). For instance in #3 of Lab 11 one seeks to minimize the variance of the return on a portfolio of investments subject to a fixed value of the expected return. The general problem considered here is to find the extrema of $f : D \subset \mathbb{R}^n \to \mathbb{R}$ subject to $\mathbf{g}(\mathbf{x}) = \mathbf{0}_{m\times 1}$ for $m < n$. E.g.

  P : Minimize $\mathbf{x}'\mathbf{A}\mathbf{x}$ subject to $\mathbf{B}\mathbf{x} = \mathbf{c}_{m\times 1}$, where $\mathbf{A} > \mathbf{0}$ and $\mathbf{B}_{m\times n}$ has rank $m < n$.

  Put

$$F(\mathbf{x}; \boldsymbol{\lambda}) = f(\mathbf{x}) + \boldsymbol{\lambda}'\mathbf{g}(\mathbf{x})$$

  for a vector $\boldsymbol{\lambda}_{m\times 1}$ of 'Lagrange multipliers'.
  Claim (proof omitted): The stationary points of F that satisfy the constraints determine the stationary points in the original problem. These points then satisfy the $m+n$ equations in the $m+n$ variables of $(\mathbf{x}, \boldsymbol{\lambda})$:

$$\boldsymbol{\nabla}_F(\mathbf{x}; \boldsymbol{\lambda}) = \mathbf{0}_{(m+n)\times 1}.$$

Equivalently (how?),

$$\begin{aligned} \boldsymbol{\nabla}_f(\mathbf{x}) + \mathbf{J}'_{\mathbf{g}}(\mathbf{x})\boldsymbol{\lambda} &= \mathbf{0}_{n\times 1}, \\ \mathbf{g}(\mathbf{x}) &= \mathbf{0}_{m\times 1}. \end{aligned}$$

- **Example**: Problem **P** above. We have

$$\begin{aligned} F(\mathbf{x};\boldsymbol{\lambda}) &= f(\mathbf{x}) + \boldsymbol{\lambda}'\mathbf{g}(\mathbf{x}) \\ &= \mathbf{x}'\mathbf{A}\mathbf{x} + \boldsymbol{\lambda}'(\mathbf{B}\mathbf{x} - \mathbf{c}) \end{aligned}$$

with (lab problem)

$$\mathbf{0}_{n\times 1} = \left(\frac{\partial F}{\partial \mathbf{x}}\right)' = 2\mathbf{A}\mathbf{x} + \mathbf{B}'\boldsymbol{\lambda},$$

implying

$$\mathbf{x} = -\frac{1}{2}\mathbf{A}^{-1}\mathbf{B}'\boldsymbol{\lambda}.$$

Combine this with $\mathbf{B}\mathbf{x} = \mathbf{c}$ to get

$$\boldsymbol{\lambda} = -2\left(\mathbf{B}\mathbf{A}^{-1}\mathbf{B}'\right)^{-1}\mathbf{c},$$

whence

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{B}'\left(\mathbf{B}\mathbf{A}^{-1}\mathbf{B}'\right)^{-1}\mathbf{c}.$$

(That $\mathbf{B}\mathbf{A}^{-1}\mathbf{B}'$ is non-singular was 'possible midterm exam question' #17.)

- The virtue of the Lagrange multiplier method is that it reduces to a small number the points that must be checked – we know that the required extrema are among them. Once we have these stationary points we must check for a minimum or maximum. An easy way to do this (if it works) is as follows. Suppose that $(\mathbf{x}_0; \boldsymbol{\lambda}_0)$ is a stationary point of $F(\mathbf{x}; \boldsymbol{\lambda})$ and that $\mathbf{x}_0$ minimizes $F(\mathbf{x}; \boldsymbol{\lambda}_0)$ *unconditionally*, i.e. that $F(\mathbf{x}_0; \boldsymbol{\lambda}_0) < F(\mathbf{x}; \boldsymbol{\lambda}_0)$ for *all* $\mathbf{x} \neq \mathbf{x}_0$. Then in the class of those $\mathbf{x}$ that *do* satisfy $\mathbf{g}(\mathbf{x}) = \mathbf{0}$ (in this class $F(\mathbf{x}; \boldsymbol{\lambda}_0) = f(\mathbf{x})$) we have

$$f(\mathbf{x}_0) = F(\mathbf{x}_0; \boldsymbol{\lambda}_0) < F(\mathbf{x}; \boldsymbol{\lambda}_0) = f(\mathbf{x}).$$

  - In Problem **P** there was only one stationary point $\mathbf{x}_0$. This furnishes the minimum since

$$F(\mathbf{x}; \boldsymbol{\lambda}_0) = \mathbf{x}'\mathbf{A}\mathbf{x} + \boldsymbol{\lambda}_0' (\mathbf{B}\mathbf{x} - \mathbf{c}),$$

    where $\boldsymbol{\lambda}_0 = -2 \left(\mathbf{B}\mathbf{A}^{-1}\mathbf{B}'\right)^{-1} \mathbf{c}$, has (lab problem) Hessian $2\mathbf{A} > \mathbf{0}$ (<u>everywhere</u> $> \mathbf{0}$), hence is minimized unconditionally at the stationary point $\mathbf{x}_0$.

32. Normal sampling distributions

- Change of variables in multivariate densities. Recall that if $Y = h(X)$, where X has a density $f(x)$ and h is invertible (i.e. increasing or decreasing, so $h' \neq 0$), then the density $g(\cdot)$ of Y is given by

$$g(y) = f(x) \left| \frac{dx}{dy} \right|,$$

with the rhs evaluated at $x = h^{-1}(y)$. There is a multivariate analogue.

- Suppose f is the p.d.f. of a r.vec. $\mathbf{X}_{n\times 1}$ (i.e. $P(\mathbf{X} \in D) = \int_D f(\mathbf{x}) d\mathbf{x}$). Put $\mathbf{Y}_{n\times 1} = \mathbf{h}(\mathbf{X})$, where $\mathbf{J_h}(\mathbf{x})$ is nonsingular. Then the p.d.f. g of $\mathbf{Y}$ is given by

$$g(\mathbf{y}) = f(\mathbf{x}(\mathbf{y})) \left| \left(\frac{\partial \mathbf{x}}{\partial \mathbf{y}} \right) \right|_+, \text{ with } \mathbf{x} = \mathbf{x}(\mathbf{y}).$$

It is sometimes easier to evaluate

$$\left(\frac{\partial \mathbf{x}}{\partial \mathbf{y}} \right)^{-1} = \left(\frac{\partial \mathbf{y}}{\partial \mathbf{x}} \right) = \mathbf{J_h}(\mathbf{x}).$$

The following result, on the joint distribution of the sample mean and variance in Normal samples, is of fundamental importance in Statistics.
Suppose that $X_1, ..., X_n$ are i.i.d. $N(\mu, \sigma^2)$ r.v.s, so that $\mathbf{X} = (X_1, ..., X_n)'$ has p.d.f.

$$\begin{aligned} f(\mathbf{x}) &= \prod_{i=1}^{n} \left\{ \left(2\pi\sigma^2\right)^{-1/2} \exp\left\{ -\frac{(x_i-\mu)^2}{2\sigma^2} \right\} \right\} \\ &= \left(2\pi\sigma^2\right)^{-n/2} \exp\left\{ -\sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2} \right\}. \end{aligned}$$

Note

$$\begin{aligned} \sum_{i=1}^{n} (x_i-\mu)^2 &= \sum_{i=1}^{n} \left[(x_i-\bar{x}) + (\bar{x}-\mu) \right]^2 \\ &= (n-1)s^2 + n\left(\bar{x}-\mu\right)^2, \end{aligned}$$

so that

$$f(\mathbf{x}) = \left(2\pi\sigma^2\right)^{-n/2} e^{-\frac{(n-1)s^2}{2\sigma^2}} e^{-\frac{n(\bar{x}-\mu)^2}{2\sigma^2}}.$$

We derive the joint p.d.f. of $(S^2, \bar{X})$. First note that $\mathbf{1}_{n\times 1}/\sqrt{n}$ has norm 1. Adjoin $n-1$ unit vectors $\mathbf{e}_i$ to get a basis for $\mathbb{R}^n$, and then apply Gram-Schmidt to get an orthonormal basis whose first member is $1/\sqrt{n}$.

This yields an orthogonal matrix $\mathbf{H}_{n\times n} = \begin{pmatrix} \mathbf{1}'/\sqrt{n} \\ \mathbf{H}_1 \end{pmatrix}$.
Put $\mathbf{Y} = \mathbf{HX}$. Then

$$\begin{aligned} Y_1 &= \mathbf{1}'\mathbf{X}/\sqrt{n} = \sqrt{n}\bar{X}; \text{ and} \\ \sum_{i=2}^{n} Y_i^2 &= \|\mathbf{Y}\|^2 - Y_1^2 \\ &= \|\mathbf{X}\|^2 - \left(\sqrt{n}\bar{X}\right)^2 \\ &= \sum_{i=1}^{n} X_i^2 - n\bar{X}^2 \\ &= (n-1)S^2. \end{aligned}$$

Note that $\left|\frac{\partial \mathbf{x}}{\partial \mathbf{y}}\right|_+ = |\mathbf{H}'|_+ = |\pm 1| = 1$, so that the p.d.f. $g(\mathbf{y})$ of $\mathbf{Y}$ is

$$f(\mathbf{x}(\mathbf{y}))\left|\frac{\partial \mathbf{x}}{\partial \mathbf{y}}\right|_+ = f(\mathbf{x}(\mathbf{y}))$$

$$\begin{aligned} &= \left(2\pi\sigma^2\right)^{-n/2} e^{-\frac{\sum_{i=2}^{n} y_i^2}{2\sigma^2}} e^{-\frac{(y_1-\sqrt{n}\mu)^2}{2\sigma^2}} \\ &= \left\{\left(2\pi\sigma^2\right)^{-1/2} e^{-\frac{(y_1-\sqrt{n}\mu)^2}{2\sigma^2}}\right\} \cdot \prod_{i=2}^{n}\left\{\left(2\pi\sigma^2\right)^{-1/2} e^{-\frac{y_i^2}{2\sigma^2}}\right\}. \end{aligned}$$

From this factorization we see that

1. $Y_1, ..., Y_n$ are independently distributed;

2. $Y_1 \sim N(\sqrt{n}\mu, \sigma^2)$, so that $\bar{X} \sim N(\mu, \sigma^2/n)$;

3. $\frac{(n-1)S^2}{\sigma^2} = \sum_{i=2}^{n} \left(\frac{Y_i}{\sigma}\right)^2 \sim \chi^2_{n-1}$, since $\frac{Y_i}{\sigma} \sim i.i.d.$ $N(0,1)$; furthermore $\bar{X}$ and S^2 are independently distributed.

- **Example**: 'Student's t_{n-1}' is $T = \frac{(\bar{X}-\mu)}{S/\sqrt{n}} \sim Z \Big/ \sqrt{\frac{\chi^2_{n-1}}{n-1}}$, where $Z \sim N(0,1)$, with d.f. $\Phi(z)$, independently of the χ^2 with density $f_{\chi^2_{n-1}}(s)$ (see #4 on Lab 11). It follows that the d.f. is (by Double Expectation Theorem + independence)

$$\begin{aligned} G_{n-1}(t) &= P(T \le t) = E[I(Z \le t\sqrt{\chi^2_{n-1}/(n-1)})] \\ &= E_{\chi^2_{n-1}} E\left[I\left(Z \le t\sqrt{\frac{s}{n-1}}\right) \Big| \chi^2_{n-1} = s\right] \\ &= \int_0^\infty \Phi\left(t\sqrt{\frac{s}{n-1}}\right) f_{\chi^2_{n-1}}(s)\, ds, \end{aligned}$$

from which the density can be obtained.

33. Maximum likelihood I: Estimation

Maximum Likelihood Estimation. This is the most common and versatile method of estimation in statistics. It almost always gives reasonable estimates, even in situations that are so intractable as to be highly resistant to other estimation methods.

- Data $\mathbf{x}$, p.d.f. $f(\mathbf{x}; \boldsymbol{\theta}_{p\times 1})$; e.g. i.i.d. $N(\mu, \sigma^2)$ observations gives

$$\begin{aligned} \boldsymbol{\theta} &= (\mu, \sigma^2)' \\ f(\mathbf{x}; \boldsymbol{\theta}) &= \prod_{i=1}^{n} \frac{1}{\sigma} \phi\left(\frac{x_i - \mu}{\sigma}\right) \\ &= \left(2\pi\sigma^2\right)^{-n/2} e^{-\frac{1}{2\sigma^2}\sum (x_i-\mu)^2}. \end{aligned}$$

  The p.d.f. evaluated at the data is the *likelihood function* $L(\boldsymbol{\theta}; \mathbf{x})$; its logarithm

$$l(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta}; \mathbf{x})$$

  is the *log-likelihood*.

- For i.i.d. observations with common p.d.f. or p.m.f. $p\left(x;\boldsymbol{\theta}_{p\times 1}\right)$ we have

$$\begin{aligned} f\left(\mathbf{x};\boldsymbol{\theta}\right) &= \prod_{i=1}^{n} p\left(x_i;\boldsymbol{\theta}\right), \text{ so} \\ l\left(\boldsymbol{\theta}\right) &= \sum_{i=1}^{n} \log p\left(x_i;\boldsymbol{\theta}\right). \end{aligned}$$

  Viewed as a r.v., $l\left(\boldsymbol{\theta}\right) = \sum_{i=1}^{n} \log p\left(X_i;\boldsymbol{\theta}\right)$ is itself a sum of i.i.d.s.

- The MLE $\hat{\boldsymbol{\theta}}$ is the maximizer of the likelihood; intuitively it makes the observed data 'most likely to have occurred'.

- A more quantitative justification for the MLE is as follows. Let $\boldsymbol{\theta}_0$ be the true value, and assume the X_i are i.i.d. We will show that

$$P_{\boldsymbol{\theta}_0}\left(L(\boldsymbol{\theta}_0;\mathbf{X}) > L(\boldsymbol{\theta};\mathbf{X})\right) \to 1 \qquad (*)$$

  as $n \to \infty$, for any $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$.

By this, for large samples and with high probability, the (random) likelihood is maximized by the true parameter value, hence the maximizer of the (observed) likelihood should be a good estimate of this true value.

Proof of (*): The inequality

$$L(\boldsymbol{\theta}_0; \mathbf{X}) = \prod_{i=1}^{n} p\left(X_i; \boldsymbol{\theta}_0\right) > \prod_{i=1}^{n} p\left(X_i; \boldsymbol{\theta}\right) = L(\boldsymbol{\theta}; \mathbf{X})$$

is equivalent to

$$-\frac{1}{n} \sum_{i=1}^{n} \log \frac{p\left(X_i; \boldsymbol{\theta}\right)}{p\left(X_i; \boldsymbol{\theta}_0\right)} > 0.$$

By the WLLN this average tends in probability to

$$\begin{aligned}
& E_{\boldsymbol{\theta}_0}\left[-\log \frac{p\left(X; \boldsymbol{\theta}\right)}{p\left(X; \boldsymbol{\theta}_0\right)}\right] \\
> \; & -\log E_{\boldsymbol{\theta}_0}\left[\frac{p\left(X; \boldsymbol{\theta}\right)}{p\left(X; \boldsymbol{\theta}_0\right)}\right] \quad \text{(why?)} \\
= \; & -\log \int \frac{p\left(x; \boldsymbol{\theta}\right)}{p\left(x; \boldsymbol{\theta}_0\right)} p\left(x; \boldsymbol{\theta}_0\right) dx \\
= \; & -\log \int p\left(x; \boldsymbol{\theta}\right) dx \\
= \; & 0.
\end{aligned}$$

How is the proof finished?

- The MLE is generally obtained as a root of the *likelihood equation*

$$\dot{l}(\boldsymbol{\theta}) = \mathbf{0},$$

where $\dot{l}(\boldsymbol{\theta}) = \boldsymbol{\nabla}_l(\boldsymbol{\theta})$ denotes the gradient. Under reasonable conditions we have that the roots $\hat{\boldsymbol{\theta}}_n$ of the likelihood equation (based on n observations) are asymptotically normal:

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\right) \xrightarrow{L} N(0, \mathbf{I}^{-1}(\boldsymbol{\theta})),$$

where (for i.i.d. data) the 'Fisher's Information matrix' is defined as

$$\mathbf{I}(\boldsymbol{\theta}) = \text{cov}\left[\left(\frac{\partial}{\partial \boldsymbol{\theta}} \log p\left(X; \boldsymbol{\theta}\right)\right)'\right] \tag{33.1}$$

The practical interpretation is that

$$\hat{\boldsymbol{\theta}}_n \overset{d}{\approx} N\left(\boldsymbol{\theta}, \text{cov} = \frac{1}{n}\mathbf{I}^{-1}(\boldsymbol{\theta})\right),$$

i.e. with I^{jk} representing the $(j,k)^{th}$ element of $\mathbf{I}^{-1}(\boldsymbol{\theta})$ (or of $\mathbf{I}^{-1}(\hat{\boldsymbol{\theta}}_n)$) we have the approximations

$$\hat{\theta}_j \overset{d}{\approx} N(\theta_j, \frac{I^{jj}}{n}), \text{cov}[\hat{\theta}_j, \hat{\theta}_k] \approx \frac{I^{jk}}{n}.$$

34. Maximum likelihood II: Optimality

- **Example**: In $N(\mu, \sigma^2)$ samples with both parameters unknown we have (lab problem)

$$\begin{aligned} l(\boldsymbol{\theta}) &= \sum_{i=1}^{n} \log p(x_i; \boldsymbol{\theta}) \\ &= -\frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum (x_i - \mu)^2 + \text{constant}, \end{aligned}$$

and this is maximized by $\hat{\boldsymbol{\theta}} = \left(\bar{x}, \frac{1}{n}\sum (x_i - \bar{x})^2\right)$, i.e. $\hat{\mu} = \bar{x}$, $\widehat{\sigma^2} = \frac{n-1}{n} S^2$.

- The MLE has attractive large-sample optimality properties. Recall that

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\right) \xrightarrow{L} N(0, \mathbf{I}^{-1}(\boldsymbol{\theta})).$$

Suppose we aim to estimate a scalar function $\tau(\boldsymbol{\theta})$ (e.g. cv$= \sigma/\mu$). The MLE $\hat{\tau}$ is defined to be $\tau(\hat{\boldsymbol{\theta}})$, where $\hat{\boldsymbol{\theta}}$ is the MLE for $\boldsymbol{\theta}$. Recall that in studying the delta method (Lecture 20) we noted that in the single-parameter case, if $\hat{\theta}$ were asymptotically normal with mean θ then so would

be $\tau(\hat{\theta})$, with a mean of $\tau(\theta)$ and a variance of $[\tau'(\theta)]^2 \cdot \left(\text{variance of } \hat{\theta}\right)$. The multi-parameter analogue for the MLE is

$$\sqrt{n}\left(\tau\left(\hat{\boldsymbol{\theta}}_n\right) - \tau\left(\boldsymbol{\theta}\right)\right) \xrightarrow{L} N(0, \dot{\tau}'\left(\boldsymbol{\theta}\right)\mathbf{I}^{-1}(\boldsymbol{\theta})\dot{\tau}\left(\boldsymbol{\theta}\right)),$$

where $\dot{\tau}\left(\boldsymbol{\theta}\right) = \boldsymbol{\nabla}_{\tau}\left(\boldsymbol{\theta}\right)$ is the gradient. We call $\dot{\tau}'\left(\boldsymbol{\theta}\right)\mathbf{I}^{-1}(\boldsymbol{\theta})\dot{\tau}\left(\boldsymbol{\theta}\right)/n$ the *asymptotic variance*.

- Under some mild 'regularity conditions', **no unbiased estimator of $\tau(\boldsymbol{\theta})$ can have a variance smaller than $\dot{\tau}'\left(\boldsymbol{\theta}\right)\mathbf{I}^{-1}(\boldsymbol{\theta})\dot{\tau}\left(\boldsymbol{\theta}\right)/n$.**

- Here is a verification of this for just one parameter θ and i.i.d. observations with density $p\left(x;\theta\right)$. Assume that we can differentiate the equation

$$1 = \int_{-\infty}^{\infty} p\left(x;\theta\right) dx \tag{34.1}$$

under the integral sign, so that

$$\begin{aligned} 0 &= \frac{\partial}{\partial\theta}\int p\left(x;\theta\right) dx = \int \frac{\partial}{\partial\theta} p\left(x;\theta\right) dx \\ &= \int \frac{\frac{\partial}{\partial\theta} p\left(x;\theta\right)}{p\left(x;\theta\right)} p\left(x;\theta\right) dx = E\left[\frac{\partial}{\partial\theta}\log p\left(X;\theta\right)\right]; \end{aligned}$$

i.e. the r.v. $\frac{\partial}{\partial\theta}\log p\left(X;\theta\right)$ has a mean of zero. Recall (33.1) - this r.v. has a variance of $I\left(\theta\right)$. Thus

$$\dot{l}(\theta)=\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log p\left(X_i;\theta\right)$$

has a mean of zero and a variance of $nI\left(\theta\right)$.

- Now suppose that a statistic $T\left(\mathbf{x}\right)$ – i.e. a function of the data $\mathbf{x}=\left(X_1,...,X_n\right)$ – is unbiased for a function $\tau\left(\theta\right)$. Then

$$\tau\left(\theta\right)=E\left[T\left(\mathbf{x}\right)\right]=\int T\left(\mathbf{x}\right)f\left(\mathbf{x};\theta\right)d\mathbf{x}$$

and under the same assumption as before,

$$\begin{aligned}\dot{\tau}\left(\theta\right) &= \frac{\partial}{\partial\theta}\int T\left(\mathbf{x}\right)f\left(\mathbf{x};\theta\right)d\mathbf{x}=\int T\left(\mathbf{x}\right)\frac{\partial}{\partial\theta}f\left(\mathbf{x};\theta\right)d\mathbf{x}\\ &= \int T\left(\mathbf{x}\right)\frac{\frac{\partial}{\partial\theta}f\left(\mathbf{x};\theta\right)}{f\left(\mathbf{x};\theta\right)}f\left(\mathbf{x};\theta\right)d\mathbf{x}.\end{aligned}$$

But $\frac{\partial}{\partial\theta}f\left(\mathbf{x};\theta\right)/f\left(\mathbf{x};\theta\right)=\cdots=\dot{l}(\theta)$, so we have

$$\dot{\tau}\left(\theta\right)=E\left[T\left(\mathbf{x}\right)\dot{l}(\theta)\right]\overset{why?}{=}\text{cov}\left[T\left(\mathbf{x}\right),\dot{l}(\theta)\right].$$

By the Cauchy-Schwarz Inequality (in the form 'corr$^2 \leq 1$') we have

$$\begin{aligned}
\left[\dot{\tau}\left(\theta\right)\right]^2 &= \left(\text{cov}\left[T\left(\mathbf{x}\right), \dot{l}(\theta)\right]\right)^2 \\
&\leq \text{var}\left[T\left(\mathbf{x}\right)\right] \text{var}\left[\dot{l}(\theta)\right] \\
&= \text{var}\left[T\left(\mathbf{x}\right)\right] \cdot nI\left(\theta\right);
\end{aligned}$$

hence the variance of any unbiased estimator $T\left(\mathbf{x}\right)$ satisfies

$$\text{var}\left[T\left(\mathbf{x}\right)\right] \geq \frac{\left[\dot{\tau}\left(\theta\right)\right]^2}{nI\left(\theta\right)} = \text{asym. var. of } \tau\left(\hat{\theta}\right),$$

as required. This is the 'Information Inequality'.

- An often easier way to get $\mathbf{I}(\boldsymbol{\theta})$: If (34.1) – or its multivariate analogue – is differentiated again, one obtains

$$\mathbf{I}(\boldsymbol{\theta}) = E\left[-\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}} \log p\left(X; \boldsymbol{\theta}\right)\right], \tag{34.2}$$

the expectation of the negative Hessian of $\log p\left(X; \boldsymbol{\theta}\right)$. [Or $E\left[-\ddot{l}(\boldsymbol{\theta})\right]/n$; in one dimension this is $E\left[-l''\left(\theta\right)\right]/n$.]

35. Numerical optimization I: Newton-Raphson

- Numerical minimization. Suppose that a function $S : \mathbb{R}^p \to \mathbb{R}$ is to be minimized.

  - **Example**: Nonlinear regression. We observe Y_i, with mean $f(\mathbf{x}_i, \boldsymbol{\theta})$ depending on known regressors $\mathbf{x}_i$ and unknown parameters $\boldsymbol{\theta}_{p\times 1}$, and variance σ^2. If $f(\mathbf{x}_i, \boldsymbol{\theta}) = \mathbf{x}_i'\boldsymbol{\theta}$, this is just the *linear* regression model. The least squares estimates of $\boldsymbol{\theta}$ are the minimizers of $\sum (y_i - f(\mathbf{x}_i, \boldsymbol{\theta}))^2$, i.e. of

    $$S(\boldsymbol{\theta}) = \|\mathbf{y} - \boldsymbol{\eta}(\boldsymbol{\theta})\|^2,$$

    where $\boldsymbol{\eta}(\boldsymbol{\theta}) = (f(\mathbf{x}_1, \boldsymbol{\theta}), \cdots, f(\mathbf{x}_n, \boldsymbol{\theta}))'$.

  - A minimizer is a critical point of S – a solution to the p equations in p unknowns $\nabla_S(\boldsymbol{\theta}) = \mathbf{0}_{p\times 1}$ – and so we start with numerical root finders.

- **Newton-Raphson method**. We want a solution to $\mathbf{g}(\mathbf{x}_{p\times 1}) = \mathbf{0}_{p\times 1}$. Expand $\mathbf{g}(\mathbf{x})$ around an initial value $\mathbf{x}_0$:

$$\mathbf{g}(\mathbf{x}) \approx \mathbf{g}(\mathbf{x}_0) + \mathbf{J}_g(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0).$$

Equate the RHS to zero, to get the next iterate:

$$\mathbf{x}_1 = \mathbf{x}_0 - \mathbf{J}_g^{-1}(\mathbf{x}_0)\mathbf{g}(\mathbf{x}_0).$$

In general,

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \mathbf{J}_g^{-1}(\mathbf{x}_k)\mathbf{g}(\mathbf{x}_k),\ k = 0, 1, 2, \ldots .$$

At convergence, with $\mathbf{x}_\infty = \lim_{k\to\infty} \mathbf{x}_k$ and assuming that $\mathbf{J}_g(\mathbf{x}_\infty)$ is non-singular,

$$\mathbf{x}_\infty = \mathbf{x}_\infty - \mathbf{J}_g^{-1}(\mathbf{x}_\infty)\mathbf{g}(\mathbf{x}_\infty),$$

so that $\mathbf{g}(\mathbf{x}_\infty) = \mathbf{0}$.

  - If this is a minimization (of $S(\mathbf{x})$) problem, so that $\mathbf{g}(\mathbf{x}) = \boldsymbol{\nabla}_S(\mathbf{x})$, then $\mathbf{J}_g(\mathbf{x}) = \left(\frac{\partial \boldsymbol{\nabla}_S(\mathbf{x})}{\partial \mathbf{x}}\right) = \mathbf{H}_S(\mathbf{x})$ and the scheme is

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \mathbf{H}_S^{-1}(\mathbf{x}_k)\boldsymbol{\nabla}_S(\mathbf{x}_k). \tag{35.1}$$

Note

$$\begin{aligned} S(\mathbf{x}_{k+1}) &\approx S(\mathbf{x}_k) + \boldsymbol{\nabla}'_S(\mathbf{x}_k)\,(\mathbf{x}_{k+1} - \mathbf{x}_k) \\ &= S(\mathbf{x}_k) - \boldsymbol{\nabla}'_S(\mathbf{x}_k)\mathbf{H}_S^{-1}(\mathbf{x}_k)\boldsymbol{\nabla}_S(\mathbf{x}_k) \\ &< S(\mathbf{x}_k), \end{aligned}$$

if $\mathbf{H}_S(\mathbf{x}_k) > \mathbf{0}$.

- **Example:** solve $g(x) = \log x - 1 = 0$:

$$\begin{aligned} x_{k+1} &= x_k - \frac{g}{g'}(x_k) \\ &= x_k - \frac{\log x_k - 1}{1/x_k} \\ &= x_k\,(2 - \log x_k)\,, \;\; k = 0, 1, 2, \ldots\,. \end{aligned}$$

This gives

$$\begin{aligned} x_0 &= 1, \\ x_1 &= 2, \\ x_2 &= 2.6137, \\ x_3 &= 2.7162, \\ x_4 &= 2.7182811; \\ e &= 2.7182818.... \end{aligned}$$

- **Example:** We say θ is a 'location parameter' if the observations X_i have a density of the form $p(x_i - \theta)$ (as for the normal, logistic, etc.). For a sample from such a family the likelihood equation is

$$0 = \sum \frac{\partial}{\partial \theta} \log p(x_i - \theta) = \sum \psi(x_i - \theta),$$

where $\psi(t) = -p'(t)/p(t)$. More generally, even when ψ is not of this form, we call a root $\hat{\theta}$ of

$$0 = \sum \psi(x_i - \theta) \tag{35.2}$$

an 'M-estimate' of θ. (If $\psi(t) = t$ then $\hat{\theta} =$?). To solve by Newton's method:

$$\theta_{k+1} = \theta_k + \frac{\sum \psi(x_i - \theta_k)}{\sum \psi'(x_i - \theta_k)},$$

starting with some θ_0 (the sample average? sample median?).

- The mle $\bar{X}$ of a Normal mean is very 'non-robust', in that just one arbitrarily bad outlier can destroy its optimality. Since it is an M-estimate with $\psi(t) = t$, a robust alternative has been proposed:

$$\psi_c(t) = \begin{cases} t, & \text{if } |t| < c, \\ sign(t) \cdot c, & \text{if } |t| \geq c. \end{cases}$$

If $c = \infty$ this gives $\sum (x_i - \theta) = 0$, so $\hat{\theta} =$? To see what happens as $c \to 0$ write (35.2) as $0 = \sum \frac{\psi_c(x_i - \theta)}{c}$ and let $c \to 0$: in the limit (35.2) is

$$0 = \sum sign(x_i - \theta)$$

so $\hat{\theta} =$ what?

Usually c is chosen ≈ 1.5.

36. Numerical optimization II: Gauss-Newton

- Since least squares can be carried out so efficiently on computer packages, it is natural to attempt to adapt it more broadly than we have up to now. We have employed it in *linear* regression to minimize the sum of squares of errors (SSE)

$$S(\boldsymbol{\theta}) = \|\mathbf{y} - \boldsymbol{\eta}(\boldsymbol{\theta})\|^2, \tag{36.1}$$

where

$$\boldsymbol{\eta}(\boldsymbol{\theta}) = (f(\mathbf{x}_1, \boldsymbol{\theta}), \cdots, f(\mathbf{x}_n, \boldsymbol{\theta}))', \tag{36.2}$$

and $f(\mathbf{x}, \boldsymbol{\theta}) = \mathbf{x}'\boldsymbol{\theta}$ ($\nabla_f(\boldsymbol{\theta})$ doesn't depend on $\boldsymbol{\theta}$).

- The **Gauss-Newton algorithm** uses least squares minimization along with a *linear approximation* of the elements of $\boldsymbol{\eta}$. A common application is *non-linear regression* ($\nabla_f(\boldsymbol{\theta})$ depends on $\boldsymbol{\theta}$), so I'll illustrate the technique there. We observe

$$y_i = f(\mathbf{x}_i, \boldsymbol{\theta}) + \varepsilon_i,\ i = 1, ..., n.$$

Define $\boldsymbol{\eta}$ as at (36.2), then the lse of $\boldsymbol{\theta}$ is the minimizer of (36.1).

- An example is a *Michaelis-Menten* response $f(x, \boldsymbol{\theta}) = \theta_1 x/(\theta_2 + x)$, $(x > 0)$, used to describe various chemical and pharmacological reactions. Note the horizontal asymptote of θ_1; $x = \theta_2$ is the 'halfway point'.

- Take an initial value $\boldsymbol{\theta}_0$, expand f around $\boldsymbol{\theta}_0$ to get

  $$y_i - f(\mathbf{x}_i, \boldsymbol{\theta}) \approx y_i - f(\mathbf{x}_i, \boldsymbol{\theta}_0) - \boldsymbol{\nabla}'_f(\mathbf{x}_i, \boldsymbol{\theta}_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0).$$

  From this,

  $$\mathbf{y} - \boldsymbol{\eta}(\boldsymbol{\theta}) \approx \mathbf{y} - \boldsymbol{\eta}(\boldsymbol{\theta}_0) - \mathbf{J}_\eta(\boldsymbol{\theta}_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0).$$

  Define $\mathbf{y}_{(1)} = \mathbf{y} - \boldsymbol{\eta}(\boldsymbol{\theta}_0)$, so that

  $$\|\mathbf{y} - \boldsymbol{\eta}(\boldsymbol{\theta})\|^2 \approx \left\|\mathbf{y}_{(1)} - \mathbf{J}_\eta(\boldsymbol{\theta}_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0)\right\|^2$$

  is to be minimized. With $\boldsymbol{\beta} = \boldsymbol{\theta} - \boldsymbol{\theta}_0$ the rhs is the SSE $S(\boldsymbol{\beta})$ in the linear regression model

  $$\mathbf{y}_{(1)} = \mathbf{J}_\eta(\boldsymbol{\theta}_0)\boldsymbol{\beta} + error,$$

and so the minimizer is

$$\boldsymbol{\theta} - \boldsymbol{\theta}_0 = \hat{\boldsymbol{\beta}} = \left[\mathbf{J}'_\eta(\boldsymbol{\theta}_0)\mathbf{J}_\eta(\boldsymbol{\theta}_0)\right]^{-1}\mathbf{J}'_\eta(\boldsymbol{\theta}_0)\mathbf{y}_{(1)}. \tag{36.3}$$

Thus the next value is $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_0 + \hat{\boldsymbol{\beta}}$, i.e.

$$\boldsymbol{\theta}_1 = \boldsymbol{\theta}_0 + \left[\mathbf{J}'_\eta(\boldsymbol{\theta}_0)\mathbf{J}_\eta(\boldsymbol{\theta}_0)\right]^{-1}\mathbf{J}'_\eta(\boldsymbol{\theta}_0)\left(\mathbf{y} - \boldsymbol{\eta}(\boldsymbol{\theta}_0)\right).$$

In general,

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k + \left[\mathbf{J}'_\eta(\boldsymbol{\theta}_k)\mathbf{J}_\eta(\boldsymbol{\theta}_k)\right]^{-1}\mathbf{J}'_\eta(\boldsymbol{\theta}_k)\left(\mathbf{y} - \boldsymbol{\eta}(\boldsymbol{\theta}_k)\right).$$

Iterate to convergence.

Summary: We are repeatedly doing least squares regressions, in the $(k+1)^{th}$ of which the residuals from the k^{th} are regressed on the columns of the Jacobian matrix, evaluated at $\boldsymbol{\theta}_k$.

- A normal approximation is generally valid:

$$\hat{\boldsymbol{\theta}} \overset{d}{\approx} N\left(\boldsymbol{\theta},\, \sigma^2_\varepsilon\left[\mathbf{J}'_\eta(\hat{\boldsymbol{\theta}})\mathbf{J}_\eta(\hat{\boldsymbol{\theta}})\right]^{-1}\right).$$

Roughly, this arises because (36.3) is like the formula for the lse in a linear model, and with $\mathbf{J}_\eta$

playing the same role as the $\mathbf{X}$-matrix. We can use this approximation to get confidence intervals on the parameters.

- **Example**: In the Michaelis-Menten model

$$y = \frac{\theta_1 x}{\theta_2 + x} + \text{error},$$

$\mathbf{J}_\eta(\boldsymbol{\theta})$: $n \times 2$ has i^{th} row $\left(\frac{x_i}{\theta_2+x_i}, \frac{-\theta_1 x_i}{(\theta_2+x_i)^2}\right)$. Do linear regression iteratively, treating the columns of $\mathbf{J}_\eta(\boldsymbol{\theta}_k)$ as the independent variables and the residuals from the previous regression as the dependent variables.
Starting values? – always 'ad hoc'. Here is one possibility. If the error is ignored we have

$$\frac{1}{y} \approx \frac{1}{\theta_1} + \frac{\theta_2}{\theta_1} \cdot \frac{1}{x},$$

i.e. if $y' = 1/y$ and $x' = 1/x$ then

$$y' = \alpha + \beta x'$$

for $\alpha = 1/\theta_1$ and $\beta = \theta_2/\theta_1$. So regress the y'_i on the x'_i to get lse's $\hat{\alpha}$ and $\hat{\beta}$, then use starting values $\theta_{0,1} = 1/\hat{\alpha}$ and $\theta_{0,2} = \hat{\beta}/\hat{\alpha}$.

37. Maximum likelihood III: Example, computations

- **Example**. Suppose $\{X_1, ..., X_n\}$ is a sample from a gamma(α, β) population, with density

$$p(x; \boldsymbol{\theta}) = \frac{\left(\frac{x}{\alpha}\right)^{\beta-1} e^{-\frac{x}{\alpha}}}{\alpha \Gamma(\beta)}, \ 0 < x < \infty.$$

If $\boldsymbol{\theta} = (\alpha, \beta) = (2, m/2)$ then this is the χ^2_m density. If $\alpha = \lambda^{-1}$, $\beta = m$ it is the 'Erlang' density - the density of the sum of m i.i.d. $\mathbb{E}(\lambda)$ r.v.s.

The log-likelihood is

$$\begin{aligned} l(\boldsymbol{\theta}) &= \sum_{i=1}^{n} \log p(x_i; \boldsymbol{\theta}) \\ &= \sum_{i=1}^{n} \begin{bmatrix} (\beta - 1)(\log x_i - \log \alpha) \\ -\frac{x_i}{\alpha} - \log \alpha - \log \Gamma(\beta) \end{bmatrix} \\ &= n \begin{bmatrix} (\beta - 1) \operatorname{aver}(\log x) - \beta \log \alpha \\ -\frac{\bar{x}}{\alpha} - \log \Gamma(\beta) \end{bmatrix}, \end{aligned}$$

with gradient

$$\dot{l}(\boldsymbol{\theta}) = n \begin{pmatrix} -\frac{\beta}{\alpha} + \frac{\bar{x}}{\alpha^2} \\ \operatorname{aver}(\log x) - \log \alpha - \psi(\beta) \end{pmatrix},$$

where $\psi(\beta) = (d/d\beta) \log \Gamma(\beta)$ $(= E[\log(X/\alpha)])$ is the 'digamma' function.

- The Newton-Raphson method for solving the likelihood equations is (recall (35.1))

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \ddot{l}^{-1}(\boldsymbol{\theta}_k)\dot{l}(\boldsymbol{\theta}_k),$$

where the Hessian is

$$\ddot{l}(\boldsymbol{\theta}) = n \begin{pmatrix} \frac{\beta}{\alpha^2} - 2\frac{\bar{x}}{\alpha^3} & -\frac{1}{\alpha} \\ -\frac{1}{\alpha} & -\psi'(\beta) \end{pmatrix}.$$

- Starting values for iterative solution to the likelihood equations. Straightforward calculations yield

$$\mu_X = \alpha\beta, \ \sigma_X^2 = \alpha^2\beta.$$

The 'method of moments' estimates $\boldsymbol{\theta}_0 = (\alpha_0, \beta_0)$ are now obtained by equating the estimates $\bar{X}$ and S^2 of μ_X and σ_X^2 to their expectations and solving for the parameters:

$$\alpha_0 = \frac{S^2}{\bar{X}}, \ \beta_0 = \frac{\bar{X}^2}{S^2} \ \left(= \frac{\bar{X}}{\alpha_0}\right).$$

- **Method of moments**: Define population moments $\mu_k = E\left[X^k\right]$ and estimates $\hat{\mu}_k = n^{-1}\sum_{i=1}^{n} x_i^k$. By the WLLN these are consistent:

$$\hat{\mu}_k \xrightarrow{pr} \mu_k \text{ as } n \to \infty.$$

Then to estimate continuous functions

$$\boldsymbol{\theta} = \mathrm{g}\left(\mu_1, ..., \mu_p\right)$$

of the population moments, the method of moments estimate

$$\hat{\boldsymbol{\theta}} = \mathrm{g}\left(\hat{\mu}_1, ..., \hat{\mu}_p\right)$$

is also consistent. The proof is the same as in the univariate case:

$$\begin{aligned} P\left(\left\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right\| \geq \varepsilon\right) &= P\left(\left\|\mathrm{g}\left(\hat{\boldsymbol{\mu}}\right) - \mathrm{g}\left(\boldsymbol{\mu}\right)\right\| \geq \varepsilon\right) \\ &\leq P\left(\left\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\right\| \geq \delta\right) \\ &\to 0; \end{aligned}$$

here $\delta > 0$ is such that

$$\left\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\right\| < \delta \Rightarrow \left\|\mathrm{g}\left(\hat{\boldsymbol{\mu}}\right) - \mathrm{g}\left(\boldsymbol{\mu}\right)\right\| < \varepsilon,$$

and its existence is guaranteed by the continuity of g.

- The limit of the NR-process is the MLE $\hat{\boldsymbol{\theta}}$, and

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right) \xrightarrow{L} N(0, \mathbf{I}^{-1}(\boldsymbol{\theta})).$$

The information matrix is (how?)

$$\mathbf{I}(\boldsymbol{\theta}) = \begin{pmatrix} \frac{\beta}{\alpha^2} & \frac{1}{\alpha} \\ \frac{1}{\alpha} & \psi'(\beta) \end{pmatrix},$$

with

$$\mathbf{I}^{-1}(\boldsymbol{\theta}) = \left(\frac{1}{\beta\psi'(\beta)-1}\right)\begin{pmatrix} \alpha^2\psi'(\beta) & -\alpha \\ -\alpha & \beta \end{pmatrix}.$$

Then, e.g., the approximation to the distribution of $\hat{\alpha}$ is

$$\hat{\alpha} \overset{d}{\approx} N\left(\alpha, \frac{I^{11}(\alpha,\beta)}{n} = \frac{\alpha^2\psi'(\beta)}{n\left(\beta\psi'(\beta)-1\right)}\right).$$

To obtain confidence intervals, we estimate the parameters in the variance, obtaining

$$\frac{\hat{\alpha}-\alpha}{\sqrt{I^{11}\left(\hat{\alpha},\hat{\beta}\right)/n}} \overset{d}{\approx} N(0,1).$$

[One can calculate that $\psi'(\beta) = \text{var}[\log(X/\alpha)] = \text{var}[\log X]$; this can be consistently estimated (here, and in $\ddot{l}(\boldsymbol{\theta})$) by the sample variance of $\{\log X_i\}_{i=1}^n$.]